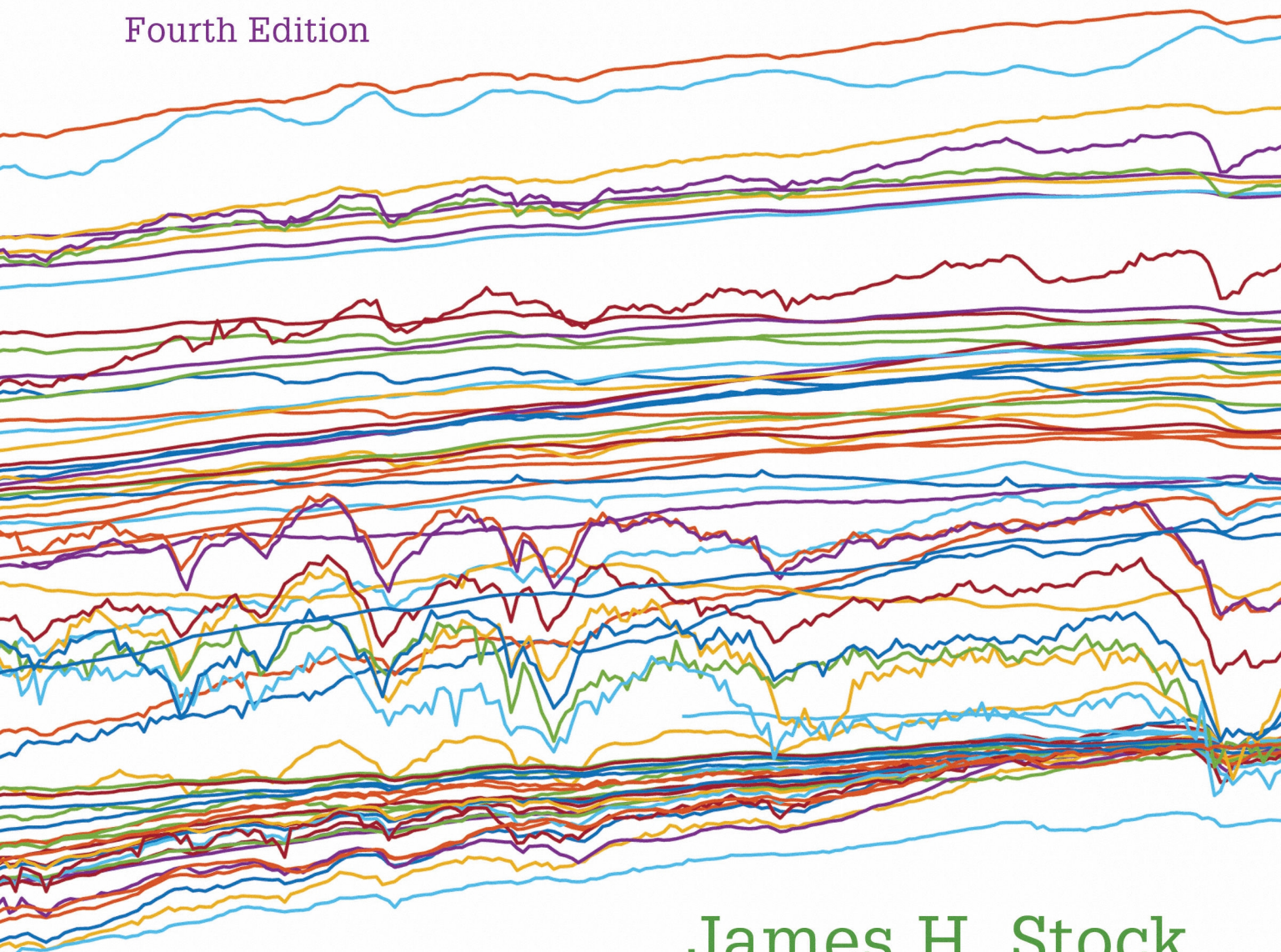# Introduction to Econometrics

Fourth Edition

James H. Stock

Mark W. Watson

Prepare, Apply, Assess and Develop Employability Skills with MyLab Economics

83%
of students said it helped them earn higher grades on homework, exams, or the course

*Source: 2016 Student Survey, n 10,263

MyLab™ **Economics** is an online homework, tutorial, and assessment program constructed to work with this text to engage students and improve results. It was designed to help students develop and assess the skills and applicable knowledge that they will need to succeed in their courses and their future careers.

See what students had to say about MyLab Economics:

"Usually when I do homework myself and don't get it I am stuck, but [MyLab Economics] provided the tools necessary to help me learn how to work my way through the trickiest problems."

— Zainul Lughmani, Binghamton University

Digital Interactives

Economic principles are not static ideas, and learning them shouldn't be either! Digital Interactives are dynamic and engaging assessment activities that promote **critical thinking** and **application** of key economic principles.

Question 2.1: Calculate Nominal GDP for 2014 (Closed Economy)
Suppose there are only three goods in this closed economy that does not trade with other countries: energy drinks (Consumption), computers (Investment), and public highways (Government). Over a two-year period, the amount spent on these items have changed. Based on information in the table below, determine nominal and real GDP for this simple economy.

Economy Final Goods and Services

Nominal GDP 2014: -

Nominal GDP 2015: -

Pearson

## Question Help

MyLab Economics homework and practice questions are correlated to the textbook, and many generate algorithmically to give students unlimited opportunity for mastery of concepts. If students get stuck, Learning Aids including Help Me Solve This and eText Pages walk them through the problem and identify helpful information in the text, giving them assistance when they need it most.
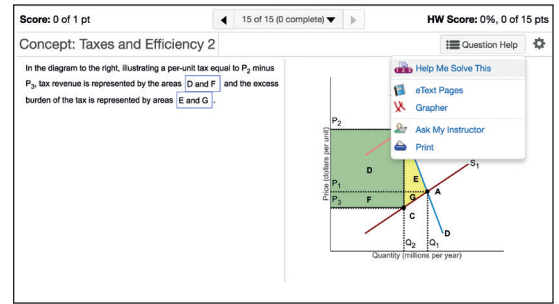
"[MyLab Economics] provides ample practice and explanation of the concepts at hand."
— Heather Burkett, University of Nebraska at Omaha

88% eText
90% Study Plan
92% Dynamic Study Modules
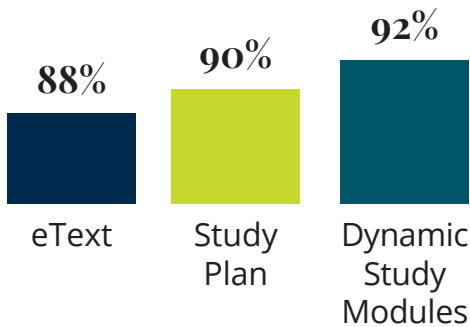
% of students who found learning tool helpful

Dynamic Study Modules help students study chapter topics effectively on their own by continuously assessing their **knowledge application** and performance in real time. These are available as prebuilt Prepare assignments, and are accessible on smartphones, tablets, and computers.

Pearson eText enhances student learning—both in and outside the classroom. Worked examples, videos, and interactive tutorials bring learning to life, while algorithmic practice and self-assessment opportunities test students' understanding of the material. Accessible anytime, anywhere via MyLab or the app.

The **MyLab Gradebook** offers an easy way for students and instructors to view course performance. Item Analysis allows instructors to quickly see trends by analyzing details like the number of students who answered correctly/incorrectly, time on task, and median time spend on a question by question basis. And because it's correlated with the AACSB Standards, instructors can track students' progress toward outcomes that the organization has deemed important in preparing students to be **leaders.**

87%

of students would tell their instructor to keep using MyLab Economics

For additional details visit: www.pearson.com/mylab/economics

Introduction to Econometrics

The Pearson Series in Economics

Abel/Bernanke/Croushore
Macroeconomics*

Acemoglu/Laibson/List
Economics*

Bade/Parkin
Foundations of Economics*

Berck/Helfand
The Economics of the Environment

Bierman/Fernandez
Game Theory with Economic Applications

Blair/Rush
The Economics of Managerial Decisions*

Blanchard
Macroeconomics*

Boyer
Principles of Transportation Economics

Branson
Macroeconomic Theory and Policy

Bruce
Public Finance and the American Economy

Carlton/Perloff
Modern Industrial Organization

Case/Fair/Oster
Principles of Economics*

Chapman
Environmental Economics: Theory, Application, and Policy

Daniels/VanHoose
International Monetary & Financial Economics

Downs
An Economic Theory of Democracy

Farnham
Economics for Managers

Froyen
Macroeconomics: Theories and Policies

Fusfeld
The Age of the Economist

Gerber
International Economics*

Gordon
Macroeconomics*

Greene
Econometric Analysis

Gregory/Stuart
Russian and Soviet Economic Performance and Structure

Hartwick/Olewiler
The Economics of Natural Resource Use

Heilbroner/Milberg
The Making of the Economic Society

Heyne/Boettke/Prychitko
The Economic Way of Thinking

Hubbard/O'Brien
Economics*

InEcon

Money, Banking, and the Financial System*

Hubbard/O'Brien/Rafferty
Macroeconomics*

Hughes/Cain
American Economic History

Husted/Melvin
International Economics

Jehle/Reny
Advanced Microeconomic Theory

Keat/Young/Erfle
Managerial Economics

Klein
Mathematical Methods for Economics

Krugman/Obstfeld/Melitz
International Economics: Theory & Policy*

Laidler
The Demand for Money

Lynn
Economic Development: Theory and Practice for a Divided World

Miller
Economics Today*

Miller/Benjamin
The Economics of Macro Issues

Miller/Benjamin/North
The Economics of Public Issues

Mishkin
The Economics of Money, Banking, and Financial Markets*

The Economics of Money, Banking, and Financial Markets, Business School Edition*

Macroeconomics: Policy and Practice*

Murray
Econometrics: A Modern Introduction

O'Sullivan/Sheffrin/Perez
Economics: Principles, Applications and Tools*

Parkin
Economics*

Perloff
Microeconomics*

Microeconomics: Theory and Applications with Calculus*

Perloff/Brander
Managerial Economics and Strategy*

Pindyck/Rubinfeld
Microeconomics*

Riddell/Shackelford/Stamos/Schneider
Economics: A Tool for Critically Understanding Society

Roberts
The Choice: A Fable of Free Trade and Protection

Scherer
Industry Structure, Strategy, and Public Policy

Schiller
The Economics of Poverty and Discrimination

Sherman
Market Regulation

Stock/Watson
Introduction to Econometrics

Studenmund
Using Econometrics: A Practical Guide

Todaro/Smith
Economic Development

Walters/Walters/Appel/Callahan/Centanni/Maex/O'Neill
Econversations: Today's Students Discuss Today's Issues

Williamson
Macroeconomics

Introduction to Econometrics

FOURTH EDITION

James H. Stock
Harvard University

Mark W. Watson
Princeton University

About the cover: The cover shows a time series plot of 72 indicators of real economic activity in the United States beginning in 1959. The plot shows the growth of these variables since 1959 and their (roughly) synchronized downturns associated with recessions. These series are a subset of the 131-variable dataset used in Chapter 17 to construct dynamic factor model forecasts of future growth in real GDP.

Brief Contents

This page intentionally left blank

Contents

Key Concepts

PART THREE   Further Topics in Regression Analysis

PART FOUR   Regression Analysis of Economic Time Series Data

PART FIVE   Regression Analysis of Economic Time Series Data

This page intentionally left blank

General Interest Boxes

This page intentionally left blank

Preface

Econometrics can be a fun course for both teacher and student. The real world of economics, business, and government is a complicated and messy place, full of competing ideas and questions that demand answers. Is it more effective to tackle drunk driving by passing tough laws or by increasing the tax on alcohol? Can you make money in the stock market by buying when prices are historically low, relative to earnings, or should you just sit tight, as the random walk theory of stock prices suggests? Can we improve elementary education by reducing class sizes, or should we simply have our children listen to Mozart for 10 minutes a day? Econometrics helps us sort out sound ideas from crazy ones and find quantitative answers to important quantitative questions. Econometrics opens a window on our complicated world that lets us see the relationships on which people, businesses, and governments base their decisions.

Introduction to Econometrics is designed for a first course in undergraduate econometrics. It is our experience that to make econometrics relevant in an introductory course, interesting applications must motivate the theory and the theory must match the applications. This simple principle represents a significant departure from the older generation of econometrics books, in which theoretical models and assumptions do not match the applications. It is no wonder that some students question the relevance of econometrics after they spend much of their time learning assumptions that they subsequently realize are unrealistic so that they must then learn "solutions" to "problems" that arise when the applications do not match the assumptions. We believe that it is far better to motivate the need for tools with a concrete application and then to provide a few simple assumptions that match the application. Because the methods are immediately relevant to the applications, this approach can make econometrics come alive.

To improve student results, we recommend pairing the text content with MyLab Economics, which is the teaching and learning platform that empowers you to reach every student. By combining trusted author content with digital tools and a flexible platform, MyLab personalizes the learning experience and will help your students learn and retain key course concepts while developing skills that future employers are seeking in their candidates. MyLab Economics helps you teach your course, your way. Learn more at www.pearson.com/mylab/economics.

New To This Edition

- New chapter on "Big Data" and machine learning
- Forecasting in time series data with large data sets

- Dynamic factor models

- Parallel treatment of prediction and causal inference using regression

- Now covers realized volatility as well as autoregressive conditional heteroskedasticity

- Updated discussion of weak instruments

Very large data sets are increasingly being used in economics and related fields. Applications include predicting consumer choices, measuring the quality of hospitals or schools, analyzing nonstandard data such as text data, and macroeconomic forecasting with many variables. The three main additions in this edition incorporate the fundamentals of this growing and exciting area of application.

First, we have a new chapter (Chapter 14) that focuses on big data and machine learning methods. Within economics, many of the applications to date have focused on the so called many-predictor problem, where the number of predictors is large relative to the sample size—perhaps even exceeding the sample size. With many predictors, ordinary least squares (OLS) provides poor predictions, and other methods, such as the LASSO, can have much lower out-of-sample prediction errors. This chapter goes over the concepts of out-of-sample prediction, why OLS performs poorly, and how shrinkage can improve upon OLS. The chapter introduces shrinkage methods and prediction using principal components, shows how to choose tuning parameters by cross-validation, and explains how these methods can be used to analyze nonstandard data such as text data. As usual, this chapter has a running empirical example, in this case, prediction of school-level test scores given school-level characteristics, for California elementary schools.

Second, in Chapter 17 (newly renumbered), we extend the many-predictor focus of Chapter 14 to time series data. Specifically, we show how the dynamic factor model can handle a very large number of time series, and show how to implement the dynamic factor model using principal components analysis. We illustrate the dynamic factor model and its use for forecasting with a 131-variable dataset of U.S. quarterly macroeconomic time series.

Third, we now lay out these two uses of regression—causal inference and prediction—up front, when regression is first introduced in Chapter 4. Regression is a statistical tool that can be used to make causal inferences or to make predictions; the two applications place different demands on how the data are collected. When the data are from a randomized controlled experiment, OLS estimates the causal effect. In observational data, if we are interested in estimating the causal effect, then the econometrician needs to use control variables and/or instruments to produce as-if randomization of the variable of interest. In contrast, for prediction, one is not interested in the causal effect so one does not need as-if random variation; however, the estimation ("training") data set must be drawn from the same population as the observations for which one wishes to make the prediction.

This edition has several smaller changes. For example, we now introduce realized volatility as a complement to the GARCH model when analyzing time series data with volatility clustering. In addition, we now extend the discussion (in a new general interest box) of the historical origins of instrumental variables regression in Chapter 12. This treatment now includes a first-ever reproduction of the original derivation of the IV estimator, which was in a letter from Philip Wright to his son Sewall in the spring of 1926, and a discussion of the first IV regression, an estimate of the elasticity of supply of flaxseed.

Solving Teaching and Learning Challenges

Introduction to Econometrics differs from other texts in three main ways. First, we integrate real-world questions and data into the development of the theory, and we take seriously the substantive findings of the resulting empirical analysis. Second, our choice of topics reflects modern theory and practice. Third, we provide theory and assumptions that match the applications. Our aim is to teach students to become sophisticated consumers of econometrics and to do so at a level of mathematics appropriate for an introductory course.

Real-World Questions and Data

We organize each methodological topic around an important real-world question that demands a specific numerical answer. For example, we teach single-variable regression, multiple regression, and functional form analysis in the context of estimating the effect of school inputs on school outputs. (Do smaller elementary school class sizes produce higher test scores?) We teach panel data methods in the context of analyzing the effect of drunk driving laws on traffic fatalities. We use possible racial discrimination in the market for home loans as the empirical application for teaching regression with a binary dependent variable (logit and probit). We teach instrumental variable estimation in the context of estimating the demand elasticity for cigarettes. Although these examples involve economic reasoning, all can be understood with only a single introductory course in economics, and many can be understood without any previous economics coursework. Thus the instructor can focus on teaching econometrics, not microeconomics or macroeconomics.

We treat all our empirical applications seriously and in a way that shows students how they can learn from data but at the same time be self-critical and aware of the limitations of empirical analyses. Through each application, we teach students to explore alternative specifications and thereby to assess whether their substantive findings are robust. The questions asked in the empirical applications are important, and we provide serious and, we think, credible answers. We encourage students and instructors to disagree, however, and invite them to reanalyze the

data, which are provided on the text's Companion Website (**www.pearsonhighered .com/stock_watson**) and in MyLab Economics.

Throughout the text, we have focused on helping students understand, retain, and apply the essential ideas. *Chapter introductions* provide real-world grounding and motivation, as well as brief road maps highlighting the sequence of the discussion. *Key terms* are boldfaced and defined in context throughout each chapter, and *Key Concept boxes* at regular intervals recap the central ideas. *General interest boxes* provide interesting excursions into related topics and highlight real-world studies that use the methods or concepts being discussed in the text. A *Summary* concluding each chapter serves as a helpful framework for reviewing the main points of coverage.

Available for student practice or instructor assignment in MyLab Economics are *Review the Concepts questions, Exercises,* and *Empirical Exercises* from the text. These questions and exercises are auto-graded, giving students practical hands-on experience with solving problems using the data sets used in the text.

- 100 percent of Review the Concepts questions are available in MyLab.

- Select Exercises and Empirical Exercises are available in MyLab. Many of the Empirical Exercises are algorithmic and based on the data sets used in the text. These exercises require students to use Excel or an econometrics software package to analyze the data and derive results.

- New to the 4th edition are concept exercises that focus on core concepts and economic interpretations. Many are algorithmic and include the Help Me Solve This learning aid.

Contemporary Choice of Topics

The topics we cover reflect the best of contemporary applied econometrics. One can only do so much in an introductory course, so we focus on procedures and tests that are commonly (or increasingly) used in practice. For example:

- *Instrumental variables regression.* We present instrumental variables regression as a general method for handling correlation between the error term and a regressor, which can arise for many reasons, including omitted variables and simultaneous causality. The two assumptions for a valid instrument—exogeneity and relevance—are given equal billing. We follow that presentation with an extended discussion of where instruments come from and with tests of overidentifying restrictions and diagnostics for weak instruments, and we explain what to do if these diagnostics suggest problems.

- *Program evaluation.* Many modern econometric studies analyze either randomized controlled experiments or quasi-experiments, also known as natural experiments. We address these topics, often collectively referred to as program

evaluation, in Chapter 13. We present this research strategy as an alternative approach to the problems of omitted variables, simultaneous causality, and selection, and we assess both the strengths and the weaknesses of studies using experimental or quasi-experimental data.

- *Prediction with "big data."* Chapter 14 takes up the opportunities and challenges posed by large cross-sectional data sets. An increasingly common application in econometrics is making predictions when the number of predictors is very large. This chapter focuses on methods designed to use many predictors in a way that produces accurate and precise out-of-sample predictions. The chapter covers some of the building blocks of machine learning, and the methods can substantially improve upon OLS when the number of predictors is large. In addition, these methods extend to nonstandard data, such as text data.

- *Forecasting.* The chapter on forecasting (Chapter 15) considers univariate (autoregressive) and multivariate forecasts using time series regression, not large simultaneous equation structural models. We focus on simple and reliable tools, such as autoregressions and model selection via an information criterion, that work well in practice. This chapter also features a practically oriented treatment of structural breaks (at known and unknown dates) and pseudo out-of-sample forecasting, all in the context of developing stable and reliable time series forecasting models.

- *Time series regression.* The chapter on causal inference using time series data (Chapter 16) pays careful attention to when different estimation methods, including generalized least squares, will or will not lead to valid causal inferences and when it is advisable to estimate dynamic regressions using OLS with heteroskedasticity- and autocorrelation-consistent standard errors.

Theory That Matches Applications

Although econometric tools are best motivated by empirical applications, students need to learn enough econometric theory to understand the strengths and limitations of those tools. We provide a modern treatment in which the fit between theory and applications is as tight as possible, while keeping the mathematics at a level that requires only algebra.

Modern empirical applications share some common characteristics: The data sets typically have many observations (hundreds or more); regressors are not fixed over repeated samples but rather are collected by random sampling (or some other mechanism that makes them random); the data are not normally distributed; and there is no *a priori* reason to think that the errors are homoskedastic (although often there are reasons to think that they are heteroskedastic).

These observations lead to important differences between the theoretical development in this text and other texts:

- *Large-sample approach.* Because data sets are large, from the outset we use large-sample normal approximations to sampling distributions for hypothesis testing and confidence intervals. In our experience, it takes less time to teach the rudiments of large-sample approximations than to teach the Student t and exact F distributions, degrees-of-freedom corrections, and so forth. This large-sample approach also saves students the frustration of discovering that, because of nonnormal errors, the exact distribution theory they just mastered is irrelevant. Once taught in the context of the sample mean, the large-sample approach to hypothesis testing and confidence intervals carries directly through multiple regression analysis, logit and probit, instrumental variables estimation, and time series methods.

- *Random sampling.* Because regressors are rarely fixed in econometric applications, from the outset we treat data on all variables (dependent and independent) as the result of random sampling. This assumption matches our initial applications to cross-sectional data, it extends readily to panel and time series data, and because of our large-sample approach, it poses no additional conceptual or mathematical difficulties.

- *Heteroskedasticity.* Applied econometricians routinely use heteroskedasticity-robust standard errors to eliminate worries about whether heteroskedasticity is present or not. In this book, we move beyond treating heteroskedasticity as an exception or a "problem" to be "solved"; instead, we allow for heteroskedasticity from the outset and simply use heteroskedasticity-robust standard errors. We present homoskedasticity as a special case that provides a theoretical motivation for OLS.

Skilled Producers, Sophisticated Consumers

We hope that students using this book will become sophisticated consumers of empirical analysis. To do so, they must learn not only how to use the tools of regression analysis but also how to assess the validity of empirical analyses presented to them.

Our approach to teaching how to assess an empirical study is threefold. First, immediately after introducing the main tools of regression analysis, we devote Chapter 9 to the threats to internal and external validity of an empirical study. This chapter discusses data problems and issues of generalizing findings to other settings. It also examines the main threats to regression analysis, including omitted variables, functional form misspecification, errors-in-variables, selection, and simultaneity—and ways to recognize these threats in practice.

Second, we apply these methods for assessing empirical studies to the empirical analysis of the ongoing examples in the book. We do so by considering alternative specifications and by systematically addressing the various threats to validity of the analyses presented in the book.

Third, to become sophisticated consumers, students need firsthand experience as producers. Active learning beats passive learning, and econometrics is an ideal course for active learning. For this reason, the MyLab Economics and text website feature data sets, software, and suggestions for empirical exercises of different scopes.

Approach to Mathematics and Level of Rigor

Our aim is for students to develop a sophisticated understanding of the tools of modern regression analysis, whether the course is taught at a "high" or a "low" level of mathematics. Parts I through IV of the text (which cover the substantive material) are written for students with only precalculus mathematics. Parts I through IV have fewer equations and more applications than many introductory econometrics books and far fewer equations than books aimed at mathematical sections of undergraduate courses. But more equations do not imply a more sophisticated treatment. In our experience, a more mathematical treatment does not lead to a deeper understanding for most students.

That said, different students learn differently, and for mathematically well-prepared students, learning can be enhanced by a more explicit mathematical treatment. The appendices in Parts I-IV therefore provide key calculations that are too involved to be included in the text. In addition, Part V contains an introduction to econometric theory that is appropriate for students with a stronger mathematical background. When the mathematical chapters in Part V are used in conjunction with the material in Parts I through IV (including appendices), this book is suitable for advanced undergraduate or master's level econometrics courses.

Developing Career Skills

For students to succeed in a rapidly changing job market, they should be aware of their career options and how to go about developing a variety of skills. Data analysis is an increasingly marketable skill. This text prepares the students for a range of data analytic applications, including causal inference and prediction. It also introduces the students to the core concepts of prediction using large data sets.

Table of Contents Overview

There are five parts to *Introduction to Econometrics*. This text assumes that the student has had a course in probability and statistics, although we review that material in Part I. We cover the core material of regression analysis in Part II. Parts III, IV, and V present additional topics that build on the core treatment in Part II.

Part I

Chapter 1 introduces econometrics and stresses the importance of providing quantitative answers to quantitative questions. It discusses the concept of causality in statistical studies and surveys the different types of data encountered in econometrics. Material from probability and statistics is reviewed in Chapters 2 and 3, respectively; whether these chapters are taught in a given course or are simply provided as a reference depends on the background of the students.

Part II

Chapter 4 introduces regression with a single regressor and ordinary least squares (OLS) estimation, and Chapter 5 discusses hypothesis tests and confidence intervals in the regression model with a single regressor. In Chapter 6, students learn how they can address omitted variable bias using multiple regression, thereby estimating the effect of one independent variable while holding other independent variables constant. Chapter 7 covers hypothesis tests, including *F*-tests, and confidence intervals in multiple regression. In Chapter 8, the linear regression model is extended to models with nonlinear population regression functions, with a focus on regression functions that are linear in the parameters (so that the parameters can be estimated by OLS). In Chapter 9, students step back and learn how to identify the strengths and limitations of regression studies, seeing in the process how to apply the concepts of internal and external validity.

Part III

Part III presents extensions of regression methods. In Chapter 10, students learn how to use panel data to control for unobserved variables that are constant over time. Chapter 11 covers regression with a binary dependent variable. Chapter 12 shows how instrumental variables regression can be used to address a variety of problems that produce correlation between the error term and the regressor, and examines how one might find and evaluate valid instruments. Chapter 13 introduces students to the analysis of data from experiments and quasi-, or natural, experiments, topics often referred to as "program evaluation." Chapter 14 turns to econometric issues that arise with large data sets, and focuses on prediction when there are very many predictors.

Part IV

Part IV takes up regression with time series data. Chapter 15 focuses on forecasting and introduces various modern tools for analyzing time series regressions, such as tests for stability. Chapter 16 discusses the use of time series data to estimate causal relations. Chapter 17 presents some more advanced tools for time series analysis, including models of volatility clustering and dynamic factor models.

Part V

Part V is an introduction to econometric theory. This part is more than an appendix that fills in mathematical details omitted from the text. Rather, it is a self-contained treatment of the econometric theory of estimation and inference in the linear regression model. Chapter 18 develops the theory of regression analysis for a single regressor; the exposition does not use matrix algebra, although it does demand a higher level of mathematical sophistication than the rest of the text. Chapter 19 presents the multiple regression model, instrumental variables regression, generalized method of moments estimation of the linear model, and principal components analysis, all in matrix form.

Prerequisites Within the Book

Because different instructors like to emphasize different material, we wrote this book with diverse teaching preferences in mind. To the maximum extent possible, the chapters in Parts III, IV, and V are "stand-alone" in the sense that they do not require first teaching all the preceding chapters. The specific prerequisites for each chapter are described in Table I. Although we have found that the sequence of topics adopted in the text works well in our own courses, the chapters are written in a way that allows instructors to present topics in a different order if they so desire.

Sample Courses

This book accommodates several different course structures.

TABLE I	Guide to Prerequisites for Special-Topic Chapters in Parts III, IV, and V								
	Prerequisite parts or chapters								
	Part I	**Part II**		**Part III**		**Part IV**			**Part V**
Chapter	**1–3**	**4–7, 9**	**8**	**10.1, 10.2**	**12.1, 12.2**	**15.1–15.4**	**15.5–15.8**	**16**	**18**
10	X[a]	X[a]	X						
11	X[a]	X[a]	X						
12.1, 12.2	X[a]	X[a]	X						
12.3–12.6	X[a]	X[a]	X	X	X				
13	X[a]	X[a]	X	X	X				
14	X[a]	X[a]	X						
15	X[a]	X[a]	b						
16	X[a]	X[a]	b			X			
17	X[a]	X[a]	b			X	X	X	
18	X	X	X						
19	X	X	X		X				X

This table shows the minimum prerequisites needed to cover the material in a given chapter. For example, estimation of dynamic causal effects with time series data (Chapter 16) first requires Part I (as needed, depending on student preparation, and except as noted in footnote a), Part II (except for Chapter 8; see footnote b), and Sections 15.1 through 15.4.

[a]Chapters 10 through 17 use exclusively large-sample approximations to sampling distributions, so the optional Sections 3.6 (the Student t distribution for testing means) and 5.6 (the Student t distribution for testing regression coefficients) can be skipped.
[b]Chapters 15 through 17 (the time series chapters) can be taught without first teaching Chapter 8 (nonlinear regression functions) if the instructor pauses to explain the use of logarithmic transformations to approximate percentage changes.

Standard Introductory Econometrics

This course introduces econometrics (Chapter 1) and reviews probability and statistics as needed (Chapters 2 and 3). It then moves on to regression with a single regressor, multiple regression, the basics of functional form analysis, and the evaluation of regression studies (all Part II). The course proceeds to cover regression with panel data (Chapter 10), regression with a limited dependent variable (Chapter 11), and instrumental variables regression (Chapter 12), as time permits. The course then

turns to experiments and quasi-experiments in Chapter 13, topics that provide an opportunity to return to the questions of estimating causal effects raised at the beginning of the semester and to recapitulate core regression methods. If there is time, the students can be introduced to big data and machine learning methods at the end (Chapter 14). *Prerequisites: Algebra II and introductory statistics*.

Introductory Econometrics with Time Series and Forecasting Applications

Like a standard introductory course, this course covers all of Part I (as needed) and Part II. Optionally, the course next provides a brief introduction to panel data (Sections 10.1 and 10.2) and takes up instrumental variables regression (Chapter 12, or just Sections 12.1 and 12.2). The course then proceeds to Chapter 14 (prediction in large cross sectional data sets). It then turns to Part IV, covering forecasting (Chapter 15) and estimation of dynamic causal effects (Chapter 16). If time permits, the course can include some advanced topics in time series analysis such as volatility clustering (Section 17.5) and forecasting with many predictors (Section 17.6). *Prerequisites: Algebra II and introductory statistics*.

Applied Time Series Analysis and Forecasting

This book also can be used for a short course on applied time series and forecasting, for which a course on regression analysis is a prerequisite. Some time is spent reviewing the tools of basic regression analysis in Part II, depending on student preparation. The course then moves directly to time series forecasting (Chapter 15), estimation of dynamic causal effects (Chapter 16), and advanced topics in time series analysis (Chapter 17), including vector autoregressions. If there is time, the course can cover prediction using large data sets (Chapter 14 and Section 17.6), An important component of this course is hands-on forecasting exercises, available as the end-of-chapter Empirical Exercises for Chapters 15 and 17. *Prerequisites: Algebra II and basic introductory econometrics or the equivalent*.

Introduction to Econometric Theory

This book is also suitable for an advanced undergraduate course in which the students have a strong mathematical preparation or for a master's level course in econometrics. The course briefly reviews the theory of statistics and probability as necessary (Part I). The course introduces regression analysis using the nonmathematical, applications-based treatment of Part II. This introduction is followed by the theoretical development in Chapters 18 and 19 (through Section 19.5). The course then takes up regression with a limited dependent variable (Chapter 11) and maximum likelihood estimation (Appendix 11.2). Next, the course optionally turns to instrumental variables regression and generalized method of moments (Chapter 12 and Section 19.7), time series methods (Chapter 15), the estimation of

causal effects using time series data and generalized least squares (Chapter 16 and Section 19.6), and/or to machine learning methods (Chapter 14 and Appendix 19.7). *Prerequisites: Calculus and introductory statistics. Chapter 18 assumes previous exposure to matrix algebra.*

Instructor Teaching Resources

This program comes with the following teaching resources:

Supplements available to instructors at www.pearsonhighered.com	Features of the Supplement
Solutions Manual	Solutions to the end-of-chapter content.
Test Bank Authored by Manfred Keil, Claremont McKenna College	1,000 multiple-choice questions, essays and longer questions, and mathematical and graphical problems with these annotations: • Type (Multiple-choice, essay, graphical)
Computerized TestGen	TestGen allows instructors to: • Customize, save, and generate classroom tests • Edit, add, or delete questions from the Test Item Files • Analyze test results • Organize a database of tests and student results.
PowerPoints	Slides include all the graphs, tables, and equations in the text. PowerPoints meet accessibility standards for students with disabilities. Features include, but not limited to: • Keyboard and Screen Reader access • Alternative text for images • High color contrast between background and foreground colors
Companion Website	The Companion Website provides a wide range of additional resources for students and faculty. These resources include more and more in depth empirical exercises, data sets for the empirical exercises, replication files for empirical results reported in the text, and EViews tutorials.

Acknowledgments

A great many people contributed to the first edition of this book. Our biggest debts of gratitude are to our colleagues at Harvard and Princeton who used early drafts of this book in their classrooms. At Harvard's Kennedy School of Government, Suzanne Cooper provided invaluable suggestions and detailed comments on multiple drafts. As a coteacher with one of the authors (Stock), she also helped vet much of the material in this book while it was being developed for a required course for master's students at the Kennedy School. We are also indebted to two other Kennedy School colleagues at the time, Alberto Abadie and Sue Dynarski, for their patient explanations of quasi-experiments and the field of program evaluation and for their detailed comments on early drafts of the text. At Princeton, Eli Tamer taught from an early draft and also provided helpful comments on the penultimate draft of the book.

We also owe much to many of our friends and colleagues in econometrics who spent time talking with us about the substance of this book and who collectively made so many helpful suggestions. Bruce Hansen (University of Wisconsin–Madison) and Bo Honore (Princeton) provided helpful feedback on very early outlines and preliminary versions of the core material in Part II. Joshua Angrist (MIT) and Guido Imbens (University of California, Berkeley) provided thoughtful suggestions about our treatment of materials on program evaluation. Our presentation of the material on time series has benefited from discussions with Yacine Ait-Sahalia (Princeton), Graham Elliott (University of California, San Diego), Andrew Harvey (Cambridge University), and Christopher Sims (Princeton). Finally, many people made helpful suggestions on parts of the manuscript close to their area of expertise: Don Andrews (Yale), John Bound (University of Michigan), Gregory Chow (Princeton), Thomas Downes (Tufts), David Drukker (StataCorp.), Jean Baldwin Grossman (Princeton), Eric Hanushek (Hoover Institution), James Heckman (University of Chicago), Han Hong (Princeton), Caroline Hoxby (Harvard), Alan Krueger (Princeton), Steven Levitt (University of Chicago), Richard Light (Harvard), David Neumark (Michigan State University), Joseph Newhouse (Harvard), Pierre Perron (Boston University), Kenneth Warner (University of Michigan), and Richard Zeckhauser (Harvard).

Many people were very generous in providing us with data. The California test score data were constructed with the assistance of Les Axelrod of the Standards and Assessments Division, California Department of Education. We are grateful to Charlie DePascale, Student Assessment Services, Massachusetts Department of Education, for his help with aspects of the Massachusetts test score data set. Christopher Ruhm (University of North Carolina, Greensboro) graciously provided us with his data set on drunk driving laws and traffic fatalities. The research department at the Federal Reserve Bank of Boston deserves thanks for putting together its data on racial discrimination in mortgage lending; we particularly thank Geoffrey Tootell for providing us with the updated version of the data set we use in Chapter 9 and Lynn Browne for explaining its policy context. We thank Jonathan Gruber (MIT) for sharing his data on cigarette sales, which we analyze in Chapter 12, and

Above all, we are indebted to our families for their endurance throughout this project. Writing this book took a long time, and for them, the project must have seemed endless. They, more than anyone else, bore the burden of this commitment, and for their help and support we are deeply grateful.

This page intentionally left blank

Economic Questions and Data

Ask a half dozen econometricians what econometrics is, and you could get a half dozen different answers. One might tell you that econometrics is the science of testing economic theories. A second might tell you that econometrics is the set of tools used for forecasting future values of economic variables, such as a firm's sales, the overall growth of the economy, or stock prices. Another might say that econometrics is the process of fitting mathematical economic models to real-world data. A fourth might tell you that it is the science and art of using historical data to make numerical, or quantitative, policy recommendations in government and business.

In fact, all these answers are right. At a broad level, econometrics is the science and art of using economic theory and statistical techniques to analyze economic data. Econometric methods are used in many branches of economics, including finance, labor economics, macroeconomics, microeconomics, marketing, and economic policy. Econometric methods are also commonly used in other social sciences, including political science and sociology.

This text introduces you to the core set of methods used by econometricians. We will use these methods to answer a variety of specific, quantitative questions from the worlds of business and government policy. This chapter poses four of those questions and discusses, in general terms, the econometric approach to answering them. The chapter concludes with a survey of the main types of data available to econometricians for answering these and other quantitative economic questions.

1.1 Economic Questions We Examine

Many decisions in economics, business, and government hinge on understanding relationships among variables in the world around us. These decisions require quantitative answers to quantitative questions.

This text examines several quantitative questions taken from current issues in economics. Four of these questions concern education policy, racial bias in mortgage lending, cigarette consumption, and macroeconomic forecasting.

Question #1: Does Reducing Class Size Improve Elementary School Education?

Proposals for reform of the U.S. public education system generate heated debate. Many of the proposals concern the youngest students, those in elementary schools. Elementary school education has various objectives, such as developing social skills,

but for many parents and educators, the most important objective is basic academic learning: reading, writing, and basic mathematics. One prominent proposal for improving basic learning is to reduce class sizes at elementary schools. With fewer students in the classroom, the argument goes, each student gets more of the teacher's attention, there are fewer class disruptions, learning is enhanced, and grades improve.

But what, precisely, is the effect on elementary school education of reducing class size? Reducing class size costs money: It requires hiring more teachers and, if the school is already at capacity, building more classrooms. A decision maker contemplating hiring more teachers must weigh these costs against the benefits. To weigh costs and benefits, however, the decision maker must have a precise quantitative understanding of the likely benefits. Is the beneficial effect on basic learning of smaller classes large or small? Is it possible that smaller class size actually has no effect on basic learning?

Although common sense and everyday experience may suggest that more learning occurs when there are fewer students, common sense cannot provide a quantitative answer to the question of what exactly is the effect on basic learning of reducing class size. To provide such an answer, we must examine empirical evidence—that is, evidence based on data—relating class size to basic learning in elementary schools.

In this text, we examine the relationship between class size and basic learning, using data gathered from 420 California school districts in 1999. In the California data, students in districts with small class sizes tend to perform better on standardized tests than students in districts with larger classes. While this fact is consistent with the idea that smaller classes produce better test scores, it might simply reflect many other advantages that students in districts with small classes have over their counterparts in districts with large classes. For example, districts with small class sizes tend to have wealthier residents than districts with large classes, so students in small-class districts could have more opportunities for learning outside the classroom. It could be these extra learning opportunities that lead to higher test scores, not smaller class sizes. In Part II, we use multiple regression analysis to isolate the effect of changes in class size from changes in other factors, such as the economic background of the students.

Question #2: Is There Racial Discrimination in the Market for Home Loans?

Most people buy their homes with the help of a mortgage, a large loan secured by the value of the home. By law, U.S. lending institutions cannot take race into account when deciding to grant or deny a request for a mortgage: Applicants who are identical in all ways except their race should be equally likely to have their mortgage applications approved. In theory, then, there should be no racial bias in mortgage lending.

In contrast to this theoretical conclusion, researchers at the Federal Reserve Bank of Boston found (using data from the early 1990s) that 28% of black applicants are

denied mortgages, while only 9% of white applicants are denied. Do these data indicate that, in practice, there is racial bias in mortgage lending? If so, how large is it?

The fact that more black than white applicants are denied in the Boston Fed data does not by itself provide evidence of discrimination by mortgage lenders because the black and white applicants differ in many ways other than their race. Before concluding that there is bias in the mortgage market, these data must be examined more closely to see if there is a difference in the probability of being denied for *otherwise identical* applicants and, if so, whether this difference is large or small. To do so, in Chapter 11 we introduce econometric methods that make it possible to quantify the effect of race on the chance of obtaining a mortgage, *holding constant* other applicant characteristics, notably their ability to repay the loan.

Question #3: How Much Do Cigarette Taxes Reduce Smoking?

Cigarette smoking is a major public health concern worldwide. Many of the costs of smoking, such as the medical expenses of caring for those made sick by smoking and the less quantifiable costs to nonsmokers who prefer not to breathe secondhand cigarette smoke, are borne by other members of society. Because these costs are borne by people other than the smoker, there is a role for government intervention in reducing cigarette consumption. One of the most flexible tools for cutting consumption is to increase taxes on cigarettes.

Basic economics says that if cigarette prices go up, consumption will go down. But by how much? If the sales price goes up by 1%, by what percentage will the quantity of cigarettes sold decrease? The percentage change in the quantity demanded resulting from a 1% increase in price is the *price elasticity of demand*. If we want to reduce smoking by a certain amount, say, 20%, by raising taxes, then we need to know the price elasticity of demand to calculate the price increase necessary to achieve this reduction in consumption. But what is the price elasticity of demand for cigarettes?

Although economic theory provides us with the concepts that help us answer this question, it does not tell us the numerical value of the price elasticity of demand. To learn the elasticity, we must examine empirical evidence about the behavior of smokers and potential smokers; in other words, we need to analyze data on cigarette consumption and prices.

The data we examine are cigarette sales, prices, taxes, and personal income for U.S. states in the 1980s and 1990s. In these data, states with low taxes, and thus low cigarette prices, have high smoking rates, and states with high prices have low smoking rates. However, the analysis of these data is complicated because causality runs both ways: Low taxes lead to high demand, but if there are many smokers in the state, then local politicians might try to keep cigarette taxes low to satisfy their smoking constituents. In Chapter 12, we study methods for handling this "simultaneous causality" and use those methods to estimate the price elasticity of cigarette demand.

Question #4: By How Much Will U.S. GDP Grow Next Year?

It seems that people always want a sneak preview of the future. What will sales be next year at a firm that is considering investing in new equipment? Will the stock market go up next month, and, if it does, by how much? Will city tax receipts next year cover planned expenditures on city services? Will your microeconomics exam next week focus on externalities or monopolies? Will Saturday be a nice day to go to the beach?

One aspect of the future in which macroeconomists are particularly interested is the growth of real economic activity, as measured by real gross domestic product (GDP), during the next year. A management consulting firm might advise a manufacturing client to expand its capacity based on an upbeat forecast of economic growth. Economists at the Federal Reserve Board in Washington, D.C., are mandated to set policy to keep real GDP near its potential in order to maximize employment. If they forecast anemic GDP growth over the next year, they might expand liquidity in the economy by reducing interest rates or other measures, in an attempt to boost economic activity.

Professional economists who rely on numerical forecasts use econometric models to make those forecasts. A forecaster's job is to predict the future by using the past, and econometricians do this by using economic theory and statistical techniques to quantify relationships in historical data.

The data we use to forecast the growth rate of GDP include past values of GDP and the so-called term spread in the United States. The *term spread* is the difference between long-term and short-term interest rates. It measures, among other things, whether investors expect short-term interest rates to rise or fall in the future. The term spread is usually positive, but it tends to fall sharply before the onset of a recession. One of the GDP growth rate forecasts we develop and evaluate in Chapter 15 is based on the term spread.

Quantitative Questions, Quantitative Answers

Each of these four questions requires a numerical answer. Economic theory provides clues about that answer—for example, cigarette consumption ought to go down when the price goes up—but the actual value of the number must be learned empirically, that is, by analyzing data. Because we use data to answer quantitative questions, our answers always have some uncertainty: A different set of data would produce a different numerical answer. Therefore, the conceptual framework for the analysis needs to provide both a numerical answer to the question and a measure of how precise the answer is.

The conceptual framework used in this text is the multiple regression model, the mainstay of econometrics. This model, introduced in Part II, provides a mathematical way to quantify how a change in one variable affects another variable, holding other things constant. For example, what effect does a change in class size have on test scores, *holding constant* or *controlling for* student characteristics (such as family

income) that a school district administrator cannot control? What effect does your race have on your chances of having a mortgage application granted, *holding constant* other factors such as your ability to repay the loan? What effect does a 1% increase in the price of cigarettes have on cigarette consumption, *holding constant* the income of smokers and potential smokers? The multiple regression model and its extensions provide a framework for answering these questions using data and for quantifying the uncertainty associated with those answers.

1.2   Causal Effects and Idealized Experiments

Like many other questions encountered in econometrics, the first three questions in Section 1.1 concern causal relationships among variables. In common usage, an action is said to cause an outcome if the outcome is the direct result, or consequence, of that action. Touching a hot stove causes you to get burned, drinking water causes you to be less thirsty, putting air in your tires causes them to inflate, putting fertilizer on your tomato plants causes them to produce more tomatoes. Causality means that a specific action (applying fertilizer) leads to a specific, measurable consequence (more tomatoes).

Estimation of Causal Effects

How best might we measure the causal effect on tomato yield (measured in kilograms) of applying a certain amount of fertilizer, say, 100 grams of fertilizer per square meter?

One way to measure this causal effect is to conduct an experiment. In that experiment, a horticultural researcher plants many plots of tomatoes. Each plot is tended identically, with one exception: Some plots get 100 grams of fertilizer per square meter, while the rest get none. Whether or not a plot is fertilized is determined randomly by a computer, ensuring that any other differences between the plots are unrelated to whether they receive fertilizer. At the end of the growing season, the horticulturalist weighs the harvest from each plot. The difference between the average yield per square meter of the treated and untreated plots is the effect on tomato production of the fertilizer treatment.

This is an example of a **randomized controlled experiment**. It is controlled in the sense that there are both a **control group** that receives no treatment (no fertilizer) and a **treatment group** that receives the treatment (100 g/m^2 of fertilizer). It is randomized in the sense that the treatment is assigned randomly. This random assignment eliminates the possibility of a systematic relationship between, for example, how sunny the plot is and whether it receives fertilizer so that the only systematic difference between the treatment and control groups is the treatment. If this experiment is properly implemented on a large enough scale, then it will yield an estimate of the causal effect on the outcome of interest (tomato production) of the treatment (applying 100 g/m^2 of fertilizer).

In this text, the **causal effect** is defined to be the effect on an outcome of a given action or treatment, as measured in an ideal randomized controlled experiment. In such an experiment, the only systematic reason for differences in outcomes between the treatment and control groups is the treatment itself.

It is possible to imagine an ideal randomized controlled experiment to answer each of the first three questions in Section 1.1. For example, to study class size, one can imagine randomly assigning "treatments" of different class sizes to different groups of students. If the experiment is designed and executed so that the only systematic difference between the groups of students is their class size, then in theory this experiment would estimate the effect on test scores of reducing class size, holding all else constant.

Experiments are used increasingly widely in econometrics. In many applications, however, they are not an option because they are unethical, impossible to execute satisfactorily, too time-consuming, or prohibitively expensive. Even with non-experimental data, the concept of an ideal randomized controlled experiment is important because it provides a definition of a causal effect.

Prediction, Forecasting, and Causality

Although the first three questions in Section 1.1, concern causal effects, the fourth—forecasting the growth rate of GDP—does not.

Forecasting is a special case of what statisticians and econometricians call **prediction**, which is using information on some variables to make a statement about the value of another variable. A **forecast** is a prediction about the value of a variable in the future, like GDP growth next year.

You do not need to know a causal relationship to make a good prediction. A good way to "predict" whether it is raining is to observe whether pedestrians are using umbrellas, but the act of using an umbrella does not cause it to rain.

When one has a small number of predictors and the data do not evolve over time, the multiple regression methods of Part II can provide reliable predictions. Predictions can often be improved, however, if there is a large number of candidate predictors. Methods for using many predictors are covered in Chapter 14.

Forecasts—that is, predictions about the future—use data on variables that evolve over time, which introduces new challenges and opportunities. As we will see in Chapter 15, multiple regression analysis allows us to quantify historical relationships, to check whether those relationships have been stable over time, to make quantitative forecasts about the future, and to assess the accuracy of those forecasts.

1.3   Data: Sources and Types

In econometrics, data come from one of two sources: experiments or nonexperimental observations of the world. This text examines both experimental and nonexperimental data sets.

Experimental versus Observational Data

Experimental data come from experiments designed to evaluate a treatment or policy or to investigate a causal effect. For example, the state of Tennessee financed a large randomized controlled experiment examining class size in the 1980s. In that experiment, which we examine in Chapter 13, thousands of students were randomly assigned to classes of different sizes for several years and were given standardized tests annually.

The Tennessee class size experiment cost millions of dollars and required the ongoing cooperation of many administrators, parents, and teachers over several years. Because real-world experiments with human subjects are difficult to administer and to control, they have flaws relative to ideal randomized controlled experiments. Moreover, in some circumstances, experiments are not only expensive and difficult to administer but also unethical. (Would it be ethical to offer randomly selected teenagers inexpensive cigarettes to see how many they buy?) Because of these financial, practical, and ethical problems, experiments in economics are relatively rare. Instead, most economic data are obtained by observing real-world behavior.

Data obtained by observing actual behavior outside an experimental setting are called **observational data**. Observational data are collected using surveys, such as telephone surveys of consumers, and administrative records, such as historical records on mortgage applications maintained by lending institutions.

Observational data pose major challenges to econometric attempts to estimate causal effects, and the tools of econometrics are designed to tackle these challenges. In the real world, levels of "treatment" (the amount of fertilizer in the tomato example, the student–teacher ratio in the class size example) are not assigned at random, so it is difficult to sort out the effect of the "treatment" from other relevant factors. Much of econometrics, and much of this text, is devoted to methods for meeting the challenges encountered when real-world data are used to estimate causal effects.

Whether the data are experimental or observational, data sets come in three main types: cross-sectional data, time series data, and panel data. In this text, you will encounter all three types.

Cross-Sectional Data

Data on different entities—workers, consumers, firms, governmental units, and so forth—for a single time period are called **cross-sectional data**. For example, the data on test scores in California school districts are cross sectional. Those data are for 420 entities (school districts) for a single time period (1999). In general, the number of entities on which we have observations is denoted n; so, for example, in the California data set, $n = 420$.

The California test score data set contains measurements of several different variables for each district. Some of these data are tabulated in Table 1.1. Each row lists data for a different district. For example, the average test score for the first

TABLE 1.1	Selected Observations on Test Scores and Other Variables for California School Districts in 1999			
Observation (District) Number	**District Average Test Score (fifth grade)**	**Student–Teacher Ratio**	**Expenditure per Pupil ($)**	**Percentage of Students Learning English**
1	690.8	17.89	$6385	0.0%
2	661.2	21.52	5099	4.6
3	643.6	18.70	5502	30.0
4	647.7	17.36	7102	0.0
5	640.8	18.67	5236	13.9
⋮	⋮	⋮	⋮	⋮
418	645.0	21.89	4403	24.3
419	672.2	20.20	4776	3.0
420	655.8	19.04	5993	5.0

Note: The California test score data set is described in Appendix 4.1.

district ("district 1") is 690.8; this is the average of the math and science test scores for all fifth-graders in that district in 1999 on a standardized test (the Stanford Achievement Test). The average student–teacher ratio in that district is 17.89; that is, the number of students in district 1 divided by the number of classroom teachers in district 1 is 17.89. Average expenditure per pupil in district 1 is $6385. The percentage of students in that district still learning English—that is, the percentage of students for whom English is a second language and who are not yet proficient in English—is 0%.

The remaining rows present data for other districts. The order of the rows is arbitrary, and the number of the district, which is called the **observation number**, is an arbitrarily assigned number that organizes the data. As you can see in the table, all the variables listed vary considerably.

With cross-sectional data, we can learn about relationships among variables by studying differences across people, firms, or other economic entities during a single time period.

Time Series Data

Time series data are data for a single entity (person, firm, country) collected at multiple time periods. Our data set on the growth rate of GDP and the term spread in the United States is an example of a time series data set. The data set contains observations on two variables (the growth rate of GDP and the term spread) for a single entity (the United States) for 232 time periods. Each time period in this data set is a quarter of a year (the first quarter is January, February, and March; the second

TABLE 1.2	**Selected Observations on the Growth Rate of GDP and the Term Spread in the United States: Quarterly Data, 1960:Q1–2017:Q4**		
Observation Number	**Date (year: quarter)**	**GDP Growth Rate (% at an annual rate)**	**Term Spread (percentage points)**
1	1960:Q1	8.8%	0.6
2	1960:Q2	−1.5	1.3
3	1960:Q3	1.0	1.5
4	1960:Q4	−4.9	1.6
5	1961:Q1	2.7	1.4
⋮	⋮	⋮	⋮
230	2017:Q2	3.0	1.4
231	2017:Q3	3.1	1.2
232	2017:Q4	2.5	1.2

Note: The United States GDP and term spread data set is described in Appendix 15.1.

quarter is April, May, and June; and so forth). The observations in this data set begin in the first quarter of 1960, which is denoted 1960:Q1, and end in the fourth quarter of 2017 (2017:Q4). The number of observations (that is, time periods) in a time series data set is denoted T. Because there are 232 quarters from 1960:Q1 to 2017:Q4, this data set contains $T = 232$ observations.

Some observations in this data set are listed in Table 1.2. The data in each row correspond to a different time period (year and quarter). In the first quarter of 1960, for example, GDP grew 8.8% at an annual rate. In other words, if GDP had continued growing for four quarters at its rate during the first quarter of 1960, the level of GDP would have increased by 8.8%. In the first quarter of 1960, the long-term interest rate was 4.5%, and the short-term interest rate was 3.9%; so their difference, the term spread, was 0.6 percentage points.

By tracking a single entity over time, time series data can be used to study the evolution of variables over time and to forecast future values of those variables.

Panel Data

Panel data, also called **longitudinal data**, are data for multiple entities in which each entity is observed at two or more time periods. Our data on cigarette consumption and prices are an example of a panel data set, and selected variables and observations in that data set are listed in Table 1.3. The number of entities in a panel data set is denoted n, and the number of time periods is denoted T. In the cigarette data set, we have observations on $n = 48$ continental U.S. states (entities) for $T = 11$ years (time periods) from 1985 to 1995. Thus, there is a total of $n \times T = 48 \times 11 = 528$ observations.

TABLE 1.3	Selected Observations on Cigarette Sales, Prices, and Taxes, by State and Year for U.S. States, 1985–1995				
Observation Number	State	Year	Cigarette Sales (packs per capita)	Average Price per Pack (including taxes)	Total Taxes (cigarette excise tax + sales tax)
1	Alabama	1985	116.5	$1.022	$0.333
2	Arkansas	1985	128.5	1.015	0.370
3	Arizona	1985	104.5	1.086	0.362
⋮	⋮	⋮	⋮	⋮	⋮
47	West Virginia	1985	112.8	1.089	0.382
48	Wyoming	1985	129.4	0.935	0.240
49	Alabama	1986	117.2	1.080	0.334
⋮	⋮	⋮	⋮	⋮	⋮
96	Wyoming	1986	127.8	1.007	0.240
97	Alabama	1987	115.8	1.135	0.335
⋮	⋮	⋮	⋮	⋮	⋮
528	Wyoming	1995	112.2	1.585	0.360

Note: The cigarette consumption data set is described in Appendix 12.1.

Some data from the cigarette consumption data set are listed in Table 1.3. The first block of 48 observations lists the data for each state in 1985, organized alphabetically from Alabama to Wyoming. The next block of 48 observations lists the data for 1986, and so forth, through 1995. For example, in 1985, cigarette sales in Arkansas were 128.5 packs per capita (the total number of packs of cigarettes sold in Arkansas in 1985 divided by the total population of Arkansas in 1985 equals 128.5). The average price of a pack of cigarettes in Arkansas in 1985, including tax, was $1.015, of which 37¢ went to federal, state, and local taxes.

Panel data can be used to learn about economic relationships from the experiences of the many different entities in the data set and from the evolution over time of the variables for each entity.

The definitions of cross-sectional data, time series data, and panel data are summarized in Key Concept 1.1.

KEY CONCEPT

1.1

Cross-Sectional, Time Series, and Panel Data

- Cross-sectional data consist of multiple entities observed at a single time period.

- Time series data consist of a single entity observed at multiple time periods.

- Panel data (also known as longitudinal data) consist of multiple entities, where each entity is observed at two or more time periods.

Summary

1. Many decisions in business and economics require quantitative estimates of how a change in one variable affects another variable.

2. Conceptually, the way to estimate a causal effect is in an ideal randomized controlled experiment, but performing experiments in economic applications can be unethical, impractical, or too expensive.

3. Econometrics provides tools for estimating causal effects using either observational (nonexperimental) data or data from real-world, imperfect experiments.

4. Econometrics also provides tools for predicting the value of a variable of interest using information in other, related variables.

5. Cross-sectional data are gathered by observing multiple entities at a single point in time; time series data are gathered by observing a single entity at multiple points in time; and panel data are gathered by observing multiple entities, each of which is observed at multiple points in time.

Key Terms

randomized controlled experiment (5)   observational data (7)
control group (5)   cross-sectional data (7)
treatment group (5)   observation number (8)
causal effect (6)   time series data (8)
prediction (6)   panel data (9)
forecast (6)   longitudinal data (9)
experimental data (7)

Review the Concepts

1.1   Describe a hypothetical ideal randomized controlled experiment to study the effect of hours spent studying on performance on microeconomics exams. Suggest some impediments to implementing this experiment in practice.

1.2   Describe a hypothetical ideal randomized controlled experiment to study the effect on highway traffic deaths of wearing seat belts. Suggest some impediments to implementing this experiment in practice.

1.3   You are asked to study the causal effect of hours spent on employee training (measured in hours per worker per week) in a manufacturing plant on the productivity of its workers (output per worker per hour). Describe:

    a.   an ideal randomized controlled experiment to measure this causal effect;

    b.   an observational cross-sectional data set with which you could study this effect;

    c.   an observational time series data set for studying this effect; and

    d.   an observational panel data set for studying this effect.

2 Review of Probability

This chapter reviews the core ideas of the theory of probability that are needed to understand regression analysis and econometrics. We assume that you have taken an introductory course in probability and statistics. If your knowledge of probability is stale, you should refresh it by reading this chapter. If you feel confident with the material, you still should skim the chapter and the terms and concepts at the end to make sure you are familiar with the ideas and notation.

Most aspects of the world around us have an element of randomness. The theory of probability provides mathematical tools for quantifying and describing this randomness. Section 2.1 reviews probability distributions for a single random variable, and Section 2.2 covers the mathematical expectation, mean, and variance of a single random variable. Most of the interesting problems in economics involve more than one variable, and Section 2.3 introduces the basic elements of probability theory for two random variables. Section 2.4 discusses three special probability distributions that play a central role in statistics and econometrics: the normal, chi-squared, and F distributions.

The final two sections of this chapter focus on a specific source of randomness of central importance in econometrics: the randomness that arises by randomly drawing a sample of data from a larger population. For example, suppose you survey ten recent college graduates selected at random, record (or "observe") their earnings, and compute the average earnings using these ten data points (or "observations"). Because you chose the sample at random, you could have chosen ten different graduates by pure random chance; had you done so, you would have observed ten different earnings, and you would have computed a different sample average. Because the average earnings vary from one randomly chosen sample to the next, the sample average is itself a random variable. Therefore, the sample average has a probability distribution, which is referred to as its sampling distribution because this distribution describes the different possible values of the sample average that would have occurred had a different sample been drawn.

Section 2.5 discusses random sampling and the sampling distribution of the sample average. This sampling distribution is, in general, complicated. When the sample size is sufficiently large, however, the sampling distribution of the sample average is approximately normal, a result known as the central limit theorem, which is discussed in Section 2.6.

2.1 Random Variables and Probability Distributions

Probabilities, the Sample Space, and Random Variables

Probabilities and outcomes. The sex of the next new person you meet, your grade on an exam, and the number of times your wireless network connection fails while you are writing a term paper all have an element of chance or randomness. In each of these examples, there is something not yet known that is eventually revealed.

The mutually exclusive potential results of a random process are called the **outcomes**. For example, while writing your term paper, the wireless connection might never fail, it might fail once, it might fail twice, and so on. Only one of these outcomes will actually occur (the outcomes are mutually exclusive), and the outcomes need not be equally likely.

The **probability** of an outcome is the proportion of the time that the outcome occurs in the long run. If the probability of your wireless connection not failing while you are writing a term paper is 80%, then over the course of writing many term papers, you will complete 80% without a wireless connection failure.

The sample space and events. The set of all possible outcomes is called the **sample space**. An **event** is a subset of the sample space; that is, an event is a set of one or more outcomes. The event "my wireless connection will fail no more than once" is the set consisting of two outcomes: "no failures" and "one failure."

Random variables. A random variable is a numerical summary of a random outcome. The number of times your wireless connection fails while you are writing a term paper is random and takes on a numerical value, so it is a random variable.

Some random variables are discrete and some are continuous. As their names suggest, a **discrete random variable** takes on only a discrete set of values, like $0, 1, 2, \ldots$, whereas a **continuous random variable** takes on a continuum of possible values.

Probability Distribution of a Discrete Random Variable

Probability distribution. The **probability distribution** of a discrete random variable is the list of all possible values of the variable and the probability that each value will occur. These probabilities sum to 1.

For example, let M be the number of times your wireless network connection fails while you are writing a term paper. The probability distribution of the random variable M is the list of probabilities of all possible outcomes: The probability that $M = 0$, denoted $\Pr(M = 0)$, is the probability of no wireless connection failures; $\Pr(M = 1)$ is the probability of a single connection failure; and so forth. An example of a probability distribution for M is given in the first row of Table 2.1. According to this distribution, the probability of no connection failures is 80%; the probability of one failure is 10%; and the probabilities of two, three, and four failures are,

TABLE 2.1 Probability of Your Wireless Network Connection Failing *M* Times					
	Outcome (number of failures)				
	0	**1**	**2**	**3**	**4**
Probability distribution	0.80	0.10	0.06	0.03	0.01
Cumulative probability distribution	0.80	0.90	0.96	0.99	1.00

respectively, 6%, 3%, and 1%. These probabilities sum to 100%. This probability distribution is plotted in Figure 2.1.

Probabilities of events. The probability of an event can be computed from the probability distribution. For example, the probability of the event of one or two failures is the sum of the probabilities of the constituent outcomes. That is, $\Pr(M = 1 \text{ or } M = 2) = \Pr(M = 1) + \Pr(M = 2) = 0.10 + 0.06 = 0.16$, or 16%.

Cumulative probability distribution. The **cumulative probability distribution** is the probability that the random variable is less than or equal to a particular value. The final row of Table 2.1 gives the cumulative probability distribution of the random variable M. For example, the probability of at most one connection failure, $\Pr(M \leq 1)$, is 90%, which is the sum of the probabilities of no failures (80%) and of one failure (10%).

  A cumulative probability distribution is also referred to as a **cumulative distribution function**, a **c.d.f.**, or a **cumulative distribution**.

FIGURE 2.1   Probability Distribution of the Number of Wireless Network Connection Failures

The height of each bar is the probability that the wireless connection fails the indicated number of times. The height of the first bar is 0.8, so the probability of 0 connection failures is 80%. The height of the second bar is 0.1, so the probability of 1 failure is 10%, and so forth for the other bars.
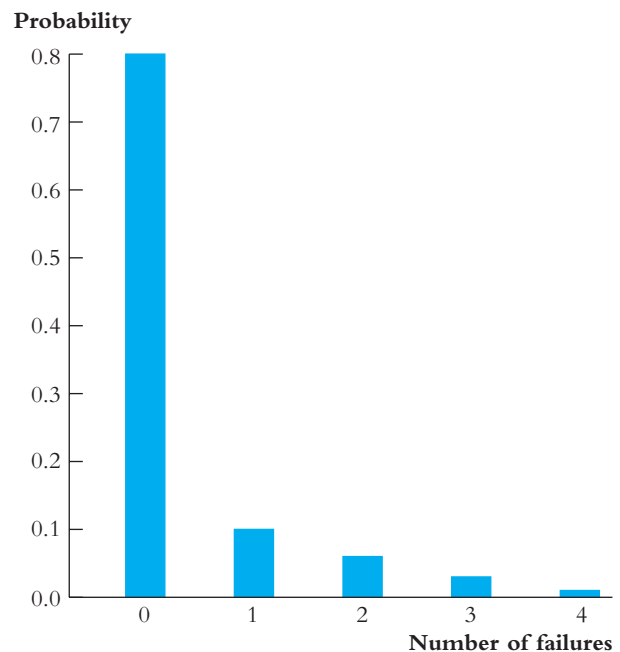
The Bernoulli distribution. An important special case of a discrete random variable is when the random variable is binary; that is, the outcome is 0 or 1. A binary random variable is called a **Bernoulli random variable** (in honor of the 17th-century Swiss mathematician and scientist Jacob Bernoulli), and its probability distribution is called the **Bernoulli distribution**.

For example, let G be the sex of the next new person you meet, where $G = 0$ indicates that the person is male and $G = 1$ indicates that the person is female. The outcomes of G and their probabilities thus are

$$G = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p, \end{cases} \tag{2.1}$$

where p is the probability of the next new person you meet being a woman. The probability distribution in Equation (2.1) is the Bernoulli distribution.

Probability Distribution of a Continuous Random Variable

Cumulative probability distribution. The cumulative probability distribution for a continuous variable is defined just as it is for a discrete random variable. That is, the cumulative probability distribution of a continuous random variable is the probability that the random variable is less than or equal to a particular value.

For example, consider a student who drives from home to school. This student's commuting time can take on a continuum of values, and because it depends on random factors such as the weather and traffic conditions, it is natural to treat it as a continuous random variable. Figure 2.2a plots a hypothetical cumulative distribution of commuting times. For example, the probability that the commute takes less than 15 minutes is 20%, and the probability that it takes less than 20 minutes is 78%.

Probability density function. Because a continuous random variable can take on a continuum of possible values, the probability distribution used for discrete variables, which lists the probability of each possible value of the random variable, is not suitable for continuous variables. Instead, the probability is summarized by the **probability density function**. The area under the probability density function between any two points is the probability that the random variable falls between those two points. A probability density function is also called a **p.d.f.**, a **density function**, or simply a **density**.

Figure 2.2b plots the probability density function of commuting times corresponding to the cumulative distribution in Figure 2.2a. The probability that the commute takes between 15 and 20 minutes is given by the area under the p.d.f. between 15 minutes and 20 minutes, which is 0.58, or 58%. Equivalently, this probability can be seen on the cumulative distribution in Figure 2.2a as the difference between the probability that the commute is less than 20 minutes (78%) and the probability that it is less than 15 minutes (20%). Thus the probability density function and the cumulative probability distribution show the same information in different formats.

FIGURE 2.2   **Cumulative Probability Distribution and Probability Density Functions of Commuting Time**

Probability

Pr (*Commuting time* ≤ 20) = 0.78

Pr (*Commuting time* ≤ 15) = 0.20

Commuting time (minutes)
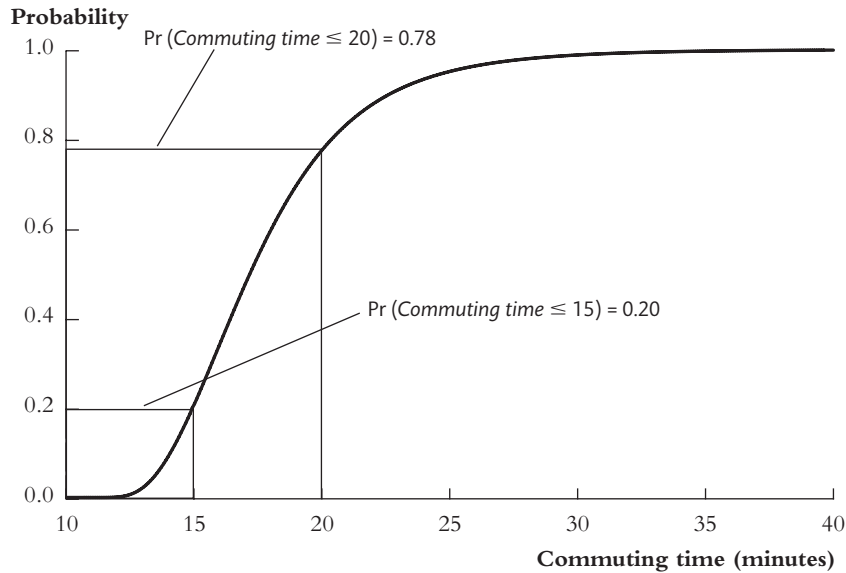
(a) Cumulative probability distribution function of commuting times

Probability density

Pr (*Commuting time* ≤ 15) = 0.20

Pr (15 < *Commuting time* ≤ 20) = 0.58

Pr (*Commuting time* > 20) = 0.22

0.58

0.20

0.22

Commuting time (minutes)
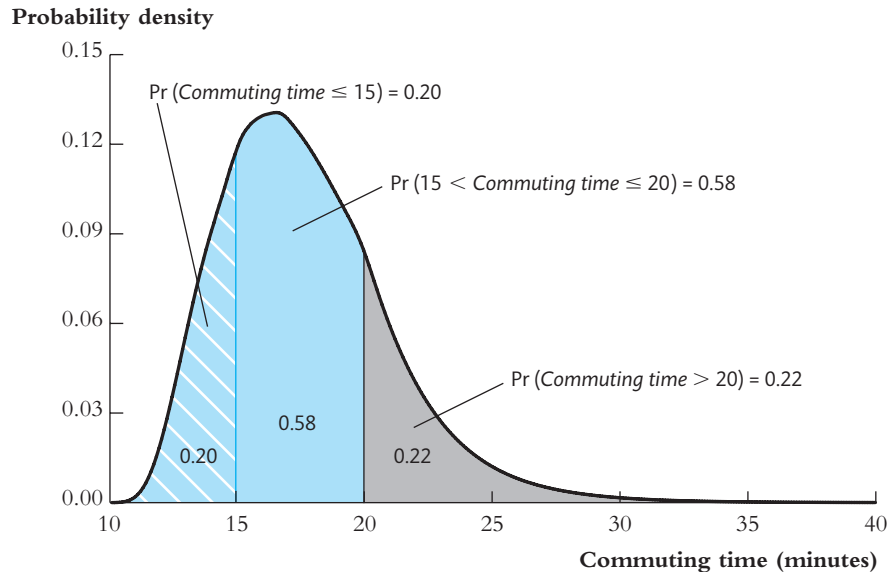
(b) Probability density function of commuting times

Figure 2.2a shows the cumulative probability distribution function (c.d.f.) of commuting times. The probability that a commuting time is less than 15 minutes is 0.20 (or 20%), and the probability that it is less than 20 minutes is 0.78 (78%). Figure 2.2b shows the probability density function (or p.d.f.) of commuting times. Probabilities are given by areas under the p.d.f. The probability that a commuting time is between 15 and 20 minutes is 0.58 (58%) and is given by the area under the curve between 15 and 20 minutes.

2.2 Expected Values, Mean, and Variance

The Expected Value of a Random Variable

Expected value. The **expected value** of a random variable Y, denoted $E(Y)$, is the long-run average value of the random variable over many repeated trials or occurrences. The expected value of a discrete random variable is computed as a weighted average of the possible outcomes of that random variable, where the weights are the probabilities of that outcome. The expected value of Y is also called the **expectation** of Y or the **mean** of Y and is denoted μ_Y.

For example, suppose you loan a friend $100 at 10% interest. If the loan is repaid, you get $110 (the principal of $100 plus interest of $10), but there is a risk of 1% that your friend will default and you will get nothing at all. Thus the amount you are repaid is a random variable that equals $110 with probability 0.99 and equals $0 with probability 0.01. Over many such loans, 99% of the time you would be paid back $110, but 1% of the time you would get nothing, so on average you would be repaid $110 \times 0.99 + \$0 \times 0.01 = \$108.90$. Thus the expected value of your repayment is $108.90.

As a second example, consider the number of wireless network connection failures M with the probability distribution given in Table 2.1. The expected value of M—that is, the mean of M—is the average number of failures over many term papers, weighted by the frequency with which a given number of failures occurs. Accordingly,

$$E(M) = 0 \times 0.80 + 1 \times 0.10 + 2 \times 0.06 + 3 \times 0.03 + 4 \times 0.01 = 0.35. \quad (2.2)$$

That is, the expected number of connection failures while writing a term paper is 0.35. Of course, the actual number of failures must always be an integer; it makes no sense to say that the wireless connection failed 0.35 times while writing a particular term paper! Rather, the calculation in Equation (2.2) means that the average number of failures over many such term papers is 0.35.

The formula for the expected value of a discrete random variable Y that can take on k different values is given in Key Concept 2.1. (Key Concept 2.1 uses summation notation, which is reviewed in Exercise 2.25.)

KEY CONCEPT

2.1

Expected Value and the Mean

Suppose that the random variable Y takes on k possible values, $y_1, \ldots, y_k$, where y_1 denotes the first value, y_2 denotes the second value, and so forth, and that the probability that Y takes on y_1 is p_1, the probability that Y takes on y_2 is p_2, and so forth. The expected value of Y, denoted $E(Y)$, is

$$E(Y) = y_1 p_1 + y_2 p_2 + \cdots + y_k p_k = \sum_{i=1}^{k} y_i p_i, \quad (2.3)$$

where the notation $\sum_{i=1}^{k} y_i p_i$ means "the sum of $y_i p_i$ for i running from 1 to k." The expected value of Y is also called the mean of Y or the expectation of Y and is denoted μ_Y.

Expected value of a Bernoulli random variable. An important special case of the general formula in Key Concept 2.1 is the mean of a Bernoulli random variable. Let G be the Bernoulli random variable with the probability distribution in Equation (2.1). The expected value of G is

$$E(G) = 0 \times (1 - p) + 1 \times p = p. \tag{2.4}$$

Thus the expected value of a Bernoulli random variable is p, the probability that it takes on the value 1.

Expected value of a continuous random variable. The expected value of a continuous random variable is also the probability-weighted average of the possible outcomes of the random variable. Because a continuous random variable can take on a continuum of possible values, the formal mathematical definition of its expectation involves calculus and its definition is given in Appendix 18.1.

The Standard Deviation and Variance

The variance and standard deviation measure the dispersion or the "spread" of a probability distribution. The **variance** of a random variable Y, denoted var(Y), is the expected value of the square of the deviation of Y from its mean: $\text{var}(Y) = E[(Y - \mu_Y)^2]$.

Because the variance involves the square of Y, the units of the variance are the units of the square of Y, which makes the variance awkward to interpret. It is therefore common to measure the spread by the **standard deviation**, which is the square root of the variance and is denoted σ_Y. The standard deviation has the same units as Y. These definitions are summarized in Key Concept 2.2.

For example, the variance of the number of connection failures M is the probability-weighted average of the squared difference between M and its mean, 0.35:

$$\text{var}(M) = (0 - 0.35)^2 \times 0.80 + (1 - 0.35)^2 \times 0.10 + (2 - 0.35)^2 \times 0.06 \\ + (3 - 0.35)^2 \times 0.03 + (4 - 0.35)^2 \times 0.01 = 0.6475. \tag{2.5}$$

The standard deviation of M is the square root of the variance, so $\sigma_M = \sqrt{0.64750} \cong 0.80$.

Variance and Standard Deviation

KEY CONCEPT

2.2

The variance of the discrete random variable Y, denoted σ_Y^2, is

$$\sigma_Y^2 = \text{var}(Y) = E[(Y - \mu_Y)^2] = \sum_{i=1}^{k} (y_i - \mu_Y)^2 p_i. \tag{2.6}$$

The standard deviation of Y is σ_Y, the square root of the variance. The units of the standard deviation are the same as the units of Y.

Variance of a Bernoulli random variable. The mean of the Bernoulli random variable G with the probability distribution in Equation (2.1) is $\mu_G = p$ [Equation (2.4)], so its variance is

$$\text{var}(G) = \sigma_G^2 = (0 - p)^2 \times (1 - p) + (1 - p)^2 \times p = p(1 - p). \quad (2.7)$$

Thus the standard deviation of a Bernoulli random variable is $\sigma_G = \sqrt{p(1 - p)}$.

Mean and Variance of a Linear Function of a Random Variable

This section discusses random variables (say, X and Y) that are related by a linear function. For example, consider an income tax scheme under which a worker is taxed at a rate of 20% on his or her earnings and then given a (tax-free) grant of $2000. Under this tax scheme, after-tax earnings Y are related to pre-tax earnings X by the equation

$$Y = 2000 + 0.8X. \quad (2.8)$$

That is, after-tax earnings Y is 80% of pre-tax earnings X, plus $2000.

Suppose an individual's pre-tax earnings next year are a random variable with mean μ_X and variance σ_X^2. Because pre-tax earnings are random, so are after-tax earnings. What are the mean and standard deviations of her after-tax earnings under this tax? After taxes, her earnings are 80% of the original pre-tax earnings, plus $2000. Thus the expected value of her after-tax earnings is

$$E(Y) = \mu_Y = 2000 + 0.8\mu_X. \quad (2.9)$$

The variance of after-tax earnings is the expected value of $(Y - \mu_Y)^2$. Because $Y = 2000 + 0.8X$, $\quad Y - \mu_Y = 2000 + 0.8X - (2000 + 0.8\mu_X) = 0.8(X - \mu_X)$. Thus $E[(Y - \mu_Y)^2] = E\{[0.8(X - \mu_X)]^2\} = 0.64E[(X - \mu_X)^2]$. It follows that $\text{var}(Y) = 0.64\text{var}(X)$, so, taking the square root of the variance, the standard deviation of Y is

$$\sigma_Y = 0.8\sigma_X. \quad (2.10)$$

That is, the standard deviation of the distribution of her after-tax earnings is 80% of the standard deviation of the distribution of her pre-tax earnings.

This analysis can be generalized so that Y depends on X with an intercept a (instead of $2000) and a slope b (instead of 0.8) so that

$$Y = a + bX. \quad (2.11)$$

Then the mean and variance of Y are

$$\mu_Y = a + b\mu_X \quad \text{and} \quad (2.12)$$
$$\sigma_Y^2 = b^2\sigma_X^2, \quad (2.13)$$

and the standard deviation of Y is $\sigma_Y = b\sigma_X$. The expressions in Equations (2.9) and (2.10) are applications of the more general formulas in Equations (2.12) and (2.13) with $a = 2000$ and $b = 0.8$.

Other Measures of the Shape of a Distribution

The mean and standard deviation measure two important features of a distribution: its center (the mean) and its spread (the standard deviation). This section discusses measures of two other features of a distribution: the skewness, which measures the lack of symmetry of a distribution, and the kurtosis, which measures how thick, or "heavy," are its tails. The mean, variance, skewness, and kurtosis are all based on what are called the **moments of a distribution**.

Skewness.   Figure 2.3 plots four distributions, two that are symmetric (Figures 2.3a and 2.3b) and two that are not (Figures 2.3c and 2.3d). Visually, the distribution in Figure 2.3d appears to deviate more from symmetry than does the distribution in

FIGURE 2.3   Four Distributions with Different Skewness and Kurtosis

(a) Skewness = 0, kurtosis = 3

(b) Skewness = 0, kurtosis = 20

(c) Skewness = −0.1, kurtosis = 5

(d) Skewness = 0.6, kurtosis = 5
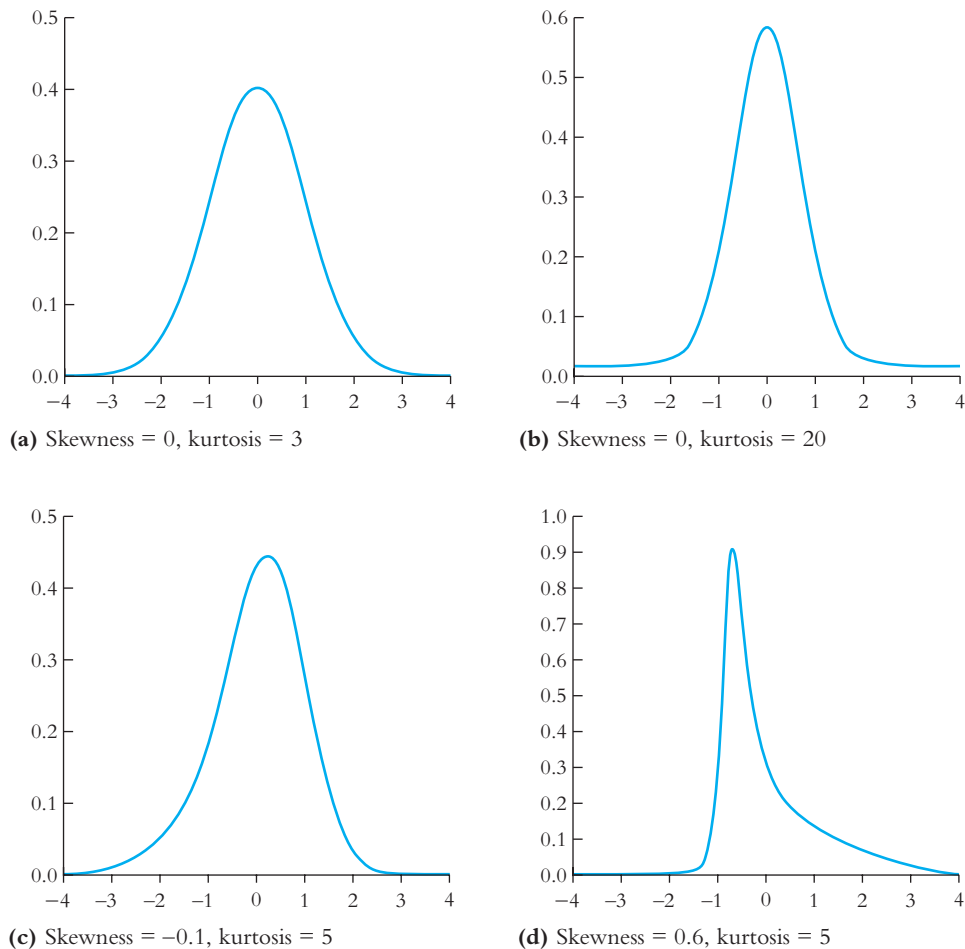
All of these distributions have a mean of 0 and a variance of 1. The distributions with skewness of 0 (a and b) are symmetric; the distributions with nonzero skewness (c and d) are not symmetric. The distributions with kurtosis exceeding 3 (b, c, and d) have heavy tails.

Figure 2.3c. The skewness of a distribution provides a mathematical way to describe how much a distribution deviates from symmetry.

The **skewness** of the distribution of a random variable Y is

$$\text{Skewness} = \frac{E[(Y - \mu_Y)^3]}{\sigma_Y^3}, \tag{2.14}$$

where σ_Y is the standard deviation of Y. For a symmetric distribution, a value of Y a given amount above its mean is just as likely as a value of Y the same amount below its mean. If so, then positive values of $(Y - \mu_Y)^3$ will be offset on average (in expectation) by equally likely negative values. Thus, for a symmetric distribution, $E(Y - \mu_Y)^3 = 0$: The skewness of a symmetric distribution is 0. If a distribution is not symmetric, then a positive value of $(Y - \mu_Y)^3$ generally is not offset on average by an equally likely negative value, so the skewness is nonzero for a distribution that is not symmetric. Dividing by σ_Y^3 in the denominator of Equation (2.14) cancels the units of Y^3 in the numerator, so the skewness is unit free; in other words, changing the units of Y does not change its skewness.

Below each of the four distributions in Figure 2.3 is its skewness. If a distribution has a long right tail, positive values of $(Y - \mu_Y)^3$ are not fully offset by negative values, and the skewness is positive. If a distribution has a long left tail, its skewness is negative.

Kurtosis. The **kurtosis** of a distribution is a measure of how much mass is in its tails and therefore is a measure of how much of the variance of Y arises from extreme values. An extreme value of Y is called an **outlier**. The greater the kurtosis of a distribution, the more likely are outliers.

The kurtosis of the distribution of Y is

$$\text{Kurtosis} = \frac{E[(Y - \mu_Y)^4]}{\sigma_Y^4}. \tag{2.15}$$

If a distribution has a large amount of mass in its tails, then some extreme departures of Y from its mean are likely, and these departures will lead to large values, on average (in expectation), of $(Y - \mu_Y)^4$. Thus, for a distribution with a large amount of mass in its tails, the kurtosis will be large. Because $(Y - \mu_Y)^4$ cannot be negative, the kurtosis cannot be negative.

The kurtosis of a normally distributed random variable is 3, so a random variable with kurtosis exceeding 3 has more mass in its tails than a normal random variable. A distribution with kurtosis exceeding 3 is called **leptokurtic** or, more simply, heavy-tailed. Like skewness, the kurtosis is unit free, so changing the units of Y does not change its kurtosis.

Below each of the four distributions in Figure 2.3 is its kurtosis. The distributions in Figures 2.3b–d are heavy-tailed.

Moments. The mean of Y, $E(Y)$, is also called the first moment of Y, and the expected value of the square of Y, $E(Y^2)$, is called the second moment of Y. In general, the

expected value of Y^r is called the rth **moment** of the random variable Y. That is, the rth moment of Y is $E(Y^r)$. The skewness is a function of the first, second, and third moments of Y, and the kurtosis is a function of the first through fourth moments of Y.

Standardized Random Variables

A random variable can be transformed into a random variable with mean 0 and variance 1 by subtracting its mean and then dividing by its standard deviation, a process called standardization. Specifically, let Y have mean μ_Y and variance σ_Y^2. Then the **standardized random variable** computed from Y is $(Y - \mu_Y)/\sigma_Y$. The mean of the standardized random variable is $E(Y - \mu_Y)/\sigma_Y = (EY - \mu_Y)/\sigma_Y = 0$, and its variance is $\text{var}[(Y - \mu_Y)/\sigma_Y] = \text{var}(Y)/\sigma_Y^2 = 1$. Standardized random variables do not have any units, such as dollars or meters, because the units of Y are canceled by dividing through by σ_Y, which also has the units of Y.

2.3   Two Random Variables

Most of the interesting questions in economics involve two or more variables. Are college graduates more likely to have a job than nongraduates? How does the distribution of income for women compare to that for men? These questions concern the distribution of two random variables, considered together (education and employment status in the first example, income and sex in the second). Answering such questions requires an understanding of the concepts of joint, marginal, and conditional probability distributions.

Joint and Marginal Distributions

Joint distribution.   The **joint probability distribution** of two discrete random variables, say X and Y, is the probability that the random variables simultaneously take on certain values, say x and y. The probabilities of all possible (x, y) combinations sum to 1. The joint probability distribution can be written as the function $\Pr(X = x, Y = y)$.

For example, weather conditions—whether or not it is raining—affect the commuting time of the student commuter in Section 2.1. Let Y be a binary random variable that equals 1 if the commute is short (less than 20 minutes) and that equals 0 otherwise, and let X be a binary random variable that equals 0 if it is raining and 1 if not. Between these two random variables, there are four possible outcomes: it rains and the commute is long ($X = 0, Y = 0$); rain and short commute ($X = 0, Y = 1$); no rain and long commute ($X = 1, Y = 0$); and no rain and short commute ($X = 1, Y = 1$). The joint probability distribution is the frequency with which each of these four outcomes occurs over many repeated commutes.

An example of a joint distribution of these two variables is given in Table 2.2. According to this distribution, over many commutes, 15% of the days have rain and a long commute ($X = 0, Y = 0$); that is, the probability of a long rainy commute is

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times			
	Rain ($X = 0$)	**No Rain ($X = 1$)**	**Total**
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

15%, or $\Pr(X = 0, Y = 0) = 0.15$. Also, $\Pr(X = 0, Y = 1) = 0.15$, $\Pr(X = 1, Y = 0) = 0.07$, and $\Pr(X = 1, Y = 1) = 0.63$. These four possible outcomes are mutually exclusive and constitute the sample space, so the four probabilities sum to 1.

Marginal probability distribution. The **marginal probability distribution** of a random variable Y is just another name for its probability distribution. This term is used to distinguish the distribution of Y alone (the marginal distribution) from the joint distribution of Y and another random variable.

The marginal distribution of Y can be computed from the joint distribution of X and Y by adding up the probabilities of all possible outcomes for which Y takes on a specified value. If X can take on l different values $x_1, \ldots, x_l$, then the marginal probability that Y takes on the value y is

$$\Pr(Y = y) = \sum_{i=1}^{l} \Pr(X = x_i, Y = y). \tag{2.16}$$

For example, in Table 2.2, the probability of a long rainy commute is 15%, and the probability of a long commute with no rain is 7%, so the probability of a long commute (rainy or not) is 22%. The marginal distribution of commuting times is given in the final column of Table 2.2. Similarly, the marginal probability that it will rain is 30%, as shown in the final row of Table 2.2.

Conditional Distributions

Conditional distribution. The distribution of a random variable Y conditional on another random variable X taking on a specific value is called the **conditional distribution** of Y given X. The conditional probability that Y takes on the value y when X takes on the value x is written $\Pr(Y = y | X = x)$.

For example, what is the probability of a long commute ($Y = 0$) if you know it is raining ($X = 0$)? From Table 2.2, the joint probability of a rainy short commute is 15%, and the joint probability of a rainy long commute is 15%, so if it is raining, a long commute and a short commute are equally likely. Thus the probability of a long commute ($Y = 0$) conditional on it being rainy ($X = 0$) is 50%, or $\Pr(Y = 0 | X = 0) = 0.50$. Equivalently, the marginal probability of rain is 30%; that is, over many commutes, it rains 30% of the time. Of this 30% of commutes, 50% of the time the commute is long ($0.15/0.30$).

| TABLE 2.3 | Joint and Conditional Distributions of Number of Wireless Connection Failures (M) and Network Age (A) |

A. Joint Distribution

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
Old network ($A = 0$)	0.35	0.065	0.05	0.025	0.01	0.50
New network ($A = 1$)	0.45	0.035	0.01	0.005	0.00	0.50
Total	0.80	0.10	0.06	0.03	0.01	1.00

B. Conditional Distributions of M given A

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
$\Pr(M \mid A = 0)$	0.70	0.13	0.10	0.05	0.02	1.00
$\Pr(M \mid A = 1)$	0.90	0.07	0.02	0.01	0.00	1.00

In general, the conditional distribution of Y given $X = x$ is

$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}. \tag{2.17}$$

For example, the conditional probability of a long commute given that it is rainy is $\Pr(Y = 0 \mid X = 0) = \Pr(X = 0, Y = 0) / \Pr(X = 0) = 0.15/0.30 = 0.50$.

As a second example, consider a modification of the network connection failure example. Suppose that half the time you write your term paper in the school library, which has a new wireless network; otherwise, you write it in your room, which has an old wireless network. If we treat the location where you write the term paper as random, then the network age A ($= 1$ if the network is new, $= 0$ if it is old) is a random variable. Suppose the joint distribution of the random variables M and A is given in Part A of Table 2.3. Then the conditional distributions of connection failures given the age of the network are shown in Part B of the table. For example, the joint probability of $M = 0$ and $A = 0$ is 0.35; because half the time you use the old network, the conditional probability of no failures given that you use the old network is $\Pr(M = 0 \mid A = 0) = \Pr(M = 0, A = 0)/\Pr(A = 0) = 0.35/0.50 = 0.70$, or 70%. In contrast, the conditional probability of no failures given that you use the new network is 90%. According to the conditional distributions in Part B of Table 2.3, the new network is less likely to fail than the old one; for example, the probability of three failures is 5% using the old network but 1% using the new network.

Conditional expectation. The **conditional expectation** of Y given X, also called the **conditional mean** of Y given X, is the mean of the conditional distribution of Y given X. That is, the conditional expectation is the expected value of Y, computed using the conditional distribution of Y given X. If Y takes on k values $y_1, \ldots, y_k$, then the conditional mean of Y given $X = x$ is

$$E(Y \mid X = x) = \sum_{i=1}^{k} y_i \Pr(Y = y_i \mid X = x). \tag{2.18}$$

For example, based on the conditional distributions in Table 2.3, the expected number of connection failures, given that the network is old, is $E(M|A = 0) = 0 \times 0.70 + 1 \times 0.13 + 2 \times 0.10 + 3 \times 0.05 + 4 \times 0.02 = 0.56$. The expected number of failures, given that the network is new, is $E(M|A = 1) = 0.14$, less than for the old network.

The conditional expectation of Y given $X = x$ is just the mean value of Y when $X = x$. In the example of Table 2.3, the mean number of failures is 0.56 for the old network, so the conditional expectation of Y given that the network is old is 0.56. Similarly, for the new network, the mean number of failures is 0.14; that is, the conditional expectation of Y given that the network is new is 0.14.

The law of iterated expectations. The mean of Y is the weighted average of the conditional expectation of Y given X, weighted by the probability distribution of X. For example, the mean height of adults is the weighted average of the mean height of men and the mean height of women, weighted by the proportions of men and women. Stated mathematically, if X takes on the l values $x_1, \ldots, x_l$, then

$$E(Y) = \sum_{i=1}^{l} E(Y|X = x_i)\Pr(X = x_i). \tag{2.19}$$

Equation (2.19) follows from Equations (2.18) and (2.17) (see Exercise 2.19).

Stated differently, the expectation of Y is the expectation of the conditional expectation of Y given X,

$$E(Y) = E[E(Y|X)], \tag{2.20}$$

where the inner expectation on the right-hand side of Equation (2.20) is computed using the conditional distribution of Y given X and the outer expectation is computed using the marginal distribution of X. Equation (2.20) is known as the **law of iterated expectations**.

For example, the mean number of connection failures M is the weighted average of the conditional expectation of M given that it is old and the conditional expectation of M given that it is new, so $E(M) = E(M|A = 0) \times \Pr(A = 0) + E(M|A = 1) \times \Pr(A = 1) = 0.56 \times 0.50 + 0.14 \times 0.50 = 0.35$. This is the mean of the marginal distribution of M, as calculated in Equation (2.2).

The law of iterated expectations implies that if the conditional mean of Y given X is 0, then the mean of Y is 0. This is an immediate consequence of Equation (2.20): if $E(Y|X) = 0$, then $E(Y) = E[E(Y|X)] = E[0] = 0$. Said differently, if the mean of Y given X is 0, then it must be that the probability-weighted average of these conditional means is 0; that is, the mean of Y must be 0.

The law of iterated expectations also applies to expectations that are conditional on multiple random variables. For example, let X, Y, and Z be random variables that are jointly distributed. Then the law of iterated expectations says that $E(Y) = E[E(Y|X, Z)]$, where $E(Y|X, Z)$ is the conditional expectation of Y

given both X and Z. For example, in the network connection illustration of Table 2.3, let P denote the number of people using the network; then $E(M|A, P)$ is the expected number of failures for a network with age A that has P users. The expected number of failures overall, $E(M)$, is the weighted average of the expected number of failures for a network with age A and number of users P, weighted by the proportion of occurrences of both A and P.

Exercise 2.20 provides some additional properties of conditional expectations with multiple variables.

Conditional variance. The variance of Y conditional on X is the variance of the conditional distribution of Y given X. Stated mathematically, the **conditional variance** of Y given X is

$$\text{var}\,(Y|X = x) = \sum_{i=1}^{k} [y_i - E(Y|X = x)]^2 \, \Pr\,(Y = y_i|X = x). \quad (2.21)$$

For example, the conditional variance of the number of failures given that the network is old is $\text{var}\,(M|A = 0) = (0 - 0.56)^2 \times 0.70 + (1 - 0.56)^2 \times 0.13 + (2 - 0.56)^2 \times 0.10 + (3 - 0.56)^2 \times 0.05 + (4 - 0.56)^2 \times 0.02 \cong 0.99$. The standard deviation of the conditional distribution of M given that $A = 0$ is thus $\sqrt{0.99} = 0.99$. The conditional variance of M given that $A = 1$ is the variance of the distribution in the second row of Part B of Table 2.3, which is 0.22, so the standard deviation of M for the new network is $\sqrt{0.22} = 0.47$. For the conditional distributions in Table 2.3, the expected number of failures for the new network (0.14) is less than that for the old network (0.56), and the spread of the distribution of the number of failures, as measured by the conditional standard deviation, is smaller for the new network (0.47) than for the old (0.99).

Bayes' rule. **Bayes' rule** says that the conditional probability of Y given X is the conditional probability of X given Y times the relative marginal probabilities of Y and X:

$$\Pr\,(Y = y|X = x) = \frac{\Pr\,(X = x|Y = y)\Pr\,(Y = y)}{\Pr(X = x)} \text{(Bayes' rule).} \quad (2.22)$$

Equation (2.22) obtains from the definition of the conditional distribution in Equation (2.17), which implies that $\Pr\,(X = x, Y = y) = \Pr\,(Y = y|X = x)\,\Pr\,(X = x)$ and that $\Pr\,(X = x, Y = y) = \Pr\,(X = x|Y = y)\Pr\,(Y = y)$; equating the second parts of these two equalities and rearranging gives Bayes' rule.

Bayes' rule can be used to deduce conditional probabilities from the reverse conditional probability, with the help of marginal probabilities. For example, suppose you told your friend that you were dropped by the network three times last night while working on your term paper and your friend knows that half the time you work in the library and half the time you work in your room. Then your friend could deduce from Table 2.3 that the probability you worked in your room last night given three network failures is 83% (Exercise 2.28).

The conditional mean is the minimum mean squared error prediction. The conditional mean plays a central role in prediction; in fact it is, in a precise sense, the optimal prediction of Y given $X = x$.

A common formulation of the statistical prediction problem is to posit that the cost of making a prediction error increases with the square of that error. The motivation for this squared-error prediction loss is that small errors in prediction might not matter much, but large errors can be very costly in real-world applications. Stated mathematically, the prediction problem thus is: what is the function $g(X)$ that minimizes the mean squared prediction error, $E\{[Y - g(X)]^2\}$? The answer is the conditional mean $E(Y|X)$: Of all possible ways to use the information X, the conditional mean minimizes the mean squared prediction error. This result is proven in Appendix 2.2.

Independence

Two random variables X and Y are **independently distributed**, or **independent**, if knowing the value of one of the variables provides no information about the other. Specifically, X and Y are independent if the conditional distribution of Y given X equals the marginal distribution of Y. That is, X and Y are independently distributed if, for all values of x and y,

$$\Pr(Y = y | X = x) = \Pr(Y = y) \text{ (independence of } X \text{ and } Y). \quad (2.23)$$

Substituting Equation (2.23) into Equation (2.17) gives an alternative expression for independent random variables in terms of their joint distribution. If X and Y are independent, then

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y). \quad (2.24)$$

That is, the joint distribution of two independent random variables is the product of their marginal distributions.

Covariance and Correlation

Covariance. One measure of the extent to which two random variables move together is their covariance. The **covariance** between X and Y is the expected value $E[(X - \mu_X)(Y - \mu_Y)]$, where μ_X is the mean of X and μ_Y is the mean of Y. The covariance is denoted $\text{cov}(X, Y)$ or σ_{XY}. If X can take on l values and Y can take on k values, then the covariance is given by the formula

$$\text{cov}(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{l}(x_j - \mu_X)(y_i - \mu_Y)\Pr(X = x_j, Y = y_i). \quad (2.25)$$

To interpret this formula, suppose that when X is greater than its mean (so that $X - \mu_X$ is positive), then Y tends be greater than its mean (so that $Y - \mu_Y$ is

positive) and that when X is less than its mean (so that $X - \mu_X < 0$), then Y tends to be less than its mean (so that $Y - \mu_Y < 0$). In both cases, the product $(X - \mu_X) \times (Y - \mu_Y)$ tends to be positive, so the covariance is positive. In contrast, if X and Y tend to move in opposite directions (so that X is large when Y is small, and vice versa), then the covariance is negative. Finally, if X and Y are independent, then the covariance is 0 (see Exercise 2.19).

Correlation. Because the covariance is the product of X and Y, deviated from their means, its units are, awkwardly, the units of X multiplied by the units of Y. This "units" problem can make numerical values of the covariance difficult to interpret.

The correlation is an alternative measure of dependence between X and Y that solves the "units" problem of the covariance. Specifically, the **correlation** between X and Y is the covariance between X and Y divided by their standard deviations:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\ \text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{2.26}$$

Because the units of the numerator in Equation (2.26) are the same as those of the denominator, the units cancel, and the correlation is unit free. The random variables X and Y are said to be **uncorrelated** if $\text{corr}(X, Y) = 0$.

The correlation always is between -1 and 1; that is, as proven in Appendix 2.1,

$$-1 \leq \text{corr}(X, Y) \leq 1 \quad (\text{correlation inequality}). \tag{2.27}$$

Correlation and conditional mean. If the conditional mean of Y does not depend on X, then Y and X are uncorrelated. That is,

$$\text{if } E(Y|X) = \mu_Y, \text{ then } \text{cov}(Y, X) = 0 \text{ and } \text{corr}(Y, X) = 0. \tag{2.28}$$

We now show this result. First, suppose Y and X have mean 0, so that $\text{cov}(Y, X) = E[(Y - \mu_Y)(X - \mu_X)] = E(YX)$. By the law of iterated expectations [Equation (2.20)], $E(YX) = E[E(YX|X)] = E[E(Y|X)X] = 0$ because $E(Y|X) = 0$, so $\text{cov}(Y, X) = 0$. Equation (2.28) follows by substituting $\text{cov}(Y, X) = 0$ into the definition of correlation in Equation (2.26). If Y and X do not have mean 0, subtract off their means, and then the preceding proof applies.

It is *not* necessarily true, however, that if X and Y are uncorrelated, then the conditional mean of Y given X does not depend on X. Said differently, it is possible for the conditional mean of Y to be a function of X but for Y and X nonetheless to be uncorrelated. An example is given in Exercise 2.23.

The Mean and Variance of Sums of Random Variables

The mean of the sum of two random variables, X and Y, is the sum of their means:

$$E(X + Y) = E(X) + E(Y) = \mu_X + \mu_Y. \tag{2.29}$$

The Distribution of Earnings in the United States in 2015

Some parents tell their children that they will be able to get a better, higher-paying job if they get a college degree than if they skip higher education. Are these parents right? Does the distribution of earnings differ between workers who are college graduates and workers who have only a high school diploma and, if so, how? Among workers with a similar education, does the distribution of earnings for men and women differ?

For example, do the best-paid college-educated women earn as much as the best-paid college-educated men?

One way to answer these questions is to examine the distribution of earnings of full-time workers, conditional on the highest educational degree achieved (high school diploma or bachelor's degree) and on sex. These four conditional distributions are shown in Figure 2.4, and the mean, standard deviation, and

---

FIGURE 2.4   **Conditional Distributions of Average Hourly Earnings of U.S. Full-Time Workers in 2015, Given Education Level and Sex**

The four distributions of earnings are for women and men, for those with only a high school diploma (a and c) and those whose highest degree is from a four-year college (b and d).
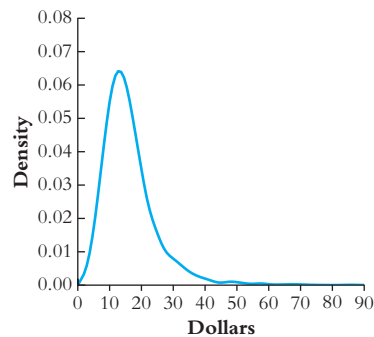
(a) Women with a high school diploma
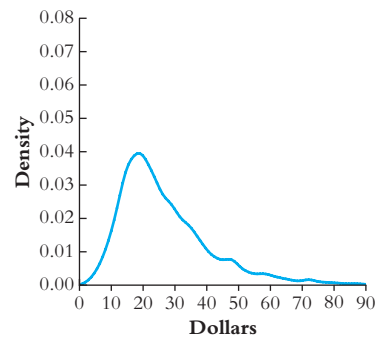
(b) Women with a four-year college degree
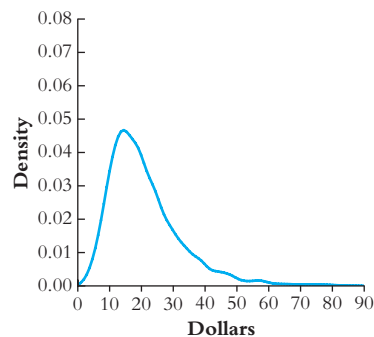
(c) Men with a high school diploma
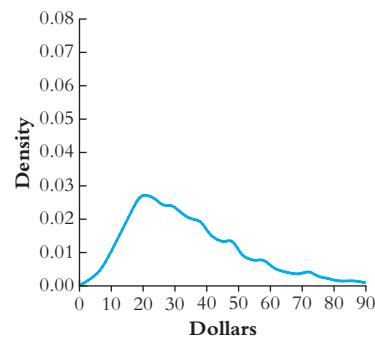
(d) Men with a four-year college degree

TABLE 2.4	Summary of the Conditional Distributions of Average Hourly Earnings of U.S. Full-Time Workers in 2015 Given Education Level and Sex						
					Percentile		
	Mean	**Standard Deviation**	**25%**	**50% (median)**	**75%**	**90%**
(a) Women with a high school diploma	$16.28	$8.91	$10.99	$14.42	$19.23	$25.64
(b) Women with a four-year college degree	27.23	16.18	16.83	23.56	33.65	47.60
(c) Men with a high school diploma	21.22	11.96	13.22	19.12	26.10	36.06
(d) Men with a four-year college degree	35.10	20.36	20.67	30.92	44.71	60.90

Average hourly earnings are the sum of annual pre-tax wages, salaries, tips, and bonuses divided by the number of hours worked annually.

some percentiles of the conditional distributions are presented in Table 2.4.[1] For example, the conditional mean of earnings for women whose highest degree is a high school diploma—that is, $E(Earnings|Highest\ degree = high\ school\ diploma, Sex = female)$—is $16.28 per hour.

The distribution of average hourly earnings for female college graduates (Figure 2.4b) is shifted to the right of the distribution for women with only a high school diploma (Figure 2.4a); the same shift can be seen for the two groups of men (Figure 2.4d and Figure 2.4c). For both men and women, mean earnings are higher for those with a college degree (Table 2.4, first numeric column). Interestingly, the spread of the distribution of earnings, as measured by the standard deviation, is greater for those with a college degree than for those with a high school diploma. In addition, for both men and women, the 90th percentile of earnings is much higher for workers with a college degree than for workers with only a high school diploma. This final comparison is consistent with the parental admonition that a college degree opens doors that remain closed to individuals with only a high school diploma.

Another feature of these distributions is that the distribution of earnings for men is shifted to the right of the distribution of earnings for women for a given level of education. This "gender gap" in earnings is an important—and, to many, troubling—aspect of the distribution of earnings. We return to this topic in later chapters.

---

[1]The distributions were estimated using data from the March 2016 Current Population Survey, which is discussed in more detail in Appendix 3.1.

KEY CONCEPT

2.3

Means, Variances, and Covariances of Sums of Random Variables

Let X, Y, and V be random variables; let μ_X and σ_X^2 be the mean and variance of X and let σ_{XY} be the covariance between X and Y (and so forth for the other variables); and let a, b, and c be constants. Equations (2.30) through (2.36) follow from the definitions of the mean, variance, and covariance:

$$E(a + bX + cY) = a + b\mu_X + c\mu_Y, \tag{2.30}$$

$$\text{var}(a + bY) = b^2\sigma_Y^2, \tag{2.31}$$

$$\text{var}(aX + bY) = a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2, \tag{2.32}$$

$$E(Y^2) = \sigma_Y^2 + \mu_Y^2, \tag{2.33}$$

$$\text{cov}(a + bX + cV, Y) = b\sigma_{XY} + c\sigma_{VY}, \tag{2.34}$$

$$E(XY) = \sigma_{XY} + \mu_X\mu_Y, \tag{2.35}$$

$$|\text{corr}(X, Y)| \leq 1 \text{ and } |\sigma_{XY}| \leq \sqrt{\sigma_X^2\sigma_Y^2} \text{ (correlation inequality)}. \tag{2.36}$$

The variance of the sum of X and Y is the sum of their variances plus two times their covariance:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}. \tag{2.37}$$

If X and Y are independent, then the covariance is 0, and the variance of their sum is the sum of their variances:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) = \sigma_X^2 + \sigma_Y^2$$
$$(\text{if } X \text{ and } Y \text{ are independent}). \tag{2.38}$$

Useful expressions for means, variances, and covariances involving weighted sums of random variables are collected in Key Concept 2.3. The results in Key Concept 2.3 are derived in Appendix 2.1.

2.4 The Normal, Chi-Squared, Student *t*, and *F* Distributions

The probability distributions most often encountered in econometrics are the normal, chi-squared, Student *t*, and *F* distributions.

The Normal Distribution

A continuous random variable with a **normal distribution** has the familiar bell-shaped probability density shown in Figure 2.5. The function defining the normal probability density is given in Appendix 18.1. As Figure 2.5 shows, the normal density with mean μ and variance σ^2 is symmetric around its mean and has 95% of its probability between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$.

Some special notation and terminology have been developed for the normal distribution. The normal distribution with mean μ and variance σ^2 is expressed concisely as $N(\mu, \sigma^2)$. The **standard normal distribution** is the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ and is denoted $N(0, 1)$. Random variables that have a $N(0, 1)$ distribution are often denoted Z, and the standard normal cumulative distribution function is denoted by the Greek letter Φ; accordingly, $\Pr(Z \leq c) = \Phi(c)$, where *c* is a constant. Values of the standard normal cumulative distribution function are tabulated in Appendix Table 1.

To look up probabilities for a normal variable with a general mean and variance, we must first standardize the variable. For example, suppose *Y* is distributed $N(1, 4)$—that is, *Y* is normally distributed with a mean of 1 and a variance of 4. What is the probability that $Y \leq 2$—that is, what is the shaded area in Figure 2.6a? The standardized version of *Y* is *Y* minus its mean, divided by its standard deviation; that is, $(Y - 1)/\sqrt{4} = \frac{1}{2}(Y - 1)$. Accordingly, the random variable $\frac{1}{2}(Y - 1)$ is normally distributed with mean 0 and variance 1 (see Exercise 2.8); it has the standard normal

---

FIGURE 2.5   **The Normal Probability Density**

The normal probability density function with mean μ and variance σ^2 is a bell-shaped curve, centered at μ. The area under the normal p.d.f. between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95. The normal distribution is denoted $N(\mu, \sigma^2)$.
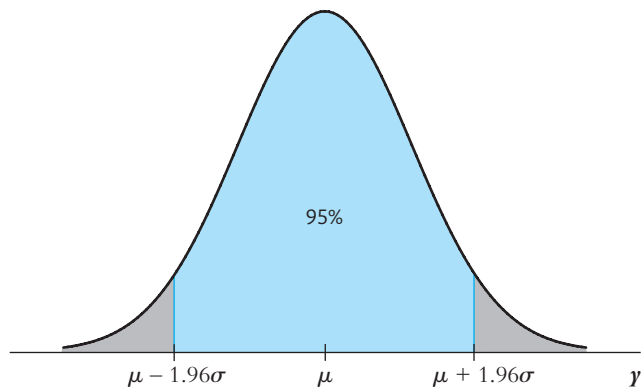
FIGURE 2.6 Calculating the Probability That $Y \leq 2$ When Y Is Distributed $N(1, 4)$

To calculate $\Pr(Y \leq 2)$, standardize Y, then use the standard normal distribution table. Y is standardized by subtracting its mean ($\mu = 1$) and dividing by its standard deviation ($\sigma = 2$). The probability that $Y \leq 2$ is shown in Figure 2.6a, and the corresponding probability after standardizing Y is shown in Figure 2.6b. Because the standardized random variable, $(Y - 1)/2$, is a standard normal (Z) random variable, $\Pr(Y \leq 2) = \Pr(\frac{Y-1}{2} \leq \frac{2-1}{2}) = \Pr(Z \leq 0.5)$. From Appendix Table 1, $\Pr(Z \leq 0.5) = \Phi(0.5) = 0.691$.

Pr($Y \leq 2$)

$N(1, 4)$ distribution

1.0 2.0

y

(a) $N(1, 4)$

0.691

Pr($Z \leq 0.5$)

$N(0, 1)$ distribution

0.0 0.5

z

(b) $N(0, 1)$
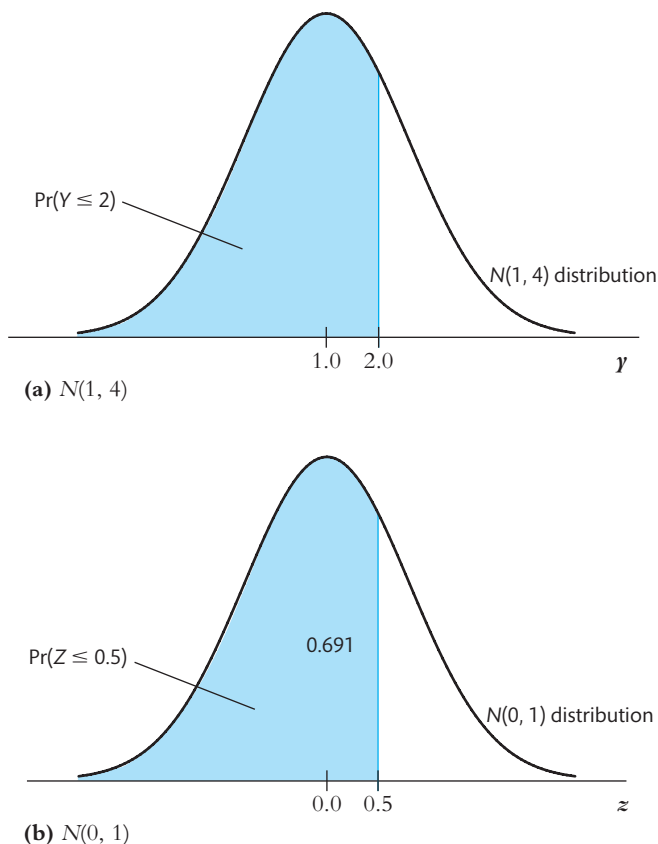
KEY CONCEPT

2.4

Computing Probabilities and Involving Normal Random Variables

Suppose Y is normally distributed with mean μ and variance σ^2; in other words, Y is distributed $N(\mu, \sigma^2)$. Then Y is standardized by subtracting its mean and dividing by its standard deviation, that is, by computing $Z = (Y - \mu)/\sigma$.

Let c_1 and c_2 denote two numbers with $c_1 < c_2$, and let $d_1 = (c_1 - \mu)/\sigma$ and $d_2 = (c_2 - \mu)/\sigma$. Then

$$\Pr(Y \leq c_2) = \Pr(Z \leq d_2) = \Phi(d_2), \tag{2.39}$$

$$\Pr(Y \geq c_1) = \Pr(Z \geq d_1) = 1 - \Phi(d_1), \tag{2.40}$$

$$\Pr(c_1 \leq Y \leq c_2) = \Pr(d_1 \leq Z \leq d_2) = \Phi(d_2) - \Phi(d_1). \tag{2.41}$$

The normal cumulative distribution function Φ is tabulated in Appendix Table 1.

distribution shown in Figure 2.6b. Now $Y \leq 2$ is equivalent to $\frac{1}{2}(Y - 1) \leq \frac{1}{2}(2 - 1)$; that is, $\frac{1}{2}(Y - 1) \leq \frac{1}{2}$. Thus

$$\Pr(Y \leq 2) = \Pr[\tfrac{1}{2}(Y - 1) \leq \tfrac{1}{2}] = \Pr(Z \leq \tfrac{1}{2}) = \Phi(0.5) = 0.691, \quad (2.42)$$

where the value 0.691 is taken from Appendix Table 1.

The same approach can be used to compute the probability that a normally distributed random variable exceeds some value or that it falls in a certain range. These steps are summarized in Key Concept 2.4. The box "A Bad Day on Wall Street" presents an unusual application of the cumulative normal distribution.

The normal distribution is symmetric, so its skewness is 0. The kurtosis of the normal distribution is 3.

The multivariate normal distribution. The normal distribution can be generalized to describe the joint distribution of a set of random variables. In this case, the distribution is called the **multivariate normal distribution** or, if only two variables are being considered, the **bivariate normal distribution**. The formula for the bivariate normal p.d.f. is given in Appendix 18.1, and the formula for the general multivariate normal p.d.f. is given in Appendix 19.2.

The multivariate normal distribution has four important properties. If X and Y have a bivariate normal distribution with covariance σ_{XY} and if a and b are two constants, then $aX + bY$ has the normal distribution:

$$aX + bY \text{ is distributed } N(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY})$$

$$(X,\ Y \text{ bivariate normal}). \quad (2.43)$$

A Bad Day on Wall Street

O n a typical day, the overall value of stocks traded on the U.S. stock market can rise or fall by 1% or even more. This is a lot—but nothing compared to what happened on Monday, October 19, 1987. On "Black Monday," the Dow Jones Industrial Average (an average of 30 large industrial stocks) fell by 22.6%! From January 1, 1980, to September 29, 2017, the standard deviation of daily percentage price changes on the Dow was 1.08%, so the drop of 22.6% was a negative return of $21(= 22.6/1.08)$ standard deviations. The enormity of this drop can be seen in Figure 2.7, a plot of the daily returns on the Dow during the 1980s.

If daily percentage price changes are normally distributed, then the probability of a change of at least 21 standard deviations is $\Pr(|Z| \geq 21) = 2 \times \Phi(-21)$. You will not find this value in Appendix Table 1, but you can calculate it using a computer (try it!). This probability is 6.6×10^{-98}—that is, $0.000\ldots00066$, where there are a total of 97 zeros!

How small is 6.6×10^{-98}? Consider the following:

• The world population is about 7.6 billion, so the probability of winning a random lottery among all living people is about 1 in 7.6 billion, or 1.3×10^{-10}.

FIGURE 2.7   Daily Percentage Changes in the Dow Jones Industrial Average in the 1980s
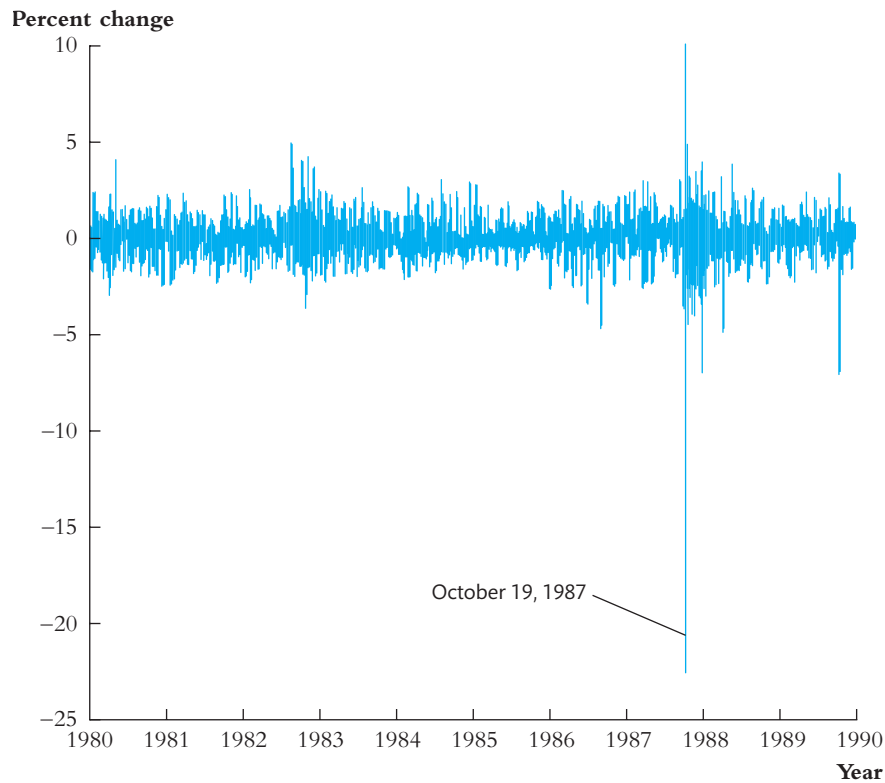
From January 1980 through September 2017, the average percentage daily change of "the Dow" index was 0.04% and its standard deviation was 1.08%. On October 19, 1987—"Black Monday"—the Dow fell 22.6%, or 21 standard deviations.

- The universe is believed to have existed for 14 billion years, or about 5×10^{17} seconds, so the probability of choosing a particular second at random from all the seconds since the beginning of time is 2×10^{-18}.
- There are approximately 10^{43} molecules of gas in the first kilometer above the earth's surface. The probability of choosing a particular molecule at random is 10^{-43}.

Although Wall Street *did* have a bad day, the fact that it happened at all suggests its probability was more than 6.6×10^{-98}. In fact, there have been many days—good and bad—with stock price changes too large to be consistent with a normal distribution with a constant variance. Table 2.5 lists the ten largest daily percentage price changes in the Dow Jones Industrial Average in the 9521 trading days between January 1, 1980 and September 29, 2017, along with the standardized change using the mean and variance over this period. All ten changes exceed 6.6 standard deviations, an extremely rare event if stock prices are normally distributed.

Clearly, stock price percentage changes have a distribution with heavier tails than the normal distribution. For this reason, finance professionals use other models of stock price changes. One such model treats stock price changes as normally distributed with a variance that evolves over time, so periods

TABLE 2.5	The Ten Largest Daily Percentage Changes in the Dow Jones Industrial Average, January 1980–September 2017, and the Normal Probability of a Change at Least as Large		
Date	**Percentage Change (x)**	**Standardized Change $z = (x - \mu)/\sigma$**	**Normal Probability of a Change at Least This Large $Pr(\lvert Z \rvert \geq \lvert z \rvert) = 2\Phi(-\lvert z \rvert)$**
October 19, 1987	−22.6	−21.0	6.6×10^{-98}
October 13, 2008	11.1	10.2	1.5×10^{-24}
October 28, 2008	10.9	10.0	1.0×10^{-23}
October 21, 1987	10.1	9.4	7.7×10^{-21}
October 26, 1987	−8.0	−7.5	7.2×10^{-14}
October 15, 2008	−7.9	−7.3	2.3×10^{-13}
December 01, 2008	−7.7	−7.2	7.4×10^{-13}
October 09, 2008	−7.3	−6.8	8.5×10^{-12}
October 27, 1997	−7.2	−6.7	2.2×10^{-11}
September 17, 2001	−7.1	−6.6	3.1×10^{-11}

like October 1987 and the financial crisis in the fall of 2008 have higher volatility than others (models with time-varying variances are discussed in Chapter 17). Other models abandon the normal distribution in favor of distributions with heavier tails, an idea popularized in Nassim Taleb's 2007 book, *The Black Swan*. These models are more consistent with the very bad— and very good—days we actually see on Wall Street.

More generally, if n random variables have a multivariate normal distribution, then any linear combination of these variables (such as their sum) is normally distributed.

Second, if a set of variables has a multivariate normal distribution, then the marginal distribution of each of the variables is normal [this follows from Equation (2.43) by setting $a = 1$ and $b = 0$].

Third, if variables with a multivariate normal distribution have covariances that equal 0, then the variables are independent. Thus, if X and Y have a bivariate normal distribution and $\sigma_{XY} = 0$, then X and Y are independent (this is shown in Appendix 18.1). In Section 2.3, it was shown that if X and Y are independent, then, regardless of their joint distribution, $\sigma_{XY} = 0$. If X and Y are jointly normally distributed, then the converse is also true. This result—that 0 covariance implies independence—is a special property of the multivariate normal distribution that is not true in general.

Fourth, if X and Y have a bivariate normal distribution, then the conditional expectation of Y given X is linear in X; that is, $E(Y|X = x) = a + bx$, where a and b are constants (Exercise 18.11). Joint normality implies linearity of conditional expectations, but linearity of conditional expectations does not imply joint normality.

The Chi-Squared Distribution

The chi-squared distribution is used when testing certain types of hypotheses in statistics and econometrics.

The **chi-squared distribution** is the distribution of the sum of m squared independent standard normal random variables. This distribution depends on m, which is called the degrees of freedom of the chi-squared distribution. For example, let Z_1, Z_2, and Z_3 be independent standard normal random variables. Then $Z_1^2 + Z_2^2 + Z_3^2$ has a chi-squared distribution with 3 degrees of freedom. The name for this distribution derives from the Greek letter used to denote it: A chi-squared distribution with m degrees of freedom is denoted χ_m^2.

Selected percentiles of the χ_m^2 distribution are given in Appendix Table 3. For example, Appendix Table 3 shows that the 95th percentile of the χ_3^2 distribution is 7.81, so $\Pr(Z_1^2 + Z_2^2 + Z_3^3 \leq 7.81) = 0.95$.

The Student *t* Distribution

The **Student *t* distribution** with m degrees of freedom is defined to be the distribution of the ratio of a standard normal random variable to the square root of an independently distributed chi-squared random variable with m degrees of freedom divided by m. That is, let Z be a standard normal random variable, let W be a random variable with a chi-squared distribution with m degrees of freedom, and let Z and W be independently distributed. Then the random variable $Z / \sqrt{W/m}$ has a Student *t* distribution (also called the ***t* distribution**) with m degrees of freedom. This distribution is denoted t_m. Selected percentiles of the Student *t* distribution are given in Appendix Table 2.

The Student *t* distribution depends on the degrees of freedom m. Thus the 95th percentile of the t_m distribution depends on the degrees of freedom m. The Student *t* distribution has a bell shape similar to that of the normal distribution, but it has more mass in the tails; that is, it is a "fatter" bell shape than the normal. When m is 30 or more, the Student *t* distribution is well approximated by the standard normal distribution, and the t_∞ distribution equals the standard normal distribution.

The *F* Distribution

The ***F* distribution** with m and n degrees of freedom, denoted $F_{m,n}$, is defined to be the distribution of the ratio of a chi-squared random variable with degrees of freedom m, divided by m, to an independently distributed chi-squared random variable with degrees of freedom n, divided by n. To state this mathematically, let W be a chi-squared random variable with m degrees of freedom and let V be a chi-squared random variable with n degrees of freedom, where W and V are independently distributed. Then $\frac{W/m}{V/n}$ has an $F_{m,n}$ distribution—that is, an F distribution with numerator degrees of freedom m and denominator degrees of freedom n.

In statistics and econometrics, an important special case of the F distribution arises when the denominator degrees of freedom is large enough that the $F_{m,n}$

distribution can be approximated by the $F_{m,\infty}$ distribution. In this limiting case, the denominator random variable V/n is the mean of infinitely many squared standard normal random variables, and that mean is 1 because the mean of a squared standard normal random variable is 1 (see Exercise 2.24). Thus the $F_{m,\infty}$ distribution is the distribution of a chi-squared random variable with m degrees of freedom divided by m: W/m is distributed $F_{m,\infty}$. For example, from Appendix Table 4, the 95th percentile of the $F_{3,\infty}$ distribution is 2.60, which is the same as the 95th percentile of the χ_3^2 distribution, 7.81 (from Appendix Table 2), divided by the degrees of freedom, which is $3(7.81/3 = 2.60)$.

The 90th, 95th, and 99th percentiles of the $F_{m,n}$ distribution are given in Appendix Table 5 for selected values of m and n. For example, the 95th percentile of the $F_{3,30}$ distribution is 2.92, and the 95th percentile of the $F_{3,90}$ distribution is 2.71. As the denominator degrees of freedom n increases, the 95th percentile of the $F_{3,n}$ distribution tends to the $F_{3,\infty}$ limit of 2.60.

2.5   Random Sampling and the Distribution of the Sample Average

Almost all the statistical and econometric procedures used in this text involve averages or weighted averages of a sample of data. Characterizing the distributions of sample averages therefore is an essential step toward understanding the performance of econometric procedures.

This section introduces some basic concepts about random sampling and the distributions of averages that are used throughout the book. We begin by discussing random sampling. The act of random sampling—that is, randomly drawing a sample from a larger population—has the effect of making the sample average itself a random variable. Because the sample average is a random variable, it has a probability distribution, which is called its sampling distribution. This section concludes with some properties of the sampling distribution of the sample average.

Random Sampling

Simple random sampling.   Suppose our commuting student from Section 2.1 aspires to be a statistician and decides to record her commuting times on various days. She selects these days at random from the school year, and her daily commuting time has the cumulative distribution function in Figure 2.2a. Because these days were selected at random, knowing the value of the commuting time on one of these randomly selected days provides no information about the commuting time on another of the days; that is, because the days were selected at random, the values of the commuting time on the different days are independently distributed random variables.

The situation described in the previous paragraph is an example of the simplest sampling scheme used in statistics, called **simple random sampling**, in which n objects are

2.5   Simple Random Sampling and i.i.d. Random Variables

In a simple random sample, n objects are drawn at random from a population, and each object is equally likely to be drawn. The value of the random variable Y for the ith randomly drawn object is denoted Y_i. Because each object is equally likely to be drawn and the distribution of Y_i is the same for all i, the random variables $Y_1, \ldots, Y_n$ are independently and identically distributed (i.i.d.); that is, the distribution of Y_i is the same for all $i = 1, \ldots, n$, and Y_1 is distributed independently of $Y_2, \ldots, Y_n$ and so forth.

selected at random from a **population** (the population of commuting days) and each member of the population (each day) is equally likely to be included in the sample.

The n observations in the sample are denoted $Y_1, \ldots, Y_n$, where Y_1 is the first observation, Y_2 is the second observation, and so forth. In the commuting example, Y_1 is the commuting time on the first of the n randomly selected days, and Y_i is the commuting time on the ith of the randomly selected days.

Because the members of the population included in the sample are selected at random, the values of the observations $Y_1, \ldots, Y_n$ are themselves random. If different members of the population are chosen, their values of Y will differ. Thus the act of random sampling means that $Y_1, \ldots, Y_n$ can be treated as random variables. Before they are sampled, $Y_1, \ldots, Y_n$ can take on many possible values; after they are sampled, a specific value is recorded for each observation.

i.i.d. draws. Because $Y_1, \ldots, Y_n$ are randomly drawn from the same population, the marginal distribution of Y_i is the same for each $i = 1, \ldots, n$; this marginal distribution is the distribution of Y in the population being sampled. When Y_i has the same marginal distribution for $i = 1, \ldots, n$, then $Y_1, \ldots, Y_n$ are said to be **identically distributed**.

Under simple random sampling, knowing the value of Y_1 provides no information about Y_2, so the conditional distribution of Y_2 given Y_1 is the same as the marginal distribution of Y_2. In other words, under simple random sampling, Y_1 is distributed independently of $Y_2, \ldots, Y_n$.

When $Y_1, \ldots, Y_n$ are drawn from the same distribution and are independently distributed, they are said to be **independently and identically distributed (i.i.d.)**.

Simple random sampling and i.i.d. draws are summarized in Key Concept 2.5.

The Sampling Distribution of the Sample Average

The **sample average** or **sample mean**, $\overline{Y}$, of the n observations $Y_1, \ldots, Y_n$ is

$$\overline{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{2.44}$$

An essential concept is that the act of drawing a random sample has the effect of making the sample average $\overline{Y}$ a random variable. Because the sample was drawn at random, the value of each Y_i is random. Because $Y_1, \ldots, Y_n$ are random, their average is random. Had a different sample been drawn, then the observations and their sample average would have been different: The value of $\overline{Y}$ differs from one randomly drawn sample to the next.

For example, suppose our student commuter selected five days at random to record her commute times, then computed the average of those five times. Had she chosen five different days, she would have recorded five different times—and thus would have computed a different value of the sample average.

Because $\overline{Y}$ is random, it has a probability distribution. The distribution of $\overline{Y}$ is called the **sampling distribution** of $\overline{Y}$ because it is the probability distribution associated with possible values of $\overline{Y}$ that could be computed for different possible samples $Y_1, \ldots, Y_n$.

The sampling distribution of averages and weighted averages plays a central role in statistics and econometrics. We start our discussion of the sampling distribution of $\overline{Y}$ by computing its mean and variance under general conditions on the population distribution of Y.

Mean and variance of $\overline{Y}$. Suppose that the observations $Y_1, \ldots, Y_n$ are i.i.d., and let μ_Y and σ_Y^2 denote the mean and variance of Y_i (because the observations are i.i.d., the mean is the same for all $i = 1, \ldots, n$, and so is the variance). When $n = 2$, the mean of the sum $Y_1 + Y_2$ is given by applying Equation (2.29): $E(Y_1 + Y_2) = \mu_Y + \mu_Y = 2\mu_Y$. Thus the mean of the sample average is $E[\frac{1}{2}(Y_1 + Y_2)] = \frac{1}{2} \times 2\mu_Y = \mu_Y$. In general,

$$E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \mu_Y. \tag{2.45}$$

The variance of $\overline{Y}$ is found by applying Equation (2.38). For example, for $n = 2$, $\mathrm{var}(Y_1 + Y_2) = 2\sigma_Y^2$, so [by applying Equation (2.32) with $a = b = \frac{1}{2}$ and $\mathrm{cov}(Y_1, Y_2) = 0$], $\mathrm{var}(\overline{Y}) = \frac{1}{2}\sigma_Y^2$. For general n, because $Y_1, \ldots, Y_n$ are i.i.d., Y_i and Y_j are independently distributed for $i \neq j$, so $\mathrm{cov}(Y_i, Y_j) = 0$. Thus

$$\mathrm{var}(\overline{Y}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(Y_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j \neq i}^{n}\mathrm{cov}(Y_i, Y_j)$$

$$= \frac{\sigma_Y^2}{n}. \tag{2.46}$$

The standard deviation of $\overline{Y}$ is the square root of the variance, $\sigma_Y/\sqrt{n}$.

Financial Diversification and Portfolios

The principle of diversification says that you can reduce your risk by holding small investments in multiple assets, compared to putting all your money into one asset. That is, you shouldn't put all your eggs in one basket.

The math of diversification follows from Equation (2.46). Suppose you divide \$1 equally among n assets. Let Y_i represent the payout in one year of \$1 invested in the ith asset. Because you invested $1/n$ dollars in each asset, the actual payoff of your portfolio after one year is $(Y_1 + Y_2 + \cdots + Y_n)/n = \overline{Y}$. To keep things simple, suppose that each asset has the same expected payout, μ_Y, the same variance, σ^2, and the same positive correlation, ρ, across assets [so that $\text{cov}(Y_i, Y_j) = \rho\sigma^2$]. Then the expected payout is $E(\overline{Y}) = \mu_Y$, and for large n, the variance of the portfolio payout is $\text{var}(\overline{Y}) = \rho\sigma^2$ (Exercise 2.26). Putting all your money into one asset or spreading it equally across all n assets has the same expected payout, but diversifying reduces the variance from σ^2 to $\rho\sigma^2$.

The math of diversification has led to financial products such as stock mutual funds, in which the fund holds many stocks and an individual owns a share of the fund, thereby owning a small amount of many stocks. But diversification has its limits: For many assets, payouts are positively correlated, so $\text{var}(\overline{Y})$ remains positive even if n is large. In the case of stocks, risk is reduced by holding a portfolio, but that portfolio remains subject to the unpredictable fluctuations of the overall stock market.

In summary, if $Y_1, \ldots, Y_n$ are i.i.d., the mean, the variance, and the standard deviation of $\overline{Y}$ are

$$E(\overline{Y}) = \mu_Y, \tag{2.47}$$

$$\text{var}(\overline{Y}) = \sigma_{\overline{Y}}^2 = \frac{\sigma_Y^2}{n}, \text{ and} \tag{2.48}$$

$$\text{std.dev}(\overline{Y}) = \sigma_{\overline{Y}} = \frac{\sigma_Y}{\sqrt{n}}. \tag{2.49}$$

These results hold whatever the distribution of Y is; that is, the distribution of Y does not need to take on a specific form, such as the normal distribution, for Equations (2.47) through (2.49) to hold.

The notation $\sigma_{\overline{Y}}^2$ denotes the variance of the sampling distribution of the sample average $\overline{Y}$. In contrast, σ_Y^2 is the variance of each individual Y_i, that is, the variance of the population distribution from which the observation is drawn. Similarly, $\sigma_{\overline{Y}}$ denotes the standard deviation of the sampling distribution of $\overline{Y}$.

Sampling distribution of $\overline{Y}$ when Y is normally distributed. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. draws from the $N(\mu_Y, \sigma_Y^2)$ distribution. As stated following Equation (2.43), the sum of n normally distributed random variables is itself normally distributed. Because the mean of $\overline{Y}$ is μ_Y and the variance of $\overline{Y}$ is σ_Y^2/n, this means that, if $Y_1, \ldots, Y_n$ are i.i.d. draws from the $N(\mu_Y, \sigma_Y^2)$ distribution, then $\overline{Y}$ is distributed $N(\mu_Y, \sigma_Y^2/n)$.

2.6 Large-Sample Approximations to Sampling Distributions

Sampling distributions play a central role in the development of statistical and econometric procedures, so it is important to know, in a mathematical sense, what the sampling distribution of $\overline{Y}$ is. There are two approaches to characterizing sampling distributions: an "exact" approach and an "approximate" approach.

The exact approach entails deriving a formula for the sampling distribution that holds exactly for any value of n. The sampling distribution that exactly describes the distribution of $\overline{Y}$ for any n is called the **exact distribution** or **finite-sample distribution** of $\overline{Y}$. For example, if Y is normally distributed and $Y_1, \ldots, Y_n$ are i.i.d., then (as discussed in Section 2.5) the exact distribution of $\overline{Y}$ is normal with mean μ_Y and variance σ_Y^2/n. Unfortunately, if the distribution of Y is not normal, then in general the exact sampling distribution of $\overline{Y}$ is very complicated and depends on the distribution of Y.

The approximate approach uses approximations to the sampling distribution that rely on the sample size being large. The large-sample approximation to the sampling distribution is often called the **asymptotic distribution**—"asymptotic" because the approximations become exact in the limit that $n \to \infty$. As we see in this section, these approximations can be very accurate even if the sample size is only $n = 30$ observations. Because sample sizes used in practice in econometrics typically number in the hundreds or thousands, these asymptotic distributions can be counted on to provide very good approximations to the exact sampling distribution.

This section presents the two key tools used to approximate sampling distributions when the sample size is large: the law of large numbers and the central limit theorem. The law of large numbers says that when the sample size is large, $\overline{Y}$ will be close to μ_Y with very high probability. The central limit theorem says that when the sample size is large, the sampling distribution of the standardized sample average, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$, is approximately normal.

Although exact sampling distributions are complicated and depend on the distribution of Y, the asymptotic distributions are simple. Moreover—remarkably—the asymptotic normal distribution of $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$ does *not* depend on the distribution of Y. This normal approximate distribution provides enormous simplifications and underlies the theory of regression used throughout this text.

The Law of Large Numbers and Consistency

The **law of large numbers** states that, under general conditions, $\overline{Y}$ will be near μ_Y with very high probability when n is large. This is sometimes called the "law of averages." When a large number of random variables with the same mean are averaged together, the large values tend to balance the small values, and their sample average is close to their common mean.

For example, consider a simplified version of our student commuter's experiment in which she simply records whether her commute was short (less than

Convergence in Probability, Consistency, and the Law of Large Numbers

The sample average $\overline{Y}$ converges in probability to μ_Y (or, equivalently, $\overline{Y}$ is consistent for μ_Y) if the probability that $\overline{Y}$ is in the range $(\mu_Y - c)$ to $(\mu_Y + c)$ becomes arbitrarily close to 1 as n increases for any constant $c > 0$. The convergence of $\overline{Y}$ to μ_Y in probability is written $\overline{Y} \xrightarrow{p} \mu_Y$.

The law of large numbers says that if $Y_1, \ldots, Y_n$ are independently and identically distributed with $E(Y_i) = \mu_Y$ and if large outliers are unlikely (technically if $\text{var}(Y_i) = \sigma_Y^2 < \infty$), then $\overline{Y} \xrightarrow{p} \mu_Y$.

20 minutes) or long. Let $Y_i = 1$ if her commute was short on the ith randomly selected day and $Y_i = 0$ if it was long. Because she used simple random sampling, $Y_1, \ldots, Y_n$ are i.i.d. Thus $Y_1, \ldots, Y_n$ are i.i.d. draws of a Bernoulli random variable, where (from Table 2.2) the probability that $Y_i = 1$ is 0.78. Because the expectation of a Bernoulli random variable is its success probability, $E(Y_i) = \mu_Y = 0.78$. The sample average $\overline{Y}$ is the fraction of days in her sample in which her commute was short.

Figure 2.8 shows the sampling distribution of $\overline{Y}$ for various sample sizes n. When $n = 2$ (Figure 2.8a), $\overline{Y}$ can take on only three values: 0, $\frac{1}{2}$, and 1 (neither commute was short, one was short, and both were short), none of which is particularly close to the true proportion in the population, 0.78. As n increases, however (Figures 2.8b–d), $\overline{Y}$ takes on more values, and the sampling distribution becomes tightly centered on μ_Y.

The property that $\overline{Y}$ is near μ_Y with probability increasing to 1 as n increases is called **convergence in probability** or, more concisely, **consistency** (see Key Concept 2.6). The law of large numbers states that under certain conditions $\overline{Y}$ converges in probability to μ_Y or, equivalently, that $\overline{Y}$ is consistent for μ_Y.

The conditions for the law of large numbers that we will use in this text are that $Y_1, \ldots, Y_n$ are i.i.d. and that the variance of Y_i, σ_Y^2, is finite. The mathematical role of these conditions is made clear in Section 18.2, where the law of large numbers is proven. If the data are collected by simple random sampling, then the i.i.d. assumption holds. The assumption that the variance is finite says that extremely large values of Y_i—that is, outliers—are unlikely and are observed infrequently; otherwise, these large values could dominate $\overline{Y}$, and the sample average would be unreliable. This assumption is plausible for the applications in this text. For example, because there is an upper limit to our student's commuting time (she could park and walk if the traffic is dreadful), the variance of the distribution of commuting times is finite.

The Central Limit Theorem

The **central limit theorem** says that, under general conditions, the distribution of $\overline{Y}$ is well approximated by a normal distribution when n is large. Recall that the mean of $\overline{Y}$ is μ_Y and its variance is $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$. According to the central limit theorem, when

FIGURE 2.8 Sampling Distribution of the Sample Average of *n* Bernoulli Random Variables
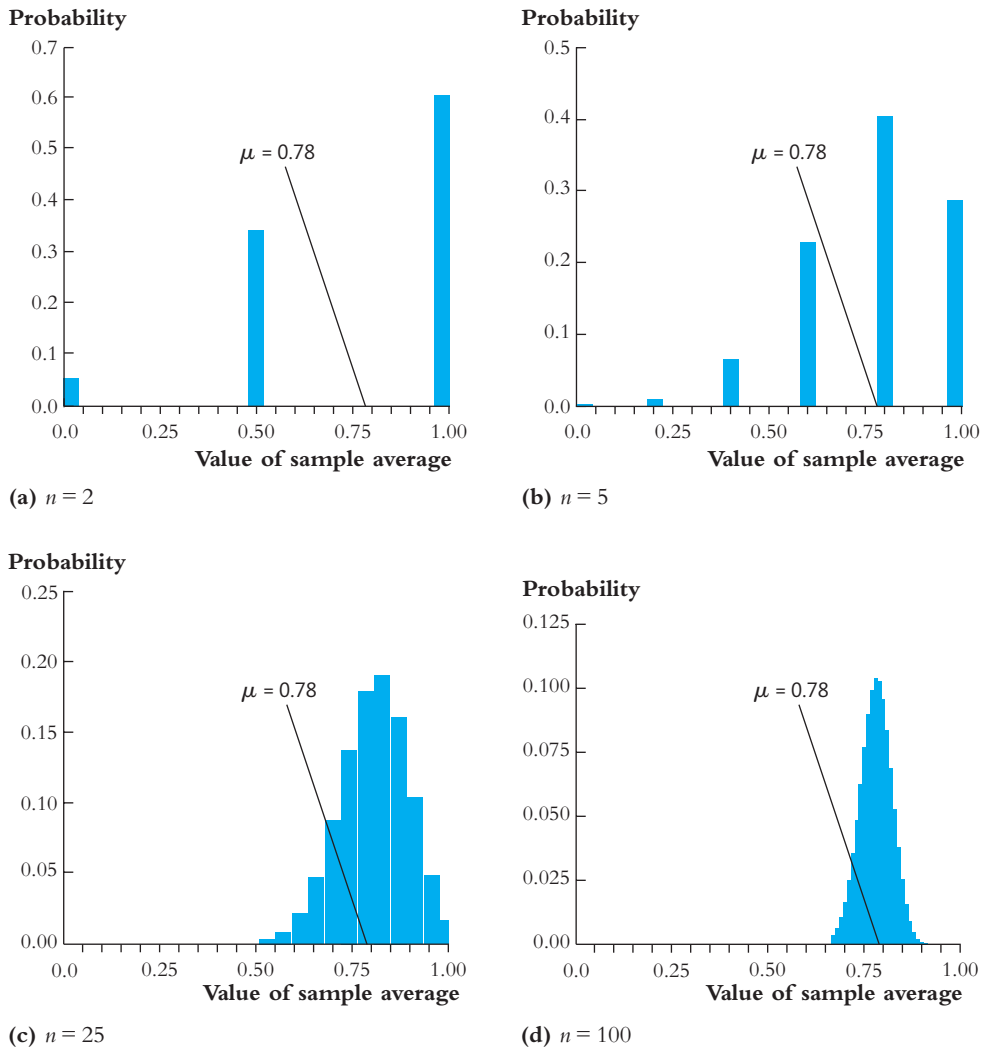
The distributions are the sampling distributions of $\overline{Y}$, the sample average of *n* independent Bernoulli random variables with $p = \Pr(Y_i = 1) = 0.78$ (the probability of a short commute is 78%). The variance of the sampling distribution of $\overline{Y}$ decreases as *n* gets larger, so the sampling distribution becomes more tightly concentrated around its mean, $\mu = 0.78$, as the sample size *n* increases.

n is large, the distribution of $\overline{Y}$ is approximately $N(\mu_Y, \sigma_{\overline{Y}}^2)$. As discussed at the end of Section 2.5, the distribution of $\overline{Y}$ is *exactly* $N(\mu_Y, \sigma_{\overline{Y}}^2)$ when the sample is drawn from a population with the normal distribution $N(\mu_Y, \sigma_Y^2)$. The central limit theorem says that this same result is *approximately* true when *n* is large even if $Y_1, \ldots, Y_n$ are not themselves normally distributed.

The convergence of the distribution of $\overline{Y}$ to the bell-shaped, normal approximation can be seen (a bit) in Figure 2.8. However, because the distribution gets quite tight for large *n*, this requires some squinting. It would be easier to see the shape of

FIGURE 2.9   Distribution of the Standardized Sample Average of *n* Bernoulli Random Variables with *p* = 0.78
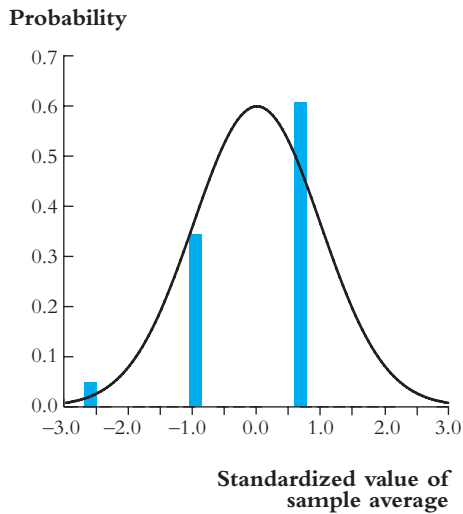
(a) *n* = 2
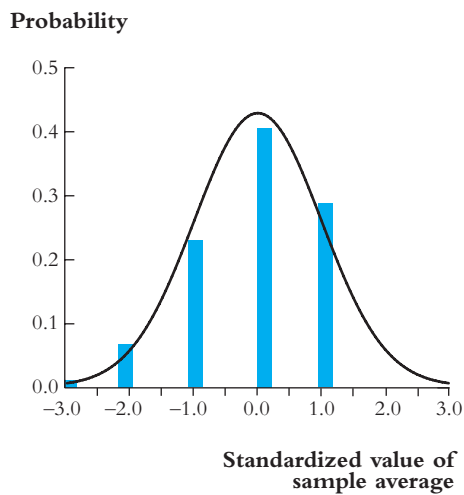
(b) *n* = 5
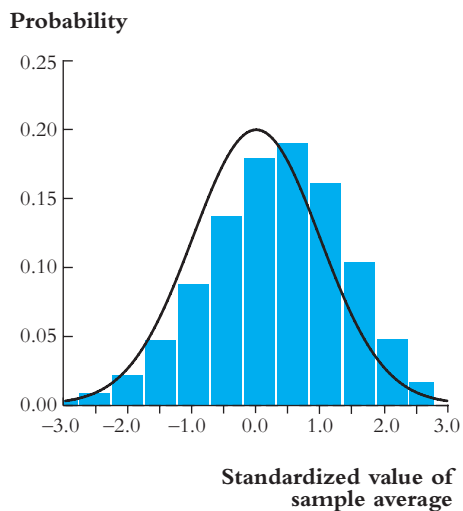
(c) *n* = 25
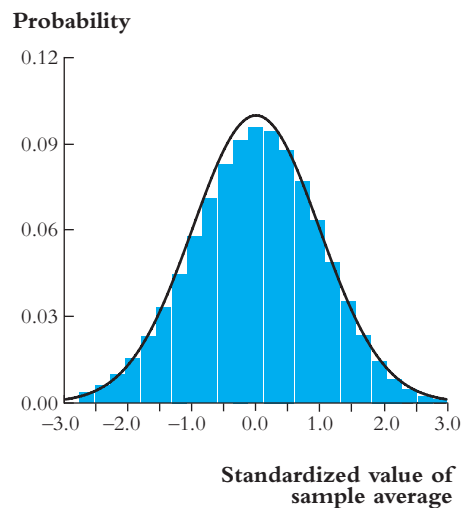
(d) *n* = 100

The sampling distributions of $\overline{Y}$ in Figure 2.8 are plotted here after standardizing $\overline{Y}$. Standardization centers the distributions in Figure 2.8 and magnifies the scale on the horizontal axis by a factor of $\sqrt{n}$. When the sample size is large, the sampling distributions are increasingly well approximated by the normal distribution (the solid line), as predicted by the central limit theorem. The normal distribution is scaled so that the height of the distribution is approximately the same in all figures.

the distribution of $\overline{Y}$ if you used a magnifying glass or had some other way to zoom in or to expand the horizontal axis of the figure.

One way to do this is to standardize $\overline{Y}$ so that it has a mean of 0 and a variance of 1. This process leads to examining the distribution of the standardized version of $\overline{Y}$, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$. According to the central limit theorem, this distribution should be well approximated by a $N(0, 1)$ distribution when n is large.

The distribution of the standardized average $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$ is plotted in Figure 2.9 for the distributions in Figure 2.8; the distributions in Figure 2.9 are exactly the same as in Figure 2.8, except that the scale of the horizontal axis is changed so that the standardized variable has a mean of 0 and a variance of 1. After this change of scale, it is easy to see that, if n is large enough, the distribution of $\overline{Y}$ is well approximated by a normal distribution.

One might ask, how large is "large enough"? That is, how large must n be for the distribution of $\overline{Y}$ to be approximately normal? The answer is, "It depends." The quality of the normal approximation depends on the distribution of the underlying Y_i that make up the average. At one extreme, if the Y_i are themselves normally distributed, then $\overline{Y}$ is exactly normally distributed for all n. In contrast, when the underlying Y_i themselves have a distribution that is far from normal, then this approximation can require $n = 30$ or even more.

This point is illustrated in Figure 2.10 for a population distribution, shown in Figure 2.10a, that is quite different from the Bernoulli distribution. This distribution has a long right tail (it is skewed to the right). The sampling distribution of $\overline{Y}$, after centering and scaling, is shown in Figures 2.10b–d for $n = 5, 25$, and 100, respectively. Although the sampling distribution is approaching the bell shape for $n = 25$, the normal approximation still has noticeable imperfections. By $n = 100$, however, the normal approximation is quite good. In fact, for $n \geq 100$, the normal approximation to the distribution of $\overline{Y}$ typically is very good for a wide variety of population distributions.

The central limit theorem is a remarkable result. While the "small n" distributions of $\overline{Y}$ in parts b and c of Figures 2.9 and 2.10 are complicated and quite different from each other, the "large n" distributions in Figures 2.9d and 2.10d are simple and, amazingly, have a similar shape. Because the distribution of $\overline{Y}$ approaches the normal as n grows large, $\overline{Y}$ is said to have an **asymptotic normal distribution**.

The convenience of the normal approximation, combined with its wide applicability because of the central limit theorem, makes it a key underpinning of applied econometrics. The central limit theorem is summarized in Key Concept 2.7.

The Central Limit Theorem

KEY CONCEPT

2.7

Suppose that $Y_1, \ldots, Y_n$ are i.i.d. with $E(Y_i) = \mu_Y$ and var $(Y_i) = \sigma_Y^2$, where $0 < \sigma_Y^2 < \infty$. As $n \rightarrow \infty$, the distribution of $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$ (where $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$) becomes arbitrarily well approximated by the standard normal distribution.

FIGURE 2.10   **Distribution of the Standardized Sample Average of *n* Draws from a Skewed Population Distribution**
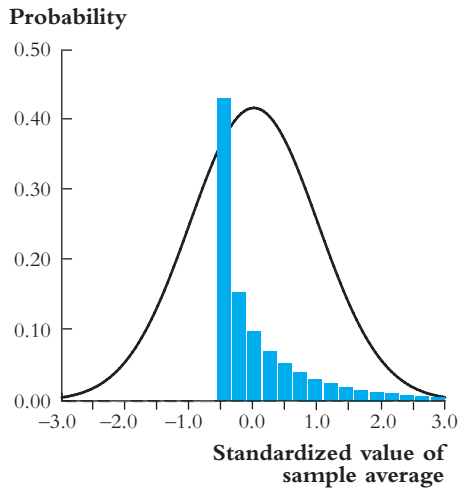
(a) *n* = 1
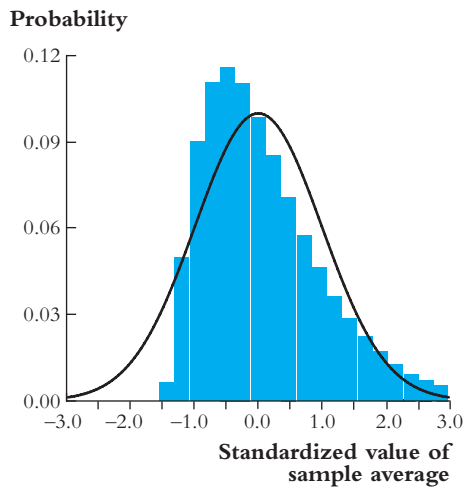
(b) *n* = 5
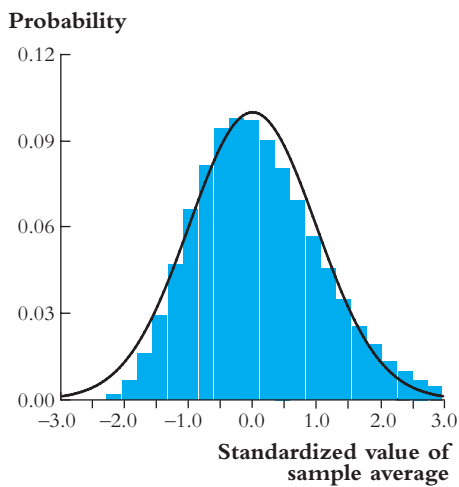
(c) *n* = 25
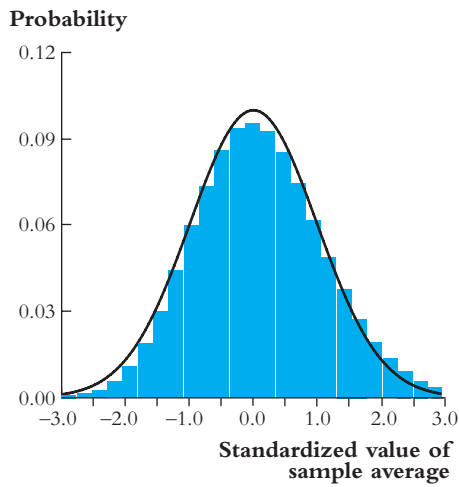
(d) *n* = 100

The figures show sampling distributions of the standardized sample average of *n* draws from the skewed (asymmetric) population distribution shown in Figure 2.10a. When *n* is small (*n* = 5), the sampling distribution, like the population distribution, is skewed. But when *n* is large (*n* = 100), the sampling distribution is well approximated by a standard normal distribution (solid line), as predicted by the central limit theorem. The normal distribution is scaled so that the height of the distribution is approximately the same in all figures.

Summary

1. The probabilities with which a random variable takes on different values are summarized by the cumulative distribution function, the probability distribution function (for discrete random variables), and the probability density function (for continuous random variables).

2. The expected value of a random variable Y (also called its mean, μ_Y), denoted $E(Y)$, is its probability-weighted average value. The variance of Y is $\sigma_Y^2 = E[(Y - \mu_Y)^2]$, and the standard deviation of Y is the square root of its variance.

3. The joint probabilities for two random variables, X and Y, are summarized by their joint probability distribution. The conditional probability distribution of Y given $X = x$ is the probability distribution of Y, conditional on X taking on the value x.

4. A normally distributed random variable has the bell-shaped probability density in Figure 2.5. To calculate a probability associated with a normal random variable, first standardize the variable, and then use the standard normal cumulative distribution tabulated in Appendix Table 1.

5. Simple random sampling produces n random observations, $Y_1, \ldots, Y_n$, that are independently and identically distributed (i.i.d.).

6. The sample average, $\overline{Y}$, varies from one randomly chosen sample to the next and thus is a random variable with a sampling distribution. If $Y_1, \ldots, Y_n$ are i.i.d., then
   a. the sampling distribution of $\overline{Y}$ has mean μ_Y and variance $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$;
   b. the law of large numbers says that $\overline{Y}$ converges in probability to μ_Y; and
   c. the central limit theorem says that the standardized version of $\overline{Y}$, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$, has a standard normal distribution [$N(0, 1)$ distribution] when n is large.

Key Terms

outcomes (14)

probability (14)

sample space (14)

event (14)

discrete random variable (14)

continuous random variable (14)

probability distribution (14)

cumulative probability distribution (15)

cumulative distribution function (c.d.f.) (15)

cumulative distribution (15)

Bernoulli random variable (16)

Bernoulli distribution (16)

probability density function (p.d.f.) (16)

density function (16)

density (16)

expected value (18)

expectation (18)

mean (18)

variance (19)

standard deviation (19)

moments of a distribution (21)

skewness (22)

kurtosis (22)

outlier (22)

leptokurtic (22)

r^{th} moment (23)

standardized random variable (23)

joint probability distribution (23)

marginal probability distribution (24)

conditional distribution (24)

conditional expectation (25)

conditional mean (25)

law of iterated expectations (26)

conditional variance (27)

Bayes' rule (27)

independently distributed (28)

independent (28)

covariance (28)

correlation (29)

uncorrelated (29)

normal distribution (33)

standard normal distribution (33)

multivariate normal distribution (35)

bivariate normal distribution (35)

chi-squared distribution (38)

Student t distribution (38)

t distribution (38)

F distribution (38)

simple random sampling (39)

population (40)

identically distributed (40)

independently and identically

  distributed (i.i.d.) (40)

sample average (40)

sample mean (40)

sampling distribution (41)

exact (finite-sample) distribution (43)

asymptotic distribution (43)

law of large numbers (43)

convergence in probability (44)

consistency (44)

central limit theorem (44)

asymptotic normal distribution (47)

MyLab Economics Can Help You Get a Better Grade

MyLab Economics   If your exam were tomorrow, would you be ready? For each chapter, **MyLab Economics** Practice Tests and Study Plan help you prepare for your exams. You can also find the Exercises and all Review the Concepts Questions available now in **MyLab Economics**. To see how it works, turn to the **MyLab Economics** spread on the inside front cover of this text and then go to **www.pearson.com/mylab/economics**.

For additional Empirical Exercises and Data Sets, log on to the Companion Website at **www.pearsonhighered.com/stock_watson**.

Review the Concepts

2.1   Examples of random variables used in this chapter included (a) the sex of the next person you meet, (b) the number of times a wireless network fails, (c) the time it takes to commute to school, and (d) whether it is raining or not. Explain why each can be thought of as random.

2.2   Suppose that the random variables X and Y are independent and you know their distributions. Explain why knowing the value of X tells you nothing about the value of Y.

2.3   Suppose that X denotes the amount of rainfall in your hometown during a randomly selected month and Y denotes the number of children born in Los Angeles during the same month. Are X and Y independent? Explain.

2.4 An econometrics class has 80 students, and the mean student weight is 145 lb. A random sample of 4 students is selected from the class, and their average weight is calculated. Will the average weight of the students in the sample equal 145 lb? Why or why not? Use this example to explain why the sample average, $\overline{Y}$, is a random variable.

2.5 Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random variables with a $N(1,4)$ distribution. Sketch the probability density of $\overline{Y}$ when $n = 2$. Repeat this for $n = 10$ and $n = 100$. In words, describe how the densities differ. What is the relationship between your answer and the law of large numbers?

2.6 Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random variables with the probability distribution given in Figure 2.10a. You want to calculate $\Pr(\overline{Y} \le 0.1)$. Would it be reasonable to use the normal approximation if $n = 5$? What about $n = 25$ or $n = 100$? Explain.

2.7 Y is a random variable with $\mu_Y = 0$, $\sigma_Y = 1$, skewness $= 0$, and kurtosis $= 100$. Sketch a hypothetical probability distribution of Y. Explain why n random variables drawn from this distribution might have some large outliers.

Exercises

2.1 Let Y denote the number of "heads" that occur when two coins are tossed.

   a. Derive the probability distribution of Y.

   b. Derive the cumulative probability distribution of Y.

   c. Derive the mean and variance of Y.

2.2 Use the probability distribution given in Table 2.2 to compute (a) $E(Y)$ and $E(X)$; (b) σ_X^2 and σ_Y^2; and (c) σ_{XY} and corr(X, Y).

2.3 Using the random variables X and Y from Table 2.2, consider two new random variables, $W = 3 + 6X$ and $V = 20 - 7Y$. Compute (a) $E(W)$ and $E(V)$; (b) σ_W^2 and σ_V^2; and (c) σ_{WV} and corr(W, V).

2.4 Suppose X is a Bernoulli random variable with $\Pr(X = 1) = p$.

   a. Show $E(X^3) = p$.

   b. Show $E(X^k) = p$ for $k > 0$.

   c. Suppose that $p = 0.3$. Compute the mean, variance, skewness, and kurtosis of X. (*Hint:* You might find it helpful to use the formulas given in Exercise 2.21.)

2.5   In September, Seattle's daily high temperature has a mean of 70° F and a standard deviation of 7° F. What are the mean, standard deviation, and variance in degrees Celsius?

2.6   The following table gives the joint probability distribution between employment status and college graduation among those either employed or looking for work (unemployed) in the working-age U.S. population for September 2017.

Joint Distribution of Employment Status and College Graduation in the U.S. Population Aged 25 and Older, September 2017			
	Unemployed ($Y = 0$)	Employed ($Y = 1$)	Total
Non–college grads ($X = 0$)	0.026	0.576	0.602
College grads ($X = 1$)	0.009	0.389	0.398
Total	0.035	0.965	1.000

   a. Compute $E(Y)$.

   b. The unemployment rate is the fraction of the labor force that is unemployed. Show that the unemployment rate is given by $1 - E(Y)$.

   c. Calculate $E(Y|X = 1)$ and $E(Y|X = 0)$.

   d. Calculate the unemployment rate for (i) college graduates and (ii) non–college graduates.

   e. A randomly selected member of this population reports being unemployed. What is the probability that this worker is a college graduate? A non–college graduate?

   f. Are educational achievement and employment status independent? Explain.

2.7   In a given population of two-earner male-female couples, male earnings have a mean of $40,000 per year and a standard deviation of $12,000. Female earnings have a mean of $45,000 per year and a standard deviation of $18,000. The correlation between male and female earnings for a couple is 0.80. Let C denote the combined earnings for a randomly selected couple.

   a. What is the mean of C?

   b. What is the covariance between male and female earnings?

   c. What is the standard deviation of C?

   d. Convert the answers to (a) through (c) from U.S. dollars ($) to euros (€).

2.8   The random variable Y has a mean of 1 and a variance of 4. Let $Z = \frac{1}{2}(Y - 1)$. Show that $\mu_Z = 0$ and $\sigma_Z^2 = 1$.

2.9 X and Y are discrete random variables with the following joint distribution:

		Value of Y				
		14	22	30	40	65
	1	0.02	0.05	0.10	0.03	0.01
Value of X	5	0.17	0.15	0.05	0.02	0.01
	8	0.02	0.03	0.15	0.10	0.09

That is, $\Pr(X = 1, \ Y = 14) = 0.02$, and so forth.

a. Calculate the probability distribution, mean, and variance of Y.

b. Calculate the probability distribution, mean, and variance of Y given $X = 8$.

c. Calculate the covariance and correlation between X and Y.

2.10 Compute the following probabilities:

a. If Y is distributed $N(1, 4)$, find $\Pr(Y \leq 3)$.

b. If Y is distributed $N(3, 9)$, find $\Pr(Y > 0)$.

c. If Y is distributed $N(50, 25)$, find $\Pr(40 \leq Y \leq 52)$.

d. If Y is distributed $N(5, 2)$, find $\Pr(6 \leq Y \leq 8)$.

2.11 Compute the following probabilities:

a. If Y is distributed χ_4^2, find $\Pr(Y \leq 7.78)$.

b. If Y is distributed χ_{10}^2, find $\Pr(Y > 18.31)$.

c. If Y is distributed $F_{10, \infty}$, find $\Pr(Y > 1.83)$.

d. Why are the answers to (b) and (c) the same?

e. If Y is distributed χ_1^2, find $\Pr(Y \leq 1.0)$. (*Hint:* Use the definition of the χ_1^2 distribution.)

2.12 Compute the following probabilities:

a. If Y is distributed t_{15}, find $\Pr(Y > 1.75)$.

b. If Y is distributed t_{90}, find $\Pr(-1.99 \leq Y \leq 1.99)$.

c. If Y is distributed $N(0, 1)$, find $\Pr(-1.99 \leq Y \leq 1.99)$.

d. Why are the answers to (b) and (c) approximately the same?

e. If Y is distributed $F_{7, 4}$, find $\Pr(Y > 4.12)$.

f. If Y is distributed $F_{7, 120}$, find $\Pr(Y > 2.79)$.

2.13 X is a Bernoulli random variable with $\Pr(X = 1) = 0.99$; Y is distributed $N(0, 1)$; W is distributed $N(0, 100)$; and X, Y, and W are independent. Let $S = XY + (1 - X)W$. (That is, $S = Y$ when $X = 1$, and $S = W$ when $X = 0$.)

   a. Show that $E(Y^2) = 1$ and $E(W^2) = 100$.

   b. Show that $E(Y^3) = 0$ and $E(W^3) = 0$. (*Hint:* What is the skewness for a symmetric distribution?)

   c. Show that $E(Y^4) = 3$ and $E(W^4) = 3 \times 100^2$. (*Hint:* Use the fact that the kurtosis is 3 for a normal distribution.)

   d. Derive $E(S)$, $E(S^2)$, $E(S^3)$, and $E(S^4)$. (*Hint:* Use the law of iterated expectations conditioning on $X = 0$ and $X = 1$.)

   e. Derive the skewness and kurtosis for S.

2.14 In a population, $\mu_Y = 100$ and $\sigma_Y^2 = 43$. Use the central limit theorem to answer the following questions:

   a. In a random sample of size $n = 100$, find $\Pr(\overline{Y} \leq 101)$.

   b. In a random sample of size $n = 165$, find $\Pr(\overline{Y} > 98)$.

   c. In a random sample of size $n = 64$, find $\Pr(101 \leq \overline{Y} \leq 103)$.

2.15 Suppose $Y_1, \ldots, Y_n$ are i.i.d. random variables, each distributed $N(10, 4)$.

   a. Compute $\Pr(9.6 \leq \overline{Y} \leq 10.4)$ when (i) $n = 20$, (ii) $n = 100$, and (iii) $n = 1000$.

   b. Suppose c is a positive number. Show that $\Pr(10 - c \leq \overline{Y} \leq 10 + c)$ becomes close to 1.0 as n grows large.

   c. Use your answer in (b) to argue that $\overline{Y}$ converges in probability to 10.

2.16 Y is distributed $N(5, 100)$, and you want to calculate $\Pr(Y < 3.6)$. Unfortunately, you do not have your textbook, and do not have access to a normal probability table like Appendix Table 1. However, you do have your computer and a computer program that can generate i.i.d. draws from the $N(5, 100)$ distribution. Explain how you can use your computer to compute an accurate approximation for $\Pr(Y < 3.6)$.

2.17 $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli random variables with $p = 0.4$. Let $\overline{Y}$ denote the sample mean.

   a. Use the central limit to compute approximations for

      i. $\Pr(\overline{Y} \geq 0.43)$ when $n = 100$.
      ii. $\Pr(\overline{Y} \leq 0.37)$ when $n = 400$.

   b. How large would n need to be to ensure that $\Pr(0.39 \leq \overline{Y} \leq 0.41) \geq 0.95$? (Use the central limit theorem to compute an approximate answer.)

2.18 In any year, the weather can inflict storm damage to a home. From year to year, the damage is random. Let Y denote the dollar value of damage in any given year. Suppose that in 95% of the years $Y = 0$ but in 5% of the years $Y = \$20,000$.

a. What are the mean and standard deviation of the damage in any year?

b. Consider an "insurance pool" of 100 people whose homes are sufficiently dispersed so that, in any year, the damage to different homes can be viewed as independently distributed random variables. Let $\overline{Y}$ denote the average damage to these 100 homes in a year. (i) What is the expected value of the average damage, $\overline{Y}$? (ii) What is the probability that $\overline{Y}$ exceeds $2000?

2.19 Consider two random variables, X and Y. Suppose that Y takes on k values $y_1, \ldots, y_k$ and that X takes on l values $x_1, \ldots, x_l$.

a. Show that $\Pr(Y = y_j) = \sum_{i=1}^{l} \Pr(Y = y_j | X = x_i) \Pr(X = x_i)$. [*Hint:* Use the definition of $\Pr(Y = y_j | X = x_i)$.]

b. Use your answer to (a) to verify Equation (2.19).

c. Suppose that X and Y are independent. Show that $\sigma_{XY} = 0$ and $\text{corr}(X, Y) = 0$.

2.20 Consider three random variables, X, Y, and Z. Suppose that Y takes on k values $y_1, \ldots, y_k$; that X takes on l values $x_1, \ldots, x_l$; and that Z takes on m values $z_1, \ldots, z_m$. The joint probability distribution of X, Y, Z is $\Pr(X = x, Y = y, Z = z)$, and the conditional probability distribution of Y given X and Z is $\Pr(Y = y | X = x, Z = z) = \frac{\Pr(Y = y, X = x, Z = z)}{\Pr(X = x, Z = z)}$.

a. Explain how the marginal probability that $Y = y$ can be calculated from the joint probability distribution. [*Hint:* This is a generalization of Equation (2.16).]

b. Show that $E(Y) = E[E(Y | X, Z)]$. [*Hint:* This is a generalization of Equations (2.19) and (2.20).]

2.21 X is a random variable with moments $E(X)$, $E(X^2)$, $E(X^3)$, and so forth.

a. Show $E(X - \mu)^3 = E(X^3) - 3[E(X^2)][E(X)] + 2[E(X)]^3$.

b. Show
$E(X - \mu)^4 = E(X^4) - 4[E(X)][E(X^3)] + 6[E(X)]^2[E(X^2)] - 3[E(X)]^4$.

2.22 Suppose you have some money to invest—for simplicity, $1—and you are planning to put a fraction w into a stock market mutual fund and the rest, $1 - w$, into a bond mutual fund. Suppose that $1 invested in a stock fund yields R_s after one year and that $1 invested in a bond fund yields R_b, suppose that R_s is random with mean 0.08 (8%) and standard deviation 0.07, and suppose that R_b is random with mean 0.05 (5%) and standard deviation 0.04. The correlation between R_s and R_b is 0.25. If you place a fraction w of your money in the stock fund and the rest, $1 - w$, in the bond fund, then the return on your investment is $R = wR_s + (1 - w)R_b$.

a. Suppose that $w = 0.5$. Compute the mean and standard deviation of R.

b. Suppose that $w = 0.75$. Compute the mean and standard deviation of R.

c. What value of w makes the mean of R as large as possible? What is the standard deviation of R for this value of w?

d. (Harder) What is the value of w that minimizes the standard deviation of R? (Show using a graph, algebra, or calculus.)

2.23 This exercise provides an example of a pair of random variables, X and Y, for which the conditional mean of Y given X depends on X but $\text{corr}(X, Y) = 0$. Let X and Z be two independently distributed standard normal random variables, and let $Y = X^2 + Z$.

a. Show that $E(Y|X) = X^2$.

b. Show that $\mu_Y = 1$.

c. Show that $E(XY) = 0$. (*Hint:* Use the fact that the odd moments of a standard normal random variable are all 0.)

d. Show that $\text{cov}(X, Y) = 0$ and thus $\text{corr}(X, Y) = 0$.

2.24 Suppose Y_i is distributed i.i.d. $N(0, \sigma^2)$ for $i = 1, 2, \ldots, n$.

a. Show that $E(Y_i^2/\sigma^2) = 1$.

b. Show that $W = (1/\sigma^2)\sum_{i=1}^{n}Y_i^2$ is distributed χ_n^2.

c. Show that $E(W) = n$. [*Hint:* Use your answer to (a).]

d. Show that $V = Y_1 \left/ \sqrt{\dfrac{\sum_{i=2}^{n}Y_i^2}{n-1}}\right.$ is distributed t_{n-1}.

2.25 (Review of summation notation) Let $x_1, \ldots, x_n$ denote a sequence of numbers; $y_1, \ldots, y_n$ denote another sequence of numbers; and $a, b,$ and c denote three constants. Show that

a. $\displaystyle\sum_{i=1}^{n}ax_i = a\sum_{i=1}^{n}x_i,$

b. $\displaystyle\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n}x_i + \sum_{i=1}^{n}y_i,$

c. $\displaystyle\sum_{i=1}^{n}a = n \times a,$ and

d. $\displaystyle\sum_{i=1}^{n}(a + bx_i + cy_i)^2 = na^2 + b^2\sum_{i=1}^{n}x_i^2 + c^2\sum_{i=1}^{n}y_i^2 + 2ab\sum_{i=1}^{n}x_i$
$\displaystyle + 2ac\sum_{i=1}^{n}y_i + 2bc\sum_{i=1}^{n}x_iy_i.$

2.26 Suppose that $Y_1, Y_2, \ldots, Y_n$ are random variables with a common mean μ_Y; a common variance σ_Y^2; and the same correlation ρ (so that the correlation between Y_i and Y_j is equal to ρ for all pairs i and j, where $i \neq j$).

a. Show that $\text{cov}(Y_i, Y_j) = \rho\sigma_Y^2$ for $i \neq j$.

b. Suppose that $n = 2$. Show that $E(\overline{Y}) = \mu_Y$ and $\text{var}(\overline{Y}) = \frac{1}{2}\sigma_Y^2 + \frac{1}{2}\rho\sigma_Y^2$.

c. For $n \geq 2$, show that $E(\overline{Y}) = \mu_Y$ and var $(\overline{Y}) = \sigma_Y^2/n +$ $[(n-1)/n]\rho\sigma_Y^2$.

d. When n is very large, show that var $(\overline{Y}) \simeq \rho\sigma_Y^2$.

2.27 Consider the problem of predicting Y using another variable, X, so that the prediction of Y is some function of X, say $g(X)$. Suppose that the quality of the prediction is measured by the squared prediction error made on average over all predictions, that is, by $E\{[Y - g(X)]^2\}$. This exercise provides a non-calculus proof that of all possible prediction functions g, the best prediction is made by the conditional expectation, $E(Y|X)$.

a. Let $\hat{Y} = E(Y|X)$, and let $u = Y - \hat{Y}$ denote its prediction error. Show that $E(u) = 0$. (*Hint:* Use the law of iterated expectations.)

b. Show that $E(uX) = 0$.

c. Let $\widetilde{Y} = g(X)$ denote a different prediction of Y using X, and let $v = Y - \widetilde{Y}$ denote its error. Show that $E[(Y - \widetilde{Y})^2] > E[(Y - \hat{Y})^2]$. [*Hint:* Let $h(X) = g(X) - E(Y|X)$, so that $v = [Y - E(Y|X)] - h(X)$. Derive $E(v^2)$.]

2.28 Refer to Part B of Table 2.3 for the conditional distribution of the number of network failures M given network age A. Let $\Pr(A = 0) = 0.5$; that is, you work in your room 50% of the time.

a. Compute the probability of three network failures, $\Pr(M = 3)$.

b. Use Bayes' rule to compute $\Pr(A = 0|M = 3)$.

c. Now suppose you work in your room one-fourth of the time, so $\Pr(A = 0) = 0.25$. Use Bayes' rule to compute $\Pr(A = 0|M = 3)$.

Empirical Exercise

E2.1 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find the spreadsheet **Age_HourlyEarnings**, which contains the joint distribution of age (*Age*) and average hourly earnings (*AHE*) for 25- to 34-year-old full-time workers in 2015 with an education level that exceeds a high school diploma. Use this joint distribution to carry out the following exercises. (*Note:* For these exercises, you need to be able to carry out calculations and construct charts using a spreadsheet.)

a. Compute the marginal distribution of *Age*.

b. Compute the mean of *AHE* for each value of *Age*; that is, compute, $E(AHE|Age = 25)$, and so forth.

c. Compute and plot the mean of *AHE* versus *Age*. Are average hourly earnings and age related? Explain.

  d. Use the law of iterated expectations to compute the mean of *AHE*; that is, compute $E(AHE)$.

  e. Compute the variance of *AHE*.

  f. Compute the covariance between *AHE* and *Age*.

  g. Compute the correlation between *AHE* and *Age*.

  h. Relate your answers in (f) and (g) to the plot you constructed in (c).

2.1 Derivation of Results in Key Concept 2.3

This appendix derives the equations in Key Concept 2.3.

Equation (2.30) follows from the definition of the expectation.

To derive Equation (2.31), use the definition of the variance to write $\text{var}(a + bY) = E\{[a + bY - E(a + bY)]^2\} = E\{[b(Y - \mu_Y)]^2\} = b^2 E[(Y - \mu_Y)^2] = b^2 \sigma_Y^2$.

To derive Equation (2.32), use the definition of the variance to write

$$
\begin{aligned}
\text{var}\,(aX + bY) &= E\{[(aX + bY) - (a\mu_X + b\mu_Y)]^2\} \\
&= E\{[a(X - \mu_X) + b(Y - \mu_Y)]^2\} \\
&= E[a^2(X - \mu_X)^2] + 2E[ab(X - \mu_X)(Y - \mu_Y)] \\
&\quad + E[b^2(Y - \mu_Y)^2] \\
&= a^2\text{var}\,(X) + 2ab\,\text{cov}(X, Y) + b^2\,\text{var}\,(Y) \\
&= a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2, \tag{2.50}
\end{aligned}
$$

where the second equality follows by collecting terms, the third equality follows by expanding the quadratic, and the fourth equality follows by the definition of the variance and covariance.

To derive Equation (2.33), write

$$E(Y^2) = E\{[(Y - \mu_Y) + \mu_Y]^2\} = E[(Y - \mu_Y)^2] + 2\,\mu_Y E(Y - \mu_Y) + \mu_Y^2 = \sigma_Y^2 + \mu_Y^2$$

because $E(Y - \mu_Y) = 0$.

To derive Equation (2.34), use the definition of the covariance to write

$$
\begin{aligned}
\text{cov}(a + bX + cV, Y) &= E\{[a + bX + cV - E(a + bX + cV)][Y - \mu_Y]\} \\
&= E\{[b(X - \mu_X) + c(V - \mu_V)][Y - \mu_Y]\} \\
&= E\{[b(X - \mu_X)][Y - \mu_Y]\} + E\{[c(V - \mu_V)][Y - \mu_Y]\} \\
&= b\sigma_{XY} + c\sigma_{VY}, \tag{2.51}
\end{aligned}
$$

which is Equation (2.34).

To derive Equation (2.35), write

$$
\begin{aligned}
E(XY) &= E\{[(X - \mu_X) + \mu_X][(Y - \mu_Y) + \mu_Y]\} \\
&= E[(X - \mu_X)(Y - \mu_Y)] + \mu_X E(Y - \mu_Y) + \mu_Y E(X - \mu_X) + \mu_X \mu_Y \\
&= \sigma_{XY} + \mu_X \mu_Y.
\end{aligned}
$$

We now prove the correlation inequality in Equation (2.36); that is, $|\mathrm{corr}(X, Y) \leq 1.|$ Let $a = -\sigma_{XY}/\sigma_X^2$ and $b = 1$. Applying Equation (2.32), we have,

$$
\begin{aligned}
\mathrm{var}(aX + Y) &= a^2 \sigma_X^2 + \sigma_Y^2 + 2a\sigma_{XY} \\
&= (-\sigma_{XY}/\sigma_X^2)^2 \sigma_X^2 + \sigma_Y^2 + 2(-\sigma_{XY}/\sigma_X^2)\sigma_{XY} \\
&= \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2.
\end{aligned}
\tag{2.52}
$$

Because $\mathrm{var}(aX + Y)$ is a variance, it cannot be negative, so from the final line of Equation (2.52), it must be that $\sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2 \geq 0$. Rearranging this inequality yields

$$
\sigma_{XY}^2 \leq \sigma_X^2 \sigma_Y^2 \text{ (covariance inequality)}.
\tag{2.53}
$$

The covariance inequality implies that $\sigma_{XY}^2/(\sigma_X^2\sigma_Y^2) \leq 1$ or, equivalently, $|\sigma_{XY}/(\sigma_X\sigma_Y)| \leq 1$, which (using the definition of the correlation) proves the correlation inequality, $|\mathrm{corr}(X\ Y)| \leq 1$.

APPENDIX

2.2 The Conditional Mean as the Minimum Mean Squared Error Predictor

At a general level, the statistical prediction problem is, how does one best use the information in a random variable X to predict the value of another random variable Y?

To answer to this question, we must first make precise mathematically what it means for one prediction to be better than another. A common way to do so is to consider the cost of making a prediction error. This cost, which is called the prediction loss, depends on the magnitude of the prediction error. For example, if your job is to predict sales so that a production supervisor can develop a production schedule, being off by a small amount is unlikely to inconvenience customers or to disrupt the production process. But if you are off by a large amount and production is set far too low, your company might lose customers who need to wait a long time to receive a product they order, or if production is far too high, the company will have costly excess inventory on its hands. In either case, a large prediction error can be disproportionately more costly than a small one.

One way to make this logic precise is to let the cost of a prediction error depend on the square of that error, so an error twice as large is four times as costly. Specifically, suppose that your prediction of Y, given the random variable X, is $g(X)$. The prediction error is $Y - g(X)$, and the quadratic loss associated with this prediction is,

$$Loss = E\{[Y - g(X)]^2\}. \tag{2.54}$$

We now show that, of all possible functions $g(X)$, the loss in Equation (2.54) is minimized by $g(X) = E(Y|X)$. We show this result using discrete random variables, however this result extends to continuous random variables. The proof here uses calculus; Exercise 2.27 works through a non-calculus proof of this result.

First consider the simpler problem of finding a number, m, that minimizies $E[(Y - m)^2]$. From the definition of the expectation, $E[(Y - m)^2] = \sum_{i=1}^{k}(Y_i - m)^2 p_i$. To find the value of m that minimizes $E[(Y - m)^2]$, take the derivative of $\sum_{i=1}^{k}(Y_i - m)^2 p_i$ with respect to m and set it to zero:

$$\frac{d}{dm}\sum_{i=1}^{k}(Y_i - m)^2 p_i = -2\sum_{i=1}^{k}(Y_i - m)p_i = -2\left(\sum_{i=1}^{k}Y_i p_i - m\sum_{i=1}^{k}p_i\right)$$

$$= -2\left(\sum_{i=1}^{k}Y_i p_i - m\right) = 0, \tag{2.55}$$

where the final equality uses the fact that probabilities sum to 1. It follow from the final equality in Equation (2.55) that the squared error prediction loss is minimized by $m = \sum_{i=1}^{k}Y_i p_i = E(Y)$, that is, by setting m equal to the mean of Y.

To find the predictor $g(X)$ that minimizes the loss in Equation (2.54), use the law of iterated expectations to write that loss as, $Loss = E\{[Y - g(X)]^2\} = E(E\{[Y - g(X)]^2|X\})$. Thus, if the function $g(X)$ minimizes $E\{[Y - g(X)]^2|X = x\}$ for each value of x, it minimizes the loss in Equation (2.54). But for a fixed value $X = x$, $g(X) = g(x)$ is a fixed number, so this problem is the same as the one just solved, and the loss is minimized by choosing $g(x)$ to be the mean of Y, given $X = x$. This is true for every value of x. Thus the squared error loss in Equation (2.54) is minimzed by $g(X) = E(Y|X)$.

3

Review of Statistics

Statistics is the science of using data to learn about the world around us. Statistical tools help us answer questions about unknown characteristics of distributions in populations of interest. For example, what is the mean of the distribution of earnings of recent college graduates? Do mean earnings differ for men and women and, if so, by how much?

These questions relate to the distribution of earnings in the population of workers. One way to answer these questions would be to perform an exhaustive survey of the population of workers, measuring the earnings of each worker and thus finding the population distribution of earnings. In practice, however, such a comprehensive survey would be extremely expensive. The only comprehensive survey of the U.S. population is the decennial census, which cost $13 billion to carry out in 2010. The process of designing the census forms, managing and conducting the surveys, and compiling and analyzing the data takes ten years. Despite this extraordinary commitment, many members of the population slip through the cracks and are not surveyed. Thus a different, more practical approach is needed.

The key insight of statistics is that one can learn about a population distribution by selecting a random sample from that population. Rather than survey the entire U.S. population, we might survey, say, 1000 members of the population, selected at random by simple random sampling. Using statistical methods, we can use this sample to reach tentative conclusions—to draw statistical inferences—about characteristics of the full population.

Three types of statistical methods are used throughout econometrics: estimation, hypothesis testing, and confidence intervals. Estimation entails computing a "best guess" numerical value for an unknown characteristic of a population distribution, such as its mean, from a sample of data. Hypothesis testing entails formulating a specific hypothesis about the population and then using sample evidence to decide whether it is true. Confidence intervals use a set of data to estimate an interval or range for an unknown population characteristic. Sections 3.1, 3.2, and 3.3 review estimation, hypothesis testing, and confidence intervals in the context of statistical inference about an unknown population mean.

Most of the interesting questions in economics involve relationships between two or more variables or comparisons between different populations. For example, is there a gap between the mean earnings for male and female recent college graduates? In Section 3.4, the methods for learning about the mean of a single population in Sections 3.1 through 3.3 are extended to compare means in two different populations.

Section 3.5 discusses how the methods for comparing the means of two populations can be used to estimate causal effects in experiments. Sections 3.2 through 3.5 focus on the use of the normal distribution for performing hypothesis tests and for constructing confidence intervals when the sample size is large. In some special circumstances, hypothesis tests and confidence intervals can be based on the Student t distribution instead of the normal distribution; these special circumstances are discussed in Section 3.6. The chapter concludes with a discussion of the sample correlation and scatterplots in Section 3.7.

3.1 Estimation of the Population Mean

Suppose you want to know the mean value of Y (that is, μ_Y) in a population, such as the mean earnings of women recently graduated from college. A natural way to estimate this mean is to compute the sample average $\overline{Y}$ from a sample of n independently and identically distributed (i.i.d.) observations, $Y_1, \ldots, Y_n$ (recall that $Y_1, \ldots, Y_n$ are i.i.d. if they are collected by simple random sampling). This section discusses estimation of μ_Y and the properties of $\overline{Y}$ as an estimator of μ_Y.

Estimators and Their Properties

Estimators. The sample average $\overline{Y}$ is a natural way to estimate μ_Y, but it is not the only way. For example, another way to estimate μ_Y is simply to use the first observation, Y_1. Both $\overline{Y}$ and Y_1 are functions of the data that are designed to estimate μ_Y; using the terminology in Key Concept 3.1, both are estimators of μ_Y. When evaluated in repeated samples, $\overline{Y}$ and Y_1 take on different values (they produce different estimates) from one sample to the next. Thus the estimators $\overline{Y}$ and Y_1 both have sampling distributions. There are, in fact, many estimators of μ_Y, of which $\overline{Y}$ and Y_1 are two examples.

There are many possible estimators, so what makes one estimator "better" than another? Because estimators are random variables, this question can be phrased more precisely: What are desirable characteristics of the sampling distribution of an

---

KEY CONCEPT

3.1

Estimators and Estimates

An **estimator** is a function of a sample of data to be drawn randomly from a population. An **estimate** is the numerical value of the estimator when it is actually computed using data from a specific sample. An estimator is a random variable because of randomness in selecting the sample, while an estimate is a nonrandom number.

estimator? In general, we would like an estimator that gets as close as possible to the unknown true value, at least in some average sense; in other words, we would like the sampling distribution of an estimator to be as tightly centered on the unknown value as possible. This observation leads to three specific desirable characteristics of an estimator: unbiasedness (a lack of bias), consistency, and efficiency.

Unbiasedness.   Suppose you evaluate an estimator many times over repeated randomly drawn samples. It is reasonable to hope that, on average, you would get the right answer. Thus a desirable property of an estimator is that the mean of its sampling distribution equals μ_Y; if so, the estimator is said to be unbiased.

　　To state this concept mathematically, let $\hat{\mu}_Y$ denote some estimator of μ_Y, such as $\overline{Y}$ or Y_1. [The caret (^) notation will be used throughout this text to denote an estimator, so $\hat{\mu}_Y$ is an estimator of μ_Y.] The estimator $\hat{\mu}_Y$ is unbiased if $E(\hat{\mu}_Y) = \mu_Y$, where $E(\hat{\mu}_Y)$ is the mean of the sampling distribution of $\hat{\mu}_Y$; otherwise, $\hat{\mu}_Y$ is biased.

Consistency.   Another desirable property of an estimator μ_Y is that when the sample size is large, the uncertainty about the value of μ_Y arising from random variations in the sample is very small. Stated more precisely, a desirable property of $\hat{\mu}_Y$ is that the probability that it is within a small interval of the true value μ_Y approaches 1 as the sample size increases; that is, $\hat{\mu}_Y$ is consistent for μ_Y (Key Concept 2.6).

Variance and efficiency.   Suppose you have two candidate estimators, $\hat{\mu}_Y$ and $\tilde{\mu}_Y$, both of which are unbiased. How might you choose between them? One way to do so is to choose the estimator with the tightest sampling distribution. This suggests choosing between $\hat{\mu}_Y$ and $\tilde{\mu}_Y$ by picking the estimator with the smallest variance. If $\hat{\mu}_Y$ has a smaller variance than $\tilde{\mu}_Y$, then $\hat{\mu}_Y$ is said to be more efficient than $\tilde{\mu}_Y$. The terminology "efficiency" stems from the notion that if $\hat{\mu}_Y$ has a smaller variance than $\tilde{\mu}_Y$, then it uses the information in the data more efficiently than does $\tilde{\mu}_Y$.

　　Bias, **consistency**, and **efficiency** are summarized in Key Concept 3.2.

Bias, Consistency, and Efficiency

KEY CONCEPT

3.2

Let $\hat{\mu}_Y$ be an estimator of μ_Y. Then:

- The *bias* of $\hat{\mu}_Y$ is $E(\hat{\mu}_Y) - \mu_Y$.
- $\hat{\mu}_Y$ is an *unbiased estimator* of μ_Y if $E(\hat{\mu}_Y) = \mu_Y$.
- $\hat{\mu}_Y$ is a *consistent estimator* of μ_Y if $\hat{\mu}_Y \xrightarrow{p} \mu_Y$.
- Let $\tilde{\mu}_Y$ be another estimator of μ_Y, and suppose that both $\hat{\mu}_Y$ and $\tilde{\mu}_Y$ are unbiased. Then $\hat{\mu}_Y$ is said to be more *efficient* than $\tilde{\mu}_Y$ if $\mathrm{var}(\hat{\mu}_Y) < \mathrm{var}(\tilde{\mu}_Y)$.

Properties of $\overline{Y}$

How does $\overline{Y}$ fare as an estimator of μ_Y when judged by the three criteria of bias, consistency, and efficiency?

Bias and consistency. The sampling distribution of $\overline{Y}$ has already been examined in Sections 2.5 and 2.6. As shown in Section 2.5, $E(\overline{Y}) = \mu_Y$, so $\overline{Y}$ is an unbiased estimator of μ_Y. Similarly, the law of large numbers (Key Concept 2.6) states that $\overline{Y} \xrightarrow{p} \mu_Y$; that is, $\overline{Y}$ is consistent.

Efficiency. What can be said about the efficiency of $\overline{Y}$? Because efficiency entails a comparison of estimators, we need to specify the estimator or estimators to which $\overline{Y}$ is to be compared.

We start by comparing the efficiency of $\overline{Y}$ to the estimator Y_1. Because $Y_1, \ldots, Y_n$ are i.i.d., the mean of the sampling distribution of Y_1 is $E(Y_1) = \mu_Y$; thus Y_1 is an unbiased estimator of μ_Y. Its variance is $\text{var}(Y_1) = \sigma_Y^2$. From Section 2.5, the variance of $\overline{Y}$ is σ_Y^2/n. Thus, for $n \geq 2$, the variance of $\overline{Y}$ is less than the variance of Y_1; that is, $\overline{Y}$ is a more efficient estimator than Y_1, so, according to the criterion of efficiency, $\overline{Y}$ should be used instead of Y_1. The estimator Y_1 might strike you as an obviously poor estimator—why would you go to the trouble of collecting a sample of n observations only to throw away all but the first?—and the concept of efficiency provides a formal way to show that $\overline{Y}$ is a more desirable estimator than Y_1.

What about a less obviously poor estimator? Consider the weighted average in which the observations are alternately weighted by $\frac{1}{2}$ and $\frac{3}{2}$:

$$\widetilde{Y} = \frac{1}{n}\left(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \frac{1}{2}Y_3 + \frac{3}{2}Y_4 + \cdots + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n\right), \tag{3.1}$$

where the number of observations n is assumed to be even for convenience. The mean of $\widetilde{Y}$ is μ_Y, and its variance is $\text{var}(\widetilde{Y}) = 1.25\, \sigma_Y^2/n$ (Exercise 3.11). Thus $\widetilde{Y}$ is unbiased, and because $\text{var}(\widetilde{Y}) \to 0$ as $n \to \infty$, $\widetilde{Y}$ is consistent. However, $\widetilde{Y}$ has a larger variance than $\overline{Y}$. Thus $\overline{Y}$ is more efficient than $\widetilde{Y}$.

The estimators $\overline{Y}$, Y_1, and $\widetilde{Y}$ have a common mathematical structure: They are weighted averages of $Y_1, \ldots, Y_n$. The comparisons in the previous two paragraphs show that the weighted averages Y_1 and $\widetilde{Y}$ have larger variances than $\overline{Y}$. In fact, these conclusions reflect a more general result: $\overline{Y}$ is the most efficient estimator of *all* unbiased estimators that are weighted averages of $Y_1, \ldots, Y_n$. Said differently, $\overline{Y}$ is the **B**est **L**inear **U**nbiased **E**stimator **(BLUE)**; that is, it is the most efficient (best) estimator among all estimators that are unbiased and are linear functions of $Y_1, \ldots, Y_n$. This result is stated in Key Concept 3.3 and is proved in Chapter 5.

Efficiency of $\overline{Y}$: $\overline{Y}$ Is BLUE

Let $\hat{\mu}_Y$ be an estimator of μ_Y that is a weighted average of $Y_1, \ldots, Y_n$; that is, $\hat{\mu}_Y = (1/n)\sum_{i=1}^{n} a_i Y_i$, where $a_1, \ldots, a_n$ are nonrandom constants. If $\hat{\mu}_Y$ is unbiased, then $\text{var}(\overline{Y}) < \text{var}(\hat{\mu}_Y)$ unless $\hat{\mu}_Y = \overline{Y}$. Thus $\overline{Y}$ is the Best Linear Unbiased Estimator (BLUE); that is, $\overline{Y}$ is the most efficient estimator of μ_Y among all unbiased estimators that are weighted averages of $Y_1, \ldots, Y_n$.

$\overline{Y}$ is the least squares estimator of μ_Y. The sample average $\overline{Y}$ provides the best fit to the data in the sense that the average squared differences between the observations and $\overline{Y}$ are the smallest of all possible estimators.

Consider the problem of finding the estimator m that minimizes

$$\sum_{i=1}^{n} (Y_i - m)^2, \tag{3.2}$$

which is a measure of the total squared gap or distance between the estimator m and the sample points. Because m is an estimator of $E(Y)$, you can think of it as a prediction of the value of Y_i, so the gap $Y_i - m$ can be thought of as a prediction mistake. The sum of squared gaps in Expression (3.2) can be thought of as the sum of squared prediction mistakes.

The estimator m that minimizes the sum of squared gaps $Y_i - m$ in Expression (3.2) is called the **least squares estimator**. One can imagine using trial and error to solve the least squares problem: Try many values of m until you are satisfied that you have the value that makes Expression (3.2) as small as possible. Alternatively, as is done in Appendix 3.2, you can use algebra or calculus to show that choosing $m = \overline{Y}$ minimizes the sum of squared gaps in Expression (3.2), so that $\overline{Y}$ is the least squares estimator of μ_Y.

The Importance of Random Sampling

We have assumed that $Y_1, \ldots, Y_n$ are i.i.d. draws, such as those that would be obtained from simple random sampling. This assumption is important because non-random sampling can result in $\overline{Y}$ being biased. Suppose that to estimate the monthly national unemployment rate, a statistical agency adopts a sampling scheme in which interviewers survey working-age adults sitting in city parks at 10 a.m. on the second Wednesday of the month. Because most employed people are at work at that hour (not sitting in the park!), the unemployed are overly represented in the sample, and an estimate of the unemployment rate based on this sampling plan would be biased. This bias arises because this sampling scheme overrepresents, or oversamples, the unemployed members of the population. This example is fictitious, but the

Landon Wins!

S hortly before the 1936 U.S. presidential election, the *Literary Gazette* published a poll indicating that Alf M. Landon would defeat the incumbent, Franklin D. Roosevelt, by a landslide—57% to 43%. The *Gazette* was right that the election was a landslide, but it was wrong about the winner: Roosevelt won by 59% to 41%!

How could the *Gazette* have made such a big mistake? The *Gazette*'s sample was chosen from telephone records and automobile registration files. But in 1936, many households did not have cars or telephones, and those that did tended to be richer—and were also more likely to be Republican. Because the telephone survey did not sample randomly from the population but instead undersampled Democrats, the estimator was biased, and the *Gazette* made an embarrassing mistake.

Political polls have since become much more sophisticated and adjust for sampling bias, but they still can make mistakes. In the 2016 presidential election of Hillary Clinton vs. Donald Trump, polls correctly predicted that Clinton would win the national vote, but state-level polls incorrectly predicted that she would win the Electoral College—which Trump won by a large margin (304 vs. 227 Electoral College votes). According to the American Association for Public Opinion Research (2017), one reason that state-level polls got it wrong was that they failed to adjust for overrepresentation of college graduates among the respondents—a mistake pollsters are unlikely to make in the future.

"Landon Wins!" box gives a real-world example of biases introduced by sampling that is not entirely random.

It is important to design sample selection schemes in a way that minimizes bias. Appendix 3.1 includes a discussion of what the Bureau of Labor Statistics actually does when it conducts the U.S. Current Population Survey (CPS), the survey it uses to estimate the monthly U.S. unemployment rate.

3.2 Hypothesis Tests Concerning the Population Mean

Many hypotheses about the world around us can be phrased as yes/no questions. Do the mean hourly earnings of recent U.S. college graduates equal $20 per hour? Are mean earnings the same for male and female college graduates? Both these questions embody specific hypotheses about the population distribution of earnings. The statistical challenge is to answer these questions based on a sample of evidence. This section describes **hypothesis tests** concerning the population mean (Does the population mean of hourly earnings equal $20?). Hypothesis tests involving two populations (Are mean earnings the same for men and women?) are taken up in Section 3.4.

Null and Alternative Hypotheses

The starting point of statistical hypotheses testing is specifying the hypothesis to be tested, called the **null hypothesis**. Hypothesis testing entails using data to compare the null hypothesis to a second hypothesis, called the **alternative hypothesis**, that holds if the null does not.

The null hypothesis is that the population mean, $E(Y)$, takes on a specific value, denoted $\mu_{Y,0}$. The null hypothesis is denoted H_0 and thus is

$$H_0: E(Y) = \mu_{Y,0}. \tag{3.3}$$

For example, the conjecture that, on average in the population, college graduates earn \$20 per hour constitutes a null hypothesis about the population distribution of hourly earnings. Stated mathematically, if Y is the hourly earnings of a randomly selected recent college graduate, then the null hypothesis is that $E(Y) = 20$; that is, $\mu_{Y,0} = 20$ in Equation (3.3).

The alternative hypothesis specifies what is true if the null hypothesis is not. The most general alternative hypothesis is that $E(Y) \neq \mu_{Y,0}$, which is called a **two-sided alternative hypothesis** because it allows $E(Y)$ to be either less than or greater than $\mu_{Y,0}$. The two-sided alternative is written as

$$H_1: E(Y) \neq \mu_{Y,0} \text{ (two-sided alternative).} \tag{3.4}$$

One-sided alternatives are also possible, and these are discussed later in this section.

The problem facing the statistician is to use the evidence in a randomly selected sample of data to decide whether to accept the null hypothesis H_0 or to reject it in favor of the alternative hypothesis H_1. If the null hypothesis is "accepted," this does not mean that the statistician declares it to be true; rather, it is accepted tentatively with the recognition that it might be rejected later based on additional evidence. For this reason, statistical hypothesis testing can be posed as either rejecting the null hypothesis or failing to do so.

The *p*-Value

In any given sample, the sample average $\overline{Y}$ will rarely be exactly equal to the hypothesized value $\mu_{Y,0}$. Differences between $\overline{Y}$ and $\mu_{Y,0}$ can arise because the true mean, in fact, does not equal $\mu_{Y,0}$ (the null hypothesis is false) or because the true mean equals $\mu_{Y,0}$ (the null hypothesis is true) but $\overline{Y}$ differs from $\mu_{Y,0}$ because of random sampling. It is impossible to distinguish between these two possibilities with certainty. Although a sample of data cannot provide conclusive evidence about the null hypothesis, it is possible to do a probabilistic calculation that permits testing the null hypothesis in a way that accounts for sampling uncertainty. This calculation involves using the data to compute the *p*-value of the null hypothesis.

The *p*-value, also called the **significance probability**, is the probability of drawing a statistic at least as adverse to the null hypothesis as the one you actually computed in your sample, assuming the null hypothesis is correct. In the case at hand, the *p*-value is the probability of drawing $\overline{Y}$ at least as far in the tails of its distribution under the null hypothesis as the sample average you actually computed.

For example, suppose that, in your sample of recent college graduates, the average wage is $22.64. The *p*-value is the probability of observing a value of $\overline{Y}$ at least as different from $20 (the population mean under the null hypothesis) as the observed value of $22.64 by pure random sampling variation, assuming that the null hypothesis is true. If this *p*-value is small (say, 0.1%), then it is very unlikely that this sample would have been drawn if the null hypothesis is true; thus it is reasonable to conclude that the null hypothesis is not true. By contrast, if this *p*-value is large (say, 40%), then it is quite likely that the observed sample average of $22.64 could have arisen just by random sampling variation if the null hypothesis is true; accordingly, the evidence against the null hypothesis is weak in this probabilistic sense, and it is reasonable not to reject the null hypothesis.

To state the definition of the *p*-value mathematically, let $\overline{Y}^{act}$ denote the value of the sample average actually computed in the data set at hand, and let $\Pr_{H_0}$ denote the probability computed under the null hypothesis (that is, computed assuming that $E(Y) = \mu_{Y,0}$). The *p*-value is

$$p\text{-value} = \Pr_{H_0}[\,|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|\,]. \tag{3.5}$$

That is, the *p*-value is the area in the tails of the distribution of $\overline{Y}$ under the null hypothesis beyond $\mu_{Y,0} \pm |\overline{Y}^{act} - \mu_{Y,0}|$. If the *p*-value is large, then the observed value $\overline{Y}^{act}$ is consistent with the null hypothesis, but if the *p*-value is small, it is not.

To compute the *p*-value, it is necessary to know the sampling distribution of $\overline{Y}$ under the null hypothesis. As discussed in Section 2.6, when the sample size is small, this distribution is complicated. However, according to the central limit theorem, when the sample size is large, the sampling distribution of $\overline{Y}$ is well approximated by a normal distribution. Under the null hypothesis the mean of this normal distribution is $\mu_{Y,0}$, so under the null hypothesis $\overline{Y}$ is distributed $N(\mu_{Y,0}, \sigma_{\overline{Y}}^2)$, where $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$. This large-sample normal approximation makes it possible to compute the *p*-value without needing to know the population distribution of Y, as long as the sample size is large. The details of the calculation, however, depend on whether σ_Y^2 is known.

Calculating the *p*-Value When σ_Y Is Known

The calculation of the *p*-value when σ_Y is known is summarized in Figure 3.1. If the sample size is large, then under the null hypothesis the sampling distribution of $\overline{Y}$ is $N(\mu_{Y,0}, \sigma_{\overline{Y}}^2)$, where $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$. Thus, under the null hypothesis, the standardized version of $\overline{Y}$, $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$, has a standard normal distribution. The *p*-value is the

FIGURE 3.1   Calculating a *p*-value

The *p*-value is the probability of drawing a value of $\overline{Y}$ that differs from $\mu_{Y,0}$ by at least as much as $\overline{Y}^{\text{act}}$. In large samples, $\overline{Y}$ is distributed $N(\mu_{Y,0}, \sigma_{\overline{Y}}^2)$ under the null hypothesis, so $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$ is distributed $N(0,1)$. Thus the *p*-value is the shaded standard normal tail probability outside $\pm |(\overline{Y}^{\text{act}} - \mu_{Y,0})/\sigma_{\overline{Y}}|$.

The *p*-value is the shaded area in the graph
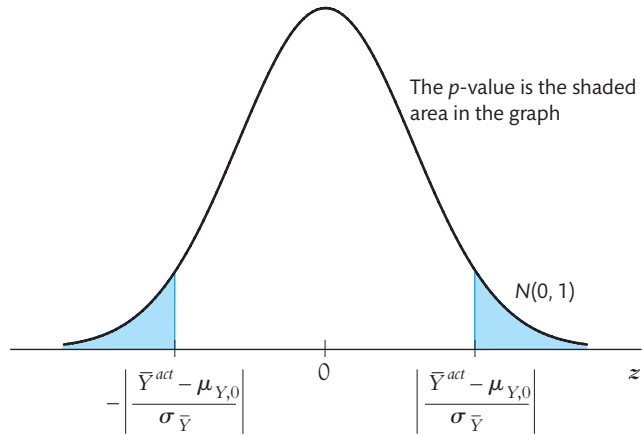
probability of obtaining a value of $\overline{Y}$ farther from $\mu_{Y,0}$ than $\overline{Y}^{\text{act}}$ under the null hypothesis or, equivalently, it is the probability of obtaining $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$ greater than $(\overline{Y}^{\text{act}} - \mu_{Y,0})/\sigma_{\overline{Y}}$ in absolute value. This probability is the shaded area shown in Figure 3.1. Written mathematically, the shaded tail probability in Figure 3.1 (that is, the *p*-value) is

$$p\text{-value} = \Pr_{H_0}\left(\left|\frac{\overline{Y} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right| > \left|\frac{\overline{Y}^{\text{act}} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right|\right) = 2\Phi\left(-\left|\frac{\overline{Y}^{\text{act}} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right|\right), \qquad (3.6)$$

where Φ is the standard normal cumulative distribution function. That is, the *p*-value is the area in the tails of a standard normal distribution outside $\pm |\overline{Y}^{\text{act}} - \mu_{Y,0}|/\sigma_{\overline{Y}}$.

The formula for the *p*-value in Equation (3.6) depends on the variance of the population distribution, σ_Y^2. In practice, this variance is typically unknown. [An exception is when Y_i is binary, so that its distribution is Bernoulli, in which case the variance is determined by the null hypothesis; see Equation (2.7) and Exercise 3.2.] Because in general σ_Y^2 must be estimated before the *p*-value can be computed, we now turn to the problem of estimating σ_Y^2.

The Sample Variance, Sample Standard Deviation, and Standard Error

The sample variance, s_Y^2, is an estimator of the population variance, σ_Y^2; the sample standard deviation, s_Y, is an estimator of the population standard deviation, σ_Y; and the standard error of the sample average, $\overline{Y}$, is an estimator of the standard deviation of the sampling distribution of $\overline{Y}$.

The sample variance and standard deviation. The **sample variance**, s_Y^2, is

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{3.7}$$

The **sample standard deviation**, s_Y, is the square root of the sample variance.

The formula for the sample variance is much like the formula for the population variance. The population variance, $E(Y - \mu_Y)^2$, is the average value of $(Y - \mu_Y)^2$ in the population distribution. Similarly, the sample variance is the sample average of $(Y_i - \mu_Y)^2$, $i = 1, \cdots, n$, with two modifications: First, μ_Y is replaced by $\overline{Y}$, and second, the average uses the divisor $n - 1$ instead of n.

The reason for the first modification—replacing μ_Y by $\overline{Y}$—is that μ_Y is unknown and thus must be estimated; the natural estimator of μ_Y is $\overline{Y}$. The reason for the second modification—dividing by $n - 1$ instead of by n—is that estimating μ_Y by $\overline{Y}$ introduces a small downward bias in $(Y_i - \overline{Y})^2$. Specifically, as is shown in Exercise 3.18, $E[(Y_i - \overline{Y})^2] = [(n-1)/n]\sigma_Y^2$. Thus $E\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = nE[(Y_i - \overline{Y})^2] = (n-1)\sigma_Y^2$. Dividing by $n - 1$ in Equation (3.7) instead of n corrects for this small downward bias, and as a result s_Y^2 is unbiased.

Dividing by $n - 1$ in Equation (3.7) instead of n is called a **degrees of freedom** correction: Estimating the mean uses up some of the information—that is, uses up 1 "degree of freedom"—in the data, so that only $n - 1$ degrees of freedom remain.

Consistency of the sample variance. The sample variance is a consistent estimator of the population variance:

$$s_Y^2 \xrightarrow{p} \sigma_Y^2. \tag{3.8}$$

In other words, the sample variance is close to the population variance with high probability when n is large.

The result in Equation (3.9) is proven in Appendix 3.3 under the assumptions that $Y_1, \ldots, Y_n$ are i.i.d. and Y_i has a finite fourth moment; that is, $E(Y_i^4) < \infty$. Intuitively, the reason that s_Y^2 is consistent is that it is a sample average, so s_Y^2 obeys the law of large numbers. For s_Y^2 to obey the law of large numbers in Key Concept 2.6, $(Y_i - \mu_Y)^2$ must have finite variance, which in turn means that $E(Y_i^4)$ must be finite; in other words, Y_i must have a finite fourth moment.

The standard error of $\overline{Y}$. Because the standard deviation of the sampling distribution of $\overline{Y}$ is $\sigma_{\overline{Y}} = \sigma_Y/\sqrt{n}$, Equation (3.9) justifies using $s_Y/\sqrt{n}$ as an estimator of $\sigma_{\overline{Y}}$. The estimator of $\sigma_{\overline{Y}}$, $s_Y/\sqrt{n}$, is called the **standard error of $\overline{Y}$** and is denoted $SE(\overline{Y})$ or $\hat{\sigma}_{\overline{Y}}$. The standard error of $\overline{Y}$ is summarized as in Key Concept 3.4.

When $Y_1, \ldots, Y_n$ are i.i.d. draws from a Bernoulli distribution with success probability p, the formula for the variance of $\overline{Y}$ simplifies to $p(1-p)/n$ (see Exercise 3.2). The formula for the standard error also takes on a simple form that depends only on $\overline{Y}$ and n: $SE(\overline{Y}) = \sqrt{\overline{Y}(1-\overline{Y})/n}$.

<div style="border:1px solid; padding:10px;">

The Standard Error of $\overline{Y}$

The standard error of $\overline{Y}$ is an estimator of the standard deviation of $\overline{Y}$. The standard error of $\overline{Y}$ is denoted $SE(\overline{Y})$ or $\hat{\sigma}_{\overline{Y}}$. When $Y_1, \ldots, Y_n$ are i.i.d.,

$$SE(\overline{Y}) = \hat{\sigma}_{\overline{Y}} = s_Y/\sqrt{n}. \tag{3.9}$$

</div>

Calculating the p-Value When σ_Y Is Unknown

Because s_Y^2 is a consistent estimator of σ_Y^2, the p-value can be computed by replacing $\sigma_{\overline{Y}}$ in Equation (3.6) by the standard error, $SE(\overline{Y}) = \hat{\sigma}_{\overline{Y}}$. That is, when σ_Y is unknown and $Y_1, \ldots, Y_n$ are i.i.d., the p-value is calculated using the formula

$$p\text{-value} = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})}\right|\right). \tag{3.10}$$

The t-Statistic

The standardized sample average $(\overline{Y} - \mu_{Y,0})/SE(\overline{Y})$ plays a central role in testing statistical hypotheses and has a special name, the **t-statistic** or **t-ratio**:

$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})}. \tag{3.11}$$

In general, a **test statistic** is a statistic used to perform a hypothesis test. The t-statistic is an important example of a test statistic.

Large-sample distribution of the t-statistic. When n is large, s_Y^2 is close to σ_Y^2 with high probability. Thus the distribution of the t-statistic is approximately the same as the distribution of $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$, which in turn is well approximated by the standard normal distribution when n is large because of the central limit theorem (Key Concept 2.7). Accordingly, under the null hypothesis,

$$t \text{ is approximately distributed } N(0, 1) \text{ for large } n. \tag{3.12}$$

The formula for the p-value in Equation (3.10) can be rewritten in terms of the t-statistic. Let t^{act} denote the value of the t-statistic actually computed:

$$t^{act} = \frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})}. \tag{3.13}$$

Accordingly, when n is large, the p-value can be calculated using

$$p\text{-value} = 2\Phi(-|t^{act}|). \tag{3.14}$$

As a hypothetical example, suppose that a sample of $n = 200$ recent college graduates is used to test the null hypothesis that the mean wage, $E(Y)$, is \$20 per hour. The sample average wage is $\overline{Y}^{act} = \$22.64$, and the sample standard deviation is $s_Y = \$18.14$. Then the standard error of $\overline{Y}$ is $s_Y / \sqrt{n} = 18.14 / \sqrt{200} = 1.28$. The value of the t-statistic is $t^{act} = (22.64 - 20) / 1.28 = 2.06$. From Appendix Table 1, the p-value is $2\Phi(-2.06) = 0.039$, or 3.9%. That is, assuming the null hypothesis to be true, the probability of obtaining a sample average at least as different from the null as the one actually computed is 3.9%.

Hypothesis Testing with a Prespecified Significance Level

When you undertake a statistical hypothesis test, you can make two types of mistakes: You can incorrectly reject the null hypothesis when it is true, or you can fail to reject the null hypothesis when it is false. Hypothesis tests can be performed without computing the p-value if you are willing to specify in advance the probability you are willing to tolerate of making the first kind of mistake—that is, of incorrectly rejecting the null hypothesis when it is true. If you choose a prespecified probability of rejecting the null hypothesis when it is true (for example, 5%), then you will reject the null hypothesis if and only if the p-value is less than 0.05. This approach gives preferential treatment to the null hypothesis, but in many practical situations, this preferential treatment is appropriate.

Hypothesis tests using a fixed significance level. Suppose it has been decided that the hypothesis will be rejected if the p-value is less than 5%. Because the area under the tails of the standard normal distribution outside ± 1.96 is 5%, this gives a simple rule:

$$\text{Reject } H_0 \text{ if } |t^{act}| > 1.96. \tag{3.15}$$

That is, reject if the absolute value of the t-statistic computed from the sample is greater than 1.96. If n is large enough, then under the null hypothesis the t-statistic has a $N(0, 1)$ distribution. Thus the probability of erroneously rejecting the null hypothesis (rejecting the null hypothesis when it is, in fact, true) is 5%.

This framework for testing statistical hypotheses has some specialized terminology, summarized in Key Concept 3.5. The significance level of the test in Equation (3.15) is 5%, the critical value of this two-sided test is 1.96, and the rejection region is the values of the t-statistic outside ± 1.96. If the test rejects at the 5% significance level, the population mean μ_Y is said to be statistically significantly different from $\mu_{Y,0}$ at the 5% significance level.

Testing hypotheses using a prespecified significance level does not require computing p-values. In the previous example of testing the hypothesis that the mean earnings of recent college graduates is \$20 per hour, the t-statistic was 2.06. This value exceeds 1.96, so the hypothesis is rejected at the 5% level. Although performing the

The Terminology of Hypothesis Testing

A statistical hypothesis test can make two types of mistakes: a **type I error**, in which the null hypothesis is rejected when in fact it is true; and a **type II error**, in which the null hypothesis is not rejected when in fact it is false. The prespecified rejection probability of a statistical hypothesis test when the null hypothesis is true—that is, the prespecified probability of a type I error—is the **significance level** of the test. The **critical value** of the test statistic is the value of the statistic for which the test just rejects the null hypothesis at the given significance level. The set of values of the test statistic for which the test rejects the null hypothesis is the **rejection region**, and the set of values of the test statistic for which it does not reject the null hypothesis is the **acceptance region**. The probability that the test actually incorrectly rejects the null hypothesis when it is true is the **size of the test**, and the probability that the test correctly rejects the null hypothesis when the alternative is true is the **power of the test**.

The **p-value** is the probability of obtaining a test statistic, by random sampling variation, at least as adverse to the null hypothesis value as is the statistic actually observed, assuming that the null hypothesis is correct. Equivalently, the *p*-value is the smallest significance level at which you can reject the null hypothesis.

test with a 5% significance level is easy, reporting only whether the null hypothesis is rejected at a prespecified significance level conveys less information than reporting the *p*-value.

What significance level should you use in practice? This is a question of active debate. Historically, statisticians and econometricians have used a 5% significance level. If you were to test many statistical hypotheses at the 5% level, you would incorrectly reject the null, on average, once in 20 cases. Whether this is a small number depends on how you look at it. If only a small fraction of all null hypotheses tested are, in fact, false, then among those tests that reject, the probability of the null actually being false can be small (Exercise 3.22). This probability—the fraction of incorrect rejections among all rejections—is called the false positive rate. The false positive rate can have great practical importance. For example, for newly reported statistically significant findings of effective medical treatments, it is the fraction for which the treatment is in fact ineffective. Concern that the false positive rate can be high when the 5% significance level is used has led some statisticians to recommend using instead a 0.5% significance level when reporting new results (Benjamin et al., 2017). Similar concerns can apply in a legal setting, where justice might aim to keep the fraction of false convictions low. Using a 0.5% significance level leads to two-sided rejection when the *t*-statistic exceeds 2.81 in absolute value. In such cases, a *p*-value

Testing the Hypothesis $E(Y) = \mu_{Y,0}$ Against the Alternative $E(Y) \neq \mu_{Y,0}$

1. Compute the standard error of $\overline{Y}$, $SE(\overline{Y})$ [Equation (3.8)].
2. Compute the t-statistic [Equation (3.13)].
3. Compute the p-value [Equation (3.14)]. Reject the hypothesis at the 5% significance level if the p-value is less than 0.05 (equivalently, if $|t^{act}| > 1.96$).

between 0.05 and 0.005 can be viewed as suggestive, but not conclusive, evidence against the null that merits further investigation.

The choice of significance level requires judgment and depends on the application. In some economic applications, a false positive might be less of a problem than in a medical context, where the false positive could lead to patients receiving ineffective treatments. In such cases, a 5% significance level could be appropriate.

Whatever the significance level, it is important to keep in mind that p-values are designed for tests of a null hypothesis, so they, like t-statistics, are useful only when the null hypothesis itself is of interest. This section uses the example of earnings. Even though many interns are unpaid, nobody thinks that, on average, workers earn nothing at all, so the null hypothesis that earnings are zero is economically uninteresting and not worth testing. In contrast, the null hypothesis that the mean earnings of men and of women are the same is interesting and of societal importance, and that null hypothesis is examined in Section 3.4.

Key Concept 3.6 summarizes hypothesis tests for the population mean against the two-sided alternative.

One-Sided Alternatives

In some circumstances, the alternative hypothesis might be that the mean exceeds $\mu_{Y,0}$. For example, one hopes that education helps in the labor market, so the relevant alternative to the null hypothesis that earnings are the same for college graduates and non–college graduates is not just that their earnings differ, but rather that graduates earn more than nongraduates. This is called a **one-sided alternative hypothesis** and can be written

$$H_1: E(Y) > \mu_{Y,0} \text{ (one-sided alternative)}. \qquad (3.16)$$

The general approach to computing p-values and to hypothesis testing is the same for one-sided alternatives as it is for two-sided alternatives, with the modification that only large positive values of the t-statistic reject the null hypothesis rather

than values that are large in absolute value. Specifically, to test the one-sided hypothesis in Equation (3.16), construct the t-statistic in Equation (3.13). The p-value is the area under the standard normal distribution to the right of the calculated t-statistic. That is, the p-value, based on the $N(0, 1)$ approximation to the distribution of the t-statistic, is

$$p\text{-value} = \Pr_{H_0}(Z > t^{act}) = 1 - \Phi(t^{act}). \qquad (3.17)$$

The $N(0, 1)$ critical value for a one-sided test with a 5% significance level is 1.64. The rejection region for this test is all values of the t-statistic exceeding 1.64.

The one-sided hypothesis in Equation (3.16) concerns values of μ_Y exceeding $\mu_{Y,0}$. If instead the alternative hypothesis is that $E(Y) < \mu_{Y,0}$, then the discussion of the previous paragraph applies except that the signs are switched; for example, the 5% rejection region consists of values of the t-statistic less than -1.64.

3.3   Confidence Intervals for the Population Mean

Because of random sampling error, it is impossible to learn the exact value of the population mean of Y using only the information in a sample. However, it is possible to use data from a random sample to construct a set of values that contains the true population mean μ_Y with a certain prespecified probability. Such a set is called a **confidence set**, and the prespecified probability that μ_Y is contained in this set is called the **confidence level**. The confidence set for μ_Y turns out to be all the possible values of the mean between a lower and an upper limit, so that the confidence set is an interval, called a **confidence interval**.

Here is one way to construct a 95% confidence set for the population mean. Begin by picking some arbitrary value for the mean; call it $\mu_{Y,0}$. Test the null hypothesis that $\mu_Y = \mu_{Y,0}$ against the alternative that $\mu_Y \neq \mu_{Y,0}$ by computing the t-statistic; if its absolute value is less than 1.96, this hypothesized value $\mu_{Y,0}$ is not rejected at the 5% level, so write down this nonrejected value $\mu_{Y,0}$. Now pick another arbitrary value of $\mu_{Y,0}$ and test it; if you cannot reject it, write down this value on your list. Do this again and again; indeed, do so for all possible values of the population mean. Continuing this process yields the set of all values of the population mean that cannot be rejected at the 5% level by a two-sided hypothesis test.

This list is useful because it summarizes the set of hypotheses you can and cannot reject (at the 5% level) based on your data: If someone walks up to you with a specific number in mind, you can tell him whether his hypothesis is rejected or not simply by looking up his number on your handy list. A bit of clever reasoning shows that this set of values has a remarkable property: The probability that it contains the true value of the population mean is 95%.

Confidence Intervals for the Population Mean

A 95% two-sided confidence interval for μ_Y is an interval constructed so that it contains the true value of μ_Y in 95% of all possible random samples. When the sample size n is large, 90%, 95%, and 99% confidence intervals for μ_Y are:

90% confidence interval for $\mu_Y = \{\overline{Y} \pm 1.64SE(\overline{Y})\}$,

95% confidence interval for $\mu_Y = \{\overline{Y} \pm 1.96SE(\overline{Y})\}$, and

99% confidence interval for $\mu_Y = \{\overline{Y} \pm 2.58SE(\overline{Y})\}$.

The clever reasoning goes like this: Suppose the true value of μ_Y is 21.5 (although we do not know this). Then $\overline{Y}$ has a normal distribution centered on 21.5, and the t-statistic testing the null hypothesis $\mu_Y = 21.5$ has a $N(0, 1)$ distribution. Thus, if n is large, the probability of rejecting the null hypothesis $\mu_Y = 21.5$ at the 5% level is 5%. But because you tested all possible values of the population mean in constructing your set, in particular you tested the true value, $\mu_Y = 21.5$. In 95% of all samples, you will correctly accept 21.5; this means that in 95% of all samples, your list will contain the true value of μ_Y. Thus the values on your list constitute a 95% confidence set for μ_Y.

This method of constructing a confidence set is impractical, for it requires you to test all possible values of μ_Y as null hypotheses. Fortunately, there is a much easier approach. According to the formula for the t-statistic in Equation (3.13), a trial value of $\mu_{Y,0}$ is rejected at the 5% level if it is more than 1.96 standard errors away from $\overline{Y}$. Thus the set of values of μ_Y that are not rejected at the 5% level consists of those values within $\pm 1.96SE(\overline{Y})$ of $\overline{Y}$; that is, a 95% confidence interval for μ_Y is $\overline{Y} - 1.96SE(\overline{Y}) \leq \mu_Y \leq \overline{Y} + 1.96SE(\overline{Y})$. Key Concept 3.7 summarizes this approach.

As an example, consider the problem of constructing a 95% confidence interval for the mean hourly earnings of recent college graduates using a hypothetical random sample of 200 recent college graduates where $\overline{Y} = \$22.64$ and $SE(\overline{Y}) = 1.28$. The 95% confidence interval for mean hourly earnings is $22.64 \pm 1.96 \times 1.28 = 22.64 \pm 2.51 = (\$20.13, \$25.15)$.

This discussion so far has focused on two-sided confidence intervals. One could instead construct a one-sided confidence interval as the set of values of μ_Y that cannot be rejected by a one-sided hypothesis test. Although one-sided confidence intervals have applications in some branches of statistics, they are uncommon in applied econometric analysis.

Coverage probabilities. The **coverage probability** of a confidence interval for the population mean is the probability, computed over all possible random samples, that it contains the true population mean.

3.4   Comparing Means from Different Populations

Do recent male and female college graduates earn the same amount on average? Answering this question involves comparing the means of two different population distributions. This section summarizes how to test hypotheses and how to construct confidence intervals for the difference in the means from two different populations.

Hypothesis Tests for the Difference Between Two Means

To illustrate a **test for the difference between two means**, let μ_w be the mean hourly earnings in the population of women recently graduated from college, and let μ_m be the population mean for recently graduated men. Consider the null hypothesis that mean earnings for these two populations differ by a certain amount, say, d_0. Then the null hypothesis and the two-sided alternative hypothesis are

$$H_0\colon \mu_m - \mu_w = d_0 \text{ vs. } H_1\colon \mu_m - \mu_w \neq d_0. \qquad (3.18)$$

The null hypothesis that men and women in these populations have the same mean earnings corresponds to H_0 in Equation (3.18) with $d_0 = 0$.

Because these population means are unknown, they must be estimated from samples of men and women. Suppose we have samples of n_m men and n_w women drawn at random from their populations. Let the sample average annual earnings be $\overline{Y}_m$ for men and $\overline{Y}_w$ for women. Then an estimator of $\mu_m - \mu_w$ is $\overline{Y}_m - \overline{Y}_w$.

To test the null hypothesis that $\mu_m - \mu_w = d_0$ using $\overline{Y}_m - \overline{Y}_w$, we need to know the sampling distribution of $\overline{Y}_m - \overline{Y}_w$. Recall that $\overline{Y}_m$ is, according to the central limit theorem, approximately distributed $N(\mu_m, \sigma_m^2/n_m)$, where σ_m^2 is the population variance of earnings for men. Similarly, $\overline{Y}_w$ is approximately distributed $N(\mu_w, \sigma_w^2/n_w)$, where σ_w^2 is the population variance of earnings for women. Also, recall from Section 2.4 that a weighted average of two normal random variables is itself normally distributed. Because $\overline{Y}_m$ and $\overline{Y}_w$ are constructed from different randomly selected samples, they are independent random variables. Thus $\overline{Y}_m - \overline{Y}_w$ is distributed $N[\mu_m - \mu_w, (\sigma_m^2/n_m) + (\sigma_w^2/n_w)]$.

If σ_m^2 and σ_w^2 are known, then this approximate normal distribution can be used to compute p-values for the test of the null hypothesis that $\mu_m - \mu_w = d_0$. In practice, however, these population variances are typically unknown, so they must be estimated. As before, they can be estimated using the sample variances, s_m^2 and s_w^2, where s_m^2 is defined as in Equation (3.7), except that the statistic is computed only for

the men in the sample, and s_w^2 is defined similarly for the women. Thus the standard error of $\overline{Y}_m - \overline{Y}_w$ is

$$SE(\overline{Y}_m - \overline{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}. \tag{3.19}$$

For a simplified version of Equation (3.19) when Y is a Bernoulli random variable, see Exercise 3.15.

The t-statistic for testing the null hypothesis is constructed analogously to the t-statistic for testing a hypothesis about a single population mean, by subtracting the null hypothesized value of $\mu_m - \mu_w$ from the estimator $\overline{Y}_m - \overline{Y}_w$ and dividing the result by the standard error of $\overline{Y}_m - \overline{Y}_w$:

$$t = \frac{(\overline{Y}_m - \overline{Y}_w) - d_0}{SE(\overline{Y}_m - \overline{Y}_w)} \quad (t\text{-statistic for comparing two means}). \tag{3.20}$$

If both n_m and n_w are large, then this t-statistic has a standard normal distribution when the null hypothesis is true.

Because the t-statistic in Equation (3.20) has a standard normal distribution under the null hypothesis when n_m and n_w are large, the p-value of the two-sided test is computed exactly as it was in the case of a single population. That is, the p-value is computed using Equation (3.14).

To conduct a test with a prespecified significance level, simply calculate the t-statistic in Equation (3.20), and compare it to the appropriate critical value. For example, the null hypothesis is rejected at the 5% significance level if the absolute value of the t-statistic exceeds 1.96.

If the alternative is one-sided rather than two-sided (that is, if the alternative is that $\mu_m - \mu_w > d_0$), then the test is modified as outlined in Section 3.2. The p-value is computed using Equation (3.17), and a test with a 5% significance level rejects when $t > 1.64$.

Confidence Intervals for the Difference Between Two Population Means

The method for constructing confidence intervals summarized in Section 3.3 extends to constructing a confidence interval for the difference between the means, $d = \mu_m - \mu_w$. Because the hypothesized value d_0 is rejected at the 5% level if $|t| > 1.96$, d_0 will be in the confidence set if $|t| \leq 1.96$. But $|t| \leq 1.96$ means that the estimated difference, $\overline{Y}_m - \overline{Y}_w$, is less than 1.96 standard errors away from d_0. Thus the 95% two-sided confidence interval for d consists of those values of d within ± 1.96 standard errors of $\overline{Y}_m - \overline{Y}_w$:

95% confidence interval for $d = \mu_m - \mu_w$ is

$$(\overline{Y}_m - \overline{Y}_w) \pm 1.96 SE(\overline{Y}_m - \overline{Y}_w). \tag{3.21}$$

With these formulas in hand, the box "The Gender Gap of Earnings of College Graduates in the United States" contains an empirical investigation of sex differences in earnings of U.S. college graduates.

3.5 Differences-of-Means Estimation of Causal Effects Using Experimental Data

Recall from Section 1.2 that a randomized controlled experiment randomly selects subjects (individuals or, more generally, entities) from a population of interest, then randomly assigns them either to a treatment group, which receives the experimental treatment, or to a control group, which does not receive the treatment. The difference between the sample means of the treatment and control groups is an estimator of the causal effect of the treatment.

The Causal Effect as a Difference of Conditional Expectations

The causal effect of a treatment is the expected effect on the outcome of interest of the treatment as measured in an ideal randomized controlled experiment. This effect can be expressed as the difference of two conditional expectations. Specifically, the **causal effect** on Y of treatment level x is the difference in the conditional expectations, $E(Y|X = x) - E(Y|X = 0)$, where $E(Y|X = x)$ is the expected value of Y for the treatment group (which receives treatment level $X = x$) in an ideal randomized controlled experiment and $E(Y|X = 0)$ is the expected value of Y for the control group (which receives treatment level $X = 0$). In the context of experiments, the causal effect is also called the **treatment effect**. If there are only two treatment levels (that is, if the treatment is binary), then we can let $X = 0$ denote the control group and $X = 1$ denote the treatment group. If the treatment is binary, then the causal effect (that is, the treatment effect) is $E(Y|X = 1) - E(Y|X = 0)$ in an ideal randomized controlled experiment.

Estimation of the Causal Effect Using Differences of Means

If the treatment in a randomized controlled experiment is binary, then the causal effect can be estimated by the difference in the sample average outcomes between the treatment and control groups. The hypothesis that the treatment is ineffective is equivalent to the hypothesis that the two means are the same, which can be tested using the *t*-statistic for comparing two means, given in Equation (3.20). A 95% confidence interval for the difference in the means of the two groups is a 95% confidence interval for the causal effect, so a 95% confidence interval for the causal effect can be constructed using Equation (3.21).

   A well-designed, well-run experiment can provide a compelling estimate of a causal effect. For this reason, randomized controlled experiments are commonly conducted in some fields, such as medicine. In economics, however, experiments tend to be expensive, difficult to administer, and, in some cases, ethically questionable, so they are used less often. For this reason, econometricians sometimes study "natural

The Gender Gap of Earnings of College Graduates in the United States

The box in Chapter 2 "The Distribution of Earnings in the United States in 2015" shows that, on average, male college graduates earn more than female college graduates. What are the recent trends in this "gender gap" in earnings? Social norms and laws governing sex discrimination in the workplace have changed substantially in the United States. Is the gender gap in earnings of college graduates stable, or has it changed over time?

Table 3.1 gives estimates of hourly earnings for college-educated full-time workers ages 25–34 in the United States in 1996, 2008, and 2015, using data collected by the Current Population Survey (CPS). Earnings for 1996 and 2008 were adjusted for inflation by putting them in 2015 dollars using the Consumer Price Index (CPI).[1] In 2015, the average hourly earnings of the 1917 men surveyed was $28.06, and the standard deviation of earnings for men was $14.37. The average hourly earnings in 2015 of the 1816 women surveyed was $23.04, and the standard deviation of earnings was $11.22. Thus the

estimate of the gender gap in earnings for 2015 is $5.02 ($=$28.06 - 23.04), with a standard error of $0.42 ($=\sqrt{14.37^2/1917 + 11.22^2/1816}$). The 95% confidence interval for the gender gap in earnings in 2015 is $5.02 \pm 1.96 \times 0.42 = (\$4.19, \$5.84)$. The null hypothesis of no gender gap is rejected at the 1% significance level (the t-statistic is $11.9 = 5.02/0.42$, which exceeds the two-sided 1% critical value of 2.58).

The results in Table 3.1 suggest four conclusions. First, the gender gap is large. An hourly gap of $5.02 adds up to more than $10,000 over a year, assuming a 40-hour workweek and 50 paid weeks per year. Second, from 1996 to 2015, the estimated gender gap increased by $1.14 per hour in real terms, from $3.88 per hour to $5.02 per hour, which is statistically significant at the 5% significance level (Exercise 3.17). Third, the gap is also large if it is measured instead in percentage terms: According to the estimates in Table 3.1, in 2015 women earned 18% less per hour than men did ($5.02/28.06$), slightly more than the

TABLE 3.1 Trends in Hourly Earnings in the United States of Working College Graduates, Ages 25–34, 1996 to 2015, in 2015 Dollars

	Men			Women			Difference, Men vs. Women		
Year	$\overline{Y}_m$	s_m	n_m	$\overline{Y}_w$	s_w	n_w	$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	95% Confidence Interval for d
1996	24.87	11.44	1387	20.99	8.93	1232	3.88**	0.40	3.09–4.66
2008	27.94	13.88	1855	23.16	11.11	1877	4.78**	0.41	3.98–5.59
2015	28.06	14.37	1917	23.04	11.22	1816	5.02**	0.42	4.19–5.84

These estimates are computed using data on all full-time workers ages 25–34 surveyed in the Current Population Survey conducted in March of the next year (for example, the data for 2015 were collected in March 2016). The difference is significantly different from 0 at the **1% significance level.

gap of 16% seen in 1996 (3.88/24.87). Fourth, the gender gap is less for young college graduates (the group analyzed in Table 3.1) than it is for all college graduates (analyzed in Table 2.4): As reported in Table 2.4, the mean earnings for all college-educated women working full-time in 2015 was $27.23, while for men this mean was $35.10, which corresponds to a gender gap of 22% $[=(35.10 - 27.23)/35.10]$ among all full-time college-educated workers.

This empirical analysis documents that the gender gap in hourly earnings is large and has been fairly stable (or perhaps has increased slightly) over the recent past. The analysis does not, however, tell us *why* this gap exists. Does it arise from sex discrimination in the labor market? Does it reflect differences in skills, experience, or education between men and women? Does it reflect differences in choice of jobs? Or is there some other cause? We return to these questions once we have in hand the tools of multiple regression analysis, the topic of Part II.

_____

[1]Because of inflation, a dollar in 1996 was worth more than a dollar in 2015, in the sense that a dollar in 1996 could buy more goods and services than a dollar in 2015 could. Thus earnings in 1996 cannot be directly compared to earnings in 2015 without adjusting for inflation. One way to make this adjustment is to use the CPI, a measure of the price of a "market basket" of consumer goods and services constructed by the Bureau of Labor Statistics. Over the 19 years from 1996 to 2015, the price of the CPI market basket rose by 51.1%; in other words, the CPI basket of goods and services that cost $100 in 1996 cost $151.10 in 2015. To make earnings in 1996 and 2015 comparable in Table 3.1, 1996 earnings are inflated by the amount of overall CPI price inflation, that is, by multiplying 1996 earnings by 1.511 to put them into "2015 dollars."

experiments," also called quasi-experiments, in which some event unrelated to the treatment or subject characteristics has the effect of assigning different treatments to different subjects *as if* they had been part of a randomized controlled experiment. The box "A Novel Way to Boost Retirement Savings" provides an example of such a quasi-experiment that yielded some surprising conclusions.

3.6   Using the *t*-Statistic When the Sample Size Is Small

In Sections 3.2 through 3.5, the *t*-statistic is used in conjunction with critical values from the standard normal distribution for hypothesis testing and for the construction of confidence intervals. The use of the standard normal distribution is justified by the central limit theorem, which applies when the sample size is large. When the sample size is small, the standard normal distribution can provide a poor approximation to the distribution of the *t*-statistic. If, however, the population distribution is itself normally distributed, then the exact distribution (that is, the finite-sample distribution; see Section 2.6) of the *t*-statistic testing the mean of a single population is the Student *t* distribution with $n - 1$ degrees of freedom, and critical values can be taken from the Student *t* distribution.

A Novel Way to Boost Retirement Savings

Many economists think that people do not save enough for retirement. Conventional methods for encouraging retirement savings focus on financial incentives, but there also has been an upsurge in interest in unconventional ways to encourage saving for retirement.

In an important study published in 2001, Brigitte Madrian and Dennis Shea considered one such unconventional method for stimulating retirement savings. Many firms offer retirement savings plans in which the firm matches, in full or in part, savings taken out of the paycheck of participating employees. Enrollment in such plans, called 401(k) plans after the applicable section of the U.S. tax code, is always optional. However, at some firms, employees are automatically enrolled in the plan, although they can opt out; at other firms, employees are enrolled only if they choose to opt in. According to conventional economic models of behavior, the method of enrollment—opt out or opt in—should not matter: The rational worker computes the optimal action, then takes it. But, Madrian and Shea wondered, could conventional economics be wrong? Does the *method of enrollment* in a savings plan directly affect its enrollment rate?

To measure the effect of the method of enrollment, Madrian and Shea studied a large firm that changed the default option for its 401(k) plan from nonparticipation to participation. They compared two groups of workers: those hired the year before the change and not automatically enrolled (but could opt in) and those hired in the year after the change and automatically enrolled (but could opt out). The financial aspects of the plan remained the same, and Madrian and Shea found no systematic differences between the workers hired before and after the change. Thus, from an econometrician's perspective, the change was like a randomly assigned treatment, so the causal effect of the change could be estimated by the difference in means between the two groups.

Madrian and Shea found that the default enrollment rule made a huge difference: The enrollment rate for the "opt-in" (control) group was 37.4% ($n = 4249$), whereas the enrollment rate for the "opt-out" (treatment) group was 85.9% ($n = 5801$). The estimate of the treatment effect is 48.5 percentage points ($= 85.9\% - 37.4\%$). Because their sample is large, the 95% confidence (computed in Exercise 3.15) for the treatment effect is tight, 46.8 to 50.2 percentage points.

How could the default choice matter so much? Maybe workers found these financial choices too confusing, or maybe they just didn't want to think about growing old. Neither explanation is economically rational—but both are consistent with the predictions of the growing field of behavioral economics, and both could lead to accepting the default enrollment option.

This research had an important practical impact. In August 2006, Congress passed the Pension Protection Act, which (among other things) encouraged firms to offer 401(k) plans in which enrollment is the default. The econometric findings of Madrian and Shea and others featured prominently in testimony on this part of the legislation.

There is increasing evidence that small "nudges," like the opt-in default, can have large effects on complicated personal financial decisions. For example, Bettinger et al. (2012) conducted a randomized controlled experiment, done in conjunction with a large tax preparation firm, in which some of the low-income parents using the firm's services were randomly provided a small amount of help in filling out the U.S. federal student aid form. They found that this "nudge" substantially increased college application rates and ultimately college attendance. To learn more about behavioral economics and the design of retirement savings plans, see Benartzi and Thaler (2007) and Beshears, Choi, Laibson, and Madrian (2008).

In 2017, Richard Thaler was awarded the Nobel Prize in Economics for his work establishing the field of behavioral economics. To learn more about the many ways that behavioral economics has influenced economic thinking and economic policy, see the Nobel committee's summary (Nobel Committee 2017).

The *t*-Statistic and the Student *t* Distribution

The t-statistic testing the mean. Consider the *t*-statistic used to test the hypothesis that the mean of Y is $\mu_{Y,0}$, using data $Y_1, \ldots, Y_n$. The formula for this statistic is given by Equation (3.10), where the standard error of $\overline{Y}$ is given by Equation (3.8). Substitution of the latter expression into the former yields the formula for the *t*-statistic:

$$t = \frac{\overline{Y} - \mu_{Y,0}}{\sqrt{s_Y^2/n}}, \tag{3.22}$$

where s_Y^2 is given in Equation (3.7).

   As discussed in Section 3.2, under general conditions the *t*-statistic has a standard normal distribution if the sample size is large and the null hypothesis is true [see Equation (3.12)]. Although the standard normal approximation to the *t*-statistic is reliable for a wide range of distributions of Y if n is large, it can be unreliable if n is small. The exact distribution of the *t*-statistic depends on the distribution of Y, and it can be very complicated. There is, however, one special case in which the exact distribution of the *t*-statistic is relatively simple: If $Y_1, \ldots, Y_n$ are i.i.d. draws from a normal distribution, then the *t*-statistic in Equation (3.22) has a Student *t* distribution with $n - 1$ degrees of freedom. (The mathematics behind this result is provided in Sections 18.4 and 19.4.)

   If the population distribution is normally distributed, then critical values from the Student *t* distribution can be used to perform hypothesis tests and to construct confidence intervals. As an example, consider a hypothetical problem in which $t^{act} = 2.15$ and $n = 8$, so that the degrees of freedom is $n - 1 = 7$. From Appendix Table 2, the 5% two-sided critical value for the t_7 distribution is 2.36. Because the *t*-statistic is smaller in absolute value than the critical value $(2.15 < 2.36)$, the null hypothesis would not be rejected at the 5% significance level against the two-sided alternative. The 95% confidence interval for μ_Y, constructed using the t_7 distribution, would be $\overline{Y} \pm 2.36SE(\overline{Y})$. This confidence interval is wider than the confidence interval constructed using the standard normal critical value of 1.96.

The t-statistic testing differences of means. The *t*-statistic testing the difference of two means, given in Equation (3.20), does not have a Student *t* distribution, even if the population distribution of Y is normal. (The Student *t* distribution does not apply here because the variance estimator used to compute the standard error in Equation (3.19) does not produce a denominator in the *t*-statistic with a chi-squared distribution.)

   A modified version of the differences-of-means *t*-statistic, based on a different standard error formula—the "pooled" standard error formula—has an exact Student *t* distribution when Y is normally distributed; however, the pooled standard error formula applies only in the special case that the two groups have the same variance or that each group has the same number of observations (Exercise 3.21). Adopt the

notation of Equation (3.19) so that the two groups are denoted as m and w. The pooled variance estimator is

$$s_{pooled}^2 = \frac{1}{n_m + n_w - 2} \left[\underbrace{\sum_{i=1}^{n_m} (Y_i - \overline{Y}_m)^2}_{\text{group } m} + \underbrace{\sum_{i=1}^{n_w} (Y_i - \overline{Y}_m)^2}_{\text{group } w} \right], \quad (3.23)$$

where the first summation is for the observations in group m and the second summation is for the observations in group w. The pooled standard error of the difference in means is $SE_{pooled}(\overline{Y}_m - \overline{Y}_w) = s_{pooled} \times \sqrt{1/n_m + 1/n_w}$, and the pooled t-statistic is computed using Equation (3.20), where the standard error is the pooled standard error, $SE_{pooled}(\overline{Y}_m - \overline{Y}_w)$.

If the population distribution of Y in group m is $N(\mu_m, \sigma_m^2)$, if the population distribution of Y in group w is $N(\mu_w, \sigma_w^2)$, *and* if the two group variances are the same (that is, $\sigma_m^2 = \sigma_w^2$), then under the null hypothesis the t-statistic computed using the pooled standard error has a Student t distribution with $n_m + n_w - 2$ degrees of freedom.

The drawback of using the pooled variance estimator s_{pooled}^2 is that it applies only if the two population variances are the same (assuming $n_m \neq n_w$). If the population variances are different, the pooled variance estimator is biased and inconsistent. If the population variances are different but the pooled variance formula is used, the null distribution of the pooled t-statistic is not a Student t distribution, even if the data are normally distributed; in fact, it does not even have a standard normal distribution in large samples. Therefore, the pooled standard error and the pooled t-statistic should not be used unless you have a good reason to believe that the population variances are the same.

Use of the Student t Distribution in Practice

For the problem of testing the mean of Y, the Student t distribution is applicable if the underlying population distribution of Y is normal. For economic variables, however, normal distributions are the exception (for example, see the boxes in Chapter 2 "The Distribution of Earnings in the United States in 2015" and "A Bad Day on Wall Street"). Even if the data are not normally distributed, the normal approximation to the distribution of the t-statistic is valid if the sample size is large. Therefore, inferences—hypothesis tests and confidence intervals—about the mean of a distribution should be based on the large-sample normal approximation.

When comparing two means, any economic reason for two groups having different means typically implies that the two groups also could have different variances. Accordingly, the pooled standard error formula is inappropriate, and the correct standard error formula, which allows for different group variances, is as given in Equation (3.19). Even if the population distributions are normal, the t-statistic computed using the standard error formula in Equation (3.19) does not have a Student

t distribution. In practice, therefore, inferences about differences in means should be based on Equation (3.19), used in conjunction with the large-sample standard normal approximation.

Even though the Student *t* distribution is rarely applicable in economics, some software uses the Student *t* distribution to compute *p*-values and confidence intervals. In practice, this does not pose a problem because the difference between the Student *t* distribution and the standard normal distribution is negligible if the sample size is large. For $n > 15$, the difference in the *p*-values computed using the Student *t* and standard normal distributions never exceeds 0.01; for $n > 80$, the difference never exceeds 0.002. In most modern applications, and in all applications in this text, the sample sizes are in the hundreds or thousands, large enough for the difference between the Student *t* distribution and the standard normal distribution to be negligible.

3.7   Scatterplots, the Sample Covariance, and the Sample Correlation

What is the relationship between age and earnings? This question, like many others, relates one variable, X (age), to another, Y (earnings). This section reviews three ways to summarize the relationship between variables: the scatterplot, the sample covariance, and the sample correlation coefficient.

Scatterplots

A **scatterplot** is a plot of n observations on X_i and Y_i, in which each observation is represented by the point (X_i, Y_i). For example, Figure 3.2 is a scatterplot of age (X) and hourly earnings (Y) for a sample of 200 managers in the information industry from the March 2016 CPS. Each dot in Figure 3.2 corresponds to an (X, Y) pair for one of the observations. For example, one of the workers in this sample is 45 years old and earns $49.15 per hour; this worker's age and earnings are indicated by the highlighted dot in Figure 3.2. The scatterplot shows a positive relationship between age and earnings in this sample: Older workers tend to earn more than younger workers. This relationship is not exact, however, and earnings could not be predicted perfectly using only a person's age.

Sample Covariance and Correlation

The covariance and correlation were introduced in Section 2.3 as two properties of the joint probability distribution of the random variables X and Y. Because the population distribution is unknown, in practice we do not know the population covariance or correlation. The population covariance and correlation can, however, be estimated by taking a random sample of n members of the population and collecting the data (X_i, Y_i), $i = 1, \ldots, n$.

FIGURE 3.2   Scatterplot of Average Hourly Earnings vs. Age
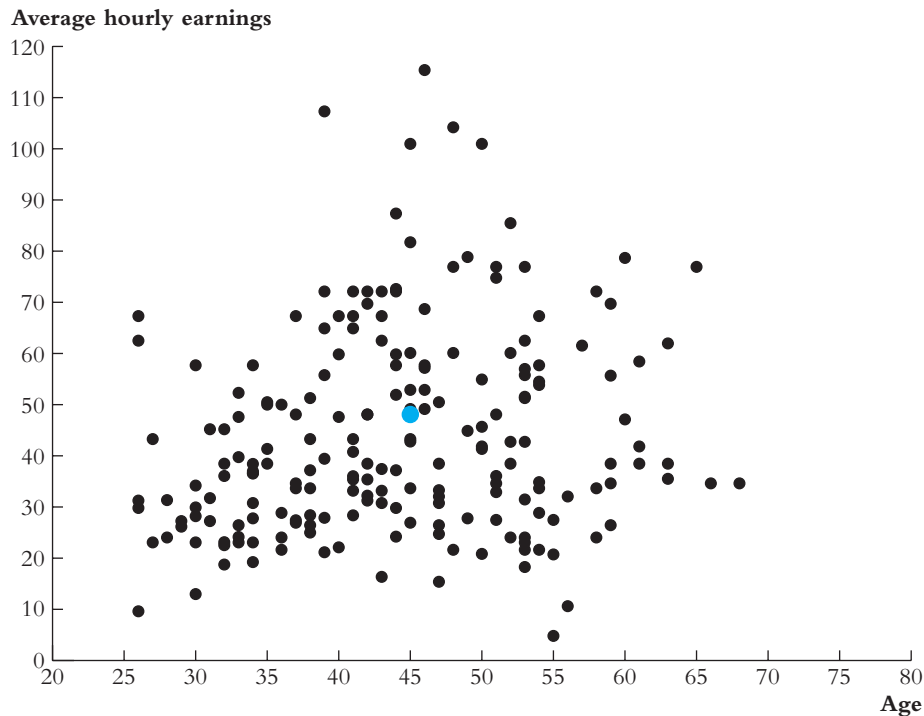
Each point in the plot represents the age and average earnings of one of the 200 workers in the sample. The highlighted dot corresponds to a 45-year-old worker who earns $49.15 per hour. The data are for computer and information systems managers from the March 2016 CPS.

The sample covariance and correlation are estimators of the population covariance and correlation. Like the estimators discussed previously in this chapter, they are computed by replacing a population mean (the expectation) with a sample mean. The **sample covariance**, denoted s_{XY}, is

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}). \tag{3.24}$$

Like the sample variance, the average in Equation (3.24) is computed by dividing by $n-1$ instead of n; here, too, this difference stems from using $\overline{X}$ and $\overline{Y}$ to estimate the respective population means. When n is large, it makes little difference whether division is by n or $n-1$.

The **sample correlation coefficient**, or **sample correlation**, is denoted r_{XY} and is the ratio of the sample covariance to the sample standard deviations:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}. \tag{3.25}$$

The sample correlation measures the strength of the linear association between X and Y in a sample of n observations. Like the population correlation, the sample correlation is unit free and lies between -1 and 1: $|r_{XY}| \leq 1$.

The sample correlation equals 1 if $X_i = Y_i$ for all i and equals -1 if $X_i = -Y_i$ for all i. More generally, the correlation is ± 1 if the scatterplot is a straight line. If the line slopes upward, then there is a positive relationship between X and Y and the correlation is 1. If the line slopes down, then there is a negative relationship and the correlation is -1. The closer the scatterplot is to a straight line, the closer the correlation is to ± 1. A high correlation coefficient does not necessarily mean that the line has a steep slope; rather, it means that the points in the scatterplot fall very close to a straight line.

Consistency of the sample covariance and correlation. Like the sample variance, the sample covariance is consistent. That is,

$$s_{XY} \xrightarrow{p} \sigma_{XY}. \tag{3.26}$$

In other words, in large samples the sample covariance is close to the population covariance with high probability.

The proof of the result in Equation (3.26) under the assumption that (X_i, Y_i) are i.i.d. and that X_i and Y_i have finite fourth moments is similar to the proof in Appendix 3.3 that the sample covariance is consistent and is left as an exercise (Exercise 3.20).

Because the sample variance and sample covariance are consistent, the sample correlation coefficient is consistent; that is, $r_{XY} \xrightarrow{p} \text{corr}(X_i, Y_i)$.

Example. As an example, consider the data on age and earnings in Figure 3.2. For these 200 workers, the sample standard deviation of age is $s_A = 9.57$ years, and the sample standard deviation of earnings is $s_E = \$19.93$ per hour. The sample covariance between age and earnings is $s_{AE} = 91.51$ (the units are years $\times$ dollars per hour, not readily interpretable). Thus the sample correlation coefficient is $r_{AE} = 91.51/(9.57 \times 19.93) = 0.48$. The correlation of 0.48 means that there is a positive relationship between age and earnings, but as is evident in the scatterplot, this relationship is far from perfect.

To verify that the correlation does not depend on the units of measurement, suppose that earnings had been reported in cents, in which case the sample standard deviation of earnings is 1993¢ per hour and the covariance between age and earnings is 9151 (units are years $\times$ cents per hour); then the correlation is $9151/(9.57 \times 1993) = 0.48$, or 48%.

Figure 3.3 gives additional examples of scatterplots and correlation. Figure 3.3a shows a strong positive linear relationship between these variables, and the sample correlation is 0.9.

Figure 3.3b shows a strong negative relationship with a sample correlation of -0.8. Figure 3.3c shows a scatterplot with no evident relationship, and the sample

FIGURE 3.3   Scatterplots for Four Hypothetical Data Sets

The scatterplots in Figures 3.3a and 3.3b show strong linear relationships between X and Y. In Figure 3.3c, X is independent of Y and the two variables are uncorrelated. In Figure 3.3d, the two variables also are uncorrelated even though they are related nonlinearly.
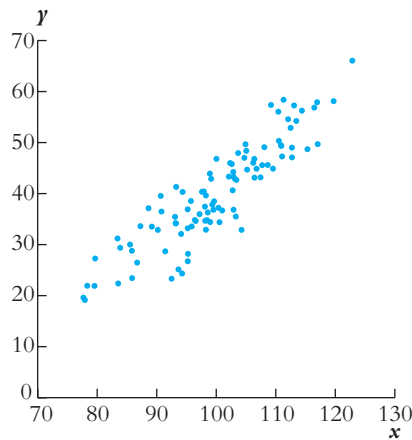
(a) Correlation = +0.9
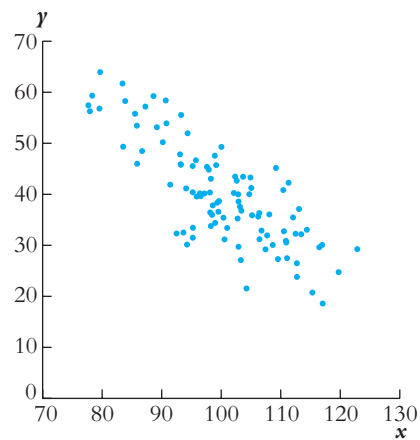
(b) Correlation = −0.8
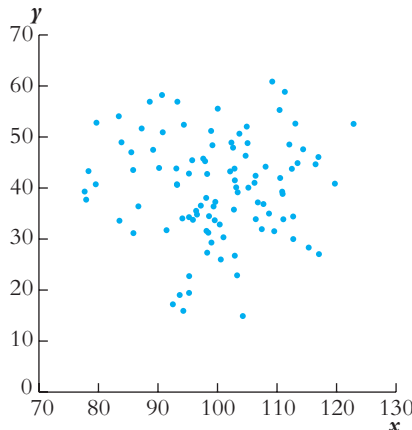
(c) Correlation = 0.0
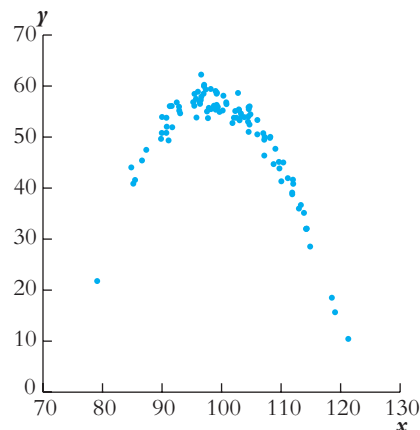
(d) Correlation = 0.0 (quadratic)

correlation is 0. Figure 3.3d shows a clear relationship: As X increases, Y initially increases but then decreases. Despite this discernable relationship between X and Y, the sample correlation is 0; the reason is that for these data small values of Y are associated with *both* large and small values of X.

This final example emphasizes an important point: The correlation coefficient is a measure of *linear* association. There is a relationship in Figure 3.3d, but it is not linear.

Summary

1. The sample average, $\overline{Y}$, is an estimator of the population mean, μ_Y. When $Y_1, \ldots, Y_n$ are i.i.d.,
   a. the sampling distribution of $\overline{Y}$ has mean μ_Y and variance $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$;
   b. $\overline{Y}$ is unbiased;
   c. by the law of large numbers, $\overline{Y}$ is consistent; and
   d. by the central limit theorem, $\overline{Y}$ has an approximately normal sampling distribution when the sample size is large.

2. The t-statistic is used to test the null hypothesis that the population mean takes on a particular value. If n is large, the t-statistic has a standard normal sampling distribution when the null hypothesis is true.

3. The t-statistic can be used to calculate the p-value associated with the null hypothesis. The p-value is the probability of drawing a statistic at least as adverse to the null hypothesis as the one you actually computed in your sample, assuming the null hypothesis is correct. A small p-value is evidence that the null hypothesis is false.

4. A 95% confidence interval for μ_Y is an interval constructed so that it contains the true value of μ_Y in 95% of all possible samples.

5. Hypothesis tests and confidence intervals for the difference in the means of two populations are conceptually similar to tests and intervals for the mean of a single population.

6. The sample correlation coefficient is an estimator of the population correlation coefficient and measures the linear relationship between two variables—that is, how well their scatterplot is approximated by a straight line.

Key Terms

estimator (62)

estimate (62)

bias (63)

consistency (63)

efficiency (63)

BLUE (Best Linear Unbiased Estimator) (64)

least squares estimator (65)

hypothesis tests (66)

null hypothesis (67)

alternative hypothesis (67)

two-sided alternative hypothesis (67)

p-value (significance probability) (68)

sample variance (70)

sample standard deviation (70)

degrees of freedom (70)

standard error of $\overline{Y}$ (70)

t-statistic (71)

t-ratio (71)

test statistic (71)

type I error (73)

type II error (73)

significance level (73)

critical value (73)

rejection region (73)

acceptance region (73)

size of a test (73)

power of a test (73)

one-sided alternative hypothesis (74)

confidence set (75)

confidence level (75)

confidence interval (75)

coverage probability (76)

test for the difference between two
means (77)

causal effect (79)

treatment effect (79)

scatterplot (85)

sample covariance (86)

sample correlation coefficient (sample
correlation) (86)

MyLab Economics Can Help You Get a Better Grade

MyLab Economics If your exam were tomorrow, would you be ready? For each chapter, **MyLab Economics** Practice Tests and Study Plan help you prepare for your exams. You can also find the Exercises and all Review the Concepts Questions available now in **MyLab Economics**. To see how it works, turn to the **MyLab Economics** spread on the inside front cover of this text and then go to **www.pearson.com/mylab/economics**.

For additional Empirical Exercises and Data Sets, log on to the Companion Website at **www.pearsonhighered.com/stock_watson**.

Review the Concepts

3.1 Explain the difference between the sample average $\overline{Y}$ and the population mean.

3.2 Explain the difference between an estimator and an estimate. Provide an example of each.

3.3 A population distribution has a mean of 10 and a variance of 16. Determine the mean and variance of $\overline{Y}$ from an i.i.d. sample from this population for (a) $n = 10$; (b) $n = 100$; and (c) $n = 1000$. Relate your answers to the law of large numbers.

3.4 What role does the central limit theorem play in statistical hypothesis testing? In the construction of confidence intervals?

3.5 What is the difference between a null hypothesis and an alternative hypothesis? Among size, significance level, and power? Between a one-sided alternative hypothesis and a two-sided alternative hypothesis?

3.6 Why does a confidence interval contain more information than the result of a single hypothesis test?

3.7 Explain why the differences-of-means estimator, applied to data from a randomized controlled experiment, is an estimator of the treatment effect.

3.8 Sketch a hypothetical scatterplot for a sample of size 10 for two random variables with a population correlation of (a) 1.0; (b) −1.0; (c) 0.9; (d) −0.5; and (e) 0.0.

Exercises

3.1 In a population, $\mu_Y = 100$ and $\sigma_Y^2 = 43$. Use the central limit theorem to answer the following questions:

    a. In a random sample of size $n = 100$, find $\Pr(\overline{Y} < 101)$.

    b. In a random sample of size $n = 64$, find $\Pr(101 < \overline{Y} < 103)$.

    c. In a random sample of size $n = 165$, find $\Pr(\overline{Y} > 98)$.

3.2 Let Y be a Bernoulli random variable with success probability $\Pr(Y = 1) = p$, and let $Y_1, \ldots, Y_n$ be i.i.d. draws from this distribution. Let $\hat{p}$ be the fraction of successes (1s) in this sample.

    a. Show that $\hat{p} = \overline{Y}$.

    b. Show that $\hat{p}$ is an unbiased estimator of p.

    c. Show that $\text{var}(\hat{p}) = p(1 - p)/n$.

3.3 In a survey of 400 likely voters, 215 responded that they would vote for the incumbent, and 185 responded that they would vote for the challenger. Let p denote the fraction of all likely voters who preferred the incumbent at the time of the survey, and let $\hat{p}$ be the fraction of survey respondents who preferred the incumbent.

    a. Use the survey results to estimate p.

    b. Use the estimator of the variance of $\hat{p}$, $\hat{p}(1 - \hat{p})/n$, to calculate the standard error of your estimator.

    c. What is the p-value for the test of H_0: $p = 0.5$ vs. H_1: $p \neq 0.5$?

    d. What is the p-value for the test of H_0: $p = 0.5$ vs. H_1: $p > 0.5$?

    e. Why do the results from (c) and (d) differ?

    f. Did the survey contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey? Explain.

3.4 Using the data in Exercise 3.3:

    a. Construct a 95% confidence interval for p.

    b. Construct a 99% confidence interval for p.

    c. Why is the interval in (b) wider than the interval in (a)?

    d. Without doing any additional calculations, test the hypothesis H_0: $p = 0.50$ vs. H_1: $p \neq 0.50$ at the 5% significance level.

3.5 A survey of 1055 registered voters is conducted, and the voters are asked to choose between candidate A and candidate B. Let p denote the fraction of voters in the population who prefer candidate A, and let $\hat{p}$ denote the fraction of voters in the sample who prefer candidate A.

    a. You are interested in the competing hypotheses H_0: $p = 0.5$ vs. H_1: $p \neq 0.5$. Suppose you decide to reject H_0 if $|\hat{p} - 0.5| > 0.02$.

    i. What is the size of this test?

    ii. Compute the power of this test if $p = 0.53$.

  b. In the survey, $\hat{p} = 0.54$.

    i. Test $H_0: p = 0.5$ vs. $H_1: p \neq 0.5$ using a 5% significance level.

    ii. Test $H_0: p = 0.5$ vs. $H_1: p > 0.5$ using a 5% significance level.

    iii. Construct a 95% confidence interval for p.

    iv. Construct a 99% confidence interval for p.

    v. Construct a 50% confidence interval for p.

  c. Suppose the survey is carried out 20 times, using independently selected voters in each survey. For each of these 20 surveys, a 95% confidence interval for p is constructed.

    i. What is the probability that the true value of p is contained in all 20 of these confidence intervals?

    ii. How many of these confidence intervals do you expect to contain the true value of p?

  d. In survey jargon, the "margin of error" is $1.96 \times SE(\hat{p})$; that is, it is half the length of the 95% confidence interval. Suppose you want to design a survey that has a margin of error of at most 1%. That is, you want $\Pr(|\hat{p} - p| > 0.01) \leq 0.05$. How large should n be if the survey uses simple random sampling?

3.6 Let $Y_1, \ldots, Y_n$ be i.i.d. draws from a distribution with mean μ. A test of $H_0: \mu = 5$ vs. $H_1: \mu \neq 5$ using the usual t-statistic yields a p-value of 0.03.

  a. Does the 95% confidence interval contain $\mu = 5$? Explain.

  b. Can you determine if $\mu = 6$ is contained in the 95% confidence interval? Explain.

3.7 In a given population, 11% of the likely voters are African American. A survey using a simple random sample of 600 landline telephone numbers finds 8% African Americans. Is there evidence that the survey is biased? Explain.

3.8 A new version of the SAT is given to 1000 randomly selected high school seniors. The sample mean test score is 1110, and the sample standard deviation is 123. Construct a 95% confidence interval for the population mean test score for high school seniors.

3.9 Suppose a lightbulb manufacturing plant produces bulbs with a mean life of 2000 hours and a standard deviation of 200 hours. An inventor claims to have developed an improved process that produces bulbs with a longer mean life and the same standard deviation. The plant manager randomly selects 100 bulbs produced by the process. She says that she will believe the inventor's claim if the sample mean life of the bulbs is greater than 2100 hours;

otherwise, she will conclude that the new process is no better than the old process. Let μ denote the mean of the new process. Consider the null and alternative hypotheses H_0: $\mu = 2000$ vs. H_1: $\mu > 2000$.

  a. What is the size of the plant manager's testing procedure?

  b. Suppose the new process is, in fact, better and has a mean bulb life of 2150 hours. What is the power of the plant manager's testing procedure?

  c. What testing procedure should the plant manager use if she wants the size of her test to be 5%?

3.10 Suppose a new standardized test is given to 100 randomly selected third-grade students in New Jersey. The sample average score, $\overline{Y}$, on the test is 58 points, and the sample standard deviation, s_Y, is 8 points.

  a. The authors plan to administer the test to all third-grade students in New Jersey. Construct a 95% confidence interval for the mean score of all New Jersey third graders.

  b. Suppose the same test is given to 200 randomly selected third graders from Iowa, producing a sample average of 62 points and a sample standard deviation of 11 points. Construct a 90% confidence interval for the difference in mean scores between Iowa and New Jersey.

  c. Can you conclude with a high degree of confidence that the population means for Iowa and New Jersey students are different? (What is the standard error of the difference in the two sample means? What is the p-value of the test of no difference in means versus some difference?)

3.11 Consider the estimator $\widetilde{Y}$, defined in Equation (3.1). Show that (a) $E(\widetilde{Y}) = \mu_Y$ and (b) $\text{var}(\widetilde{Y}) = 1.25\sigma_Y^2/n$.

3.12 To investigate possible sex discrimination in a firm, a sample of 100 men and 64 women with similar job descriptions is selected at random. A summary of the resulting monthly salaries follows:

	Average Salary ($\overline{Y}$)	Standard Deviation (s_Y)	n
Men	$3100	$200	100
Women	$2900	$320	64

  a. What do these data suggest about wage differences in the firm? Do they represent statistically significant evidence that average wages of men and women are different? (To answer this question, first, state the null and alternative hypotheses; second, compute the relevant t-statistic; third, compute the p-value associated with the t-statistic; and, finally, use the p-value to answer the question.)

  b. Do these data suggest that the firm is guilty of sex discrimination in its compensation policies? Explain.

3.13 Data on fifth-grade test scores (reading and mathematics) for 420 school districts in California yield average score $\overline{Y} = 654.2$ and standard deviation $s_Y = 19.1$.

  a. Construct a 95% confidence interval for the mean test score in the population.

  b. When the districts were divided into those with small classes (< 20 students per teacher) and those with large classes (≥ 20 students per teacher), the following results were found:

Class Size	Average Score ($\overline{Y}$)	Standard Deviation (s_Y)	n
Small	657.4	19.4	238
Large	650.0	17.9	182

  Is there statistically significant evidence that the districts with smaller classes have higher average test scores? Explain.

3.14 Values of height in inches (X) and weight in pounds (Y) are recorded from a sample of 300 male college students. The resulting summary statistics are $\overline{X} = 70.5$ in., $\overline{Y} = 158$ lb, $s_X = 1.8$ in., $s_Y = 14.2$ lb, $s_{XY} = 21.73$ in. $\times$ lb, and $r_{XY} = 0.85$. Convert these statistics to the metric system (meters and kilograms).

3.15 Y_a and Y_b are Bernoulli random variables from two different populations, denoted a and b. Suppose $E(Y_a) = p_a$ and $E(Y_b) = p_b$. A random sample of size n_a is chosen from population a, with a sample average denoted $\hat{p}_a$, and a random sample of size n_b is chosen from population b, with a sample average denoted $\hat{p}_b$. Suppose the sample from population a is independent of the sample from population b.

  a. Show that $E(\hat{p}_a) = p_a$ and $\text{var}(\hat{p}_a) = p_a(1 - p_a)/n_a$. Show that $E(\hat{p}_b) = p_b$ and $\text{var}(\hat{p}_b) = p_b(1 - p_b)/n_b$.

  b. Show that $\text{var}(\hat{p}_a - \hat{p}_b) = \dfrac{p_a(1 - p_a)}{n_a} + \dfrac{p_b(1 - p_b)}{n_b}$.

  (*Hint:* Remember that the samples are independent.)

  c. Suppose n_a and n_b are large. Show that a 95% confidence interval for $p_a - p_b$ is given by $(\hat{p}_a - \hat{p}_b) \pm 1.96\sqrt{\dfrac{\hat{p}_a(1 - \hat{p}_a)}{n_a} + \dfrac{\hat{p}_b(1 - \hat{p}_b)}{n_b}}$. How would you construct a 90% confidence interval for $p_a - p_b$?

  d. Read the box "A Novel Way to Boost Retirement Savings" in Section 3.5. Let population a denote the opt-out (treatment) group and population b denote the opt-in (control) group. Construct a 95% confidence interval for the treatment effect, $p_a - p_b$.

3.16 Grades on a standardized test are known to have a mean of 1000 for students in the United States. The test is administered to 453 randomly selected students in Florida; in this sample, the mean is 1013, and the standard deviation (s) is 108.

  a. Construct a 95% confidence interval for the average test score for Florida students.

  b. Is there statistically significant evidence that Florida students perform differently than other students in the United States?

  c. Another 503 students are selected at random from Florida. They are given a three-hour preparation course before the test is administered. Their average test score is 1019, with a standard deviation of 95.

   i. Construct a 95% confidence interval for the change in mean test scores associated with the prep course.

   ii. Is there statistically significant evidence that the prep course helped?

  d. The original 453 students are given the prep course and then are asked to take the test a second time. The average change in their test scores is 9 points, and the standard deviation of the change is 60 points.

   i. Construct a 95% confidence interval for the change in average test scores.

   ii. Is there statistically significant evidence that students will perform better on their second attempt, after taking the prep course?

   iii. Students may have performed better in their second attempt because they took the prep course or because they gained test-taking experience in their first attempt. Describe an experiment that would quantify these two effects.

3.17 Read the box "The Gender Gap of Earnings of College Graduates in the United States" in Section 3.4.

  a. Construct a 95% confidence interval for the change in men's average hourly earnings between 1996 and 2015.

  b. Construct a 95% confidence interval for the change in women's average hourly earnings between 1996 and 2015.

  c. Construct a 95% confidence interval for the change in the gender gap in average hourly earnings between 1996 and 2015. (*Hint:* $\overline{Y}_{m,1996} - \overline{Y}_{w,1996}$ is independent of $\overline{Y}_{m,2015} - \overline{Y}_{w,2015}$.)

3.18 This exercise shows that the sample variance is an unbiased estimator of the population variance when $Y_1, \ldots, Y_n$ are i.i.d. with mean μ_Y and variance σ_Y^2.

  a. Use Equation (2.32) to show that
  $E(Y_i - \overline{Y})^2 = \text{var}(Y_i) - 2\text{cov}(Y_i, \overline{Y}) + \text{var}(\overline{Y})$.

  b. Use Equation (2.34) to show that $\text{cov}(\overline{Y}, Y_i) = \sigma_Y^2/n$.

  c. Use the results in (a) and (b) to show that $E(s_Y^2) = \sigma_Y^2$.

3.19 a. $\overline{Y}$ is an unbiased estimator of μ_Y. Is $\overline{Y}^2$ an unbiased estimator of μ_Y^2?

b. $\overline{Y}$ is a consistent estimator of μ_Y. Is $\overline{Y}^2$ a consistent estimator of μ_Y^2?

3.20 Suppose (X_i, Y_i) are i.i.d. with finite fourth moments. Prove that the sample covariance is a consistent estimator of the population covariance; that is, $s_{XY} \xrightarrow{P} \sigma_{XY}$, where s_{XY} is defined in Equation (3.24). (*Hint:* Use the strategy of Appendix 3.3.)

3.21 Show that the pooled standard error $[SE_{pooled}(\overline{Y}_m - \overline{Y}_w)]$ given following Equation (3.23) equals the usual standard error for the difference in means in Equation (3.19) when the two group sizes are the same $(n_m = n_w)$.

3.22 Suppose $Y_i \sim$ i.i.d.$N(\mu_Y, \sigma_Y^2)$ for $i = 1, \ldots, n$. With σ_Y^2 known, the t-statistic for testing H_0: $\mu_Y = 0$ vs. H_1: $\mu_Y > 0$ is $t = (\overline{Y} - 0)/SE(\overline{Y})$, where $SE(\overline{Y}) = \sigma_Y/\sqrt{n}$. Suppose $\sigma_Y = 10$ and $n = 100$, so that $SE(\overline{Y}) = 1$. Using a test with a size of 5%, the null hypothesis is rejected if $t > 1.64$.

a. Suppose $\mu_Y = 0$, so the null hypothesis is true. What is the probability that the null hypothesis is rejected?

b. Suppose $\mu_Y = 2$, so the alternative hypothesis is true. What is the probability that the null hypothesis is rejected?

c. Suppose that in 90% of cases the data are drawn from a population where the null is true ($\mu_Y = 0$) and in 10% of cases the data come from a population where the alternative is true and $\mu_Y = 2$. Your data came from either the first or the second population, but you don't know which.

   i. You compute the t-statistic. What is the probability that $t > 1.64$—that is, that you reject the null hypothesis?

   ii. Suppose you reject the null hypothesis; that is, $t > 1.64$. What is the probability that the sample data were drawn from the $\mu_Y = 0$ population?

d. It is hard to discover a new effective drug. Suppose 90% of new drugs are ineffective and only 10% are effective. Let Y denote the drop in the level of a specific blood toxin for a patient taking a new drug. If the drug is ineffective, $\mu_Y = 0$ and $\sigma_Y = 10$; if the drug is effective, $\mu_Y = 2$ and $\sigma_Y = 10$.

   i. A new drug is tested on a random sample of $n = 100$ patients, data are collected, and the resulting t-statistic is found to be greater than 1.64. What is the probability that the drug is ineffective (i.e., what is the false positive rate for the test using $t > 1.64$)?

   ii. Suppose the one-sided test uses instead the 0.5% significance level. What is the probability that the drug is ineffective (i.e., what is the false positive rate)?

Empirical Exercises

E3.1 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find the data file **CPS96_15**, which contains an extended version of the data set used in Table 3.1 of the text for the years 1996 and 2015. It contains data on full-time workers, ages 25–34, with a high school diploma or a B.A./B.S. as their highest degree. A detailed description is given in **CPS96_15_Description**, available on the website. Use these data to complete the following.

a. i. Compute the sample mean for average hourly earnings (*AHE*) in 1996 and 2015.

  ii. Compute the sample standard deviation for *AHE* in 1996 and 2015.

  iii. Construct a 95% confidence interval for the population means of *AHE* in 1996 and 2015.

  iv. Construct a 95% confidence interval for the change in the population means of *AHE* between 1996 and 2015.

b. In 2015, the value of the Consumer Price Index (CPI) was 237.0. In 1996, the value of the CPI was 156.9. Repeat (a), but use *AHE* measured in real 2015 dollars ($2015); that is, adjust the 1996 data for the price inflation that occurred between 1996 and 2015.

c. If you were interested in the change in workers' purchasing power from 1996 to 2015, would you use the results from (a) or (b)? Explain.

d. Using the data for 2015:

  i. Construct a 95% confidence interval for the mean of *AHE* for high school graduates.

  ii. Construct a 95% confidence interval for the mean of *AHE* for workers with a college degree.

  iii. Construct a 95% confidence interval for the difference between the two means.

e. Repeat (d) using the 1996 data expressed in $2015.

f. Using appropriate estimates, confidence intervals, and test statistics, answer the following questions:

  i. Did real (inflation-adjusted) wages of high school graduates increase from 1996 to 2015?

  ii. Did real wages of college graduates increase?

  iii. Did the gap between earnings of college and high school graduates increase? Explain.

g. Table 3.1 presents information on the gender gap for college graduates. Prepare a similar table for high school graduates, using the 1996 and 2015 data. Are there any notable differences between the results for high school and college graduates?

E3.2   A consumer is given the chance to buy a baseball card for $1, but he declines the trade. If the consumer is now given the baseball card, will he be willing to sell it for $1? Standard consumer theory suggests yes, but behavioral economists have found that "ownership" tends to increase the value of goods to consumers. That is, the consumer may hold out for some amount more than $1 (for example, $1.20) when selling the card, even though he was willing to pay only some amount less than $1 (for example, $0.88) when buying it. Behavioral economists call this phenomenon the "endowment effect." John List investigated the endowment effect in a randomized experiment involving sports memorabilia traders at a sports-card show. Traders were randomly given one of two sports collectibles, say good A or good B, that had approximately equal market value.[1] Those receiving good A were then given the option of trading good A for good B with the experimenter; those receiving good B were given the option of trading good B for good A with the experimenter. Data from the experiment and a detailed description can be found on the text website, **http://www.pearsonhighered.com/stock_watson/**, in the files **Sportscards** and **Sportscards_Description**.[2]

a.   i.   Suppose that, absent any endowment effect, all the subjects prefer good A to good B. What fraction of the experiment's subjects would you expect to trade the good that they were given for the other good? (*Hint:* Because of random assignment of the two treatments, approximately 50% of the subjects received good A, and 50% received good B.)

   ii.   Suppose that, absent any endowment effect, 50% of the subjects prefer good A to good B, and the other 50% prefer good B to good A. What fraction of the subjects would you expect to trade the good they were given for the other good?

   iii.   Suppose that, absent any endowment effect, X% of the subjects prefer good A to good B, and the other $(100 - X)$% prefer good B to good A. Show that you would expect 50% of the subjects to trade the good they were given for the other good.

b.   Using the sports-card data, what fraction of the subjects traded the good they were given? Is the fraction significantly different from 50%? Is there evidence of an endowment effect? (*Hint:* Review Exercises 3.2 and 3.3.)

c.   Some have argued that the endowment effect may be present but that it is likely to disappear as traders gain more trading experience. Half of the experimental subjects were dealers, and the other half were nondealers. Dealers have more experience than nondealers. Repeat (b) for dealers and nondealers. Is there a significant difference in their behavior?

---

[1]Good A was a ticket stub from the game in which Cal Ripken, Jr., set the record for consecutive games played, and good B was a souvenir from the game in which Nolan Ryan won his 300th game.

[2]These data were provided by Professor John List of the University of Chicago and were used in his paper "Does Market Experience Eliminate Market Anomalies," *Quarterly Journal of Economics*, 2003, 118(1): 41–71.

Is the evidence consistent with the hypothesis that the endowment effect disappears as traders gain more experience? (*Hint:* Review Exercise 3.15.)

3.1 The U.S. Current Population Survey

Each month the U.S. Census Bureau and the U.S. Bureau of Labor Statistics conduct the Current Population Survey (CPS), which provides data on labor force characteristics of the population, including the levels of employment, unemployment, and earnings. Approximately 54,000 U.S. households are surveyed each month. The sample is chosen by randomly selecting addresses from a database of addresses from the most recent decennial census augmented with data on new housing units constructed after the last census. The exact random sampling scheme is rather complicated (first, small geographical areas are randomly selected; then housing units within these areas are randomly selected); details can be found in the *Handbook of Labor S*t*atistics* and on the Bureau of Labor Statistics website (www.bls.gov).

The survey conducted each March is more detailed than those in other months and asks questions about earnings during the previous year. The statistics in Tables 2.4 and 3.1 were computed using the March surveys. The CPS earnings data are for full-time workers, defined to be persons employed more than 35 hours per week for at least 48 weeks in the previous year.

More details on the data can be found in the replication materials for this chapter, available at http://www.princeton.edu/~mwatson/Stock-Watson_4E.

3.2 Two Proofs That $\overline{Y}$ Is the Least Squares Estimator of μ_Y

This appendix provides two proofs, one using calculus and one not, that $\overline{Y}$ minimizes the sum of squared prediction mistakes in Equation (3.2)—that is, that $\overline{Y}$ is the least squares estimator of $E(Y)$.

Calculus Proof

To minimize the sum of squared prediction mistakes, take its derivative and set it to 0:

$$\frac{d}{dm} \sum_{i=1}^{n} (Y_i - m)^2 = -2 \sum_{i=1}^{n} (Y_i - m) = -2 \sum_{i=1}^{n} Y_i + 2nm = 0. \qquad (3.27)$$

Solving for the final equation for m shows that $\sum_{i=1}^{n} (Y_i - m)^2$ is minimized when $m = \overline{Y}$.

Noncalculus Proof

The strategy is to show that the difference between the least squares estimator and $\overline{Y}$ must be 0, from which it follows that $\overline{Y}$ is the least squares estimator. Let $d = \overline{Y} - m$, so that $m = \overline{Y} - d$. Then $(Y_i - m)^2 = (Y_i - [\overline{Y} - d])^2 = ([Y_i - \overline{Y}] + d)^2 = (Y_i - \overline{Y})^2 + 2d(Y_i - \overline{Y}) + d^2$. Thus the sum of squared prediction mistakes [Equation (3.2)] is

$$\sum_{i=1}^{n}(Y_i - m)^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 + 2d\sum_{i=1}^{n}(Y_i - \overline{Y}) + nd^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 + nd^2, \quad (3.28)$$

where the second equality uses the fact that $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$. Because both terms in the final line of Equation (3.28) are nonnegative and because the first term does not depend on d, $\sum_{i=1}^{n}(Y_i - m)^2$ is minimized by choosing d to make the second term, nd^2, as small as possible. This is done by setting $d = 0$—that is, by setting $m = \overline{Y}$—so that $\overline{Y}$ is the least squares estimator of $E(Y)$.

3.3 A Proof That the Sample Variance Is Consistent

This appendix uses the law of large numbers to prove that the sample variance, s_Y^2, is a consistent estimator of the population variance, σ_Y^2, as stated in Equation (3.9), when $Y_1, \ldots, Y_n$ are i.i.d. and $E(Y_i^4) < \infty$.

First, consider a version of the sample variance that uses n instead of $n - 1$ as a divisor:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - 2\overline{Y}\frac{1}{n}\sum_{i=1}^{n}Y_i + \overline{Y}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \overline{Y}^2$$

$$\xrightarrow{p} (\sigma_Y^2 + \mu_Y^2) - \mu_Y^2$$

$$= \sigma_Y^2, \quad (3.29)$$

where the first equality uses $(Y_i - \overline{Y})^2 = Y_i^2 - 2\overline{Y}Y_i + \overline{Y}^2$ and the second uses $\frac{1}{n}\sum_{i=1}^{n}Y_i = \overline{Y}$.

The convergence in the third line follows from (i) applying the law of large numbers to $\frac{1}{n}\sum_{i=1}^{n}Y_i^2 \xrightarrow{p} E(Y^2)$ (which follows because Y_i^2 are i.i.d. and have finite variance because $E(Y_i^4)$ is finite), (ii) recognizing that $E(Y_i^2) = \sigma_Y^2 + \mu_Y^2$ (Key Concept 2.3), and (iii) noting $\overline{Y} \xrightarrow{p} \mu_Y$, so that $\overline{Y}^2 \xrightarrow{p} \mu_Y^2$. Finally, $s_Y^2 = (\frac{n}{n-1})(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2) \xrightarrow{p} \sigma_Y^2$ follows from Equation (3.29) and $(\frac{n}{n-1}) \to 1$.

Linear Regression with One Regressor

The superintendent of an elementary school district must decide whether to hire additional teachers, and she wants your advice. Hiring the teachers will reduce the number of students per teacher (the student–teacher ratio) by two but will increase the district's expenses. So she asks you: If she cuts class sizes by two, what will the effect be on student performance, as measured by scores on standardized tests?

Now suppose a father tells you that his family wants to move to a town with a good school system. He is interested in a specific school district: Test scores for this district are not publicly available, but the father knows its class size, based on the district's student–teacher ratio. So he asks you: if he tells you the district's class size, could you predict that district's standardized test scores?

These two questions are clearly related: They both pertain to the relation between class size and test scores. Yet they are different. To answer the superintendent's question, you need an estimate of the causal effect of a change in one variable (the student–teacher ratio, X) on another (test scores, Y). To answer the father's question, you need to know how X relates to Y, on average, across school districts so you can use this relation to predict Y given X in a specific district.

These two questions are examples of two different types of questions that arise in econometrics. The first type of questions pertains to **causal inference**: using data to estimate the effect on an outcome of interest of an intervention that changes the value of another variable. The second type of questions concerns **prediction**: using the observed value of some variable to predict the value of another variable.

This chapter introduces the linear regression model relating one variable, X, to another, Y. This model postulates a linear relationship between X and Y. Just as the mean of Y is an unknown characteristic of the population distribution of Y, the intercept and slope of the line relating X and Y are unknown characteristics of the population joint distribution of X and Y. The econometric problem is to estimate the intercept and slope using a sample of data on these two variables.

Like the differences in means, linear regression is a statistical procedure that can be used for causal inference and for prediction. The two uses, however, place different requirements on the data. Section 3.5 explained how a difference in mean outcomes between a treatment and a control group estimates the causal effect of the treatment when the treatment is randomly assigned in an experiment. When X is continuous, computing differences-in-means no longer works because there are many values X can take on, not just two. If, however, we make the additional assumption that the relation between X and Y is linear, then if X is randomly assigned, we can use linear regression to estimate the causal effect on Y of an intervention that changes X. Even if X is not randomly assigned,

however, linear regression gives us a way to predict the value of Y given X by modeling the conditional mean of Y given X as a linear function of X. As long as the observation for which Y is to be predicted is drawn from the same population as the data used to estimate the linear regression, the regression line provides a way to predict Y given X.

Sections 4.1–4.3 lay out the linear regression model and the least squares estimators of its slope and intercept. In Section 4.4, we turn to requirements on the data for estimation of a causal effect. In essence, the key requirement is that either X is set at random in an experiment or X is as-if randomly set.

Our focus on causal inference continues through Chapter 13. We return to the prediction problem in Chapter 14.

4.1 The Linear Regression Model

Return to the father's question: If he tells you the district's class size, could you predict that district's standardized test scores? In Chapter 2, we used the notation $E(Y|X = x)$ to denote the mean of Y given that X takes on the value x—that is, the conditional expectation of Y given $X = x$. The easiest starting point for modeling a function of X, when X can take on multiple values, is to suppose that it is linear. In the case of test scores and class size, this linear function can be written

$$E(TestScore|ClassSize) = \beta_0 + \beta_{ClassSize} \times ClassSize, \tag{4.1}$$

where β is the Greek letter beta, β_0 is the intercept, and $\beta_{ClassSize}$ is the slope.

If you were lucky enough to know β_0 and $\beta_{ClassSize}$, you could use Equation (4.1) to answer the father's question. For example, suppose he was looking at a district with a class size of 20 and that $\beta_0 = 720$ and $\beta_{ClassSize} = -0.6$. Then you could answer his question: Given that the class size is 20, you would predict test scores to be $720 - 0.6 \times 20 = 708$.

Equation (4.1) tells you what the test score will be, on average, for districts with class sizes of that value; it does not tell you what specifically the test score will be in any one district. Districts with the same class sizes can nevertheless differ in many ways and in general will have different values of test scores. As a result, if we use Equation (4.1) to make a prediction for a given district, we know that prediction will not be exactly right: The prediction will have an error. Stated mathematically, for any given district the imperfect relationship between class size and test score can be written

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize + error. \tag{4.2}$$

Equation (4.2) expresses the test score for the district in terms of one component, $\beta_0 + \beta_{ClassSize} \times ClassSize$, that represents the average relationship between class

size and scores in the population of school districts, and a second component that represents the error made using the prediction in Equation (4.1).

Although this discussion has focused on test scores and class size, the idea expressed in Equation (4.2) is much more general, so it is useful to introduce more general notation. Suppose you have a sample of n districts. Let Y_i be the average test score in the ith district, and let X_i be the average class size in the ith district, so that Equation (4.1) becomes $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$. Let u_i denote the error made by predicting Y_i using its conditional mean. Then Equation (4.2) can be written more generally as

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{4.3}$$

for each district (that is, $i = 1, \ldots, n$), where β_0 is the intercept of this line and β_1 is the slope. The general notation β_1 is used for the slope in Equation (4.3) instead of $\beta_{ClassSize}$ because this equation is written in terms of a general variable X.

Equation (4.3) is the **linear regression model with a single regressor**, in which Y is the **dependent variable** and X is the **independent variable** or the **regressor**.

The first part of Equation (4.3), $\beta_0 + \beta_1 X_i$, is the **population regression line** or the **population regression function**. This is the relationship that holds between Y and X, on average, over the population. Thus, given the value of X, according to this population regression line you would predict the value of the dependent variable, Y, to be its conditional mean given X. That conditional mean is given by Equation (4.1) which, in the more general notation of Equation (4.3), is $E(Y|X) = \beta_0 + \beta_1 X$.

The **intercept** β_0 and the **slope** β_1 are the **coefficients** of the population regression line, also known as the **parameters** of the population regression line. The slope β_1 is the difference in Y associated with a unit difference in X. The intercept is the value of the population regression line when $X = 0$; it is the point at which the population regression line intersects the Y axis. In some econometric applications, the intercept has a meaningful economic interpretation. In other applications, the intercept has no real-world meaning; for example, when X is the class size, strictly speaking the intercept is the expected value of test scores when there are no students in the class! When the real-world meaning of the intercept is nonsensical, it is best to think of it simply as the coefficient that determines the level of the regression line.

The term u_i in Equation (4.3) is the **error term**. In the context of the prediction problem, u_i is the difference between Y_i and its predicted value using the population regression line.

The linear regression model and its terminology are summarized in Key Concept 4.1.

Figure 4.1 summarizes the linear regression model with a single regressor for seven hypothetical observations on test scores (Y) and class size (X). The population regression line is the straight line $\beta_0 + \beta_1 X$. The population regression line slopes

Terminology for the Linear Regression Model with a Single Regressor

The linear regression model is

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

where

the subscript i runs over observations, $i = 1, \ldots, n$;

Y_i is the *dependent variable*, the *regressand*, or simply the *left-hand variable*;

X_i is the *independent variable*, the *regressor*, or simply the *right-hand variable*;

$\beta_0 + \beta_1 X$ is the *population regression line* or the *population regression function*;

β_0 is the *intercept* of the population regression line;

β_1 is the *slope* of the population regression line; and

u_i is the *error term*.

FIGURE 4.1 Scatterplot of Test Score vs. Student–Teacher Ratio (Hypothetical Data)

The scatterplot shows hypothetical observations for seven school districts. The population regression line is $\beta_0 + \beta_1 X$. The vertical distance from the ith point to the population regression line is $Y_i - (\beta_0 + \beta_1 X_i)$, which is the population error term u_i for the ith observation.
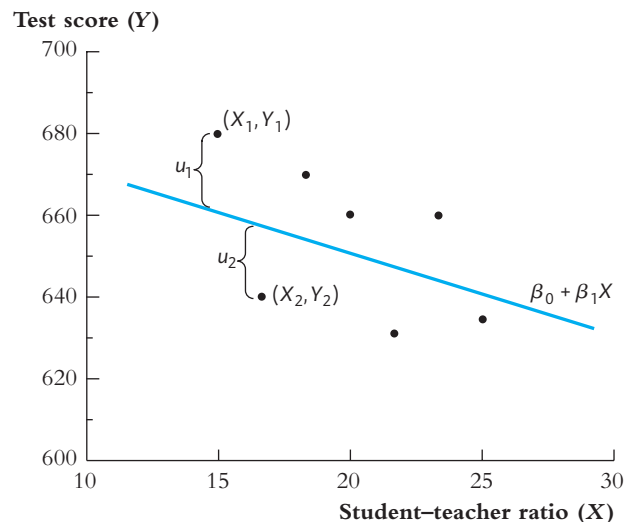
down ($\beta_1 < 0$), which means that districts with lower student–teacher ratios (smaller classes) tend to have higher test scores. The intercept β_0 has a mathematical meaning as the value of the Y axis intersected by the population regression line, but, as mentioned earlier, it has no real-world meaning in this example.

The hypothetical observations in Figure 4.1 do not fall exactly on the population regression line. For example, the value of Y for district 1, Y_1, is above the population regression line. This means that test scores in district 1 were better than predicted by the population regression line, so the error term for that district, u_1, is positive. In contrast, Y_2 is below the population regression line, so test scores for that district were worse than predicted and $u_2 < 0$.

4.2   Estimating the Coefficients of the Linear Regression Model

In a practical situation such as the application to class size and test scores, the intercept β_0 and the slope β_1 of the population regression line are unknown. Therefore, we must use data to estimate these unknown coefficients.

This estimation problem is similar to those faced in Chapter 3. For example, suppose you want to compare the mean earnings of men and women who recently graduated from college. Although the population mean earnings are unknown, we can estimate the population means using a random sample of male and female college graduates. Then the natural estimator of the unknown population mean earnings for women, for example, is the average earnings of the female college graduates in the sample.

The same idea extends to the linear regression model. We do not know the population value of $\beta_{ClassSize}$, the slope of the unknown population regression line relating X (class size) and Y (test scores). But just as it was possible to learn about the population mean using a sample of data drawn from that population, so is it possible to learn about the population slope $\beta_{ClassSize}$ using a sample of data.

The data we analyze here consist of test scores and class sizes in 1999 in 420 California school districts that serve kindergarten through eighth grade. The test score is the districtwide average of reading and math scores for fifth graders. Class size can be measured in various ways. The measure used here is one of the broadest, which is the number of students in the district divided by the number of teachers—that is, the districtwide student–teacher ratio. These data are described in more detail in Appendix 4.1.

Table 4.1 summarizes the distributions of test scores and class sizes for this sample. The average student–teacher ratio is 19.6 students per teacher, and the standard deviation is 1.9 students per teacher. The 10th percentile of the distribution of

TABLE 4.1	Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1999								
						Percentile			
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

FIGURE 4.2   **Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)**

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.

---
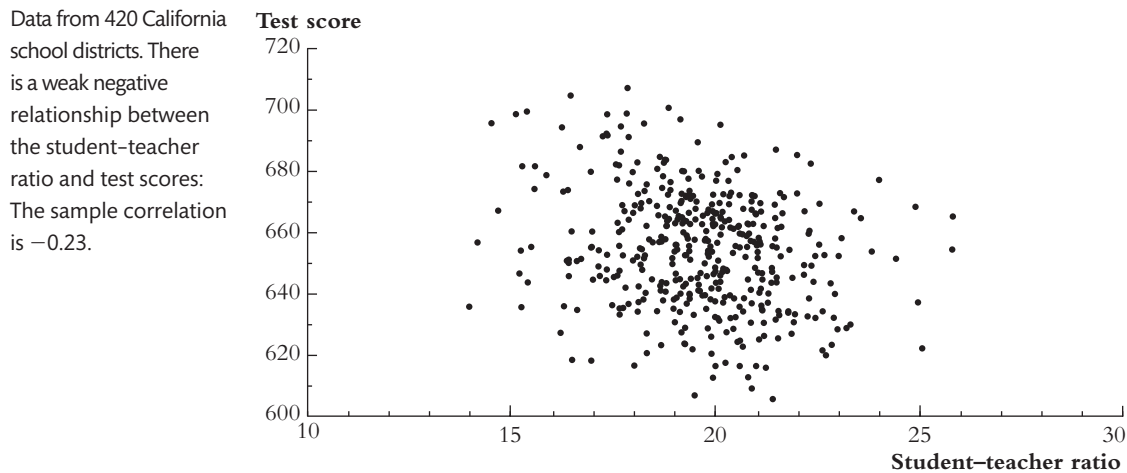
the student–teacher ratio is 17.3 (that is, only 10% of districts have student–teacher ratios below 17.3), while the district at the 90th percentile has a student–teacher ratio of 21.9.

A scatterplot of these 420 observations on test scores and student–teacher ratios is shown in Figure 4.2. The sample correlation is −0.23, indicating a weak negative relationship between the two variables. Although larger classes in this sample tend to have lower test scores, there are other determinants of test scores that keep the observations from falling perfectly along a straight line.

Despite this low correlation, if one could somehow draw a straight line through these data, then the slope of this line would be an estimate of $\beta_{ClassSize}$ based on these data. One way to draw the line would be to take out a pencil and a ruler and to "eyeball" the best line you could. While this method is easy, it is unscientific, and different people would create different estimated lines.

How, then, should you choose among the many possible lines? By far the most common way is to choose the line that produces the "least squares" fit to these data—that is, to use the ordinary least squares (OLS) estimator.

The Ordinary Least Squares Estimator

The OLS estimator chooses the regression coefficients so that the estimated regression line is as close as possible to the observed data, where closeness is measured by the sum of the squared mistakes made in predicting Y given X.

As discussed in Section 3.1, the sample average, $\overline{Y}$, is the least squares estimator of the population mean, $E(Y)$; that is, $\overline{Y}$ minimizes the total squared estimation mistakes $\sum_{i=1}^{n}(Y_i - m)^2$ among all possible estimators m [see Expression (3.2)].

The OLS estimator extends this idea to the linear regression model. Let b_0 and b_1 be some estimators of β_0 and β_1. The regression line based on these estimators is $b_0 + b_1 X$, so the value of Y_i predicted using this line is $b_0 + b_1 X_i$. Thus the mistake made in predicting the ith observation is $Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i$. The sum of these squared prediction mistakes over all n observations is

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2. \tag{4.4}$$

The sum of the squared mistakes for the linear regression model in Expression (4.4) is the extension of the sum of the squared mistakes for the problem of estimating the mean in Expression (3.2). In fact, if there is no regressor, then b_1 does not enter Expression (4.4), and the two problems are identical except for the different notation [m in Expression (3.2), b_0 in Expression (4.4)]. Just as there is a unique estimator, $\overline{Y}$, that minimizes Expression (3.2), so there is a unique pair of estimators of β_0 and β_1 that minimizes Expression (4.4).

The estimators of the intercept and slope that minimize the sum of squared mistakes in Expression (4.4) are called the **ordinary least squares (OLS) estimators** of β_0 and β_1.

OLS has its own special notation and terminology. The OLS estimator of β_0 is denoted $\hat{\beta}_0$, and the OLS estimator of β_1 is denoted $\hat{\beta}_1$. The **OLS regression line**, also called the **sample regression line** or **sample regression function**, is the straight line constructed using the OLS estimators: $\hat{\beta}_0 + \hat{\beta}_1 X$. The **predicted value** of Y_i given X_i, based on the OLS regression line, is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. The **residual** for the ith observation is the difference between Y_i and its predicted value: $\hat{u}_i = Y_i - \hat{Y}_i$.

The OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, are sample counterparts of the population coefficients, β_0 and β_1. Similarly, the OLS regression line, $\hat{\beta}_0 + \hat{\beta}_1 X$, is the sample counterpart of the population regression line, $\beta_0 + \beta_1 X$; and the OLS residuals, $\hat{u}_i$, are sample counterparts of the population errors, u_i.

You could compute the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ by trying different values of b_0 and b_1 repeatedly until you find those that minimize the total squared mistakes in Expression (4.4); they are the least squares estimates. This method would be tedious, however. Fortunately, there are formulas, derived by minimizing Expression (4.4) using calculus, that streamline the calculation of the OLS estimators.

The OLS formulas and terminology are collected in Key Concept 4.2. These formulas, which are derived in Appendix 4.2, are implemented in virtually all statistical and spreadsheet software.

OLS Estimates of the Relationship Between Test Scores and the Student–Teacher Ratio

When OLS is used to estimate a line relating the student–teacher ratio to test scores using the 420 observations in Figure 4.2, the estimated slope is −2.28, and

The OLS Estimator, Predicted Values, and Residuals

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.5}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{4.6}$$

The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n \tag{4.7}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n. \tag{4.8}$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{u}_i$) are computed from a sample of n observations of X_i and Y_i, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

the estimated intercept is 698.9. Accordingly, the OLS regression line for these 420 observations is

$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \tag{4.9}$$

where *TestScore* is the average test score in the district and *STR* is the student–teacher ratio. The "^" over *TestScore* in Equation (4.9) indicates that it is the predicted value based on the OLS regression line. Figure 4.3 plots this OLS regression line superimposed over the scatterplot of the data previously shown in Figure 4.2.

The slope of −2.28 means that when comparing two districts with class sizes that differ by one student per class (that is, *STR* differs by 1), the district with the larger class size has, on average, test scores that are lower by 2.28 points. A difference in the student–teacher ratio of two students per class is, on average, associated with a difference in test scores of 4.56 points $[= -2 \times (-2.28)]$. The negative slope indicates that districts with more students per teacher (larger classes) tend to do worse on the test.

It is now possible to predict the districtwide test score given a value of the student–teacher ratio. For example, for a district with 20 students per teacher, the predicted

The estimated regression line shows a negative relationship between test scores and the student–teacher ratio. For two districts with class sizes that differ by one student per class, the district with the larger class has, on average, test scores that are lower by 2.28 points.

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$
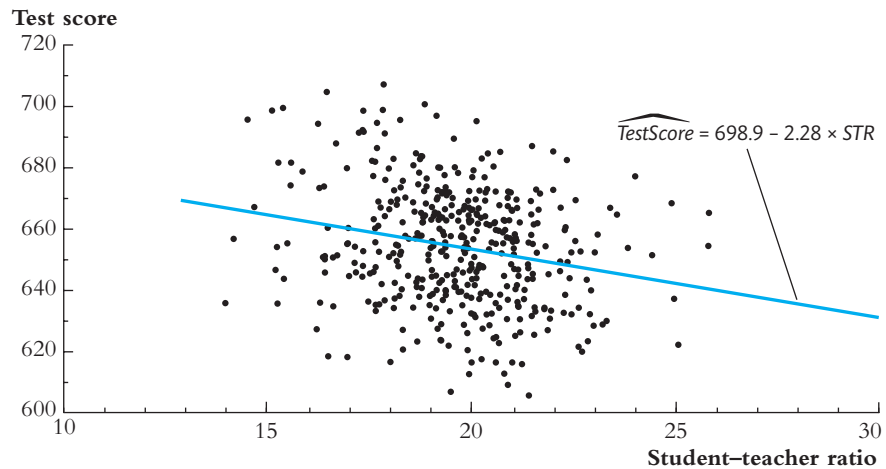
test score is $698.9 - 2.28 \times 20 = 653.3$. Of course, this prediction will not be exactly right because of the other factors that determine a district's performance. But the regression line does give a prediction (the OLS prediction) of what test scores would be for that district, based on its student–teacher ratio, absent those other factors.

Is the estimated slope large or small? According to Equation (4.9), for two districts with student-teacher ratios that differ by 2, the predicted value of test scores would differ by 4.56 points. For the California data, this difference of two students per class is large: It is roughly the difference between the median and the 10th percentile in Table 4.1. The associated difference in predicted test scores, however, is small compared to the spread of test scores in the data: 4.56 is slightly less than the difference between the median and the 60th percentile of test scores. In other words, a difference in class size that is large among these schools is associated with a relatively small difference in predicted test scores.

Why Use the OLS Estimator?

There are both practical and theoretical reasons to use the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Because OLS is the dominant method used in practice, it has become the common language for regression analysis throughout economics, finance (see "The 'Beta' of a Stock" box), and the social sciences more generally. Presenting results using OLS (or its variants discussed later in this text) means that you are "speaking the same language" as other economists and statisticians. The OLS formulas are built into virtually all spreadsheet and statistical software packages, making OLS easy to use.

The "Beta" of a Stock

A fundamental idea of modern finance is that an investor needs a financial incentive to take a risk. Said differently, the expected return[1] on a risky investment, R, must exceed the return on a safe, or risk-free, investment, R_f. Thus the expected excess return, $R - R_f$, on a risky investment, like owning stock in a company, should be positive.

At first, it might seem like the risk of a stock should be measured by its variance. Much of that risk, however, can be reduced by holding other stocks in a "portfolio"—in other words, by diversifying your financial holdings. This means that the right way to measure the risk of a stock is not by its *variance* but rather by its *covariance* with the market.

The capital asset pricing model (CAPM) formalizes this idea. According to the CAPM, the expected excess return on an asset is proportional to the expected excess return on a portfolio of all available assets (the market portfolio). That is, the CAPM says that

$$R - R_f = \beta(R_m - R_f), \qquad (4.10)$$

where R_m is the expected return on the market portfolio and β is the coefficient in the population regression of $R - R_f$ on $R_m - R_f$. In practice, the risk-free return is often taken to be the rate of interest on short-term U.S. government debt. According to the CAPM, a stock with a $\beta < 1$ has less risk than the market portfolio and therefore has a lower expected excess return than the market portfolio. In contrast, a stock with a $\beta > 1$ is riskier than the market portfolio and thus commands a higher expected excess return.

The "beta" of a stock has become a workhorse of the investment industry, and you can obtain estimated betas for hundreds of stocks on investment firm websites. Those betas typically are estimated by OLS regression of the actual excess return on the stock against the actual excess return on a broad market index.

The table below gives estimated betas for seven U.S. stocks. Low-risk sellers and producers of consumer staples like Wal-Mart and Coca-Cola have stocks with low betas; riskier stocks have high betas.

Company	Estimated β
Wal-Mart (discount retailer)	0.1
Coca-Cola (soft drinks)	0.6
Verizon (telecommunications)	0.7
Google (information technology)	1.0
General Electric (industrial)	1.1
Boeing (aircraft)	1.3
Bank of America (bank)	1.7

Source: finance.yahoo.com.

---

[1] The return on an investment is the change in its price plus any payout (dividend) from the investment as a percentage of its initial price. For example, a stock bought on January 1 for $100, which then paid a $2.50 dividend during the year and sold on December 31 for $105, would have a return of $R = [(\$105 - \$100) + \$2.50]/\$100 = 7.5\%$.

The OLS estimators also have desirable theoretical properties. They are analogous to the desirable properties, studied in Section 3.1, of $\overline{Y}$ as an estimator of the population mean. Under the assumptions introduced in Section 4.4, the OLS estimator is unbiased and consistent. The OLS estimator is also efficient among a certain class of unbiased estimators; however, this efficiency result holds under some additional special conditions, and further discussion of this result is deferred until Section 5.5.

4.3 Measures of Fit and Prediction Accuracy

Having estimated a linear regression, you might wonder how well that regression line describes the data. Does the regressor account for much or for little of the variation in the dependent variable? Are the observations tightly clustered around the regression line, or are they spread out?

The R^2 and the standard error of the regression measure how well the OLS regression line fits the data. The R^2 ranges between 0 and 1 and measures the fraction of the variance of Y_i that is explained by X_i. The standard error of the regression measures how far Y_i typically is from its predicted value.

The R^2

The **regression R^2** is the fraction of the sample variance of Y explained by (or predicted by) X. The definitions of the predicted value and the residual (see Key Concept 4.2) allow us to write the dependent variable Y_i as the sum of the predicted value, $\hat{Y}_i$, plus the residual $\hat{u}_i$:

$$Y_i = \hat{Y}_i + \hat{u}_i. \tag{4.11}$$

In this notation, the R^2 is the ratio of the sample variance of $\hat{Y}$ to the sample variance of Y.

Mathematically, the R^2 can be written as the ratio of the explained sum of squares to the total sum of squares. The **explained sum of squares (*ESS*)** is the sum of squared deviations of the predicted value, $\hat{Y}_i$, from its average, and the **total sum of squares (*TSS*)** is the sum of squared deviations of Y_i from its average:

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 \tag{4.12}$$

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{4.13}$$

Equation (4.12) uses the fact that the sample average OLS predicted value equals $\overline{Y}$ (proven in Appendix 4.3).

The R^2 is the ratio of the explained sum of squares to the total sum of squares:

$$R^2 = \frac{ESS}{TSS}. \tag{4.14}$$

Alternatively, the R^2 can be written in terms of the fraction of the variance of Y_i *not* explained by X_i. The **sum of squared residuals (*SSR*)** is the sum of the squared OLS residuals:

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2. \tag{4.15}$$

It is shown in Appendix 4.3 that $TSS = ESS + SSR$. Thus the R^2 also can be expressed as 1 minus the ratio of the sum of squared residuals to the total sum of squares:

$$R^2 = 1 - \frac{SSR}{TSS}. \tag{4.16}$$

Finally, the R^2 of the regression of Y on the single regressor X is the square of the correlation coefficient between Y and X (Exercise 4.12).

The R^2 ranges between 0 and 1. If $\hat{\beta}_1 = 0$, then X_i explains none of the variation of Y_i, and the predicted value of Y_i is $\hat{Y}_i = \hat{\beta}_0 = \overline{Y}$ [from Equation (4.6)]. In this case, the explained sum of squares is 0 and the sum of squared residuals equals the total sum of squares; thus the R^2 is 0. In contrast, if X_i explains all of the variation of Y_i, then $Y_i = \hat{Y}_i$ for all i, and every residual is 0 (that is, $\hat{u}_i = 0$), so that $ESS = TSS$ and $R^2 = 1$. In general, the R^2 does not take on the extreme value of 0 or 1 but falls somewhere in between. An R^2 near 1 indicates that the regressor is good at predicting Y_i, while an R^2 near 0 indicates that the regressor is not very good at predicting Y_i.

The Standard Error of the Regression

The **standard error of the regression (*SER*)** is an estimator of the standard deviation of the regression error u_i. The units of u_i and Y_i are the same, so the *SER* is a measure of the spread of the observations around the regression line, measured in the units of the dependent variable. For example, if the units of the dependent variable are dollars, then the *SER* measures the magnitude of a typical deviation from the regression line—that is, the magnitude of a typical regression error—in dollars.

Because the regression errors $u_1, \ldots, u_n$ are unobserved, the *SER* is computed using their sample counterparts, the OLS residuals $\hat{u}_1, \ldots, \hat{u}_n$. The formula for the *SER* is

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2}, \text{ where } s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-2}, \tag{4.17}$$

where the formula for $s_{\hat{u}}^2$ uses the fact (proven in Appendix 4.3 that the sample average of the OLS residuals is 0.

The formula for the *SER* in Equation (4.17) is similar to the formula for the sample standard deviation of Y given in Equation (3.7) in Section 3.2, except that $Y_i - \overline{Y}$ in Equation (3.7) is replaced by $\hat{u}_i$ and the divisor in Equation (3.7) is $n - 1$, whereas here it is $n - 2$. The reason for using the divisor $n - 2$ here (instead of n) is the same as the reason for using the divisor $n - 1$ in Equation (3.7): It corrects for a slight downward bias introduced because two regression coefficients were estimated. This is called a "degrees of freedom" correction because when two coefficients were estimated (β_0 and β_1), two "degrees of freedom" of the data were lost, so the divisor in this factor is $n - 2$. (The mathematics behind this is discussed in Section 5.6.) When n is large, the difference among dividing by n, by $n - 1$, or by $n - 2$ is negligible.

Prediction Using OLS

The predicted value $\hat{Y}_i$ for the ith observation is the value of Y predicted by the OLS regression line when X takes on its value X_i for that observation. This is called an **in-sample prediction** because the observation for which the prediction is made was also used to estimate the regression coefficients.

In practice, prediction methods are used to predict Y when X is known but Y is not. Such observations are not in the data set used to estimate the coefficients. Prediction for observations *not* in the estimation sample is called **out-of-sample prediction**.

The goal of prediction is to provide accurate out-of-sample predictions. For example, in the father's prediction problem, he was interested in predicting test scores for a district that had not reported them, using that district's student–teacher ratio. In the linear regression model with a single regressor, the predicted value for an out-of-sample observation that takes on the value X is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Because no prediction is perfect, a prediction should be accompanied by an estimate of its accuracy—that is, by an estimate of how accurate the prediction might reasonably be expected to be. A natural measure of that accuracy is the standard deviation of the out-of-sample prediction error, $Y - \hat{Y}$. Because Y is not known, this out-of-sample standard deviation cannot be estimated directly. If, however, the observation being predicted is drawn from the same population as the data used to estimate the regression coefficients, then the standard deviation of the out-of-sample prediction error can be estimated using the sample standard deviation of the in-sample prediction error, which is the standard error of the regression. A common way to report a prediction and its accuracy is as the prediction $\pm$ the SER—that is, $\hat{Y} \pm s_{\hat{u}}$. More refined measures of prediction accuracy are introduced in Chapter 14.

Application to the Test Score Data

Equation (4.9) reports the regression line, estimated using the California test score data, relating the standardized test score (*TestScore*) to the student–teacher ratio (*STR*). The R^2 of this regression is 0.051, or 5.1%, and the SER is 18.6.

The R^2 of 0.051 means that the regressor STR explains 5.1% of the variance of the dependent variable *TestScore*. Figure 4.3 superimposes the sample regression line on the scatterplot of the *TestScore* and *STR* data. As the scatterplot shows, the student–teacher ratio explains some of the variation in test scores, but much variation remains unaccounted for.

The SER of 18.6 means that the standard deviation of the regression residuals is 18.6, where the units are points on the standardized test. Because the standard deviation is a measure of spread, the SER of 18.6 means that there is a large spread of the scatterplot in Figure 4.3 around the regression line as measured in points on the test. This large spread means that predictions of test scores made using only the student–teacher ratio for that district will often be wrong by a large amount.

What should we make of this low R^2 and large *SER*? The fact that the R^2 of this regression is low (and the *SER* is large) does not, by itself, imply that this regression is either "good" or "bad." What the low R^2 *does* tell us is that other important factors influence test scores. These factors could include differences in the student body across districts, differences in school quality unrelated to the student–teacher ratio, or luck on the test. The low R^2 and high *SER* do not tell us what these factors are, but they do indicate that the student–teacher ratio alone explains only a small part of the variation in test scores in these data.

4.4 The Least Squares Assumptions for Causal Inference

In the test score example, the sample regression line, estimated using California district-level data, provides an answer to the father's problem of predicting the test score in a district when he knows its student–teacher ratio but not its test score.

The superintendent, however, is not interested in predicting test scores: She wants to improve them in her district. For that purpose, she needs to know the causal effect on test scores if she were to reduce the student–teacher ratio. Said differently, the superintendent has in mind a very particular definition of β_1: the causal effect on test scores of an intervention that changes the student–teacher ratio.

When β_1 is defined to be the causal effect, whether it is well estimated by OLS depends on the nature of the data. As discussed in Section 3.5, the difference in means between the treatment and control groups in an ideal randomized experiment is an unbiased estimator of the causal effect of a binary treatment; that is, if X is randomly assigned, the causal effect of the treatment is $E(Y|X = 1) - E(Y|X = 0)$. The difference in means is a workhorse statistical tool that can be used for many purposes; when X is randomly assigned, it provides an unbiased estimate of the causal effect of a binary treatment. This logic extends to the linear regression model and the least squares estimator.

In this section, we define β_1 to be the causal effect of a unit change in X. Because X can take on multiple values, the causal effect of a given change in X, Δx, is $\beta_1 \Delta x$, where the Greek letter Δ (delta) stands for "change in." This definition of the coefficient on the variable of interest (for example, *STR*) as its causal effect is maintained through Chapter 13.

This section lays out three mathematical assumptions under which OLS estimates the causal effect. The first assumption translates the idea that X is randomly assigned, or as-if randomly assigned, into the language of linear regression. The other two assumptions are technical ones under which the sampling distributions of the OLS estimators can be approximated by a normal distribution in large samples. These latter two assumptions are extensions of the two assumptions underlying the weak law of large numbers (Key Concept 2.6) and central limit theorem (Key Concept 2.7) for the sample mean $\overline{Y}$: that the data are i.i.d. and that outliers are unlikely.

Assumption 1: The Conditional Distribution of u_i Given X_i Has a Mean of Zero

The first least squares assumption translates into the language of regression analysis the requirement that, for estimation of the causal effect, X must be randomly assigned or as-if randomly assigned. To make this translation, we first need to be more specific about what the error term u_i is.

In the test score example, class size is just one of many facets of elementary education. One district might have better teachers, or it might use better textbooks. Two districts with comparable class sizes, teachers, and textbooks still might have very different student populations; perhaps one district has more immigrants (and thus fewer native English speakers) or wealthier families. Finally, even if two districts are the same in all these ways, they might have different test scores for essentially random reasons having to do with the performance of the individual students on the day of the test or errors in recording their scores. The error term in the class size regression represents the contribution to test scores made by all these other, omitted factors.

The first **least squares assumption** is that the conditional distribution of u_i given X_i has a mean of 0. This assumption is a formal mathematical statement about the other factors contained in u_i and asserts that these other factors are unrelated to X_i in the sense that, given a value of X_i, the mean of the distribution of these other factors is 0.

The conditional mean of u in a randomized controlled experiment. In a randomized controlled experiment with binary treatment, subjects are randomly assigned to the treatment group $(X = 1)$ or to the control group $(X = 0)$. When random assignment is done using a computer program that uses no information about the subject, X is distributed independently of the subject's personal characteristics, including those that determine Y. Because of random assignment, the conditional mean of u given X is 0. Because regression analysis models the conditional mean, X does not need to be distributed independently of all the other factors comprising u. However, the mean of u cannot be related to X; that is, $E(u_i | X_i) = 0$.

In observational data, X is not randomly assigned in an experiment. Instead, the best that can be hoped for is that X is *as if* randomly assigned, in the precise sense that $E(u_i | X_i) = 0$. Whether this assumption holds in a given empirical application with observational data requires careful thought and judgment, and we return to this issue repeatedly.

Correlation and conditional mean. Recall from Section 2.3 that if the conditional mean of one random variable given another is 0, then the two random variables have 0 covariance and thus are uncorrelated [Equation (2.28)]. Thus the conditional mean assumption $E(u_i | X_i) = 0$ implies that X_i and u_i are uncorrelated, or $\text{corr}(X_i, u_i) = 0$. Because correlation is a measure of linear association, this implication does not go the other way; even if X_i and u_i are uncorrelated, the conditional mean of u_i given X_i might be nonzero (see Figure 3.3). However, if X_i and u_i are correlated, then it must

be the case that $E(u_i|X_i)$ is nonzero. It is therefore often convenient to discuss the conditional mean assumption in terms of possible correlation between X_i and u_i. If X_i and u_i are correlated, then the conditional mean assumption is violated.

Assumption 2: (X_i, Y_i), i = 1, . . . , n, Are Independently and Identically Distributed

The second least squares assumption is that (X_i, Y_i), $i = 1, \ldots, n$, are independently and identically distributed (i.i.d.) across observations. As discussed in Section 2.5 (Key Concept 2.5), this assumption is a statement about how the sample is drawn. If the observations are drawn by simple random sampling from a single large population, then (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. For example, let X be the age of a worker and Y be his or her earnings, and imagine drawing a person at random from the population of workers. That randomly drawn person will have a certain age and earnings (that is, X and Y will take on some values). If a sample of n workers is drawn from this population, then (X_i, Y_i), $i = 1, \ldots, n$, necessarily have the same distribution. If they are drawn at random, they are also distributed independently from one observation to the next; that is, they are i.i.d.

The i.i.d. assumption is a reasonable one for many data collection schemes. For example, survey data from a randomly chosen subset of the population typically can be treated as i.i.d.

Not all sampling schemes produce i.i.d. observations on (X_i, Y_i). One example is when the values of X are not drawn from a random sample of the population but rather are set by a researcher as part of an experiment. For example, suppose a horticulturalist wants to study the effects of different organic weeding methods (X) on tomato production (Y) and accordingly grows different plots of tomatoes using different organic weeding techniques. If she picks the technique (the level of X) to be used on the ith plot and applies the same technique to the ith plot in all repetitions of the experiment, then the value of X_i does not change from one sample to the next. Said differently, X is fixed in repeated experiments—that is, repeated draws of the sample. Thus X_i is nonrandom (although the outcome Y_i is random), so the sampling scheme is not i.i.d. The results presented in this chapter developed for i.i.d. regressors are also true if the regressors are nonrandom. The case of a nonrandom regressor is, however, quite special. For example, modern experimental protocols would have the horticulturalist assign the level of X to the different plots using a computerized random number generator, thereby circumventing any possible bias by the horticulturalist (she might use her favorite weeding method for the tomatoes in the sunniest plot). When this modern experimental protocol is used, the level of X is random, and (X_i, Y_i) are i.i.d.

Another example of non-i.i.d. sampling is when observations refer to the same unit of observation over time. For example, we might have data on inventory levels (Y) at a firm and the interest rate at which the firm can borrow (X), where these data are collected over time from a specific firm; for example, they might be recorded four

times a year (quarterly) for 30 years. This is an example of time series data, and a key feature of time series data is that observations falling close to each other in time are not independent but rather tend to be correlated with each other: If interest rates are low now, they are likely to be low next quarter. This pattern of correlation violates the "independence" part of the i.i.d. assumption. Time series data introduce a set of complications that are best handled after developing the basic tools of regression analysis, so we postpone discussion of time series data until Chapter 15.

Assumption 3: Large Outliers Are Unlikely

The third least squares assumption is that large outliers—that is, observations with values of X_i, Y_i, or both that are far outside the usual range of the data—are unlikely. Large outliers can make OLS regression results misleading. This potential sensitivity of OLS to extreme outliers is illustrated in Figure 4.4 using hypothetical data.

In this book, the assumption that large outliers are unlikely is made mathematically precise by assuming that X and Y have nonzero finite fourth moments: $0 < E(X_i^4) < \infty$ and $0 < E(Y_i^4) < \infty$. Another way to state this assumption is that X and Y have finite kurtosis.

The assumption of finite kurtosis is used in the mathematics that justify the large-sample approximations to the distributions of the OLS test statistics. For example, we encountered this assumption in Chapter 3 when discussing the consistency of the sample variance. Specifically, Equation (3.9) states that the sample variance is a consistent estimator of the population variance σ_Y^2 ($s_Y^2 \xrightarrow{p} \sigma_Y^2$). If $Y_1, \ldots, Y_n$ are i.i.d. and the

---

FIGURE 4.4   The Sensitivity of OLS to Large Outliers

This hypothetical data set has one outlier. The OLS regression line estimated with the outlier shows a strong positive relationship between X and Y, but the OLS regression line estimated without the outlier shows no relationship.
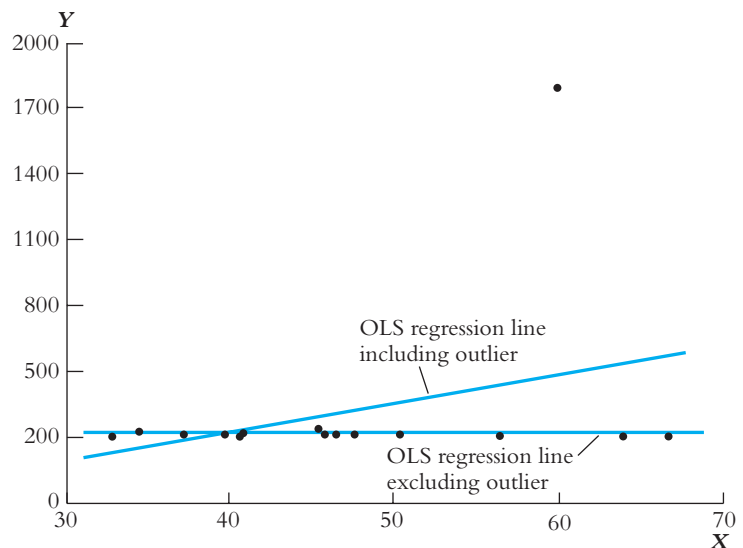
fourth moment of Y_i is finite, then the law of large numbers in Key Concept 2.6 applies to the average, $\frac{1}{n}\sum_{i=1}^{n}Y_i^2$, a key step in the proof in Appendix 3.3 showing that s_Y^2 is consistent.

One source of large outliers is data entry errors, such as a typographical error or incorrectly using different units for different observations. Imagine collecting data on the height of students in meters but inadvertently recording one student's height in centimeters instead. This would create a large outlier in the sample. One way to find outliers is to plot your data. If you decide that an outlier is due to a data entry error, then you can either correct the error or, if that is impossible, drop the observation from your data set.

Data entry errors aside, the assumption of finite kurtosis is a plausible one in many applications with economic data. Class size is capped by the physical capacity of a classroom; the best you can do on a standardized test is to get all the questions right, and the worst you can do is to get all the questions wrong. Because class size and test scores have a finite range, they necessarily have finite kurtosis. More generally, commonly used distributions such as the normal distribution have four moments. Still, as a mathematical matter, some distributions have infinite fourth moments, and this assumption rules out those distributions. If the assumption of finite fourth moments holds, then it is unlikely that statistical inferences using OLS will be dominated by a few observations.

Use of the Least Squares Assumptions

The three least squares assumptions for the linear regression model are summarized in Key Concept 4.3. The least squares assumptions play twin roles, and we return to them repeatedly throughout this text.

Their first role is mathematical: If these assumptions hold, then, as is shown in the next section, in large samples the OLS estimators are consistent and have sampling distributions that are normal. This large-sample normal distribution underpins methods for testing hypotheses and constructing confidence intervals using the OLS estimators.

KEY CONCEPT

4.3

The Least Squares Assumptions for Causal Inference

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n,$$

where β_1 is the causal effect on Y of X, and:

1. The error term u_i has conditional mean 0 given X_i: $E(u_i | X_i) = 0$;

2. $(X_i, Y_i), i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) draws from their joint distribution; and

3. Large outliers are unlikely: X_i and Y_i have nonzero finite fourth moments.

Their second role is to organize the circumstances that pose difficulties for OLS estimation of the causal effect β_1. As we will see, the first least squares assumption is the most important to consider in practice. One reason why the first least squares assumption might not hold in practice is discussed in Chapter 6, and additional reasons are discussed in Section 9.2.

It is also important to consider whether the second assumption holds in an application. Although it plausibly holds in many cross-sectional data sets, the independence assumption is inappropriate for panel and time series data. In those settings, some of the regression methods developed under assumption 2 require modifications. Those modifications are developed in Chapters 10 and 15–17.

The third assumption serves as a reminder that OLS, just like the sample mean, can be sensitive to large outliers. If your data set contains outliers, you should examine them carefully to make sure those observations are correctly recorded and belong in the data set.

The assumptions in Key Concept 4.3 apply when the aim is to estimate the causal effect—that is, when β_1 is the causal effect. Appendix 4.4 lays out a parallel set of least squares assumptions for prediction and discusses their relation to the assumptions in Key Concept 4.3.

4.5   The Sampling Distribution of the OLS Estimators

Because the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed from a randomly drawn sample, the estimators themselves are random variables with a probability distribution—the sampling distribution—that describes the values they could take over different possible random samples. In small samples, these sampling distributions are complicated, but in large samples, they are approximately normal because of the central limit theorem.

Review of the sampling distribution of $\overline{Y}$. Recall the discussion in Sections 2.5 and 2.6 about the sampling distribution of the sample average, $\overline{Y}$, an estimator of the unknown population mean of Y, μ_Y. Because $\overline{Y}$ is calculated using a randomly drawn sample, $\overline{Y}$ is a random variable that takes on different values from one sample to the next; the probability of these different values is summarized in its sampling distribution. Although the sampling distribution of $\overline{Y}$ can be complicated when the sample size is small, it is possible to make certain statements about it that hold for all n. In particular, the mean of the sampling distribution is μ_Y, that is, $E(\overline{Y}) = \mu_Y$, so $\overline{Y}$ is an unbiased estimator of μ_Y. If n is large, then more can be said about the sampling distribution. In particular, the central limit theorem (Section 2.6) states that this distribution is approximately normal.

The sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$. These ideas carry over to the OLS estimators β_0 and β_1 of the unknown intercept β_0 and slope β_1 of the population regression line. Because the OLS estimators are calculated using a random sample, $\hat{\beta}_0$ and $\hat{\beta}_1$ are

random variables that take on different values from one sample to the next; the probability of these different values is summarized in their sampling distributions.

Although the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ can be complicated when the sample size is small, it is possible to make certain statements about it that hold for all n. In particular, the means of the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are β_0 and β_1. In other words, under the least squares assumptions in Key Concept 4.3,

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1; \qquad (4.18)$$

that is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1. The proof that $\hat{\beta}_1$ is unbiased is given in Appendix 4.3, and the proof that $\hat{\beta}_0$ is unbiased is left as Exercise 4.7.

If the sample is sufficiently large, by the central limit theorem the joint sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is well approximated by the bivariate normal distribution (Section 2.4). This implies that the marginal distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal in large samples.

This argument invokes the central limit theorem. Technically, the central limit theorem concerns the distribution of averages (like $\overline{Y}$). If you examine the numerator in Equation (4.5) for $\hat{\beta}_1$, you will see that it, too, is a type of average—not a simple average, like $\overline{Y}$, but an average of the product, $(Y_i - \overline{Y})(X_i - \overline{X})$. As discussed further in Appendix 4.3, the central limit theorem applies to this average, so that, like the simpler average $\overline{Y}$, it is normally distributed in large samples.

The normal approximation to the distribution of the OLS estimators in large samples is summarized in Key Concept 4.4. (Appendix 4.3 summarizes the derivation of these formulas.) A relevant question in practice is how large n must be for these approximations to be reliable. In Section 2.6, we suggested that $n = 100$ is sufficiently large for the sampling distribution of $\overline{Y}$ to be well approximated by a normal distribution, and sometimes a smaller n suffices. This criterion carries over to the more complicated averages appearing in regression analysis. In virtually all modern

---

KEY CONCEPT

4.4

Large-Sample Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

If the least squares assumptions in Key Concept 4.3 hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a jointly normal sampling distribution. The large-sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$, where the variance of this distribution, $\sigma_{\hat{\beta}_1}^2$, is

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n} \frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}. \qquad (4.19)$$

The large-sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma_{\hat{\beta}_0}^2)$, where

$$\sigma_{\hat{\beta}_0}^2 = \frac{1}{n} \frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right] X_i. \qquad (4.20)$$

econometric applications, $n > 100$, so we will treat the normal approximations to the distributions of the OLS estimators as reliable unless there are good reasons to think otherwise.

The results in Key Concept 4.4 imply that the OLS estimators are consistent; that is, when the sample size is large and the least squares assumptions hold, $\hat{\beta}_0$ and $\hat{\beta}_1$ will be close to the true population coefficients β_0 and β_1 with high probability. This is because the variances $\sigma^2_{\hat{\beta}_0}$ and $\sigma^2_{\hat{\beta}_1}$ of the estimators decrease to 0 as n increases (n appears in the denominator of the formulas for the variances), so the distribution of the OLS estimators will be tightly concentrated around their means, β_0 and β_1, when n is large.

Another implication of the distributions in Key Concept 4.4 is that, in general, the larger is the variance of X_i, the smaller is the variance $\sigma^2_{\hat{\beta}_1}$ of $\hat{\beta}_1$. Mathematically, this implication arises because the variance of $\hat{\beta}_1$ in Equation (4.19) is inversely proportional to the square of the variance of X_i: the larger is $\text{var}(X_i)$, the larger is the denominator in Equation (4.19) so the smaller is $\sigma^2_{\hat{\beta}_1}$. To get a better sense of why this is so, look at Figure 4.5, which presents a scatterplot of 150 artificial data points on X and Y. The data points indicated by the colored dots are the 75 observations closest to $\overline{X}$. Suppose you were asked to draw a line as accurately as possible through *either* the colored or the black dots—which would you choose? It would be easier to draw a precise line through the black dots, which have a larger variance than the colored dots. Similarly, the larger the variance of X, the more precise is $\hat{\beta}_1$.

The distributions in Key Concept 4.4 also imply that the smaller is the variance of the error u_i, the smaller is the variance of $\hat{\beta}_1$. This can be seen mathematically in

FIGURE 4.5   The Variance of $\hat{\beta}_1$ and the Variance of X

The colored dots represent a set of X_i's with a small variance. The black dots represent a set of X_i's with a large variance. The regression line can be estimated more accurately with the black dots than with the colored dots.
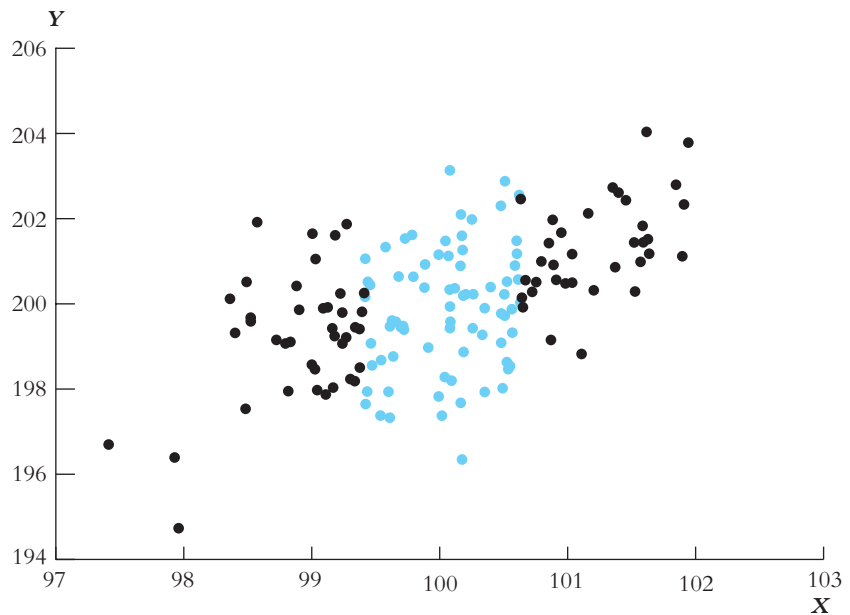
Equation (4.19) because u_i enters the numerator, but not denominator, of $\sigma^2_{\hat{\beta}_1}$: If all u_i were smaller by a factor of one-half but the X's did not change, then $\sigma_{\hat{\beta}_1}$ would be smaller by a factor of one-half and $\sigma^2_{\hat{\beta}_1}$ would be smaller by a factor of one-fourth (Exercise 4.13). Stated less mathematically, if the errors are smaller (holding the X's fixed), then the data will have a tighter scatter around the population regression line, so its slope will be estimated more precisely.

The normal approximation to the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is a powerful tool. With this approximation in hand, we are able to develop methods for making inferences about the true population values of the regression coefficients using only a sample of data.

4.6   Conclusion

This chapter has focused on the use of ordinary least squares to estimate the intercept and slope of a population regression line using a sample of n observations on a dependent variable, Y, and a single regressor, X. The sample regression line, estimated by OLS, can be used to predict Y given a value of X. When β_1 is defined to be the causal effect on Y of a unit change in X and the least squares assumptions for causal inference (Key Concept 4.3) hold, then the OLS estimators of the slope and intercept are unbiased, are consistent, and have a sampling distribution with a variance that is inversely proportional to the sample size n. Moreover, if n is large, then the sampling distribution of the OLS estimator is normal.

The first least squares assumption for causal inference is that the error term in the linear regression model has a conditional mean of 0 given the regressor X. This assumption holds if X is randomly assigned in an experiment or is as-if randomly assigned in observational data. Under this assumption, the OLS estimator is an unbiased estimator of the causal effect β_1.

The second least squares assumption is that (X_i, Y_i) are i.i.d., as is the case if the data are collected by simple random sampling. This assumption yields the formula, presented in Key Concept 4.4, for the variance of the sampling distribution of the OLS estimator.

The third least squares assumption is that large outliers are unlikely. Stated more formally, X and Y have finite fourth moments (finite kurtosis). This assumption is needed because OLS can be unreliable if there are large outliers. Taken together, the three least squares assumptions imply that the OLS estimator is normally distributed in large samples as described in Key Concept 4.4.

The results in this chapter describe the sampling distribution of the OLS estimator. By themselves, however, these results are not sufficient to test a hypothesis about the value of β_1 or to construct a confidence interval for β_1. Doing so requires an estimator of the standard deviation of the sampling distribution—that is, the standard error of the OLS estimator. This step—moving from the sampling distribution of $\hat{\beta}_1$ to its standard error, hypothesis tests, and confidence intervals—is taken in the next chapter.

Summary

1. The population regression line, $\beta_0 + \beta_1 X$, is the mean of Y as a function of the value of X. The slope, β_1, is the expected difference in Y between two observations with X values that differ by one unit. The intercept, β_0, determines the level (or height) of the regression line. Key Concept 4.1 summarizes the terminology of the population linear regression model.

2. The population regression line can be estimated using sample observations $(Y_i, X_i), i = 1, \ldots, n$, by ordinary least squares (OLS). The OLS estimators of the regression intercept and slope are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$. The predicted value of Y given X is $\hat{\beta}_0 + \hat{\beta}_1 X$.

3. The R^2 and standard error of the regression (SER) are measures of how close the values of Y_i are to the estimated regression line. The R^2 is between 0 and 1, with a larger value indicating that the Y_i's are closer to the line. The standard error of the regression estimates the standard deviation of the regression error.

4. There are three key assumptions for estimating causal effects using the linear regression model: (1) The regression errors, u_i, have a mean of 0, conditional on the regressors X_i; (2) the sample observations are i.i.d. random draws from the population; and (3) large outliers are unlikely. If these assumptions hold, the OLS estimator $\hat{\beta}_1$ is (1) an unbiased estimator of the causal effect β_1, (2) consistent, and (3) normally distributed when the sample is large.

Key Terms

causal inference (101)
prediction (101)
linear regression model with a single
    regressor (103)
dependent variable (103)
independent variable (103)
regressor (103)
population regression line (103)
population regression function (103)
intercept (103)
slope (103)
coefficients (103)
parameters (103)
error term (103)
ordinary least squares (OLS)
    estimators (107)

OLS regression line (107)
sample regression line (107)
sample regression function (107)
predicted value (107)
residual (107)
regression R^2 (111)
explained sum of squares
    (ESS) (111)
total sum of squares (TSS) (111)
sum of squared residuals
    (SSR) (111)
standard error of the regression
    (SER) (112)
in-sample prediction (113)
out-of-sample prediction (113)
least squares assumptions (115)

Review the Concepts

4.1 Explain the difference between $\hat{\beta}_1$ and β_1; between the residual $\hat{u}_i$ and the regression error u_i; and between the OLS predicted value $\hat{Y}_i$ and $E(Y_i | X_i)$.

4.2 For each least squares assumption, provide an example in which the assumption is valid, and then provide an example in which the assumption fails.

4.3 SER and R^2 are "measures of fit" for a regression. Explain how SER measures the fit of a regression. What are the units of SER? Explain how R^2 measures the fit of a regression. What are the units of R^2?

4.4 Sketch a hypothetical scatterplot of data for an estimated regression with $R^2 = 0.9$. Sketch a hypothetical scatterplot of data for a regression with $R^2 = 0.5$.

Exercises

4.1 A researcher, using data on class size (CS) and average test scores from 100 third-grade classes, estimates the OLS regression:

$$\widehat{TestScore} = 520.4 - 5.82 \times CS, R^2 = 0.08, SER = 11.5.$$

a. A classroom has 22 students. What is the regression's prediction for that classroom's average test score?

b. Last year a classroom had 19 students, and this year it has 23 students. What is the regression's prediction for the change in the classroom average test score?

c. The sample average class size across the 100 classrooms is 21.4. What is the sample average of the test scores across the 100 classrooms? (*Hint:* Review the formulas for the OLS estimators.)

d. What is the sample standard deviation of test scores across the 100 classrooms? (*Hint:* Review the formulas for the R^2 and SER.)

4.2 A random sample of 200 20-year-old men is selected from a population and these men's height and weight are recorded. A regression of weight on height yields

$$\widehat{Weight} = -99.41 + 3.94 \times Height, R^2 = 0.81, SER = 10.2,$$

where *Weight* is measured in pounds and *Height* is measured in inches.

  a. What is the regression's weight prediction for someone who is 70 in. tall? 65 in. tall? 74 in. tall?

  b. A man has a late growth spurt and grows 1.5 in. over the course of a year. What is the regression's prediction for the increase in this man's weight?

  c. Suppose that instead of measuring weight and height in pounds and inches, these variables are measured in centimeters and kilograms. What are the regression estimates from this new kilogram–centimeter regression? (Give all results, estimated coefficients, R^2, and *SER*.)

4.3 A regression of average weekly earnings (*AWE*, measured in dollars) on age (measured in years) using a random sample of college-educated full-time workers aged 25–65 yields the following:

$$\widehat{AWE} = 696.7 + 9.6 \times Age, R^2 = 0.023, SER = 624.1.$$

  a. Explain what the coefficient values 696.7 and 9.6 mean.

  b. The standard error of the regression (*SER*) is 624.1. What are the units of measurement for the *SER*? (Dollars? Years? Or is *SER* unit free?)

  c. The regression R^2 is 0.023. What are the units of measurement for the R^2? (Dollars? Years? Or is R^2 unit free?)

  d. What does the regression predict will be the earnings for a 25-year-old worker? For a 45-year-old worker?

  e. Will the regression give reliable predictions for a 99-year-old worker? Why or why not?

  f. Given what you know about the distribution of earnings, do you think it is plausible that the distribution of errors in the regression is normal? (*Hint:* Do you think that the distribution is symmetric or skewed? What is the smallest value of earnings, and is it consistent with a normal distribution?)

  g. The average age in this sample is 41.6 years. What is the average value of *AWE* in the sample? (*Hint:* Review Key Concept 4.2.)

4.4 Read the box "The 'Beta' of a Stock" in Section 4.2.

  a. Suppose the value of β is greater than 1 for a particular stock. Show that the variance of $(R - R_f)$ for this stock is greater than the variance of $(R_m - R_f)$.

b. Suppose the value of β is less than 1 for a particular stock. Is it possible that the variance of $(R - R_f)$ for this stock is greater than the variance of $(R_m - R_t)$? (*Hint:* Don't forget the regression error.)

c. In a given year, the rate of return on 3-month Treasury bills is 2.0% and the rate of return on a large diversified portfolio of stocks (the S&P 500) is 5.3%. For each company listed in the table in the box, use the estimated value of β to estimate the stock's expected rate of return.

4.5 A professor decides to run an experiment to measure the effect of time pressure on final exam scores. He gives each of the 400 students in his course the same final exam, but some students have 90 minutes to complete the exam, while others have 120 minutes. Each student is randomly assigned one of the examination times, based on the flip of a coin. Let Y_i denote the number of points scored on the exam by the ith student ($0 \le Y_i \le 100$), let X_i denote the amount of time that the student has to complete the exam ($X_i = 90$ or 120), and consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

a. Explain what the term u_i represents. Why will different students have different values of u_i?

b. Explain why $E(u_i | X_i) = 0$ for this regression model.

c. Are the other assumptions in Key Concept 4.3 satisfied? Explain.

d. The estimated regression is $\hat{Y}_i = 49 + 0.24 X_i$.

  i. Compute the estimated regression's prediction for the average score of students given 90 minutes to complete the exam. Repeat for 120 minutes and 150 minutes.

  ii. Compute the estimated gain in score for a student who is given an additional 10 minutes on the exam.

4.6 Show that the first least squares assumption, $E(u_i | X_i) = 0$, implies that $E(Y_i | X_i) = \beta_0 + \beta_1 X_i$.

4.7 Show that $\hat{\beta}_0$ is an unbiased estimator of β_0. (*Hint:* Use the fact that $\hat{\beta}_1$ is unbiased, which is shown in Appendix 4.3.)

4.8 Suppose all of the regression assumptions in Key Concept 4.3 are satisfied except that the first assumption is replaced with $E(u_i | X_i) = 2$. Which parts of Key Concept 4.4 continue to hold? Which change? Why? (Is $\hat{\beta}_1$ normally distributed in large samples with mean and variance given in Key Concept 4.4? What about $\hat{\beta}_0$?)

4.9 **a.** A linear regression yields $\hat{\beta}_1 = 0$. Show that $R^2 = 0$.

  b. A linear regression yields $R^2 = 0$. Does this imply that $\hat{\beta}_1 = 0$?

4.10 Suppose $Y_i = \beta_0 + \beta_1 X_i + u_i$, where (X_i, u_i) are i.i.d. and X_i is a Bernoulli random variable with $\Pr(X = 1) = 0.20$. When $X = 1$, u_i is $N(0, 4)$; when $X = 0$, u_i is $N(0, 1)$.

    a. Show that the regression assumptions in Key Concept 4.3 are satisfied.

    b. Derive an expression for the large-sample variance of $\hat{\beta}_1$. [*Hint:* Evaluate the terms in Equation (4.19).]

4.11 Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

    a. Suppose you know that $\beta_0 = 0$. Derive a formula for the least squares estimator of β_1.

    b. Suppose you know that $\beta_0 = 4$. Derive a formula for the least squares estimator of β_1.

4.12 a. Show that the regression R^2 in the regression of Y on X is the squared value of the sample correlation between X and Y. That is, show that $R^2 = r_{XY}^2$.

    b. Show that the R^2 from the regression of Y on X is the same as the R^2 from the regression of X on Y.

    c. Show that $\hat{\beta}_1 = r_{XY}(s_Y/s_X)$, where r_{XY} is the sample correlation between X and Y and s_X and s_Y are the sample standard deviations of X and Y.

4.13 Suppose $Y_i = \beta_0 + \beta_1 X_i + \kappa u_i$, where κ is a nonzero constant and (Y_i, X_i) satisfy the three least squares assumptions. Show that the large-sample variance of $\hat{\beta}_1$ is given by $\sigma_{\hat{\beta}_1}^2 = \kappa^2 \frac{1}{n} \frac{\mathrm{var}[(X_i - \mu_X)u_i]}{[\mathrm{var}(X_i)^2]}$. [*Hint:* This equation is the variance given in Equation (4.19) multiplied by κ^2.]

4.14 Show that the sample regression line passes through the point $(\overline{X}, \overline{Y})$.

4.15 (Requires Appendix 4.4) A sample (X_i, Y_i), $i = 1, \ldots, n$, is collected from a population with $E(Y|X) = \beta_0 + \beta_1 X$ and used to compute the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. You are interested in predicting the value of Y^{oos} from a randomly chosen out-of-sample observation with $X^{oos} = x^{oos}$.

    a. Suppose the out-of-sample observation is from the same population as the in-sample observations (X_i, Y_i) and is chosen independently of the in-sample observations.

        i. Explain why $E(Y^{oos}|X^{oos} = x^{oos}) = \beta_0 + \beta_1 x^{oos}$.

        ii. Let $\hat{Y}^{oos} = \hat{\beta}_0 + \hat{\beta}_1 x^{oos}$. Show that
$E(\hat{Y}^{oos}|X^{oos} = x^{oos}) = \beta_0 + \beta_1 x^{oos}$.

        iii. Let $u^{oos} = Y^{oos} - (\beta_0 + \beta_1 X^{oos})$ and $\hat{u}^{oos} = Y^{oos} - (\hat{\beta}_0 + \hat{\beta}_1 X^{oos})$. Show that $\mathrm{var}(\hat{u}^{oos}) = \mathrm{var}(u^{oos}) + \mathrm{var}(\hat{\beta}_0 + \hat{\beta}_1 X^{oos})$.

    b. Suppose the out-of-sample observation is drawn from a different population than the in-sample population and that the joint distributions of X and Y differ for the two populations. Continue to let β_0 and β_1

be the coefficients of the population regression line for the in-sample population.

  i. Does $E(Y^{oos}|X^{oos} = x^{oos}) = \beta_0 + \beta_1 x^{oos}$?

  ii. Does $E(\hat{Y}^{oos}|X^{oos} = x^{oos}) = \beta_0 + \beta_1 x^{oos}$?

Empirical Exercises

E4.1 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find the data file **Growth**, which contains data on average growth rates from 1960 through 1995 for 65 countries, along with variables that are potentially related to growth.[1] A detailed description is given in **Growth_Description**, also available on the website. In this exercise, you will investigate the relationship between growth and trade.

    a. Construct a scatterplot of average annual growth rate (*Growth*) on the average trade share (*TradeShare*). Does there appear to be a relationship between the variables?

    b. One country, Malta, has a trade share much larger than the other countries. Find Malta on the scatterplot. Does Malta look like an outlier?

    c. Using all observations, run a regression of *Growth* on *TradeShare*. What is the estimated slope? What is the estimated intercept? Use the regression to predict the growth rate for a country with a trade share of 0.5 and for another with a trade share equal to 1.0.

    d. Estimate the same regression, excluding the data from Malta. Answer the same questions in (c).

    e. Plot the estimated regression functions from (c) and (d). Using the scatterplot in (a), explain why the regression function that includes Malta is steeper than the regression function that excludes Malta.

    f. Where is Malta? Why is the Malta trade share so large? Should Malta be included or excluded from the analysis?

E4.2 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find the data file **Earnings_and_Height**, which contains data on earnings, height, and other characteristics of a random sample of U.S. workers.[2]

[1]These data were provided by Professor Ross Levine of the University of California at Berkeley and were used in his paper with Thorsten Beck and Norman Loayza, "Finance and the Sources of Growth," *Journal of Financial Economics*, 2000, 58: 261–300.

[2] These data were provided by Professors Anne Case (Princeton University) and Christina Paxson (Brown University) and were used in their paper "Stature and Status: Height, Ability, and Labor Market Outcomes," *Journal of Political Economy, 2008*, 116(3): 499–532.

A detailed description is given in **Earnings_and_Height_Description**, also available on the website. In this exercise, you will investigate the relationship between earnings and height.

a. What is the median value of height in the sample?

b. i. Estimate average earnings for workers whose height is at most 67 inches.

   ii. Estimate average earnings for workers whose height is greater than 67 inches.

   iii. On average, do taller workers earn more than shorter workers? How much more? What is a 95% confidence interval for the difference in average earnings?

c. Construct a scatterplot of annual earnings (*Earnings*) on height (*Height*). Notice that the points on the plot fall along horizontal lines. (There are only 23 distinct values of *Earnings*). Why? (*Hint:* Carefully read the detailed data description.)

d. Run a regression of *Earnings* on *Height*.

   i. What is the estimated slope?

   ii. Use the estimated regression to predict earnings for a worker who is 67 inches tall, for a worker who is 70 inches tall, and for a worker who is 65 inches tall.

e. Suppose height were measured in centimeters instead of inches. Answer the following questions about the *Earnings* on *Height* (in cm) regression.

   i. What is the estimated slope of the regression?

   ii. What is the estimated intercept?

   iii. What is the R^2?

   iv. What is the standard error of the regression?

f. Run a regression of *Earnings* on *Height*, using data for female workers only.

   i. What is the estimated slope?

   ii. A randomly selected woman is 1 inch taller than the average woman in the sample. Would you predict her earnings to be higher or lower than the average earnings for women in the sample? By how much?

g. Repeat (f) for male workers.

h. Do you think that height is uncorrelated with other factors that cause earning? That is, do you think that the regression error term, u_i has a conditional mean of 0 given *Height* (X_i)? (You will investigate this more in the *Earnings* and *Height* exercises in later chapters.)

4.1 The California Test Score Data Set

The California Standardized Testing and Reporting data set contains data on test performance, school characteristics, and student demographic backgrounds. The data used here are from all 420 K–6 and K–8 districts in California with data available for 1999. Test scores are the average of the reading and math scores on the Stanford 9 Achievement Test, a standardized test administered to fifth-grade students. School characteristics (averaged across the district) include enrollment, number of teachers (measured as "full-time equivalents"), number of computers per classroom, and expenditures per student. The student–teacher ratio used here is the number of students in the district divided by the number of full-time equivalent teachers. Demographic variables for the students also are averaged across the district. The demographic variables include the percentage of students who are in the public assistance program CalWorks (formerly AFDC), the percentage of students who qualify for a reduced-price lunch, and the percentage of students who are English learners (that is, students for whom English is a second language). All of these data were obtained from the California Department of Education (**www.cde.ca.gov**).

4.2 Derivation of the OLS Estimators

This appendix uses calculus to derive the formulas for the OLS estimators given in Key Concept 4.2. To minimize the sum of squared prediction mistakes $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ [Equation (4.4)], first take the partial derivatives with respect to b_0 and b_1:

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) \text{ and} \tag{4.21}$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)X_i. \tag{4.22}$$

The OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, are the values of b_0 and b_1 that minimize $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ or, equivalently, the values of b_0 and b_1 for which the derivatives in Equations (4.21) and (4.22) equal 0. Accordingly, setting these derivatives equal to 0, collecting terms, and dividing by n shows that the OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy the two equations

$$\overline{Y} - \hat{\beta}_0 - \hat{\beta}_1\overline{X} = 0 \text{ and} \tag{4.23}$$

$$\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \hat{\beta}_0\overline{X} - \hat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}X_i^2 = 0. \tag{4.24}$$

Solving this pair of equations for $\hat{\beta}_0$ and $\hat{\beta}_1$ yields

$$\hat{\beta}_1 = \frac{\dfrac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}}{\dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{4.25}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.26}$$

Equations (4.25) and (4.26) are the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ given in Key Concept 4.2; the formula $\hat{\beta}_1 = s_{XY}/s_X^2$ is obtained by dividing the numerator and denominator in Equation (4.25) by $n-1$.

4.3 Sampling Distribution of the OLS Estimator

In this appendix, we show that the OLS estimator $\hat{\beta}_1$ is unbiased and, in large samples, has the normal sampling distribution given in Key Concept 4.4.

Representation of $\hat{\beta}_1$ in Terms of the Regressors and Errors

We start by providing an expression for $\hat{\beta}_1$ in terms of the regressors and errors. Because $Y_i = \beta_0 + \beta_1 X_i + u_i$, $Y_i - \overline{Y} = \beta_1(X_i - \overline{X}) + u_i - \overline{u}$, so the numerator of the formula for $\hat{\beta}_1$ in Equation (4.25) is

$$\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n}(X_i - \overline{X})[\beta_1(X_i - \overline{X}) + (u_i - \overline{u})]$$

$$= \beta_1 \sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}). \tag{4.27}$$

Now $\sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}) = \sum_{i=1}^{n}(X_i - \overline{X})u_i - \sum_{i=1}^{n}(X_i - \overline{X})\overline{u} = \sum_{i=1}^{n}(X_i - \overline{X})u_i$, where the final equality follows from the definition of $\overline{X}$, which implies that $\sum_{i=1}^{n}(X_i - \overline{X})\overline{u} = (\sum_{i=1}^{n}X_i - n\overline{X})\overline{u} = 0$. Substituting $\sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}) = \sum_{i=1}^{n}(X_i - \overline{X})u_i$ into the final expression in Equation (4.27) yields $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \beta_1 \sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{n}(X_i - \overline{X})u_i$. Substituting this expression in turn into the formula for $\hat{\beta}_1$ in Equation (4.25) yields

$$\hat{\beta}_1 = \beta_1 + \frac{\dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{4.28}$$

Proof That $\hat{\beta}_1$ Is Unbiased

The argument that $\hat{\beta}_1$ is unbiased under the first least squares assumption uses the law of iterated expectations [Equation (2.20)]. First, obtain $E(\hat{\beta}_1 | X_1, \ldots, X_n)$ by taking the conditional expectation of both sides of Equation (4.28):

$$
E(\hat{\beta}_1 | X_1, \ldots, X_n) = \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \middle| X_1, \ldots, X_n \right]
$$

$$
= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})E(u_i | X_i, \ldots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{4.29}
$$

By the second least squares assumption, u_i is distributed independently of X for all observations other than i, so $E(u_i | X_1, \ldots, X_n) = E(u_i | X_i)$. By the first least squares assumption, however, $E(u_i | X_i) = 0$. Thus the second term in the final line of Equation (4.29) is 0, from which it follows that $E(\hat{\beta}_1 | X_1, \ldots, X_n) = \beta_1$.

Because $\hat{\beta}_1$ is unbiased given $X_1, \ldots, X_n$, it is unbiased after averaging over all samples $X_1, \ldots, X_n$. Thus the unbiasedness of $\hat{\beta}_1$ follows Equation (4.29) and the law of iterated expectations: $E(\hat{\beta}_1) = E[E(\hat{\beta}_1 | X_1, \ldots, X_n)] = \beta_1$.

Large-Sample Normal Distribution of the OLS Estimator

The large-sample normal approximation to the limiting distribution of $\hat{\beta}_1$ (Key Concept 4.4) is obtained by considering the behavior of the final term in Equation (4.28).

First, consider the numerator of this term. Because $\overline{X}$ is consistent, if the sample size is large, $\overline{X}$ is nearly equal to μ_X. Thus, to a close approximation, the term in the numerator of Equation (4.28) is the sample average $\overline{v}$, where $v_i = (X_i - \mu_X)u_i$. By the first least squares assumption, v_i has a mean of 0. By the second least squares assumption, v_i is i.i.d. The variance of v_i is $\sigma_v^2 = [\text{var}(X_i - \mu_X)u_i]$, which, by the third least squares assumption, is nonzero and finite. Therefore, $\overline{v}$ satisfies all the requirements of the central limit theorem (Key Concept 2.7). Thus $\overline{v}/\sigma_{\overline{v}}$ is, in large samples, distributed $N(0, 1)$, where $\sigma_{\overline{v}}^2 = \sigma_v^2/n$. Therefore the distribution of $\overline{v}$ is well approximated by the $N(0, \sigma_v^2/n)$ distribution.

Next consider the expression in the denominator in Equation (4.28); this is the sample variance of X (except dividing by n rather than $n - 1$, which is inconsequential if n is large). As discussed in Section 3.2 [Equation (3.8)], the sample variance is a consistent estimator of the population variance, so in large samples it is arbitrarily close to the population variance of X.

Combining these two results, we have that, in large samples, $\hat{\beta}_1 - \beta_1 \cong \overline{v}/\text{var}(X_i)$, so that the sampling distribution of $\hat{\beta}_1$ is, in large samples, $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$, where $\sigma_{\hat{\beta}_1}^2 = \text{var}(\overline{v})/[\text{var}(X_i)]^2 = \text{var}[(X_i - \mu_X)u_i]/\{n[\text{var}(X_i)]^2\}$, which is the expression in Equation (4.19).

Some Additional Algebraic Facts About OLS

The OLS residuals and predicted values satisfy

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0, \tag{4.30}$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \overline{Y}, \tag{4.31}$$

$$\sum_{i=1}^{n}\hat{u}_i X_i = 0 \text{ and } s_{\hat{u}X} = 0, \text{ and} \tag{4.32}$$

$$TSS = SSR + ESS. \tag{4.33}$$

Equations (4.30) through (4.33) say that the sample average of the OLS residuals is 0; the sample average of the OLS predicted values equals $\overline{Y}$; the sample covariance $s_{\hat{u}X}$ between the OLS residuals and the regressors is 0; and the total sum of squares is the sum of squared residuals and the explained sum of squares. [The *ESS*, *TSS*, and *SSR* are defined in Equations (4.12), (4.13), and (4.15).]

To verify Equation (4.30), note that the definition of $\hat{\beta}_0$ lets us write the OLS residuals as $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})$; thus

$$\sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}(Y_i - \overline{Y}) - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \overline{X}).$$

But the definitions of $\overline{Y}$ and $\overline{X}$ imply that $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$ and $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$, so $\sum_{i=1}^{n}\hat{u}_i = 0$.

To verify Equation (4.31), note that $Y_i = \hat{Y}_i + \hat{u}_i$, so $\sum_{i=1}^{n}Y_i = \sum_{i=1}^{n}\hat{Y}_i + \sum_{i=1}^{n}\hat{u}_1 = \sum_{i=1}^{n}\hat{Y}_i$, where the second equality is a consequence of Equation (4.30).

To verify Equation (4.32), note that $\sum_{i=1}^{n}\hat{u}_i = 0$ implies $\sum_{i=1}^{n}\hat{u}_i X_i = \sum_{i=1}^{n}\hat{u}_i(X_i - \overline{X})$, so

$$\sum_{i=1}^{n}\hat{u}_i X_i = \sum_{i=1}^{n}[(Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})](X_i - \overline{X})$$

$$= \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}) - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \overline{X})^2 = 0, \tag{4.34}$$

where the final equality in Equation (4.34) is obtained using the formula for $\hat{\beta}_1$ in Equation (4.25). This result, combined with the preceding results, implies that $s_{\hat{u}X} = 0$.

Equation (4.33) follows from the previous results and some algebra:

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2$$

$$= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y})$$

$$= SSR + ESS + 2\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = SSR + ESS, \tag{4.35}$$

where the final equality follows from $\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = \sum_{i=1}^{n}\hat{u}_i(\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0\sum_{i=1}^{n}\hat{u}_i + \hat{\beta}_1\sum_{i=1}^{n}\hat{u}_i X_i = 0$ by the previous results.

4.4 The Least Squares Assumptions for Prediction

Section 4.4 provides the least squares assumptions for estimation of a causal effect. There is a parallel set of least squares assumptions for prediction. The difference between the two stems from the difference between the two problems. For estimation of a causal effect, X must be randomly assigned or as-if randomly assigned, which leads to least squares assumption 1 in Key Concept 4.3. In contrast, as discussed in Section 4.3, the goal of prediction is to provide accurate out-of-sample predictions. To do so, the estimated regression line must be relevant to the observation being predicted. This is the case if the data used for estimation and the observation being predicted are drawn from the same population distribution.

For example, return to the superintendent's and father's problems. The superintendent is interested in the causal effect on *TestScore* of a change in *STR*. Ideally, to answer her question we would have data from an experiment in which students were randomly assigned to different size classes. Absent such an experiment, she may or may not be satisfied with the regression of *TestScore* on *STR* using California data—that depends on whether least squares assumption 1 is satisfied where β_1 is defined to be the causal effect.

In contrast, the father is interested in predicting test scores in a California district that did not report its test scores, so for his purposes he is interested in the population regression line relating *TestScore* and *STR* in California, the slope of which may or may not be the causal effect.

To make this precise, we introduce some additional notation. Let (X^{oos}, Y^{oos}) denote the out-of-sample ("oos") observation for which the prediction is to be made, and continue to let $(X_i, Y_i), i = 1, \ldots, n$, be the data used to estimate the regression coefficients. The least squares assumptions for prediction are

$$E(Y|X) = \beta_0 + \beta_1 X \text{ and } u = Y - E(Y|X), \text{ where}$$

1. (X^{oos}, Y^{oos}) are randomly drawn from the same population distribution as $(X_i, Y_i), i = 1, \ldots, n$;
2. $(X_i, Y_i), i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) draws from their joint distribution; and
3. Large outliers are unlikely: X_i and Y_i have nonzero finite fourth moments.

There are two differences between these assumptions and the assumptions in Key Concept 4.3. The first is the definition of β_1. The best predictor is given by $E(Y|X)$ (where the best predictor is defined in terms of the mean squared prediction error; see Appendix 2.2). With the assumption of linearity, for prediction β_1 is defined to be the slope of this conditional expectation, which may or may not be the causal effect. Second, because the regression line is estimated using in-sample observations but is used to predict an out-of-sample observation, the first assumption is that these are drawn from the same population.

The second and third assumptions are the same as those for estimation of causal effects in Section 4.4. They ensure that the OLS estimators are consistent for the coefficients of the population prediction line and are normally distributed when n is large.

Under the least squares assumptions for prediction, the OLS predicted value of Y^{oos} is unbiased:

$$E(\hat{Y}^{oos}|X^{oos} = x^{oos}) = E(\hat{\beta}_0 + \hat{\beta}_0 X^{oos}|X^{oos} = x^{oos})$$

$$= E(\hat{\beta}_0) + E(\hat{\beta}_1)x^{oos} \qquad (4.36)$$

where the second equality follows because (X^{oos}, Y^{oos}) are independent of the observations used to compute the OLS estimators. For the prediction problem, u is defined to be $u = Y - E(Y|X)$, so by definition $E(u_i|X_i) = 0$ and the algebra in Appendix 4.3 applies directly. Thus $E(\hat{\beta}_0) + E(\hat{\beta}_1)x^{oos} = \beta_0 + \beta_1 x^{oos} = E(Y^{oos}|X^{oos} = x^{oos})$. Combining this expression with the first expression in Equation (4.36), we have that $E(Y^{oos} - \hat{Y}^{oos}|X^{oos} = x^{oos}) = 0$; that is, the OLS prediction is unbiased.

The least squares assumptions for prediction also ensure that the regression *SER* estimates the variance of the out-of-sample prediction error, $\hat{u}^{oos} = Y^{oos} - \hat{Y}^{oos}$. To show this, it is useful to write the out-of-sample prediction error as the sum of two terms: the error that would be made were the regression coefficients known and the error made by needing to estimate them. Write $\hat{u}^{oos} = Y^{oos} - (\hat{\beta}_0 + \hat{\beta}_1 X^{oos}) = \beta_0 + \beta_1 X^{oos} + u^{oos} - (\hat{\beta}_0 + \hat{\beta}_1 X^{oos}) = u^{oos} - [(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)X^{oos}]$. Thus $\text{var}(\hat{u}^{oos}) = \text{var}(u^{oos}) + \text{var}(\hat{\beta}_0 + \hat{\beta}_1 X^{oos})$ (Exercise 4.15). The second term in this final expression is the contribution of the estimation error to the out-of-sample prediction error. When the sample size is large, the first term in this final expression is much larger than the second term. Because the in- and out-of-sample observations are from the same population, $\text{var}(u^{oos}) = \text{var}(u_i) = \sigma_u^2$, so the standard deviation of $\hat{u}^{oos}$ is estimated by the *SER*.

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals

This chapter continues the treatment of linear regression with a single regressor. Chapter 4 explained how the OLS estimator $\hat{\beta}_1$ of the slope coefficient β_1 differs from one sample to the next—that is, how $\hat{\beta}_1$ has a sampling distribution. In this chapter, we show how knowledge of this sampling distribution can be used to make statements about β_1 that accurately summarize the sampling uncertainty. The starting point is the standard error of the OLS estimator, which measures the spread of the sampling distribution of $\hat{\beta}_1$. Section 5.1 provides an expression for this standard error (and for the standard error of the OLS estimator of the intercept) and then shows how to use $\hat{\beta}_1$ and its standard error to test hypotheses. Section 5.2 explains how to construct confidence intervals for β_1. Section 5.3 takes up the special case of a binary regressor.

Sections 5.1 through 5.3 assume that the three least squares assumptions of Key Concept 4.3 hold. If, in addition, some stronger technical conditions hold, then some stronger results can be derived regarding the distribution of the OLS estimator. One of these stronger conditions is that the errors are homoskedastic, a concept introduced in Section 5.4. Section 5.5 presents the Gauss–Markov theorem, which states that, under certain conditions, OLS is efficient (has the smallest variance) among a certain class of estimators. Section 5.6 discusses the distribution of the OLS estimator when the population distribution of the regression errors is normal.

5.1 Testing Hypotheses About One of the Regression Coefficients

Your client, the superintendent, calls you with a problem. She has an angry taxpayer in her office who asserts that cutting class size will not help boost test scores, so hiring more teachers is a waste of money. Class size, the taxpayer claims, has no effect on test scores.

The taxpayer's claim can be restated in the language of regression analysis: The taxpayer is asserting that the true causal effect on test scores of a change in class size is 0; that is, $\beta_{ClassSize} = 0$.

You already provided the superintendent with an estimate of $\beta_{ClassSize}$ using your sample of 420 observations on California school districts, under the assumption that the least squares assumptions of Key Concept 4.3 hold. Is there, the superintendent asks, evidence in your data this slope is nonzero? Can you reject the taxpayer's hypothesis that $\beta_{ClassSize} = 0$, or should you accept it, at least tentatively pending further new evidence?

General Form of the *t*-Statistic

In general, the *t*-statistic has the form

$$t = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error of the estimator}}. \qquad (5.1)$$

This section discusses tests of hypotheses about the population coefficients β_0 and β_1. We start by discussing two-sided tests of β_1 in detail, then turn to one-sided tests and to tests of hypotheses regarding the intercept β_0.

Two-Sided Hypotheses Concerning β_1

The general approach to testing hypotheses about the coefficient β_1 is the same as to testing hypotheses about the population mean, so we begin with a brief review.

Testing hypotheses about the population mean. Recall from Section 3.2 that the null hypothesis that the mean of Y is a specific value $\mu_{Y,0}$ can be written as H_0: $E(Y) = \mu_{Y,0}$, and the two-sided alternative is H_1: $E(Y) \neq \mu_{Y,0}$.

The test of the null hypothesis H_0 against the two-sided alternative proceeds as in the three steps summarized in Key Concept 3.6. The first is to compute the standard error of $\overline{Y}$, $SE(\overline{Y})$, which is an estimator of the standard deviation of the sampling distribution of $\overline{Y}$. The second step is to compute the *t*-statistic, which has the general form given in Key Concept 5.1; applied here, the *t*-statistic is $t = (\overline{Y} - \mu_{Y,0})/SE(\overline{Y})$.

The third step is to compute the *p*-value, which is the smallest significance level at which the null hypothesis could be rejected, based on the test statistic actually observed; equivalently, the *p*-value is the probability of obtaining a statistic, by random sampling variation, at least as different from the null hypothesis value as is the statistic actually observed, assuming that the null hypothesis is correct (Key Concept 3.5). Because the *t*-statistic has a standard normal distribution in large samples under the null hypothesis, the *p*-value for a two-sided hypothesis test is $2\Phi(-|t^{act}|)$, where t^{act} is the value of the *t*-statistic actually computed and Φ is the cumulative standard normal distribution tabulated in Appendix Table 1. Alternatively, the third step can be replaced by simply comparing the *t*-statistic to the critical value appropriate for the test with the desired significance level. For example, a two-sided test with a 5% significance level would reject the null hypothesis if $|t^{act}| > 1.96$. In this case, the population mean is said to be statistically significantly different from the hypothesized value at the 5% significance level.

Testing hypotheses about the slope β_1. At a theoretical level, the critical feature justifying the foregoing testing procedure for the population mean is that, in large samples, the sampling distribution of $\overline{Y}$ is approximately normal. Because $\hat{\beta}_1$ also has a normal sampling distribution in large samples, hypotheses about the true value of the slope β_1 can be tested using the same general approach.

The null and alternative hypotheses need to be stated precisely before they can be tested. The angry taxpayer's hypothesis is that $\beta_{ClassSize} = 0$. More generally, under the null hypothesis the true population coefficient β_1 takes on some specific value, $\beta_{1,0}$. Under the two-sided alternative, β_1 does not equal $\beta_{1,0}$. That is, the **null hypothesis** and the **two-sided alternative hypothesis** are

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq \beta_{1,0} \text{ (two-sided alternative)}. \tag{5.2}$$

To test the null hypothesis H_0, we follow the same three steps as for the population mean.

The first step is to compute the **standard error of $\hat{\beta}_1$**, $SE(\hat{\beta}_1)$. The standard error of $\hat{\beta}_1$ is an estimator of $\sigma_{\hat{\beta}_1}$, the standard deviation of the sampling distribution of $\hat{\beta}_1$. Specifically,

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}, \tag{5.3}$$

where

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\dfrac{1}{n-2}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 \hat{u}_i^2}{\left[\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}. \tag{5.4}$$

The estimator of the variance in Equation (5.4) is discussed in Appendix 5.1. Although the formula for $\hat{\sigma}^2_{\hat{\beta}_1}$ is complicated, in applications the standard error is computed by regression software so that it is easy to use in practice.

The second step is to compute the *t*-statistic,

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}. \tag{5.5}$$

The third step is to compute the *p*-value, the probability of observing a value of $\hat{\beta}_1$ at least as different from $\beta_{1,0}$ as the estimate actually computed ($\hat{\beta}_1^{act}$), assuming that the null hypothesis is correct. Stated mathematically,

$$p\text{-value} = \Pr\nolimits_{H_0}\left[|\hat{\beta}_1 - \beta_1| > |\hat{\beta}_1^{act} - \beta_{1,0}|\right]$$

$$= \Pr\nolimits_{H_0}\left[\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1^{act} - \beta_{1,0}}{SE(\hat{\beta}_1)}\right|\right] = \Pr\nolimits_{H_0}(|t| > |t^{act}|), \tag{5.6}$$

Testing the Hypothesis $\beta_1 = \beta_{1,0}$ Against the Alternative $\beta_1 \neq \beta_{1,0}$

1. Compute the standard error of $\hat{\beta}_1$, $SE(\hat{\beta}_1)$ [Equation (5.3)].
2. Compute the *t*-statistic [Equation (5.5)].
3. Compute the *p*-value [Equation (5.7)]. Reject the hypothesis at the 5% significance level if the *p*-value is less than 0.05 or, equivalently, if $|t^{act}| > 1.96$.

The standard error and (typically) the *t*-statistic and *p*-value testing $\beta_1 = 0$ are computed automatically by regression software.

where Pr_{H_0} denotes the probability computed under the null hypothesis, the second equality follows by dividing by $SE(\hat{\beta}_1)$, and t^{act} is the value of the *t*-statistic actually computed. Because $\hat{\beta}_1$ is approximately normally distributed in large samples, under the null hypothesis the *t*-statistic is approximately distributed as a standard normal random variable, so in large samples

$$p\text{-value} = \mathrm{Pr}(|Z| > |t^{act}|) = 2\Phi(-|t^{act}|). \tag{5.7}$$

A *p*-value of less than 5% provides evidence against the null hypothesis in the sense that, under the null hypothesis, the probability of obtaining a value of $\hat{\beta}_1$ at least as far from the null as that actually observed is less than 5%. If so, the null hypothesis is rejected at the 5% significance level.

Alternatively, the hypothesis can be tested at the 5% significance level simply by comparing the absolute value of the *t*-statistic to 1.96, the critical value for a two-sided test, and rejecting the null hypothesis at the 5% level if $|t^{act}| > 1.96$.

These steps are summarized in Key Concept 5.2.

Reporting regression equations and application to test scores. The OLS regression of the test score against the student–teacher ratio, reported in Equation (4.9), yielded $\hat{\beta}_0 = 698.9$ and $\hat{\beta}_1 = -2.28$. The standard errors of these estimates are $SE(\hat{\beta}_0) = 10.4$ and $SE(\hat{\beta}_1) = 0.52$.

Because of the importance of the standard errors, by convention they are included when reporting the estimated OLS coefficients. One compact way to report the standard errors is to place them in parentheses below the respective coefficients of the OLS regression line:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR, R^2 = 0.051, SER = 18.6. \tag{5.8}$$
$$(10.4) \quad (0.52)$$

Equation (5.8) also reports the regression R^2 and the standard error of the regression (SER) following the estimated regression line. Thus Equation (5.8) provides the estimated regression line, estimates of the sampling uncertainty of the slope and the

---

FIGURE 5.1 Calculating the *p*-Value of a Two-Sided Test When $t^{act} = -4.38$

The *p*-value of a two-sided test is the probability that $|Z| > |t^{act}|$, where Z is a standard normal random variable and t^{act} is the value of the *t*-statistic calculated from the sample. When $t^{act} = -4.38$, the *p*-value is only 0.00001.

$N(0, 1)$

−4.38   0   4.38 z

The *p*-value is the area
to the left of –4.38
+
the area to the right of +4.38.

---
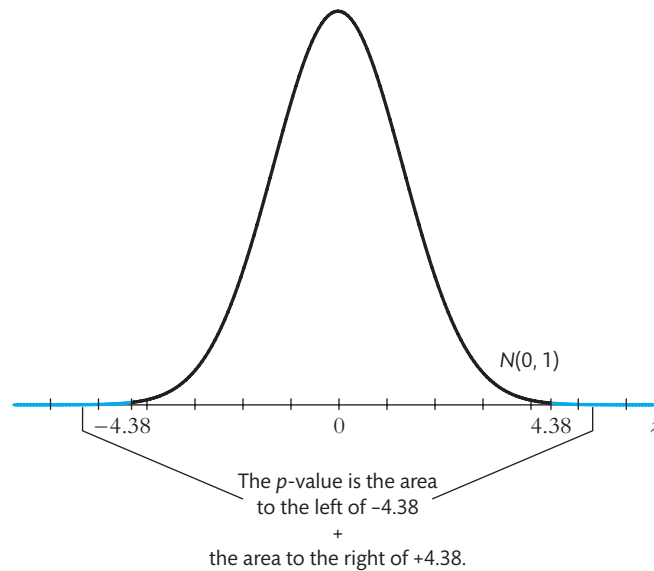
intercept (the standard errors), and two measures of the fit of this regression line (the R^2 and the *SER*). This is a common format for reporting a single regression equation, and it will be used throughout the rest of this text.

Suppose you wish to test the null hypothesis that the slope β_1 is 0 in the population counterpart of Equation (5.8) at the 5% significance level. To do so, construct the *t*-statistic, and compare its absolute value to 1.96, the 5% (two-sided) critical value taken from the standard normal distribution. The *t*-statistic is constructed by substituting the hypothesized value of β_1 under the null hypothesis (0), the estimated slope, and its standard error from Equation (5.8) into the general formula in Equation (5.5); the result is $t^{act} = (-2.280)/0.52 = -4.38$. The absolute value of this *t*-statistic exceeds the 5% two-sided critical value of 1.96, so the null hypothesis is rejected in favor of the two-sided alternative at the 5% significance level.

Alternatively, we can compute the *p*-value associated with $t^{act} = -4.38$. This probability is the area in the tails of the standard normal distribution, as shown in Figure 5.1. This probability is extremely small, approximately 0.00001, or 0.001%. That is, if the null hypothesis $\beta_{ClassSize} = 0$ is true, the probability of obtaining a value of $\hat{\beta}_1$ as far from the null as the value we actually obtained is extremely small, less than 0.001%. Because this event is so unlikely, it is reasonable to conclude that the null hypothesis is false.

One-Sided Hypotheses Concerning β_1

The discussion so far has focused on testing the hypothesis that $\beta_1 = \beta_{1,0}$ against the hypothesis that $\beta_1 \neq \beta_{1,0}$. This is a two-sided hypothesis test because, under the

alternative, β_1 could be either larger or smaller than $\beta_{1,0}$. Sometimes, however, it is appropriate to use a one-sided hypothesis test. For example, in the student–teacher ratio/test score problem, many people think that smaller classes provide a better learning environment. Under that hypothesis, β_1 is negative: Smaller classes lead to higher scores. It might make sense therefore to test the null hypothesis that $\beta_1 = 0$ (no effect) against the one-sided alternative that $\beta_1 < 0$.

For a one-sided test, the null hypothesis and the one-sided alternative hypothesis are

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 < \beta_{1,0} \text{ (one-sided alternative)}, \tag{5.9}$$

where $\beta_{1,0}$ is the value of β_1 under the null (0 in the student–teacher ratio example) and the alternative is that β_1 is less than $\beta_{1,0}$. If the alternative is that β_1 is greater than $\beta_{1,0}$, the inequality in Equation (5.9) is reversed.

Because the null hypothesis is the same for a one- and a two-sided hypothesis test, the construction of the t-statistic is the same. The only difference between a one- and a two-sided hypothesis test is how you interpret the t-statistic. For the one-sided alternative in Equation (5.9), the null hypothesis is rejected against the one-sided alternative for large negative values, but not large positive values, of the t-statistic: Instead of rejecting if $|t^{act}| > 1.96$, the hypothesis is rejected at the 5% significance level if $t^{act} < -1.64$.

The p-value for a one-sided test is obtained from the cumulative standard normal distribution as

$$p\text{-value} = \Pr(Z < t^{act}) = \Phi(t^{act}) \text{ } (p\text{-value, one-sided left-tail test}). \tag{5.10}$$

If the alternative hypothesis is that β_1 is greater than $\beta_{1,0}$, the inequalities in Equations (5.9) and (5.10) are reversed, so the p-value is the right-tail probability, $\Pr(Z > t^{act})$.

When should a one-sided test be used?   In practice, one-sided alternative hypotheses should be used only when there is a clear reason for doing so. This reason could come from economic theory, prior empirical evidence, or both. However, even if it initially seems that the relevant alternative is one-sided, upon reflection this might not necessarily be so. A newly formulated drug undergoing clinical trials actually could prove harmful because of previously unrecognized side effects. In the class size example, we are reminded of the graduation joke that a university's secret of success is to admit talented students and then make sure that the faculty stays out of their way and does as little damage as possible. In practice, such ambiguity often leads econometricians to use two-sided tests.

Application to test scores.   The t-statistic testing the hypothesis that there is no effect of class size on test scores [so $\beta_{1,0} = 0$ in Equation (5.9)] is $t^{act} = -4.38$. This value is less than -2.33 (the critical value for a one-sided test with a 1% significance level),

so the null hypothesis is rejected against the one-sided alternative at the 1% level. In fact, the *p*-value is less than 0.0006%. Based on these data, you can reject the angry taxpayer's assertion that the negative estimate of the slope arose purely because of random sampling variation at the 1% significance level.

Testing Hypotheses About the Intercept β_0

This discussion has focused on testing hypotheses about the slope β_1. Occasionally, however, the hypothesis concerns the intercept β_0. The null hypothesis concerning the intercept and the two-sided alternative are

$$H_0: \beta_0 = \beta_{0,0} \text{ vs. } H_1: \beta_0 \neq \beta_{0,0} \text{ (two-sided alternative)}. \qquad (5.11)$$

The general approach to testing this null hypothesis consists of the three steps in Key Concept 5.2 applied to β_0 (the formula for the standard error of $\hat{\beta}_0$ is given in Appendix 5.1). If the alternative is one-sided, this approach is modified as was discussed in the previous subsection for hypotheses about the slope.

Hypothesis tests are useful if you have a specific null hypothesis in mind (as did our angry taxpayer). Being able to accept or reject this null hypothesis based on the statistical evidence provides a powerful tool for coping with the uncertainty inherent in using a sample to learn about the population. Yet there are many times that no single hypothesis about a regression coefficient is dominant, and instead one would like to know a range of values of the coefficient that are consistent with the data. This calls for constructing a confidence interval.

5.2   Confidence Intervals for a Regression Coefficient

Because any statistical estimate of the slope β_1 necessarily has sampling uncertainty, we cannot determine the true value of β_1 exactly from a sample of data. It is possible, however, to use the OLS estimator and its standard error to construct a confidence interval for the slope β_1 or for the intercept β_0.

Confidence interval for β_1.   Recall from the discussion of confidence intervals in Section 3.3 that a 95% **confidence interval for β_1** has two equivalent definitions. First, it is the set of values that cannot be rejected using a two-sided hypothesis test with a 5% significance level. Second, it is an interval that has a 95% probability of containing the true value of β_1; that is, in 95% of possible samples that might be drawn, the confidence interval will contain the true value of β_1. Because this interval contains the true value in 95% of all samples, it is said to have a **confidence level** of 95%.

The reason these two definitions are equivalent is as follows. A hypothesis test with a 5% significance level will, by definition, reject the true value of β_1 in only 5%

Confidence Interval for β_1

A 95% two-sided confidence interval for β_1 is an interval that contains the true value of β_1 with a 95% probability; that is, it contains the true value of β_1 in 95% of all possible randomly drawn samples. Equivalently, it is the set of values of β_1 that cannot be rejected by a 5% two-sided hypothesis test. When the sample size is large, it is constructed as

$$95\% \text{ confidence interval for } \beta_1 = [\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)]. \quad (5.12)$$

of all possible samples; that is, in 95% of all possible samples, the true value of β_1 will *not* be rejected. Because the 95% confidence interval (as defined in the first definition) is the set of all values of β_1 that are *not* rejected at the 5% significance level, it follows that the true value of β_1 will be contained in the confidence interval in 95% of all possible samples.

As in the case of a confidence interval for the population mean (Section 3.3), in principle a 95% confidence interval can be computed by testing all possible values of β_1 (that is, testing the null hypothesis $\beta_1 = \beta_{1,0}$ for all values of $\beta_{1,0}$) at the 5% significance level using the t-statistic. The 95% confidence interval is then the collection of all the values of β_1 that are not rejected. But constructing the t-statistic for all values of β_1 would take forever.

An easier way to construct the confidence interval is to note that the t-statistic will reject the hypothesized value $\beta_{1,0}$ whenever $\beta_{1,0}$ is outside the range $\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$. This implies that the 95% confidence interval for β_1 is the interval $[\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)]$. This argument parallels the argument used to develop a confidence interval for the population mean.

The construction of a confidence interval for β_1 is summarized as Key Concept 5.3.

Confidence interval for β_0. A 95% confidence interval for β_0 is constructed as in Key Concept 5.3, with $\hat{\beta}_0$ and $SE(\hat{\beta}_0)$ replacing $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$.

Application to test scores. The OLS regression of the test score against the student–teacher ratio, reported in Equation (5.8), yielded $\hat{\beta}_1 = -2.28$ and $SE(\hat{\beta}_1) = 0.52$. The 95% two-sided confidence interval for β_1 is $\{-2.28 \pm 1.96 \times 0.52\}$, or $-3.30 \le \beta_1 \le -1.26$. The value $\beta_1 = 0$ is not contained in this confidence interval, so (as we knew already from Section 5.1) the hypothesis $\beta_1 = 0$ can be rejected at the 5% significance level.

Confidence intervals for predicted effects of changing X. The 95% confidence interval for β_1 can be used to construct a 95% confidence interval for the predicted effect of a general change in X.

Consider changing X by a given amount, Δx. The expected change in Y associated with this change in X is $\beta_1 \Delta x$. The population slope β_1 is unknown, but because we can construct a confidence interval for β_1, we can construct a confidence interval for the expected effect $\beta_1 \Delta x$. Because one end of a 95% confidence interval for β_1 is $\hat{\beta}_1 - 1.96SE(\hat{\beta}_1)$, the predicted effect of the change Δx using this estimate of β_1 is $[\hat{\beta}_1 - 1.96SE(\hat{\beta}_1)] \times \Delta x$. The other end of the confidence interval is $\hat{\beta}_1 + 1.96SE(\hat{\beta}_1)$, and the predicted effect of the change using that estimate is $[\hat{\beta}_1 + 1.96SE(\hat{\beta}_1)] \times \Delta x$. Thus a 95% confidence interval for the effect of changing X by the amount Δx can be expressed as

$$95\% \text{ confidence interval for } \beta_1 \Delta x$$
$$= [(\hat{\beta}_1 - 1.96SE(\hat{\beta}_1))\Delta x, (\hat{\beta}_1 + 1.96SE(\hat{\beta}_1))\Delta x]. \tag{5.13}$$

For example, our hypothetical superintendent is contemplating reducing the student–teacher ratio by 2. Because the 95% confidence interval for β_1 is $[-3.30, -1.26]$, the effect of reducing the student–teacher ratio by 2 could be as great as $-3.30 \times (-2) = 6.60$ or as little as $-1.26 \times (-2) = 2.52$. Thus decreasing the student–teacher ratio by 2 is estimated to increase test scores by between 2.52 and 6.60 points, with a 95% confidence level.

5.3 Regression When X Is a Binary Variable

The discussion so far has focused on the case that the regressor is a continuous variable. Regression analysis can also be used when the regressor is binary—that is, when it takes on only two values, 0 and 1. For example, X might be a worker's sex ($= 1$ if female, $= 0$ if male), whether a school district is urban or rural ($= 1$ if urban, $= 0$ if rural), or whether the district's class size is small or large ($= 1$ if small, $= 0$ if large). A binary variable is also called an **indicator variable** or sometimes a **dummy variable**.

Interpretation of the Regression Coefficients

The mechanics of regression with a binary regressor are the same as if it is continuous. The interpretation of β_1, however, is different, and it turns out that regression with a binary variable is equivalent to performing a difference of means analysis, as described in Section 3.4.

To see this, suppose you have a variable D_i that equals either 0 or 1, depending on whether the student–teacher ratio is less than 20:

$$D_i = \begin{cases} 1 \text{ if the student–teacher ratio in } i^{\text{th}} \text{ district} < 20 \\ 0 \text{ if the student–teacher ratio in } i^{\text{th}} \text{ district} \geq 20 \end{cases}. \tag{5.14}$$

The population regression model with D_i as the regressor is

$$Y_i = \beta_0 + \beta_1 D_i + u_i, i = 1, \ldots, n. \qquad (5.15)$$

This is the same as the regression model with the continuous regressor X_i except that now the regressor is the binary variable D_i. Because D_i is not continuous, it is not useful to think of β_1 as a slope; indeed, because D_i can take on only two values, there is no "line," so it makes no sense to talk about a slope. Thus we will not refer to β_1 as the slope in Equation (5.15); instead, we will simply refer to β_1 as the **coefficient multiplying D_i** in this regression or, more compactly, the **coefficient on D_i.**

If β_1 in Equation (5.15) is not a slope, what is it? The best way to interpret β_0 and β_1 in a regression with a binary regressor is to consider, one at a time, the two possible cases, $D_i = 0$ and $D_i = 1$. If the student–teacher ratio is high, then $D_i = 0$, and Equation (5.15) becomes

$$Y_i = \beta_0 + u_i \quad (D_i = 0). \qquad (5.16)$$

Because $E(u_i|D_i) = 0$, the conditional expectation of Y_i when $D_i = 0$ is $E(Y_i \,|\, D_i = 0) = \beta_0$; that is, β_0 is the population mean value of test scores when the student–teacher ratio is high. Similarly, when $D_i = 1$,

$$Y_i = \beta_0 + \beta_1 + u_i \quad (D_i = 1). \qquad (5.17)$$

Thus, when $D_i = 1, E(Y_i \,|\, D_i = 1) = \beta_0 + \beta_1$; that is, $\beta_0 + \beta_1$ is the population mean value of test scores when the student–teacher ratio is low.

Because $\beta_0 + \beta_1$ is the population mean of Y_i when $D_i = 1$ and β_0 is the population mean of Y_i when $D_i = 0$, the difference $(\beta_0 + \beta_1) - \beta_0 = \beta_1$ is the difference between these two means. In other words, β_1 is the difference between the conditional expectation of Y_i when $D_i = 1$ and when $D_i = 0$, or $\beta_1 = E(Y_i \,|\, D_i = 1) - E(Y_i \,|\, D_i = 0)$. In the test score example, β_1 is the difference between the mean test score in districts with low student–teacher ratios and the mean test score in districts with high student–teacher ratios.

Because β_1 is the difference in the population means, it makes sense that the OLS estimator β_1 is the difference between the sample averages of Y_i in the two groups, and, in fact, this is the case.

Hypothesis tests and confidence intervals. If the two population means are the same, then β_1 in Equation (5.15) is 0. Thus the null hypothesis that the two population means are the same can be tested against the alternative hypothesis that they differ by testing the null hypothesis $\beta_1 = 0$ against the alternative $\beta_1 \neq 0$. This hypothesis can be tested using the procedure outlined in Section 5.1. Specifically, the null hypothesis can be rejected at the 5% level against the two-sided alternative when the OLS *t*-statistic $t = \hat{\beta}_1/SE(\hat{\beta}_1)$ exceeds 1.96 in absolute value. Similarly, a 95% confidence interval for β_1, constructed as $\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$ as described in Section 5.2, provides a 95% confidence interval for the difference between the two population means.

Application to test scores. As an example, a regression of the test score against the student–teacher ratio binary variable D defined in Equation (5.14) estimated by OLS using the 420 observations in Figure 4.2 yields

$$\widehat{TestScore} = 650.0 + 7.4D, R^2 = 0.037, SER = 18.7,$$
$$(1.3) \quad (1.8) \tag{5.18}$$

where the standard errors of the OLS estimates of the coefficients β_0 and β_1 are given in parentheses below the OLS estimates. Thus the average test score for the sub-sample with student–teacher ratios greater than or equal to 20 (that is, for which $D = 0$) is 650.0, and the average test score for the subsample with student–teacher ratios less than 20 (so $D = 1$) is $650.0 + 7.4 = 657.4$. The difference between the sample average test scores for the two groups is 7.4. This is the OLS estimate of β_1, the coefficient on the student–teacher ratio binary variable D.

Is the difference in the population mean test scores in the two groups statistically significantly different from 0 at the 5% level? To find out, construct the t-statistic on β_1: $t = 7.4/1.8 = 4.04$. This value exceeds 1.96 in absolute value, so the hypothesis that the population mean test scores in districts with high and low student–teacher ratios are the same can be rejected at the 5% significance level.

The OLS estimator and its standard error can be used to construct a 95% confidence interval for the true difference in means. This is $7.4 \pm 1.96 \times 1.8 = (3.9, 10.9)$. This confidence interval excludes $\beta_1 = 0$, so that (as we know from the previous paragraph) the hypothesis $\beta_1 = 0$ can be rejected at the 5% significance level.

5.4 Heteroskedasticity and Homoskedasticity

Our only assumption about the distribution of u_i conditional on X_i is that it has a mean of 0 (the first least squares assumption). If, furthermore, the *variance* of this conditional distribution does not depend on X_i, then the errors are said to be homoskedastic. This section discusses homoskedasticity, its theoretical implications, the simplified formulas for the standard errors of the OLS estimators that arise if the errors are homoskedastic, and the risks you run if you use these simplified formulas in practice.

What Are Heteroskedasticity and Homoskedasticity?

Definitions of heteroskedasticity and homoskedasticity. The error term u_i is **homoskedastic** if the variance of the conditional distribution of u_i given X_i is constant for $i = 1, \ldots, n$ and in particular does not depend on X_i. Otherwise, the error term is **heteroskedastic**.

Homoskedasticity and heteroskedasticity are illustrated in Figure 5.2. The distribution of the errors u_i is shown for various values of x. Because this distribution applies specifically for the indicated value of x, this is the conditional distribution of u_i given $X_i = x$; by the first least squares assumption, this distribution has mean 0 for all x. In Figure 5.2(a), all these conditional distributions have the same spread; more

FIGURE 5.2   Homoskedasticity and Heteroskedasticity

The figure plots the conditional distribution of test scores for three different class sizes (x). In figure (a), the spread of these distributions does not depend on x; that is, $\text{var}(u\,|\,X = x)$ does not depend on x, so the errors are homoskedastic. In figure (b), these distributions become more spread out (have a larger variance) for larger class sizes, so $\text{var}(u\,|\,X = x)$ depends on x and the u is heteroskedastic.

(a) The errors are homoscedastic

(b) The errors are heteroskedastic
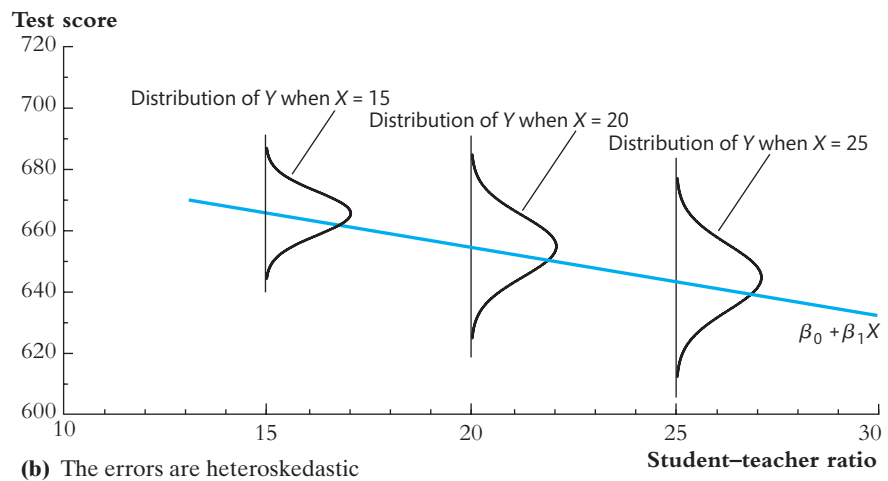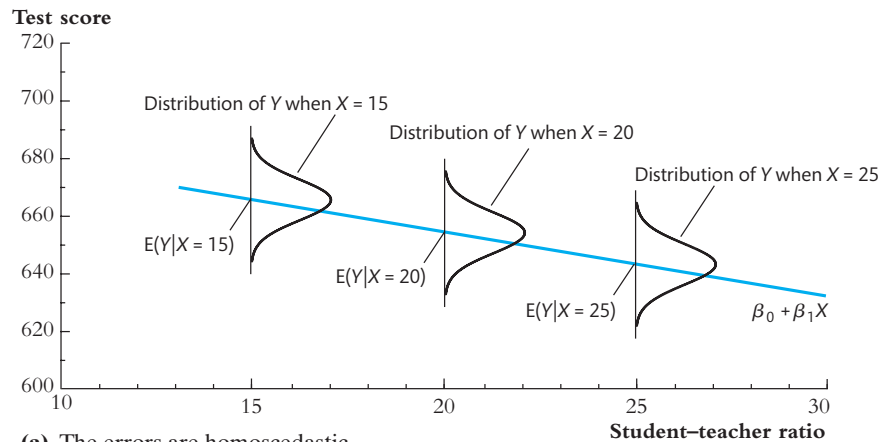
precisely, the variance of these distributions is the same for the various values of x. That is, in Figure 5.2(a), the conditional variance of u_i given $X_i = x$ does not depend on x, so the errors illustrated in Figure 5.2(a) are homoskedastic.

In contrast, Figure 5.2(b) illustrates a case in which the conditional distribution of u_i spreads out as x increases. For small values of x, this distribution is tight, but for larger values of x, it has a greater spread. Thus in Figure 5.2 the variance of u_i given $X_i = x$ increases with x, so that the errors in Figure 5.2 are heteroskedastic.

The definitions of heteroskedasticity and homoskedasticity are summarized in Key Concept 5.4.

Example. These terms are a mouthful, and the definitions might seem abstract. To help clarify them with an example, we digress from the student–teacher ratio/test score problem and instead return to the example of earnings of male versus female college

Heteroskedasticity and Homoskedasticity

The error term u_i is homoskedastic if the variance of the conditional distribution of u_i given X_i, $var(u_i|X_i = x)$, is constant for $i = 1, \ldots, n$ and in particular does not depend on x. Otherwise, the error term is heteroskedastic.

graduates considered in the box in Chapter 3 titled "The Gender Gap in Earnings of College Graduates in the United States." Let $MALE_i$ be a binary variable that equals 1 for male college graduates and equals 0 for female graduates. The binary variable regression model relating a college graduate's earnings to his or her sex is

$$Earnings_i = \beta_0 + \beta_1 MALE_i + u_i \qquad (5.19)$$

for $i = 1, \ldots, n$. Because the regressor is binary, β_1 is the difference in the population means of the two groups—in this case, the difference in mean earnings between men and women who graduated from college.

The definition of homoskedasticity states that the variance of u_i does not depend on the regressor. Here the regressor is $MALE_i$, so at issue is whether the variance of the error term depends on $MALE_i$. In other words, is the variance of the error term the same for men and for women? If so, the error is homoskedastic; if not, it is heteroskedastic.

Deciding whether the variance of u_i depends on $MALE_i$ requires thinking hard about what the error term actually is. In this regard, it is useful to write Equation (5.19) as two separate equations, one for each sex:

$$Earnings_i = \beta_0 + u_i \quad \text{(females) and} \qquad (5.20)$$

$$Earnings_i = \beta_0 + \beta_1 + u_i \quad \text{(males)}. \qquad (5.21)$$

Thus, for women, u_i is the deviation of the ith woman's earnings from the population mean earnings for women (β_0), and for men, u_i is the deviation of the ith man's earnings from the population mean earnings for men ($\beta_0 + \beta_1$). It follows that the statement "the variance of u_i does not depend on $MALE$" is equivalent to the statement "the variance of earnings is the same for males as it is for females." In other words, in this example, the error term is homoskedastic if the variance of the population distribution of earnings is the same for both sexes; if these variances differ, the error term is heteroskedastic.

Mathematical Implications of Homoskedasticity

The OLS estimators remain unbiased and asymptotically normal. Because the least squares assumptions in Key Concept 4.3 place no restrictions on the conditional variance, they apply to both the general case of heteroskedasticity and the special case of homoskedasticity. Therefore, the OLS estimators remain unbiased and consistent even if the errors are homoskedastic. In addition, the OLS estimators have sampling distributions that are normal in large samples even if the errors are homoskedastic. Whether the errors are homoskedastic or heteroskedastic, the OLS estimator is unbiased, consistent, and asymptotically normal.

Efficiency of the OLS estimator when the errors are homoskedastic. If the least squares assumptions in Key Concept 4.3 hold and the errors are homoskedastic, then the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are efficient among all estimators that are linear in $Y_1, \ldots, Y_n$ and are unbiased, conditional on $X_1, \ldots, X_n$. This result, which is called the Gauss–Markov theorem, is discussed in Section 5.5.

Homoskedasticity-only variance formula. If the error term is homoskedastic, then the formulas for the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ in Key Concept 4.4 simplify. Consequently, if the errors are homoskedastic, then there is a specialized formula that can be used for the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. The **homoskedasticity-only standard error** of $\hat{\beta}_1$, derived in Appendix 5.1, is $SE(\hat{\beta}_1) = \sqrt{\tilde{\sigma}^2_{\hat{\beta}_1}}$, where $\tilde{\sigma}^2_{\hat{\beta}_1}$ is the homoskedasticity-only estimator of the variance of $\hat{\beta}_1$:

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \frac{s^2_{\hat{u}}}{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad (\text{homoskedasticity-only}), \tag{5.22}$$

where $s^2_{\hat{u}}$ is given in Equation (4.17). The homoskedasticity-only formula for the standard error of $\hat{\beta}_0$ is given in Appendix 5.1. In the special case that X is a binary variable, the estimator of the variance of $\hat{\beta}_1$ under homoskedasticity (that is, the square of the standard error of $\hat{\beta}_1$ under homoskedasticity) is the so-called pooled variance formula for the difference in means given in Equation (3.23).

Because these alternative formulas are derived for the special case that the errors are homoskedastic and do not apply if the errors are heteroskedastic, they will be referred to as the "homoskedasticity-only" formulas for the variance and standard error of the OLS estimators. As the name suggests, if the errors are heteroskedastic, then the homoskedasticity-only standard errors are inappropriate. Specifically, if the errors are heteroskedastic, then the t-statistic computed using the homoskedasticity-only standard error does not have a standard normal distribution, even in large samples. In fact, the correct critical values to use for this homoskedasticity-only t-statistic depend on the precise nature of the heteroskedasticity, so those critical values cannot be tabulated. Similarly, if the errors are heteroskedastic but a confidence interval is constructed as ± 1.96 homoskedasticity-only standard errors, in general the probability that this interval contains the true value of the coefficient is not 95%, even in large samples.

In contrast, because homoskedasticity is a special case of heteroskedasticity, the estimators $\hat{\sigma}^2_{\hat{\beta}_1}$ and $\hat{\sigma}^2_{\hat{\beta}_0}$ of the variances of $\hat{\beta}_1$ and $\hat{\beta}_0$ given in Equations (5.4) and (5.26) produce valid statistical inferences whether the errors are heteroskedastic or homoskedastic. Thus hypothesis tests and confidence intervals based on those standard errors are valid whether or not the errors are heteroskedastic. Because the standard errors we have used so far [that is, those based on Equations (5.4) and (5.26)] lead to statistical inferences that are valid whether or not the errors are heteroskedastic, they are called **heteroskedasticity-robust standard errors**. Because such formulas were proposed by Eicker (1967), Huber (1967), and White (1980), they are also referred to as Eicker–Huber–White standard errors.

What Does This Mean in Practice?

Which is more realistic, heteroskedasticity or homoskedasticity? The answer to this question depends on the application. However, the issues can be clarified by returning to the example of the gender gap in earnings among college graduates. Familiarity with how people are paid in the world around us gives some clues as to which assumption is more sensible. For many years—and, to a lesser extent, even today—women were not found in the top-paying jobs: There have always been poorly paid men, but there have rarely been highly paid women. This suggests that the distribution of earnings is tighter for women than for men (see the box in Chapter 3 "The Gender Gap in Earnings of College Graduates in the United States"). In other words, the variance of the error term in Equation (5.20) for women is plausibly less than the variance of the error term in Equation (5.21) for men. Thus the still-thin presence of women in top-paying jobs suggests that the error term in the binary variable regression model in Equation (5.19) is heteroskedastic. Unless there are compelling reasons to the contrary—and we can think of none—it makes sense to treat the error term in this example as heteroskedastic.

As the example of earnings illustrates, heteroskedasticity arises in many econometric applications. At a general level, economic theory rarely gives any reason to believe that the errors are homoskedastic. It therefore is prudent to assume that the errors might be heteroskedastic unless you have compelling reasons to believe otherwise.

Practical implications. The main issue of practical relevance in this discussion is whether one should use heteroskedasticity-robust or homoskedasticity-only standard errors. In this regard, it is useful to imagine computing both, then choosing between them. If the homoskedasticity-only and heteroskedasticity-robust standard errors are the same, nothing is lost by using the heteroskedasticity-robust standard errors; if they differ, however, then you should use the more reliable ones that allow for heteroskedasticity. The simplest thing, then, is always to use the heteroskedasticity-robust standard errors.

For historical reasons, many software programs report homoskedasticity-only standard errors as their default setting, so it is up to the user to specify the option of heteroskedasticity-robust standard errors. The details of how to implement heteroskedasticity-robust standard errors depend on the software package you use.

All of the empirical examples in this book employ heteroskedasticity-robust standard errors unless explicitly stated otherwise.[1]

---

[1] In case this book is used in conjunction with other texts, it might be helpful to note that some textbooks add homoskedasticity to the list of least squares assumptions. As just discussed, however, this additional assumption is not needed for the validity of OLS regression analysis as long as heteroskedasticity-robust standard errors are used.

The Economic Value of a Year of Education: Homoskedasticity or Heteroskedasticity?

On average, workers with more education have higher earnings than workers with less education. But if the best-paying jobs mainly go to the college educated, it might also be that the *spread* of the distribution of earnings is greater for workers with more education. Does the distribution of earnings spread out as education increases?

This is an empirical question, so answering it requires analyzing data. Figure 5.3 is a scatterplot of the hourly earnings and the number of years of education for a sample of 2731 full-time workers in the United States in 2015, ages 29 and 30, with between 8 and 18 years of education. The data come from the March 2016 Current Population Survey, which is described in Appendix 3.1.

Figure 5.3 has two striking features. The first is that the mean of the distribution of earnings increases with the number of years of education. This increase is summarized by the OLS regression line,

$$\widehat{Earnings} = -12.12 + 2.37\ Years\ Education,$$
$$(1.36)\quad (0.10)$$
$$R^2 = 0.185,\ SER = 11.24. \tag{5.23}$$

This line is plotted in Figure 5.3. The coefficient of 2.37 in the OLS regression line means that, on average, hourly earnings increase by $2.37 for each additional year of education. The 95% confidence interval for this coefficient is $2.37 \pm 1.96 \times 0.10$, or $2.17 to $2.57.

The second striking feature of Figure 5.3 is that the spread of the distribution of earnings increases with the years of education. While some workers with many years of education have low-paying jobs, very few workers with low levels of education have high-paying jobs. This can be quantified by looking at the spread of the residuals around the OLS regression line. For workers with ten years of education, the standard deviation of the residuals is $6.31; for workers with a high school diploma, this standard deviation is $8.54; and for workers with a college degree, this standard deviation increases to $13.55. Because these standard deviations differ for different levels of education, the variance of the residuals in the regression of Equation (5.23) depends on the value of the regressor (the years of education); in other words, the regression errors are heteroskedastic. In real-world terms, not all college graduates will be earning $75 per hour by the time they are 29, but some will, and workers with only ten years of education have no shot at those jobs.

FIGURE 5.3   **Scatterplot of Hourly Earnings and Years of Education for 29- to 30-Year-Olds in the United States in 2015**

Hourly earnings are plotted against years of education for 2731 full-time 29- to 30-year-old workers. The spread around the regression line increases with the years of education, indicating that the regression errors are heteroskedastic.
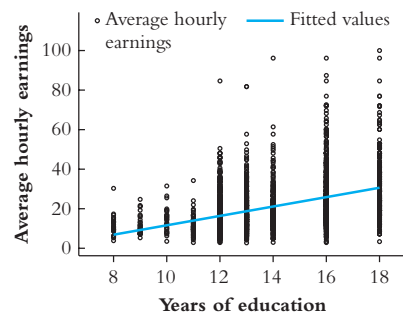
*5.5 The Theoretical Foundations of Ordinary Least Squares

As discussed in Section 4.5, the OLS estimator is unbiased, is consistent, has a variance that is inversely proportional to n, and has a normal sampling distribution when the sample size is large. In addition, under certain conditions the OLS estimator is more efficient than some other candidate estimators. Specifically, if the least squares assumptions hold and if the errors are homoskedastic, then the OLS estimator has the smallest variance of all conditionally unbiased estimators that are linear functions of $Y_1, \ldots, Y_n$. This section explains and discusses this result, which is a consequence of the Gauss–Markov theorem. The section concludes with a discussion of alternative estimators that are more efficient than OLS when the conditions of the Gauss–Markov theorem do not hold.

Linear Conditionally Unbiased Estimators and the Gauss–Markov Theorem

If the three least squares assumptions in Key Concept 4.3 hold and if the error is homoskedastic, then the OLS estimator has the smallest variance, conditional on $X_1, \ldots, X_n$, among all estimators in the class of linear conditionally unbiased estimators. In other words, the OLS estimator is the **B**est **L**inear conditionally **U**nbiased **E**stimator—that is, it is BLUE. This result is an extension of the result, summarized in Key Concept 3.3, that the sample average $\overline{Y}$ is the most efficient estimator of the population mean in the class of all estimators that are unbiased and are linear functions (weighted averages) of $Y_1, \ldots, Y_n$.

Linear conditionally unbiased estimators.   The class of linear conditionally unbiased estimators consists of all estimators of β_1 that are linear functions of $Y_1, \ldots, Y_n$ and that are unbiased, conditional on $X_1, \ldots, X_n$. That is, if $\widetilde{\beta}_1$ is a linear estimator, then it can be written as

$$\widetilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i \quad (\widetilde{\beta}_1 \text{ is linear}), \tag{5.24}$$

where the weights $a_1, \ldots, a_n$ can depend on $X_1, \ldots, X_n$ but *not* on $Y_1, \ldots, Y_n$. The estimator $\widetilde{\beta}_1$ is conditionally unbiased if the mean of its conditional sampling distribution given $X_1, \ldots, X_n$ is β_1. That is, the estimator $\widetilde{\beta}_1$ is conditionally unbiased if

$$E(\widetilde{\beta}_1 | X_1, \ldots, X_n) = \beta_1 \ (\widetilde{\beta}_1 \text{ is conditionally unbiased}). \tag{5.25}$$

The estimator $\widetilde{\beta}_1$ is a linear conditionally unbiased estimator if it can be written in the form of Equation (5.24) (it is linear) and if Equation (5.25) holds (it is conditionally unbiased). It is shown in Appendix 5.2 that the OLS estimator is linear and conditionally unbiased.

---

* This section is optional and is not used in later chapters.

The Gauss–Markov Theorem for $\hat{\beta}_1$

If the three least squares assumptions in Key Concept 4.3 hold *and* if errors are homoskedastic, then the OLS estimator $\hat{\beta}_1$ is the **B**est (most efficient) **L**inear conditionally **U**nbiased **E**stimator (**BLUE**).

The Gauss–Markov theorem. The **Gauss–Markov theorem** states that, under a set of conditions known as the Gauss–Markov conditions, the OLS estimator $\hat{\beta}_1$ has the smallest conditional variance given $X_1, \ldots, X_n$ of all linear conditionally unbiased estimators of β_1; that is, the OLS estimator is BLUE. The Gauss–Markov conditions, which are stated in Appendix 5.2, are implied by the three least squares assumptions plus the assumption that the errors are homoskedastic. Consequently, if the three least squares assumptions hold and the errors are homoskedastic, then OLS is BLUE. The Gauss–Markov theorem is stated in Key Concept 5.5 and proven in Appendix 5.2.

Limitations of the Gauss–Markov theorem. The Gauss–Markov theorem provides a theoretical justification for using OLS. However, the theorem has two important limitations. First, its conditions might not hold in practice. In particular, if the error term is heteroskedastic—as it often is in economic applications—then the OLS estimator is no longer BLUE. As discussed in Section 5.4, the presence of heteroskedasticity does not pose a threat to inference based on heteroskedasticity-robust standard errors, but it does mean that OLS is no longer the efficient linear conditionally unbiased estimator. An alternative to OLS when there is heteroskedasticity of a known form, called the weighted least squares estimator, is discussed below.

The second limitation of the Gauss–Markov theorem is that even if the conditions of the theorem hold, there are other candidate estimators that are not linear and conditionally unbiased; under some conditions, these other estimators are more efficient than OLS.

Regression Estimators Other Than OLS

Under certain conditions, some regression estimators are more efficient than OLS.

The weighted least squares estimator. If the errors are heteroskedastic, then OLS is no longer BLUE. If the nature of the heteroskedasticity is known—specifically, if the conditional variance of u_i given X_i is known up to a constant factor of proportionality—then it is possible to construct an estimator that has a smaller variance than the OLS estimator. This method, called **weighted least squares (WLS)**, weights the ith observation by the inverse of the square root of the conditional variance of u_i given X_i. Because of this weighting, the errors in this weighted regression are homoskedastic, so OLS, when applied to the weighted data, is BLUE. Although theoretically elegant, the practical problem with weighted least squares is that you must know how

the conditional variance of u_i depends on X_i, something that is rarely known in econometric applications. Weighted least squares is therefore used far less frequently than OLS, and further discussion of WLS is deferred to Chapter 18.

The least absolute deviations estimator. As discussed in Section 4.3, the OLS estimator can be sensitive to outliers. If extreme outliers are not rare, then other estimators can be more efficient than OLS and can produce inferences that are more reliable. One such estimator is the least absolute deviations (LAD) estimator, in which the regression coefficients β_0 and β_1 are obtained by solving a minimization problem like that in Equation (4.4) except that the absolute value of the prediction "mistake" is used instead of its square. That is, the LAD estimators of β_0 and β_1 are the values of b_0 and b_1 that minimize $\sum_{i=1}^{n}|Y_i - b_0 - b_1 X_i|$. The LAD estimator is less sensitive to large outliers in u than is OLS.

In many economic data sets, severe outliers in u are rare, so use of the LAD estimator, or other estimators with reduced sensitivity to outliers, is uncommon in applications. Thus the treatment of linear regression throughout the remainder of this text focuses exclusively on least squares methods.

*5.6 Using the *t*-Statistic in Regression When the Sample Size Is Small

When the sample size is small, the exact distribution of the *t*-statistic is complicated and depends on the unknown population distribution of the data. If, however, the three least squares assumptions hold, the regression errors are homoskedastic, *and* the regression errors are normally distributed, then the OLS estimator is normally distributed and the homoskedasticity-only *t*-statistic has a Student *t* distribution. These five assumptions—the three least squares assumptions, that the errors are homoskedastic, and that the errors are normally distributed—are collectively called the **homoskedastic normal regression assumptions**.

The *t*-Statistic and the Student *t* Distribution

Recall from Section 2.4 that the Student *t* distribution with m degrees of freedom is defined to be the distribution of $Z/\sqrt{W/m}$, where Z is a random variable with a standard normal distribution, W is a random variable with a chi-squared distribution with m degrees of freedom, and Z and W are independent. Under the null hypothesis, the *t*-statistic computed using the homoskedasticity-only standard error can be written in this form.

The details of the calculation are presented in Sections 18.4 and 19.4; the main ideas are as follows. The homoskedasticity-only *t*-statistic testing $\beta_1 = \beta_{1,0}$ is $\widetilde{t} = (\hat{\beta}_1 - \beta_{1,0})/\widetilde{\sigma}_{\hat{\beta}_1}$, where $\widetilde{\sigma}_{\hat{\beta}_1}^2$ is defined in Equation (5.22). Under the homoskedastic

---

* This section is optional and is not used in later chapters.

normal regression assumptions, Y_i has a normal distribution, conditional on $X_1, \ldots, X_n$. As discussed in Section 5.5, the OLS estimator is a weighted average of $Y_1, \ldots, Y_n$, where the weights depend on $X_1, \ldots, X_n$ [see Equation (5.32) in Appendix 5.2]. Because a weighted average of independent normal random variables is normally distributed, $\hat{\beta}_1$ has a normal distribution, conditional on $X_1, \ldots, X_n$. Thus $\hat{\beta}_1 - \beta_{1,0}$ has a normal distribution with mean 0 under the null hypothesis, conditional on $X_1, \ldots, X_n$. In addition, Sections 18.4 and 19.4 show that the (normalized) homoskedasticity-only variance estimator has a chi-squared distribution with $n - 2$ degrees of freedom, divided by $n - 2$, and $\tilde{\sigma}^2_{\hat{\beta}_1}$ and $\hat{\beta}_1$ are independently distributed. Consequently, the homoskedasticity-only t-statistic has a Student t distribution with $n - 2$ degrees of freedom.

This result is closely related to a result discussed in Section 3.5 in the context of testing for the equality of the means in two samples. In that problem, if the two population distributions are normal with the same variance and if the t-statistic is constructed using the pooled standard error formula [Equation (3.23)], then the (pooled) t-statistic has a Student t distribution. When X is binary, the homoskedasticity-only standard error for $\hat{\beta}_1$ simplifies to the pooled standard error formula for the difference of means. It follows that the result of Section 3.5 is a special case of the result that if the homoskedastic normal regression assumptions hold, then the homoskedasticity-only regression t-statistic has a Student t distribution (see Exercise 5.10).

Use of the Student t Distribution in Practice

If the regression errors are homoskedastic and normally distributed and if the homoskedasticity-only t-statistic is used, then critical values should be taken from the Student t distribution (Appendix Table 2) instead of the standard normal distribution. Because the difference between the Student t distribution and the normal distribution is negligible if n is moderate or large, this distinction is relevant only if the sample size is small.

In econometric applications, there is rarely a reason to believe that the errors are homoskedastic and normally distributed. Because sample sizes typically are large, however, inference can proceed as described in Sections 5.1 and 5.2—that is, by first computing heteroskedasticity-robust standard errors and then by using the standard normal distribution to compute p-values, hypothesis tests, and confidence intervals.

5.7   Conclusion

Return for a moment to the problem of the superintendent who is considering hiring additional teachers to cut the student–teacher ratio. What have we learned that she might find useful?

Our regression analysis, based on the 420 observations in the California test score data set, showed that there was a negative relationship between the student–teacher ratio and test scores: Districts with smaller classes have higher test scores.

The coefficient is moderately large, in a practical sense: Districts with two fewer students per teacher have, on average, test scores that are 4.6 points higher. This corresponds to moving a district at the 50th percentile of the distribution of test scores to approximately the 60th percentile.

The coefficient on the student–teacher ratio is statistically significantly different from 0 at the 5% significance level. The population coefficient might be 0, and we might simply have estimated our negative coefficient by random sampling variation. However, the probability of doing so (and of obtaining a t-statistic on β_1 as large as we did) purely by random variation over potential samples is exceedingly small, approximately 0.001%. A 95% confidence interval for β_1 is $-3.30 \leq \beta_1 \leq -1.26$.

These results represent progress toward answering the superintendent's question—yet a nagging concern remains. There is a negative relationship between the student–teacher ratio and test scores, but is this relationship the *causal* one that the superintendent needs to make her decision? Districts with lower student–teacher ratios have, on average, higher test scores. But does this mean that reducing the student–teacher ratio will, in fact, increase scores?

The question of whether OLS applied to the California data estimates the causal effect of class size on test scores can be sharpened by returning to the least squares assumptions of Key Concept 4.3. The first least squares assumption requires that, when β_1 is defined to be the causal effect, the distribution of the errors has conditional mean 0. This requirement has the interpretation of, in effect, requiring X (class size) to be randomly assigned or as-if randomly assigned. Because the California data are observational, class size was not randomly assigned. So the question is: In the California data, is class size as-if randomly assigned, in the sense that $E(u|X) = 0$?

There is, in fact, reason to worry that it might not be. Hiring more teachers, after all, costs money, so wealthier school districts can better afford smaller classes. But students at wealthier schools also have other advantages over their poorer neighbors, including better facilities, newer books, and better-paid teachers. Moreover, students at wealthier schools tend themselves to come from more affluent families and thus have other advantages not directly associated with their school. For example, California has a large immigrant community; these immigrants tend to be poorer than the overall population, and in many cases, their children are not native English speakers. It thus might be that our negative estimated relationship between test scores and the student–teacher ratio is a consequence of large classes being found in conjunction with many other factors that are, in fact, the real reason for the lower test scores.

These other factors, or "omitted variables," could mean that the OLS analysis done so far has little value to the superintendent. Indeed, it could be misleading: Changing the student–teacher ratio alone would not change these other factors that determine a child's performance at school. To address this problem, we need a method that will allow us to isolate the effect on test scores of changing the student–teacher ratio, *holding these other factors constant*. That method is multiple regression analysis, the topic of Chapters 6 and 7.

Summary

1. Hypothesis testing for regression coefficients is analogous to hypothesis testing for the population mean: Use the *t*-statistic to calculate the *p*-values and either accept or reject the null hypothesis. Like a confidence interval for the population mean, a 95% confidence interval for a regression coefficient is computed as the estimator ± 1.96 standard errors.

2. When X is binary, the regression model can be used to estimate and test hypotheses about the difference between the population means of the "$X = 0$" group and the "$X = 1$" group.

3. In general, the error u_i is heteroskedastic; that is, the variance of u_i at a given value of X_i, $\text{var}(u_i \mid X_i = x)$, depends on x. A special case is when the error is homoskedastic; that is, when $\text{var}(u_i \mid X_i = x)$ is constant. Homoskedasticity-only standard errors do not produce valid statistical inferences when the errors are heteroskedastic, but heteroskedasticity-robust standard errors do.

4. If the three least squares assumption hold *and* if the regression errors are homoskedastic, then, as a result of the Gauss–Markov theorem, the OLS estimator is BLUE.

5. If the three least squares assumptions hold, if the regression errors are homoskedastic, *and* if the regression errors are normally distributed, then the OLS *t*-statistic computed using homoskedasticity-only standard errors has a Student *t* distribution when the null hypothesis is true. The difference between the Student *t* distribution and the normal distribution is negligible if the sample size is moderate or large.

Key Terms

null hypothesis (138)

two-sided alternative hypothesis (138)

standard error of $\hat{\beta}_1$ (138)

t-statistic (138)

p-value (138)

confidence interval for β_1 (142)

confidence level (142)

indicator variable (144)

dummy variable (144)

coefficient multiplying D_i (145)

coefficient on D_i (145)

homoskedasticity and
    heteroskedasticity (146)

homoskedasticity-only standard
    error (149)

heteroskedasticity-robust standard
    error (149)

Gauss–Markov theorem (164)

best linear unbiased estimator
    (BLUE) (153)

weighted least squares (WLS) (153)

homoskedastic normal regression
    assumptions (154)

Gauss–Markov conditions (166)

Review the Concepts

5.1 Outline the procedures for computing the p-value of a two-sided test of $H_0: \mu_Y = 0$ using an i.i.d. set of observations $Y_i, i = 1, \ldots, n$. Outline the procedures for computing the p-value of a two-sided test of $H_0: \beta_1 = 0$ in a regression model using an i.i.d. set of observations $(Y_i, X_i), i = 1, \ldots, n$.

5.2 Explain how you could use a regression model to estimate the wage gender gap using the data on earnings of men and women. What are the dependent and independent variables?

5.3 Define *homoskedasticity* and *heteroskedasticity*. Provide a hypothetical empirical example in which you think the errors would be heteroskedastic, and explain your reasoning.

5.4 Consider the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, where Y_i denotes a worker's average hourly earnings (measured in dollars) and X_i is a binary (or indicator) variable that is equal to 1 if the worker has a college degree and is equal to 0 otherwise. Suppose $\beta_1 = 8.1$. Explain what this value means. Include the units of β_1 in your answer.

Exercises

5.1 Suppose a researcher, using data on class size (CS) and average test scores from 100 third-grade classes, estimates the OLS regression

$$\widehat{TestScore} = 520.4 - 5.82 \times CS, R^2 = 0.08, SER = 11.5.$$
$$\quad\quad (20.4) \quad (2.21)$$

a. Construct a 95% confidence interval for β_1, the regression slope coefficient.

b. Calculate the p-value for the two-sided test of the null hypothesis $H_0: \beta_1 = 0$. Do you reject the null hypothesis at the 5% level? At the 1% level?

   c. Calculate the p-value for the two-sided test of the null hypothesis $H_0: \beta_1 = -5.6$. Without doing any additional calculations, determine whether -5.6 is contained in the 95% confidence interval for β_1.

   d. Construct a 99% confidence interval for β_0.

5.2 Suppose a researcher, using wage data on 250 randomly selected male workers and 280 female workers, estimates the OLS regression

$$\widehat{Wage} = 12.52 + 2.12 \times Male, R^2 = 0.06, SER = 4.2,$$
$$\qquad\quad (0.23) \quad (0.36)$$

where *Wage* is measured in dollars per hour and *Male* is a binary variable that is equal to 1 if the person is a male and 0 if the person is a female. Define the wage gender gap as the difference in mean earnings between men and women.

   a. What is the estimated gender gap?

   b. Is the estimated gender gap significantly different from 0? (Compute the p-value for testing the null hypothesis that there is no gender gap.)

   c. Construct a 95% confidence interval for the gender gap.

   d. In the sample, what is the mean wage of women? Of men?

   e. Another researcher uses these same data but regresses *Wages* on *Female*, a variable that is equal to 1 if the person is female and 0 if the person a male. What are the regression estimates calculated from this regression?

$$\widehat{Wage} = \underline{\quad} + \underline{\quad} \times Female, R^2 = \underline{\quad}, SER = \underline{\quad}.$$

5.3 Suppose a random sample of 200 20-year-old men is selected from a population and their heights and weights are recorded. A regression of weight on height yields

$$\widehat{Weight} = -99.41 + 3.94 \times Height, R^2 = 0.81, SER = 10.2,$$
$$\qquad\qquad (2.15) \quad (0.31)$$

where *Weight* is measured in pounds and *Height* is measured in inches. Two of your classmates differ in height by 1.5 inches. Construct a 99% confidence interval for the difference in their weights.

5.4 Read the box "The Economic Value of a Year of Education: Homoskedasticity or Heteroskedasticity?" in Section 5.4. Use the regression reported in Equation (5.23) to answer the following.

   a. A randomly selected 30-year-old worker reports an education level of 16 years. What is the worker's expected average hourly earnings?

b. A high school graduate (12 years of education) is contemplating going to a community college for a 2-year degree. How much are this worker's average hourly earnings expected to increase?

c. A high school counselor tells a student that, on average, college graduates earn $10 per hour more than high school graduates. Is this statement consistent with the regression evidence? What range of values is consistent with the regression evidence?

5.5 In the 1980s, Tennessee conducted an experiment in which kindergarten students were randomly assigned to "regular" and "small" classes and given standardized tests at the end of the year. (Regular classes contained approximately 24 students, and small classes contained approximately 15 students.) Suppose, in the population, the standardized tests have a mean score of 925 points and a standard deviation of 75 points. Let *SmallClass* denote a binary variable equal to 1 if the student is assigned to a small class and equal to 0 otherwise. A regression of *TestScore* on *SmallClass* yields

$$TestScore = 918.0 + 13.9 \times SmallClass, R^2 = 0.01, SER = 74.6.$$
$$\quad\;\;(1.6)\quad(2.5)$$

a. Do small classes improve test scores? By how much? Is the effect large? Explain.

b. Is the estimated effect of class size on test scores statistically significant? Carry out a test at the 5% level.

c. Construct a 99% confidence interval for the effect of *SmallClass* on *TestScore*.

d. Does least squares assumption 1 plausibly hold for this regression? Explain.

5.6 Refer to the regression described in Exercise 5.5.

a. Do you think that the regression errors are plausibly homoskedastic? Explain.

b. $SE(\hat{\beta}_1)$ was computed using Equation (5.3). Suppose the regression errors were homoskedastic. Would this affect the validity of the confidence interval constructed in Exercise 5.5(c)? Explain.

5.7 Suppose (Y_i, X_i) satisfy the least squares assumptions in Key Concept 4.3. A random sample of size $n = 250$ is drawn and yields

$$\hat{Y} = 5.4 + 3.2X, R^2 = 0.26, SER = 6.2.$$
$$\quad\;(3.1)\quad(1.5)$$

a. Test $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$ at the 5% level.

b. Construct a 95% confidence interval for β_1.

c. Suppose you learned that Y_i and X_i were independent. Would you be surprised? Explain.

d. Suppose Y_i and X_i are independent and many samples of size $n = 250$ are drawn, regressions estimated, and (a) and (b) answered. In what fraction of the samples would H_0 from (a) be rejected? In what fraction of samples would the value $\beta_1 = 0$ be included in the confidence interval from (b)?

5.8 Suppose (Y_i, X_i) satisfy the least squares assumptions in Key Concept 4.3 and, in addition, u_i is $N(0, \sigma_u^2)$ and is independent of X_i. A sample of size $n = 30$ yields

$$\hat{Y} = 43.2 + 61.5X, R^2 = 0.54, SER = 1.52,$$
$$\quad (10.2) \quad (7.4)$$

where the numbers in parentheses are the homoskedastic-only standard errors for the regression coefficients.

a. Construct a 95% confidence interval for β_0.

b. Test $H_0: \beta_1 = 55$ vs. $H_1: \beta_1 \neq 55$ at the 5% level.

c. Test $H_0: \beta_1 = 55$ vs. $H_1: \beta_1 > 55$ at the 5% level.

5.9 Consider the regression model

$$Y_i = \beta X_i + u_i,$$

where u_i and X_i satisfy the least squares assumptions in Key Concept 4.3. Let $\bar{\beta}$ denote an estimator of β that is constructed as $\bar{\beta} = \bar{Y}/\bar{X}$, where $\bar{Y}$ and $\bar{X}$ are the sample means of Y_i and X_i, respectively.

a. Show that $\bar{\beta}$ is a linear function of $Y_1, Y_2, \ldots, Y_n$.

b. Show that $\bar{\beta}$ is conditionally unbiased.

5.10 Let X_i denote a binary variable, and consider the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$. Let $\bar{Y}_0$ denote the sample mean for observations with $X = 0$, and let $\bar{Y}_1$ denote the sample mean for observations with $X = 1$. Show that $\hat{\beta}_0 = \bar{Y}_0, \hat{\beta}_0 + \hat{\beta}_1 = \bar{Y}_1$, and $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$.

5.11 A random sample of workers contains $n_m = 120$ men and $n_w = 131$ women. The sample average of men's weekly earnings $[\bar{Y}_m = (1/n_m)\sum_{i=1}^{n_m} Y_{m,i}]$ is \$523.10, and the sample standard deviation $[s_m = \sqrt{\frac{1}{n_m - 1}\sum_{i=1}^{n_m}(Y_{m,i} - \bar{Y}_m)^2}]$ is \$68.10. The corresponding values for women are $\bar{Y}_w = \$485.10$ and $s_w = \$51.10$. Let *Women* denote an indicator variable that is equal to 1 for women and 0 for men, and suppose that all 251 observations are used in the regression $Y_i = \beta_0 + \beta_1 Women_i + u_i$. Find the OLS estimates of β_0 and β_1 and their corresponding standard errors.

5.12 Starting from Equation (4.20), derive the variance of $\hat{\beta}_0$ under homoskedasticity given in Equation (5.28) in Appendix 5.1.

5.13 Suppose (Y_i, X_i) satisfy the least squares assumptions in Key Concept 4.3 and, in addition, u_i is distributed $N(0, \sigma_u^2)$ and is independent of X_i.

a. Is $\hat{\beta}_1$ conditionally unbiased?

b. Is $\hat{\beta}_1$ the best linear conditionally unbiased estimator of β_1?

c. How would your answers to (a) and (b) change if you assumed only that (Y_i, X_i) satisfied the least squares assumptions in Key Concept 4.3 and $\text{var}(u_i | X_i = x)$ is constant?

d. How would your answers to (a) and (b) change if you assumed only that (Y_i, X_i) satisfied the least squares assumptions in Key Concept 4.3?

5.14 Suppose $Y_i = \beta X_i + u_i$, where (u_i, X_i) satisfy the Gauss–Markov conditions given in Equation (5.31).

a. Derive the least squares estimator of β, and show that it is a linear function of $Y_1, \ldots, Y_n$.

b. Show that the estimator is conditionally unbiased.

c. Derive the conditional variance of the estimator.

d. Prove that the estimator is BLUE.

5.15 A researcher has two independent samples of observations on (Y_i, X_i). To be specific, suppose Y_i denotes earnings, X_i denotes years of schooling, and the independent samples are for men and women. Write the regression for men as $Y_{m,i} = \beta_{m,0} + \beta_{m,1} X_{m,i} + u_{m,i}$ and the regression for women as $Y_{w,i} = \beta_{w,0} + \beta_{w,1} X_{w,i} + u_{w,i}$. Let $\hat{\beta}_{m,1}$ denote the OLS estimator constructed using the sample of men, $\hat{\beta}_{w,1}$ denote the OLS estimator constructed from the sample of women, and $SE(\hat{\beta}_{m,1})$ and $SE(\hat{\beta}_{w,1})$ denote the corresponding standard errors. Show that the standard error of $\hat{\beta}_{m,1} - \hat{\beta}_{w,1}$ is given by $SE(\hat{\beta}_{m,1} - \hat{\beta}_{w,1}) = \sqrt{[SE(\hat{\beta}_{m,1})]^2 + [SE(\hat{\beta}_{w,1})]^2}$.

Empirical Exercises

(Only three empirical exercises for this chapter are given in the text, but you can find more on the text website, http://www.pearsonhighered.com/stock_watson/.)

E5.1 Use the data set **Earnings_and_Height** described in Empirical Exercise 4.2 to carry out the following exercises.

a. Run a regression of *Earnings* on *Height*.

    i. Is the estimated slope statistically significant?

    ii. Construct a 95% confidence interval for the slope coefficient.

b. Repeat (a) for women.

c. Repeat (a) for men.

d. Test the null hypothesis that the effect of height on earnings is the same for men and women. (*Hint:* See Exercise 5.15.)

    e. One explanation for the effect of height on earnings is that some professions require strength, which is correlated with height. Does the effect of height on earnings disappear when the sample is restricted to occupations in which strength is unlikely to be important?

E5.2 Using the data set **Growth** described in Empirical Exercise 4.1, but excluding the data for Malta, run a regression of *Growth* on *TradeShare*.

    a. Is the estimated regression slope statistically significant? That is, can you reject the null hypothesis H_0: $\beta_1 = 0$ vs. a two-sided alternative hypothesis at the 10%, 5%, or 1% significance level?

    b. What is the *p*-value associated with the coefficient's *t*-statistic?

    c. Construct a 90% confidence interval for β_1.

E5.3 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find the data file **Birthweight_Smoking**, which contains data for a random sample of babies born in Pennsylvania in 1989. The data include the baby's birth weight together with various characteristics of the mother, including whether she smoked during the pregnancy.[2] A detailed description is given in **Birthweight_Smoking_Description**, also available on the website. In this exercise, you will investigate the relationship between birth weight and smoking during pregnancy.

    a. In the sample:

        i. What is the average value of *Birthweight* for all mothers?

        ii. For mothers who smoke?

        iii. For mothers who do not smoke?

    b. **i.** Use the data in the sample to estimate the difference in average birth weight for smoking and nonsmoking mothers.

        ii. What is the standard error for the estimated difference in (i)?

        iii. Construct a 95% confidence interval for the difference in the average birth weight for smoking and nonsmoking mothers.

    c. Run a regression of *Birthweight* on the binary variable *Smoker*.

        i. Explain how the estimated slope and intercept are related to your answers in parts (a) and (b).

        ii. Explain how the $SE(\hat{\beta}_1)$ is related to your answer in b(ii).

        iii. Construct a 95% confidence interval for the effect of smoking on birth weight.

---

[2] These data were provided by Professors Douglas Almond (Columbia University), Ken Chay (Brown University), and David Lee (Princeton University) and were used in their paper "The Costs of Low Birth Weight," *Quarterly Journal of Economics*, August 2005, 120(3): 1031–1083.

>   **d.** Do you think smoking is uncorrelated with other factors that cause
>   low birth weight? That is, do you think that the regression error term—
>   say, u_i—has a conditional mean of 0 given *Smoking* (X_i)?(You will
>   investigate this further in *Birthweight* and *Smoking* exercises in later
>   chapters.)

5.1 Formulas for OLS Standard Errors

This appendix discusses the formulas for OLS standard errors. These are first presented under
the least squares assumptions in Key Concept 4.3, which allow for heteroskedasticity; these are
the "heteroskedasticity-robust" standard errors. Formulas for the variance of the OLS estimators
and the associated standard errors are then given for the special case of homoskedasticity.

Heteroskedasticity-Robust Standard Errors

The estimator $\hat{\sigma}^2_{\hat{\beta}_1}$ defined in Equation (5.4) is obtained by replacing the population variances
in Equation (4.19) by the corresponding sample variances, with a modification. The variance
in the numerator of Equation (4.19) is estimated by $\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2$, where the divisor
$n - 2$ (instead of n) incorporates a degrees-of-freedom adjustment to correct for downward
bias, analogously to the degrees-of-freedom adjustment used in the definition of the *SER* in
Section 4.3. The variance in the denominator is estimated by $(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2$. Replacing
$\text{var}[(X_i - \mu_X)u_i]$ and $\text{var}(X_i)$ in Equation (4.19) by these two estimators yields $\hat{\sigma}^2_{\hat{\beta}_1}$ in
Equation (5.4). The consistency of heteroskedasticity-robust standard errors is discussed in
Section 18.3.

   The estimator of the variance of $\hat{\beta}_0$ is

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n} \times \frac{\dfrac{1}{n-2}\sum_{i=1}^{n}\hat{H}_i^2\hat{u}_i^2}{\left(\dfrac{1}{n}\sum_{i=1}^{n}\hat{H}_i^2\right)^2}, \tag{5.26}$$

where $\hat{H}_i = 1 - (\overline{X}/\frac{1}{n}\sum_{i=1}^{n}X_i^2)X_i$. The standard error of $\hat{\beta}_0$ is $SE(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_0}}$. The reasoning
behind the estimator $\hat{\sigma}^2_{\hat{\beta}_0}$ is the same as behind $\hat{\sigma}^2_{\hat{\beta}_1}$ and stems from replacing population expec-
tations with sample averages.

Homoskedasticity-Only Variances

Under homoskedasticity, the conditional variance of u_i given X_i is a constant: $\text{var}(u_i \mid X_i) = \sigma_u^2$.
If the errors are homoskedastic, the formulas in Key Concept 4.4 simplify to

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{n\sigma^2_X} \text{ and} \tag{5.27}$$

$$\sigma^2_{\hat{\beta}_0} = \frac{E(X_i^2)}{n\sigma^2_X}\sigma^2_u. \tag{5.28}$$

To derive Equation (5.27), write the numerator in Equation (4.19) as $\text{var}[(X_i - \mu_X)u_i] = E\{(X_i-\mu_X)u_i - E[(X_i-\mu_X)u_i]\}^2 = E\{[(X_i-\mu_X)u_i]^2\} = E[(X_i-\mu_X)^2u_i^2] = E[(X_i-\mu_X)^2 \text{var}(u_i | X_i)]$, where the second equality follows because $E[(X_i - \mu_X)u_i] = 0$ (by the first least squares assumption) and where the final equality follows from the law of iterated expectations (Section 2.3). If u_i is homoskedastic, then $\text{var}(u_i | X_i) = \sigma^2_u$, so $E[(X_i - \mu_X)^2\text{var}(u_i | X_i)] = \sigma^2_u E[(X_i - \mu_X)^2] = \sigma^2_u\sigma^2_X$. The result in Equation (5.27) follows by substituting this expression into the numerator of Equation (4.19) and simplifying. A similar calculation yields Equation (5.28).

Homoskedasticity-Only Standard Errors

The homoskedasticity-only standard errors are obtained by substituting sample means and variances for the population means and variances in Equations (5.27) and (5.28) and by estimating the variance of u_i by the square of the *SER*. The homoskedasticity-only estimators of these variances are

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \frac{s^2_{\hat{u}}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad \text{(homoskedasticity-only) and} \tag{5.29}$$

$$\tilde{\sigma}^2_{\hat{\beta}_0} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)s^2_{\hat{u}}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad \text{(homoskedasticity-only)} \tag{5.30}$$

where $s^2_{\hat{u}}$ is given in Equation (4.17). The homoskedasticity-only standard errors are the square roots of $\tilde{\sigma}^2_{\hat{\beta}_0}$ and $\tilde{\sigma}^2_{\hat{\beta}_1}$.

APPENDIX

5.2 The Gauss–Markov Conditions and a Proof of the Gauss–Markov Theorem

As discussed in Section 5.5, the Gauss–Markov theorem states that if the Gauss–Markov conditions hold, then the OLS estimator is the best (most efficient) conditionally linear unbiased estimator (is BLUE). This appendix begins by stating the Gauss–Markov conditions and showing that they are implied by the three least squares assumptions plus homoskedasticity. We next show that the OLS estimator is a linear conditionally unbiased estimator. Finally, we turn to the proof of the theorem.

The Gauss–Markov Conditions

The three Gauss–Markov conditions are

(i) $E(u_i | X_1, \ldots, X_n) = 0$
(ii) $\text{var}(u_i | X_1, \ldots, X_n) = \sigma_u^2, \quad 0 < \sigma_u^2 < \infty$
(iii) $E(u_i u_j | X_1, \ldots, X_n) = 0, i \neq j,$ (5.31)

where the conditions hold for $i, j = 1, \ldots, n$. The three conditions, respectively, state that u_i has a conditional mean of 0, that u_i has a constant variance, and that the errors are uncorrelated for different observations, where all these statements hold conditionally on all observed X's $(X_1, \ldots, X_n)$.

The **Gauss–Markov conditions** are implied by the three least squares assumptions (Key Concept 4.3), plus the additional assumption that the errors are homoskedastic. Because the observations are i.i.d. (assumption 2), $E(u_i | X_1, \ldots, X_n) = E(u_i | X_i)$, and by assumption 1, $E(u_i | X_i) = 0$; thus condition (i) holds. Similarly, by assumption 2, $\text{var}(u_i | X_1, \ldots, X_n) = \text{var}(u_i | X_i)$, and because the errors are assumed to be homoskedastic, $\text{var}(u_i | X_i) = \sigma_u^2$, which is constant. Assumption 3 (nonzero finite fourth moments) ensures that $0 < \sigma_u^2 < \infty$, so condition (ii) holds. To show that condition (iii) is implied by the least squares assumptions, note that $E(u_i u_j | X_1, \ldots, X_n) = E(u_i u_j | X_i, X_j)$ because (X_i, Y_i) are i.i.d. by assumption 2. Assumption 2 also implies that $E(u_i u_j | X_i, X_j) = E(u_i | X_i) E(u_j | X_j)$ for $i \neq j$; because $E(u_i | X_i) = 0$ for all i, it follows that $E(u_i u_j | X_1, \ldots, X_n) = 0$ for all $i \neq j$, so condition (iii) holds. Thus the least squares assumptions in Key Concept 4.3, plus homoskedasticity of the errors, imply the Gauss–Markov conditions in Equation (5.31).

The OLS Estimator $\hat{\beta}_1$ Is a Linear Conditionally Unbiased Estimator

To show that $\hat{\beta}_1$ is linear, first note that because $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$ (by the definition of $\overline{X}$), $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n}(X_i - \overline{X})Y_i - \overline{Y}\sum_{i=1}^{n}(X_i - \overline{X}) = \sum_{i=1}^{n}(X_i - \overline{X})Y_i$. Substituting this result into the formula for $\hat{\beta}_1$ in Equation (4.5) yields

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{j=1}^{n}(X_j - \overline{X})^2} = \sum_{i=1}^{n}\hat{a}_i Y_i, \quad \text{where } \hat{a}_i = \frac{(X_i - \overline{X})}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \quad (5.32)$$

Because the weights $\hat{a}_i, i = 1, \ldots, n$, in Equation (5.32) depend on $X_1, \ldots, X_n$ but not on $Y_1, \ldots, Y_n$, the OLS estimator $\hat{\beta}_1$ is a linear estimator.

Under the Gauss–Markov conditions, $\hat{\beta}_1$ is conditionally unbiased, and the variance of the conditional distribution of $\hat{\beta}_1$ given $X_1, \ldots, X_n$ is

$$\text{var}(\hat{\beta}_1 | X_1, \ldots, X_n) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}. \quad (5.33)$$

The result that $\hat{\beta}_1$ is conditionally unbiased was previously shown in Appendix 4.3.

Proof of the Gauss–Markov Theorem

We start by deriving some facts that hold for all linear conditionally unbiased estimators—that is, for all estimators $\widetilde{\beta}_1$ satisfying Equations (5.24) and (5.25). Substituting $Y_i = \beta_0 + \beta_1 X_i + u_i$ into $\widetilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$ and collecting terms, we have that

$$\widetilde{\beta}_1 = \beta_0 \left(\sum_{i=1}^{n} a_i \right) + \beta_1 \left(\sum_{i=1}^{n} a_i X_i \right) + \sum_{i=1}^{n} a_i u_i. \tag{5.34}$$

By the first Gauss–Markov condition, $E(\sum_{i=1}^{n} a_i u_i | X_1, \ldots, X_n) = \sum_{i=1}^{n} a_i E(u_i | X_1, \ldots, X_n) = 0$; thus taking conditional expectations of both sides of Equation (5.34) yields $E(\widetilde{\beta}_1 | X_1, \ldots, X_n) = \beta_0 (\sum_{i=1}^{n} a_i) + \beta_1 (\sum_{i=1}^{n} a_i X_i)$. Because $\widetilde{\beta}_1$ is conditionally unbiased by assumption, it must be that $\beta_0 (\sum_{i=1}^{n} a_i) + \beta_1 (\sum_{i=1}^{n} a_i X_i) = \beta_1$, but for this equality to hold for all values of β_0 and β_1, it must be the case that, for $\widetilde{\beta}_1$ to be conditionally unbiased,

$$\sum_{i=1}^{n} a_i = 0 \text{ and } \sum_{i=1}^{n} a_i X_i = 1. \tag{5.35}$$

Under the Gauss–Markov conditions, the variance of $\widetilde{\beta}_1$, conditional on $X_1, \ldots, X_n$, has a simple form. Substituting Equation (5.35) into Equation (5.34) yields $\widetilde{\beta}_1 - \beta_1 = \sum_{i=1}^{n} a_i u_i$. Thus $\text{var}(\widetilde{\beta}_1 | X_1, \ldots, X_n) = \text{var}(\sum_{i=1}^{n} a_i u_i | X_1, \ldots, X_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{cov}(u_i, u_j | X_1, \ldots, X_n)$; applying the second and third Gauss–Markov conditions, the cross terms in the double summation vanish, and the expression for the conditional variance simplifies to

$$\text{var}(\widetilde{\beta}_1 | X_1, \ldots, X_n) = \sigma_u^2 \sum_{i=1}^{n} a_i^2. \tag{5.36}$$

Note that Equations (5.35) and (5.36) apply to $\hat{\beta}_1$ with weights $a_i = \hat{a}_i$, given in Equation (5.32).

We now show that the two restrictions in Equation (5.35) and the expression for the conditional variance in Equation (5.36) imply that the conditional variance of $\widetilde{\beta}_1$ exceeds the conditional variance of $\hat{\beta}_1$ unless $\widetilde{\beta}_1 = \hat{\beta}_1$. Let $a_i = \hat{a}_i + d_i$, so $\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} (\hat{a}_i + d_i)^2 = \sum_{i=1}^{n} \hat{a}_i^2 + 2 \sum_{i=1}^{n} \hat{a}_i d_i + \sum_{i=1}^{n} d_i^2$. Using the definition of $\hat{a}_i$ in Equation (5.32), we have that

$$\sum_{i=1}^{n} \hat{a}_i d_i = \frac{\sum_{i=1}^{n} (X_i - \overline{X}) d_i}{\sum_{j=1}^{n} (X_j - \overline{X})^2} = \frac{\sum_{i=1}^{n} d_i X_i - \overline{X} \sum_{i=1}^{n} d_i}{\sum_{j=1}^{n} (X_j - \overline{X})^2}$$

$$= \frac{\left(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} \hat{a}_i X_i \right) - \overline{X} \left(\sum_{i=1}^{n} a_i - \sum_{i=1}^{n} \hat{a}_i \right)}{\sum_{j=1}^{n} (X_j - \overline{X})^2} = 0,$$

where the penultimate equality follows from $d_i = a_i - \hat{a}_i$ and the final equality follows from Equation (5.35) (which holds for both a_i and $\hat{a}_i$). Thus $\sigma_u^2 \sum_{i=1}^{n} a_i^2 = \sigma_u^2 \sum_{i=1}^{n} \hat{a}_i^2 + \sigma_u^2 \sum_{i=1}^{n} d_i^2 = \text{var}(\hat{\beta}_1 | X_1, \ldots, X_n) + \sigma_u^2 \sum_{i=1}^{n} d_i^2$; substituting this result into Equation (5.36) yields

$$\text{var}(\widetilde{\beta}_1 | X_1, \ldots, X_n) - \text{var}(\hat{\beta}_1 | X_1, \ldots, X_n) = \sigma_u^2 \sum_{i=1}^{n} d_i^2. \tag{5.37}$$

Thus $\widetilde{\beta}_1$ has a greater conditional variance than $\hat{\beta}_1$ if d_i is nonzero for any $i = 1, \ldots, n$. But if $d_i = 0$ for all i, then $a_i = \hat{a}_i$ and $\widetilde{\beta}_1 = \hat{\beta}_1$, which proves that OLS is BLUE.

The Gauss–Markov Theorem When *X* Is Nonrandom

With a minor change in interpretation, the Gauss–Markov theorem also applies to nonrandom regressors; that is, it applies to regressors that do not change their values over repeated samples. Specifically, if the second least squares assumption is replaced by the assumption that $X_1, \ldots, X_n$ are nonrandom (fixed over repeated samples) and $u_1, \ldots, u_n$ are i.i.d., then the foregoing statement and proof of the Gauss–Markov theorem apply directly, except that all of the "conditional on $X_1, \ldots, X_n$" statements are unnecessary because $X_1, \ldots, X_n$ take on the same values from one sample to the next.

The Sample Average Is the Efficient Linear Estimator of *E(Y)*

An implication of the Gauss–Markov theorem is that the sample average, $\overline{Y}$, is the most efficient linear estimator of $E(Y_i)$ when $Y_1, \ldots, Y_n$ are i.i.d. To see this, consider the case of regression without an "X," so that the only regressor is the constant regressor $X_{0i} = 1$. Then the OLS estimator $\hat{\beta}_0 = \overline{Y}$. It follows that, under the Gauss–Markov assumptions, $\overline{Y}$ is BLUE. Note that the Gauss–Markov requirement that the error be homoskedastic is automatically satisfied in this case because there is no regressor, so it follows that $\overline{Y}$ is BLUE if $Y_1, \ldots, Y_n$ are i.i.d. This result was stated previously in Key Concept 3.3.

Linear Regression with Multiple Regressors

Chapter 5 ended on a worried note. Although school districts with lower student–teacher ratios tend to have higher test scores in the California data set, perhaps students from districts with small classes have other advantages that help them perform well on standardized tests. Could this have produced a misleading estimate of the causal effect of class size on test scores, and, if so, what can be done?

Omitted factors, such as student characteristics, can, in fact, make the ordinary least squares (OLS) estimator of the effect of class size on test scores misleading or, more precisely, biased. This chapter explains this "omitted variable bias" and introduces multiple regression, a method that can eliminate omitted variable bias. The key idea of multiple regression is that if we have data on these omitted variables, then we can include them as additional regressors and thereby estimate the causal effect of one regressor (the student–teacher ratio) while holding constant the other variables (such as student characteristics).

Alternatively, if one is interested not in causal inference but in prediction, the multiple regression model makes it possible to use multiple variables as regressors—that is, multiple predictors—to improve upon predictions made using a single regressor.

This chapter explains how to estimate the coefficients of the multiple linear regression model. Many aspects of multiple regression parallel those of regression with a single regressor, studied in Chapters 4 and 5. The coefficients of the multiple regression model can be estimated from data using OLS; the OLS estimators in multiple regression are random variables because they depend on data from a random sample; and in large samples, the sampling distributions of the OLS estimators are approximately normal.

6.1 Omitted Variable Bias

By focusing only on the student–teacher ratio, the empirical analysis in Chapters 4 and 5 ignored some potentially important determinants of test scores by collecting their influences in the regression error term. These omitted factors include school characteristics, such as teacher quality and computer usage, and student characteristics, such as family background. We begin by considering an omitted student characteristic that is particularly relevant in California because of its large immigrant population: the prevalence in the school district of students who are still learning English.

By ignoring the percentage of English learners in the district, the OLS estimator of the effect on test scores of the student–teacher ratio could be biased; that is, the mean of the sampling distribution of the OLS estimator might not equal the true causal

effect on test scores of a unit change in the student–teacher ratio. Here is the reasoning. Students who are still learning English might perform worse on standardized tests than native English speakers. If districts with large classes also have many students still learning English, then the OLS regression of test scores on the student–teacher ratio could erroneously find a correlation and produce a large estimated coefficient, when in fact the true causal effect of cutting class sizes on test scores is small, even zero. Accordingly, based on the analysis of Chapters 4 and 5, the superintendent might hire enough new teachers to reduce the student–teacher ratio by 2, but her hoped-for improvement in test scores will fail to materialize if the true coefficient is small or zero.

A look at the California data lends credence to this concern. The correlation between the student–teacher ratio and the percentage of English learners (students who are not native English speakers and who have not yet mastered English) in the district is 0.19. This small but positive correlation suggests that districts with more English learners tend to have a higher student–teacher ratio (larger classes). If the student–teacher ratio were unrelated to the percentage of English learners, then it would be safe to ignore English proficiency in the regression of test scores against the student–teacher ratio. But because the student–teacher ratio and the percentage of English learners are correlated, it is possible that the OLS coefficient in the regression of test scores on the student–teacher ratio reflects that influence.

Definition of Omitted Variable Bias

If the regressor (the student–teacher ratio) is correlated with a variable that has been omitted from the analysis (the percentage of English learners) and that determines, in part, the dependent variable (test scores), then the OLS estimator will have **omitted variable bias**.

Omitted variable bias occurs when two conditions are true: (1) the omitted variable is correlated with the included regressor and (2) the omitted variable is a determinant of the dependent variable. To illustrate these conditions, consider three examples of variables that are omitted from the regression of test scores on the student–teacher ratio.

Example 1: Percentage of English learners. Because the percentage of English learners is correlated with the student–teacher ratio, the first condition for omitted variable bias holds. It is plausible that students who are still learning English will do worse on standardized tests than native English speakers, in which case the percentage of English learners is a determinant of test scores and the second condition for omitted variable bias holds. Thus the OLS estimator in the regression of test scores on the student–teacher ratio could incorrectly reflect the influence of the omitted variable, the percentage of English learners. That is, omitting the percentage of English learners may introduce omitted variable bias.

Example 2: Time of day of the test. Another variable omitted from the analysis is the time of day that the test was administered. For this omitted variable, it is plausible that the first condition for omitted variable bias does not hold but that the second

Omitted Variable Bias in Regression with a Single Regressor

Omitted variable bias is the bias in the OLS estimator of the causal effect of X on Y that arises when the regressor, X, is correlated with an omitted variable. For omitted variable bias to occur, two conditions must be true:

1. X is correlated with the omitted variable.

2. The omitted variable is a determinant of the dependent variable, Y.

condition does. If the time of day of the test varies from one district to the next in a way that is unrelated to class size, then the time of day and class size would be uncorrelated, so the first condition does not hold. Conversely, the time of day of the test could affect scores (alertness varies through the school day), so the second condition holds. However, because in this example the time of day the test is administered is uncorrelated with the student–teacher ratio, the student–teacher ratio could not be incorrectly picking up the "time of day" effect. Thus omitting the time of day of the test does not result in omitted variable bias.

Example 3: Parking lot space per pupil. Another omitted variable is parking lot space per pupil (the area of the teacher parking lot divided by the number of students). This variable satisfies the first but not the second condition for omitted variable bias. Specifically, schools with more teachers per pupil probably have more teacher parking space, so the first condition would be satisfied. However, under the assumption that learning takes place in the classroom, not the parking lot, parking lot space has no direct effect on learning; thus the second condition does not hold. Because parking lot space per pupil is not a determinant of test scores, omitting it from the analysis does not lead to omitted variable bias.

Omitted variable bias is summarized in Key Concept 6.1.

Omitted variable bias and the first least squares assumption. Omitted variable bias means that the first least squares assumption for causal inference—that $E(u_i | X_i) = 0$, as listed in Key Concept 4.3—does not hold. To see why, recall that the error term u_i in the linear regression model with a single regressor represents all factors, other than X_i, that are determinants of Y_i. If one of these other factors is correlated with X_i, this means that the error term (which contains this factor) is correlated with X_i. In other words, if an omitted variable is a determinant of Y_i, then it is in the error term, and if it is correlated with X_i, then the error term is correlated with X_i. Because u_i and X_i are correlated, the conditional mean of u_i given X_i is nonzero. This correlation therefore violates the first least squares assumption, and the consequence is serious: The OLS estimator is biased. This bias does not vanish even in very large samples, and the OLS estimator is inconsistent.

A Formula for Omitted Variable Bias

The discussion of the previous section about omitted variable bias can be summarized mathematically by a formula for this bias. Let the correlation between X_i and u_i be $\text{corr}(X_i, u_i) = \rho_{Xu}$. Suppose that the second and third least squares assumptions hold, but the first does not because ρ_{Xu} is nonzero. Then the OLS estimator has the limit (derived in Appendix 6.1)

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}. \tag{6.1}$$

That is, as the sample size increases, $\hat{\beta}_1$ is close to $\beta_1 + \rho_{Xu}(\sigma_u/\sigma_X)$ with increasingly high probability.

The formula in Equation (6.1) summarizes several of the ideas discussed above about omitted variable bias:

1. Omitted variable bias is a problem whether the sample size is large or small. Because $\hat{\beta}_1$ does not converge in probability to the true value β_1, $\hat{\beta}_1$ is biased and inconsistent; that is, $\hat{\beta}_1$ is not a consistent estimator of β_1 when there is omitted variable bias. The term $\rho_{Xu}(\sigma_u/\sigma_X)$ in Equation (6.1) is the bias in $\hat{\beta}_1$ that persists even in large samples.

The Mozart Effect: Omitted Variable Bias?

A study published in *Nature* in 1993 (Rauscher, Shaw, and Ky 1993) suggested that listening to Mozart for 10 to 15 minutes could temporarily raise your IQ by 8 or 9 points. That study made big news—and politicians and parents saw an easy way to make their children smarter. For a while, the state of Georgia even distributed classical music CDs to all infants in the state.

What is the evidence for the "Mozart effect"? A review of dozens of studies found that students who take optional music or arts courses in high school do, in fact, have higher English and math test scores than those who don't.[1] A closer look at these studies, however, suggests that the real reason for the better test performance has little to do with those courses. Instead, the authors of the review suggested that the correlation between testing well and taking art or music could arise from any number of things. For example, the academically better students might have more time to take optional music courses or more interest in doing so, or those schools with a deeper music curriculum might just be better schools across the board.

In the terminology of regression, the estimated relationship between test scores and taking optional music courses appears to have omitted variable bias. By omitting factors such as the student's innate ability or the overall quality of the school, studying music appears to have an effect on test scores when in fact it has none.

So is there a Mozart effect? One way to find out is to do a randomized controlled experiment. Randomized controlled experiments eliminate omitted variable bias by randomly assigning participants to treatment and control groups. Taken together, the many controlled experiments on the Mozart effect fail to show that listening to Mozart improves IQ or general test performance. For reasons not fully understood, however, it seems that listening to classical music *does* help temporarily in one narrow area: folding paper and visualizing shapes. So the next time you cram for an origami exam, try to fit in a little Mozart, too.

[1]See the fall/winter 2000 issue of *Journal of Aesthetic Education* 34, especially the article by Ellen Winner and Monica Cooper (pp. 11–76) and the one by Lois Hetland (pp. 105–148).

2. Whether this bias is large or small in practice depends on the correlation ρ_{Xu} between the regressor and the error term. The larger $|\rho_{Xu}|$ is, the larger the bias.

3. The direction of the bias in $\hat{\beta}_1$ depends on whether X and u are positively or negatively correlated. For example, we speculated that the percentage of students learning English has a *negative* effect on district test scores (students still learning English have lower scores), so that the percentage of English learners enters the error term with a negative sign. In our data, the fraction of English learners is *positively* correlated with the student–teacher ratio (districts with more English learners have larger classes). Thus the student–teacher ratio (X) would be *negatively* correlated with the error term (u), so $\rho_{Xu} < 0$ and the coefficient on the student–teacher ratio $\hat{\beta}_1$ would be biased toward a negative number. In other words, having a small percentage of English learners is associated with both *high* test scores and *low* student–teacher ratios, so one reason that the OLS estimator suggests that small classes improve test scores may be that the districts with small classes have fewer English learners.

Addressing Omitted Variable Bias by Dividing the Data into Groups

What can you do about omitted variable bias? In the test score example, class size is correlated with the fraction of English learners. One way to address this problem is to select a subset of districts that have the same fraction of English learners but have different class sizes: For that subset of districts, class size cannot be picking up the English learner effect because the fraction of English learners is held constant. More generally, this observation suggests estimating the effect of the student–teacher ratio on test scores, *holding constant* the percentage of English learners.

Table 6.1 reports evidence on the relationship between class size and test scores within districts with comparable percentages of English learners. Districts are divided into eight groups. First, the districts are broken into four categories that correspond to the quartiles

TABLE 6.1   Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High Student–Teacher Ratio	
	Average Test Score	n	Average Test Score	n	Difference	t-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

of the distribution of the percentage of English learners across districts. Second, within each of these four categories, districts are further broken down into two groups, depending on whether the student–teacher ratio is small ($STR < 20$) or large ($STR \geq 20$).

The first row in Table 6.1 reports the overall difference in average test scores between districts with low and high student–teacher ratios—that is, the difference in test scores between these two groups without breaking them down further into the quartiles of English learners. (Recall that this difference was previously reported in regression form in Equation (5.18) as the OLS estimate of the coefficient on D_i in the regression of *TestScore* on D_i, where D_i is a binary regressor that equals 1 if $STR_i < 20$ and equals 0 otherwise.) Over the full sample of 420 districts, the average test score is 7.4 points higher in districts with a low student–teacher ratio than a high one; the *t*-statistic is 4.04, so the null hypothesis that the mean test score is the same in the two groups is rejected at the 1% significance level.

The final four rows in Table 6.1 report the difference in test scores between districts with low and high student–teacher ratios, broken down by the quartile of the percentage of English learners. This evidence presents a different picture. Of the districts with the fewest English learners ($< 1.9\%$), the average test score for those 76 with low student–teacher ratios is 664.5, and the average for the 27 with high student–teacher ratios is 665.4. Thus, for the districts with the fewest English learners, test scores were, on average, 0.9 points *lower* in the districts with low student–teacher ratios! In the second quartile, districts with low student–teacher ratios had test scores that averaged 3.3 points higher than those with high student–teacher ratios; this gap was 5.2 points for the third quartile and only 1.9 points for the quartile of districts with the most English learners. Once we hold the percentage of English learners constant, the difference in performance between districts with high and low student–teacher ratios is perhaps half (or less) of the overall estimate of 7.4 points.

At first, this finding might seem puzzling. How can the overall effect of test scores be twice the effect of test scores within any quartile? The answer is that the districts with the most English learners tend to have *both* the highest student–teacher ratios *and* the lowest test scores. The difference in the average test scores between districts in the lowest and highest quartiles of the percentage of English learners is large, approximately 30 points. The districts with few English learners tend to have lower student–teacher ratios: 74% (76 of 103) of the districts in the first quartile of English learners have small classes ($STR < 20$), while only 42% (44 of 105) of the districts in the quartile with the most English learners have small classes. So the districts with the most English learners have both lower test scores and higher student–teacher ratios than the other districts.

This analysis reinforces the superintendent's worry that omitted variable bias is present in the regression of test scores against the student–teacher ratio. By looking within quartiles of the percentage of English learners, the test score differences in the second part of Table 6.1 improve on the simple difference-of-means analysis in the first line of Table 6.1. Still, this analysis does not yet provide the superintendent with a useful estimate of the effect on test scores of changing class size, holding constant the fraction of English learners. Such an estimate can be provided, however, using the method of multiple regression.

6.2 The Multiple Regression Model

The **multiple regression model** extends the single variable regression model of Chapters 4 and 5 to include additional variables as regressors. When used for causal inference, this model permits estimating the effect on Y_i of changing one variable (X_{1i}) while holding the other regressors (X_{2i}, X_{3i}, and so forth) constant. In the class size problem, the multiple regression model provides a way to isolate the effect on test scores (Y_i) of the student–teacher ratio (X_{1i}) while holding constant the percentage of students in the district who are English learners (X_{2i}). When used for prediction, the multiple regression model can improve predictions by using multiple variables as predictors.

As in Chapter 4, we introduce the terminology and statistics of multiple regression in the context of prediction. Section 6.5 returns to causal inference and formalizes the requirements for multiple regression to eliminate omitted variable bias in the estimation of a causal effect.

The Population Regression Line

Suppose for the moment that there are only two independent variables, X_{1i} and X_{2i}. In the linear multiple regression model, the average relationship between these two independent variables and the dependent variable, Y, is given by the linear function

$$E(Y_i | X_{1i} = x_1, X_{2i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{6.2}$$

where $E(Y_i | X_{1i} = x_1, X_{2i} = x_2)$ is the conditional expectation of Y_i given that $X_{1i} = x_1$ and $X_{2i} = x_2$. That is, if the student–teacher ratio in the ith district (X_{1i}) equals some value x_1 and the percentage of English learners in the ith district (X_{2i}) equals x_2, then the expected value of Y_i given the student–teacher ratio and the percentage of English learners is given by Equation (6.2).

Equation (6.2) is the **population regression line** or **population regression function** in the multiple regression model. The coefficient β_0 is the **intercept**; the coefficient β_1 is the **slope coefficient of X_{1i}** or, more simply, the **coefficient on X_{1i}**; and the coefficient β_2 is the **slope coefficient of X_{2i}** or, more simply, the **coefficient on X_{2i}**.

The interpretation of the coefficient β_1 in Equation (6.2) is different than it was when X_{1i} was the only regressor: In Equation (6.2), β_1 is the predicted difference in Y between two observations with a unit difference in X_1, **holding X_2 constant** or **controlling for X_2**.

This interpretation of β_1 follows from comparing the predictions (conditional expectations) for two observations with the same value of X_2 but with values of X_1 that differ by ΔX_1, so that the first observation has X values (X_1, X_2) and the second observation has X values ($X_1 + \Delta X_1, X_2$). For the first observation, the predicted value of Y is given by Equation (6.2); write this as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. For the second observation, the predicted value of Y is $Y + \Delta Y$, where

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2. \tag{6.3}$$

An equation for ΔY in terms of ΔX_1 is obtained by subtracting the equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ from Equation (6.3), yielding $\Delta Y = \beta_1 \Delta X_1$. Rearranging this equation shows that

$$\beta_1 = \frac{\Delta Y}{\Delta X_1}, \text{holding } X_2 \text{ constant.} \tag{6.4}$$

Thus the coefficient β_1 is the difference in the predicted values of Y (the difference in the conditional expectations of Y) between two observations with a unit difference in X_1, holding X_2 fixed. Another term used to describe β_1 is the **partial effect** on Y of X_1, holding X_2 fixed.

The interpretation of the intercept in the multiple regression model, β_0, is similar to the interpretation of the intercept in the single-regressor model: It is the expected value of Y_i when X_{1i} and X_{2i} are 0. Simply put, the intercept β_0 determines how far up the Y axis the population regression line starts.

The Population Multiple Regression Model

The population regression line in Equation (6.2) is the relationship between Y and X_1 and X_2 that holds, on average, in the population. Just as in the case of regression with a single regressor, however, this relationship does not hold exactly because many other factors influence the dependent variable. In addition to the student–teacher ratio and the fraction of students still learning English, for example, test scores are influenced by school characteristics, other student characteristics, and luck. Thus the population regression function in Equation (6.2) needs to be augmented to incorporate these additional factors.

Just as in the case of regression with a single regressor, the factors that determine Y_i in addition to X_{1i} and X_{2i} are incorporated into Equation (6.2) as an "error" term u_i. Accordingly, we have

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, i = 1, \ldots, n, \tag{6.5}$$

where the subscript i indicates the ith of the n observations (districts) in the sample.

Equation (6.5) is the **population multiple regression model** when there are two regressors, X_{1i} and X_{2i}.

It can be useful to treat β_0 as the coefficient on a regressor that always equals 1; think of β_0 as the coefficient on X_{0i}, where $X_{0i} = 1$ for $i = 1, \ldots, n$. Accordingly, the population multiple regression model in Equation (6.5) can alternatively be written as

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \text{where } X_{0i} = 1, i = 1, \ldots, n. \tag{6.6}$$

The variable X_{0i} is sometimes called the **constant regressor** because it takes on the same value—the value 1—for all observations. Similarly, the intercept, β_0, is sometimes called the **constant term** in the regression.

The two ways of writing the population regression model, Equations (6.5) and (6.6), are equivalent.

The Multiple Regression Model

The multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, i = 1, \ldots, n, \qquad (6.7)$$

where

- Y_i is i^{th} observation on the dependent variable; $X_{1i}, X_{2i}, \ldots, X_{ki}$ are the ith observations on each of the k regressors; and u_i is the error term.

- The population regression line is the relationship that holds between Y and the X's, on average, in the population:

$$E(Y \mid X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ki} = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

- β_1 is the slope coefficient on X_1, β_2 is the slope coefficient on X_2, and so on. The coefficient β_1 is the expected difference in Y_i associated with a unit difference in X_1, holding constant the other regressors, $X_2, \ldots, X_k$. The coefficients on the other X's are interpreted similarly.

- The intercept β_0 is the expected value of Y when all the X's equal 0. The intercept can be thought of as the coefficient on a regressor, X_0, that equals 1 for all i.

The discussion so far has focused on the case of a single additional variable, X_2. In applications, it is common to have more than two regressors. This reasoning leads us to consider a model that includes k regressors. The multiple regression model with k regressors, $X_{1i}, X_{2i}, \ldots, X_{ki}$, is summarized as Key Concept 6.2.

The definitions of homoskedasticity and heteroskedasticity in the multiple regression model extend their definitions in the single-regressor model. The error term u_i in the multiple regression model is **homoskedastic** if the variance of the conditional distribution of u_i given $X_{1i}, \ldots, X_{ki}$, $\text{var}(u_i \mid X_{1i}, \ldots, X_{ki})$, is constant for $i = 1, \ldots, n$, and thus does not depend on the values of $X_{1i}, \ldots, X_{ki}$. Otherwise, the error term is **heteroskedastic**.

6.3   The OLS Estimator in Multiple Regression

To be of practical value, we need to estimate the unknown population coefficients $\beta_0, \ldots, \beta_k$ using a sample of data. As in regression with a single regressor, these coefficients can be estimated using ordinary least squares.

The OLS Estimator

Section 4.2 shows how to estimate the intercept and slope coefficients in the single-regressor model by applying OLS to a sample of observations of Y and X. The key idea is that these coefficients can be estimated by minimizing the sum of squared prediction mistakes—that is, by choosing the estimators b_0 and b_1 so as to minimize $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$. The estimators that do so are the OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$.

The method of OLS also can be used to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ in the multiple regression model. Let $b_0, b_1, \ldots, b_k$ be estimates of $\beta_0, \beta_1, \ldots, \beta_k$. The predicted value of Y_i, calculated using these estimates, is $b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}$, and the mistake in predicting Y_i is $Y_i - (b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}) = Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki}$. The sum of these squared prediction mistakes over all n observations is thus

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2. \tag{6.8}$$

The sum of the squared mistakes for the linear regression model in Expression (6.8) is the extension of the sum of the squared mistakes given in Equation (4.4) for the linear regression model with a single regressor.

The estimators of the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ that minimize the sum of squared mistakes in Expression (6.8) are called the **ordinary least squares (OLS) estimators of** $\beta_0, \beta_1, \ldots, \beta_k$. The OLS estimators are denoted $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

The terminology of OLS in the linear multiple regression model is the same as in the linear regression model with a single regressor. The **OLS regression line** is the straight line constructed using the OLS estimators: $\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$. The **predicted value** of Y_i given $X_{1i}, \ldots, X_{ki}$, based on the OLS regression line, is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_1 X_{ki}$. The **OLS residual** for the ith observation is the difference between Y_i and its OLS predicted value; that is, the OLS residual is $\hat{u}_i = Y_i - \hat{Y}_i$.

The OLS estimators could be computed by trial and error, repeatedly trying different values of $b_0, \ldots, b_k$ until you are satisfied that you have minimized the total sum of squares in Expression (6.8). It is far easier, however, to use explicit formulas for the OLS estimators that are derived using calculus. The formulas for the OLS estimators in the multiple regression model are similar to those in Key Concept 4.2 for the single-regressor model. These formulas are incorporated into modern statistical software. In the multiple regression model, the formulas are best expressed and discussed using matrix notation, so their presentation is deferred to Section 19.1.

The definitions and terminology of OLS in multiple regression are summarized in Key Concept 6.3.

The OLS Estimators, Predicted Values, and Residuals in the Multiple Regression Model

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the values of $b_0, b_1, \ldots, b_k$ that minimize the sum of squared prediction errors $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})^2$. The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_{1i} + \cdots + \hat{\beta}_kX_{ki}, i = 1, \ldots, n, \text{and} \tag{6.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, i = 1, \ldots, n. \tag{6.10}$$

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ and residual $\hat{u}_i$ are computed from a sample of n observations of $(X_{1i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$. These are estimators of the unknown true population coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and error term u_i.

Application to Test Scores and the Student–Teacher Ratio

In Section 4.2, we used OLS to estimate the intercept and slope coefficient of the regression relating test scores (*TestScore*) to the student–teacher ratio (*STR*), using our 420 observations for California school districts. The estimated OLS regression line, reported in Equation (4.9), is

$$\widehat{TestScore} = 698.9 - 2.28 \times STR. \tag{6.11}$$

From the perspective of the father looking for a way to predict test scores, this relation is not very satisfying: its R^2 is only 0.051; that is, the student–teacher ratio explains only 5.1% of the variation in test scores. Can this prediction be made more precise by including additional regressors?

To find out, we estimate a multiple regression with test scores as the dependent variable (Y_i) and with two regressors: the student–teacher ratio (X_{1i}) and the percentage of English learners in the school district (X_{2i}). The OLS regression line, estimated using our 420 districts ($i = 1, \ldots, 420$), is

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL, \tag{6.12}$$

where *PctEL* is the percentage of students in the district who are English learners. The OLS estimate of the intercept ($\hat{\beta}_0$) is 686.0, the OLS estimate of the coefficient on the student–teacher ratio ($\hat{\beta}_1$) is −1.10, and the OLS estimate of the coefficient on the percentage English learners ($\hat{\beta}_2$) is −0.65.

The coefficient on the student–teacher ratio in the multiple regression is approximately half as large as when the student–teacher ratio is the only regressor, −1.10 vs. −2.28. This difference occurs because the coefficient on *STR* in the multiple

regression holds constant (or controls for) *PctEL*, whereas in the single-regressor regression, *PctEL* is not held constant.

The decline in the magnitude of the coefficient on the student–teacher ratio, once one controls for *PctEL*, parallels the findings in Table 6.1. There we saw that, among schools within the same quartile of percentage of English learners, the difference in test scores between schools with a high vs. a low student–teacher ratio is less than the difference if one does not hold constant the percentage of English learners. As in Table 6.1, this strongly suggests that, from the perspective of causal inference, the original estimate of the effect of the student–teacher ratio on test scores in Equation (6.11) is subject to omitted variable bias.

Equation (6.12) provides multiple regression estimates that the father can use for prediction, now using two predictors; we have not yet, however, answered his question as to whether the quality of that prediction has been improved. To do so, we need to extend the measures of fit in the single-regressor model to multiple regression.

6.4 Measures of Fit in Multiple Regression

Three commonly used summary statistics in multiple regression are the standard error of the regression, the regression R^2, and the adjusted R^2 (also known as $\overline{R}^2$). All three statistics measure how well the OLS estimate of the multiple regression line describes, or "fits," the data.

The Standard Error of the Regression (*SER*)

The standard error of the regression (*SER*) estimates the standard deviation of the error term u_i. Thus the *SER* is a measure of the spread of the distribution of Y around the regression line. In multiple regression, the *SER* is

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2}, \text{ where } s_{\hat{u}}^2 = \frac{1}{n - k - 1} \sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{n - k - 1} \qquad (6.13)$$

and where *SSR* is the sum of squared residuals, $SSR = \sum_{i=1}^{n} \hat{u}_i^2$.

The only difference between the definition of the *SER* in Equation (6.13) and the definition of the *SER* in Section 4.3 for the single-regressor model is that here the divisor is $n - k - 1$ rather than $n - 2$. In Section 4.3, the divisor $n - 2$ (rather than n) adjusts for the downward bias introduced by estimating two coefficients (the slope and intercept of the regression line). Here, the divisor $n - k - 1$ adjusts for the downward bias introduced by estimating $k + 1$ coefficients (the k slope coefficients plus the intercept). As in Section 4.3, using $n - k - 1$ rather than n is called a degrees-of-freedom adjustment. If there is a single regressor, then $k = 1$, so the formula in Section 4.3 is the same as that in Equation (6.13). When n is large, the effect of the degrees-of-freedom adjustment is negligible.

The R^2

The regression R^2 is the fraction of the sample variance of Y_i explained by (or predicted by) the regressors. Equivalently, the R^2 is 1 minus the fraction of the variance of Y_i *not* explained by the regressors.

The mathematical definition of the R^2 is the same as for regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}, \tag{6.14}$$

where the explained sum of squares is $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ and the total sum of squares is $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

In multiple regression, the R^2 increases whenever a regressor is added unless the estimated coefficient on the added regressor is exactly 0. To see this, think about starting with one regressor and then adding a second. When you use OLS to estimate the model with both regressors, OLS finds the values of the coefficients that minimize the sum of squared residuals. If OLS happens to choose the coefficient on the new regressor to be exactly 0, then the SSR will be the same whether or not the second variable is included in the regression. But if OLS chooses any value other than 0, then it must be that this value reduced the SSR relative to the regression that excludes this regressor. In practice, it is extremely unusual for an estimated coefficient to be exactly 0, so in general the SSR will decrease when a new regressor is added. But this means that the R^2 generally increases (and never decreases) when a new regressor is added.

The Adjusted R^2

Because the R^2 increases when a new variable is added, an increase in the R^2 does not mean that adding a variable actually improves the fit of the model. In this sense, the R^2 gives an inflated estimate of how well the regression fits the data. One way to correct for this is to deflate or reduce the R^2 by some factor, and this is what the adjusted R^2, or $\overline{R}^2$, does.

The **adjusted R^2**, or $\overline{R}^2$, is a modified version of the R^2 that does not necessarily increase when a new regressor is added. The $\overline{R}^2$ is

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}. \tag{6.15}$$

The difference between this formula and the second definition of the R^2 in Equation (6.14) is that the ratio of the sum of squared residuals to the total sum of squares is multiplied by the factor $(n-1)/(n-k-1)$. As the second expression in Equation (6.15) shows, this means that the adjusted R^2 is 1 minus the ratio of the sample variance of the OLS residuals [with the degrees-of-freedom correction in Equation (6.13)] to the sample variance of Y.

There are three useful things to know about the $\overline{R}^2$. First, $(n-1)/(n-k-1)$ is always greater than 1, so $\overline{R}^2$ is always less than R^2.

Second, adding a regressor has two opposite effects on the $\overline{R}^2$. On the one hand, the SSR falls, which increases the $\overline{R}^2$. On the other hand, the factor $(n-1)/(n-k-1)$ increases. Whether the $\overline{R}^2$ increases or decreases depends on which of these two effects is stronger.

Third, the $\overline{R}^2$ can be negative. This happens when the regressors, taken together, reduce the sum of squared residuals by such a small amount that this reduction fails to offset the factor $(n-1)/(n-k-1)$.

Application to Test Scores

Equation (6.12) reports the estimated regression line for the multiple regression relating test scores (*TestScore*) to the student–teacher ratio (*STR*) and the percentage of English learners (*PctEL*). The R^2 for this regression line is $R^2 = 0.426$, the adjusted R^2 is $\overline{R}^2 = 0.424$, and the standard error of the regression is $SER = 14.5$.

Comparing these measures of fit with those for the regression in which *PctEL* is excluded [Equation (5.8)] shows that including *PctEL* in the regression increases the R^2 from 0.051 to 0.426. When the only regressor is *STR*, only a small fraction of the variation in *TestScore* is explained; however, when *PctEL* is added to the regression, more than two-fifths (42.6%) of the variation in test scores is explained. In this sense, including the percentage of English learners substantially improves the fit of the regression. Because *n* is large and only two regressors appear in Equation (6.12), the difference between R^2 and adjusted R^2 is very small ($R^2 = 0.426$ vs. $\overline{R}^2 = 0.424$).

The *SER* for the regression excluding *PctEL* is 18.6; this value falls to 14.5 when *PctEL* is included as a second regressor. The units of the *SER* are points on the standardized test. The reduction in the *SER* tells us that predictions about standardized test scores are substantially more precise if they are made using the regression with both *STR* and *PctEL* than if they are made using the regression with only *STR* as a regressor.

***Using the* R^2 *and adjusted* R^2.** The $\overline{R}^2$ is useful because it quantifies the extent to which the regressors account for, or explain, the variation in the dependent variable. Nevertheless, heavy reliance on the $\overline{R}^2$ (or R^2) can be a trap.

In applications in which the goal is to produce reliable out-of-sample predictions, including many regressors can produce a good in-sample fit but can degrade the out-of-sample performance. Although the $\overline{R}^2$ improves upon the R^2 for this purpose, simply maximizing the $\overline{R}^2$ still can produce poor out-of-sample forecasts. We return to this issue in Chapter 14.

In applications in which the goal is causal inference, the decision about whether to include a variable in a multiple regression should be based on whether including that variable allows you better to estimate the causal effect of interest. The least

squares assumptions for causal inference in multiple regression make precise the requirements for an included variable to eliminate omitted variable bias, and we now turn to those assumptions.

6.5 The Least Squares Assumptions for Causal Inference in Multiple Regression

In this section, we make precise the requirements for OLS to provide valid inferences about causal effects. We consider the case in which we are interested in knowing the causal effects of all k regressors in the multiple regression model; that is, all the coefficients $\beta_1, \ldots, \beta_k$ are causal effects of interest. Section 6.8 presents the least squares assumptions that apply when only some of the coefficients are causal effects, while the rest are coefficients on variables included to control for omitted factors and do not necessarily have a causal interpretation. Appendix 6.4 provides the least squares assumptions for prediction with multiple regression.

There are four least squares assumptions for causal inference in the multiple regression model. The first three are those of Section 4.3 for the single-regressor model (Key Concept 4.3) extended to allow for multiple regressors, and they are discussed here only briefly. The fourth assumption is new and is discussed in more detail.

Assumption 1: The Conditional Distribution of u_i Given $X_{1i}, X_{2i}, \ldots, X_{ki}$ Has a Mean of 0

The first assumption is that the conditional distribution of u_i given $X_{1i}, \ldots, X_{ki}$ has a mean of 0. This assumption extends the first least squares assumption with a single regressor to multiple regressors. This assumption is implied if $X_{1i}, \ldots, X_{ki}$ are randomly assigned or are as-if randomly assigned; if so, for any value of the regressors, the expected value of u_i is 0. As is the case for regression with a single regressor, this is the key assumption that makes the OLS estimators unbiased.

Assumption 2: $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, Are i.i.d.

The second assumption is that $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, are independently and identically distributed (i.i.d.) random variables. This assumption holds automatically if the data are collected by simple random sampling. The comments on this assumption appearing in Section 4.3 for a single regressor also apply to multiple regressors.

Assumption 3: Large Outliers Are Unlikely

The third least squares assumption is that large outliers—that is, observations with values far outside the usual range of the data—are unlikely. This assumption serves as a reminder that, as in the single-regressor case, the OLS estimator of the coefficients in the multiple regression model can be sensitive to large outliers.

The assumption that large outliers are unlikely is made mathematically precise by assuming that $X_{1i}, \ldots, X_{ki}$ and Y_i have nonzero finite fourth moments: $0 < E(X_{1i}^4) < \infty, \ldots, 0 < E(X_{ki}^4) < \infty$ and $0 < E(Y_i^4) < \infty$. Another way to state this assumption is that the dependent variable and regressors have finite kurtosis. This assumption is used to derive the properties of OLS regression statistics in large samples.

Assumption 4: No Perfect Multicollinearity

The fourth assumption is new to the multiple regression model. It rules out an inconvenient situation called perfect multicollinearity, in which it is impossible to compute the OLS estimator. The regressors are said to exhibit **perfect multicollinearity** (or to be perfectly multicollinear) if one of the regressors is a perfect linear function of the other regressors. The fourth least squares assumption is that the regressors are not perfectly multicollinear.

Why does perfect multicollinearity make it impossible to compute the OLS estimator? Suppose you want to estimate the coefficient on *STR* in a regression of *TestScore_i* on *STR_i* and *PctEL_i* but you make a typographical error and accidentally type in *STR_i* a second time instead of *PctEL_i*; that is, you regress *TestScore_i* on *STR_i* and *STR_i*. This is a case of perfect multicollinearity because one of the regressors (the first occurrence of *STR*) is a perfect linear function of another regressor (the second occurrence of *STR*). Depending on how your software package handles perfect multicollinearity, if you try to estimate this regression, the software will do one of two things: Either it will drop one of the occurrences of *STR*, or it will refuse to calculate the OLS estimates and give an error message. The mathematical reason for this failure is that perfect multicollinearity produces division by 0 in the OLS formulas.

At an intuitive level, perfect multicollinearity is a problem because you are asking the regression to answer an illogical question. In multiple regression, the coefficient on one of the regressors is the effect of a change in that regressor, holding the other regressors constant. In the hypothetical regression of *TestScore* on *STR* and *STR*, the coefficient on the first occurrence of *STR* is the effect on test scores of a change in *STR*, holding constant *STR*. This makes no sense, and OLS cannot estimate this nonsensical partial effect.

The solution to perfect multicollinearity in this hypothetical regression is simply to correct the typo and to replace one of the occurrences of *STR* with the variable you originally wanted to include. This example is typical: When perfect multicollinearity occurs, it often reflects a logical mistake in choosing the regressors or some previously unrecognized feature of the data set. In general, the solution to perfect multicollinearity is to modify the regressors to eliminate the problem.

Additional examples of perfect multicollinearity are given in Section 6.7, which also defines and discusses imperfect multicollinearity.

The least squares assumptions for the multiple regression model are summarized in Key Concept 6.4.

The Least Squares Assumptions for Causal Inference in the Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, i = 1, \ldots, n,$$

where $\beta_1, \ldots, \beta_k$ are causal effects and

1. u_i has a conditional mean of 0 given $X_{1i}, X_{2i}, \ldots, X_{ki}$; that is,

$$E(u_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}) = 0.$$

2. $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n,$ are independently and identically distributed (i.i.d.) draws from their joint distribution.

3. Large outliers are unlikely: $X_{1i}, \ldots, X_{ki}$ and Y_i have nonzero finite fourth moments.

4. There is no perfect multicollinearity.

6.6   The Distribution of the OLS Estimators in Multiple Regression

Because the data differ from one sample to the next, different samples produce different values of the OLS estimators. This variation across possible samples gives rise to the uncertainty associated with the OLS estimators of the population regression coefficients, $\beta_0, \beta_1, \ldots, \beta_k$. Just as in the case of regression with a single regressor, this variation is summarized in the sampling distribution of the OLS estimators.

Recall from Section 4.4 that, under the least squares assumptions, the OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) are unbiased and consistent estimators of the unknown coefficients (β_0 and β_1) in the linear regression model with a single regressor. In addition, in large samples, the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is well approximated by a bivariate normal distribution.

These results carry over to multiple regression analysis. That is, under the least squares assumptions of Key Concept 6.4, the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are unbiased and consistent estimators of $\beta_0, \beta_1, \ldots, \beta_k$ in the linear multiple regression model. In large samples, the joint sampling distribution of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ is well approximated by a multivariate normal distribution, which is the extension of the bivariate normal distribution to the general case of two or more jointly normal random variables (Section 2.4).

Although the algebra is more complicated when there are multiple regressors, the central limit theorem applies to the OLS estimators in the multiple regression model for the same reason that it applies to $\overline{Y}$ and to the OLS estimators when there

Large-Sample Distribution of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$

If the least squares assumptions (Key Concept 6.4) hold, then in large samples the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are jointly normally distributed, and each $\hat{\beta}_j$ is distributed $N(\beta_j, \sigma^2_{\hat{\beta}_j})$, $j = 0, \ldots, k$.

is a single regressor: The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are averages of the randomly sampled data, and if the sample size is sufficiently large, the sampling distribution of those averages becomes normal. Because the multivariate normal distribution is best handled mathematically using matrix algebra, the expressions for the joint distribution of the OLS estimators are deferred to Chapter 19.

Key Concept 6.5 summarizes the result that, in large samples, the distribution of the OLS estimators in multiple regression is approximately jointly normal. In general, the OLS estimators are correlated; this correlation arises from the correlation between the regressors. The joint sampling distribution of the OLS estimators is discussed in more detail for the case where there are two regressors and homoskedastic errors in Appendix 6.2, and the general case is discussed in Section 19.2.

6.7 Multicollinearity

As discussed in Section 6.5, perfect multicollinearity arises when one of the regressors is a perfect linear combination of the other regressors. This section provides some examples of perfect multicollinearity and discusses how perfect multicollinearity can arise, and can be avoided, in regressions with multiple binary regressors. Imperfect multicollinearity arises when one of the regressors is very highly correlated—but not perfectly correlated—with the other regressors. Unlike perfect multicollinearity, imperfect multicollinearity does not prevent estimation of the regression, nor does it imply a logical problem with the choice of regressors. However, it does mean that one or more regression coefficients could be estimated imprecisely.

Examples of Perfect Multicollinearity

We continue the discussion of perfect multicollinearity from Section 6.5 by examining three additional hypothetical regressions. In each, a third regressor is added to the regression of $TestScore_i$ on STR_i and $PctEL_i$ in Equation (6.12).

Example 1: Fraction of English learners. Let $FracEL_i$ be the fraction of English learners in the ith district, which varies between 0 and 1. If the variable $FracEL_i$ were included as a third regressor in addition to STR_i and $PctEL_i$, the regressors would be

perfectly multicollinear. The reason is that *PctEL* is the *percentage* of English learners, so that $PctEL_i = 100 \times FracEL_i$ for every district. Thus one of the regressors ($PctEL_i$) can be written as a perfect linear function of another regressor ($FracEL_i$).

Because of this perfect multicollinearity, it is impossible to compute the OLS estimates of the regression of $TestScore_i$ on STR_i, $PctEL_i$, and $FracEL_i$. At an intuitive level, OLS fails because you are asking, What is the effect of a unit change in the *percentage* of English learners, holding constant the *fraction* of English learners? Because the percentage of English learners and the fraction of English learners move together in a perfect linear relationship, this question makes no sense, and OLS cannot answer it.

Example 2: "Not very small" classes.   Let NVS_i be a binary variable that equals 1 if the student–teacher ratio in the ith district is "not very small"; specifically, NVS_i equals 1 if $STR_i \geq 12$ and equals 0 otherwise. This regression also exhibits perfect multicollinearity, but for a more subtle reason than the regression in the previous example. There are, in fact, no districts in our data set with $STR_i < 12$; as you can see in the scatterplot in Figure 4.2, the smallest value of STR is 14. Thus $NVS_i = 1$ for all observations. Now recall that the linear regression model with an intercept can equivalently be thought of as including a regressor, X_{0i}, that equals 1 for all i, as shown in Equation (6.6). Thus we can write $NVS_i = 1 \times X_{0i}$ for all the observations in our data set; that is, NVS_i can be written as a perfect linear combination of the regressors; specifically, it equals X_{0i}.

This illustrates two important points about perfect multicollinearity. First, when the regression includes an intercept, then one of the regressors that can be implicated in perfect multicollinearity is the constant regressor X_{0i}. Second, perfect multicollinearity is a statement about the data set you have on hand. While it is possible to imagine a school district with fewer than 12 students per teacher, there are no such districts in our data set, so we cannot analyze them in our regression.

Example 3: Percentage of English speakers.   Let $PctES_i$ be the percentage of English speakers in the ith district, defined to be the percentage of students who are not English learners. Again the regressors will be perfectly multicollinear. Like the previous example, the perfect linear relationship among the regressors involves the constant regressor X_{0i}: For every district, $PctES_i = 100 - PctEL_i = 100 \times X_{0i} - PctEL_i$ because $X_{0i} = 1$ for all i.

This example illustrates another point: Perfect multicollinearity is a feature of the entire set of regressors. If either the intercept (that is, the regressor X_{0i}) or $PctEL_i$ were excluded from this regression, the regressors would not be perfectly multicollinear.

The dummy variable trap.   Another possible source of perfect multicollinearity arises when multiple binary, or dummy, variables are used as regressors. For example, suppose you have partitioned the school districts into three categories: rural,

suburban, and urban. Each district falls into one (and only one) category. Let these binary variables be $Rural_i$, which equals 1 for a rural district and equals 0 otherwise; $Suburban_i$; and $Urban_i$. If you include all three binary variables in the regression along with a constant, the regressors will be perfectly multicollinear: Because each district belongs to one and only one category, $Rural_i + Suburban_i + Urban_i = 1 = X_{0i}$, where X_{0i} denotes the constant regressor introduced in Equation (6.6). Thus, to estimate the regression, you must exclude one of these four variables, either one of the binary indicators or the constant term. By convention, the constant term is typically retained, in which case one of the binary indicators is excluded. For example, if $Rural_i$ were excluded, then the coefficient on $Suburban_i$ would be the average difference between test scores in suburban and rural districts, holding constant the other variables in the regression.

In general, if there are G binary variables, if each observation falls into one and only one category, if there is an intercept in the regression, and if all G binary variables are included as regressors, then the regression will fail because of perfect multicollinearity. This situation is called the **dummy variable trap**. The usual way to avoid the dummy variable trap is to exclude one of the binary variables from the multiple regression, so only $G - 1$ of the G binary variables are included as regressors. In this case, the coefficients on the included binary variables represent the incremental effect of being in that category, relative to the base case of the omitted category, holding constant the other regressors. Alternatively, all G binary regressors can be included if the intercept is omitted from the regression.

Solutions to perfect multicollinearity.   Perfect multicollinearity typically arises when a mistake has been made in specifying the regression. Sometimes the mistake is easy to spot (as in the first example), but sometimes it is not (as in the second example). In one way or another, your software will let you know if you make such a mistake because it cannot compute the OLS estimator if you have.

When your software lets you know that you have perfect multicollinearity, it is important that you modify your regression to eliminate it. You should understand the source of the multicollinearity. Some software is unreliable when there is perfect multicollinearity, and at a minimum, you will be ceding control over your choice of regressors to your computer if your regressors are perfectly multicollinear.

Imperfect Multicollinearity

Despite its similar name, imperfect multicollinearity is conceptually quite different from perfect multicollinearity. **Imperfect multicollinearity** means that two or more of the regressors are highly correlated in the sense that there is a linear function of the regressors that is highly correlated with another regressor. Imperfect multicollinearity does not pose any problems for the theory of the OLS estimators; on the contrary, one use of OLS is to sort out the independent influences of the various regressors when the regressors are correlated.

If the regressors are imperfectly multicollinear, then the coefficients on at least one individual regressor will be imprecisely estimated. For example, consider the regression of *TestScore* on *STR* and *PctEL*. Suppose we were to add a third regressor, the percentage of the district's residents who are first-generation immigrants. First-generation immigrants often speak English as a second language, so the variables *PctEL* and percentage immigrants will be highly correlated: Districts with many recent immigrants will tend to have many students who are still learning English. Because these two variables are highly correlated, it would be difficult to use these data to estimate the coefficient on *PctEL*, holding constant the percentage of immigrants. In other words, the data set provides little information about what happens to test scores when the percentage of English learners is low but the fraction of immigrants is high, or vice versa. As a result, the OLS estimator of the coefficient on *PctEL* in this regression will have a larger variance than if the regressors *PctEL* and percentage immigrants were uncorrelated.

The effect of imperfect multicollinearity on the variance of the OLS estimators can be seen mathematically by inspecting Equation (6.20) in Appendix 6.2, which is the variance of $\hat{\beta}_1$ in a multiple regression with two regressors (X_1 and X_2) for the special case of a homoskedastic error. In this case, the variance of $\hat{\beta}_1$ is inversely proportional to $1 - \rho_{X_1, X_2}^2$, where ρ_{X_1, X_2} is the correlation between X_1 and X_2. The larger the correlation between the two regressors, the closer this term is to 0, and the larger is the variance of $\hat{\beta}_1$. More generally, when multiple regressors are imperfectly multicollinear, the coefficients on one or more of these regressors will be imprecisely estimated; that is, they will have a large sampling variance.

Perfect multicollinearity is a problem that often signals the presence of a logical error. In contrast, imperfect multicollinearity is not necessarily an error but rather just a feature of OLS, your data, and the question you are trying to answer. If the variables in your regression are the ones you meant to include—the ones you chose to address the potential for omitted variable bias—then imperfect multicollinearity implies that it will be difficult to estimate precisely one or more of the partial effects using the data at hand.

6.8 Control Variables and Conditional Mean Independence

In the test score example, we included the percentage of English learners in the regression to address omitted variable bias in the estimate of the effect of class size. Specifically, by including percent English learners in the regression, we were able to estimate the effect of class size, controlling for the percent English learners.

In this section, we make explicit the distinction between a regressor for which we wish to estimate a causal effect—that is, a variable of interest—and control variables. A **control variable** is not the object of interest in the study; rather, it is a regressor included to hold constant factors that, if neglected, could lead the estimated causal

effect of interest to suffer from omitted variable bias. This distinction leads to a modification of the first least squares assumption in Key Concept 6.4, in which some of the variables are control variables. If this alternative assumption holds, the OLS estimator of the effect of interest is unbiased, but the OLS coefficients on control variables are, in general, biased and do not have a causal interpretation.

For example, consider the potential omitted variable bias arising from omitting outside learning opportunities from a test score regression. Although "outside learning opportunities" is a broad concept that is difficult to measure, those opportunities are correlated with the students' economic background, which can be measured. Thus a measure of economic background can be included in a test score regression to control for omitted income-related determinants of test scores, like outside learning opportunities. To this end, we augment the regression of test scores on STR and $PctEL$ with the percentage of students receiving a free or subsidized school lunch ($LchPct$). Students are eligible for this program if their family income is less than a certain threshold (approximately 150% of the poverty line), so $LchPct$ measures the fraction of economically disadvantaged children in the district. The estimated regression is

$$\widehat{TestScore} = 700.2 - 1.00 \times STR - 0.122 \times PctEL - 0.547 \times LchPct. \quad (6.16)$$

In this regression, the coefficient on the student–teacher ratio is the effect of the student–teacher ratio on test scores, controlling for the percentage of English learners and the percentage eligible for a reduced-price lunch. Including the control variable $LchPct$ does not substantially change any conclusions about the class size effect: The coefficient on STR changes only slightly from its value of -1.10 in Equation (6.12) to -1.00 in Equation (6.16).

What does one make of the coefficient on $LchPct$ in Equation (6.16)? That coefficient is very large: The difference in test scores between a district with $LchPct = 0\%$ and one with $LchPct = 50\%$ is estimated to be 27.4 points $[= 0.547 \times (50 - 0)]$, approximately the difference between the 75th and 25th percentiles of test scores in Table 4.1. Does this coefficient have a causal interpretation? Suppose that upon seeing Equation (6.16) the superintendent proposed eliminating the reduced-price lunch program so that, for her district, $LchPct$ would immediately drop to 0. Would eliminating the lunch program boost her district's test scores? Common sense suggests that the answer is no; in fact, by leaving some students hungry, eliminating the reduced-price lunch program might well have the opposite effect. But does it make sense to treat as causal the coefficient on the variable of interest STR but not the coefficient on the control variable $LchPct$?

Control Variables and Conditional Mean Independence

To distinguish between variables of interest and control variables, we modify the notation of the linear regression model to include k variables of interest, denoted by

The Least Squares Assumptions for Causal Inference in the Multiple Regression Model with Control Variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, i = 1, \ldots, n,$$

where $\beta_1, \ldots, \beta_k$ are causal effects; the W's are control variables; and

1. u_i has a conditional mean that does not depend on the X's given the W's; that is,

$$E(u_i | X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}) = E(u_i | W_{1i}, \ldots, W_{ri})$$

(conditional mean independence). $\qquad$ (6.17)

2. $(X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Y_i), i = 1, \ldots, n,$ are independently and identically distributed (i.i.d.) draws from their joint distribution.

3. Large outliers are unlikely: $X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri},$ and Y_i have nonzero finite fourth moments.

4. There is no perfect multicollinearity.

X, and r control variables, denoted by W. Accordingly, the **multiple regression model with control variables** is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i, i = 1, \ldots, n. \quad (6.18)$$

The coefficients on the X's, $\beta_1, \ldots, \beta_k,$ are causal effects of interest.

The reason for including control variables in multiple regression is to make the variables of interest no longer correlated with the error term, once the control variables are held constant. This idea is made precise by replacing assumption 1 in Key Concept 6.4 with an assumption called conditional mean independence. **Conditional mean independence** requires that the conditional expectation of u_i given the variable of interest and the control variables does not depend on (is independent of) the variable of interest, although it can depend on control variables.

The least squares assumptions for causal inference with control variables are summarized in Key Concept 6.6. The first of these assumptions is a mathematical statement of the conditional mean independence requirement. The remaining three assumptions are extensions of their counterparts in Key Concept 6.4.

The idea of conditional mean independence is that once you control for the W's, the X's can be treated as if they were randomly assigned, in the sense that the conditional mean of the error term no longer depends on X. Controlling for W makes the X's uncorrelated with the error term, so that OLS can estimate the causal effects on Y of a change in each of the X's. The control variables, however, remain correlated with the error term, so the coefficients on the control variables are subject to omitted variable bias and do not have a causal interpretation. The mathematics of this

interpretation is laid out in Appendix 6.5, where it is shown that if conditional mean independence holds, then the OLS estimators of the coefficients on the X's are unbiased estimators of the causal effects of the X's, but the OLS estimators of the coefficients on the W's are in general biased. This bias does not pose a problem because we are interested in the coefficients on the X's, not on the W's.

In the class size example, *LchPct* can be correlated with factors, such as learning opportunities outside school, that enter the error term; indeed, it is *because* of this correlation that *LchPct* is a useful control variable. This correlation between *LchPct* and the error term means that the estimated coefficient on *LchPct* does not have a causal interpretation. What the conditional mean independence assumption requires is that, given the control variables in the regression (*PctEL* and *LchPct*), the mean of the error term does not depend on the student–teacher ratio. Said differently, conditional mean independence says that among schools with the same values of *PctEL* and *LchPct*, class size is "as-if" randomly assigned: Including *PctEL* and *LchPct* in the regression controls for omitted factors so that *STR* is uncorrelated with the error term. If so, the coefficient on the student–teacher ratio has a causal interpretation even though the coefficient on *LchPct* does not.

The first least squares assumption for multiple regression with control variables makes precise the requirement needed to eliminate the omitted variable bias with which this chapter began: Given, or holding constant, the values of the control variables, the variable of interest is as-if randomly assigned in the sense that the mean of the error term no longer depends on X given the control variables. This requirement serves as a useful guide for choosing of control variables and for judging their adequacy.

6.9 Conclusion

Regression with a single regressor is vulnerable to omitted variable bias: If an omitted variable is a determinant of the dependent variable and is correlated with the regressor, then the OLS estimator of the causal effect will be biased and will reflect both the effect of the regressor and the effect of the omitted variable. Multiple regression makes it possible to mitigate or eliminate omitted variable bias by including the omitted variable in the regression. The coefficient on a regressor, X_1, in multiple regression is the partial effect of a change in X_1, holding constant the other included regressors. In the test score example, including the percentage of English learners as a regressor made it possible to estimate the effect on test scores of a change in the student–teacher ratio, holding constant the percentage of English learners. Doing so reduced by half the estimated effect on test scores of a change in the student–teacher ratio.

The statistical theory of multiple regression builds on the statistical theory of regression with a single regressor. The least squares assumptions for multiple regression are extensions of the three least squares assumptions for regression with a single

regressor, plus a fourth assumption ruling out perfect multicollinearity. Because the regression coefficients are estimated using a single sample, the OLS estimators have a joint sampling distribution and therefore have sampling uncertainty. This sampling uncertainty must be quantified as part of an empirical study, and the ways to do so in the multiple regression model are the topic of the next chapter.

Summary

1. Omitted variable bias occurs when an omitted variable (a) is correlated with an included regressor and (b) is a determinant of Y.

2. The multiple regression model is a linear regression model that includes multiple regressors, $X_1, X_2, \ldots, X_k$. Associated with each regressor is a regression coefficient, $\beta_1, \beta_2, \ldots, \beta_k$. The coefficient β_1 is the expected difference in Y associated with a one-unit difference in X_1, holding the other regressors constant. The other regression coefficients have an analogous interpretation.

3. The coefficients in multiple regression can be estimated by OLS. When the four least squares assumptions in Key Concept 6.4 are satisfied, the OLS estimators of the causal effect are unbiased, consistent, and normally distributed in large samples.

4. The role of control variables is to hold constant omitted factors so that the variable of interest is no longer correlated with the error term. Properly chosen control variables can eliminate omitted variable bias in the OLS estimate of the causal effect of interest.

5. Perfect multicollinearity, which occurs when one regressor is an exact linear function of the other regressors, usually arises from a mistake in choosing which regressors to include in a multiple regression. Solving perfect multicollinearity requires changing the set of regressors.

6. The standard error of the regression, the R^2, and the $\overline{R}^2$ are measures of fit for the multiple regression model.

Key Terms

omitted variable bias (170)

multiple regression model (175)

population regression line (175)

population regression function (175)

intercept (175)

slope coefficient of X_{1i} (175)

coefficient on X_{1i} (175)

slope coefficient of X_{2i} (175)

coefficient on X_{2i} (175)

holding X_2 constant (175)

controlling for X_2 (175)

partial effect (176)

population multiple regression
    model (176)

constant regressor (176)

constant term (176)

homoskedastic (177)

heteroskedastic (177)

ordinary least squares (OLS)
estimators of $\beta_0, \beta_1, \ldots, \beta_k$ (178)
OLS regression line (178)
predicted value (178)
OLS residual (178)
R^2 (181)
adjusted $R^2(\overline{R}^2)$ (181)
perfect multicollinearity (184)

dummy variable trap (188)
imperfect multicollinearity (188)
control variable (189)
multiple regression model with control
variables (191)
conditional mean independence (191)

Review the Concepts

6.1   A researcher is interested in the effect on test scores of computer usage. Using
school district data like that used in this chapter, she regresses district average
test scores on the number of computers per student. Will $\hat{\beta}_1$ be an unbiased
estimator of the effect on test scores of increasing the number of comput-
ers per student? Why or why not? If you think $\hat{\beta}_1$ is biased, is it biased up or
down? Why?

6.2   A multiple regression includes two regressors: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$.
What is the expected change in Y if X_1 increases by 3 units and X_2 is
unchanged? What is the expected change in Y if X_2 decreases by 5 units and
X_1 is unchanged? What is the expected change in Y if X_1 increases by 3 units
and X_2 decreases by 5 units?

6.3   How does $\overline{R}^2$ differ from R^2? Why is $\overline{R}^2$ useful in a regression model with
multiple regressors?

6.4   Explain why two perfectly multicollinear regressors cannot be included in a
linear multiple regression. Give two examples of a pair of perfectly multicol-
linear regressors.

6.5   Explain why it is difficult to estimate precisely the partial effect of X_1, holding
X_2 constant, if X_1 and X_2 are highly correlated.

Exercises

The first four exercises refer to the table of estimated regressions on page 196, computed using data for 2015 from the Current Population Survey. The data set consists of information on 7178 full-time, full-year workers. The highest educational achievement for each worker was either a high school diploma or a bachelor's degree. The workers' ages ranged from 25 to 34 years. The data set also contains information on the region of the country where the person lived, marital status, and number of children. For the purposes of these exercises, let

AHE = average hourly earnings
$College$ = binary variable (1 if college, 0 if high school)
$Female$ = binary variable (1 if female, 0 if male)
Age = age (in years)
$Northeast$ = binary variable (1 if Region = Northeast, 0 otherwise)
$Midwest$ = binary variable (1 if Region = Midwest, 0 otherwise)
$South$ = binary variable (1 if Region = South, 0 otherwise)
$West$ = binary variable (1 if Region = West, 0 otherwise)

6.1 Compute $\overline{R}^2$ for each of the regressions.

6.2 Using the regression results in column (1):

    a. Do workers with college degrees earn more, on average, than workers with only high school diplomas? How much more?

    b. Do men earn more than women, on average? How much more?

6.3 Using the regression results in column (2):

    a. Is age an important determinant of earnings? Explain.

    b. Sally is a 29-year-old female college graduate. Betsy is a 34-year-old female college graduate. Predict Sally's and Betsy's earnings.

6.4 Using the regression results in column (3):

    a. Do there appear to be important regional differences?

    b. Why is the regressor *West* omitted from the regression? What would happen if it were included?

    c. Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

6.5 Data were collected from a random sample of 220 home sales from a community in 2013. Let *Price* denote the selling price (in $1000s), *BDR* denote the number of bedrooms, *Bath* denote the number of bathrooms, *Hsize* denote the size of the house (in square feet), *Lsize* denote the lot size (in square feet),

Results of Regressions of Average Hourly Earnings on Sex and Education Binary Variables and Other Characteristics, Using 2015 Data from the Current Population Survey

Dependent variable: average hourly earnings (AHE).

Regressor	(1)	(2)	(3)
College (X_1)	10.47	10.44	10.42
Female (X_2)	−4.69	−4.56	−4.57
Age (X_3)		0.61	0.61
Northeast (X_4)			0.74
Midwest (X_5)			−1.54
South (X_6)			−0.44
Intercept	18.15	0.11	0.33
Summary Statistics			
SER	12.15	12.03	12.01
R^2	0.165	0.182	0.185
$\overline{R}^2$			
n	7178	7178	7178

Age denote the age of the house (in years), and *Poor* denote a binary variable that is equal to 1 if the condition of the house is reported as "poor." An estimated regression yields

$$\widehat{Price} = 119.2 + 0.485BDR + 23.4Bath + 0.156Hsize + 0.002Lsize$$
$$+ 0.090Age - 48.8Poor, \overline{R}^2 = 0.72, SER = 41.5.$$

a. Suppose a homeowner converts part of an existing family room in her house into a new bathroom. What is the expected increase in the value of the house?

b. Suppose a homeowner adds a new bathroom to her house, which increases the size of the house by 100 square feet. What is the expected increase in the value of the house?

c. What is the loss in value if a homeowner lets his house run down, so that its condition becomes "poor"?

d. Compute the R^2 for the regression.

6.6 A researcher plans to study the causal effect of police on crime, using data from a random sample of U.S. counties. He plans to regress the county's crime rate on the (per capita) size of the county's police force.

    a. Explain why this regression is likely to suffer from omitted variable bias. Which variables would you add to the regression to control for important omitted variables?

    b. Use your answer to (a) and the expression for omitted variable bias given in Equation (6.1) to determine whether the regression will likely over- or underestimate the effect of police on the crime rate. That is, do you think that $\hat{\beta}_1 > \beta_1$ or $\hat{\beta}_1 < \beta_1$?

6.7 Critique each of the following proposed research plans. Your critique should explain any problems with the proposed research and describe how the research plan might be improved. Include a discussion of any additional data that need to be collected and the appropriate statistical techniques for analyzing those data.

    a. A researcher is interested in determining whether a large aerospace firm is guilty of sex bias in setting wages. To determine potential bias, the researcher collects data on salary and sex for all of the firm's engineers. The researcher then plans to conduct a difference-in-means test to determine whether the average salary for women is significantly less than the average salary for men.

    b. A researcher is interested in determining whether time spent in prison has a permanent effect on a person's wage rate. He collects data on a random sample of people who have been out of prison for at least 15 years. He collects similar data on a random sample of people who have never served time in prison. The data set includes information on each person's current wage, education, age, ethnicity, sex, tenure (time in current job), occupation, and union status, as well as whether the person has ever been incarcerated. The researcher plans to estimate the effect of incarceration on wages by regressing wages on an indicator variable for incarceration, including in the regression the other potential determinants of wages (education, tenure, union status, and so on).

6.8 A recent study found that the death rate for people who sleep 6 to 7 hours per night is lower than the death rate for people who sleep 8 or more hours. The 1.1 million observations used for this study came from a random survey of Americans aged 30 to 102. Each survey respondent was tracked for 4 years. The death rate for people sleeping 7 hours was calculated as the ratio of the number of deaths over the span of the study among people sleeping 7 hours to the total number of survey respondents who slept 7 hours. This calculation was then repeated for people sleeping 6 hours and so on. Based on this summary, would you recommend that Americans who sleep 9 hours per night consider reducing their sleep to 6 or 7 hours if they want to prolong their lives? Why or why not? Explain.

6.9 (Y_i, X_{1i}, X_{2i}) satisfy the assumptions in Key Concept 6.4. You are interested in β_1, the causal effect of X_1 on Y. Suppose X_1 and X_2 are uncorrelated. You estimate β_1 by regressing Y onto X_1 (so that X_2 is not included in the regression). Does this estimator suffer from omitted variable bias? Explain.

6.10 (Y_i, X_{1i}, X_{2i}) satisfy the assumptions in Key Concept 6.4; in addition, $\text{var}(u_i \mid X_{1i}, X_{2i}) = 4$ and $\text{var}(X_{1i}) = 6$. A random sample of size $n = 400$ is drawn from the population.

a. Assume that X_1 and X_2 are uncorrelated. Compute the variance of $\hat{\beta}_1$. [*Hint:* Look at Equation (6.20) in Appendix 6.2.]

b. Assume that $\text{corr}(X_1, X_2) = 0.5$. Compute the variance of $\hat{\beta}_1$.

c. Comment on the following statements: "When X_1 and X_2 are correlated, the variance of $\hat{\beta}_1$ is larger than it would be if X_1 and X_2 were uncorrelated. Thus, if you are interested in β_1, it is best to leave X_2 out of the regression if it is correlated with X_1."

6.11 (Requires calculus) Consider the regression model

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

for $i = 1, \ldots, n$. (Notice that there is no constant term in the regression.) Following analysis like that used in Appendix 4.2:

a. Specify the least squares function that is minimized by OLS.

b. Compute the partial derivatives of the objective function with respect to b_1 and b_2.

c. Suppose that $\sum_{i=1}^{n} X_{1i} X_{2i} = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^{n} X_{1i} Y_i / \sum_{i=1}^{n} X_{1i}^2$.

d. Suppose that $\sum_{i=1}^{n} X_{1i} X_{2i} \neq 0$. Derive an expression for $\hat{\beta}_1$ as a function of the data (Y_i, X_{1i}, X_{2i}), $i = 1, \ldots, n$.

e. Suppose that the model includes an intercept: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Show that the least squares estimators satisfy $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$.

f. As in (e), suppose that the model contains an intercept. Also suppose that $\sum_{i=1}^{n} (X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2) = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^{n} (X_{1i} - \overline{X}_1)(Y_i - \overline{Y}) / \sum_{i=1}^{n} (X_{1i} - \overline{X}_1)^2$. How does this compare to the OLS estimator of β_1 from the regression that omits X_2?

6.12 A school district undertakes an experiment to estimate the effect of class size on test scores in second-grade classes. The district assigns 50% of its previous year's first graders to small second-grade classes (18 students per classroom) and 50% to regular-size classes (21 students per classroom). Students new to the district are handled differently: 20% are randomly assigned to small classes and 80% to regular-size classes. At the end of the second-grade school year, each student is given a standardized exam. Let Y_i denote the exam score for the ith student, X_i denote a binary variable that equals 1 if the student is assigned to a small class, and W_i denote a binary variable that equals 1 if the student is newly enrolled. Let β_1 denote the causal effect on test scores of reducing class size from regular to small.

a. Consider the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$. Do you think that $E(u_i | X_i) = 0$? Is the OLS estimator of β_1 unbiased and consistent? Explain.

b. Consider the regression $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$. Do you think that $E(u_i | X_i, W_i)$ depends on X_i? Is the OLS estimator of β_1 unbiased and consistent? Explain. Do you think that $E(u_i | X_i, W_i)$ depends on W_i? Will the OLS estimator of β_2 provide an unbiased and consistent estimate of the causal effect of transferring to a new school (that is, being a newly enrolled student)? Explain.

Empirical Exercises

(Only two empirical exercises for this chapter are given in the text, but you can find more on the text website, **http://www.pearsonhighered.com/stock_watson/**.)

E6.1   Use the **Birthweight_Smoking** data set introduced in Empirical Exercise E5.3 to answer the following questions.

    a. Regress *Birthweight* on *Smoker*. What is the estimated effect of smoking on birth weight?

    b. Regress *Birthweight* on *Smoker*, *Alcohol*, and *Nprevist*.

      i. Using the two conditions in Key Concept 6.1, explain why the exclusion of *Alcohol* and *Nprevist* could lead to omitted variable bias in the regression estimated in (a).

      ii. Is the estimated effect of smoking on birth weight substantially different from the regression that excludes *Alcohol* and *Nprevist*? Does the regression in (a) seem to suffer from omitted variable bias?

      iii. Jane smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of Jane's child.

      iv. Compute R^2 and $\bar{R}^2$. Why are they so similar?

      v. How should you interpret the coefficient on *Nprevist*? Does the coefficient measure a causal effect of prenatal visits on birth weight? If not, what does it measure?

    c. Estimate the coefficient on *Smoking* for the multiple regression model in (b), using the three-step process in Appendix 6.3 (the Frisch–Waugh theorem). Verify that the three-step process yields the same estimated coefficient for *Smoking* as that obtained in (b).

    d. An alternative way to control for prenatal visits is to use the binary variables *Tripre0* through *Tripre3*. Regress *Birthweight* on *Smoker*, *Alcohol*, *Tripre0*, *Tripre2*, and *Tripre3*.

    i. Why is *Tripre1* excluded from the regression? What would happen if you included it in the regression?

   ii. The estimated coefficient on *Tripre0* is large and negative. What does this coefficient measure? Interpret its value.

  iii. Interpret the value of the estimated coefficients on *Tripre2* and *Tripre3*.

  iv. Does the regression in (d) explain a larger fraction of the variance in birth weight than the regression in (b)?

E6.2   Using the data set **Growth** described in Empirical Exercise E4.1, but excluding the data for Malta, carry out the following exercises.

  a. Construct a table that shows the sample mean, standard deviation, and minimum and maximum values for the series *Growth*, *TradeShare*, *YearsSchool*, *Oil*, *Rev_Coups*, *Assassinations*, and *RGDP60*. Include the appropriate units for all entries.

  b. Run a regression of *Growth* on *TradeShare*, *YearsSchool*, *Rev_Coups*, *Assassinations*, and *RGDP60*. What is the value of the coefficient on *Rev_Coups*? Interpret the value of this coefficient. Is it large or small in a real-world sense?

  c. Use the regression to predict the average annual growth rate for a country that has average values for all regressors.

  d. Repeat (c), but now assume that the country's value for *TradeShare* is one standard deviation above the mean.

  e. Why is *Oil* omitted from the regression? What would happen if it were included?

APPENDIX

6.1 Derivation of Equation (6.1)

This appendix presents a derivation of the formula for omitted variable bias in Equation (6.1). Equation (4.28) in Appendix 4.3 states

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{6.19}$$

Under the last two assumptions in Key Concept 4.3, $(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2 \xrightarrow{p} \sigma_X^2$ and $(1/n)\sum_{i=1}^{n}(X_i - \overline{X})u_i \xrightarrow{p} \text{cov}(u_i, X_i) = \rho_{Xu}\sigma_u\sigma_X$. Substitution of these limits into Equation (6.19) yields Equation (6.1).

6.2 Distribution of the OLS Estimators When There Are Two Regressors and Homoskedastic Errors

Although the general formula for the variance of the OLS estimators in multiple regression is complicated, if there are two regressors ($k = 2$) and the errors are homoskedastic, then the formula simplifies enough to provide some insights into the distribution of the OLS estimators.

Because the errors are homoskedastic, the conditional variance of u_i can be written as $\text{var}(u_i \mid X_{1i}, X_{2i}) = \sigma_u^2$. When there are two regressors, X_{1i} and X_{2i}, and the error term is homoskedastic, in large samples the sampling distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$, where the variance of this distribution, $\sigma_{\hat{\beta}_1}^2$, is

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n}\left(\frac{1}{1 - \rho_{X_1, X_2}^2}\right)\frac{\sigma_u^2}{\sigma_{X_1}^2}, \tag{6.20}$$

where ρ_{X_1, X_2} is the population correlation between the two regressors X_1 and X_2 and $\sigma_{X_1}^2$ is the population variance of X_1.

The variance $\sigma_{\hat{\beta}_1}^2$ of the sampling distribution of $\hat{\beta}_1$ depends on the squared correlation between the regressors. If X_1 and X_2 are highly correlated, either positively or negatively, then ρ_{X_1, X_2}^2 is close to 1, so the term $1 - \rho_{X_1, X_2}^2$ in the denominator of Equation (6.20) is small and the variance of $\hat{\beta}_1$ is larger than it would be if ρ_{X_1, X_2} were close to 0.

Another feature of the joint normal large-sample distribution of the OLS estimators is that $\hat{\beta}_1$ and $\hat{\beta}_2$ are, in general, correlated. When the errors are homoskedastic, the correlation between the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ is the negative of the correlation between the two regressors (see Exercise 19.18):

$$\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = -\rho_{X_1, X_2}. \tag{6.21}$$

6.3 The Frisch–Waugh Theorem

The OLS estimator in multiple regression can be computed by a sequence of shorter regressions. Consider the multiple regression model in Equation (6.7). The OLS estimator of β_1 can be computed in three steps:

1. Regress X_1 on $X_2, X_3, \ldots, X_k$, and let $\widetilde{X}_1$ denote the residuals from this regression;

2. Regress Y on $X_2, X_3, \ldots, X_k$, and let $\widetilde{Y}$ denote the residuals from this regression; and

3. Regress $\widetilde{Y}$ on $\widetilde{X}_1$,

where the regressions include a constant term (intercept). The Frisch–Waugh theorem states that the OLS coefficient in step 3 equals the OLS coefficient on X_1 in the multiple regression model [Equation (6.7)].

This result provides a mathematical statement of how the multiple regression coefficient $\hat{\beta}_1$ estimates the effect on Y of X_1, controlling for the other X's: Because the first two regressions (steps 1 and 2) remove from Y and X_1 their variation associated with the other X's, the third regression estimates the effect on Y of X_1 using what is left over after removing (controlling for) the effect of the other X's. The Frisch–Waugh theorem is proven in Exercise 19.17.

This theorem suggests how Equation (6.20) can be derived from Equation (5.27). Because $\hat{\beta}_1$ is the OLS regression coefficient from the regression of $\tilde{Y}$ onto $\tilde{X}_1$, Equation (5.27) suggests that the homoskedasticity-only variance of $\hat{\beta}_1$ is $\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{n\sigma^2_{\tilde{X}_1}}$, where $\sigma^2_{\tilde{X}_1}$ is the variance of $\tilde{X}_1$. Because $\tilde{X}_1$ is the residual from the regression of X_1 onto X_2 (recall that Equation (6.20) pertains to the model with $k = 2$ regressors), Equation (6.15) implies that $s^2_{\tilde{X}_1} = (1 - \overline{R}^2_{X_1, X_2})s^2_{X_1}$, where $\overline{R}^2_{X_1, X_2}$ is the adjusted R^2 from the regression of X_1 onto X_2. Equation (6.20) follows from $s^2_{\tilde{X}_1} \xrightarrow{p} \sigma^2_{\tilde{X}_1}$, $\overline{R}^2_{X_1, X_2} \xrightarrow{p} \rho^2_{X_1, X_2}$, and $s^2_{X_1} \xrightarrow{p} \sigma^2_{X_1}$.

APPENDIX

6.4 The Least Squares Assumptions for Prediction with Multiple Regressors

This appendix extends the least squares assumptions for prediction with a single regressor in Appendix 4.4 to multiple regressors. It then discusses the unbiasedness of the OLS estimator of the population regression line and the unbiasedness of the forecasts.

Adopt the notation of the least square assumptions for prediction with a single regressor in Appendix 4.4, so that the out-of-sample ("oos") observation is $(X_1^{oos}, \ldots, X_k^{oos}, Y^{oos})$. The aim is to predict Y^{oos} given $X_1^{oos}, \ldots, X_k^{oos}$. Let $(X_{1i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$, be the data used to estimate the regression coefficients. The least squares assumptions for prediction with multiple regressors are

$$E(Y|X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \text{ and } u = Y - E(Y|X_1, \ldots, X_k), \text{ where}$$

1. $(X_1^{oos}, \ldots, X_k^{oos}, Y^{oos})$ are randomly drawn from the same population distribution as $(X_{1i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$.

2. $(X_{1i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$, are i.i.d. draws from their joint distribution.

3. Large outliers are unlikely: $X_{1i}, \ldots, X_{ki}$ and Y_i have nonzero finite fourth moments.

4. There is no perfect multicollinearity.

As in the case of a single X in Appendix 4.4, for prediction the β's are defined to be the coefficients of the population conditional expectation. These β's may or may not have a causal interpretation. Assumption 1 ensures that this conditional expectation, estimated using the in-sample data, is the same as the conditional expectation that applies to the out-of-sample

prediction observation. The remaining assumptions are technical assumptions that play the same role as they do for causal inference.

Under the definition that the β's are the coefficients of the linear conditional expectation, the error u necessarily has a conditional mean of 0, so that $E(u_i | X_{1i}, \ldots, X_{ki}) = 0$. Thus the calculations in Chapter 19 show that the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are unbiased for the respective population slope coefficients. Under the additional technical conditions of assumptions 2–4, the OLS estimators are consistent for these conditional expectation slope coefficients and are normally distributed in large samples.

The unbiasedness of the out-of-sample forecast follows from the unbiasedness of the OLS estimators and the first prediction assumption, which ensures that the out-of-sample observation and in-sample observations are independently drawn from the same distribution. Specifically,

$$
\begin{aligned}
E(&\hat{Y}^{oos} | X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}) \\
&= E(\hat{\beta}_0 + \hat{\beta}_1 X_1^{oos} + \cdots + \hat{\beta}_k X_k^{oos} | X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}) \\
&= E(\hat{\beta}_0 | X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}) + E(\hat{\beta}_1 X_1^{oos} | X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}) \\
&\quad + \cdots + E(\hat{\beta}_k X_1^{oos} | X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}) \\
&= \beta_0 + \beta_1 x_1^{oos} + \cdots + \beta_k x_k^{oos} \\
&= E(Y^{oos} | X_1^{oos} = x_1^{oos}, \ldots, X_k^{oos} = x_k^{oos}),
\end{aligned}
\tag{6.22}
$$

where the third equality follows from the independence of the out-of-sample and in-sample observations and from the unbiasedness of the OLS estimators for the population slope coefficients of the in-sample conditional expectation, and where the final equality follows from the in- and out-of-sample observations being drawn from the same distribution.

6.5 Distribution of OLS Estimators in Multiple Regression with Control Variables

This appendix shows that under least squares assumption 1 for multiple regression with control variables [Equation (6.18)], the OLS coefficient estimator is unbiased for the causal effect of the variables of interest. Moreover, with the addition of technical assumptions 2–4 in Key Concept 6.6, the OLS estimator is a consistent estimator of the causal effect and has a normal distribution in large samples. The OLS estimator of the coefficients on the control variables estimates the slope coefficient in a conditional expectation and is normally distributed in large samples around that slope coefficient; however, that slope coefficient does not, in general, have a causal interpretation.

As we have throughout, assume that conditional expectations are linear, so that the conditional mean independence assumption is

$$
E(u_i | X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}) = E(u_i | W_{1i}, \ldots, W_{ri}) = \gamma_0 + \gamma_1 W_{1i} + \cdots + \gamma_k W_{ki}, \tag{6.23}
$$

where the γ's are coefficients. Then the conditional expectation of Y_i is

$E(Y_i | X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri})$
$= E(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i | X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri})$
$= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + E(u_i | X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri})$
$= (\beta_0 + \gamma_0) + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + (\beta_{k+1} + \gamma_1) W_{1i} + \cdots + (\beta_{k+r} + \gamma_r) W_{ri}$
$= \delta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \delta_1 W_{1i} + \cdots + \delta_r W_{ri},$ \hfill (6.24)

where the first equality uses Equation (6.17), the second equality distributes the conditional expectation, the third equality uses Equation (6.23), and the fourth equality defines $\delta_0 = \beta_0 + \gamma_0$ and $\delta_j = \beta_{k+j} + \gamma_j, j = 1, \ldots, r$.

It follows from Equation (6.24) that we can rewrite the multiple regression model with control variables as

$$Y = \delta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \delta_1 W_{1i} + \cdots + \delta_r W_{ri} + v_i, \hfill (6.25)$$

where the error term v_i has a conditional mean of 0: $E(v_i | X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}) = 0$. Thus, for this rewritten regression, the least squares assumptions in Key Concept 6.4 apply, with the reinterpretation of the coefficients as being those of Equation (6.24).

Three conclusions follow from the rewritten form of the multiple regression model with control variables given in Equation (6.25). First, OLS provides unbiased estimators for the β's and δ's in Equation (6.25), and under the additional assumptions 2–4 of Key Concept 6.6, the OLS estimators are consistent and have a normal distribution in large samples. Second, under the conditional mean independence assumption, the OLS estimators of the coefficients on the X's have a causal interpretation; that is, they are unbiased for the causal effects $\beta_1, \ldots, \beta_k$. Third, the coefficients on the control variables do not, in general, have a causal interpretation. The reason is that those coefficients estimate any direct causal effect of the control variables, plus a term (the γ's) arising because of correlation between u_i and the control variable. Thus, under conditional mean independence, the OLS estimator of the coefficients on the control variables, in general, suffer from omitted variable bias, even though the coefficients on the variables of interest do not.

Hypothesis Tests and Confidence Intervals in Multiple Regression

As discussed in Chapter 6, multiple regression analysis provides a way to mitigate the problem of omitted variable bias by including additional regressors, thereby controlling for the effects of those additional regressors. The coefficients of the multiple regression model can be estimated by OLS. Like all estimators, the OLS estimator has sampling uncertainty because its value differs from one sample to the next.

This chapter presents methods for quantifying the sampling uncertainty of the OLS estimator through the use of standard errors, statistical hypothesis tests, and confidence intervals. One new possibility that arises in multiple regression is a hypothesis that simultaneously involves two or more regression coefficients. The general approach to testing such "joint" hypotheses involves a new test statistic, the F-statistic.

Section 7.1 extends the methods for statistical inference in regression with a single regressor to multiple regression. Sections 7.2 and 7.3 show how to test hypotheses that involve two or more regression coefficients. Section 7.4 extends the notion of confidence intervals for a single coefficient to confidence sets for multiple coefficients. Deciding which variables to include in a regression is an important practical issue, so Section 7.5 discusses ways to approach this problem. In Section 7.6, we apply multiple regression analysis to obtain improved estimates of the causal effect on test scores of a reduction in the student–teacher ratio using the California test score data set.

7.1 Hypothesis Tests and Confidence Intervals for a Single Coefficient

This section describes how to compute the standard error, how to test hypotheses, and how to construct confidence intervals for a single coefficient in a multiple regression equation.

Standard Errors for the OLS Estimators

Recall that, in the case of a single regressor, it was possible to estimate the variance of the OLS estimator by substituting sample averages for expectations, which led to the estimator $\hat{\sigma}^2_{\hat{\beta}_1}$ given in Equation (5.4). Under the least squares assumptions, the law of large numbers implies that these sample averages converge to their population counterparts, so, for example, $\hat{\sigma}^2_{\hat{\beta}_1}/\sigma^2_{\hat{\beta}_1} \xrightarrow{p} 1$. The square root of $\hat{\sigma}^2_{\hat{\beta}_1}$ is the standard error of $\hat{\beta}_1$, $SE(\hat{\beta}_1)$, an estimator of the standard deviation of the sampling distribution of $\hat{\beta}_1$.

All this extends directly to multiple regression. The OLS estimator $\hat{\beta}_j$ of the jth regression coefficient has a standard deviation, and this standard deviation is estimated by its standard error, $SE(\hat{\beta}_j)$. The formula for the standard error is best stated using matrices (see Section 19.2). The important point is that, as far as standard errors are concerned, there is nothing conceptually different between the single- and multiple-regressor cases. The key ideas—the large-sample normality of the estimators and the ability to estimate consistently the standard deviation of their sampling distribution—are the same whether there are one, two, or a dozen regressors.

Hypothesis Tests for a Single Coefficient

Suppose that you want to test the hypothesis that a change in the student–teacher ratio has no effect on test scores, holding constant the percentage of English learners in the district. This corresponds to hypothesizing that the true coefficient β_1 on the student–teacher ratio is 0 in the population regression of test scores on *STR* and *PctEL*. More generally, we might want to test the hypothesis that the true coefficient β_j on the jth regressor takes on some specific value, $\beta_{j,0}$. The null value $\beta_{j,0}$ comes either from economic theory or, as in the student–teacher ratio example, from the decision-making context of the application. If the alternative hypothesis is two-sided, then the two hypotheses can be written mathematically as

$$H_0 : \beta_j = \beta_{j,0} \text{ vs. } H_1 : \beta_j \neq \beta_{j,0} \quad \text{(two-sided alternative)}. \tag{7.1}$$

For example, if the first regressor is *STR*, then the null hypothesis that changing the student–teacher ratio has no effect on test scores corresponds to the null hypothesis that $\beta_1 = 0$ (so $\beta_{1,0} = 0$). Our task is to test the null hypothesis H_0 against the alternative H_1 using a sample of data.

Key Concept 5.2 gives a procedure for testing this null hypothesis when there is a single regressor. The first step in this procedure is to calculate the standard error of the coefficient. The second step is to calculate the *t*-statistic using the general formula in Key Concept 5.1. The third step is to compute the *p*-value of the test using the cumulative normal distribution in Appendix Table 1 or, alternatively, to compare the *t*-statistic to the critical value corresponding to the desired significance level of the test. The theoretical underpinnings of this procedure are that the OLS estimator has a large-sample normal distribution that, under the null hypothesis, has as its mean the hypothesized true value and that the variance of this distribution can be estimated consistently.

These underpinnings are present in multiple regression as well. As stated in Key Concept 6.5, the sampling distribution of $\hat{\beta}_j$ is approximately normal. Under the null hypothesis, the mean of this distribution is $\beta_{j,0}$. The variance of this distribution can be estimated consistently. Therefore we can simply follow the same procedure as in the single-regressor case to test the null hypothesis in Equation (7.1).

Testing the Hypothesis $\beta_j = \beta_{j,0}$ Against the Alternative $\beta_j \neq \beta_{j,0}$

1. Compute the standard error of $\hat{\beta}_j$, $SE(\hat{\beta}_j)$.

2. Compute the *t*-statistic:

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}. \tag{7.2}$$

3. Compute the *p*-value:

$$p\text{-value} = 2\Phi(-|t^{act}|), \tag{7.3}$$

where t^{act} is the value of the *t*-statistic actually computed. Reject the hypothesis at the 5% significance level if the *p*-value is less than 0.05 or, equivalently, if $|t^{act}| > 1.96$.

The standard error and (typically) the *t*-statistic and *p*-value testing $\beta_j = 0$ are computed automatically by regression software.

The procedure for testing a hypothesis on a single coefficient in multiple regression is summarized as Key Concept 7.1. The *t*-statistic actually computed is denoted t^{act} in this box. However, it is customary to denote this simply as *t*, and we adopt this simplified notation for the rest of the book.

Confidence Intervals for a Single Coefficient

The method for constructing a confidence interval in the multiple regression model is also the same as in the single-regressor model. This method is summarized as Key Concept 7.2.

The method for conducting a hypothesis test in Key Concept 7.1 and the method for constructing a confidence interval in Key Concept 7.2 rely on the large-sample normal approximation to the distribution of the OLS estimator $\hat{\beta}_j$. Accordingly, it should be kept in mind that these methods for quantifying the sampling uncertainty are only guaranteed to work in large samples.

Application to Test Scores and the Student–Teacher Ratio

Can we reject the null hypothesis that a change in the student–teacher ratio has no effect on test scores, once we control for the percentage of English learners in the district? What is a 95% confidence interval for the effect on test scores of a change in the student–teacher ratio, controlling for the percentage of English learners? We are now able to find out. The regression of test scores against *STR* and *PctEL*,

Confidence Intervals for a Single Coefficient in Multiple Regression

7.2

A 95% two-sided confidence interval for the coefficient β_j is an interval that contains the true value of β_j with a 95% probability; that is, it contains the true value of β_j in 95% of all possible randomly drawn samples. Equivalently, it is the set of values of β_j that cannot be rejected by a 5% two-sided hypothesis test. When the sample size is large, the 95% confidence interval is

$$95\% \text{ confidence interval for } \beta_j = [\hat{\beta}_j - 1.96\, SE(\hat{\beta}_j),\ \hat{\beta}_j + 1.96\, SE(\hat{\beta}_j)]. \quad (7.4)$$

A 90% confidence interval is obtained by replacing 1.96 in Equation (7.4) with 1.64.

estimated by OLS, was given in Equation (6.12) and is restated here with standard errors in parentheses below the coefficients:

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.650 \times PctEL. \quad (7.5)$$
$$(8.7) \quad (0.43) \quad\quad\quad (0.031)$$

To test the hypothesis that the true coefficient on STR is 0, we first need to compute the t-statistic in Equation (7.2). Because the null hypothesis says that the true value of this coefficient is 0, the t-statistic is $t = (-1.10 - 0)/0.43 = -2.54$. The associated p-value is $2\Phi(-2.54) = 1.1\%$; that is, the smallest significance level at which we can reject the null hypothesis is 1.1%. Because the p-value is less than 5%, the null hypothesis can be rejected at the 5% significance level (but not quite at the 1% significance level).

A 95% confidence interval for the population coefficient on STR is $-1.10 \pm 1.96 \times 0.43 = (-1.95, -0.26)$; that is, we can be 95% confident that the true value of the coefficient is between -1.95 and -0.26. Interpreted in the context of the superintendent's interest in decreasing the student–teacher ratio by 2, the 95% confidence interval for the effect on test scores of this reduction is $(-0.26 \times -2, -1.95 \times -2) = (0.52, 3.90)$.

Adding expenditures per pupil to the equation. Your analysis of the multiple regression in Equation (7.5) has persuaded the superintendent that, based on the evidence so far, reducing class size will improve test scores in her district. Now, however, she moves on to a more nuanced question. If she is to hire more teachers, she can pay for those teachers either by making cuts elsewhere in the budget (no new computers, reduced maintenance, and so on) or by asking for an increase in her budget, which taxpayers do not favor. What, she asks, is the effect on test scores of reducing the student–teacher ratio, holding expenditures per pupil (and the percentage of English learners) constant?

This question can be addressed by estimating a regression of test scores on the student–teacher ratio, total spending per pupil, and the percentage of English learners. The OLS regression line is

$$\widehat{TestScore} = 649.6 - 0.29 \times STR + 3.87 \times Expn - 0.656 \times PctEL, \quad (7.6)$$
$$(15.5) \quad (0.48) \qquad\quad (1.59) \qquad\qquad (0.032)$$

where *Expn* is total annual expenditures per pupil in the district in thousands of dollars.

The result is striking. Holding expenditures per pupil and the percentage of English learners constant, changing the student–teacher ratio is estimated to have a very small effect on test scores: The estimated coefficient on *STR* is -1.10 in Equation (7.5), but after adding *Expn* as a regressor in Equation (7.6), it is only -0.29. Moreover, the *t*-statistic for testing that the true value of the coefficient is 0 is now $t = (-0.29 - 0)/0.48 = -0.60$, so the hypothesis that the population value of this coefficient is indeed 0 cannot be rejected even at the 10% significance level $(|-0.60| < 1.64)$. Thus Equation (7.6) provides no evidence that hiring more teachers improves test scores if overall expenditures per pupil are held constant.

One interpretation of the regression in Equation (7.6) is that, in these California data, school administrators allocate their budgets efficiently. Suppose, counterfactually, that the coefficient on *STR* in Equation (7.6) were negative and large. If so, school districts could raise their test scores simply by decreasing funding for other purposes (textbooks, technology, sports, and so on) and using those funds to hire more teachers, thereby reducing class sizes while holding expenditures constant. However, the small and statistically insignificant coefficient on *STR* in Equation (7.6) indicates that this transfer would have little effect on test scores. Put differently, districts are already allocating their funds efficiently.

Note that the standard error on *STR* increased when *Expn* was added, from 0.43 in Equation (7.5) to 0.48 in Equation (7.6). This illustrates the general point, introduced in Section 6.7 in the context of imperfect multicollinearity, that correlation between regressors (the correlation between *STR* and *Expn* is -0.62) can make the OLS estimators less precise.

What about our angry taxpayer? He asserts that the population values of *both* the coefficient on the student–teacher ratio (β_1) *and* the coefficient on spending per pupil (β_2) are 0; that is, he hypothesizes that both $\beta_1 = 0$ and $\beta_2 = 0$. Although it might seem that we can reject this hypothesis because the *t*-statistic testing $\beta_2 = 0$ in Equation (7.6) is $t = 3.87/1.59 = 2.43$, this reasoning is flawed. The taxpayer's hypothesis is a joint hypothesis, and to test it we need a new tool, the *F*-statistic.

7.2   Tests of Joint Hypotheses

This section describes how to formulate joint hypotheses on multiple regression coefficients and how to test them using an *F*-statistic.

Testing Hypotheses on Two or More Coefficients

Joint null hypotheses. Consider the regression in Equation (7.6) of the test score against the student–teacher ratio, expenditures per pupil, and the percentage of English learners. Our angry taxpayer hypothesizes that neither the student–teacher ratio nor expenditures per pupil have an effect on test scores, once we control for the percentage of English learners. Because *STR* is the first regressor in Equation (7.6) and *Expn* is the second, we can write this hypothesis mathematically as

$$H_0\colon \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ vs. } H_1\colon \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0. \tag{7.7}$$

The hypothesis that *both* the coefficient on the student–teacher ratio (β_1) *and* the coefficient on expenditures per pupil (β_2) are 0 is an example of a joint hypothesis on the coefficients in the multiple regression model. In this case, the null hypothesis restricts the value of two of the coefficients, so as a matter of terminology we can say that the null hypothesis in Equation (7.7) imposes two **restrictions** on the multiple regression model: $\beta_1 = 0$ *and* $\beta_2 = 0$.

In general, a **joint hypothesis** is a hypothesis that imposes two or more restrictions on the regression coefficients. We consider joint null and alternative hypotheses of the form

$$H_0\colon \beta_j = \beta_{j,0},\ \beta_m = \beta_{m,0},\ \ldots,\ \text{for a total of } q \text{ restrictions, vs.}$$
$$H_1\colon \text{one or more of the } q \text{ restrictions under } H_0 \text{ does not hold}, \tag{7.8}$$

where β_j, β_m, $\ldots$, refer to different regression coefficients and $\beta_{j,0}$, $\beta_{m,0}$, $\ldots$, refer to the values of these coefficients under the null hypothesis. The null hypothesis in Equation (7.7) is an example of Equation (7.8). Another example is that, in a regression with $k = 6$ regressors, the null hypothesis is that the coefficients on the second, fourth, and fifth regressors are 0; that is, $\beta_2 = 0$, $\beta_4 = 0$, and $\beta_5 = 0$, so that there are $q = 3$ restrictions. In general, under the null hypothesis H_0, there are q such restrictions.

If at least one of the equalities comprising the null hypothesis H_0 in Equation (7.8) is false, then the joint null hypothesis itself is false. Thus the alternative hypothesis is that at least one of the equalities in the null hypothesis H_0 does not hold.

Why can't I just test the individual coefficients one at a time? Although it seems it should be possible to test a joint hypothesis by using the usual *t*-statistics to test the restrictions one at a time, the following calculation shows that this approach is unreliable. Specifically, suppose you are interested in testing the joint null hypothesis in Equation (7.6) that $\beta_1 = 0$ and $\beta_2 = 0$. Let t_1 be the *t*-statistic for testing the null hypothesis that $\beta_1 = 0$, and let t_2 be the *t*-statistic for testing the null hypothesis that $\beta_2 = 0$. What happens when you use the "one-at-a-time" testing procedure: Reject the joint null hypothesis if either t_1 or t_2 exceeds 1.96 in absolute value?

Because this question involves the two random variables t_1 and t_2, answering it requires characterizing the joint sampling distribution of t_1 and t_2. As mentioned in Section 6.6, in large samples, $\hat{\beta}_1$ and $\hat{\beta}_2$ have a joint normal distribution, so under the joint null hypothesis the t-statistics t_1 and t_2 have a bivariate normal distribution, where each t-statistic has a mean equal to 0 and variance equal to 1.

First, consider the special case in which the t-statistics are uncorrelated and thus are independent in large samples. What is the size of the one-at-a-time testing procedure; that is, what is the probability that you will reject the null hypothesis when it is true? More than 5%! In this special case, we can calculate the rejection probability of this method exactly. The null is *not* rejected only if both $|t_1| \leq 1.96$ and $|t_2| \leq 1.96$. Because the t-statistics are independent, $Pr(|t_1| \leq 1.96 \text{ and } |t_2| \leq 1.96) = Pr(|t_1| \leq 1.96) \times Pr(|t_2| \leq 1.96) = 0.95^2 = 0.9025 = 90.25\%$. So the probability of rejecting the null hypothesis when it is true is $1 - 0.95^2 = 9.75\%$. This one-at-a-time method rejects the null too often because it gives you too many chances: If you fail to reject using the first t-statistic, you get to try again using the second.

If the regressors are correlated, the situation is more complicated. The size of the one-at-a-time procedure depends on the value of the correlation between the regressors. Because the one-at-a-time testing approach has the wrong size—that is, its rejection rate under the null hypothesis does not equal the desired significance level—a new approach is needed.

One approach is to modify the one-at-a-time method so that it uses different critical values that ensure that its size equals its significance level. This method, called the Bonferroni method, is described in Appendix 7.1. The advantage of the Bonferroni method is that it applies very generally. Its disadvantage is that it can have low power: It frequently fails to reject the null hypothesis when, in fact, the alternative hypothesis is true.

Fortunately, there is another approach to testing joint hypotheses that is more powerful, especially when the regressors are highly correlated. That approach is based on the F-statistic.

The *F*-Statistic

The **F-statistic** is used to test a joint hypothesis about regression coefficients. The formulas for the F-statistic are integrated into modern regression software. We first discuss the case of two restrictions then turn to the general case of q restrictions.

The F-statistic with q = 2 restrictions. When the joint null hypothesis has the two restrictions that $\beta_1 = 0$ and $\beta_2 = 0$, the F-statistic combines the two t-statistics t_1 and t_2 using the formula

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2}t_1t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right), \tag{7.9}$$

where $\hat{\rho}_{t_1,t_2}$ is an estimator of the correlation between the two t-statistics.

To understand the F-statistic in Equation (7.9), first suppose we know that the t-statistics are uncorrelated, so we can drop the terms involving $\hat{\rho}_{t_1, t_2}$. If so, Equation (7.9) simplifies, and $F = \frac{1}{2}(t_1^2 + t_2^2)$; that is, the F-statistic is the average of the squared t-statistics. Under the null hypothesis, t_1 and t_2 are independent standard normal random variables (because the t-statistics are uncorrelated by assumption), so under the null hypothesis F has an $F_{2,\infty}$ distribution (Section 2.4). Under the alternative hypothesis that either β_1 is nonzero or β_2 is nonzero (or both), then either t_1^2 or t_2^2 (or both) will be large, leading the test to reject the null hypothesis.

In general, the t-statistics are correlated, and the formula for the F-statistic in Equation (7.9) adjusts for this correlation. This adjustment is made so that under the null hypothesis the F-statistic has an $F_{2,\infty}$ distribution in large samples whether or not the t-statistics are correlated.

The F-statistic with q restrictions. The formula for the heteroskedasticity-robust F-statistic testing the q restrictions of the joint null hypothesis in Equation (7.8) is given in Section 19.3. This formula is incorporated into regression software, making the F-statistic easy to compute in practice.

Under the null hypothesis, the F-statistic has a sampling distribution that, in large samples, is given by the $F_{q,\infty}$ distribution. That is, in large samples, under the null hypothesis

$$\text{the } F\text{-statistic is distributed } F_{q,\infty}. \tag{7.10}$$

Thus the critical values for the F-statistic can be obtained from the tables of the $F_{q,\infty}$ distribution in Appendix Table 4 for the appropriate value of q and the desired significance level.

Computing the heteroskedasticity-robust F-statistic in statistical software. If the F-statistic is computed using the general heteroskedasticity-robust formula, its large-n distribution under the null hypothesis is $F_{q,\infty}$ regardless of whether the errors are homoskedastic or heteroskedastic. As discussed in Section 5.4, for historical reasons, most statistical software computes homoskedasticity-only standard errors by default. Consequently, in some software packages you must select a "robust" option so that the F-statistic is computed using heteroskedasticity-robust standard errors (and, more generally, a heteroskedasticity-robust estimate of the "covariance matrix"). The homoskedasticity-only version of the F-statistic is discussed at the end of this section.

Computing the p-value using the F-statistic. The p-value of the F-statistic can be computed using the large-sample $F_{q,\infty}$ approximation to its distribution. Let F^{act} denote the value of the F-statistic actually computed. Because the F-statistic has a large-sample $F_{q,\infty}$ distribution under the null hypothesis, the p-value is

$$p\text{-value} = \Pr\left[F_{q,\infty} > F^{act}\right]. \tag{7.11}$$

The *p*-value in Equation (7.11) can be evaluated using a table of the $F_{q,\infty}$ distribution (or, alternatively, a table of the χ_q^2 distribution because a χ_q^2-distributed random variable is *q* times an $F_{q,\infty}$-distributed random variable). Alternatively, the *p*-value can be evaluated using a computer because formulas for the cumulative chi-squared and *F* distributions have been incorporated into most modern statistical software.

The overall regression F-statistic. The overall regression *F*-statistic tests the joint hypothesis that *all* the slope coefficients are 0. That is, the null and alternative hypotheses are

$$H_0: \beta_1 = 0, \ \beta_2 = 0, \ldots, \ \beta_k = 0 \text{ vs. } H_1: \beta_j \neq 0, \text{ at least one } j, j = 1, \ldots, k. \quad (7.12)$$

Under this null hypothesis, none of the regressors explains any of the variation in Y_i, although the intercept (which under the null hypothesis is the mean of Y_i) can be nonzero. The null hypothesis in Equation (7.12) is a special case of the general null hypothesis in Equation (7.8), and the overall regression *F*-statistic is the *F*-statistic computed for the null hypothesis in Equation (7.12). In large samples, the overall regression *F*-statistic has an $F_{k,\infty}$ distribution when the null hypothesis is true.

The F-statistic when q = 1. When $q = 1$, the *F*-statistic tests a single restriction. Then the joint null hypothesis reduces to the null hypothesis on a single regression coefficient, and the *F*-statistic is the square of the *t*-statistic.

Application to Test Scores and the Student–Teacher Ratio

We are now able to test the null hypothesis that the coefficients on *both* the student–teacher ratio *and* expenditures per pupil are 0 against the alternative that at least one coefficient is nonzero, controlling for the percentage of English learners in the district.

To test this hypothesis, we need to compute the heteroskedasticity-robust *F*-statistic testing the null hypothesis that $\beta_1 = 0$ and $\beta_2 = 0$ using the regression of *TestScore* on *STR*, *Expn*, and *PctEL* reported in Equation (7.6). This *F*-statistic is 5.43. Under the null hypothesis, in large samples this statistic has an $F_{2,\infty}$ distribution. The 5% critical value of the $F_{2,\infty}$ distribution is 3.00 (Appendix Table 4), and the 1% critical value is 4.61. The value of the *F*-statistic computed from the data, 5.43, exceeds 4.61, so the null hypothesis is rejected at the 1% level. It is very unlikely that we would have drawn a sample that produced an *F*-statistic as large as 5.43 if the null hypothesis really were true (the *p*-value is 0.005). Based on the evidence in Equation (7.6) as summarized in this *F*-statistic, we can reject the taxpayer's hypothesis that *neither* the student–teacher ratio *nor* expenditures per pupil have an effect on test scores (holding constant the percentage of English learners).

The Homoskedasticity-Only *F*-Statistic

One way to restate the question addressed by the *F*-statistic is to ask whether relaxing the q restrictions that constitute the null hypothesis improves the fit of the regression by enough that this improvement is unlikely to be the result merely of random sampling variation if the null hypothesis is true. This restatement suggests that there is a link between the *F*-statistic and the regression R^2: A large *F*-statistic should, it seems, be associated with a substantial increase in the R^2. In fact, if the error u_i is homoskedastic, this intuition has an exact mathematical expression. Specifically, if the error term is homoskedastic, the *F*-statistic can be written in terms of the improvement in the fit of the regression as measured either by the decrease in the sum of squared residuals or by the increase in the regression R^2. The resulting *F*-statistic is referred to as the homoskedasticity-only *F*-statistic because it is valid only if the error term is homoskedastic. In contrast, the heteroskedasticity-robust *F*-statistic computed using the formula in Section 19.3 (and reported above) is valid whether the error term is homoskedastic or heteroskedastic. Despite this significant limitation of the homoskedasticity-only *F*-statistic, its simple formula sheds light on what the *F*-statistic is doing. In addition, the simple formula can be computed using standard regression output, such as might be reported in a table that includes regression R^2's but not *F*-statistics.

The homoskedasticity-only *F*-statistic is computed using a simple formula based on the sum of squared residuals from two regressions. In the first regression, called the **restricted regression**, the null hypothesis is forced to be true. When the null hypothesis is of the type in Equation (7.8), where all the hypothesized values are 0, the restricted regression is the regression in which those coefficients are set to 0; that is, the relevant regressors are excluded from the regression. In the second regression, called the **unrestricted regression**, the alternative hypothesis is allowed to be true. If the sum of squared residuals is sufficiently smaller in the unrestricted than in the restricted regression, then the test rejects the null hypothesis.

The **homoskedasticity-only *F*-statistic** is given by the formula

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n - k_{unrestricted} - 1)}, \tag{7.13}$$

where $SSR_{restricted}$ is the sum of squared residuals from the restricted regression, $SSR_{unrestricted}$ is the sum of squared residuals from the unrestricted regression, q is the number of restrictions under the null hypothesis, and $k_{unrestricted}$ is the number of regressors in the unrestricted regression. An alternative equivalent formula for the homoskedasticity-only *F*-statistic is based on the R^2 of the two regressions:

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}. \tag{7.14}$$

If the errors are homoskedastic, then the difference between the homoskedasticity-only *F*-statistic computed using Equation (7.13) or (7.14) and the heteroskedasticity-robust *F*-statistic vanishes as the sample size n increases. Thus, if the errors are

homoskedastic, the sampling distribution of the homoskedasticity-only F-statistic under the null hypothesis is, in large samples, $F_{q,\infty}$.

These formulas are easy to compute and have an intuitive interpretation in terms of how well the unrestricted and restricted regressions fit the data. Unfortunately, the formulas apply only if the errors are homoskedastic. Because homoskedasticity is a special case that cannot be counted on in applications with economic data—or more generally with data sets typically found in the social sciences—in practice the homoskedasticity-only F-statistic is not a satisfactory substitute for the heteroskedasticity-robust F-statistic.

Using the homoskedasticity-only F-statistic when n is small. If the errors are i.i.d., homoskedastic, and normally distributed, then the homoskedasticity-only F-statistic defined in Equations (7.13) and (7.14) has an $F_{q,n-k_{unrestricted}-1}$ distribution under the null hypothesis (see Section 19.4). Critical values for this distribution, which depend on both q and $n - k_{unrestricted} - 1$, are given in Appendix Table 5. As discussed in Section 2.4, the $F_{q,n-k_{unrestricted}-1}$ distribution converges to the $F_{q,\infty}$ distribution as n increases; for large sample sizes, the differences between the two distributions are negligible. For small samples, however, the two sets of critical values differ.

Application to test scores and the student–teacher ratio. To test the null hypothesis that the population coefficients on *STR* and *Expn* are 0, controlling for *PctEL*, we need to compute the R^2 (or *SSR*) for the restricted and unrestricted regressions. The unrestricted regression has the regressors *STR*, *Expn*, and *PctEL* and is given in Equation (7.6). Its R^2 is 0.4366; that is, $R^2_{unrestricted} = 0.4366$. The restricted regression imposes the joint null hypothesis that the true coefficients on *STR* and *Expn* are 0; that is, under the null hypothesis *STR* and *Expn* do not enter the population regression, although *PctEL* does (the null hypothesis does not restrict the coefficient on *PctEL*). The restricted regression, estimated by OLS, is

$$\widehat{TestScore} = 664.7 - 0.671 \times PctEL, \ R^2 = 0.4149, \qquad (7.15)$$
$$\qquad\qquad\quad (1.0) \quad (0.032)$$

so $R^2_{restricted} = 0.4149$. The number of restrictions is $q = 2$, the number of observations is $n = 420$, and the number of regressors in the unrestricted regression is $k = 3$. The homoskedasticity-only F-statistic, computed using Equation (7.14), is

$$F = \frac{(0.4366 - 0.4149)/2}{(1 - 0.4366)/(420 - 3 - 1)} = 8.01.$$

Because 8.01 exceeds the 1% critical value of 4.61, the hypothesis is rejected at the 1% level using the homoskedasticity-only test.

This example illustrates the advantages and disadvantages of the homoskedasticity-only F-statistic. An advantage is that it can be computed using a calculator. Its main disadvantage is that the values of the homoskedasticity-only and heteroskedasticity-robust F-statistics can be very different: The heteroskedasticity-robust F-statistic

testing this joint hypothesis is 5.43, quite different from the less reliable homoskedasticity-only value of 8.01.

7.3 Testing Single Restrictions Involving Multiple Coefficients

Sometimes economic theory suggests a single restriction that involves two or more regression coefficients. For example, theory might suggest a null hypothesis of the form $\beta_1 = \beta_2$; that is, the effects of the first and second regressors are the same. In this case, the task is to test this null hypothesis against the alternative that the two coefficients differ:

$$H_0: \beta_1 = \beta_2 \text{ vs. } H_1: \beta_1 \neq \beta_2. \tag{7.16}$$

This null hypothesis has a single restriction, so $q = 1$, but that restriction involves multiple coefficients (β_1 and β_2). We need to modify the methods presented so far to test this hypothesis. There are two approaches; which is easier depends on your software.

Approach 1: Test the restriction directly. Some statistical packages have a specialized command designed to test restrictions like Equation (7.16), and the result is an *F*-statistic that, because $q = 1$, has an $F_{1,\infty}$ distribution under the null hypothesis. (Recall from Section 2.4 that the square of a standard normal random variable has an $F_{1,\infty}$ distribution, so the 95% percentile of the $F_{1,\infty}$ distribution is $1.96^2 = 3.84$.)

Approach 2: Transform the regression. If your statistical package cannot test the restriction directly, the hypothesis in Equation (7.16) can be tested using a trick in which the original regression equation is rewritten to turn the restriction in Equation (7.16) into a restriction on a single regression coefficient. To be concrete, suppose there are only two regressors, X_{1i} and X_{2i}, in the regression, so the population regression has the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i. \tag{7.17}$$

Here is the trick: By subtracting and adding $\beta_2 X_{1i}$, we have that $\beta_1 X_{1i} + \beta_2 X_{2i} = \beta_1 X_{1i} - \beta_2 X_{1i} + \beta_2 X_{1i} + \beta_2 X_{2i} = (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) = \gamma_1 X_{1i} + \beta_2 V_i$, where $\gamma_1 = \beta_1 - \beta_2$ and $V_i = X_{1i} + X_{2i}$. Thus the population regression in Equation (7.17) can be rewritten as

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 V_i + u_i. \tag{7.18}$$

Because the coefficient γ_1 in this equation is $\gamma_1 = \beta_1 - \beta_2$, under the null hypothesis in Equation (7.16) $\gamma_1 = 0$, while under the alternative $\gamma_1 \neq 0$. Thus, by turning Equation (7.17) into Equation (7.18), we have turned a restriction on two regression coefficients into a restriction on a single regression coefficient.

Because the restriction now involves the single coefficient γ_1, the null hypothesis in Equation (7.16) can be tested using the t-statistic method of Section 7.1. In practice, this is done by first constructing the new regressor V_i as the sum of the two original regressors, then estimating the regression of Y_i on X_{1i} and V_i. A 95% confidence interval for the difference in the coefficients $\beta_1 - \beta_2$ can be calculated as $\hat{\gamma}_1 \pm 1.96\, SE(\hat{\gamma}_1)$.

This method can be extended to other restrictions on regression equations using the same trick (see Exercise 7.9).

The two methods (approaches 1 and 2) are equivalent in the sense that the F-statistic from the first method equals the square of the t-statistic from the second method.

Extension to $q > 1$.   In general, it is possible to have q restrictions under the null hypothesis in which some or all of these restrictions involve multiple coefficients. The F-statistic of Section 7.2 extends to this type of joint hypothesis. The F-statistic can be computed by either of the two methods just discussed for $q = 1$. Precisely how best to do this in practice depends on the specific regression software being used.

7.4   Confidence Sets for Multiple Coefficients

This section explains how to construct a confidence set for two or more regression coefficients. The method is conceptually similar to the method in Section 7.1 for constructing a confidence set for a single coefficient using the t-statistic except that the confidence set for multiple coefficients is based on the F-statistic.

A **95% confidence set** for two or more coefficients is a set that contains the true population values of these coefficients in 95% of randomly drawn samples. Thus a confidence set is the generalization to two or more coefficients of a confidence interval for a single coefficient.

Recall that a 95% confidence interval is computed by finding the set of values of the coefficients that are not rejected using a t-statistic at the 5% significance level. This approach can be extended to the case of multiple coefficients. To make this concrete, suppose you are interested in constructing a confidence set for two coefficients, β_1 and β_2. Section 7.2 showed how to use the F-statistic to test a joint null hypothesis that $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$. Suppose you were to test every possible value of $\beta_{1,0}$ and $\beta_{2,0}$ at the 5% level. For each pair of candidates $(\beta_{1,0}, \beta_{2,0})$, you compute the F-statistic and reject it if it exceeds the 5% critical value of 3.00. Because the test has a 5% significance level, the true population values of β_1 and β_2 will not be rejected in 95% of all samples. Thus the set of values not rejected at the 5% level by this F-statistic constitutes a 95% confidence set for β_1 and β_2.

Although this method of trying all possible values of $\beta_{1,0}$ and $\beta_{2,0}$ works in theory, in practice it is much simpler to use an explicit formula for the confidence set. This formula for the confidence set for an arbitrary number of coefficients is obtained

FIGURE 7.1 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)

The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (β_2) is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the *F*-statistic at the 5% significance level. The point $(\beta_1, \beta_2) = (0, 0)$ is not contained in the confidence set, so the null hypothesis $H_0: \beta_1 = 0$ and $\beta_2 = 0$ is rejected at the 5% significance level.
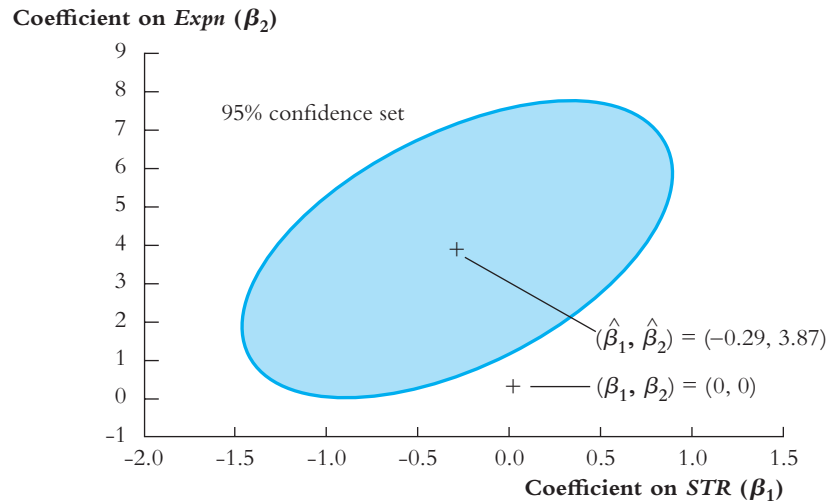
using the formula for the *F*-statistic given in Section 19.3. When there are two coefficients, the resulting confidence sets are ellipses.

As an illustration, Figure 7.1 shows a 95% confidence set (confidence ellipse) for the coefficients on the student–teacher ratio and expenditures per pupil, holding constant the percentage of English learners, based on the estimated regression in Equation (7.6). This ellipse does not include the point (0, 0). This means that the null hypothesis that these two coefficients are both 0 is rejected using the *F*-statistic at the 5% significance level, which we already knew from Section 7.2. The confidence ellipse is a fat sausage with the long part of the sausage oriented in the lower-left/ upper-right direction. The reason for this orientation is that the estimated correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ is positive, which in turn arises because the correlation between the regressors *STR* and *Expn* is negative (schools that spend more per pupil tend to have fewer students per teacher).

7.5 Model Specification for Multiple Regression

When estimating a causal effect, the job of determining which variables to include in multiple regression—that is, the problem of choosing a regression specification—can be quite challenging, and no single rule applies in all situations. But do not despair, because some useful guidelines are available. The starting point for choosing a regression specification is thinking through the possible sources of omitted variable bias. It is important to rely on your expert knowledge of the empirical problem and to focus on obtaining an unbiased estimate of the causal effect of interest; do not rely primarily on purely statistical measures of fit such as the R^2 or $\overline{R}^2$.

Model Specification and Choosing Control Variables

Multiple regression makes it possible to control for factors that could lead to omitted variable bias in the estimate of the effect of interest. But how does one determine the "right" set of control variables?

At a general level, this question is answered by the conditional mean independence condition of Key Concept 6.5. That is, to eliminate omitted variables bias, a set of control variables must satisfy $E(u_i|X_i,W_i) = E(u_i|W_i)$, where X_i denotes the variable or variables of interest and W_i denotes one or more control variables. This condition requires that, among observations with the same values of the control variables, the variable of interest is randomly assigned or as-if randomly assigned in the sense that the mean of u no longer depends on X. If this condition fails, then there remain omitted determinants of Y that are correlated with X, even after holding W constant, and the result is omitted variable bias.

In practice, determining which control variables to include requires thinking through the application and using judgment. For example, economic conditions could vary substantially across school districts with the same percentage of English learners. Because the budget of a school district depends in part on the affluence of the district, more affluent districts would be expected to have lower class sizes, even among districts with the same percentage of English learners. Moreover, more affluent families tend to have more access to outside learning opportunities. If so, the affluence of the district satisfies the two conditions for omitted variable bias in Key Concept 6.1, even after controlling for the percentage of English learners. This logic leads to including one or more additional control variables in the test score regressions, where the additional control variables measure economic conditions of the district.

Our approach to the challenge of choosing control variables is twofold. First, a core or base set of regressors should be chosen using a combination of expert judgment, economic theory, and knowledge of how the data were collected; the regression using this base set of regressors is sometimes referred to as a **base specification**. This base specification should contain the variables of primary interest and the control variables suggested by expert judgment and economic theory. Expert judgment and economic theory are rarely decisive, however, and often the variables suggested by economic theory are not the ones on which you have data. Therefore the next step is to develop a list of candidate **alternative specifications**—that is, alternative sets of regressors. If the estimates of the coefficients of interest are numerically similar across the alternative specifications, then this provides evidence that the estimates from your base specification are reliable. If, on the other hand, the estimates of the coefficients of interest change substantially across specifications, this often provides evidence that the original specification had omitted variable bias and heightens the concern that so might your alternative specifications. We elaborate on this approach to model specification in Section 9.2 after studying some additional tools for specifying regressions.

Interpreting the R^2 and the Adjusted R^2 in Practice

An R^2 or an $\overline{R}^2$ near 1 means that the regressors are good at predicting the values of the dependent variable in the sample, and an R^2 or an $\overline{R}^2$ near 0 means that they are not. This makes these statistics useful summaries of the predictive ability of the regression. However, it is easy to read more into them than they deserve.

There are four potential pitfalls to guard against when using the R^2 or $\overline{R}^2$:

1. **An increase in the R^2 or $\overline{R}^2$ does not necessarily mean that an added variable is statistically significant.** The R^2 increases whenever you add a regressor, whether or not it is statistically significant. The $\overline{R}^2$ does not always increase, but if it does, this does not necessarily mean that the coefficient on that added regressor is statistically significant. To ascertain whether an added variable is statistically significant, you need to perform a hypothesis test using the t-statistic.

2. **A high R^2 or $\overline{R}^2$ does not mean that the regressors are a true cause of the dependent variable.** Imagine regressing test scores against parking lot area per pupil. Parking lot area is correlated with the student–teacher ratio, with whether the school is in a suburb or a city, and possibly with district income—all things that are correlated with test scores. Thus the regression of test scores on parking lot area per pupil could have a high R^2 and $\overline{R}^2$, but the relationship is not causal (try telling the superintendent that the way to increase test scores is to increase parking space!).

3. **A high R^2 or $\overline{R}^2$ does not mean that there is no omitted variable bias.** Recall the discussion of Section 6.1, which concerned omitted variable bias in the regression of test scores on the student–teacher ratio. The R^2 of the regression was not mentioned because it played no logical role in this discussion. Omitted variable bias can occur in regressions with a low R^2, a moderate R^2, or a high R^2. Conversely, a low R^2 does not imply that there necessarily is omitted variable bias.

4. **A high R^2 or $\overline{R}^2$ does not necessarily mean that you have the most appropriate set of regressors, nor does a low R^2 or $\overline{R}^2$ necessarily mean that you have an inappropriate set of regressors.** The question of what constitutes the right set of regressors in multiple regression is difficult, and we return to it throughout this textbook. Decisions about the regressors must weigh issues of omitted variable bias, data availability, data quality, and, most importantly, economic theory and the nature of the substantive questions being addressed. None of these questions can be answered simply by having a high (or low) regression R^2 or $\overline{R}^2$.

These points are summarized in Key Concept 7.3.

7.6 Analysis of the Test Score Data Set

This section presents an analysis of the effect on test scores of the student–teacher ratio using the California data set. This analysis illustrates how multiple regression analysis can be used to mitigate omitted variable bias. It also shows how to use a table to summarize regression results.

R^2 and $\overline{R}^2$: What They Tell You—and What They Don't

The R^2 and $\overline{R}^2$ tell you whether the regressors are good at predicting, or "explaining," the values of the dependent variable in the sample of data on hand. If the R^2 (or $\overline{R}^2$) is nearly 1, then the regressors produce good predictions of the dependent variable in that sample in the sense that the variance of the OLS residual is small compared to the variance of the dependent variable. If the R^2 (or $\overline{R}^2$) is nearly 0, the opposite is true.

   The R^2 and $\overline{R}^2$ do NOT tell you whether

1. An included variable is statistically significant,

2. The regressors are a true cause of the dependent variable,

3. There is omitted variable bias, or

4. You have chosen the most appropriate set of regressors.

Discussion of the base and alternative specifications. This analysis focuses on estimating the effect on test scores of a change in the student–teacher ratio, controlling for factors that otherwise could lead to omitted variable bias. Many factors potentially affect the average test score in a district. Some of these factors are correlated with the student–teacher ratio, so omitting them from the regression results in omitted variable bias. Because these factors, such as outside learning opportunities, are not directly measured, we include control variables that are correlated with these omitted factors. If the control variables are adequate in the sense that the conditional mean independence assumption holds, then the coefficient on the student–teacher ratio is the effect of a change in the student–teacher ratio, holding constant these other factors. Said differently, our aim is to include control variables such that, once they are held constant, the student-teacher ratio is as-if randomly assigned.

    Here we consider three variables that control for background characteristics of the students that could affect test scores: the fraction of students who are still learning English, the percentage of students who are eligible to receive a subsidized or free lunch at school, and a new variable, the percentage of students in the district whose families qualify for a California income assistance program. Eligibility for this income assistance program depends in part on family income, with a higher (stricter) threshold than the subsidized lunch program. The final two variables thus are different measures of the fraction of economically disadvantaged children in the district (their correlation coefficient is 0.74). Theory and expert judgment do not tell us which of these two variables to use to control for determinants of test scores related to economic background. For our base specification, we use the percentage eligible

FIGURE 7.2 Scatterplots of Test Scores vs. Three Student Characteristics

(a) Percentage of English learners

(b) Percentage eligible for subsidized lunch

(c) Percentage qualifying for income assistance
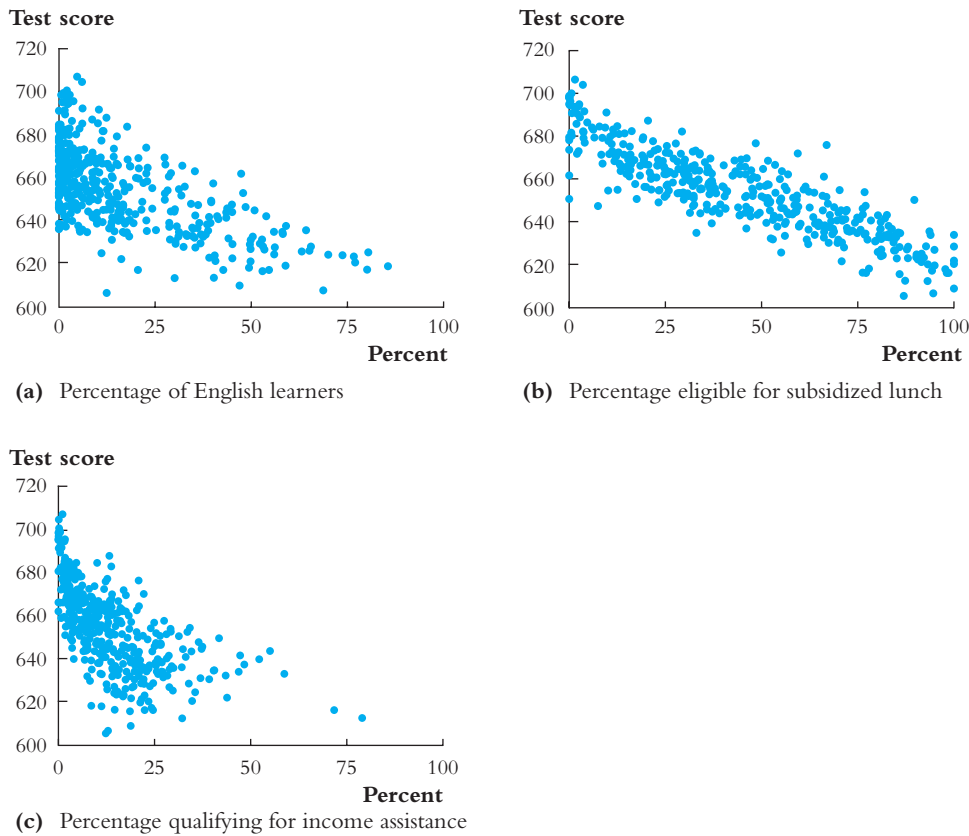
The scatterplots show a negative relationship between test scores and (a) the percentage of English learners (correlation = −0.64), (b) the percentage of students eligible for a subsidized lunch (correlation = −0.87); and (c) the percentage of students qualifying for income assistance (correlation = −0.63).

for a subsidized lunch, but we also consider an alternative specification that uses the fraction eligible for the income assistance program.

Scatterplots of tests scores and these variables are presented in Figure 7.2. Each of these variables exhibits a negative correlation with test scores. The correlation between test scores and the percentage of English learners is −0.64, between test scores and the percentage eligible for a subsidized lunch is −0.87, and between test scores and the percentage qualifying for income assistance is −0.63.

What scale should we use for the regressors? A practical question that arises in regression analysis is what scale you should use for the regressors. In Figure 7.2, the units of the variables are percentages, so the maximum possible range of the data is 0 to 100. Alternatively, we could have defined these variables to be a *decimal fraction*

rather than a percentage; for example, *PctEL* could be replaced by the *fraction* of English learners, $FracEL(=PctEL/100)$, which would range between 0 and 1 instead of between 0 and 100. More generally, in regression analysis some decision usually needs to be made about the scale of both the dependent and the independent variables. How, then, should you choose the scale, or units, of the variables?

The general answer to the question of choosing the scale of the variables is to make the regression results easy to read and to interpret. In the test score application, the natural unit for the dependent variable is the score of the test itself. In the regression of *TestScore* on *STR* and *PctEL* reported in Equation (7.5), the coefficient on *PctEL* is −0.650. If instead the regressor had been *FracEL*, the regression would have had an identical R^2 and *SER*; however, the coefficient on *FracEL* would have been −65.0. In the specification with *PctEL*, the coefficient is the predicted change in test scores for a 1-percentage-point increase in English learners, holding *STR* constant; in the specification with *FracEL*, the coefficient is the predicted change in test scores for an increase by 1 in the fraction of English learners—that is, for a 100-percentage-point-increase—holding *STR* constant. Although these two specifications are mathematically equivalent, for the purposes of interpretation the one with *PctEL* seems, to us, more natural.

Another consideration when deciding on a scale is to choose the units of the regressors so that the resulting regression coefficients are easy to read. For example, if a regressor is measured in dollars and has a coefficient of 0.00000356, it is easier to read if the regressor is converted to millions of dollars and the coefficient 3.56 is reported.

Tabular presentation of result.   We are now faced with a communication problem. What is the best way to show the results from several multiple regressions that contain different subsets of the possible regressors? So far, we have presented regression results by writing out the estimated regression equations, as in Equations (7.6) and (7.19). This works well when there are only a few regressors and only a few equations, but with more regressors and equations, this method of presentation can be confusing. A better way to communicate the results of several regressions is in a table.

Table 7.1 summarizes the results of regressions of the test score on various sets of regressors. Each column presents a separate regression. Each regression has the same dependent variable, test score. The first row reports statistics that provide information about the causal effect of interest, the effect of the student–teacher ratio on test scores. The first entry is the OLS estimate, below which is its standard error (in parentheses). Below the standard error in brackets is a 95% confidence interval for the population coefficient. Although a reader could take out his or her calculator and compute the confidence interval from the estimate and its standard error, doing so is inconvenient, so the table provides this information for the reader. A reader interested in testing the null hypothesis that the coefficient takes on some particular value, for example 0, at the 5% significance level can do so by checking whether that value is included in the 95% confidence interval.

| TABLE 7.1 | Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts |

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28	−1.10	−1.00	−1.31	−1.01
	(0.52)	(0.43)	(0.27)	(0.34)	(0.27)
	[−3.30, −1.26]	[−1.95, −0.25]	[−1.53, −0.47]	[−1.97, −0.64]	[−1.54, −0.49]
Control variables					
Percentage English learners (X_2)		−0.650	−0.122	−0.488	−0.130
		(0.031)	(0.033)	(0.030)	(0.036)
Percentage eligible for subsidized lunch (X_3)			−0.547		−0.529
			(0.024)		(0.038)
Percentage qualifying for income assistance (X_4)				−0.790	0.048
				(0.068)	(0.059)
Intercept	698.9	686.0	700.2	698.0	700.4
	(10.4)	(8.7)	(5.6)	(6.9)	(5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\overline{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. For the variable of interest, the student–teacher ratio, the 95% confidence interval is given in brackets below the standard error.

The remaining variables are control variables and the constant term (intercept); for these, only the OLS estimate and its standard error are reported. Because the coefficients on the control variables do not, in general, have a causal interpretation, these coefficient estimates are often of limited independent interest, so no confidence interval is reported, although a reader who wants a confidence interval for one of those coefficients can compute it using the information provided. In cases in which there are many control variables, as there are in regressions later in this text, sometimes a table will report no information at all about their coefficients or standard errors and will simply list the included control variables. Similarly, the value of the intercept often is of limited interest, so it, too, might not be reported.

The final three rows contain summary statistics for the regression (the standard error of the regression, *SER*, and the $\overline{R}^2$) and the sample size (which is the same for all of the regressions, 420 observations).

All the information that we have presented so far in equation format appears in this table. For example, consider the regression of the test score against the student–teacher ratio, with no control variables. In equation form, this regression is

$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \overline{R}^2 = 0.049, SER = 18.58, n = 420. \quad (7.21)$$
$$(10.4) \quad (0.52)$$

All this information appears in column (1) of Table 7.1. The estimated coefficient on the student–teacher ratio (-2.28) appears in the first row of numerical entries, and its standard error (0.52) appears in parentheses just below the estimated coefficient. The table augments the information in Equation (7.21) by reporting the 95% confidence interval. The intercept (698.9) and its standard error (10.4) are given in the row labeled "Intercept." (Sometimes you will see this row labeled "Constant" because, as discussed in Section 6.2, the intercept can be viewed as the coefficient on a regressor that is always equal to 1.) Similarly, the $\overline{R}^2$ (0.049), the SER (18.58), and the sample size n (420) appear in the final rows. The blank entries in the rows of the other regressors indicate that those regressors are not included in this regression.

Although the table does not report t-statistics, they can be computed from the information provided; for example, the t-statistic testing the hypothesis that the coefficient on the student–teacher ratio in column (1) is 0 is $-2.28/0.52 = -4.38$. This hypothesis is rejected at the 1% level.

Regressions that include the control variables measuring student characteristics are reported in columns (2) through (5). Column (2), which reports the regression of test scores on the student–teacher ratio and on the percentage of English learners, was previously stated as Equation (7.5).

Column (3) presents the base specification, in which the regressors are the student–teacher ratio and two control variables, the percentage of English learners and the percentage of students eligible for a subsidized lunch.

Columns (4) and (5) present alternative specifications that examine the effect of changes in the way the economic background of the students is measured. In column (4), the percentage of students qualifying for income assistance is included as a regressor, and in column (5), both of the economic background variables are included.

Discussion of empirical results. These results suggest three conclusions:

1. Controlling for these student characteristics cuts the estimated effect of the student–teacher ratio on test scores approximately in half. This estimated effect is not very sensitive to which specific control variables are included in the regression. In all cases, the hypothesis that the coefficient on the student–teacher ratio is 0 can be rejected at the 5% level. In the four specifications with control variables, regressions (2) through (5), reducing the student–teacher ratio by one student per teacher is estimated to increase average test scores by approximately 1 point, holding constant student characteristics.

2. The student characteristic variables are potent predictors of test scores. The student–teacher ratio alone explains only a small fraction of the variation in test scores: The $\overline{R}^2$ in column (1) is 0.049. The $\overline{R}^2$ jumps, however, when the student characteristic variables are added. For example, the $\overline{R}^2$ in the base specification, regression (3), is 0.773. The signs of the coefficients on the student demographic variables are consistent with the patterns seen in Figure 7.2: Districts with many English learners and districts with many poor children have lower test scores.

3. In contrast to the other two control variables, the percentage qualifying for income assistance appears to be redundant. As reported in regression (5), adding it to regression (3) has a negligible effect on the estimated coefficient on the student-teacher ratio or its standard error.

7.7 Conclusion

Chapter 6 began with a concern: In the regression of test scores against the student–teacher ratio, omitted student characteristics that influence test scores might be correlated with the student–teacher ratio in the district, and, if so, the student–teacher ratio in the district would pick up the effect on test scores of these omitted student characteristics. Thus the OLS estimator would have omitted variable bias. To mitigate this potential omitted variable bias, we augmented the regression by including variables that control for various student characteristics (the percentage of English learners and two measures of student economic background). Doing so cuts the estimated effect of a unit change in the student–teacher ratio in half, although it remains possible to reject the null hypothesis that the population effect on test scores, holding these control variables constant, is 0 at the 5% significance level. Because they eliminate omitted variable bias arising from these student characteristics, these multiple regression estimates, hypothesis tests, and confidence intervals are much more useful for advising the superintendent than are the single-regressor estimates of Chapters 4 and 5.

The analysis in this and the preceding chapter has presumed that the population regression function is linear in the regressors—that is, that the conditional expectation of Y_i given the regressors is a straight line. There is, however, no particular reason to think this is so. In fact, the effect of reducing the student–teacher ratio might be quite different in districts with large classes than in districts that already have small classes. If so, the population regression line is not linear in the X's but rather is a nonlinear function of the X's. To extend our analysis to regression functions that are nonlinear in the X's, however, we need the tools developed in the next chapter.

Summary

1. Hypothesis tests and confidence intervals for a single regression coefficient are carried out using essentially the same procedures used in the one-variable linear regression model of Chapter 5. For example, a 95% confidence interval for β_1 is given by $\hat{\beta}_1 \pm 1.96\, SE(\hat{\beta}_1)$.
2. Hypotheses involving more than one restriction on the coefficients are called joint hypotheses. Joint hypotheses can be tested using an F-statistic.
3. Regression specification proceeds by first determining a base specification chosen to address concern about omitted variable bias. The base specification can be modified by including additional regressors that control for other potential sources of omitted variable bias. Simply choosing the specification with the highest R^2 can lead to regression models that do not estimate the causal effect of interest.

Key Terms

restrictions (210)

joint hypothesis (210)

F-statistic (211)

restricted regression (214)

unrestricted regression (214)

homoskedasticity-only *F*-statistic (214)

95% confidence set (217)

base specification (219)

alternative specifications (219)

Bonferroni test (233)

MyLab Economics Can Help You Get a Better Grade

MyLab Economics If your exam were tomorrow, would you be ready? For each chapter, **MyLab Economics** Practice Tests and Study Plan help you prepare for your exams. You can also find the Exercises and all Review the Concepts Questions available now in **MyLab Economics**. To see how it works, turn to the **MyLab Economics** spread on the inside front cover of this text and then go to **www.pearson.com/mylab/economics**.

For additional Empirical Exercises and Data Sets, log on to the Companion Website at **www.pearsonhighered.com/stock_watson.**

Review the Concepts

7.1 Explain how you would test the null hypothesis that $\beta_1 = 0$ in the multiple regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Explain how you would test the null hypothesis that $\beta_2 = 0$. Explain how you would test the joint hypothesis that $\beta_1 = 0$ and $\beta_2 = 0$. Why isn't the result of the joint test implied by the results of the first two tests?

7.2 Provide an example of a regression that arguably would have a high value of R^2 but would produce biased and inconsistent estimators of a causal effect. Explain why the R^2 is likely to be high. Explain why the OLS estimators would be biased and inconsistent.

7.3 What is a control variable, and how does it differ from a variable of interest? Looking at Table 7.1, for what factors are the control variables controlling? Do coefficients on control variables measure causal effects? Explain.

Exercises

The first six exercises refer to the table of estimated regressions on page 228, computed using data for 2015 from the Current Population Survey. The data set consists of information on 7178 full-time, full-year workers. The highest educational achievement for each worker was either a high school diploma or a bachelor's degree. The workers' ages ranged from 25 to 34 years. The data set also contains information on the region of the country where the person lived, marital status, and number of children. For the purposes of these exercises, let

AHE = average hourly earnings

$College$ = binary variable (1 if college, 0 if high school)

$Female$ = binary variable (1 if female, 0 if male)

Age = (in years)

$Northeast$ = binary variable (1 if Region = Northeast, 0 otherwise)

$Midwest$ = binary variable (1 if Region = Midwest, 0 otherwise)

$South$ = binary variable (1 if Region = South, 0 otherwise)

$West$ = binary variable (1 if Region = West, 0 otherwise)

Results of Regressions of Average Hourly Earnings on Sex and Education Binary Variables and Other Characteristics Using 2015 Data from the Current Population Survey

Dependent variable: average hourly earnings (*AHE*).

Regressor	(1)	(2)	(3)
College (X_1)	10.47 (0.29)	10.44 (0.29)	10.42 (0.29)
Female (X_2)	−4.69 (0.29)	−4.56 (0.29)	−4.57 (0.29)
Age (X_3)		0.61 (0.05)	0.61 (0.05)
Northeast (X_4)			0.74 (0.47)
Midwest (X_5)			−1.54 (0.40)
South (X_6)			−0.44 (0.37)
Intercept	18.15 (0.19)	0.11 (1.46)	0.33 (1.47)
Summary Statistics and Joint Tests			
F-statistic testing regional effects = 0			9.32
SER	12.15	12.03	12.01
R^2	0.165	0.182	0.185
n	7178	7178	7178

7.1 For each of the three regressions, compute the *p*-value of the test of the hypothesis that the coefficient on *College* is 0 against the 2-sided alternative that it is nonzero.

7.2   Using the regression results in column (1):

    a. Is the college–high school earnings difference estimated from this regression statistically significant at the 5% level? Construct a 95% confidence interval of the difference.

    b. Is the male–female earnings difference estimated from this regression statistically significant at the 5% level? Construct a 95% confidence interval for the difference.

7.3   Using the regression results in column (2):

    a. Is age an important determinant of earnings? Use an appropriate statistical test and/or confidence interval to explain your answer.

    b. Sally is a 29-year-old female college graduate. Betsy is a 34-year-old female college graduate. Construct a 95% confidence interval for the expected difference between their earnings.

7.4   Using the regression results in column (3):

    a. Do there appear to be important regional differences? Use an appropriate hypothesis test to explain your answer.

    b. Juanita is a 28-year-old female college graduate from the South. Molly is a 28-year-old female college graduate from the West. Jennifer is a 28-year-old female college graduate from the Midwest.

       i. Construct a 95% confidence interval for the difference in expected earnings between Juanita and Molly.

      ii. Explain how you would construct a 95% confidence interval for the difference in expected earnings between Juanita and Jennifer. (*Hint:* What would happen if you included *West* and excluded *Midwest* from the regression?)

7.5   The regression shown in column (2) was estimated again, this time using data from 1992 (4000 observations selected at random from the March 1993 Current Population Survey, converted into 2015 dollars using the Consumer Price Index). The results are

$$\widehat{AHE} = 1.30 + 8.94\ College - 4.38\ Female + 0.67Age,\ SER = 9.88,\ \overline{R}^2 = 0.21.$$
$$\phantom{\widehat{AHE} = }(1.65)\ \ (0.34)\ \ \ \ \ \ \ \ \ (0.30)\ \ \ \ \ \ \ \ \ \ (0.05)$$

Comparing this regression to the regression for 2015 shown in column (2), was there a statistically significant change in the coefficient on *College*?

7.6   Evaluate the following statement: "In all of the regressions, the coefficient on *Female* is negative, large, and statistically significant. This provides strong statistical evidence of sex discrimination in the U.S. labor market."

7.7 Question 6.5 reported the following regression (where standard errors have been added):

$$\widehat{Price} = 119.2 + 0.485BDR + 23.4Bath + 0.156Hsize + 0.002Lsize$$
$$\quad\quad (23.9) \quad (2.61) \quad\quad\quad (8.94) \quad\quad (0.011) \quad\quad (0.00048)$$

$$+ 0.090Age - 48.8Poor, \; \overline{R}^2 = 0.72, \; SER = 41.5$$
$$(0.311) \quad\quad (10.5)$$

a. Is the coefficient on *BDR* statistically significantly different from 0?

b. Typically, five-bedroom houses sell for much more than two-bedroom houses. Is this consistent with your answer to (a) and with the regression more generally?

c. A homeowner purchases 2000 square feet from an adjacent lot. Construct a 99% confident interval for the change in the value of her house.

d. Lot size is measured in square feet. Do you think that another scale might be more appropriate? Why or why not?

e. The *F*-statistic for omitting *BDR* and *Age* from the regression is $F = 0.08$. Are the coefficients on *BDR* and *Age* statistically different from 0 at the 10% level?

7.8 Referring to Table 7.1 in the text:

a. Calculate the R^2 for each of the regressions.

b. Calculate the homoskedasticity-only *F*-statistic for testing $\beta_3 = \beta_4 = 0$ in the regression shown in column (5). Is the statistic significant at the 5% level?

c. Test $\beta_3 = \beta_4 = 0$ in the regression shown in column (5) using the Bonferroni test discussed in Appendix 7.1.

d. Construct a 99% confidence interval for β_1 for the regression in column (5).

7.9 Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Use approach 2 from Section 7.3 to transform the regression so that you can use a *t*-statistic to test

a. $\beta_1 = \beta_2$.

b. $\beta_1 + 2\beta_2 = 0$.

c. $\beta_1 + \beta_2 = 1$. (*Hint:* You must redefine the dependent variable in the regression.)

7.10 Equations (7.13) and (7.14) show two formulas for the homoskedasticity-only *F*-statistic. Show that the two formulas are equivalent.

Empirical Exercises

E7.1 Use the **Birthweight_Smoking** data set introduced in Empirical Exercise E5.3 to answer the following questions. To begin, run three regressions:

(1) *Birthweight* on *Smoker*

(2) *Birthweight* on *Smoker*, *Alcohol*, and *Nprevist*

(3) *Birthweight* on *Smoker*, *Alcohol*, *Nprevist*, and *Unmarried*

a. What is the value of the estimated effect of smoking on birth weight in each of the regressions?

b. Construct a 95% confidence interval for the effect of smoking on birth weight, using each of the regressions.

c. Does the coefficient on *Smoker* in regression (1) suffer from omitted variable bias? Explain.

d. Does the coefficient on *Smoker* in regression (2) suffer from omitted variable bias? Explain.

e. Consider the coefficient on *Unmarried* in regression (3).

   i. Construct a 95% confidence interval for the coefficient.

   ii. Is the coefficient statistically significant? Explain.

   iii. Is the magnitude of the coefficient large? Explain.

   iv. A family advocacy group notes that the large coefficient suggests that public policies that encourage marriage will lead, on average, to healthier babies. Do you agree? (*Hint:* Review the discussion of control variables in Section 6.8. Discuss some of the various factors that *Unmarried* may be controlling for and how this affects the interpretation of its coefficient.)

f. Consider the various other control variables in the data set. Which do you think should be included in the regression? Using a table like Table 7.1, examine the robustness of the confidence interval you constructed in (b). What is a reasonable 95% confidence interval for the effect of smoking on birth weight?

E7.2 In the empirical exercises on earning and height in Chapters 4 and 5, you estimated a relatively large and statistically significant effect of a worker's height on his or her earnings. One explanation for this result is omitted variable bias: Height is correlated with an omitted factor that affects earnings. For example, Case and Paxson (2008) suggest that cognitive ability (or intelligence) is the omitted factor. The mechanism they describe is straightforward: Poor nutrition and other harmful environmental factors in utero and in early childhood have, on average, deleterious effects on both cognitive and physical development. Cognitive ability affects earnings later in life and thus is an omitted variable in the regression.

a. Suppose that the mechanism described above is correct. Explain how this leads to omitted variable bias in the OLS regression of *Earnings* on *Height*. Does the bias lead the estimated slope to be too large or too small? [*Hint:* Review Equation (6.1).]

If the mechanism described above is correct, the estimated effect of height on earnings should disappear if a variable measuring cognitive ability is included in the regression. Unfortunately, there isn't a direct measure of cognitive ability in the data set, but the data set does include years of education for each individual. Because students with higher cognitive ability are more likely to attend school longer, years of education might serve as a control variable for cognitive ability; in this case, including education in the regression will eliminate, or at least attenuate, the omitted variable bias problem.

Use the years of education variable (*educ*) to construct four indicator variables for whether a worker has less than a high school diploma (*LT_HS* = 1 if *educ* < 12, 0 otherwise), a high school diploma (*HS* = 1 if *educ* = 12, 0 otherwise), some college (*Some_Col* = 1 if 12 < *educ* < 16, 0 otherwise), or a bachelor's degree or higher (*College* = 1 if *educ* ≥ 16, 0 otherwise).

b. Focusing first on women only, run a regression of (1) *Earnings* on *Height* and (2) *Earnings* on *Height*, including *LT_HS*, *HS*, and *Some_Col* as control variables.

   i. Compare the estimated coefficient on *Height* in regressions (1) and (2). Is there a large change in the coefficient? Has it changed in a way consistent with the cognitive ability explanation? Explain.

   ii. The regression omits the control variable *College*. Why?

   iii. Test the joint null hypothesis that the coefficients on the education variables are equal to 0.

   iv. Discuss the values of the estimated coefficients on *LT_HS*, *HS*, and *Some_Col*. (Each of the estimated coefficients is negative, and the coefficient on *LT_HS* is more negative than the coefficient on *HS*, which in turn is more negative than the coefficient on *Some_Col*. Why? What do the coefficients measure?)

c. Repeat (b), using data for men.

APPENDIX

7.1 The Bonferroni Test of a Joint Hypothesis

The method of Section 7.2 is the preferred way to test joint hypotheses in multiple regression. However, if the author of a study presents regression results but did not test a joint restriction in which you are interested and if you do not have the original data, then you will not be able to compute the F-statistic as in Section 7.2. This appendix describes a way to test joint hypotheses that can be used when you have only a table of regression results. This method is an application of a very general testing approach based on Bonferroni's inequality.

The Bonferroni test is a test of a joint hypothesis based on the t-statistics for the individual hypotheses; that is, the Bonferroni test is the one-at-a-time t-statistic test of Section 7.2 done properly. The **Bonferroni test** of the joint null hypothesis $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$, based on the critical value $c > 0$, uses the following rule:

$$\text{Accept if } |t_1| \leq c \text{ and if } |t_2| \leq c; \text{ otherwise, reject} \tag{7.22}$$
$$\text{(Bonferroni one-at-a-time } t\text{-statistic test)}$$

where t_1 and t_2 are the t-statistics that test the restrictions on β_1 and β_2, respectively.

The trick is to choose the critical value c in such a way that the probability that the one-at-a-time test rejects when the null hypothesis is true is no more than the desired significance level—say, 5%. This is done by using Bonferroni's inequality to choose the critical value c to allow both for the fact that two restrictions are being tested and for any possible correlation between t_1 and t_2.

Bonferroni's Inequality

Bonferroni's inequality is a basic result of probability theory. Let A and B be events. Let $A \cap B$ be the event "both A and B" (the intersection of A and B), and let $A \cup B$ be the event "A or B or both" (the union of A and B). Then $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$. Because $\Pr(A \cap B) \geq 0$, it follows that $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$.[1] Now let A be the event that $|t_1| > c$ and B be the event that $|t_2| > c$. Then the inequality $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$ yields

$$\Pr(|t_1| > c \text{ or } |t_2| > c \text{ or both}) \leq \Pr(|t_1| > c) + \Pr(|t_2| > c). \tag{7.23}$$

Bonferroni Tests

Because the event "$|t_1| > c$ or $|t_2| > c$ or both" is the rejection region of the one-at-a-time test, Equation (7.23) leads to a valid critical value for the one-at-a-time test. Under the null hypothesis in large samples, $\Pr(|t_1| > c) = \Pr(|t_2| > c) = \Pr(|Z| > c)$. Thus Equation (7.23) implies that in large samples the probability that the one-at-a-time test rejects under the null is

$$\Pr_{H_0}(\text{one-at-a-time test rejects}) \leq 2\Pr(|Z| > c). \tag{7.24}$$

The inequality in Equation (7.24) provides a way to choose a critical value c so that the probability of the rejection under the null hypothesis equals the desired significance level. The Bonferroni approach can be extended to more than two coefficients; if there are q restrictions under the null, the factor of 2 on the right-hand side in Equation (7.24) is replaced by q.

---

[1]This inequality can be used to derive other interesting inequalities. For example, it implies that $1 - \Pr(A \cup B) \geq 1 - [\Pr(A) + \Pr(B)]$. Let A^c and B^c be the complements of A and B-that is, the events "not A" and "not B." Because the complement of $A \cup B$ is $A^c \cap B^c$, $1 - \Pr(A \cup B) = \Pr(A^c \cap B^c)$, which yields Bonferroni's inequality, $\Pr(A^c \cap B^c) \geq 1 - [\Pr(A) + \Pr(B)]$.

Table 7.2 presents critical values c for the one-at-a-time Bonferroni test for various significance levels and $q = 2, 3,$ and 4. For example, suppose the desired significance level is 5% and $q = 2$. According to Table 7.2, the critical value c is 2.241. This critical value is the 1.25 percentile of the standard normal distribution, so $\Pr(|Z| > 2.241) = 2.5\%$. Thus Equation (7.24) tells us that in large samples the one-at-a-time test in Equation (7.22) will reject at most 5% of the time under the null hypothesis.

TABLE 7.2	**Bonferroni Critical Values c for the One-at-a-Time t-Statistic Test of a Joint Hypothesis**		
	Significance Level		
Number of Restrictions (q)	**10%**	**5%**	**1%**
2	1.960	2.241	2.807
3	2.128	2.394	2.935
4	2.241	2.498	3.023

The critical values in Table 7.2 are larger than the critical values for testing a single restriction. For example, with $q = 2$, the one-at-a-time test rejects if at least one t-statistic exceeds 2.241 in absolute value. This critical value is greater than 1.96 because it properly corrects for the fact that, by looking at two t-statistics, you get a second chance to reject the joint null hypothesis, as discussed in Section 7.2.

If the individual t-statistics are based on heteroskedasticity-robust standard errors, then the Bonferroni test is valid whether or not there is heteroskedasticity, but if the t-statistics are based on homoskedasticity-only standard errors, the Bonferroni test is valid only under homoskedasticity.

Application to Test Scores

The t-statistics testing the joint null hypothesis that the true coefficients on test scores and expenditures per pupil in Equation (7.6) are, respectively, $t_1 = -0.60$ and $t_2 = 2.43$. Although $|t_1| < 2.241$, because $|t_2| > 2.241$ we can reject the joint null hypothesis at the 5% significance level using the Bonferroni test. However, both t_1 and t_2 are less than 2.807 in absolute value, so we cannot reject the joint null hypothesis at the 1% significance level using the Bonferroni test. In contrast, using the F-statistic in Section 7.2, we were able to reject this hypothesis at the 1% significance level.

8 Nonlinear Regression Functions

I n Chapters 4 through 7, the population regression function was assumed to be linear; that is, it has a constant slope. In the context of causal inference, this constant slope corresponds to the effect on Y of a unit change in X being the same for all values of the regressors. But what if the effect on Y of a change in X in fact depends on the value of one or more of the regressors? If so, the population regression function is nonlinear.

This chapter develops two groups of methods for detecting and modeling nonlinear population regression functions. The methods in the first group are useful when the relationship between Y and an independent variable, X_1, depends on the value of X_1 itself. For example, reducing class sizes by one student per teacher might have a greater effect if class sizes are already manageably small than if they are so large that the teacher can do little more than keep the class under control. If so, the test score (Y) is a nonlinear function of the student–teacher ratio (X_1), where this function is steeper when X_1 is small. An example of a nonlinear regression function with this feature is shown in Figure 8.1. Whereas the linear population regression function in Figure 8.1(a) has a constant slope, the nonlinear population regression function in Figure 8.1(b) has a steeper slope when X_1 is small than when it is large. This first group of methods is presented in Section 8.2.

The methods in the second group are useful when the effect on Y of a change in X_1 depends on the value of another independent variable—say, X_2. For example, students still learning English might especially benefit from having more one-on-one attention; if so, the effect on test scores of reducing the student–teacher ratio will be greater in districts with many students still learning English than in districts with few English learners. In this example, the effect on test scores (Y) of a reduction in the student–teacher ratio (X_1) depends on the percentage of English learners in the district (X_2). As shown in Figure 8.1(c), the slope of this type of population regression function depends on the value of X_2. This second group of methods is presented in Section 8.3.

In the models of Sections 8.2 and 8.3, the population regression function is a nonlinear function of the independent variables. Although they are nonlinear in the X's, these models are linear functions of the unknown coefficients (or parameters) of the population regression model and thus are versions of the multiple regression model of Chapters 6 and 7. Therefore, the unknown parameters of these nonlinear regression functions can be estimated and tested using OLS and the methods of Chapters 6 and 7. In some applications, the regression function is a nonlinear function of the X's *and* of the parameters. If so, the parameters cannot be estimated by OLS, but they can be estimated using nonlinear least squares. Appendix 8.1 provides examples of such functions and describes the nonlinear least squares estimator.

Sections 8.1 and 8.2 introduce nonlinear regression functions in the context of regression with a single independent variable, and Section 8.3 extends this to two

FIGURE 8.1   **Population Regression Functions with Different Slopes**

(a) Constant slope

(b) Slope depends on the value of X_1

(c) Slope depends on the value of X_2
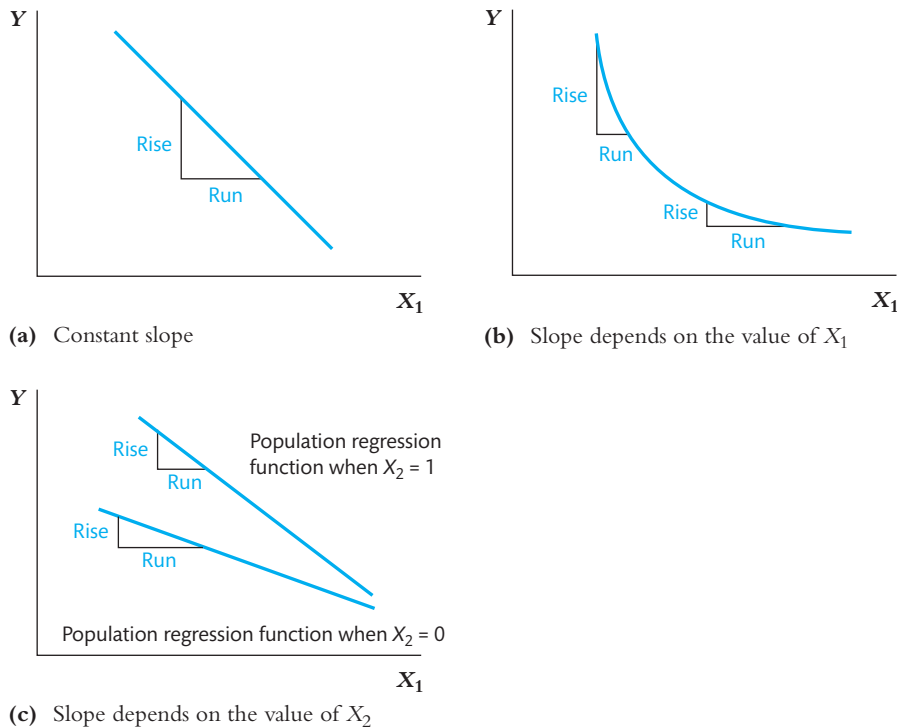
In Figure 8.1(a), the population regression function has a constant slope. In Figure 8.1(b), the slope of the population regression function depends on the value of X_1. In Figure 8.1(c), the slope of the population regression function depends on the value of X_2.

independent variables. To keep things simple, additional regressors are omitted in the empirical examples of Sections 8.1 through 8.3. In practice, however, if the aim is to use the nonlinear model to estimate causal effects, it remains important to control for omitted factors by including control variables as well. In Section 8.4, we combine nonlinear regression functions and additional control variables when we take a close look at possible nonlinearities in the relationship between test scores and the student–teacher ratio, holding student characteristics constant.

The aim of this chapter is to explain the main methods for modeling nonlinear regression functions. In Sections 8.1–8.3, we assume that the least squares assumptions for causal inference in multiple regression (Key Concept 6.4) hold, modified for a nonlinear regression function. Under those assumptions, the slopes of the nonlinear regression functions can be interpreted as causal effects. The methods of this chapter also can be used to model nonlinear population regression functions when some of the regressors are control variables (the assumptions in Key Concept 6.6) and when these functions are used for prediction (the assumptions in Appendix 6.4).

8.1 A General Strategy for Modeling Nonlinear Regression Functions

This section lays out a general strategy for modeling nonlinear population regression functions. In this strategy, the nonlinear models are extensions of the multiple regression model and therefore can be estimated and tested using the tools of Chapters 6 and 7. First, however, we return to the California test score data and consider the relationship between test scores and district income.

Test Scores and District Income

In Chapter 7, we found that the economic background of the students is an important factor in explaining performance on standardized tests. That analysis used two economic background variables (the percentage of students qualifying for a subsidized lunch and the percentage of students whose families qualify for income assistance) to measure the fraction of students in the district coming from poor families. A different, broader measure of economic background is the average annual per capita income in the school district ("district income"). The California data set includes district income measured in thousands of 1998 dollars. The sample contains a wide range of income levels: For the 420 districts in our sample, the median district income is 13.7 (that is, $13,700 per person), and it ranges from 5.3 ($5300 per person) to 55.3 ($55,300 per person).

Figure 8.2 shows a scatterplot of fifth-grade test scores against district income for the California data set, along with the OLS regression line relating these two variables. Test scores and district income are strongly positively correlated, with a

---

FIGURE 8.2   Scatterplot of Test Scores vs. District Income with a Linear OLS Regression Function

There is a positive correlation between test scores and district income (correlation = 0.71), but the linear OLS regression line does not adequately describe the relationship between these variables.
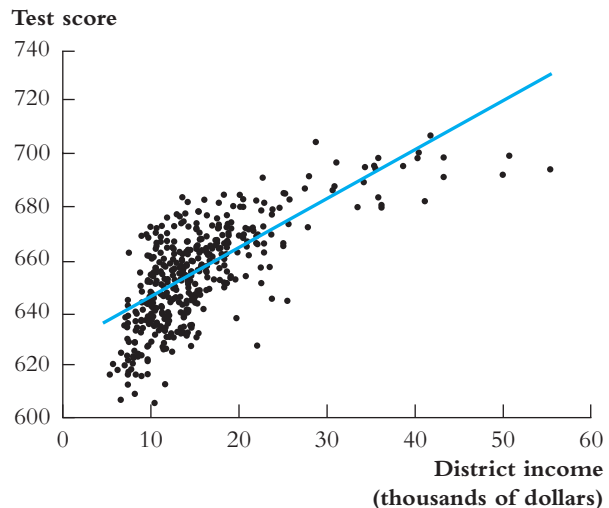
correlation coefficient of 0.71; students from affluent districts do better on the tests than students from poor districts. But this scatterplot has a peculiarity: Most of the points are below the OLS line when income is very low (under $10,000) or very high (over $40,000), but they are above the line when income is between $15,000 and $30,000. There seems to be some curvature in the relationship between test scores and district income that is not captured by the linear regression.

In short, it seems that the relationship between district income and test scores is not a straight line. Rather, it is nonlinear. A nonlinear function is a function with a slope that is not constant: The function $f(X)$ is linear if the slope of $f(X)$ is the same for all values of X, but if the slope depends on the value of X, then $f(X)$ is nonlinear.

If a straight line is not an adequate description of the relationship between district income and test scores, what is? Imagine drawing a curve that fits the points in Figure 8.2. This curve would be steep for low values of district income and then would flatten out as district income gets higher. One way to approximate such a curve mathematically is to model the relationship as a quadratic function. That is, we could model test scores as a function of income *and* the square of income.

A quadratic population regression model relating test scores and income is written mathematically as

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i, \qquad (8.1)$$

where β_0, β_1, and β_2 are coefficients; $Income_i$ is the income in the ith district; $Income_i^2$ is the square of income in the ith district; and u_i is an error term that, as usual, represents all the other factors that determine test scores. Equation (8.1) is called the **quadratic regression model** because the population regression function, $E(TestScore_i | Income_i) = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2$ is a quadratic function of the independent variable, *Income*.

If you knew the population coefficients β_0, β_1, and β_2 in Equation (8.1), you could predict the test score of a district based on its average income. But these population coefficients are unknown and therefore must be estimated using a sample of data.

At first, it might seem difficult to find the coefficients of the quadratic function that best fits the data in Figure 8.2. If you compare Equation (8.1) with the multiple regression model in Key Concept 6.2, however, you will see that Equation (8.1) is, in fact, a version of the multiple regression model with two regressors: The first regressor is *Income*, and the second regressor is $Income^2$. Mechanically, you can create this second regressor by generating a new variable that equals the square of *Income*—for example, as an additional column in a spreadsheet. Thus, after defining the regressors as *Income* and $Income^2$, the nonlinear model in Equation (8.1) is simply a multiple regression model with two regressors!

Because the quadratic regression model is a variant of multiple regression, its unknown population coefficients can be estimated and tested using the OLS methods described in Chapters 6 and 7. Estimating the coefficients of Equation (8.1) using OLS for the 420 observations in Figure 8.2 yields

$$\widehat{TestScore} = 607.3 + 3.85\,Income - 0.0423\,Income^2, \overline{R}^2 = 0.554, \qquad (8.2)$$
$$\phantom{\widehat{TestScore} = 607.3}(2.9)\quad(0.27)(0.0048)$$

where, as usual, standard errors of the estimated coefficients are given in parentheses. The estimated regression function of Equation (8.2) is plotted in Figure 8.3, superimposed over the scatterplot of the data. The quadratic function captures the curvature in the scatterplot: It is steep for low values of district income but flattens out when district income is high. In short, the quadratic regression function seems to fit the data better than the linear one.

We can go one step beyond this visual comparison and formally test the hypothesis that the relationship between district income and test scores is linear against the alternative that it is nonlinear. If the relationship is linear, then the regression function is correctly specified as Equation (8.1) except that the regressor $Income^2$ is absent; that is, if the relationship is linear, then Equation (8.1) holds with $\beta_2 = 0$. Thus we can test the null hypothesis that the population regression function is linear against the alternative that it is quadratic by testing the null hypothesis that $\beta_2 = 0$ against the alternative that $\beta_2 \neq 0$.

Because Equation (8.1) is just a variant of the multiple regression model, the null hypothesis that $\beta_2 = 0$ can be tested by constructing the t-statistic for this hypothesis. This t-statistic is $t = (\hat{\beta}_2 - 0)/SE(\hat{\beta}_2)$, which from Equation (8.2) is $t = -0.0423/0.0048 = -8.81$. In absolute value, this exceeds the 5% critical value of this test (which is 1.96). Indeed, the p-value for the t-statistic is less than 0.01%, so we can reject the hypothesis that $\beta_2 = 0$ at all conventional significance levels. Thus this formal hypothesis test supports our informal inspection of Figures 8.2 and 8.3: The quadratic model fits the data better than the linear model.

FIGURE 8.3   Scatterplot of Test Scores vs. District Income with Linear and Quadratic Regression Functions

The quadratic OLS regression function fits the data better than the linear OLS regression function.
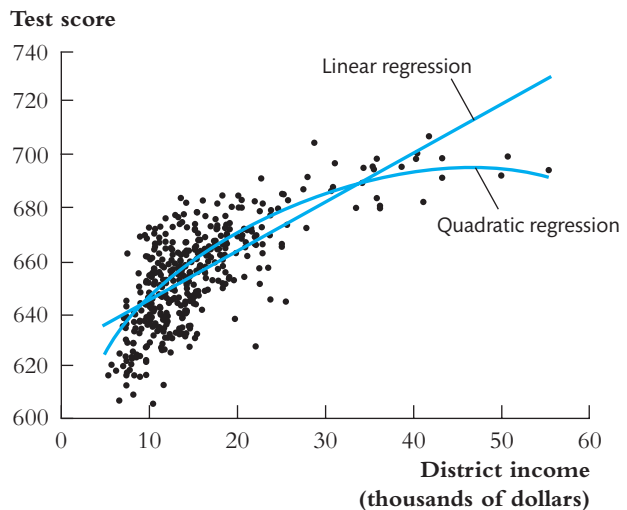
The Effect on *Y* of a Change in *X* in Nonlinear Specifications

Put aside the test score example for a moment, and consider a general problem. You want to know how the dependent variable Y is expected to change when the independent variable X_1 changes by the amount ΔX_1, holding constant other independent variables $X_2, \ldots, X_k$. When the population regression function is linear, this effect is easy to calculate: As shown in Equation (6.4), the expected change in Y is $\Delta Y = \beta_1 \Delta X$, where β_1 is the population regression coefficient multiplying X_1. When the regression function is nonlinear, however, the expected change in Y is more complicated to calculate because it can depend on the values of the independent variables.

A general formula for a nonlinear population regression function.[1] The nonlinear population regression models considered in this chapter are of the form

$$Y_i = f(X_{1i}, X_{2i}, \ldots, X_{ki}) + u_i, i = 1, \ldots, n, \tag{8.3}$$

where $f(X_{1i}, X_{2i}, \ldots, X_{ki})$ is the population **nonlinear regression function**, a possibly nonlinear function of the independent variables $X_{1i}, X_{2i}, \ldots, X_{ki}$, and u_i is the error term. For example, in the quadratic regression model in Equation (8.1), only one independent variable is present, so X_1 is *Income* and the population regression function is $f(Income_i) = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2$.

Because the population regression function is the conditional expectation of Y_i given $X_{1i}, X_{2i}, \ldots, X_{ki}$, in Equation (8.3) we allow for the possibility that this conditional expectation is a nonlinear function of X_1; that is, $E(Y_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = f(X_{1i}, X_{2i}, \ldots, X_{ki})$, where f can be a nonlinear function. If the population regression function is linear, then $f(X_{1i}, X_{2i}, \ldots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$, and Equation (8.3) becomes the linear regression model in Key Concept 6.2. However, Equation (8.3) allows for nonlinear regression functions as well.

The effect on Y of a change in X_1. Suppose an experiment is conducted on individuals with the same values of $X_2, \ldots, X_k$, and participants are randomly assigned treatment levels $X_1 = x_1$ or $X_1 + \Delta X_1 = x_1 + \Delta x_1$. Then the expected difference in outcomes is the causal effect of the treatment, holding constant $X_2, \ldots, X_k$. In the nonlinear regression model of Equation (8.3), this effect on Y is $\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k)$. In the context of prediction,

---

[1]The term *nonlinear regression* applies to two conceptually different families of models. In the first family, the population regression function is a nonlinear function of the *X*'s but is a linear function of the unknown parameters (the β's). In the second family, the population regression function is a nonlinear function of the unknown parameters and may or may not be a nonlinear function of the *X*'s. The models in the body of this chapter are all in the first family. Appendix 8.1 takes up models from the second family.

KEY CONCEPT

8.1

The Expected Change in Y from a Change in X_1 in the Nonlinear Regression Model [Equation (8.3)]

The expected change in Y, ΔY, associated with the change in X_1, ΔX_1, holding $X_2, \ldots, X_k$ constant, is the difference between the value of the population regression function before and after changing X_1, holding $X_2, \ldots, X_k$ constant. That is, the expected change in Y is the difference:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k). \qquad (8.4)$$

The estimator of this unknown population difference is the difference between the predicted values for these two cases. Let $\hat{f}(X_1, X_2, \ldots, X_k)$ be the predicted value of Y based on the estimator $\hat{f}$ of the population regression function. Then the predicted change in Y is

$$\Delta \hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k). \qquad (8.5)$$

$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k)$ is the predicted difference in Y for two observations, both with the same values of $X_2, \ldots, X_k$, but with different values of X_1, specifically $X_1 + \Delta X_1$ and X_1.

Because the regression function f is unknown, this population causal effect is also unknown. To estimate this effect, first estimate the regression function f. At a general level, denote this estimated function by $\hat{f}$; an example of such an estimated function is the estimated quadratic regression function in Equation (8.2). The estimated effect on Y (denoted $\Delta \hat{Y}$) of the change in X_1 is the difference between the predicted value of Y when the independent variables take on the values $X_1 + \Delta X_1, X_2, \ldots, X_k$ and the predicted value of Y when they take on the values $X_1, X_2, \ldots, X_k$.

The method for calculating the predicted change in Y associated with a change in X_1 is summarized in Key Concept 8.1. The computational method in Key Concept 8.1 always works, whether ΔX_1 is large or small and whether the regressors are continuous or discrete. Appendix 8.2 shows how to evaluate the slope using calculus for the special case of a single continuous regressor when ΔX_1 small.

Application to test scores and district income. What is the predicted change in test scores associated with a change in district income of $1000, based on the estimated quadratic regression function in Equation (8.2)? Because that regression function is quadratic, this effect depends on the initial district income. We therefore consider two cases: an increase in district income from 10 to 11 (i.e., from $10,000 per capita to $11,000 per capita) and an increase in district income from 40 to 41 (i.e., from $40,000 per capita to $41,000 per capita).

To compute $\Delta\hat{Y}$ associated with the change in income from 10 to 11, we can apply the general formula in Equation (8.5) to the quadratic regression model. Doing so yields

$$\Delta\hat{Y} = (\hat{\beta}_0 + \hat{\beta}_1 \times 11 + \hat{\beta}_2 \times 11^2) - (\hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 10^2), \quad (8.6)$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are the OLS estimators.

The term in the first set of parentheses in Equation (8.6) is the predicted value of Y when $Income = 11$, and the term in the second set of parentheses is the predicted value of Y when $Income = 10$. These predicted values are calculated using the OLS estimates of the coefficients in Equation (8.2). Accordingly, when $Income = 10$, the predicted value of test scores is $607.3 + 3.85 \times 10 - 0.0423 \times 10^2 = 641.57$. When $Income = 11$, the predicted value is $607.3 + 3.85 \times 11 - 0.0423 \times 11^2 = 644.53$. The difference in these two predicted values is $\Delta\hat{Y} = 644.53 - 641.57 = 2.96$ points; that is, the predicted difference in test scores between a district with average income of \$11,000 and one with average income of \$10,000 is 2.96 points.

In the second case, when income changes from \$40,000 to \$41,000, the difference in the predicted values in Equation (8.6) is $\Delta\hat{Y} = (607.3 + 3.85 \times 41 - 0.0423 \times 41^2) - (607.3 + 3.85 \times 40 - 0.0423 \times 40^2) = 694.04 - 693.62 = 0.42$ points. Thus a change of income of \$1000 is associated with a larger change in predicted test scores if the initial income is \$10,000 than if it is \$40,000 (the predicted changes are 2.96 points versus 0.42 points). Said differently, the slope of the estimated quadratic regression function in Figure 8.3 is steeper at low values of income (like \$10,000) than at the higher values of income (like \$40,000).

Standard errors of estimated effects. The estimate of the effect on Y of changing X depends on the estimate of the population regression function, $\hat{f}$, which varies from one sample to the next. Therefore, the estimated effect contains a sampling error. One way to quantify the sampling uncertainty associated with the estimated effect is to compute a confidence interval for the true population effect. To do so, we need to compute the standard error of $\Delta\hat{Y}$ in Equation (8.5).

It is easy to compute a standard error for $\Delta\hat{Y}$ when the regression function is linear. The estimated effect of a change in X_1 is $\hat{\beta}_1\Delta X_1$, so the standard error of $\Delta\hat{Y}$ is $SE(\Delta\hat{Y}) = SE(\hat{\beta}_1)\Delta X_1$ and a 95% confidence interval for the estimated change is $\hat{\beta}_1\Delta X_1 \pm 1.96\ SE(\hat{\beta}_1)\Delta X_1$.

In the nonlinear regression models of this chapter, the standard error of $\Delta\hat{Y}$ can be computed using the tools introduced in Section 7.3 for testing a single restriction involving multiple coefficients. To illustrate this method, consider the estimated change in test scores associated with a change in income from 10 to 11 in Equation (8.6), which is $\Delta\hat{Y} = \hat{\beta}_1 \times (11 - 10) + \hat{\beta}_2 \times (11^2 - 10^2) = \hat{\beta}_1 + 21\hat{\beta}_2$. The standard error of the predicted change therefore is

$$SE(\Delta\hat{Y}) = SE(\hat{\beta}_1 + 21\hat{\beta}_2). \quad (8.7)$$

Thus, if we can compute the standard error of $\hat{\beta}_1 + 21\hat{\beta}_2$, then we have computed the standard error of $\Delta\hat{Y}$.

Some regression software has a specialized command for computing the standard error in Equation (8.7) directly. If not, there are two other ways to compute it; these correspond to the two approaches in Section 7.3 for testing a single restriction on multiple coefficients.

The first method is to use approach 1 of Section 7.3, which is to compute the F-statistic testing the hypothesis that $\beta_1 + 21\beta_2 = 0$. The standard error of $\Delta\hat{Y}$ is then given by[2]

$$SE(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}. \tag{8.8}$$

When applied to the quadratic regression in Equation (8.2), the F-statistic testing the hypothesis that $\beta_1 + 21\beta_2 = 0$ is $F = 299.94$. Because $\Delta\hat{Y} = 2.96$, applying Equation (8.8) gives $SE(\Delta\hat{Y}) = 2.96/\sqrt{299.94} = 0.17$. Thus a 95% confidence interval for the change in the expected value of Y is $2.96 \pm 1.96 \times 0.17$ or $(2.63, 3.29)$.

The second method is to use approach 2 of Section 7.3, which entails transforming the regressors so that, in the transformed regression, one of the coefficients is $\beta_1 + 21\beta_2$. Doing this transformation is left as an exercise (Exercise 8.9).

A comment on interpreting coefficients in nonlinear specifications. In the multiple regression model of Chapters 6 and 7, the regression coefficients had a natural interpretation. For example, β_1 is the expected change in Y associated with a change in X_1, holding the other regressors constant. But as we have seen, this is not generally the case in a nonlinear model. That is, it is not very helpful to think of β_1 in Equation (8.1) as being the effect of changing the district income, holding the square of the district income constant. In nonlinear models, the regression function is best interpreted by graphing it and by calculating the predicted effect on Y of changing one or more of the independent variables.

A General Approach to Modeling Nonlinearities Using Multiple Regression

The general approach to modeling nonlinear regression functions taken in this chapter has five elements:

1. *Identify a possible nonlinear relationship.* The best thing to do is to use economic theory and what you know about the application to suggest a possible nonlinear relationship. Before you even look at the data, ask yourself whether the slope of the regression function relating Y and X might reasonably depend

---

[2]Equation (8.8) is derived by noting that the F-statistic is the square of the t-statistic testing this hypothesis—that is, $F = t^2 = [(\hat{\beta}_1 + 21\hat{\beta}_2)/SE(\hat{\beta}_1 + 21\hat{\beta}_1)]^2 = [\Delta\hat{Y}/SE(\Delta\hat{Y})]^2$—and solving for $SE(\Delta\hat{Y})$.

on the value of X or on another independent variable. Why might such nonlinear dependence exist? What nonlinear shapes does this suggest? For example, thinking about classroom dynamics with 11-year-olds suggests that cutting class size from 18 students to 17 could have a greater effect than cutting it from 30 to 29.

2. ***Specify a nonlinear function, and estimate its parameters by OLS.*** Sections 8.2 and 8.3 contain various nonlinear regression functions that can be estimated by OLS. After working through these sections, you will understand the characteristics of each of these functions.

3. ***Determine whether the nonlinear model improves upon a linear model.*** Just because you think a regression function is nonlinear does not mean it really is! You must determine empirically whether your nonlinear model is appropriate. Most of the time you can use t-statistics and F-statistics to test the null hypothesis that the population regression function is linear against the alternative that it is nonlinear.

4. ***Plot the estimated nonlinear regression function.*** Does the estimated regression function describe the data well? Looking at Figures 8.2 and 8.3 suggests that the quadratic model fits the data better than the linear model.

5. ***Estimate the effect on Y of a change in X.*** The final step is to use the estimated regression to calculate the effect on Y of a change in one or more regressors X using the method in Key Concept 8.1.

8.2 Nonlinear Functions of a Single Independent Variable

This section provides two methods for modeling a nonlinear regression function. To keep things simple, we develop these methods for a nonlinear regression function that involves only one independent variable, X. As we see in Section 8.5, however, these models can be modified to include multiple independent variables.

The first method discussed in this section is polynomial regression, an extension of the quadratic regression used in the last section to model the relationship between test scores and district income. The second method uses logarithms of X, of Y, or of both X and Y. Although these methods are presented separately, they can be used in combination.

Appendix 8.2 provides a calculus-based treatment of the models in this section.

Polynomials

One way to specify a nonlinear regression function is to use a polynomial in X. In general, let r denote the highest power of X that is included in the regression. The **polynomial regression model** of degree r is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i. \tag{8.9}$$

When $r = 2$, Equation (8.9) is the quadratic regression model discussed in Section 8.1. When $r = 3$, so that the highest power of X included is X^3, Equation (8.9) is called the **cubic regression model**.

The polynomial regression model is similar to the multiple regression model of Chapter 6 except that in Chapter 6 the regressors were distinct independent variables, whereas here the regressors are powers of the same dependent variable, X; that is, the regressors are X, X^2, X^3, and so on. Thus the techniques for estimation and inference developed for multiple regression can be applied here. In particular, the unknown coefficients β_0, β_1, ..., β_r in Equation (8.9) can be estimated by OLS regression of Y_i against X_i, X_i^2, ..., X_i^r.

Testing the null hypothesis that the population regression function is linear.   If the population regression function is linear, then the quadratic and higher-degree terms do not enter the population regression function. Accordingly, the null hypothesis (H_0) that the regression is linear and the alternative (H_1) that it is a polynomial of degree up to r correspond to

$$H_0\colon \beta_2 = 0,\ \beta_3 = 0,\ \ldots,\ \beta_r = 0 \text{ vs. } H_1\colon \text{at least one } \beta_j \neq 0,\ j = 2,\ \ldots, r. \quad (8.10)$$

The null hypothesis that the population regression function is linear can be tested against the alternative that it is a polynomial of degree up to r by testing H_0 against H_1 in Equation (8.10). Because H_0 is a joint null hypothesis with $q = r - 1$ restrictions on the coefficients of the population polynomial regression model, it can be tested using the F-statistic as described in Section 7.2.

Which degree polynomial should I use?   That is, how many powers of X should be included in a polynomial regression? The answer balances a trade-off between flexibility and statistical precision. Increasing the degree r introduces more flexibility into the regression function and allows it to match more shapes; a polynomial of degree r can have up to $r - 1$ bends (that is, inflection points) in its graph. But increasing r means adding more regressors, which can reduce the precision of the estimated coefficients.

Thus the answer to the question of how many terms to include is that you should include enough to model the nonlinear regression function adequately—but no more. Unfortunately, this answer is not very useful in practice!

A practical way to determine the degree of the polynomial is to ask whether the coefficients in Equation (8.9) associated with largest values of r are 0. If so, then these terms can be dropped from the regression. This procedure, which is called sequential hypothesis testing because individual hypotheses are tested sequentially, is summarized in the following steps:

1. Pick a maximum value of r, and estimate the polynomial regression for that r.

2. Use the *t*-statistic to test the hypothesis that the coefficient on X^r, β_r in Equation (8.9), is 0. If you reject this hypothesis, then X^r belongs in the regression, so use the polynomial of degree *r*.

3. If you do not reject $\beta_r = 0$ in step 2, eliminate X^r from the regression, and estimate a polynomial regression of degree $r - 1$. Test whether the coefficient on X^{r-1} is 0. If you reject, use the polynomial of degree $r - 1$.

4. If you do not reject $\beta_{r-1} = 0$ in step 3, continue this procedure until the coefficient on the highest power in your polynomial is statistically significant.

This recipe has one missing ingredient: the initial degree *r* of the polynomial. In many applications involving economic data, the nonlinear functions are smooth; that is, they do not have sharp jumps, or "spikes." If so, then it is appropriate to choose a small maximum degree for the polynomial, such as 2, 3, or 4—that is, to begin with $r = 2$ or 3 or 4 in step 1.

Application to district income and test scores.   The estimated cubic regression function relating district income to test scores is

$$\widehat{TestScore} = 600.1 + 5.02\, Income - 0.096\, Income^2 + 0.00069\, Income^3,$$
$$\quad\quad\quad (5.1)\quad (0.71)\quad\quad\quad (0.029)\quad\quad\quad\quad (0.00035)$$

$$\bar{R}^2 = 0.555. \tag{8.11}$$

The *t*-statistic on $Income^3$ is 1.97, so the null hypothesis that the regression function is a quadratic is rejected against the alternative that it is a cubic at the 5% level. Moreover, the *F*-statistic testing the joint null hypothesis that the coefficients on $Income^2$ and $Income^3$ are both 0 is 37.7, with a *p*-value less than 0.01%, so the null hypothesis that the regression function is linear is rejected against the alternative that it is either a quadratic or a cubic.

Interpretation of coefficients in polynomial regression models.   The coefficients in polynomial regressions do not have a simple interpretation. The best way to interpret polynomial regressions is to plot the estimated regression function and calculate the estimated effect on Y associated with a change in X for one or more values of X.

Logarithms

Another way to specify a nonlinear regression function is to use the natural logarithm of Y and/or X. Logarithms convert changes in variables into percentage changes, and many relationships are naturally expressed in terms of percentages. Here are some examples:

- A box in Chapter 3, "The Gender Gap of Earnings of College Graduates in the United States," examined the wage gap between male and female college graduates. In that discussion, the wage gap was measured in terms of dollars. However, it is easier to compare wage gaps across professions and over time when they are expressed in percentage terms.

- In Section 8.1, we found that district income and test scores were nonlinearly related. Would this relationship be linear using percentage changes? That is, might it be that a change in district income of 1%—rather than $1000—is associated with a change in test scores that is approximately constant for different values of income?

- In the economic analysis of consumer demand, it is often assumed that a 1% increase in price leads to a certain *percentage* decrease in the quantity demanded. The percentage decrease in demand resulting from a 1% increase in price is called the price **elasticity**.

Regression specifications that use natural logarithms allow regression models to estimate percentage relationships such as these. Before introducing those specifications, we review the exponential and natural logarithm functions.

The exponential function and the natural logarithm. The exponential function and its inverse, the natural logarithm, play an important role in modeling nonlinear regression functions. The **exponential function** of x is e^x (that is, e raised to the power x), where e is the constant $2.71828\ldots$; the exponential function is also written as $\exp(x)$. The **natural logarithm** is the inverse of the exponential function; that is, the natural logarithm is the function for which $x = \ln(e^x)$ or, equivalently, $x = \ln[\exp(x)]$. The base of the natural logarithm is e. Although there are logarithms in other bases, such as base 10, in this text we consider only logarithms in base e—that is, the natural logarithm—so when we use the term *logarithm*, we always mean *natural logarithm*.

The logarithm function $y = \ln(x)$ is graphed in Figure 8.4. Note that the logarithm function is defined only for positive values of x. The logarithm function has a slope that is steep at first and then flattens out (although the function continues to increase). The slope of the logarithm function $\ln(x)$ is $1/x$.

The logarithm function has the following useful properties:

$$\ln(1/x) = -\ln(x); \tag{8.12}$$

$$\ln(ax) = \ln(a) + \ln(x); \tag{8.13}$$

$$\ln(x/a) = \ln(x) - \ln(a); \text{ and} \tag{8.14}$$

$$\ln(x^a) = a \ln(x). \tag{8.15}$$

Logarithms and percentages. The link between the logarithm and percentages relies on a key fact: When Δx is small, the difference between the logarithm of

---

FIGURE 8.4   The Logarithm Function, $y = \ln(x)$

The logarithmic function $y = \ln(x)$ is steeper for small than for large values of x, is defined only for $x > 0$, and has slope $1/x$.
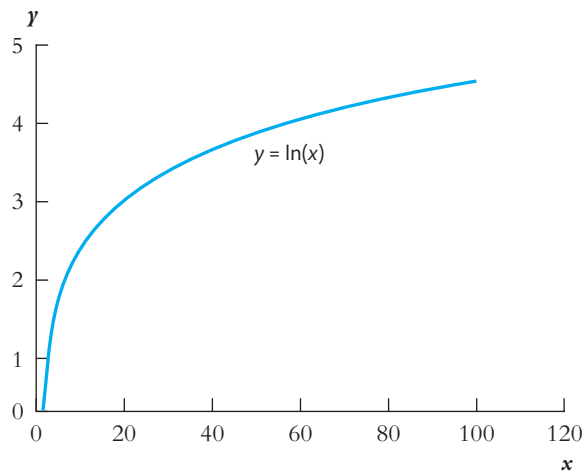
$x + \Delta x$ and the logarithm of x is approximately $\Delta x/x$, the percentage change in x divided by 100. That is,

$$\ln(x + \Delta x) - \ln(x) \cong \frac{\Delta x}{x} \quad \left(\text{when } \frac{\Delta x}{x} \text{ is small}\right), \tag{8.16}$$

where "$\cong$" means "approximately equal to." The derivation of this approximation relies on calculus, but it is readily demonstrated by trying out some values of x and Δx. For example, when $x = 100$ and $\Delta x = 1$, then $\Delta x/x = 1/100 = 0.01$ (or 1%), while $\ln(x + \Delta x) - \ln(x) = \ln(101) - \ln(100) = 0.00995$ (or 0.995%). Thus $\Delta x/x$ (which is 0.01) is very close to $\ln(x + \Delta x) - \ln(x)$ (which is 0.00995). When $\Delta x = 5$, $\Delta x/x = 5/100 = 0.05$, while $\ln(x + \Delta x) - \ln(x) = \ln(105) - \ln(100) = 0.04879$.

The three logarithmic regression models. There are three different cases in which logarithms might be used: when X is transformed by taking its logarithm but Y is not; when Y is transformed to its logarithm but X is not; and when both Y and X are transformed to their logarithms. The interpretation of the regression coefficients is different in each case. We discuss these three cases in turn.

Case I: X is in logarithms, Y is not. In this case, the regression model is

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i, \ i = 1, \ldots, n. \tag{8.17}$$

Because Y is not in logarithms but X is, this is sometimes referred to as a **linear-log model**.

In the linear-log model, a 1% change in X is associated with a change in Y of $0.01\beta_1$. To see this, consider the differences inbetween the population regression function at

values of X that differ by ΔX: This is $[\beta_0 + \beta_1 \ln(X + \Delta X)] - [\beta_0 + \beta_1 \ln(X)] = \beta_1[\ln(X + \Delta X) - \ln(X)] \cong \beta_1(\Delta X/X)$, where the final step uses the approximation in Equation (8.16). If X changes by 1%, then $\Delta X/X = 0.01$; thus in this model a 1% change in X is associated with a change of Y of $0.01\beta_1$.

The only difference between the regression model in Equation (8.17) and the regression model of Chapter 4 with a single regressor is that the right-hand variable is now the logarithm of X rather than X itself. To estimate the coefficients β_0 and β_1 in Equation (8.17), first compute a new variable, $\ln(X)$, which is readily done using a spreadsheet or statistical software. Then β_0 and β_1 can be estimated by the OLS regression of Y_i on $\ln(X_i)$, hypotheses about β_1 can be tested using the t-statistic, and a 95% confidence interval for β_1 can be constructed as $\hat{\beta}_1 \pm 1.96\ SE(\hat{\beta}_1)$.

As an example, return to the relationship between district income and test scores. Instead of the quadratic specification, we could use the linear-log specification in Equation (8.17). Estimating this regression by OLS yields

$$\widehat{TestScore} = 557.8 + 36.42\ln(Income),\ \overline{R}^2 = 0.561.$$
$$\qquad\qquad\quad (3.8)\quad\ (1.40) \tag{8.18}$$

According to Equation (8.18), a 1% increase in income is associated with an increase in test scores of $0.01 \times 36.42 = 0.36$ points.

To estimate the effect on Y of a change in X in its original units of thousands of dollars (not in logarithms), we can use the method in Key Concept 8.1. For example, what is the predicted difference in test scores for districts with average incomes of \$10,000 versus \$11,000? The estimated value of ΔY is the difference between the predicted values: $\Delta\hat{Y} = [557.8 + 36.42\ln(11)] - [557.8 + 36.42\ln(10)] = 36.42 \times [\ln(11) - \ln(10)] = 3.47$. Similarly, the predicted difference between a district with average income of \$40,000 and a district with average income of \$41,000 is $36.42 \times [\ln(41) - \ln(40)] = 0.90$. Thus, like the quadratic specification, this regression predicts that a \$1000 increase in income has a larger effect on test scores in poor districts than it does in affluent districts.

The estimated linear-log regression function in Equation (8.18) is plotted in Figure 8.5. Because the regressor in Equation (8.18) is the natural logarithm of income rather than income, the estimated regression function is not a straight line. Like the quadratic regression function in Figure 8.3, it is initially steep but then flattens out for higher levels of income.

Case II: Y is in logarithms, X is not. In this case, the regression model is

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i. \tag{8.19}$$

Because Y is in logarithms but X is not, this is referred to as a **log-linear model**.

In the log-linear model, a one-unit change in $X(\Delta X = 1)$ is associated with a $(100 \times \beta_1)\%$ change in Y. To see this, compare the expected values of $\ln(Y)$ for values of X that differ by ΔX. The expected value of $\ln(Y)$ given X is $\ln(Y) = \beta_0 + \beta_1 X$.

FIGURE 8.5 **The Linear-Log Regression Function**

The estimated linear-log regression function $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln(X)$ captures much of the nonlinear relation between test scores and district income.
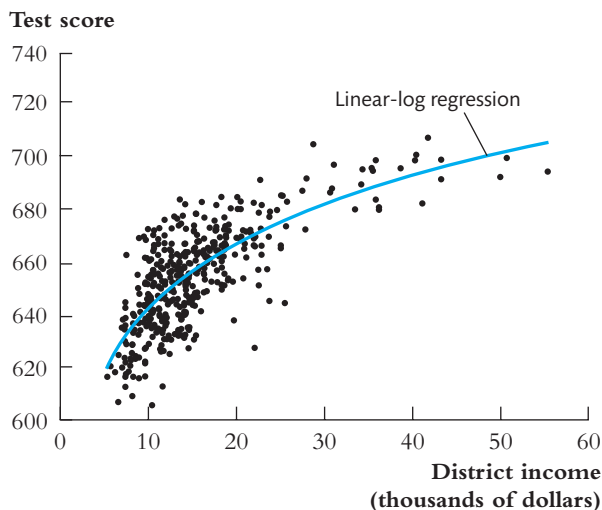
For $X + \Delta X$, the expected value is given by $\ln(Y + \Delta Y) = \beta_0 + \beta_1(X + \Delta X)$. Thus the difference between these expected values is $\ln(Y + \Delta Y) - \ln(Y) = [\beta_0 + \beta_1(X + \Delta X)] - [\beta_0 + \beta_1 X] = \beta_1 \Delta X$. From the approximation in Equation (8.16), however, if $\beta_1 \Delta X$ is small, then $\ln(Y + \Delta Y) - \ln(Y) \cong \Delta Y/Y$. Thus $\Delta Y/Y \cong \beta_1 \Delta X$. If $\Delta X = 1$, so that X changes by one unit, then $\Delta Y/Y$ changes by β_1. Translated into percentages, a one-unit change in X is associated with a $(100 \times \beta_1)\%$ change in Y.

As an illustration, we return to the empirical example of Section 3.7, the relationship between age and earnings of college graduates. Some employment contracts specify that, for each additional year of service, a worker gets a certain percentage increase in his or her wage. This percentage relationship suggests estimating the log-linear specification in Equation (8.19) so that each additional year of age (X) is, on average, associated with some constant percentage increase in earnings (Y). By first computing the new dependent variable, $\ln(Earnings_i)$, the unknown coefficients β_0 and β_1 can be estimated by the OLS regression of $\ln(Earnings_i)$ against Age_i. When estimated using the 13,872 observations on college graduates in the March 2016 Current Population Survey (the data are described in Appendix 3.1), this relationship is

$$\widehat{\ln(Earnings)} = 2.876 + 0.0095\,Age, \overline{R}^2 = 0.033.$$
$$\quad\quad\quad\quad\quad (0.019)\quad(0.0004) \tag{8.20}$$

According to this regression, earnings are predicted to increase by 0.95% $[(100 \times 0.0095)\%]$ for each additional year of age.

Case III: Both X and Y are in logarithms. In this case, the regression model is

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i. \tag{8.21}$$

Because both Y and X are specified in logarithms, this is referred to as a **log-log model**.

In the log-log model, a 1% change in X is associated with a β_1% change in Y. Thus in this specification β_1 is the elasticity of Y with respect to X. To see this, again apply Key Concept 8.1; thus $\ln(Y + \Delta Y) - \ln(Y) = [\beta_0 + \beta_1 \ln(X + \Delta X)] - [\beta_0 + \beta_1 \ln(X)] = \beta_1 [\ln(X + \Delta X) - \ln(X)]$. Application of the approximation in Equation (8.16) to both sides of this equation yields

$$\frac{\Delta Y}{Y} \cong \beta_1 \frac{\Delta X}{X} \text{ or}$$

$$\beta_1 = \frac{\Delta Y/Y}{\Delta X/X} = \frac{100 \times (\Delta Y/Y)}{100 \times (\Delta X/X)} = \frac{\text{percentage change in } Y}{\text{percentage change in } X}. \tag{8.22}$$

Thus in the log-log specification β_1 is the ratio of the percentage change in Y associated with the percentage change in X. If the percentage change in X is 1% (that is, if $\Delta X = 0.01X$), then β_1 is the percentage change in Y associated with a 1% change in X. That is, β_1 is the elasticity of Y with respect to X.

As an illustration, return to the relationship between district income and test scores. When this relationship is specified in this form, the unknown coefficients are estimated by a regression of the logarithm of test scores against the logarithm of district income. The resulting estimated equation is

$$\widehat{\ln(TestScore)} = 6.336 + 0.0554 \ln(Income), \overline{R}^2 = 0.557. \tag{8.23}$$
$$\phantom{\widehat{\ln(TestScore)} = } (0.006) \quad (0.0021)$$

According to this estimated regression function, a 1% increase in income is estimated to correspond to a 0.0554% increase in test scores.

The estimated log-log regression function in Equation (8.23) is plotted in Figure 8.6. Because Y is in logarithms, the vertical axis in Figure 8.6 is the logarithm of the test score, and the scatterplot is the logarithm of test scores versus district income. For comparison purposes, Figure 8.6 also shows the estimated regression function for a log-linear specification, which is

$$\widehat{\ln(TestScore)} = 6.439 + 0.00284 \, Income, \overline{R}^2 = 0.497. \tag{8.24}$$
$$\phantom{\widehat{\ln(TestScore)} = } (0.003) \quad (0.00018)$$

Because the vertical axis is in logarithms, the regression function in Equation (8.24) is the straight line in Figure 8.6.

In the log-linear regression function, ln(Y) is a linear function of X. In the log-log regression function, ln(Y) is a linear function of ln(X).
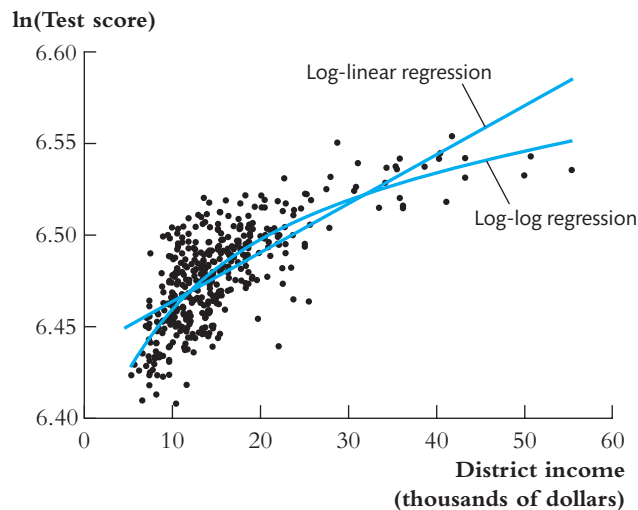
As you can see in Figure 8.6, the log-log specification fits better than the log-linear specification. This is consistent with the higher $\overline{R}^2$ for the log-log regression (0.557) than for the log-linear regression (0.497). Even so, the log-log specification does not fit the data especially well: At the lower values of income, most of the observations fall below the log-log curve, while in the middle income range most of the observations fall above the estimated regression function.

The three logarithmic regression models are summarized in Key Concept 8.2.

A difficulty with comparing logarithmic specifications. Which of the log regression models best fits the data? As we saw in the discussion of Equations (8.23) and (8.24), the $\overline{R}^2$ can be used to compare the log-linear and log-log models; as it happened, the log-log model had the higher $\overline{R}^2$. Similarly, the $\overline{R}^2$ can be used to compare the linear-log regression in Equation (8.18) and the linear regression of Y against X. In the test score and district income regression, the linear-log regression has an $\overline{R}^2$ of 0.561, while the linear regression has an $\overline{R}^2$ of 0.508, so the linear-log model fits the data better.

How can we compare the linear-log model and the log-log model? Unfortunately, the $\overline{R}^2$ *cannot* be used to compare these two regressions because their dependent variables are different [one is Y, the other is $\ln(Y)$]. Recall that the $\overline{R}^2$ measures the fraction of the variance of the dependent variable explained by the regressors. Because the dependent variables in the log-log and linear-log models are different, it does not make sense to compare their $\overline{R}^2$'s.

Because of this problem, the best thing to do in a particular application is to decide, using economic theory and either your or other experts' knowledge of the problem, whether it makes sense to specify Y in logarithms. For example, labor economists typically model earnings using logarithms because wage comparisons, contract

Logarithms in Regression: Three Cases

Logarithms can be used to transform the dependent variable Y, an independent variable X, or both (but the variable being transformed must be positive). The following table summarizes these three cases and the interpretation of the regression coefficient β_1. In each case, β_1 can be estimated by applying OLS after taking the logarithm of the dependent and/or independent variable.

Case	Regression Specification	Interpretation of β_1
I	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$	A 1% change in X is associated with a change in Y of $0.01\beta_1$.
II	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$	A change in X by one unit ($\Delta X = 1$) is associated with a $100\beta_1\%$ change in Y.
III	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$	A 1% change in X is associated with a $\beta_1\%$ change in Y, so β_1 is the elasticity of Y with respect to X.

wage increases, and so forth are often most naturally discussed in percentage terms. In modeling test scores, it seems natural (to us, anyway) to discuss test results in terms of points on the test rather than percentage increases in the test scores, so we focus on models in which the dependent variable is the test score rather than its logarithm.

Computing predicted values of Y when Y is in logarithms.[3] If the dependent variable Y has been transformed by taking logarithms, the estimated regression can be used to compute directly the predicted value of $\ln(Y)$. However, it is a bit trickier to compute the predicted value of Y itself.

To see this, consider the log-linear regression model in Equation (8.19), and rewrite it so that it is specified in terms of Y rather than $\ln(Y)$. To do so, take the exponential function of both sides of Equation (8.19); the result is

$$Y_i = \exp(\beta_0 + \beta_1 X_i + u_i) = e^{\beta_0 + \beta_1 X_i} e^{u_i}. \tag{8.25}$$

The expected value of Y_i given X_i is $E(Y_i|X_i) = E(e^{\beta_0 + \beta_1 X_i} e^{u_i}|X_i) = e^{\beta_0 + \beta_1 X_i} E(e^{u_i}|X_i)$. The problem is that even if $E(u_i|X_i) = 0$, $E(e^{u_i}|X_i) \neq 1$. Thus the appropriate predicted value of Y_i is not simply obtained by taking the exponential function of $\hat{\beta}_0 + \hat{\beta}_1 X_i$—that is, by setting $\hat{Y}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}$. This predicted value is biased because of the missing factor $E(e^{u_i}|X_i)$.

One solution to this problem is to estimate the factor $E(e^{u_i}|X_i)$ and use this estimate when computing the predicted value of Y. Exercise 17.12 works through

---

[3]This material is more advanced and can be skipped without loss of continuity.

several ways to estimate $E(e^{u_i} | X_i)$, but this gets complicated, particularly if u_i is heteroskedastic, and we do not pursue it further.

Another solution, which is the approach used in this text, is to compute predicted values of the logarithm of Y but not transform them to their original units. In practice, this is often acceptable because when the dependent variable is specified as a logarithm, it is often most natural just to use the logarithmic specification (and the associated percentage interpretations) throughout the analysis.

Polynomial and Logarithmic Models of Test Scores and District Income

In practice, economic theory or expert judgment might suggest a functional form to use, but in the end, the true form of the population regression function is unknown. In practice, fitting a nonlinear function therefore entails deciding which method or combination of methods works best. As an illustration, we compare polynomial and logarithmic models of the relationship between district income and test scores.

Polynomial specifications. We considered two polynomial specifications, quadratic [Equation (8.2)] and cubic [Equation (8.11)]. Because the coefficient on $Income^3$ in Equation (8.11) was significant at the 5% level, the cubic specification provided an improvement over the quadratic, so we select the cubic model as the preferred polynomial specification.

Logarithmic specifications. The logarithmic specification in Equation (8.18) seemed to provide a good fit to these data, but we did not test this formally. One way to do so is to augment it with higher powers of the logarithm of income. If these additional terms are not statistically different from 0, then we can conclude that the specification in Equation (8.18) is adequate in the sense that it cannot be rejected against a polynomial function of the logarithm. Accordingly, the estimated cubic regression (specified in powers of the logarithm of income) is

$$\widehat{TestScore} = 486.1 + 113.4 \ln(Income) - 26.9[\ln(Income)]^2$$
$$\phantom{\widehat{TestScore} =} (79.4)\quad (87.9)\qquad\qquad (31.7)$$
$$+ 3.06[\ln(Income)]^3,\ \overline{R}^2 = 0.560. \qquad\qquad (8.26)$$
$$(3.74)$$

The t-statistic on the coefficient on the cubic term is 0.818, so the null hypothesis that the true coefficient is 0 is not rejected at the 10% level. The F-statistic testing the joint hypothesis that the true coefficients on the quadratic and cubic term are both 0 is 0.44, with a p-value of 0.64, so this joint null hypothesis is not rejected at the 10% level. Thus the cubic logarithmic model in Equation (8.26) does not provide a statistically significant improvement over the model in Equation (8.18), which is linear in the logarithm of income.

> **FIGURE 8.7**   **The Linear-Log and Cubic Regression Functions**
>
> The estimated cubic regression function [Equation (8.11)] and the estimated linear-log regression function [Equation (8.18)] are nearly identical in this sample.
>
>
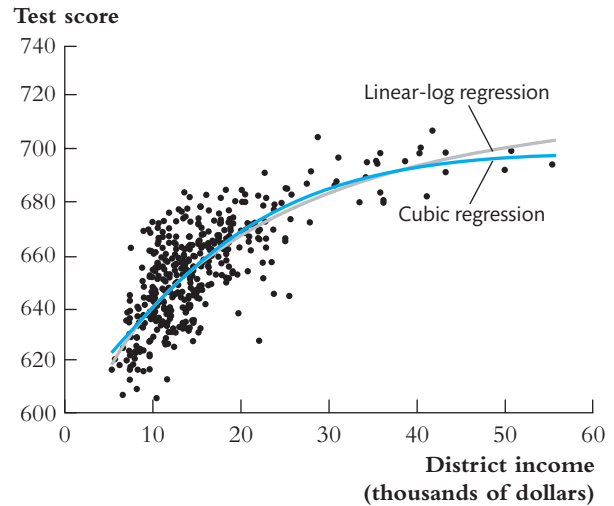
Comparing the cubic and linear-log specifications. Figure 8.7 plots the estimated regression functions from the cubic specification in Equation (8.11) and the linear-log specification in Equation (8.18). The two estimated regression functions are quite similar. One statistical tool for comparing these specifications is the $\overline{R}^2$. The $\overline{R}^2$ of the logarithmic regression is 0.561, and for the cubic regression, it is 0.555. Because the logarithmic specification has a slight edge in terms of the $\overline{R}^2$ and because this specification does not need higher-degree polynomials in the logarithm of income to fit these data, we adopt the logarithmic specification in Equation (8.18).

8.3   Interactions Between Independent Variables

In the introduction to this chapter, we wondered whether reducing the student–teacher ratio might have a bigger effect on test scores in districts where many students are still learning English than in those with few still learning English. This could arise, for example, if students who are still learning English benefit differentially from one-on-one or small-group instruction. If so, the presence of many English learners in a district would interact with the student–teacher ratio in such a way that the effect on test scores of a change in the student–teacher ratio would depend on the fraction of English learners.

This section explains how to incorporate such interactions between two independent variables into the multiple regression model. The possible interaction between the student–teacher ratio and the fraction of English learners is an example of the more general situation in which the effect on Y of a change in one independent variable depends on the value of another independent variable. We consider three cases: when both independent variables are binary, when one is binary and the other is continuous, and when both are continuous.

Interactions Between Two Binary Variables

Consider the population regression of log earnings [Y_i, where $Y_i = \ln(Earnings_i)$] against two binary variables: whether a worker has a college degree (D_{1i}, where $D_{1i} = 1$ if the ith person graduated from college) and the worker's sex (D_{2i}, where $D_{2i} = 1$ if the ith person is female). The population linear regression of Y_i on these two binary variables is

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i. \tag{8.27}$$

In this regression model, β_1 is the effect on log earnings of having a college degree, holding sex constant, and β_2 is the mean difference between female and male earnings, holding schooling constant.

The specification in Equation (8.27) has an important limitation: The effect of having a college degree in this specification, holding constant sex, is the same for men and women. There is, however, no reason that this must be so. Phrased mathematically, the effect on Y_i of D_{1i}, holding D_{2i} constant, could depend on the value of D_{2i}. In other words, there could be an interaction between having a college degree and sex, so that the value in the job market of a degree is different for men and women.

Although the specification in Equation (8.27) does not allow for this interaction between having a college degree and sex, it is easy to modify the specification so that it does by introducing another regressor, the product of the two binary variables, $D_{1i} \times D_{2i}$. The resulting regression is

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3(D_{1i} \times D_{2i}) + u_i. \tag{8.28}$$

The new regressor, the product $D_{1i} \times D_{2i}$, is called an **interaction term** or an **interacted regressor**, and the population regression model in Equation (8.28) is called a binary variable **interaction regression model**.

The interaction term in Equation (8.28) allows the population effect on log earnings (Y_i) of having a college degree (changing D_{1i} from $D_{1i} = 0$ to $D_{1i} = 1$) to depend on sex (D_{2i}). To show this mathematically, calculate the population effect of a change in D_{1i} using the general method laid out in Key Concept 8.1. The first step is to compute the conditional expectation of Y_i for $D_{1i} = 0$ given a value of D_{2i}; this is $E(Y_i | D_{1i} = 0, D_{2i} = d_2) = \beta_0 + \beta_1 \times 0 + \beta_2 \times d_2 + \beta_3 \times (0 \times d_2) = \beta_0 + \beta_2 d_2$, where we use the conditional mean zero assumption, $E(u_i | D_{1i}, D_{2i}) = 0$. The next step is to compute the conditional expectation of Y_i after the change—that is, for $D_{1i} = 1$—given the same value of D_{2i}; this is $E(Y_i | D_{1i} = 1, D_{2i} = d_2) = \beta_0 + \beta_1 \times 1 + \beta_2 \times d_2 + \beta_3 \times (1 \times d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2$. The effect of this change is the difference of expected values [that is, the difference in Equation (8.4)], which is

$$E(Y_i | D_{1i} = 1, D_{2i} = d_2) - E(Y_i | D_{1i} = 0, D_{2i} = d_2) = \beta_1 + \beta_3 d_2. \tag{8.29}$$

A Method for Interpreting Coefficients in Regressions with Binary Variables

First, compute the expected values of Y for each possible case described by the set of binary variables. Next compare these expected values. Each coefficient can then be expressed either as an expected value or as the difference between two or more expected values.

Thus, in the binary variable interaction specification in Equation (8.28), the effect of acquiring a college degree (a unit change in D_{1i}) depends on the person's sex [the value of D_{2i}, which is d_2 in Equation (8.29)]. If the person is male ($d_2 = 0$), the effect of acquiring a college degree is β_1, but if the person is female ($d_2 = 1$), the effect is $\beta_1 + \beta_3$. The coefficient β_3 on the interaction term is the difference in the effect of acquiring a college degree for women versus that for men.

Although this example was phrased using log earnings, having a college degree, and sex, the point is a general one. The binary variable interaction regression allows the effect of changing one of the binary independent variables to depend on the value of the other binary variable.

The method we used here to interpret the coefficients was, in effect, to work through each possible combination of the binary variables. This method, which applies to all regressions with binary variables, is summarized in Key Concept 8.3.

Application to the student–teacher ratio and the percentage of English learners. Let $HiSTR_i$ be a binary variable that equals 1 if the student–teacher ratio is 20 or more and that equals 0 otherwise, and let $HiEL_i$ be a binary variable that equals 1 if the percentage of English learners is 10% or more and that equals 0 otherwise. The interacted regression of test scores against $HiSTR_i$ and $HiEL_i$ is

$$\widehat{TestScore} = 664.1 - 1.9\,HiSTR - 18.2\,HiEL - 3.5(HiSTR \times HiEL),$$
$$\qquad\quad (1.4)\quad (1.9)\qquad\quad (2.3)\qquad\quad (3.1)$$
$$\overline{R}^2 = 0.290. \tag{8.30}$$

The predicted effect of moving from a district with a low student–teacher ratio to one with a high student–teacher ratio, holding constant whether the percentage of English learners is high or low, is given by Equation (8.29), with estimated coefficients replacing the population coefficients. According to the estimates in Equation (8.30), this effect thus is $-1.9 - 3.5HiEL$. That is, if the fraction of English learners is low ($HiEL = 0$), then the effect on test scores of moving from $HiSTR = 0$ to $HiSTR = 1$ is for test scores to decline by 1.9 points. If the fraction of English learners is high, then test scores are estimated to decline by $1.9 + 3.5 = 5.4$ points.

The estimated regression in Equation (8.30) also can be used to estimate the mean test scores for each of the four possible combinations of the binary variables. This is done using the procedure in Key Concept 8.3. Accordingly, the sample average test score for districts with $HiSTR_i = 0$ (low student–teacher ratios) and $HiEL_i = 0$ (low fractions of English learners) is 664.1. For districts with $HiSTR_i = 1$ (high student–teacher ratios) and $HiEL_i = 0$ (low fractions of English learners), the sample average is 662.2 (= 664.1 − 1.9). When $HiSTR_i = 0$ and $HiEL_i = 1$, the sample average is 645.9 (= 664.1 − 18.2), and when $HiSTR_i = 1$ and $HiEL_i = 1$, the sample average is 640.5 (= 664.1 − 1.9 − 18.2 − 3.5).

Interactions Between a Continuous and a Binary Variable

Next consider the population regression of log earnings $[\, Y_i = \ln(Earnings_i)\,]$ against one continuous variable, the individual's years of work experience (X_i), and one binary variable, whether the worker has a college degree $(D_i$, where $D_i = 1$ if the ith person is a college graduate). As shown in Figure 8.8, the population regression line relating Y and the continuous variable X can depend on the binary variable D in three different ways.

In Figure 8.8(a), the two regression lines differ only in their intercept. The corresponding population regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i. \qquad (8.31)$$

This is the familiar multiple regression model with a population regression function that is linear in X_i and D_i. When $D_i = 0$, the population regression function is $\beta_0 + \beta_1 X_i$, so the intercept is β_0 and the slope is β_1. When $D_i = 1$, the population regression function is $\beta_0 + \beta_1 X_i + \beta_2$, so the slope remains β_1 but the intercept is $\beta_0 + \beta_2$. Thus β_2 is the difference between the intercepts of the two regression lines, as shown in Figure 8.8(a). Stated in terms of the earnings example, β_1 is the effect on log earnings of an additional year of work experience, holding college degree status constant, and β_2 is the effect of a college degree on log earnings, holding years of experience constant. In this specification, the effect of an additional year of work experience is the same for college graduates and nongraduates; that is, the two lines in Figure 8.8(a) have the same slope.

In Figure 8.8(b), the two lines have different slopes and intercepts. The different slopes permit the effect of an additional year of work to differ for college graduates and nongraduates. To allow for different slopes, add an interaction term to Equation (8.31):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i, \qquad (8.32)$$

where $X_i \times D_i$ is a new variable, the product of X_i and D_i. To interpret the coefficients of this regression, apply the procedure in Key Concept 8.3. Doing so shows that if

---

FIGURE 8.8   Regression Functions Using Binary and Continuous Variables

(a) Different intercepts, same slope

(b) Different intercepts, different slopes

(c) Same intercept, different slopes
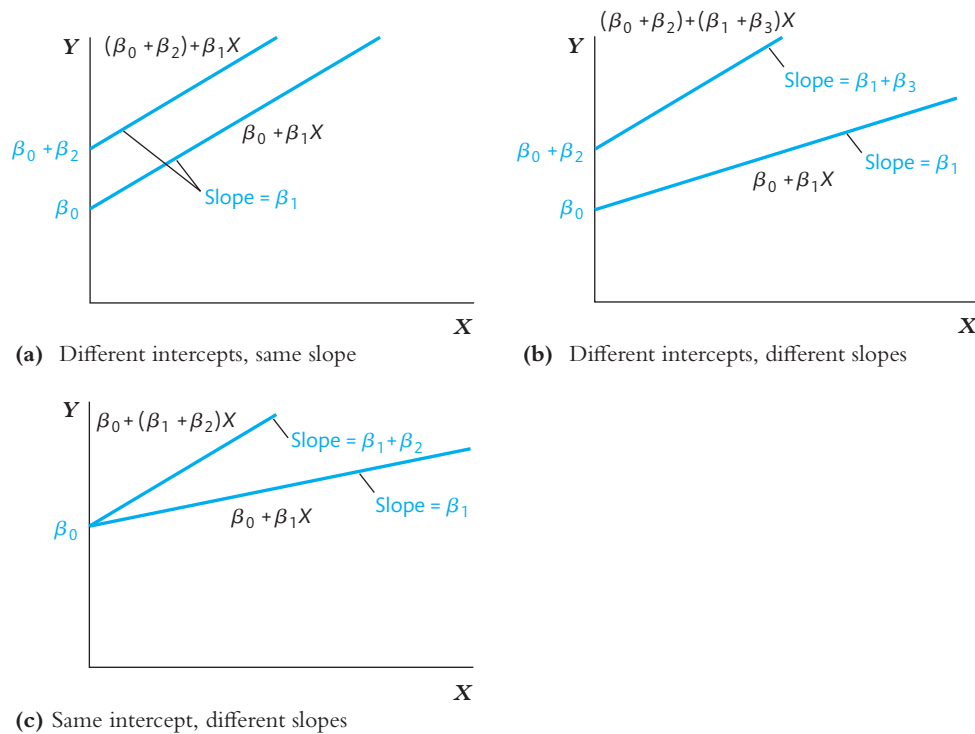
Interactions of binary variables and continuous variables can produce three different population regression functions: (a) $\beta_0 + \beta_1 X + \beta_2 D$ allows for different intercepts but has the same slope, (b) $\beta_0 + \beta_1 X + \beta_2 D + \beta_3(X \times D)$ allows for different intercepts and different slopes, and (c) $\beta_0 + \beta_1 X + \beta_2(X \times D)$ has the same intercept but allows for different slopes.

$D_i = 0$, the population regression function is $\beta_0 + \beta_1 X_i$, whereas if $D_i = 1$, the population regression function is $(\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i$. Thus this specification allows for two different population regression functions relating Y_i and X_i, depending on the value of D_i, as is shown in Figure 8.8(b). The difference between the two intercepts is β_2, and the difference between the two slopes is β_3. In the earnings example, β_1 is the effect of an additional year of work experience for nongraduates ($D_i = 0$), and $\beta_1 + \beta_3$ is this effect for graduates, so β_3 is the *difference* in the effect of an additional year of work experience for college graduates versus that for nongraduates.

A third possibility, shown in Figure 8.8(c), is that the two lines have different slopes but the same intercept. The interacted regression model for this case is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2(X_i \times D_i) + u_i. \tag{8.33}$$

The coefficients of this specification also can be interpreted using Key Concept 8.3. In terms of the earnings example, this specification allows for different effects of

Interactions Between Binary and Continuous Variables

Through the use of the interaction term $X_i \times D_i$, the population regression line relating Y_i and the continuous variable X_i can have a slope that depends on the binary variable D_i. There are three possibilities:

1. Different intercepts, same slope (Figure 8.8a):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i;$$

2. Different intercepts and slopes (Figure 8.8b):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3(X_i \times D_i) + u_i;$$

3. Same intercept, different slopes (Figure 8.8c):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2(X_i \times D_i) + u_i.$$

experience on log earnings between college graduates and nongraduates, but it requires that expected log earnings be the same for both groups when they have no prior experience. Said differently, this specification corresponds to the population mean entry-level wage being the same for college graduates and nongraduates. This does not make much sense in this application, and in practice, this specification is used less frequently than Equation (8.32), which allows for different intercepts and slopes.

All three specifications—Equations (8.31), (8.32), and (8.33)—are versions of the multiple regression model of Chapter 6, and once the new variable $X_i \times D_i$ is created, the coefficients of all three can be estimated by OLS.

The three regression models with a binary and a continuous independent variable are summarized in Key Concept 8.4.

Application to the student–teacher ratio and the percentage of English learners. Does the effect on test scores of cutting the student–teacher ratio depend on whether the percentage of students still learning English is high or low? One way to answer this question is to use a specification that allows for two different regression lines, depending on whether there is a high or a low percentage of English learners. This is achieved using the different intercept/different slope specification:

$$\widehat{TestScore} = 682.2 - 0.97\,STR + 5.6\,HiEL - 1.28(STR \times HiEL),$$
$$\phantom{\widehat{TestScore} = }(11.9)\quad(0.59)\qquad(19.5)\qquad(0.97)$$
$$\overline{R}^2 = 0.305, \tag{8.34}$$

where the binary variable $HiEL_i$ equals 1 if the percentage of students still learning English in the district is greater than 10% and equals 0 otherwise.

For districts with a low fraction of English learners $(HiEL_i = 0)$, the estimated regression line is $682.2 - 0.97STR_i$. For districts with a high fraction of English learners $(HiEL_i = 1)$, the estimated regression line is $682.2 + 5.6 - 0.97STR_i - 1.28STR_i = 687.8 - 2.25STR_i$. According to these estimates, reducing the student–teacher ratio by 1 is predicted to increase test scores by 0.97 points in districts with low fractions of English learners but by 2.25 points in districts with high fractions of English learners. The difference between these two effects, 1.28 points, is the coefficient on the interaction term in Equation (8.34).

The interaction regression model in Equation (8.34) allows us to estimate the effect of more nuanced policy interventions than the across-the-board class size reduction considered so far. For example, suppose the state considered a policy to reduce the student–teacher ratio by 2 in districts with a high fraction of English learners $(HiEL_i = 1)$ but to leave class size unchanged in other districts. Applying the method of Key Concept 8.1 to Equations (8.32) and (8.34) shows that the estimated effect of this reduction for the districts for which $HiEL = 1$ is $-2(\hat{\beta}_1 + \hat{\beta}_3) = 4.50$. The standard error of this estimated effect is $SE(-2\hat{\beta}_1 - 2\hat{\beta}_3) = 1.53$, which can be computed using Equation (8.8) and the methods of Section 7.3.

The OLS regression in Equation (8.34) can be used to test several hypotheses about the population regression line. First, the hypothesis that the two lines are, in fact, the same can be tested by computing the F-statistic testing the joint hypothesis that the coefficient on $HiEL_i$ and the coefficient on the interaction term $STR_i \times HiEL_i$ are both 0. This F-statistic is 89.9, which is significant at the 1% level.

Second, the hypothesis that two lines have the same slope can be tested by testing whether the coefficient on the interaction term is 0. The t-statistic, $-1.28/0.97 = -1.32$, is less than 1.64 in absolute value, so the null hypothesis that the two lines have the same slope cannot be rejected using a two-sided test at the 10% significance level.

Third, the hypothesis that the lines have the same intercept corresponds to the restriction that the population coefficient on $HiEL$ is 0. The t-statistic testing this restriction is $t = 5.6/19.5 = 0.29$, so the hypothesis that the lines have the same intercept cannot be rejected at the 5% level.

These three tests produce seemingly contradictory results: The joint test using the F-statistic rejects the joint hypothesis that the slope and the intercept are the same, but the tests of the individual hypotheses using the t-statistic fail to reject. The reason is that the regressors, $HiEL$ and $STR \times HiEL$, are highly correlated. This results in large standard errors on the individual coefficients. Even though it is impossible to tell which of the coefficients is nonzero, there is strong evidence against the hypothesis that *both* are 0.

Finally, the hypothesis that the student–teacher ratio does not enter this specification can be tested by computing the F-statistic for the joint hypothesis that the coefficients on STR and on the interaction term are both 0. This F-statistic is 5.64, which has a p-value of 0.004. Thus the coefficients on the student–teacher ratio are jointly statistically significant at the 1% significance level.

The Return to Education and the Gender Gap

In addition to its intellectual pleasures, education has economic rewards. As the boxes in Chapters 3 and 5 show, workers with more education tend to earn more than their counterparts with less education. The analysis in those boxes was incomplete, however, for at least three reasons. First, it failed to control for other determinants of earnings that might be correlated with educational achievement, so the OLS estimator of the coefficient on education could have omitted variable bias. Second, the functional form used in Chapter 5—a simple linear relation—implies that earnings change by a constant dollar amount for each additional year of education, whereas one might suspect that the dollar change in earnings is actually larger at higher levels of education. Third, the box in Chapter 5 ignores the sex differences in earnings highlighted in the box in Chapter 3.

These limitations can be addressed by a multiple regression analysis that controls for determinants of earnings that, if omitted, could cause omitted variable bias and that uses a nonlinear functional form relating education and earnings. Table 8.1 summarizes regressions estimated using data on full-time workers, ages 30 through 64, from the Current Population Survey (the CPS data are described in Appendix 3.1). The dependent variable is the logarithm of hourly earnings, so another year of education is associated with a constant percentage increase (not a dollar increase) in earnings.

Table 8.1 has four salient results. First, the omission of sex in regression (1) does not result in substantial omitted variable bias: Even though sex enters regression (2) significantly and with a large coefficient, sex and years of education are nearly uncorrelated: On average, men and women have nearly the same levels of education. Second, the returns to education are economically and statistically significantly different for men and women: In regression (3), the t-statistic testing the hypothesis that they are the same is 3.42

($= 0.006/0.0018$). As the tight confidence intervals attest, the return to education is precisely estimated both for men and for women. Third, regression (4) controls for the region of the country in which the individual lives, thereby addressing potential omitted variable bias that might arise if years of education differ systematically by region. Controlling for region makes a small difference to the estimated coefficients on the education terms relative to those reported in regression (3). Fourth, regression (4) controls for the potential experience of the worker, as measured by years since completion of schooling. The estimated coefficients imply a declining marginal value for each year of potential experience.

The estimated economic return to education in regression (4) is 11.14% for each year of education for men and 11.96% ($= 0.1114 + 0.0082$, in percent) for women. Because the regression functions for men and women have different slopes, the gender gap depends on the years of education. For 12 years of education, the gender gap is estimated to be 27.0% ($= 0.0082 \times 12 - 0.368$, in percent); for 16 years of education, the gender gap is less in percentage terms, 23.7%.

These estimates of the return to education and the gender gap still have limitations, including the possibility of other omitted variables, notably the native ability of the worker, and potential problems associated with the way variables are measured in the CPS. Nevertheless, the estimates in Table 8.1 are consistent with those obtained by economists who carefully address these limitations. A survey by the econometrician David Card (1999) of dozens of empirical studies concludes that labor economists' best estimates of the return to education generally fall between 8% and 11% and that the return depends on the quality of the education. If you are interested in learning more about the economic return to education, see Card (1999).

TABLE 8.1	**The Return to Education and the Gender Gap: Regression Results for the United States in 2015**			

Dependent variable: logarithm of *Hourly earnings*.

Regressor	(1)	(2)	(3)	(4)
Years of education	0.1056 (0.0009)	0.1089 (0.0009)	0.1063 (0.0018)	0.1114 (0.0013)
Female		−0.252 (0.005)	−0.342 (0.026)	−0.368 (0.026)
Female × Years of education			0.0063 (0.0018)	0.0082 (0.0018)
Potential experience				0.0147 (0.0013)
Potential experience²				−0.000183 (0.000024)
a. Regional control variables?	No	No	No	Yes
95% confidence interval for return to education				
Combined men & women	[0.104, 0.107]	[0.107, 0.111]		
For men			[0.104, 0.109]	[0.109, 0.114]
For women			[0.110, 0.115]	[0.117, 0.122]
$\overline{R}^2$	0.209	0.251	0.251	0.262

The data are from the March 2016 Current Population Survey (see Appendix 3.1). The sample size is $n = 47{,}233$ observations for each regression. *Female* is an indicator variable that equals 1 for women and 0 for men. *Potential experience* equals the number of years since completion of schooling. The regional control variables included in regression (4) are *Midwest*, *South*, and *West* which are indicator variables denoting the region of the United States in which the worker lives (the omitted region is *Northeast*). Heteroskedasticity-robust standard errors are reported in parentheses below the estimated coefficients.

Interactions Between Two Continuous Variables

Now suppose that both independent variables (X_{1i} and X_{2i}) are continuous. An example is when Y_i is log earnings of the ith worker, X_{1i} is his or her years of work experience, and X_{2i} is the number of years he or she went to school. If the population regression function is linear, the effect on wages of an additional year of experience does not depend on the number of years of education, or, equivalently, the effect of an additional year of education does not depend on the number of years of work experience. In reality, however, there might be an interaction between these two variables, so that the effect on wages of an additional year of experience depends on the number of years of education. This interaction can be modeled by

augmenting the linear regression model with an interaction term that is the product of X_{1i} and X_{2i}:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i. \qquad (8.35)$$

The interaction term allows the effect of a unit change in X_1 to depend on X_2. To see this, apply the general method for computing effects in nonlinear regression models in Key Concept 8.1. The difference in Equation (8.4), computed for the interacted regression function in Equation (8.35), is $\Delta Y = (\beta_1 + \beta_3 X_2) \Delta X_1$ [Exercise 8.10(a)]. Thus the effect on Y of a change in X_1, holding X_2 constant, is

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2, \qquad (8.36)$$

which depends on X_2. For example, in the earnings example, if β_3 is positive, then the effect on log earnings of an additional year of experience is greater, by the amount β_3, for each additional year of education the worker has.

A similar calculation shows that the effect on Y of a change ΔX_2 in X_2, holding X_1 constant, is $\Delta Y / \Delta X_2 = (\beta_2 + \beta_3 X_1)$.

Putting these two effects together shows that the coefficient β_3 on the interaction term is the effect of a unit increase in X_1 and X_2, above and beyond the sum of the effects of a unit increase in X_1 alone and a unit increase in X_2 alone. That is, if X_1 changes by ΔX_1 and X_2 changes by ΔX_2, then the expected change in Y is $\Delta Y = (\beta_1 + \beta_3 X_2) \Delta X_1 + (\beta_2 + \beta_3 X_1) \Delta X_2 + \beta_3 \Delta X_1 \Delta X_2$ [Exercise 8.10(c)]. The first term is the effect from changing X_1, holding X_2 constant; the second term is the effect from changing X_2, holding X_1 constant; and the final term, $\beta_3 \Delta X_1 \Delta X_2$, is the extra effect from changing both X_1 and X_2.

Interactions between two variables are summarized as Key Concept 8.5.

When interactions are combined with logarithmic transformations, they can be used to estimate price elasticities when the price elasticity depends on the characteristics of the good (see the box "The Demand for Economics Journals" for an example).

KEY CONCEPT

8.5

Interactions in Multiple Regression

The interaction term between the two independent variables X_1 and X_2 is their product $X_1 \times X_2$. Including this interaction term allows the effect on Y of a change in X_1 to depend on the value of X_2 and, conversely, allows the effect of a change in X_2 to depend on the value of X_1.

The coefficient on $X_1 \times X_2$ is the effect of a one-unit increase in X_1 *and* X_2, above and beyond the sum of the individual effects of a unit increase in X_1 alone and a unit increase in X_2 alone. This is true whether X_1 and/or X_2 is continuous or binary.

The Demand for Economics Journals

Professional economists follow the most recent research in their areas of specialization. Most research in economics first appears in economics journals, so economists—or their libraries—subscribe to economics journals.

How elastic is the demand by libraries for economics journals? To find out, we analyzed the relationship between the number of subscriptions to a journal at U.S. libraries (Y_i) and the journal's library subscription price using data for the year 2000 for 180 economics journals. Because the product of a journal is the ideas it contains, its price is logically measured not in dol-

lars per year or dollars per page but instead in dollars per idea. Although we cannot measure "ideas" directly, a good indirect measure is the number of times that articles in a journal are subsequently cited by other researchers. Accordingly, we measure price as the "price per citation" in the journal. The price range is enormous, from $\frac{1}{2}$¢ per citation (the *American Economic Review*) to 20¢ per citation or more. Some journals are expensive per citation because they have few citations and others because their library subscription price per year is very high. In 2017, a library print subscription to the *Journal of Econometrics*

FIGURE 8.9   **Library Subscriptions and Prices of Economics Journals**
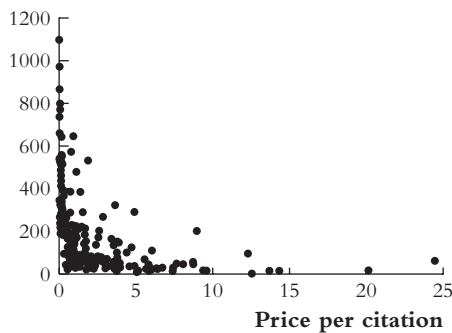
(a)   Subscriptions and price per citation
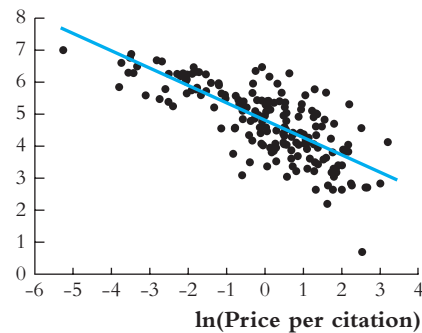
(b)   ln(Subscriptions) and ln(Price per citation)
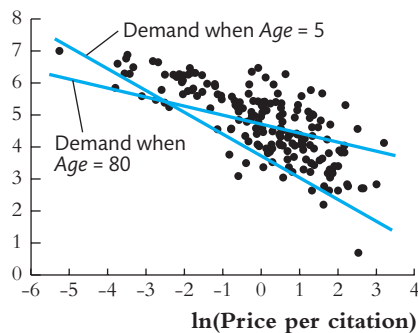
(c)   ln(Subscriptions) and ln(Price per citation)

There is a nonlinear inverse relation between the number of U.S. library subscriptions (quantity) and the library price per citation (price), as shown in Figure 8.9a for 180 economics journals in 2000. But as seen in Figure 8.9b, the relation between log quantity and log price appears to be approximately linear. Figure 8.9c shows that demand is more elastic for young journals (Age = 5) than for old journals (Age = 80).

TABLE 8.2 Estimates of the Demand for Economics Journals				
Dependent variable: logarithm of subscriptions at U.S. libraries in the year 2000; 180 observations.				
Regressor	**(1)**	**(2)**	**(3)**	**(4)**
$\ln(Price\ per\ citation)$	−0.533 (0.034)	−0.408 (0.044)	−0.961 (0.160)	−0.899 (0.145)
$[\ln(Price\ per\ citation)]^2$			0.017 (0.025)	
$[\ln(Price\ per\ citation)]^3$			0.0037 (0.0055)	
$\ln(Age)$		0.424 (0.119)	0.373 (0.118)	0.374 (0.118)
$\ln(Age) \times \ln(Price\ per\ citation)$			0.156 (0.052)	0.141 (0.040)
$\ln(Characters \div 1,000,000)$		0.206 (0.098)	0.235 (0.098)	0.229 (0.096)
F-Statistics and Summary Statistics				
F-statistic testing coefficients on quadratic and cubic terms (p-value)			0.25 (0.779)	
SER	0.750	0.705	0.691	0.688
$\overline{R}^2$	0.555	0.607	0.622	0.626

The F-statistic tests the hypothesis that the coefficients on $[\ln(Price\ per\ citation)]^2$ and $[\ln(Price\ per\ citation)]^3$ are both 0. All regressions include an intercept (not reported in the table). Standard errors are given in parentheses under coefficients, and p-values are given in parentheses under F-statistics.

cost \$5363, compared to only \$940 for a bundled subscription to all eight journals published by the American Economics Association, including the *American Economic Review*!

Because we are interested in estimating elasticities, we use a log-log specification (Key Concept 8.2). The scatterplots in Figures 8.9a and 8.9b provide empirical support for this transformation. Because some of the oldest and most prestigious journals are the cheapest per citation, a regression of log quantity against log price could have omitted variable bias. Our regressions therefore include two control variables: the logarithm of age and the logarithm of the number of characters per year in the journal.

The regression results are summarized in Table 8.2. Those results yield the following conclusions (see if you can find the basis for these conclusions in the table!):

1. Demand is less elastic for older than for newer journals.

2. The evidence supports a linear, rather than a cubic, function of log price.

3. Demand is greater for journals with more characters, holding price and age constant.

So what is the elasticity of demand for economics journals? It depends on the age of the journal. Demand curves for an 80-year-old journal and a 5-year-old upstart are superimposed on the scatterplot in Figure 8.9c; the older journal's demand elasticity is -0.28 ($SE = 0.06$), while the younger journal's is -0.67($SE = 0.08$).

This demand is very inelastic: Demand is very insensitive to price, especially for older journals. For libraries, having the most recent research on hand is a necessity, not a luxury. By way of comparison, experts estimate the demand elasticity for cigarettes to be in the range of -0.3 to -0.5. Economics journals are, it seems, as addictive as cigarettes but a lot better for your health![1]

---

[1]These data were graciously provided by Professor Theodore Bergstrom of the Department of Economics at the University of California, Santa Barbara. If you are interested in learning more about the economics of economics journals, see Bergstrom (2001).

Application to the student–teacher ratio and the percentage of English learners. The previous examples considered interactions between the student–teacher ratio and a binary variable indicating whether the percentage of English learners is large or small. A different way to study this interaction is to examine the interaction between the student–teacher ratio and the continuous variable, the percentage of English learners ($PctEL$). The estimated interaction regression is

$$\widehat{TestScore} = 686.3 - 1.12\,STR - 0.67PctEL + 0.0012(STR \times PctEL),$$
$$\qquad\quad (11.8)\quad (0.59)\qquad (0.37)\qquad\quad (0.019)$$
$$\overline{R}^2 = 0.422. \tag{8.37}$$

When the percentage of English learners is at the median ($PctEL = 8.85$), the slope of the line relating test scores and the student–teacher ratio is estimated to be -1.11 ($= -1.12 + 0.0012 \times 8.85$). When the percentage of English learners is at the 75th percentile ($PctEL = 23.0$), this line is estimated to be slightly flatter, with a slope of -1.09 ($= -1.12 + 0.0012 \times 23.0$). That is, for a district with 8.85% English learners, the estimated effect of a one-unit reduction in the student–teacher ratio is to increase test scores by 1.11 points, but for a district with 23.0% English learners, reducing the student–teacher ratio by one unit is predicted to increase test scores by only 1.09 points. The difference between these estimated effects is not statistically significant, however: The t-statistic testing whether the coefficient on the interaction term is 0 is $t = 0.0012/0.019 = 0.06$, which is not significant at the 10% level.

To keep the discussion focused on nonlinear models, the specifications in Sections 8.1 through 8.3 exclude additional control variables such as the students' economic background. Consequently, these results arguably are subject to omitted variable bias. To draw substantive conclusions about the effect on test scores of reducing the student–teacher ratio, these nonlinear specifications must be augmented with control variables, and it is to such an exercise that we now turn.

8.4 Nonlinear Effects on Test Scores of the Student–Teacher Ratio

This section addresses three specific questions about test scores and the student–teacher ratio. First, after controlling for differences in economic characteristics of different districts, does the effect on test scores of reducing the student–teacher ratio depend on the fraction of English learners? Second, does this effect depend on the value of the student–teacher ratio? Third, and most important, after taking economic factors and nonlinearities into account, what is the estimated effect on test scores of reducing the student–teacher ratio by two students per teacher, as our superintendent from Chapter 4 proposes to do?

We answer these questions by considering nonlinear regression specifications of the type discussed in Sections 8.2 and 8.3, extended to include two measures of the economic background of the students: the percentage of students eligible for a subsidized lunch and the logarithm of average district income. The logarithm of district income is used because the empirical analysis of Section 8.2 suggests that this specification captures the nonlinear relationship between test scores and district income. As in Section 7.6, we do not include expenditures per pupil as a regressor, and in so doing, we are considering the effect of decreasing the student–teacher ratio, while allowing expenditures per pupil to increase (that is, we are not holding expenditures per pupil constant).

Discussion of Regression Results

The OLS regression results are summarized in Table 8.3. The columns labeled (1) through (7) each report separate regressions. The entries in the table are the coefficients, standard errors, certain F-statistics and their p-values, and summary statistics, as indicated by the description in each row. In addition, the middle block presents 95% confidence intervals for the estimated effect of reducing the class size by two, the question asked by the superintendent. Because some of the specifications are nonlinear, the confidence intervals are worked out for various cases, including reducing the size of a larger class (22 to 20) or of a moderately-sized class (20 to 18), and for the case of high or low fractions of English learners, where the specific cases depend on the specifications.

The first column of regression results, labeled regression (1) in the table, is regression (3) in Table 7.1 repeated here for convenience. This regression does not control for district income, so the first thing we do is check whether the results change substantially when log income is included as an additional economic control variable. The results are given in regression (2) in Table 8.3. The log of income is statistically significant at the 1% level, and the coefficient on the student–teacher ratio becomes somewhat closer to 0, falling from -1.00 to -0.73, although it remains statistically significant at the 1% level. The change in the coefficient on STR is large enough

TABLE 8.3	Nonlinear Regression Models of Test Scores

Dependent variable: average test score in district; 420 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student–teacher ratio (STR)	−1.00 (0.27)	−0.73 (0.26)	−0.97 (0.59)	−0.53 (0.34)	64.33 (24.86)	83.70 (28.50)	65.29 (25.26)
STR^2					−3.42 (1.25)	−4.38 (1.44)	−3.47 (1.27)
STR^3					0.059 (0.021)	0.075 (0.024)	0.060 (0.021)
% English learners	−0.122 (0.033)	−0.176 (0.034)					−0.166 (0.034)
% English learners ≥ 10%? (Binary, $HiEL$)			5.64 (19.51)	5.50 (9.80)	−5.47 (1.03)	816.1 (327.7)	
$HiEL \times STR$			−1.28 (0.97)	−0.58 (0.50)		−123.3 (50.2)	
$HiEL \times STR^2$						6.12 (2.54)	
$HiEL \times STR^3$						−0.101 (0.043)	
Included Economic Control Variables							
% eligible for subsidized lunch	Y	Y	N	Y	Y	Y	Y
Average district income (logarithm)	N	Y	N	Y	Y	Y	Y
95% Confidence Intervals for the Effect of Reducing STR by 2							
No $HiEL$ interaction	[0.93,3.06]	[0.46,2.48]					
22 to 20					[0.61, 3.25]		[0.54, 3.26]
20 to 18					[1.64, 4.36]		[1.55, 4.30]
$HiEL = 0$			[−0.38, 4.25]	[−0.28, 2.41]			
22 to 20						[0.40, 3.98]	
20 to 18						[1.22, 4.99]	
$HiEL = 1$			[1.48, 7.50]	[0.80, 3.63]			
22 to 20						[−0.98, 2.91]	
20 to 18						[−0.72, 4.01]	
F-Statistics and p-Values on Joint Hypotheses							
All STR variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (< 0.001)	4.96 (< 0.001)	5.91 (0.001)
STR^2, $STR^3 = 0$					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)

continued on next page

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$HiEL \times STR, HiEL \times STR^2,$ $HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\overline{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Regressions include an intercept and the economic control variables indicated by "Y" or exclude them if indicated by "N" (coefficients not shown in the table). Standard errors are given in parentheses under coefficients, and p-values are given in parentheses under F-statistics.

between regressions (1) and (2) to warrant additionally controlling for the logarithm of income in the remaining regressions as a deterrent to omitted variable bias.

Regression (3) in Table 8.3 is the interacted regression in Equation (8.34) with the binary variable for a high or low percentage of English learners but with no economic control variables. When the economic control variables (percentage eligible for subsidized lunch and log income) are added [regression (4) in the table], the class size effect is reduced for both high and low English learner classes; however, the confidence intervals are wide in both cases in both regressions. Based on the evidence in regression (4), the hypothesis that the effect of *STR* is the same for districts with low and high percentages of English learners cannot be rejected at the 5% level (the *t*-statistic is $t = -0.58/0.50 = -1.16$).

Regression (5) examines whether the effect of changing the student–teacher ratio depends on the value of the student–teacher ratio by including a cubic specification in *STR,* controlling for the economic variables in regression (4) [the interaction term, $HiEL \times STR$, is not included in regression (5) because it was not significant in regression (4) at the 10% level]. The estimates in regression (5) are consistent with the student–teacher ratio having a nonlinear effect. The null hypothesis that the relationship is linear is rejected at the 1% significance level against the alternative that it is a polynomial up to degree 3 (the *F*-statistic testing the hypothesis that the true coefficients on STR^2 and STR^3 are 0 is 6.17, with a *p*-value of < 0.001). The effect of reducing the class size from 20 to 18 is estimated to be greater than if it is reduced from 22 to 20.

Regression (6) further examines whether the effect of the student–teacher ratio depends not just on the value of the student–teacher ratio but also on the fraction of English learners. By including interactions between *HiEL* and *STR*, STR^2, and STR^3, we can check whether the (possibly cubic) population regressions functions relating test scores and *STR* are different for low and high percentages of English learners. To do so, we test the restriction that the coefficients on the three interaction terms are 0. The resulting *F*-statistic is 2.69, which has a *p*-value of 0.046 and thus is significant at the 5% but not at the 1% significance level. This provides tentative evidence that the regression functions are different for districts with high and low percentages of English learners; however, comparing regressions (6) and (4) makes it clear that

FIGURE 8.10   Three Regression Functions Relating Test Scores and Student–Teacher Ratio

The cubic regressions from columns (5) and (7) of Table 8.3 are nearly identical. They indicate a small amount of nonlinearity in the relation between test scores and student–teacher ratio.
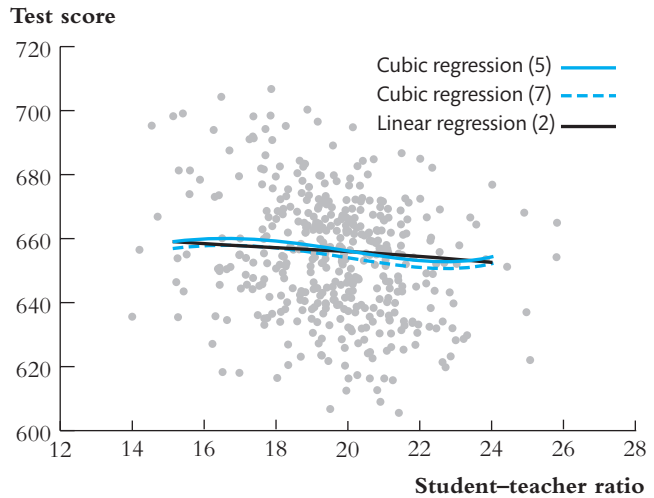
these differences are associated with the quadratic and cubic terms. Moreover, the confidence intervals are quite wide in all cases for regression (6).

Regression (7) is a modification of regression (5), in which the continuous variable *PctEL* is used instead of the binary variable *HiEL* to control for the percentage of English learners in the district. The coefficients on the other regressors do not change substantially when this modification is made, indicating that the results in regression (5) are not sensitive to what measure of the percentage of English learners is actually used in the regression.

In all the specifications, the hypothesis that the student–teacher ratio does not enter the regressions is rejected at the 1% level.

The nonlinear specifications in Table 8.3 are most easily interpreted graphically. Figure 8.10 graphs the estimated regression functions relating test scores and the student–teacher ratio for the linear specification (2) and the cubic specifications (5) and (7), along with a scatterplot of the data.[4] These estimated regression functions show the predicted value of test scores as a function of the student–teacher ratio, holding fixed other values of the independent variables in the regression. The estimated regression functions are all close to one another, although the cubic regressions flatten out for large values of the student–teacher ratio.

Regression (6) suggests that the cubic regression functions relating test scores and *STR* might depend on whether the percentage of English learners in the district is large or small. Figure 8.11 graphs these two estimated regression functions so that

---

[4]For each curve, the predicted value was computed by setting each independent variable, other than *STR*, to its sample average value and computing the predicted value by multiplying these fixed values of the independent variables by the respective estimated coefficients from Table 8.3. This was done for various values of *STR*, and the graph of the resulting adjusted predicted values is the estimated regression function relating test scores and the *STR*, holding the other variables constant at their sample averages.

FIGURE 8.11 Regression Functions for Districts with High and Low Percentages of English Learners

Districts with low percentages of English learners (*HiEL* = 0) are shown by gray dots, and districts with *HiEL* = 1 are shown by colored dots. The cubic regression function for *HiEL* = 1 from regression (6) in Table 8.3 is approximately 10 points below the cubic regression function for *HiEL* = 0 for $17 \le STR \le 23$, but otherwise the two functions have similar shapes and slopes in this range. The slopes of the regression functions differ most for very large and small values of *STR*, for which there are few observations.

---
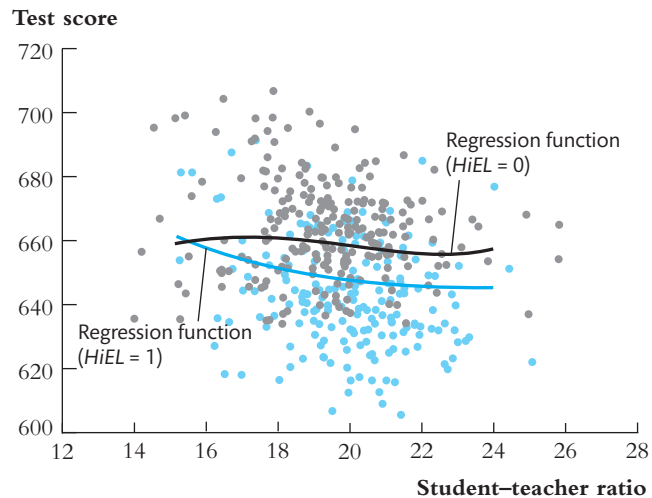
we can see whether this difference, in addition to being statistically significant, is of practical importance. As Figure 8.11 shows, for student–teacher ratios between 17 and 23—a range that includes 88% of the observations—the two functions are separated by approximately 10 points but otherwise are very similar; that is, for *STR* between 17 and 23, districts with a lower percentage of English learners do better, holding constant the student–teacher ratio, but the effect of a change in the student–teacher ratio is essentially the same for the two groups. The two regression functions are different for student–teacher ratios below 16.5, but we must be careful not to read more into this than is justified. The districts with $STR < 16.5$ constitute only 6% of the observations, so the differences between the nonlinear regression functions are reflecting differences in these very few districts with very low student–teacher ratios. Thus, based on Figure 8.11, we conclude that the effect on test scores of a change in the student–teacher ratio does not depend on the percentage of English learners for the range of student–teacher ratios for which we have the most data.

Summary of Findings

These results let us answer the three questions raised at the start of this section.

First, after controlling for economic background, there is at most weak evidence that the effect of a class size reduction depends on whether there are many or few English learners in the district. While a class size reduction is estimated to be more effective in districts with a high fraction of English learners, the difference in effects between high and low English learner districts is imprecisely estimated. Moreover, as shown in Figure 8.11, the estimated regression functions have similar slopes in the range of student–teacher ratios containing most of the data.

Second, after controlling for economic background, there is evidence of a non-linear effect on test scores of the student–teacher ratio. The nonlinear estimates suggest that the effect of reducing the student-teacher ratio is greatest in moderately sized classes and is less for very small or very large classes. The null hypothesis of linearity can be rejected at the 1% level.

Third, we now can return to the superintendent's problem that opened Chapter 4. She wants to know the effect on test scores of reducing the student–teacher ratio by two students per teacher. In the linear specification (2), this effect does not depend on the student–teacher ratio itself, and the estimated effect of this reduction is to improve test scores by $1.46 \ (= -0.73 \times -2)$ points. In the nonlinear specifications, this effect depends on the value of the student–teacher ratio. If her district currently has a student–teacher ratio of 20 and she is considering cutting it to 18, then based on regression (5), the estimated effect of this reduction is to improve test scores by 3.00 points, with a 95% confidence interval of (1.64, 4.36). If her district currently has a student–teacher ratio of 22 and she is considering cutting it to 20, then based on regression (5), the estimated effect of this reduction is to improve test scores by 1.93 points, with a 95% confidence interval of (0.61, 3.25). [Similar results obtain from regression (7).] These estimates from the nonlinear specifications thus allow a more nuanced answer to her question, based on the characteristics of her district.

8.5   Conclusion

This chapter presented several ways to model nonlinear regression functions. Because these models are variants of the multiple regression model, the unknown coefficients can be estimated by OLS, and hypotheses about their values can be tested using t- and F-statistics as described in Chapter 7. In these models, the expected effect on Y of a change in one of the independent variables, X_1, holding the other independent variables $X_2, \ldots, X_k$ constant, in general, depends on the values of $X_1, X_2, \ldots, X_k$.

There are many different models in this chapter, and you could not be blamed for being a bit bewildered about which to use in a given application. How should you analyze possible nonlinearities in practice? Section 8.1 laid out a general approach for such an analysis, but this approach requires you to make decisions and exercise judgment along the way. It would be convenient if there were a single recipe you could follow that would always work in every application, but in practice data analysis is rarely that simple.

The single most important step in specifying nonlinear regression functions is to "use your head." Before you look at the data, can you think of a reason, based on economic theory or expert judgment, why the slope of the population regression function might depend on the value of that, or another, independent variable? If so, what sort of dependence might you expect? And, most important, which nonlinearities (if any) could have major implications for the substantive issues addressed by your study? Answering these questions carefully will focus your analysis. In the test score application, for example, such reasoning led us to investigate whether hiring more teachers might have a greater effect

in districts with a large percentage of students still learning English, perhaps because those students would differentially benefit from more personal attention. By making the question precise, we were able to find a precise answer: After controlling for the economic background of the students, the estimated effect of reducing class size effectively does not depend on whether there are many or few English learners in the class.

Summary

1. In a nonlinear regression, the slope of the population regression function depends on the value of one or more of the independent variables.
2. The effect on Y of a change in the independent variable(s) can be computed by evaluating the regression function at two values of the independent variable(s). The procedure is summarized in Key Concept 8.1.
3. A polynomial regression includes powers of X as regressors. A quadratic regression includes X and X^2, and a cubic regression includes X, X^2, and X^3.
4. Small changes in logarithms can be interpreted as proportional or percentage changes in a variable. Regressions involving logarithms are used to estimate proportional changes and elasticities.
5. The product of two variables is called an interaction term. When interaction terms are included as regressors, they allow the regression slope of one variable to depend on the value of another variable.

Key Terms

quadratic regression model (238)
nonlinear regression function (240)
polynomial regression model (244)
cubic regression model (245)
elasticity (247)
exponential function (247)
natural logarithm (247)
linear-log model (248)

log-linear model (249)
log-log model (251)
interaction term (256)
interacted regressor (256)
interaction regression model (256)
nonlinear least squares (285)
nonlinear least squares
    estimators (285)

Review the Concepts

8.1 Sketch a regression function that is increasing (has a positive slope) and is steep for small values of X but less steep for large values of X. Explain how you would specify a nonlinear regression to model this shape. Can you think of an economic relationship with a shape like this?

8.2 A Cobb–Douglas production function relates production (Q) to factors of production—capital (K), labor (L), and raw materials (M)—and an error term u using the equation $Q = \lambda K^{\beta_1} L^{\beta_2} M^{\beta_3} e^u$, where λ, β_1, β_2, and β_3 are production parameters. Suppose you have data on production and the factors of production from a random sample of firms with the same Cobb–Douglas production function. How would you use regression analysis to estimate the production parameters?

8.3 Can you use $\overline{R}^2$ to compare the fit of a log-log and a log-linear regression? Why or why not? Can you use $\overline{R}^2$ to compare the fit of a log-log and a linear-log regression? Why or why not?

8.4 Suppose the regression in Equation (8.30) is estimated using $LoSTR$ and $LoEL$ in place of $HiSTR$ and $HiEL$, where $LoSTR = 1 - HiSTR$ is an indicator for a low-class-size district and $LoEL = 1 - HiEL$ is an indicator for a district with a low percentage of English learners. What are the values of the estimated regression coefficients?

8.5 Suppose that in Exercise 8.2 you thought that the value of β_2 was not constant but rather increased when K increased. How could you use an interaction term to capture this effect?

8.6 You have estimated a linear regression model relating Y to X. Your professor says, "I think that the relationship between Y and X is nonlinear." Explain how you would test the adequacy of your linear regression.

Exercises

8.1 Sales in a company are $196 million in 2013 and increase to $198 million in 2014.

a. Compute the percentage increase in sales, using the usual formula $100 \times \frac{(Sales_{2014} - Sales_{2013})}{Sales_{2013}}$. Compare this value to the approximation $100 \times [\ln(Sales_{2014}) - \ln(Sales_{2013})]$.

b. Repeat (a), assuming that $Sales_{2014} = 205$, $Sales_{2014} = 250$, and $Sales_{2014} = 500$.

c. How good is the approximation when the change is small? Does the quality of the approximation deteriorate as the percentage change increases?

8.2   Suppose a researcher collects data on houses that have sold in a particular neighborhood over the past year and obtains the regression results in the table shown below.

   a. Using the results in column (1), what is the expected change in price of building a 500-square-foot addition to a house? Construct a 95% confidence interval for the percentage change in price.

   b. Comparing columns (1) and (2), is it better to use *Size* or ln(*Size*) to explain house prices?

   c. Using column (2), what is the estimated effect of a pool on price? (Make sure you get the units right.) Construct a 95% confidence interval for this effect.

   d. The regression in column (3) adds the number of bedrooms to the regression. How large is the estimated effect of an additional bedroom? Is the effect statistically significant? Why do you think the estimated effect is so small? (*Hint:* Which other variables are being held constant?)

   e. Is the quadratic term $\ln(Size)^2$ important?

   f. Use the regression in column (5) to compute the expected change in price when a pool is added to a house that doesn't have a view. Repeat the exercise for a house that has a view. Is there a large difference? Is the difference statistically significant?

8.3   After reading this chapter's analysis of test scores and class size, an educator comments, "In my experience, student performance depends on class size, but not in the way your regressions say. Rather, students do well when class size is less than 20 students and do very poorly when class size is greater than 25. There are no gains from reducing class size below 20 students, the relationship is constant in the intermediate region between 20 and 25 students, and there is no loss to increasing class size when it is already greater than 25." The educator is describing a *threshold effect*, in which performance is constant for class sizes less than 20, jumps and is constant for class sizes between 20 and 25, and then jumps again for class sizes greater than 25. To model these threshold effects, define the binary variables

$$STRsmall = 1 \text{ if } STR < 20, \text{ and } STRsmall = 0 \text{ otherwise;}$$

$$STRmoderate = 1 \text{ if } 20 \leq STR \leq 25, \text{ and } STRmoderate = 0 \text{ otherwise; and}$$

$$STRlarge = 1 \text{ if } STR > 25, \text{ and } STRlarge = 0 \text{ otherwise.}$$

   a. Consider the regression $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRlarge_i + u_i$. Sketch the regression function relating *TestScore* to *STR* for hypothetical values of the regression coefficients that are consistent with the educator's statement.

Regression Results for Exercise 8.2

Dependent variable: ln(Price)

Regressor	(1)	(2)	(3)	(4)	(5)
Size	0.00042 (0.000038)				
ln(*Size*)		0.69 (0.054)	0.68 (0.087)	0.57 (2.03)	0.69 (0.055)
$[\ln(Size)]^2$				0.0078 (0.14)	
Bedrooms			0.0036 (0.037)		
Pool	0.082 (0.032)	0.071 (0.034)	0.071 (0.034)	0.071 (0.036)	0.071 (0.035)
View	0.037 (0.029)	0.027 (0.028)	0.026 (0.026)	0.027 (0.029)	0.027 (0.030)
Pool × *View*					0.0022 (0.10)
Condition	0.13 (0.045)	0.12 (0.035)	0.12 (0.035)	0.12 (0.036)	0.12 (0.035)
Intercept	10.97 (0.069)	6.60 (0.39)	6.63 (0.53)	7.02 (7.50)	6.60 (0.40)
Summary Statistics					
SER	0.102	0.098	0.099	0.099	0.099
$\overline{R}^2$	0.72	0.74	0.73	0.73	0.73

Variable definitions: *Price* = sale price ($); *Size* = house size (in square feet); *Bedrooms* = number of bedrooms; *Pool* = binary variable (1 if house has a swimming pool, 0 otherwise); *View* = binary variable (1 if house has a nice view, 0 otherwise); *Condition* = binary variable (1 if real estate agent reports house is in excellent condition, 0 otherwise).

b. A researcher tries to estimate the regression $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRmoderate_i + \beta_3 STRlarge_i + u_i$ and finds that the software gives an error message. Why?

8.4 Read the box "The Return to Education and the Gender Gap" in Section 8.3.

a. Consider a male with 16 years of education and 2 years of experience. Use the results from column (4) of Table 8.1 and the method in Key Concept 8.1 to estimate the expected change in the logarithm of average hourly earnings (*AHE*) associated with an additional year of experience.

b. Explain why your answer to (a) does not depend on the region he is from.

c. Repeat (a), assuming 10 years of experience.

    d. Explain why the answers to (a) and (b) are different.

    e. Is the difference in the answers to (a) and (b) statistically significant at the 5% level? Explain.

    f. Would your answers to (a) through (d) change if the person were female? Explain.

    g. How would you change the regression if you suspected that the effect of experience on earnings was different for men than for women?

8.5 Read the box "The Demand for Economics Journals" in Section 8.3.

    a. The box reaches three conclusions. Looking at the results in the table, what is the basis for each of these conclusions?

    b. Using the results in regression (4), the box reports that the elasticity of demand for an 80-year-old journal is −0.28.

        i. How was this value determined from the estimated regression?

        ii. The box reports that the standard error for the estimated elasticity is 0.06. How would you calculate this standard error? (*Hint:* See the discussion in "Standard errors of estimated effects" on page 242.)

    c. Suppose the variable *Characters* had been divided by 1000 instead of 1,000,000. How would the results in column (4) change?

8.6 Refer to Table 8.3.

    a. A researcher suspects that the effect of *%Eligible for subsidized lunch* has a nonlinear effect on test scores. In particular, he conjectures that increases in this variable from 10% to 20% have little effect on test scores but that changes from 50% to 60% have a much larger effect.

        i. Describe a nonlinear specification that can be used to model this form of nonlinearity.

        ii. How would you test whether the researcher's conjecture was better than the linear specification in column (7) of Table 8.3?

    b. A researcher suspects that the effect of income on test scores is different in districts with small classes than in districts with large classes.

        i. Describe a nonlinear specification that can be used to model this form of nonlinearity.

        ii. How would you test whether the researcher's conjecture was better than the linear specification in column (7) of Table 8.3?

8.7 This problem is inspired by a study of the gender gap in earnings in top corporate jobs (Bertrand and Hallock, 2001). The study compares total compensation among top executives in a large set of U.S. public corporations in the 1990s. (Each year these publicly traded corporations must report total compensation levels for their top five executives.)

a. Let *Female* be an indicator variable that is equal to 1 for females and 0 for males. A regression of the logarithm of earnings on *Female* yields

$$\widehat{\ln{(Earnings)}} = 6.48 - 0.44\ Female,\ SER = 2.65.$$
$$(0.01)\ \ (0.05)$$

  i. The estimated coefficient on *Female* is −0.44. Explain what this value means.

  ii. The *SER* is 2.65. Explain what this value means.

  iii. Does this regression suggest that female top executives earn less than top male executives? Explain.

  iv. Does this regression suggest that there is sex discrimination? Explain.

b. Two new variables, the market value of the firm (a measure of firm size, in millions of dollars) and stock return (a measure of firm performance, in percentage points), are added to the regression:

$$\widehat{\ln{(Earnings)}} = 3.86 - 0.28\ Female + 0.37\ln{(MarketValue)} + 0.004\ Return,$$
$$(0.03)\ \ \ (0.04)\ \ \ \ \ \ \ \ \ (0.004)\ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ (0.003)$$
$$n = 46{,}670,\ \overline{R}^2 = 0.345.$$

  i. The coefficient on ln(*MarketValue*) is 0.37. Explain what this value means.

  ii. The coefficient on *Female* is now −0.28. Explain why it has changed from the regression in (a).

c. Are large firms more likely than small firms to have female top executives? Explain.

8.8 X is a continuous variable that takes on values between 5 and 100. Z is a binary variable. Sketch the following regression functions (with values of X between 5 and 100 on the horizontal axis and values of $\hat{Y}$ on the vertical axis):

a. $\hat{Y} = 2.0 + 3.0 \times \ln(X)$.

b. $\hat{Y} = 2.0 - 3.0 \times \ln(X)$.

c. i. $\hat{Y} = 2.0 + 3.0 \times \ln(X) + 4.0Z$, with $Z = 1$.

  ii. Same as (i), but with $Z = 0$.

d. i. $\hat{Y} = 2.0 + 3.0 \times \ln(X) + 4.0Z - 1.0 \times Z \times \ln(X)$, with $Z = 1$.

  ii. Same as (i), but with $Z = 0$.

e. $\hat{Y} = 1.0 + 125.0X - 0.01X^2$.

8.9 Explain how you would use approach 2 from Section 7.3 to calculate the confidence interval discussed below Equation (8.8). [*Hint:* This requires estimating

a new regression using a different definition of the regressors and the dependent variable. See Exercise (7.9).]

8.10 Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3(X_{1i} \times X_{2i}) + u_i$. Use Key Concept 8.1 to show that

a. $\Delta Y/\Delta X_1 = \beta_1 + \beta_3 X_2$ (effect of change in X_1, holding X_2 constant).

b. $\Delta Y/\Delta X_2 = \beta_2 + \beta_3 X_1$ (effect of change in X_2, holding X_1 constant).

c. If X_1 changes by ΔX_1 and X_2 changes by ΔX_2, then $\Delta Y = (\beta_1 + \beta_3 X_2)\Delta X_1 + (\beta_2 + \beta_3 X_1)\Delta X_2 + \beta_3 \Delta X_1 \Delta X_2$.

8.11 Derive the expressions for the elasticities given in Appendix 8.2 for the linear and log-log models. (*Hint:* For the log-log model, assume that u and X are independent, as is done in Appendix 8.2 for the log-linear model.)

8.12 The discussion following Equation (8.28) interprets the coefficient on interacted binary variables using the conditional mean zero assumption. This exercise shows that this interpretation also applies under conditional mean independence. Consider the hypothetical experiment in Exercise 7.11.

a. Suppose you estimate the regression $Y_i = \gamma_0 + \gamma_1 X_{1i} + u_i$ using only the data on returning students. Show that γ_1 is the class size effect for returning students—that is, that $\gamma_1 = E(Y_i | X_{1i} = 1, X_{2i} = 0) - E(Y_i | X_{1i} = 0, X_{2i} = 0)$. Explain why $\hat{\gamma}_1$ is an unbiased estimator of γ_1.

b. Suppose you estimate the regression $Y_i = \delta_0 + \delta_1 X_{1i} + u_i$ using only the data on new students. Show that δ_1 is the class size effect for new students—that is, that $\delta_1 = E(Y_i | X_{1i} = 1, X_{2i} = 1) - E(Y_i | X_{1i} = 0, X_{2i} = 1)$. Explain why $\hat{\delta}_1$ is an unbiased estimator of δ_1.

c. Consider the regression for both returning and new students, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3(X_{1i} \times X_{2i}) + u_i$. Use the conditional mean independence assumption $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$ to show that $\beta_1 = \gamma_1$, $\beta_1 + \beta_3 = \delta_1$, and $\beta_3 = \delta_1 - \gamma_1$ (the difference in the class size effects).

d. Suppose you estimate the interaction regression in (c) using the combined data and $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$. Show that $\hat{\beta}_1$ and $\hat{\beta}_3$ are unbiased but that $\hat{\beta}_2$ is, in general, biased.

Empirical Exercises

E8.1 Lead is toxic, particularly for young children, and for this reason, government regulations severely restrict the amount of lead in our environment. But this was not always the case. In the early part of the 20th century, the underground water pipes in many U.S. cities contained lead, and lead from these pipes leached into drinking water. In this exercise, you will investigate the

effect of these lead water pipes on infant mortality. On the text website **http://www.pearsonhighered.com/stock_watson/**, you will find the data file **Lead_Mortality**, which contains data on infant mortality, type of water pipes (lead or nonlead), water acidity (pH), and several demographic variables for 172 U.S. cities in 1900.[5] A detailed description is given in **Lead_Mortality_Description**, also available on the website.

a. Compute the average infant mortality rate (*Inf*) for cities with lead pipes and for cities with nonlead pipes. Is there a statistically significant difference in the averages?

b. The amount of lead leached from lead pipes depends on the chemistry of the water running through the pipes. The more acidic the water is (that is, the lower its pH), the more lead is leached. Run a regression of *Inf* on *Lead*, *pH*, and the interaction term *Lead* $\times$ *pH*.

   i. The regression includes four coefficients (the intercept and the three coefficients multiplying the regressors). Explain what each coefficient measures.

   ii. Plot the estimated regression function relating *Inf* to *pH* for *Lead* $= 0$ and for *Lead* $= 1$. Describe the differences in the regression functions, and relate these differences to the coefficients you discussed in (i).

   iii. Does *Lead* have a statistically significant effect on infant mortality? Explain.

   iv. Does the effect of *Lead* on infant mortality depend on *pH*? Is this dependence statistically significant?

   v. What is the average value of *pH* in the sample? At this *pH* level, what is the estimated effect of *Lead* on infant mortality? What is the standard deviation of *pH*? Suppose the *pH* level is one standard deviation lower than the average level of *pH* in the sample: What is the estimated effect of *Lead* on infant mortality? What if *pH* is one standard deviation higher than the average value?

   vi. Construct a 95% confidence interval for the effect of *Lead* on infant mortality when *pH* $= 6.5$.

c. The analysis in (b) may suffer from omitted variable bias because it neglects factors that affect infant mortality and that might potentially be correlated with *Lead* and *pH*. Investigate this concern, using the other variables in the data set.

E8.2 On the text website **http://www.pearsonhighered.com/stock_watson/**, you will find a data file **CPS2015**, which contains data for full-time, full-year workers,

---

[5]These data were provided by Professor Karen Clay of Carnegie Mellon University and were used in her paper with Werner Troesken and Michael Haines, "Lead and Mortality," *Review of Economics and Statistics*, 2014, 96(3).

ages 25–34, with a high school diploma or B.A./B.S. as their highest degree. A detailed description is given in **CPS2015_Description**, also available on the website. (These are the same data as in **CPS96_15**, used in Empirical Exercise 3.1, but are limited to the year 2015.) In this exercise, you will investigate the relationship between a worker's age and earnings. (Generally, older workers have more job experience, leading to higher productivity and higher earnings.)

a. Run a regression of average hourly earnings (*AHE*) on age (*Age*), sex (*Female*), and education (*Bachelor*). If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

b. Run a regression of the logarithm of average hourly earnings, ln(*AHE*), on *Age*, *Female*, and *Bachelor*. If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

c. Run a regression of the logarithm of average hourly earnings, ln(*AHE*), on ln(*Age*), *Female*, and *Bachelor*. If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

d. Run a regression of the logarithm of average hourly earnings, ln(*AHE*), on *Age*, Age^2, *Female*, and *Bachelor*. If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

e. Do you prefer the regression in (c) to the regression in (b)? Explain.

f. Do you prefer the regression in (d) to the regression in (b)? Explain.

g. Do you prefer the regression in (d) to the regression in (c)? Explain.

h. Plot the regression relation between *Age* and ln(*AHE*) from (b), (c), and (d) for males with a high school diploma. Describe the similarities and differences between the estimated regression functions. Would your answer change if you plotted the regression function for females with college degrees?

i. Run a regression of ln(*AHE*) on *Age*, Age^2, *Female*, *Bachelor*, and the interaction term *Female* $\times$ *Bachelor*. What does the coefficient on the interaction term measure? Alexis is a 30-year-old female with a bachelor's degree. What does the regression predict for her value of ln(*AHE*)? Jane is a 30-year-old female with a high school diploma. What does the regression predict for her value of ln(*AHE*)? What is the predicted difference between Alexis's and Jane's earnings? Bob is a 30-year-old male with a bachelor's degree. What does the regression predict for his value of ln(*AHE*)? Jim is a 30-year-old male with a high school diploma. What does the regression predict for his value of ln(*AHE*)? What is the predicted difference between Bob's and Jim's earnings?

    j. Is the effect of *Age* on earnings different for men than for women? Specify and estimate a regression that you can use to answer this question.

    k. Is the effect of *Age* on earnings different for high school graduates than for college graduates? Specify and estimate a regression that you can use to answer this question.

    l. After running all these regressions (and any others that you want to run), summarize the effect of age on earnings for young workers.

APPENDIX

8.1 Regression Functions That Are Nonlinear in the Parameters

The nonlinear regression functions considered in Sections 8.2 and 8.3 are nonlinear functions of the X's but are linear functions of the unknown parameters. Because they are linear in the unknown parameters, those parameters can be estimated by OLS after defining new regressors that are nonlinear transformations of the original X's. This family of nonlinear regression functions is both rich and convenient to use. In some applications, however, economic reasoning leads to regression functions that are not linear in the parameters. Although such regression functions cannot be estimated by OLS, they can be estimated using an extension of OLS called nonlinear least squares.

Functions That Are Nonlinear in the Parameters

We begin with two examples of functions that are nonlinear in the parameters. We then provide a general formulation.

Logistic curve. Suppose you are studying the market penetration of a technology, such as the adoption of machine learning software in different industries. The dependent variable is the fraction of firms in the industry that have adopted the software, a single independent variable X describes an industry characteristic, and you have data on n industries. The dependent variable is between 0 (no adopters) and 1 (100% adoption). Because a linear regression model could produce predicted values less than 0 or greater than 1, it makes sense to use instead a function that produces predicted values between 0 and 1.

    The logistic function smoothly increases from a minimum of 0 to a maximum of 1. The logistic regression model with a single X is

$$Y_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}} + u_i. \tag{8.38}$$

The logistic function with a single X and positive values of β_0 and β_1 is graphed in Figure 8.12a. As can be seen in the graph, the logistic function has an elongated "S" shape. For small values
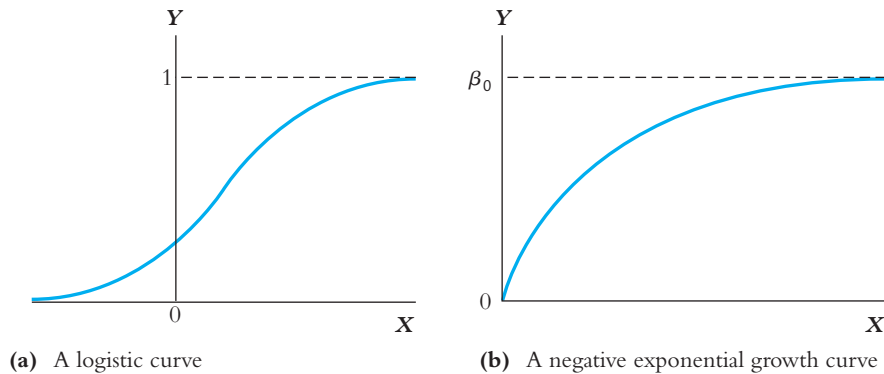

FIGURE 8.12 **Two Functions That Are Nonlinear in Their Parameters**

(a) A logistic curve

(b) A negative exponential growth curve

Figure 8.12a plots the logistic function of Equation (8.38), which has predicted values that lie between 0 and 1. Figure 8.12b plots the negative exponential growth function of Equation (8.39), which has a slope that is always positive and decreases as X increases and an asymptote at β_0 as X tends to infinity.

---

of X, the value of the function is nearly 0, and the slope is flat; the curve is steeper for moderate values of X; and for large values of X, the function approaches 1, and the slope is flat again.

Negative exponential growth. The functions used in Section 8.2 to model the relation between test scores and income have some deficiencies. For example, the polynomial models can produce a negative slope for some values of income, which is implausible. The logarithmic specification has a positive slope for all values of income; however, as income gets very large, the predicted values increase without bound, so for some incomes the predicted value for a district will exceed the maximum possible score on the test.

The negative exponential growth model provides a nonlinear specification that has a positive slope for all values of income, has a slope that is greatest at low values of income and decreases as income rises, and has an upper bound (that is, an asymptote as income increases to infinity). The negative exponential growth regression model is

$$Y_i = \beta_0 \left[1 - e^{-\beta_1(X_i - \beta_2)} \right] + u_i. \tag{8.39}$$

The negative exponential growth function with positive values of β_0 and β_1 is graphed in Figure 8.12b. The slope is steep for low values of X, but as X increases, it reaches an asymptote of β_0.

General functions that are nonlinear in the parameters. The logistic and negative exponential growth regression models are special cases of the general nonlinear regression model

$$Y_i = f(X_{1i}, \ldots, X_{ki}; \beta_0, \ldots, \beta_m) + u_i, \tag{8.40}$$

in which there are k independent variables and $m + 1$ parameters, $\beta_0, \ldots, \beta_m$. In the models of Sections 8.2 and 8.3, the X's entered this function nonlinearly, but the parameters entered linearly. In the examples of this appendix, the parameters enter nonlinearly as well. If the

parameters are known, then predicted effects can be computed using the method described in Section 8.1. In applications, however, the parameters are unknown and must be estimated from the data. Parameters that enter nonlinearly cannot be estimated by OLS, but they can be estimated by nonlinear least squares.

Nonlinear Least Squares Estimation

Nonlinear least squares is a general method for estimating the unknown parameters of a regression function when those parameters enter the population regression function nonlinearly.

Recall the discussion in Section 5.3 of the OLS estimator of the coefficients of the linear multiple regression model. The OLS estimator minimizes the sum of squared prediction mistakes in Equation (5.8), $\sum_{i=1}^{n}[Y_i - (b_0 + b_1 X_{1i} + \cdots + b_k X_{ki})]^2$. In principle, the OLS estimator can be computed by checking many trial values of $b_0, \ldots, b_k$ and settling on the values that minimize the sum of squared mistakes.

This same approach can be used to estimate the parameters of the general nonlinear regression model in Equation (8.40). Because the regression function is nonlinear in the coefficients, this method is called **nonlinear least squares**. For a set of trial parameter values $b_0, b_1, \ldots, b_m$, construct the sum of squared prediction mistakes:

$$\sum_{i=1}^{n} [Y_i - f(X_{1i}, \ldots, X_{ki}, b_1, \ldots, b_m)]^2. \tag{8.41}$$

The **nonlinear least squares estimators** of $\beta_0, \beta_1, \ldots, \beta_m$ are the values of $b_0, b_1, \ldots, b_m$ that minimize the sum of squared prediction mistakes in Equation (8.41).

In linear regression, a relatively simple formula expresses the OLS estimator as a function of the data. Unfortunately, no such general formula exists for nonlinear least squares, so the nonlinear least squares estimator must be found numerically using a computer. Regression software incorporates algorithms for solving the nonlinear least squares minimization problem, which simplifies the task of computing the nonlinear least squares estimator in practice.

Under general conditions on the function f and the X's, the nonlinear least squares estimator shares two key properties with the OLS estimator in the linear regression model: It is consistent, and it is normally distributed in large samples. In regression software that supports nonlinear least squares estimation, the output typically reports standard errors for the estimated parameters. As a consequence, inference concerning the parameters can proceed as usual; in particular, t-statistics can be constructed using the general approach in Key Concept 5.1, and a 95% confidence interval can be constructed as the estimated coefficient, plus or minus 1.96 standard errors. Just as in linear regression, the error term in the nonlinear regression model can be heteroskedastic, so heteroskedasticity-robust standard errors should be used.

Application to the Test Score–District Income Relation

A negative exponential growth model, fit to district income (X) and test scores (Y), has the desirable features of a slope that is always positive [if β_1 in Equation (8.39) is positive] and an asymptote of β_0 as income increases to infinity. Estimating β_0, β_1, and β_2 in Equation (8.39) using the California test score data yields $\hat{\beta}_0 = 703.2$ (heteroskedasticity-robust standard error $= 4.44$),

---

The negative exponential growth regression function [Equation (8.42)] and the linear-log regression function [Equation (8.18)] both capture the nonlinear relation between test scores and district income. One difference between the two functions is that the negative exponential growth model has an asymptote as *Income* increases to infinity, but the linear-log regression function does not.
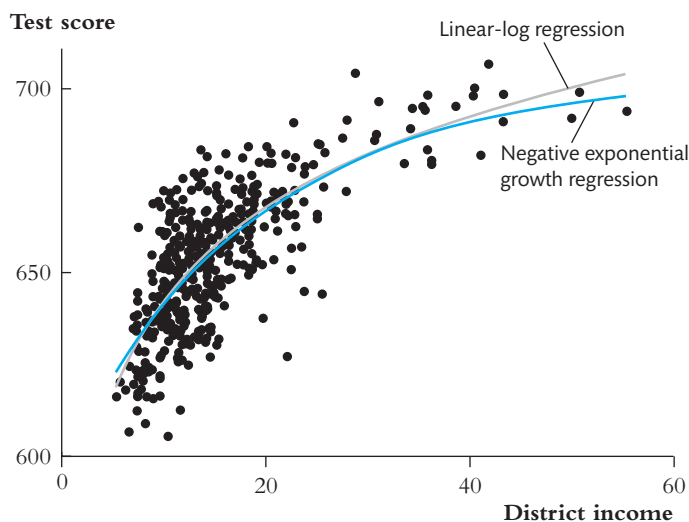
$\hat{\beta}_1 = 0.0552 \ (SE = 0.0068)$, and $\hat{\beta}_2 = -34.0 \ (SE = 4.48)$. Thus the estimated nonlinear regression function (with standard errors reported below the parameter estimates) is

$$\widehat{TestScore} = 703.2\left[1 - e^{-0.0552(Income + 34.0)}\right].$$
$$\qquad\quad (4.44) \qquad (0.0068) \quad (4.48)$$

(8.42)

This estimated regression function is plotted in Figure 8.13, along with the logarithmic regression function and a scatterplot of the data. The two specifications are, in this case, quite similar. One difference is that the negative exponential growth curve flattens out at the highest levels of income, consistent with having an asymptote.

APPENDIX

8.2 Slopes and Elasticities for Nonlinear Regression Functions

This appendix uses calculus to evaluate slopes and elasticities of nonlinear regression functions with continuous regressors. We focus on the case of Section 8.2, in which there is a single X. This approach extends to multiple X's, using partial derivatives.

Consider the nonlinear regression model, $Y_i = f(X_i) + u_i$, with $E(u_i \mid X_i) = 0$. The slope of the population regression function, $f(X)$, evaluated at the point $X = x$, is the derivative of f; that is, $df(X)/dX\mid_{X=x}$. For the polynomial regression function in Equation (8.9), $f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_r X^r$ and $dX^a/dX = aX^{a-1}$ for any constant a, so

$df(X)/dX\,|_{X=x} = \beta_1 + 2\beta_2 x + \cdots + r\beta_r x^{r-1}$. The estimated slope at x is $d\hat{f}(X)/dX\,|_{X=x} = \hat{\beta}_1 + 2\hat{\beta}_2 x + \cdots + r\hat{\beta}_r x^{r-1}$. The standard error of the estimated slope is $SE(\hat{\beta}_1 + 2\hat{\beta}_2 x + \cdots + r\hat{\beta}_r x^{r-1})$; for a given value of x, this is the standard error of a weighted sum of regression coefficients, which can be computed using the methods of Section 7.3 and Equation (8.8).

The elasticity of Y with respect to X is the percentage change in Y for a given percentage change in X. Formally, this definition applies in the limit that the percentage change in X goes to 0, so the slope appearing in the definition in Equation (8.22) is replaced by the derivative and the elasticity is

$$\text{elasticity of } Y \text{ with respect to } X = \frac{dY}{dX} \times \frac{X}{Y} = \frac{d \ln Y}{d \ln X}.$$

In a regression model, Y depends both on X and on the error term u. It is conventional to evaluate the elasticity as the percentage change not of Y but of the predicted component of Y—that is, the percentage change in $E(Y\,|\,X)$. Accordingly, the elasticity of $E(Y\,|\,X)$ with respect to X is

$$\frac{dE(Y\,|\,X)}{dX} \times \frac{X}{E(Y\,|\,X)} = \frac{d \ln E(Y\,|\,X)}{d \ln X}.$$

The elasticities for the linear model and for the three logarithmic models summarized in Key Concept 8.2 are given in the table below.

Case	Population Regression Model	Elasticity of $E(Y\,|\,X)$ with Respect to X
linear	$Y = \beta_0 + \beta_1 X + u$	$\dfrac{\beta_1 X}{\beta_0 + \beta_1 X}$
linear-log	$Y = \beta_0 + \beta_1 \ln(X) + u$	$\dfrac{\beta_1}{\beta_0 + \beta_1 \ln(X)}$
log-linear	$\ln(Y) = \beta_0 + \beta_1 X + u$	$\beta_1 X$
log-log	$\ln(Y) = \beta_0 + \beta_1 \ln(X) + u$	β_1

The log-log specification has a constant elasticity, but in the other three specifications, the elasticity depends on X.

We now derive the expressions for the linear-log and log-linear models. For the linear-log model, $E(Y\,|\,X) = \beta_0 + \beta_1 \ln(X)$. Because $d\ln(X)/dX = 1/X$, applying the chain rule yields $dE(Y\,|\,X)/dX = \beta_1/X$. Thus the elasticity is $dE(Y\,|\,X)/dX \times X/E(Y\,|\,X) = (\beta_1/X) \times X/[\beta_0 + \beta_1 \ln(X)] = \beta_1/[\beta_0 + \beta_1 \ln(X)]$, as is given in the table. For the log-linear model, it is conventional to make the additional assumption that u and X are independently distributed, so the expression for $E(Y\,|\,X)$ given following Equation (8.25) becomes $E(Y\,|\,X) = ce^{\beta_0 + \beta_1 X}$, where $c = E(e^u)$ is a constant that does not depend on X because of the additional assumption that u and X are independent. Thus $dE(Y\,|\,X)/dX = ce^{\beta_0 + \beta_1 X}\beta_1$, and the elasticity is $dE(Y\,|\,X)/dX \times X/E(Y\,|\,X) = ce^{\beta_0 + \beta_1 X}\beta_1 \times X/(ce^{\beta_0 + \beta_1 X}) = \beta_1 X$. The derivations for the linear and log-log models are left as Exercise 8.11.

9

Assessing Studies Based on Multiple Regression

The preceding five chapters explain how to use multiple regression to analyze the relationship among variables in a data set. In this chapter, we step back and ask, What makes a study that uses multiple regression reliable or unreliable? We focus on statistical studies that have the objective of estimating the causal effect of a change in some independent variable, such as class size, on a dependent variable, such as test scores. For such studies, when will multiple regression provide a useful estimate of the causal effect, and, just as importantly, when will it fail to do so?

To answer these questions, this chapter presents a framework for assessing statistical studies in general, whether or not they use regression analysis. This framework relies on the concepts of internal and external validity. A study is internally valid if its statistical inferences about causal effects are valid for the population and setting studied; it is externally valid if its inferences can be generalized to other populations and settings. In Sections 9.1 and 9.2, we discuss internal and external validity, list a variety of possible threats to internal and external validity, and discuss how to identify those threats in practice. The discussion in Sections 9.1 and 9.2 focuses on the estimation of causal effects from observational data. Section 9.3 returns to the prediction problem and discusses threats to the validity of predictions made using regression models.

As an illustration of the framework of internal and external validity, in Section 9.4 we assess the internal and external validity of the study of the effect on test scores of cutting the student–teacher ratio presented in Chapters 4 through 8.

9.1 Internal and External Validity

The concepts of internal and external validity, defined in Key Concept 9.1, provide a framework for evaluating whether a statistical or econometric study is useful for answering a specific question of interest.

Internal and external validity distinguish between the population and setting studied and the population and setting to which the results are generalized. The **population studied** is the population of entities—people, companies, school districts, and so forth—from which the sample was drawn. The population to which the results are generalized, or the **population of interest**, is the population of entities to which the causal inferences from the study are to be applied. For example, a high school (grades 9 through 12) principal might want to generalize our findings on class sizes and test scores in California elementary school districts (the population studied) to the population of high schools (the population of interest).

Internal and External Validity

9.1

A statistical analysis is said to have **internal validity** if the statistical inferences about causal effects are valid for the population being studied. The analysis is said to have **external validity** if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.

By *setting*, we mean the institutional, legal, social, physical, and economic environment. For example, it would be important to know whether the findings of a laboratory experiment assessing methods for growing organic tomatoes could be generalized to the field—that is, whether the organic methods that work in the setting of a laboratory also work in the setting of the real world. We provide other examples of differences in populations and settings later in this section.

Threats to Internal Validity

Internal validity has two components. First, the estimator of the causal effect should be unbiased and consistent. For example, if $\hat{\beta}_{STR}$ is the OLS estimator of the effect on test scores of a unit change in the student–teacher ratio in a certain regression, then $\hat{\beta}_{STR}$ should be an unbiased and consistent estimator of the population causal effect of a change in the student–teacher ratio, β_{STR}.

Second, hypothesis tests should have the desired significance level (the actual rejection rate of the test under the null hypothesis should equal its desired significance level), and confidence intervals should have the desired confidence level. For example, if a confidence interval is constructed as $\hat{\beta}_{STR} \pm 1.96 \, SE(\hat{\beta}_{STR})$, this confidence interval should contain the true population causal effect, β_{STR}, with 95% probability over repeated samples drawn from the population being studied.

In regression analysis, causal effects are estimated using the estimated regression function, and hypothesis tests are performed using the estimated regression coefficients and their standard errors. Accordingly, in a study based on OLS regression, the requirements for internal validity are that the OLS estimator is unbiased and consistent and that standard errors are computed in a way that makes confidence intervals have the desired confidence level. For various reasons, these requirements might not be met, and these reasons constitute threats to internal validity. These threats lead to failures of one or more of the least squares assumptions in Key Concept 6.4. For example, one threat that we have discussed at length is omitted variable bias; it leads to correlation between one or more regressors and the error term, which violates the first least squares assumption. If data are available on the omitted variable or on an adequate control variable, then this threat can be avoided by including that variable as an additional regressor.

Section 9.2 provides a detailed discussion of the various threats to internal validity in multiple regression analysis and suggests how to mitigate them.

Threats to External Validity

Potential threats to external validity arise from differences between the population and setting studied and the population and setting of interest.

Differences in populations. Differences between the population studied and the population of interest can pose a threat to external validity. For example, laboratory studies of the toxic effects of chemicals typically use animal populations like mice (the population studied), but the results are used to write health and safety regulations for human populations (the population of interest). Whether mice and men differ sufficiently to threaten the external validity of such studies is a matter of debate.

More generally, the true causal effect might not be the same in the population studied and the population of interest. This could be because the population was chosen in a way that makes it different from the population of interest, because of differences in characteristics of the populations, because of geographical differences, or because the study is out of date.

Differences in settings. Even if the population being studied and the population of interest are identical, it might not be possible to generalize the study results if the settings differ. For example, a study of the effect on college binge drinking of an antidrinking advertising campaign might not generalize to another, identical group of college students if the legal penalties for drinking at the two colleges differ. In this case, the legal setting in which the study was conducted differs from the legal setting to which its results are applied.

More generally, examples of differences in settings include differences in the institutional environment (public universities versus religious universities), differences in laws (differences in legal penalties), and differences in the physical environment (tailgate-party binge drinking in southern California versus Fairbanks, Alaska).

Application to test scores and the student–teacher ratio. Chapters 7 and 8 reported statistically significant, but substantively small, estimated improvements in test scores resulting from reducing the student–teacher ratio. This analysis was based on test results for California school districts. Suppose for the moment that these results are internally valid. To what other populations and settings of interest could this finding be generalized?

The closer the population and setting of the study are to those of interest, the stronger the case is for external validity. For example, college students and college instruction are very different from elementary school students and instruction, so it is implausible that the effect of reducing class sizes estimated using the California

elementary school district data would generalize to colleges. On the other hand, elementary school students, curriculum, and organization are broadly similar throughout the United States, so it is plausible that the California results might generalize to performance on standardized tests in other U.S. elementary school districts.

How to assess the external validity of a study.   External validity must be judged using specific knowledge of the populations and settings studied and those of interest. Important differences between the two will cast doubt on the external validity of the study.

Sometimes there are two or more studies on different but related populations. If so, the external validity of both studies can be checked by comparing their results. For example, in Section 9.4, we analyze test score and class size data for elementary school districts in Massachusetts and compare the Massachusetts and California results. In general, similar findings in two or more studies bolster claims to external validity, while differences in their findings that are not readily explained cast doubt on their external validity.[1]

How to design an externally valid study.   Because threats to external validity stem from a lack of comparability of populations and settings, these threats are best minimized at the early stages of a study, before the data are collected. Study design is beyond the scope of this textbook, and the interested reader is referred to Shadish, Cook, and Campbell (2002).

9.2   Threats to Internal Validity of Multiple Regression Analysis

Studies based on regression analysis are internally valid if the estimated regression coefficients are unbiased and consistent for the causal effect of interest and if their standard errors yield confidence intervals with the desired confidence level. This section surveys five reasons why the OLS estimator of the multiple regression coefficients might be biased, even in large samples: omitted variables, misspecification of the functional form of the regression function, imprecise measurement of the independent variables ("errors in variables"), sample selection, and simultaneous causality. All five sources of bias arise because the regressor is correlated with the error term in the population regression, violating the first least squares assumption in

---

[1]A comparison of many related studies on the same topic is called a meta-analysis. The discussion in the box "The Mozart Effect: Omitted Variable Bias?" in Chapter 6 is based on a meta-analysis, for example. Performing a meta-analysis of many studies has its own challenges. How do you sort the good studies from the bad? How do you compare studies when the dependent variables differ? Should you put more weight on studies with larger samples? A discussion of meta-analysis and its challenges goes beyond the scope of this text. The interested reader is referred to Hedges and Olkin (1985), Cooper and Hedges (1994), and, for more recent work that interprets *p*-values from published studies, Simonsohn, Nelson, and Simmons (2014).

Key Concept 6.4. For each, we discuss what can be done to reduce this bias. The section concludes with a discussion of circumstances that lead to inconsistent standard errors and what can be done about it.

Omitted Variable Bias

Recall that omitted variable bias arises when a variable that both determines Y and is correlated with one or more of the included regressors is omitted from the regression. This bias persists even in large samples, so the OLS estimator is inconsistent. How best to minimize omitted variable bias depends on whether or not variables that adequately control for the potential omitted variable are available.

Solutions to omitted variable bias when the variable is observed or there are adequate control variables. If you have data on the omitted variable, then you can include that variable in a multiple regression, thereby addressing the problem. Alternatively, if you have data on one or more control variables and if these control variables are adequate in the sense that they lead to conditional mean independence [Equation (6.18)], then including those control variables eliminates the potential bias in the coefficient on the variable of interest.

Adding a variable to a regression has both costs and benefits. On the one hand, omitting the variable could result in omitted variable bias. On the other hand, including the variable when it does not belong (that is, when its population regression coefficient is 0) reduces the precision of the estimators of the other regression coefficients. In other words, the decision whether to include a variable involves a trade-off between bias and variance of the coefficient of interest. In practice, there are four steps that can help you decide whether to include a variable or set of variables in a regression.

The first step is to identify the key coefficient or coefficients of interest in your regression. In the test score regressions, this is the coefficient on the student–teacher ratio because the question originally posed concerns the effect on test scores of reducing the student–teacher ratio.

The second step is to ask yourself: What are the most likely sources of important omitted variable bias in this regression? Answering this question requires applying economic theory and expert knowledge, and should occur before you actually run any regressions; because this step is done before analyzing the data, it is referred to as *a priori* ("before the fact") reasoning. In the test score example, this step entails identifying those determinants of test scores that, if ignored, could bias our estimator of the class size effect. The results of this step are a base regression specification, the starting point for your empirical regression analysis, and a list of additional, "questionable" control variables that might help to mitigate possible omitted variable bias.

The third step is to augment your base specification with the additional, questionable control variables identified in the second step. If the coefficients on the

Omitted Variable Bias: Should I Include More Variables in My Regression?

If you include another variable in your multiple regression, you will eliminate the possibility of omitted variable bias from excluding that variable, but the variance of the estimator of the coefficients of interest can increase. Here are some guidelines to help you decide whether to include an additional variable:

1. Be specific about the coefficient or coefficients of interest.

2. Use *a-priori* reasoning to identify the most important potential sources of omitted variable bias, leading to a base specification and some "questionable" variables.

3. Test whether additional, "questionable" control variables have nonzero coefficients, and assess whether including a questionable control variable makes a meaningful change in the coefficient of interest.

4. Provide "full disclosure" representative tabulations of your results so that others can see the effect of including the questionable variables on the coefficient(s) of interest.

additional control variables are statistically significant and/or if the estimated coefficients of interest change appreciably when the additional variables are included, then they should remain in the specification and you should modify your base specification. If not, then these variables can be excluded from the regression.

The fourth step is to present an accurate summary of your results in tabular form. This provides "full disclosure" to a potential skeptic, who can then draw his or her own conclusions. Tables 7.1 and 8.3 are examples of this strategy. For example, in Table 8.3, we could have presented only the regression in column (7) because that regression summarizes the relevant effects and nonlinearities in the other regressions in that table. Presenting the other regressions, however, permits the skeptical reader to draw his or her own conclusions.

These steps are summarized in Key Concept 9.2.

Solutions to omitted variable bias when adequate control variables are not available. Adding an omitted variable to a regression is not an option if you do not have data on that variable and if there are no adequate control variables. Still, there are three other ways to solve omitted variable bias. Each of these three solutions circumvents omitted variable bias through the use of different types of data.

The first solution is to use data in which the same observational unit is observed at different points in time. For example, test score and related data might be collected for the same districts in 1995 and again in 2000. Data in this form are called panel data. As explained in Chapter 10, panel data make it possible to control for unobserved omitted variables as long as those omitted variables do not change over time.

Functional Form Misspecification

Functional form misspecification arises when the functional form of the esti-mated regression function differs from the functional form of the population regression function. If the functional form is misspecified, then the estimator of the partial effect of a change in one of the variables will, in general, be biased. Functional form misspecification often can be detected by plotting the data and the estimated regression function, and it can be corrected by using a different functional form.

The second solution is to use instrumental variables regression. This method relies on a new variable, called an instrumental variable. Instrumental variables regression is discussed in Chapter 12.

The third solution is to use a study design in which the effect of interest (for example, the effect of reducing class size on student achievement) is studied using a randomized controlled experiment. Randomized controlled experiments are discussed in Chapter 13.

Misspecification of the Functional Form of the Regression Function

If the true population regression function is nonlinear but the estimated regression is linear, then this **functional form misspecification** makes the OLS estimator biased. This bias is a type of omitted variable bias, in which the omitted variables are the terms that reflect the missing nonlinear aspects of the regression function. For example, if the population regression function is a quadratic polynomial, then a regression that omits the square of the independent variable would suffer from omitted variable bias. Bias arising from functional form misspecification is summarized in Key Concept 9.3.

Solutions to functional form misspecification. When the dependent variable is continu-ous (like test scores), this problem of potential nonlinearity can be solved using the meth-ods of Chapter 8. If, however, the dependent variable is discrete or binary (for example, if Y_i equals 1 if the ith person attended college and equals 0 otherwise), things are more complicated. Regression with a discrete dependent variable is discussed in Chapter 11.

Measurement Error and Errors-in-Variables Bias

Suppose that in our regression of test scores against the student–teacher ratio we had inadvertently mixed up our data, so that we ended up regressing test scores for fifth graders on the student–teacher ratio for tenth graders in that district. Although the student–teacher ratio for elementary school students and tenth graders might be

correlated, they are not the same, so this mix-up would lead to bias in the estimated coefficient. This is an example of **errors-in-variables bias** because its source is an error in the measurement of the independent variable. This bias persists even in very large samples, so the OLS estimator is inconsistent if there is measurement error.

There are many possible sources of measurement error. If the data are collected through a survey, a respondent might give the wrong answer. For example, one question in the Current Population Survey involves last year's earnings. A respondent might not know his or her exact earnings or might misstate the amount for some other reason. If instead the data are obtained from computerized administrative records, there might have been errors when the data were first entered.

To see that errors in variables can result in correlation between the regressor and the error term, suppose there is a single regressor X_i (say, actual earnings) which is measured imprecisely by $\tilde{X}_i$ (the respondent's stated earnings). Because $\tilde{X}_i$, not X_i, is observed, the regression equation actually estimated is the one based on $\tilde{X}_i$. Written in terms of the imprecisely measured variable $\tilde{X}_i$, the population regression equation $Y_i = \beta_0 + \beta_1 X_i + u_i$ is

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1(X_i - \tilde{X}_i) + u_i] \\
&= \beta_0 + \beta_1 \tilde{X}_i + v_i,
\end{aligned}
\tag{9.1}
$$

where $v_i = \beta_1(X_i - \tilde{X}_i) + u_i$. Thus the population regression equation written in terms of $\tilde{X}_i$ has an error term that contains the measurement error, the difference between $\tilde{X}_i$ and X_i. If this difference is correlated with the measured value $\tilde{X}_i$, then the regressor $\tilde{X}_i$ will be correlated with the error term, and $\hat{\beta}_1$ will be biased and inconsistent.

The precise size and direction of the bias in $\hat{\beta}_1$ depend on the correlation between $\tilde{X}_i$ and the measurement error, $\tilde{X}_i - X_i$. This correlation depends in turn on the specific nature of the measurement error.

For example, suppose the measured value, $\tilde{X}_i$, equals the actual, unmeasured value, X_i, plus a purely random component, w_i, which has mean 0 and variance σ_w^2. Because the error is purely random, we might suppose that w_i is uncorrelated with X_i and with the regression error u_i. This assumption constitutes the **classical measurement error model**, in which $\tilde{X}_i = X_i + w_i$, where $\text{corr}(w_i, X_i) = 0$ and $\text{corr}(w_i, u_i) = 0$. Under the classical measurement error model, a bit of algebra[2] shows that $\hat{\beta}_1$ has the probability limit

$$
\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1.
\tag{9.2}
$$

---

[2]Under this measurement error assumption, $v_i = \beta_1(X_i - \tilde{X}_i) + u_i = -\beta_1 w_i + u_i$, $\text{cov}(X_i, u_i) = 0$, and $\text{cov}(\tilde{X}_1, w_i) = \text{cov}(X_i + w_i, w_i) = \sigma_w^2$, so $\text{cov}(\tilde{X}_i, v_i) = -\beta_1\text{cov}(\tilde{X}_1, w_i) + \text{cov}(\tilde{X}_i, u_i) = -\beta_1\sigma_w^2$. Thus, from Equation (6.1), $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1\sigma_w^2/\sigma_{\tilde{X}}^2$. Now $\sigma_{\tilde{X}}^2 = \sigma_X^2 + \sigma_w^2$, so $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1\sigma_w^2/(\sigma_X^2 + \sigma_w^2) = [\sigma_X^2/(\sigma_X^2 + \sigma_w^2)]\beta_1$.

Errors-in-Variables Bias

Errors-in-variables bias in the OLS estimator arises when an independent variable is measured imprecisely. This bias depends on the nature of the measurement error and persists even if the sample size is large. If the measured variable equals the actual value plus a mean 0, independently distributed measurement error, then the OLS estimator in a regression with a single right-hand variable is biased toward 0, and its probability limit is given in Equation (9.2).

That is, if the measurement error has the effect of simply adding a random element to the actual value of the independent variable, then $\hat{\beta}_1$ is inconsistent. Because the ratio $\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}$ is less than 1, $\hat{\beta}_1$ will be biased toward 0, even in large samples. In the extreme case that the measurement error is so large that essentially no information about X_i remains, the ratio of the variances in the final expression in Equation (9.2) is 0, and $\hat{\beta}_1$ converges in probability to 0. In the other extreme, when there is no measurement error, $\sigma_w^2 = 0$, so $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

A different model of measurement error supposes that the respondent makes his or her best estimate of the true value. In this "best guess" model, the response $\tilde{X}_i$ is modeled as the conditional mean of X_i given the information available to the respondent. Because $\tilde{X}_i$ is the best guess, the measurement error $\tilde{X}_i - X_i$ is uncorrelated with the response $\tilde{X}_i$ (if the measurement error were correlated with $\tilde{X}_i$, then that would be useful information for predicting X_i, in which case $\tilde{X}_i$ would not have been the best guess of X_i). That is, $E[(\tilde{X}_i - X_i)\tilde{X}_i] = 0$, and if the respondent's information is uncorrelated with u_i, then $\tilde{X}_i$ is uncorrelated with the error term v_i. Thus, in this "best guess" measurement error model, $\hat{\beta}_1$ is consistent, but because $\text{var}(v_i) > \text{var}(u_i)$, the variance of $\hat{\beta}_1$ is larger than it would be absent measurement error. The "best guess" measurement error model is examined further in Exercise 9.12.

Problems created by measurement error can be even more complicated if there is intentional misreporting. For example, suppose that survey respondents provide the income reported on their income taxes but intentionally underreport their true taxable income by not including cash payments. If, for example, all respondents report only 90% of income, then $\tilde{X}_i = 0.90X_i$, and $\hat{\beta}_1$ will be biased *up* by 10%.

Although the result in Equation (9.2) is specific to classical measurement error, it illustrates the more general proposition that if the independent variable is measured imprecisely, then the OLS estimator may be biased, even in large samples. Errors-in-variables bias is summarized in Key Concept 9.4.

Measurement error in Y. The effect of measurement error in Y is different from that of measurement error in X. If Y has classical measurement error, then this measurement error increases the variance of the regression and of $\hat{\beta}_1$ but does not induce bias

in $\hat{\beta}_1$. To see this, suppose that measured Y_i is $\tilde{Y}_i$, which equals true Y_i plus random measurement error w_i. Then the regression model estimated is $\tilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$, where $v_i = w_i + u_i$. If w_i is truly random, then w_i and X_i are independently distributed, so that $E(w_i | X_i) = 0$, in which case $E(v_i | X_i) = 0$, so $\hat{\beta}_1$ is unbiased. However, because $\text{var}(v_i) > \text{var}(u_i)$, the variance of $\hat{\beta}_1$ is larger than it would be without measurement error. In the test score/class size example, suppose test scores have purely random grading errors that are independent of the regressors; then the classical measurement error model of this paragraph applies to $\tilde{Y}_i$, and $\hat{\beta}_1$ is unbiased. More generally, measurement error in Y that has conditional mean 0 given the regressors will not induce bias in the OLS coefficients.

Solutions to errors-in-variables bias. The best way to solve the errors-in-variables problem is to get an accurate measure of X. If this is impossible, however, econometric methods can be used to mitigate errors-in-variables bias.

One such method is instrumental variables regression. It relies on having another variable (the instrumental variable) that is correlated with the actual value X_i but is uncorrelated with the measurement error. This method is studied in Chapter 12.

A second method is to develop a mathematical model of the measurement error and, if possible, to use the resulting formulas to adjust the estimates. For example, if a researcher believes that the classical measurement error model applies and if she knows or can estimate the ratio σ_w^2 / σ_X^2, then she can use Equation (9.2) to compute an estimator of β_1 that corrects for the downward bias. Because this approach requires specialized knowledge about the nature of the measurement error, the details typically are specific to a given data set and its measurement problems, and we shall not pursue this approach further in this text.

Missing Data and Sample Selection

Missing data are a common feature of economic data sets. Whether missing data pose a threat to internal validity depends on why the data are missing. We consider three cases: when the data are missing completely at random, when the data are missing based on X, and when the data are missing because of a selection process that is related to Y beyond depending on X.

When the data are missing completely at random—that is, for random reasons unrelated to the values of X or Y—the effect is to reduce the sample size but not introduce bias. For example, suppose you conduct a simple random sample of 100 classmates, then randomly lose half the records. It would be as if you had never surveyed those individuals. You would be left with a simple random sample of 50 classmates, so randomly losing the records does not introduce bias.

When the data are missing based on the value of a regressor, the effect also is to reduce the sample size but not to introduce bias. For example, in the class size/student–teacher ratio example, suppose we used only the districts in which the student–teacher ratio exceeds 20. Although we would not be able to draw conclusions

Sample Selection Bias

Sample selection bias arises when a selection process influences the availability of data and that process is related to the dependent variable beyond depending on the regressors. Such sample selection induces correlation between one or more regressors and the error term, leading to bias and inconsistency of the OLS estimator.

about what happens when $STR \leq 20$, this would not introduce bias into our analysis of the class size effect for districts with $STR > 20$.

In contrast to the first two cases, if the data are missing because of a selection process that is related to the value of the dependent variable (Y) beyond depending on the regressors (X), then this selection process can introduce correlation between the error term and the regressors. The resulting bias in the OLS estimator is called **sample selection bias**. An example of sample selection bias in polling was given in the box "Landon Wins!" in Section 3.1. In that example, the sample selection method (randomly selecting phone numbers of automobile owners) was related to the dependent variable (who the individual supported for president in 1936) because in 1936 car owners with phones were more likely to be Republicans. The sample selection problem can be cast either as a consequence of nonrandom sampling or as a missing data problem. In the 1936 polling example, the sample was a random sample of car owners with phones, not a random sample of voters. Alternatively, this example can be cast as a missing data problem by imagining a random sample of voters but with missing data for those without cars and phones. The mechanism by which the data are missing is related to the dependent variable, leading to sample selection bias.

Sample selection bias is summarized in Key Concept 9.5.[3]

Solutions to selection bias. The best solution to sample selection bias is to avoid it by the design of your study. If you want to estimate the mean height of undergraduates for your statistics course, do so by using a random sample of all undergraduates—not by sampling students as they enter a basketball court. The box "Do Stock Mutual Funds Outperform the Market?" describes a way to select a sample of funds to avoid a more subtle form of sample selection bias. If your data do have sample selection bias, it cannot be eliminated using the methods we have discussed so far. Methods for estimating models with sample selection are beyond the scope of this text. Some of those methods build on the techniques introduced in Chapter 11, where further references are provided.

---

[3]Exercise 19.16 provides a mathematical treatment of the three missing data cases discussed here.

Do Stock Mutual Funds Outperform the Market?

S tock mutual funds are investment vehicles that hold a portfolio of stocks. By purchasing shares in a mutual fund, a small investor can hold a broadly diversified portfolio without the hassle and expense (transaction cost) of buying and selling shares in individual companies. Some mutual funds simply track the market (for example, by holding the stocks in the S&P 500), whereas others are actively managed by full-time professionals whose job is to make the fund earn a better return than the over-all market—and competitors' funds. But do these actively managed funds achieve this goal? Do some mutual funds consistently beat other funds and the market?

One way to answer these questions is to com-pare future returns on mutual funds that had high returns over the past year to future returns on other funds and on the market as a whole. In making such comparisons, financial economists know that it is important to select the sample of mutual funds care-fully. This task is not as straightforward as it seems, however. Some databases include historical data on funds currently available for purchase, but this approach means that the dogs—the most poorly performing funds—are omitted from the data set because they went out of business or were merged into other funds. For this reason, a study using data on historical performance of currently available funds is subject to sample selection bias: The sample is selected based on the value of the dependent vari-able, returns, because funds with the lowest returns are eliminated. The mean return of all funds (includ-ing the defunct) over a ten-year period will be less than the mean return of those funds still in existence at the end of those ten years, so a study of only the latter funds will overstate performance. Financial economists refer to this selection bias as *survivor-ship bias* because only the better funds survive to be in the data set.

When financial econometricians correct for survivorship bias by incorporating data on defunct funds, the results do not paint a flattering portrait of mutual fund managers. Corrected for survivor-ship bias, the econometric evidence indicates that actively managed stock mutual funds do not outper-form the market, on average, and that past good per-formance does not predict future good performance. For further reading on mutual funds and survivor-ship bias, see Malkiel (2016), Chapter 7, and Carhart (1997). The problem of survivorship bias also arises in evaluating hedge fund performance; for further reading, see Aggarwal and Jorion (2010).

Simultaneous Causality

So far, we have assumed that causality runs from the regressors to the dependent vari-able (X causes Y). But what if causality also runs from the dependent variable to one or more regressors (Y causes X)? If so, causality runs "backward" as well as forward; that is, there is **simultaneous causality**. If there is simultaneous causality, an OLS regression picks up both effects, so the OLS estimator is biased and inconsistent.

For example, our study of test scores focused on the effect on test scores of reducing the student–teacher ratio, so causality is presumed to run from the student–teacher ratio to test scores. Suppose, however, a government initiative subsidized hiring teachers in school districts with poor test scores. If so, causality would run in both directions: For the usual educational reasons, low student–teacher ratios would

arguably lead to high test scores, but because of the government program, low test scores would lead to low student–teacher ratios.

Simultaneous causality leads to correlation between the regressor and the error term. In the test score example, suppose there is an omitted factor that leads to poor test scores; because of the government program, this factor that produces low scores in turn results in a low student–teacher ratio. Thus a negative error term in the population regression of test scores on the student–teacher ratio reduces test scores, but because of the government program, it also leads to a decrease in the student–teacher ratio. In other words, the student–teacher ratio is positively correlated with the error term in the population regression. This in turn leads to simultaneous causality bias and inconsistency of the OLS estimator.

This correlation between the error term and the regressor can be made mathematically precise by introducing an additional equation that describes the reverse causal link. For convenience, consider just the two variables X and Y, and ignore other possible regressors. Accordingly, there are two equations, one in which X causes Y and one in which Y causes X:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \text{ and} \tag{9.3}$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i. \tag{9.4}$$

Equation (9.3) is the familiar one in which β_1 is the effect on Y of a change in X, where u represents other factors. Equation (9.4) represents the reverse causal effect of Y on X. In the test score problem, Equation (9.3) represents the educational effect of class size on test scores, while Equation (9.4) represents the reverse causal effect of test scores on class size induced by the government program.

Simultaneous causality leads to correlation between X_i and the error term u_i in Equation (9.3). To see this, imagine that u_i is positive, which increases Y_i. However, this higher value of Y_i affects the value of X_i through the second of these equations, and if γ_1 is positive, a high value of Y_i will lead to a high value of X_i. In general, if γ_1 is nonzero, X_i and u_i will be correlated.[4]

Because it can be expressed mathematically using two simultaneous equations, simultaneous causality bias is sometimes called **simultaneous equations bias**. Simultaneous causality bias is summarized in Key Concept 9.6.

Solutions to simultaneous causality bias. There are two ways to mitigate simultaneous causality bias. One is to use instrumental variables regression, the topic of Chapter 12. The second is to design and implement a randomized controlled experiment in which the reverse causality channel is nullified, and such experiments are discussed in Chapter 13.

---

[4]To show this mathematically, note that Equation (9.4) implies that $\text{cov}(X_i, u_i) = \text{cov}(\gamma_0 + \gamma_1 Y_i + v_i, u_i) = \gamma_1 \text{cov}(Y_i, u_i) + \text{cov}(v_i, u_i)$. Assuming that $\text{cov}(v_i, u_i) = 0$, by Equation (9.3) this in turn implies that $\text{cov}(X_i, u_i) = \gamma_1 \text{cov}(\beta_0 + \beta_1 X_i + u_i, u_i) = \gamma_1 \beta_1 \text{cov}(X_i, u_i) + \gamma_1 \sigma_u^2$. Solving for $\text{cov}(X_i, u_i)$ then yields the result $\text{cov}(X_i, u_i) = \gamma_1 \sigma_u^2 / (1 - \gamma_1 \beta_1)$.

Simultaneous Causality Bias

Simultaneous causality bias, also called simultaneous equations bias, arises in a regression of Y on X when, in addition to the causal link of interest from X to Y, there is a causal link from Y to X. This reverse causality makes X correlated with the error term in the population regression of interest.

Sources of Inconsistency of OLS Standard Errors

Inconsistent standard errors pose a different threat to internal validity. Even if the OLS estimator is consistent and the sample is large, inconsistent standard errors will produce hypothesis tests with size that differs from the desired significance level and "95%" confidence intervals that fail to include the true value in 95% of repeated samples.

There are two main reasons for inconsistent standard errors: improperly handled heteroskedasticity and correlation of the error term across observations.

Heteroskedasticity. As discussed in Section 5.4, for historical reasons, some regression software reports homoskedasticity-only standard errors. If, however, the regression error is heteroskedastic, those standard errors are not a reliable basis for hypothesis tests and confidence intervals. The solution to this problem is to use heteroskedasticity-robust standard errors and to construct F-statistics using a heteroskedasticity-robust variance estimator. Heteroskedasticity-robust standard errors are provided as an option in modern software packages.

Correlation of the error term across observations. In some settings, the population regression error can be correlated across observations. This will not happen if the data are obtained by sampling at random from the population because the randomness of the sampling process ensures that the errors are independently distributed from one observation to the next. Sometimes, however, sampling is only partially random. The most common circumstance is when the data are repeated observations on the same entity over time, such as the same school district for different years. If the omitted variables that constitute the regression error are persistent (like district demographics), "serial" correlation is induced in the regression error over time. Serial correlation in the error term can arise in panel data (e.g., data on multiple districts for multiple years) and in time series data (e.g., data on a single district for multiple years).

Another situation in which the error term can be correlated across observations is when sampling is based on a geographical unit. If there are omitted variables that reflect geographic influences, these omitted variables could result in correlation of the regression errors for adjacent observations.

Correlation of the regression error across observations does not make the OLS estimator biased or inconsistent, but it does violate the second least squares

Threats to the Internal Validity of a Multiple Regression Study

There are five primary threats to the internal validity of a multiple regression study:

1. Omitted variables

2. Functional form misspecification

3. Errors in variables (measurement error in the regressors)

4. Sample selection

5. Simultaneous causality.

Each of these, if present, results in failure of the first least squares assumption in Key Concept 6.4 (or, if there are control variables, in Key Concept 6.6), which in turn means that the OLS estimator is biased and inconsistent.

Incorrect calculation of the standard errors also poses a threat to internal validity. Homoskedasticity-only standard errors are invalid if heteroskedasticity is present. If the variables are not independent across observations, as can arise in panel and time series data, then a further adjustment to the standard error formula is needed to obtain valid standard errors.

Applying this list of threats to a multiple regression study provides a systematic way to assess the internal validity of that study.

assumption in Key Concept 6.4. The consequence is that the OLS standard errors—both homoskedasticity-only *and* heteroskedasticity-robust—are incorrect in the sense that they do not produce confidence intervals with the desired confidence level.

In many cases, this problem can be fixed by using an alternative formula for standard errors. We provide formulas for computing standard errors that are robust to both heteroskedasticity and serial correlation in Chapter 10 (regression with panel data) and in Chapter 16 (regression with time series data).

Key Concept 9.7 summarizes the threats to internal validity of a multiple regression study.

9.3 Internal and External Validity When the Regression Is Used for Prediction

When regression models are used for prediction, concerns about external validity are very important, but concerns about unbiased estimation of causal effects are not.

Chapter 4 began by considering two problems. A school superintendent wants to know how much test scores will increase if she reduces class sizes in her school district; that is, the superintendent wants to know the causal effect on test scores of

a change in class size. A father, considering moving to a school district for which test scores are not publicly available, wants a reliable prediction about test scores in that district, based on data to which he has access. The father does not need to know the causal effect on test scores of class size—or, for that matter, of any variable. What matters to him is that the prediction equation estimated using the California district-level data provides an accurate and reliable prediction of test scores for the district to which the father is considering moving.

Reliable prediction using multiple regression has three requirements. The first requirement is that the data used to estimate the prediction model and the observation for which the prediction is to be made are drawn from the same distribution. This requirement is formalized as the first least squares assumption for prediction, given in Appendix 6.4 for the case of multiple predictors. If the estimation and prediction observations are drawn from the same population, then the estimated conditional expectation of Y given X generalizes to the out-of-sample observation to be predicted. This requirement is a mathematical statement of external validity in the prediction context. In the test score example, if the estimated regression line is useful for other districts in California, it could well be useful for elementary school districts in other states, but it is unlikely to be useful for colleges.

The second requirement involves the list of predictors. When the aim is to estimate a causal effect, it is important to choose control variables to reduce the threat of omitted variable bias. In contrast, for prediction the aim is to have an accurate out-of-sample forecast. For this purpose, the predictors should be ones that substantially contribute to explaining the variation in Y, whether or not they have any causal interpretation. The question of choice of predictor is further complicated when there are time series data, for then there is the opportunity to exploit correlation over time (serial correlation) to make forecasts—that is, predictions of future values of variables. The use of multiple regression for time series forecasting is taken up in Chapters 15 and 17.

The third requirement concerns the estimator itself. So far, we have focused on OLS for estimating multiple regression. In some prediction applications, however, there are very many predictors; indeed, in some applications the number of predictors can exceed the sample size. If there are very many predictors, then there are—surprisingly—some estimators that can provide more accurate out-of-sample predictions than OLS. Chapter 14 takes up prediction with many predictors and explains these specialized estimators.

9.4   Example: Test Scores and Class Size

The framework of internal and external validity helps us to take a critical look at what we have learned—and what we have not—from our analysis of the California test score data.

External Validity

Whether the California analysis can be generalized—that is, whether it is externally valid—depends on the population and setting to which the generalization is made. Here, we consider whether the results can be generalized to performance on other standardized tests in other elementary public school districts in the United States.

Section 9.1 noted that having more than one study on the same topic provides an opportunity to assess the external validity of both studies by comparing their results. In the case of test scores and class size, other comparable data sets are, in fact, available. In this section, we examine a different data set, based on standardized test results for fourth graders in 220 public school districts in Massachusetts in 1998. Both the Massachusetts and California tests are broad measures of student knowledge and academic skills, although the details differ. Similarly, the organization of classroom instruction is broadly similar at the elementary school level in the two states (as it is in most U.S. elementary school districts), although aspects of elementary school funding and curriculum differ. Thus finding similar results about the effect of the student–teacher ratio on test performance in the California and Massachusetts data would be evidence of external validity of the findings in California. Conversely, finding different results in the two states would raise questions about the internal or external validity of at least one of the studies.

Comparison of the California and Massachusetts data. Like the California data, the Massachusetts data are at the school district level. The definitions of the variables in the Massachusetts data set are the same as those in the California data set, or nearly so. More information on the Massachusetts data set, including definitions of the variables, is given in Appendix 9.1.

Table 9.1 presents summary statistics for the California and Massachusetts samples. The average test score is higher in Massachusetts, but the test is different, so a

	California		Massachusetts	
TABLE 9.1 Summary Statistics for California and Massachusetts Test Score Data Sets				
	Average	**Standard Deviation**	**Average**	**Standard Deviation**
Test scores	654.1	19.1	709.8	15.1
Student–teacher ratio	19.6	1.9	17.3	2.3
% English learners	15.8%	18.3%	1.1%	2.9%
% receiving subsidized lunch	44.7%	27.1%	15.3%	15.1%
Average district income ($)	$15,317	$7226	$18,747	$5808
Number of observations	420		220	
Year	1999		1998	

direct comparison of scores is not appropriate. The average student–teacher ratio is higher in California than in Massachusetts (19.6 versus 17.3). Average district income is 20% higher in Massachusetts, but the standard deviation of district income is greater in California; that is, there is a greater spread in average district income in California than in Massachusetts. The average percentage of students still learning English and the average percentage of students receiving subsidized lunches are both much higher in the California districts than in the Massachusetts districts.

Test scores and average district income. To save space, we do not present scatterplots of all the Massachusetts data. Because it was a focus in Chapter 8, however, it is interesting to examine the relationship between test scores and average district income in Massachusetts. This scatterplot is presented in Figure 9.1. The general pattern of this scatterplot is similar to that in Figure 8.2 for the California data: The relationship between district income and test scores appears to be steep for low values of income and flatter for high values. Evidently, the linear regression plotted in the figure misses this apparent nonlinearity. Cubic and logarithmic regression functions are also plotted in Figure 9.1. The cubic regression function has a slightly higher $\overline{R}^2$ than the logarithmic specification (0.486 versus 0.455). Comparing Figures 8.7 and 9.1 shows that the general pattern of nonlinearity found in the California district income and test score data is also present in the Massachusetts data. The precise functional forms that best describe this

FIGURE 9.1   Test Scores vs. District Income for Massachusetts Data

The estimated linear regression function does not capture the nonlinear relation between district income and test scores in the Massachusetts data. The estimated linear-log and cubic regression functions are similar for district incomes between $13,000 and $30,000, the region containing most of the observations.
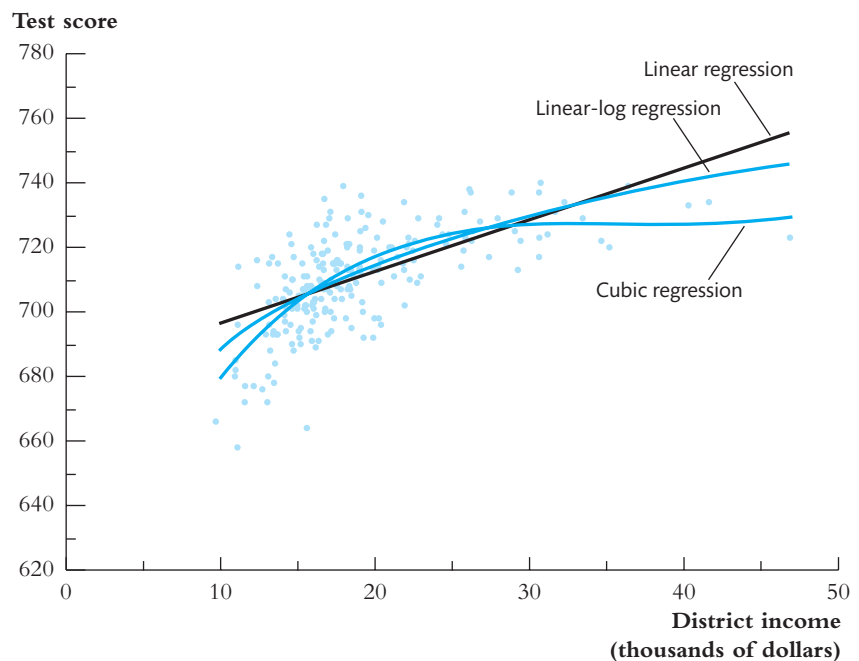
nonlinearity differ, however, with the cubic specification fitting best in Massachusetts but the linear-log specification fitting best in California.

Multiple regression results. Regression results for the Massachusetts data are presented in Table 9.2. The first regression, reported in column (1) in the table, has only the student–teacher ratio as a regressor. The slope is negative (-1.72), and the hypothesis that the coefficient is 0 can be rejected at the 1% significance level $(t = -1.72/0.50 = -3.44)$.

The remaining columns report the results of including additional variables that control for student characteristics and of introducing nonlinearities into the estimated regression function. Controlling for the percentage of English learners, the percentage of students eligible for a subsidized lunch, and the average district income reduces the estimated coefficient on the student–teacher ratio by 60%, from -1.72 in regression (1) to -0.69 in regression (2) and -0.64 in regression (3).

Comparing the $\overline{R}^2$'s of regressions (2) and (3) indicates that the cubic specification (3) provides a better model of the relationship between test scores and district income than does the logarithmic specification (2), even holding constant the student–teacher ratio. There is no statistically significant evidence of a nonlinear relationship between test scores and the student–teacher ratio: The F-statistic in regression (4) testing whether the population coefficients on STR^2 and STR^3 are 0 has a p-value of 0.641. The estimates in regression (5) suggest that a class size reduction is less effective when there are many English learners, the opposite finding from the California data; however, as in the California data, this interaction effect is imprecisely estimated and is not statistically significant at the 10% level [the t-statistic on $HiEL \times STR$ in regression (5) is $0.80/0.56 = 1.43$]. Finally, regression (6) shows that the estimated coefficient on the student–teacher ratio does not change substantially when the percentage of English learners [which is insignificant in regression (3)] is excluded. In short, the results in regression (3) are not sensitive to the changes in functional form and specification considered in regressions (4) through (6) in Table 9.2. Therefore, we adopt regression (3) as our base estimate of the effect on test scores of a change in the student–teacher ratio based on the Massachusetts data.

Comparison of Massachusetts and California results. For the California data, we found the following:

1. Adding variables that control for student background characteristics reduced the coefficient on the student–teacher ratio from -2.28 [Table 7.1, regression (1)] to -0.73 [Table 8.3, regression (2)], a reduction of 68%.
2. The hypothesis that the true coefficient on the student–teacher ratio is 0 was rejected at the 1% significance level, even after adding variables that control for student background and district economic characteristics.

| TABLE 9.2 | Multiple Regression Estimates of the Student–Teacher Ratio and Test Scores: Data from Massachusetts |

Dependent variable: average combined English, math, and science test score in the school district, fourth grade; 220 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
Student–teacher ratio (STR)	−1.72 (0.50) [−2.70, −0.73]	−0.69 (0.27) [−1.22, −0.16]	−0.64 (0.27) [−1.17, −0.11]	12.4 (14.0)	−1.02 (0.37)	−0.67 (0.27) [−1.21, −0.14]
STR^2				−0.680 (0.737)		
STR^3				0.011 (0.013)		
% English learners		−0.411 (0.306)	−0.437 (0.303)	−0.434 (0.300)		
% English learners > median? (Binary, $HiEL$)					−12.6 (9.8)	
$HiEL \times STR$					0.80 (0.56)	
% eligible for free lunch		−0.521 (0.077)	−0.582 (0.097)	−0.587 (0.104)	−0.709 (0.091)	−0.653 (0.72)
District income (logarithm)		16.53 (3.15)				
District income			−3.07 (2.35)	−3.38 (2.49)	−3.87 (2.49)	−3.22 (2.31)
District income^2			0.164 (0.085)	0.174 (0.089)	0.184 (0.090)	0.165 (0.085)
District income^3			−0.0022 (0.0010)	−0.0023 (0.0010)	−0.0023 (0.0010)	−0.0022 (0.0010)

F-Statistics and p-Values Testing Exclusion of Groups of Variables

	(1)	(2)	(3)	(4)	(5)	(6)
All STR variables and interactions $= 0$				2.86 (0.038)	4.01 (0.020)	
$STR^2, STR^3 = 0$				0.45 (0.641)		
$Income^2, Income^3$			7.74 (< 0.001)	7.75 (< 0.001)	5.85 (0.003)	6.55 (0.002)
$HiEL, HiEL \times STR$					1.58 (0.208)	
SER	14.64	8.69	8.61	8.63	8.62	8.64
$\overline{R}^2$	0.063	0.670	0.676	0.675	0.675	0.674

These regressions were estimated using the data on Massachusetts elementary school districts described in Appendix 9.1. All regressions include an intercept (not reported). Standard errors are given in parentheses under the coefficients, and p-values are given in parentheses under the F-statistics. 95% confidence intervals for the coefficient on the student-teacher ratio are presented in brackets for regressions (1), (2), (3), and (6), but not for the regressions with nonlinear terms in STR.

3. The effect of cutting the student–teacher ratio did not depend in a statistically significant way on the percentage of English learners in the district.

4. There is some evidence that the relationship between test scores and the student–teacher ratio is nonlinear.

Do we find the same things in Massachusetts? For findings (1), (2), and (3), the answer is yes. Including the additional control variables reduces the coefficient on the student–teacher ratio from -1.72 [Table 9.2, regression (1)] to -0.69 [Table 9.2, regression (2)], a reduction of 60%. The coefficients on the student–teacher ratio remain significant after adding the control variables. Those coefficients are significant only at the 5% level in the Massachusetts data, whereas they are significant at the 1% level in the California data. However, there are nearly twice as many observations in the California data, so it is not surprising that the California estimates are more precise. As in the California data, there is no statistically significant evidence in the Massachusetts data of an interaction between the student–teacher ratio and the binary variable indicating a large percentage of English learners in the district.

Finding (4), however, does not hold up in the Massachusetts data: The hypothesis that the relationship between the student–teacher ratio and test scores is linear cannot be rejected at the 5% significance level when tested against a cubic specification.

Because the two standardized tests are different, the coefficients themselves cannot be compared directly: One point on the Massachusetts test is not the same as one point on the California test. If, however, the test scores are put into the same units, then the estimated class size effects can be compared. One way to do this is to transform the test scores by standardizing them: Subtract the sample average and divide by the standard deviation so that they have a mean of 0 and a variance of 1. The slope coefficients in the regression with the standardized test score equal the slope coefficients in the original regression divided by the standard deviation of the test. Thus the coefficient on the student–teacher ratio divided by the standard deviation of test scores can be compared across the two data sets.

This comparison is undertaken in Table 9.3. The first column reports the OLS estimates of the coefficient on the student–teacher ratio in a regression with the percentage of English learners, the percentage of students eligible for a subsidized lunch, and the average district income included as control variables. The second column reports the standard deviation of the test scores across districts. The final two columns report the estimated effect on test scores of reducing the student–teacher ratio by two students per teacher (our superintendent's proposal), first in the units of the test and second in standard deviation units. For the linear specification, the OLS coefficient estimate using California data is -0.73, so cutting the student–teacher ratio by two is estimated to increase district test scores by $-0.73 \times (-2) = 1.46$ points. Because the standard deviation of test scores is 19.1 points, this corresponds to $1.46/19.1 = 0.076$ standard deviation units of the

TABLE 9.3	Student–Teacher Ratios and Test Scores: Comparing the Estimates from California and Massachusetts			
			Estimated Effect of Two Fewer Students per Teacher, in Units of:	
	OLS Estimate $\hat{\beta}_{STR}$	**Standard Deviation of Test Scores Across Districts**	**Points on the Test**	**Standard Deviations**
California				
Linear: Table 8.3(2)	−0.73 (0.26)	19.1	1.46 (0.52) [0.46, 2.48]	0.076 (0.027) [0.024, 0.130]
Cubic: Table 8.3(7) *Reduce STR from 20 to 18*	–	19.1	2.93 (0.70) [1.56, 4.30]	0.153 (0.037) [0.081, 0.226]
Cubic: Table 8.3(7) *Reduce STR from 22 to 20*	–	19.1	1.90 (0.69) [0.54, 3.26]	0.099 (0.036) [0.028, 0.171]
Massachusetts				
Linear: Table 9.2(3)	−0.64 (0.27)	15.1	1.28 (0.54) [0.22, 2.34]	0.085 (0.036) [0.015, 0.154]

Standard errors are given in parentheses. 95% confidence intervals for the effect of a two-student reduction are given in brackets.

distribution of test scores across districts. The standard error of this estimate is $0.26 \times 2/19.1 = 0.027$. The estimated effects for the nonlinear models and their standard errors were computed using the method described in Section 8.1.

Based on the linear model using California data, a reduction of two students per teacher is estimated to increase test scores by 0.076 standard deviation units, with a standard error of 0.027. The nonlinear models for California data suggest a somewhat larger effect, with the specific effect depending on the initial student–teacher ratio. Based on the Massachusetts data, this estimated effect is 0.085 standard deviation units, with a standard error of 0.036.

These estimates are essentially the same. The 95% confidence interval for Massachusetts contains the 95% confidence interval for the California linear specification. Cutting the student–teacher ratio is predicted to raise test scores, but the predicted improvement is small. In the California data, for example, the difference in test scores between the median district and a district at the 75th percentile is 12.2 test score points (Table 4.1), or 0.64 ($= 12.2/19.1$) standard deviation units. The estimated effect from the linear model is just over one-tenth this size; in other words, according to this estimate, cutting the student teacher–ratio by two would move a

district only one-tenth of the way from the median to the 75th percentile of the distribution of test scores across districts. Reducing the student–teacher ratio by two is a large change for a district, but the estimated benefits shown in Table 9.3, while nonzero, are small.

This analysis of Massachusetts data suggests that the California results are externally valid, at least when generalized to elementary school districts elsewhere in the United States.

Internal Validity

The similarity of the results for California and Massachusetts does not ensure their *internal* validity. Section 9.2 listed five possible threats to internal validity that could induce bias in the estimated effect on test scores of class size. We consider these threats in turn.

Omitted variables. The multiple regressions reported in this and previous chapters control for a student characteristic (the percentage of English learners), a family economic characteristic (the percentage of students receiving a subsidized lunch), and a broader measure of the affluence of the district (average district income).

If these control variables are adequate, then for the purpose of regression analysis it is as if the student–teacher ratio is randomly assigned among districts with the same values of these control variables, in which case the conditional mean independence assumption holds. There still could be, however, some omitted factors for which these three variables might not be adequate controls. For example, if the student–teacher ratio is correlated with teacher quality even among districts with the same fraction of immigrants and the same socioeconomic characteristics (perhaps because better teachers are attracted to schools with smaller student–teacher ratios) and if teacher quality affects test scores, then omission of teacher quality could bias the coefficient on the student–teacher ratio. Similarly, among districts with the same socioeconomic characteristics, districts with a low student–teacher ratio might have families that are more committed to enhancing their children's learning at home. Such omitted factors could lead to omitted variable bias.

One way to eliminate omitted variable bias, at least in theory, is to conduct an experiment. For example, students could be randomly assigned to different size classes, and their subsequent performance on standardized tests could be compared. Such a study was, in fact, conducted in Tennessee, and we examine it in Chapter 13.

Functional form. The analysis here and in Chapter 8 explored a variety of functional forms. We found that some of the possible nonlinearities investigated were not statistically significant, while those that were did not substantially alter the estimated effect of reducing the student–teacher ratio. Although further functional form analysis could be carried out, this suggests that the main findings of these studies are unlikely to be sensitive to using different nonlinear regression specifications.

Errors in variables. The average student–teacher ratio in the district is a broad and potentially inaccurate measure of class size. For example, because students move in and out of districts, the student–teacher ratio might not accurately represent the actual class sizes experienced by the students taking the test, which in turn could lead to the estimated class size effect being biased toward 0. Another variable with potential measurement error is average district income. Those data were taken from the 1990 Census, while the other data pertain to 1998 (Massachusetts) or 1999 (California). If the economic composition of the district changed substantially over the 1990s, this would be an imprecise measure of the actual average district income.

Sample selection. The California and the Massachusetts data cover all the public elementary school districts in the state that satisfy minimum size restrictions, so there is no reason to believe that sample selection is a problem here.

Simultaneous causality. Simultaneous causality would arise if the performance on standardized tests affected the student–teacher ratio. This could happen, for example, if there is a bureaucratic or political mechanism for increasing the funding of poorly performing schools or districts that in turn resulted in hiring more teachers. In Massachusetts, no such mechanism for equalization of school financing was in place during the time of these tests. In California, a series of court cases led to some equalization of funding, but this redistribution of funds was not based on student achievement. Thus in neither Massachusetts nor California does simultaneous causality appear to be a problem.

Heteroskedasticity and correlation of the error term across observations. All the results reported here and in earlier chapters use heteroskedastic-robust standard errors, so heteroskedasticity does not threaten internal validity. Correlation of the error term across observations, however, could threaten the consistency of the standard errors because simple random sampling was not used (the sample consists of all elementary school districts in the state). Although there are alternative standard error formulas that could be applied to this situation, the details are complicated and specialized, and we leave them to more advanced texts.

Discussion and Implications

The similarity between the Massachusetts and California results suggests that these studies are externally valid in the sense that the main findings can be generalized to performance on standardized tests at other elementary school districts in the United States.

Some of the most important potential threats to internal validity have been addressed by controlling for student background, family economic background, and district affluence and by checking for nonlinearities in the regression function. Still, some potential threats to internal validity remain. A leading candidate is omitted variable bias, perhaps arising because the control variables do not capture other characteristics of the school districts or extracurricular learning opportunities.

Based on both the California and the Massachusetts data, we are able to answer the superintendent's question from Section 4.1: After controlling for family economic background, student characteristics, and district affluence and after modeling nonlinearities in the regression function, cutting the student–teacher ratio by two students per teacher is predicted to increase test scores by approximately 0.08 standard deviations of the distribution of test scores across districts. This effect is statistically significant, but it is quite small. This small estimated effect is in line with the results of the many studies that have investigated the effects on test scores of class size reductions.[5]

The superintendent can now use this estimate to help her decide whether to reduce class sizes. In making this decision, she will need to weigh the costs of the proposed reduction against the benefits. The costs include teacher salaries and expenses for additional classrooms. The benefits include improved academic performance, which we have measured by performance on standardized tests, but there are other potential benefits that we have not studied, including lower dropout rates and enhanced future earnings. The estimated effect of the proposal on standardized test performance is one important input into her calculation of costs and benefits.

9.5 Conclusion

The concepts of internal and external validity provide a framework for assessing what has been learned from an econometric study of causal effects.

A study based on multiple regression is internally valid if the estimated coefficients are unbiased and consistent and if standard errors are consistent. Threats to the internal validity of such a study include omitted variables, misspecification of functional form (nonlinearities), imprecise measurement of the independent variables (errors in variables), sample selection, and simultaneous causality. Each of these introduces correlation between the regressor and the error term, which in turn makes OLS estimators biased and inconsistent. If the errors are correlated across observations, as they can be with time series data, or if they are heteroskedastic but the standard errors are computed using the homoskedasticity-only formula, then internal validity is compromised because the standard errors will be inconsistent. These latter problems can be addressed by computing the standard errors properly.

A study using regression analysis, like any statistical study, is externally valid if its findings can be generalized beyond the population and setting studied. Sometimes it can help to compare two or more studies on the same topic. Whether or not there are two or more such studies, however, assessing external validity requires making judgments about the similarities of the population and setting studied and the population and setting to which the results are being generalized.

---

[5]If you are interested in learning more about the relationship between class size and test scores, see the reviews by Ehrenberg et al. (2001a, 2001b).

The next two parts of this text develop ways to address threats to internal validity that cannot be mitigated by multiple regression analysis alone. Part III extends the multiple regression model in ways designed to mitigate all five sources of potential bias in the OLS estimator. Part III also discusses a different approach to obtaining internal validity, randomized controlled experiments, and it returns to the prediction problem when there are many predictors. Part IV develops methods for analyzing time series data and for using time series data to estimate so-called dynamic causal effects, which are causal effects that vary over time.

Summary

1. Statistical studies are evaluated by asking whether the analysis is internally and externally valid. A study is internally valid if the statistical inferences about causal effects are valid for the population being studied. A study is externally valid if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.
2. In regression estimation of causal effects, there are two types of threats to internal validity. First, OLS estimators are biased and inconsistent if the regressors and error terms are correlated. Second, confidence intervals and hypothesis tests are not valid when the standard errors are incorrect.
3. Regressors and error terms may be correlated when there are omitted variables, an incorrect functional form is used, one or more of the regressors are measured with error, the sample is chosen nonrandomly from the population, or there is simultaneous causality between the regressors and dependent variables.
4. Standard errors are incorrect when the errors are heteroskedastic and the computer software uses the homoskedasticity-only standard errors or when the error term is correlated across different observations.
5. When regression models are used solely for prediction, it is not necessary for the regression coefficients to be unbiased estimates of causal effects. It is critical, however, that the regression model be externally valid for the prediction application at hand.

Key Terms

population studied (288)
population of interest (288)
internal validity (289)
external validity (289)
functional form misspecification (294)
errors-in-variables bias (295)

classical measurement error
   model (295)
sample selection bias (298)
simultaneous causality (299)
simultaneous equations bias (300)

Review the Concepts

9.1   What is the difference between internal validity and external validity? Between the population studied and the population of interest?

9.2   Key Concept 9.2 describes the problem of variable selection in terms of a trade-off between bias and variance. What is this trade-off? Why could including an additional control variable decrease bias? Increase variance?

9.3   Economic variables are often measured with error. Does this mean that regression analysis is unreliable? Explain.

9.4   Suppose that a state offered voluntary standardized tests to all its third graders and that the resulting data were used in a study of class size on student performance. Explain how sample selection bias might invalidate the results.

9.5   A researcher estimates the effect on crime rates of spending on police by using city-level data. Explain how simultaneous causality might invalidate the results.

9.6   A researcher estimates a regression using two different software packages. The first uses the homoskedasticity-only formula for standard errors. The second uses the heteroskedasticity-robust formula. The standard errors are very different. Which should the researcher use? Why?

Exercises

9.1   Suppose you have just read a careful statistical study of the effect of advertising on the demand for cigarettes. Using data from New York during the 1970s, the study concluded that advertising on buses and subways was more effective than print advertising. Use the concept of external validity to determine if these results are likely to apply to Boston in the 1970s, Los Angeles in the 1970s, and New York in 2018.

9.2   Consider the one-variable regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, and suppose it satisfies the least squares assumptions in Key Concept 4.3. Suppose Y_i is measured with error, so the data are $\widetilde{Y}_i = Y_i + w_i$, where w_i is the

measurement error, which is i.i.d. and independent of Y_i and X_i. Consider the population regression $\tilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$, where v_i is the regression error, using the mismeasured dependent variable, $\tilde{Y}_i$.

a. Show that $v_i = u_i + w_i$.

b. Show that the regression $\tilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$ satisfies the least squares assumptions in Key Concept 4.3. (Assume that w_i is independent of Y_j and X_j for all values of i and j and has a finite fourth moment.)

c. Are the OLS estimators consistent?

d. Can confidence intervals be constructed in the usual way?

e. Evaluate these statements: "Measurement error in the X's is a serious problem. Measurement error in Y is not."

9.3 Labor economists studying the determinants of women's earnings discovered a puzzling empirical result. Using randomly selected employed women, they regressed earnings on the women's number of children and a set of control variables (age, education, occupation, and so forth). They found that women with more children had higher wages, controlling for these other factors. Explain how sample selection might be the cause of this result. (*Hint:* Notice that women who do not work outside the home are missing from the sample.) [This empirical puzzle motivated James Heckman's research on sample selection that led to his 2000 Nobel Prize in Economics. See Heckman (1974).]

9.4 For the linear regressions in Table 9.3, compute the estimated effects, standard errors, and confidence intervals for a reduction of three students per class.

9.5 The demand for a commodity is given by $Q = \beta_0 + \beta_1 P + u$, where Q denotes quantity, P denotes price, and u denotes factors other than price that determine demand. Supply for the commodity is given by $Q = \gamma_0 + \gamma_1 P + v$, where v denotes factors other than price that determine supply. Suppose u and v both have a mean of 0, have variances σ_u^2 and σ_v^2, and are mutually uncorrelated.

a. Solve the two simultaneous equations to show how Q and P depend on u and v.

b. Derive the means of P and Q.

c. Derive the variance of P, the variance of Q, and the covariance between Q and P.

d. A random sample of observations of (Q_i, P_i) is collected, and Q_i is regressed on P_i. (That is, Q_i is the regressand, and P_i is the regressor.) Suppose the sample is very large.

  i. Use your answers to (b) and (c) to derive values of the regression coefficients. [*Hint*: Use Equations (4.7) and (4.8).]

  ii. A researcher uses the slope of this regression as an estimate of the slope of the demand function (β_1). Is the estimated slope too large

or too small? (*Hint:* Remember that demand curves slope down and supply curves slope up.)

9.6 Suppose that $n = 100$ i.i.d. observations for (Y_i, X_i) yield the following regression results:

$$\hat{Y} = 32.1 + 66.8X, SER = 15.1, R^2 = 0.81.$$
$$\phantom{\hat{Y} = } (15.1) \ (12.2)$$

Another researcher is interested in the same regression, but he makes an error when he enters the data into his regression program: He enters each observation twice, so he has 200 observations (with observation 1 entered twice, observation 2 entered twice, and so forth).

a. Using these 200 observations, what results will be produced by his regression program? (*Hint:* Write the "incorrect" values of the sample means, variances, and covariances of Y and X as functions of the "correct" values. Use these to determine the regression statistics.)

$$\hat{Y} = \underline{} + \underline{}X, SER = \underline{}, R^2 = \underline{}.$$
$$\phantom{\hat{Y} = } (\underline{}) \ (\underline{})$$

b. Which (if any) of the internal validity conditions are violated?

9.7 Are the following statements true or false? Explain your answers.

a. "An ordinary least squares regression of Y on X will not be internally valid if X is correlated with the error term."

b. "Each of the five primary threats to internal validity implies that X is correlated with the error term."

9.8 Would the regression in Equation (4.9) in chapter 4 be useful for predicting test scores in a school district in Massachusetts? Why or why not?

9.9 Consider the linear regression of *TestScore* on *Income* shown in Figure 8.2 and the nonlinear regression in Equation (8.18). Would either of these regressions provide a reliable estimate of the causal effect of income on test scores? Would either of these regressions provide a reliable method for predicting test scores? Explain.

9.10 Read the box "The Return to Education and the Gender Gap" in Section 8.3. Discuss the internal and external validity of the estimated effect of education on earnings.

9.11 Read the box "The Demand for Economics Journals" in Section 8.3. Discuss the internal and external validity of the estimated effect of price per citation on subscriptions.

9.12 Consider the one-variable regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, and suppose it satisfies the least squares assumptions in Key Concept 4.3. The regressor X_i

is missing, but data on a related variable, Z_i, are available, and the value of X_i is estimated using $\tilde{X}_i = E(X_i | Z_i)$. Let $w_i = \tilde{X}_i - X_i$.

a. Show that $\tilde{X}_i$ is the minimum mean square error estimator of X_i using Z_i. That is, let $\hat{X}_i = g(Z_i)$ be some other guess of X_i based on Z_i, and show that $E[(\hat{X}_i - X_i)^2] \geq E[(\tilde{X}_i - X_i)^2]$. (*Hint*: Review Exercise 2.27.)

b. Show that $E(w_i | \tilde{X}_i) = 0$.

c. Suppose that $E(u_i | Z_i) = 0$ and that $\tilde{X}_i$ is used as the regressor in place of X_i. Show that $\hat{\beta}_1$ is consistent. Is $\hat{\beta}_0$ consistent?

9.13 Assume that the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ satisfies the least squares assumptions in Key Concept 4.3. You and a friend collect a random sample of 300 observations on Y and X.

a. Your friend reports that he inadvertently scrambled the X observations for 20% of the sample. For these scrambled observations, the value of X does not correspond to X_i for the ith observation; rather, it corresponds to the value of X for some other observation. In the notation of Section 9.2, the measured value of the regressor, $\tilde{X}_i$, is equal to X_i for 80% of the observations, but it is equal to a randomly selected X_j for the remaining 20% of the observations. You regress Y_i on $\tilde{X}_i$. Show that $E(\hat{\beta}_1) = 0.8\beta_1$.

b. Explain how you could construct an unbiased estimate of β_1 using the OLS estimator in (a).

c. Suppose now your friend tells you that the X's were scrambled for the first 60 observations but that the remaining 240 observations are correct. You estimate β_1 by regressing Y on X, using only the correctly measured 240 observations. Is this estimator of β_1 better than the estimator you proposed in (b)? Explain.

Empirical Exercises

E9.1 Use the data set **CPS2015**, described in Empirical Exercise 8.2, to answer the following questions.

a. Discuss the internal validity of the regressions that you used to answer Empirical Exercise 8.2(l). Include a discussion of possible omitted variable bias, misspecification of the functional form of the regression, errors in variables, sample selection, simultaneous causality, and inconsistency of the OLS standard errors.

b. The data set **CPS96_15** described in Empirical Exercise 3.1 includes data from 1996 and 2015. Use these data to investigate the (temporal) external validity of the conclusions that you reached in Empirical Exercise 8.2(l). [*Note:* Remember to adjust for inflation, as explained in Empirical Exercise 3.1(b).]

E9.2   Use the data set **Birthweight_Smoking** introduced in Empirical Exercise 5.3 to answer the following questions.

   a. In Empirical Exercise 7.1(f), you estimated several regressions and were asked: "What is a reasonable 95% confidence interval for the effect of smoking on birth weight?"

   i. In Chapter 8, you learned about nonlinear regressions. Can you think of any nonlinear regressions that can potentially improve your answer to Empirical Exercise 7.1(f)? After estimating these additional regressions, what is a reasonable 95% confidence interval for the effect of smoking on birth weight?

   ii. Discuss the internal validity of the regressions you used to construct the confidence interval. Include a discussion of possible omitted variable bias, misspecification of the functional form of the regression, errors in variables, sample selection, simultaneous causality, and inconsistency of the OLS standard errors.

   b. The data set **Birthweight_Smoking** includes babies born in Pennsylvania in 1989. Discuss the external validity of your analysis for (i) California in 1989, (ii) Illinois in 2019, and (iii) South Korea in 2019.

APPENDIX

9.1 The Massachusetts Elementary School Testing Data

The Massachusetts data are district-wide averages for public elementary school districts in 1998. The test score is taken from the Massachusetts Comprehensive Assessment System (MCAS) test administered to all fourth graders in Massachusetts public schools in the spring of 1998. The test is sponsored by the Massachusetts Department of Education and is mandatory for all public schools. The data analyzed here are the overall total score, which is the sum of the scores on the English, math, and science portions of the test.

Data on the student–teacher ratio, the percentage of students receiving a subsidized lunch, and the percentage of students still learning English are averages for each elementary school district for the 1997–1998 school year and were obtained from the Massachusetts Department of Education. Data on average district income were obtained from the 1990 U.S. Census.

10 Regression with Panel Data

Multiple regression is a powerful tool for controlling for the effect of variables on which we have data. If data are not available for some of the variables, however, they cannot be included in the regression, and the OLS estimators of the regression coefficients could have omitted variable bias.

This chapter describes a method for controlling for some types of omitted variables without actually observing them. This method requires a specific type of data, called panel data, in which each observational unit, or entity, is observed at two or more time periods. By studying *changes* in the dependent variable over time, it is possible to eliminate the effect of omitted variables that differ across entities but are constant over time.

The empirical application in this chapter concerns drunk driving: What are the effects of alcohol taxes and drunk driving laws on traffic fatalities? We address this question using data on traffic fatalities, alcohol taxes, drunk driving laws, and related variables for the 48 contiguous U.S. states for each of the seven years from 1982 to 1988. This panel data set lets us control for unobserved variables that differ from one state to the next, such as prevailing cultural attitudes toward drinking and driving, but do not change over time. It also allows us to control for variables that vary through time, like improvements in the safety of new cars, but do not vary across states.

Section 10.1 describes the structure of panel data and introduces the drunk driving data set. Fixed effects regression, the main tool for regression analysis of panel data, is an extension of multiple regression that exploits panel data to control for variables that differ across entities but are constant over time. Fixed effects regression is introduced in Sections 10.2 and 10.3, first for the case of only two time periods and then for multiple time periods. In Section 10.4, these methods are extended to incorporate so-called time fixed effects, which control for unobserved variables that are constant across entities but change over time. Section 10.5 discusses the panel data regression assumptions and standard errors for panel data regression. In Section 10.6, we use these methods to study the effect of alcohol taxes and drunk driving laws on traffic deaths.

Notation for Panel Data

Panel data consist of observations on the same n entities at two or more time periods T, as is illustrated in Table 1.3. If the data set contains observations on the variables X and Y, then the data are denoted

$$(X_{it}, Y_{it}), i = 1, \ldots, n, \text{ and } t = 1, \ldots, T, \tag{10.1}$$

where the first subscript, i, refers to the entity being observed and the second subscript, t, refers to the date at which it is observed.

10.1 Panel Data

Recall from Section 1.3 that **panel data** (also called longitudinal data) refers to data for n different entities observed at T different time periods. The state traffic fatality data studied in this chapter are panel data. Those data are for $n = 48$ entities (states), where each entity is observed in $T = 7$ time periods (each of the years 1982, . . . , 1988), for a total of $7 \times 48 = 336$ observations.

When describing cross-sectional data, it was useful to use a subscript to denote the entity; for example, Y_i referred to the variable Y for the ith entity. When describing panel data, we need some additional notation to keep track of both the entity and the time period. We do so by using two subscripts rather than one: The first, i, refers to the entity, and the second, t, refers to the time period of the observation. Thus Y_{it} denotes the variable Y observed for the ith of n entities in the tth of T periods. This notation is summarized in Key Concept 10.1.

Some additional terminology associated with panel data describes whether some observations are missing. A **balanced panel** has all its observations; that is, the variables are observed for each entity and each time period. A panel that has some missing data for at least one time period for at least one entity is called an **unbalanced panel**. The traffic fatality data set has data for all 48 contiguous U.S. states for all seven years, so it is balanced. If, however, some data were missing (for example, if we did not have data on fatalities for some states in 1983), then the data set would be unbalanced. The methods presented in this chapter are described for a balanced panel; however, all these methods can be used with an unbalanced panel, although precisely how to do so in practice depends on the regression software being used.

Example: Traffic Deaths and Alcohol Taxes

There are approximately 40,000 highway traffic fatalities each year in the United States. Approximately one-fourth of fatal crashes involve a driver who was drinking, and this fraction rises during peak drinking periods. One study (Levitt and Porter,

2001) estimates that as many as 25% of drivers on the road between 1 a.m. and 3 a.m. have been drinking and that a driver who is legally drunk is at least 13 times as likely to cause a fatal crash as a driver who has not been drinking.

In this chapter, we study how effective various government policies designed to discourage drunk driving actually are in reducing traffic deaths. The panel data set contains variables related to traffic fatalities and alcohol, including the number of traffic fatalities in each state in each year, the type of drunk driving laws in each state in each year, and the tax on beer in each state. The measure of traffic deaths we use is the fatality rate, which is the number of annual traffic deaths per 10,000 people in the population in the state. The measure of alcohol taxes we use is the "real" tax on a case of beer, which is the beer tax, put into 1988 dollars by adjusting for inflation.[1] The data are described in more detail in Appendix 10.1.

Figure 10.1a is a scatterplot of the data for 1982 on two of these variables, the fatality rate and the real tax on a case of beer. A point in this scatterplot represents the fatality rate in 1982 and the real beer tax in 1982 for a given state. The OLS regression line obtained by regressing the fatality rate on the real beer tax is also plotted in the figure; the estimated regression line is

$$\widehat{FatalityRate} = 2.01 + 0.15\,BeerTax \quad (1982\ \text{data}).$$
$$\qquad\qquad\quad (0.15)\ \ (0.13)$$

$$(10.2)$$

The coefficient on the real beer tax is positive but not statistically significant at the 10% level.

Because we have data for more than one year, we can reexamine this relationship for another year. This is done in Figure 10.1b, which is the same scatterplot as before except that it uses the data for 1988. The OLS regression line through these data is

$$\widehat{FatalityRate} = 1.86 + 0.44\,BeerTax \quad (1988\ \text{data}).$$
$$\qquad\qquad\quad (0.11)\ \ (0.13)$$

$$(10.3)$$

In contrast to the regression using the 1982 data, the coefficient on the real beer tax is statistically significant at the 1% level (the t-statistic is 3.43). Curiously, the estimated coefficients for the 1982 and the 1988 data are *positive*: Taken literally, higher real beer taxes are associated with *more*, not fewer, traffic fatalities.

Should we conclude that an increase in the tax on beer leads to more traffic deaths? Not necessarily, because these regressions could have substantial omitted variable bias. Many factors affect the fatality rate, including the quality of the automobiles driven in the state, whether the state highways are in good repair, whether most driving is rural or urban, the density of cars on the road, and whether it is socially acceptable to drink and drive. Any of these factors may be correlated

---

[1] To make the taxes comparable over time, they are put into 1988 dollars using the Consumer Price Index (CPI). For example, because of inflation, a tax of $1 in 1982 corresponds to a tax of $1.23 in 1988 dollars.

FIGURE 10.1 The Traffic Fatality Rate and the Tax on Beer

Figure 10.1a is a scatterplot of traffic fatality rates and the real tax on a case of beer (in 1988 dollars) for 48 states in 1982. Figure 10.1b shows the data for 1988. Both plots show a positive relationship between the fatality rate and the real beer tax.

$\widehat{FatalityRate}$ = 2.01 + 0.15 BeerTax

(a) 1982 data

$\widehat{FatalityRate}$ = 1.86 + 0.44 BeerTax
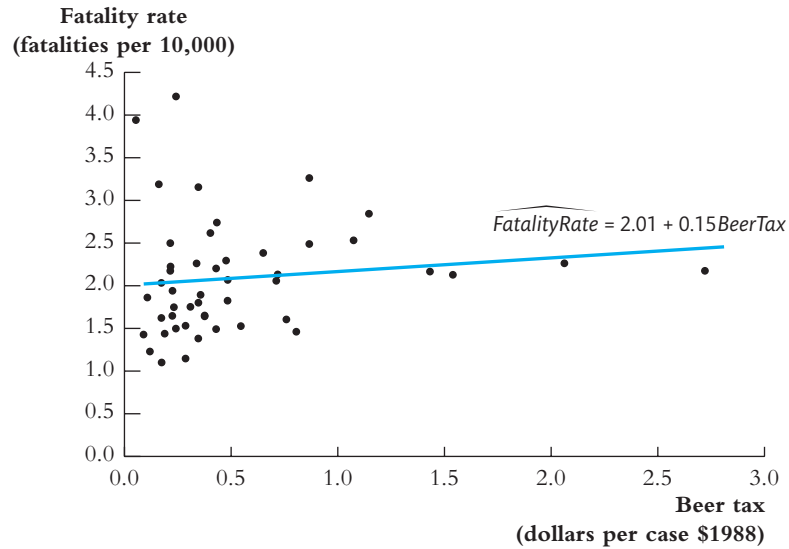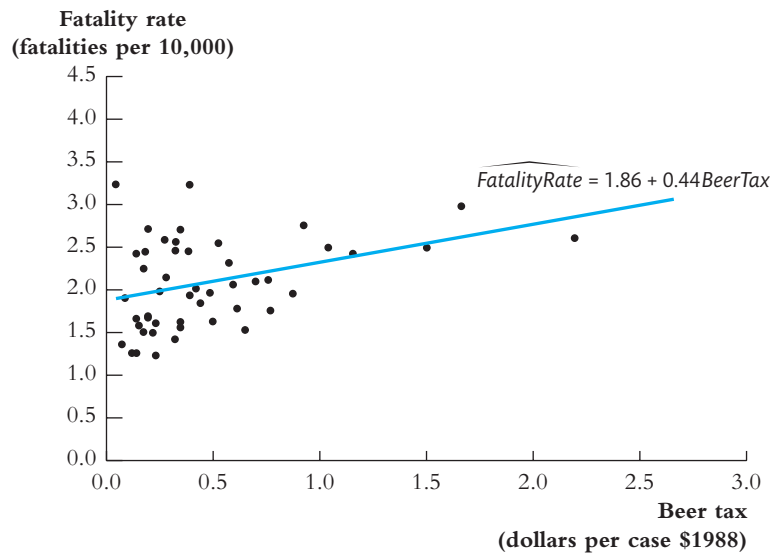
(b) 1988 data

---

with alcohol taxes, and if so, this will lead to omitted variable bias. One approach to these potential sources of omitted variable bias would be to collect data on all these variables and add them to the annual cross-sectional regressions in Equations (10.2) and (10.3). Unfortunately, some of these variables, such as the cultural acceptance of drinking and driving, might be very hard or even impossible to measure.

If these factors remain constant over time in a given state, however, then another route is available. Because we have panel data, we can, in effect, hold these factors constant even though we cannot measure them. To do so, we use OLS regression with fixed effects.

10.2   Panel Data with Two Time Periods: "Before and After" Comparisons

When data for each state are obtained for $T = 2$ time periods, it is possible to compare values of the dependent variable in the second period to values in the first period. By focusing on *changes* in the dependent variable, this "before and after" or "differences" comparison, in effect, holds constant the unobserved factors that differ from one state to the next but do not change over time within the state.

Let Z_i be a variable that determines the fatality rate in the ith state but does not change over time (so the t subscript is omitted). For example, Z_i might be the local cultural attitude toward drinking and driving, which changes slowly and thus could be considered to be constant between 1982 and 1988. Accordingly, the population linear regression relating Z_i and the real beer tax to the fatality rate is

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it}, \qquad (10.4)$$

where u_{it} is the error term, $i = 1, \ldots, n$, and $t = 1, \ldots, T$.

Because Z_i does not change over time, in the regression model in Equation (10.4) it will not produce any *change* in the fatality rate between 1982 and 1988. Thus, in this regression model, the influence of Z_i can be eliminated by analyzing the change in the fatality rate between the two periods. To see this mathematically, consider Equation (10.4) for each of the two years 1982 and 1988:

$$FatalityRate_{i1982} = \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982}, \qquad (10.5)$$

$$FatalityRate_{i1988} = \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988}. \qquad (10.6)$$

Subtracting Equation (10.5) from Equation (10.6) eliminates the effect of Z_i:

$$FatalityRate_{i1988} - FatalityRate_{i1982}$$
$$= \beta_1 (BeerTax_{i1988} - BeerTax_{i1982}) + u_{i1988} - u_{i1982}. \qquad (10.7)$$

This specification has an intuitive interpretation. Cultural attitudes toward drinking and driving affect the level of drunk driving and thus the traffic fatality rate in a state. If, however, they did not change between 1982 and 1988, then they did not produce any *change* in fatalities in the state. Rather, any changes in traffic fatalities over time must have arisen from other sources. In Equation (10.7), these other sources are

changes in the tax on beer and changes in the error term (which captures changes in other factors that determine traffic deaths).

Specifying the regression in changes in Equation (10.7) eliminates the effect of the unobserved variables Z_i that are constant over time. In other words, analyzing changes in Y and X has the effect of controlling for variables that are constant over time, thereby eliminating this source of omitted variable bias.

Figure 10.2 presents a scatterplot of the *change* in the fatality rate between 1982 and 1988 against the *change* in the real beer tax between 1982 and 1988 for the 48 states in our data set. A point in Figure 10.2 represents the change in the fatality rate and the change in the real beer tax between 1982 and 1988 for a given state. The OLS regression line, estimated using these data and plotted in the figure, is

$$\overline{FatalityRate_{1988} - FatalityRate_{1982}} = -0.072 - 1.04(BeerTax_{1988} - BeerTax_{1982}).$$
$$(0.065) \quad (0.36) \tag{10.8}$$

Including an intercept in Equation (10.8) allows for the possibility that the mean change in the fatality rate, in the absence of a change in the real beer tax, is nonzero. For example, the negative intercept (-0.072) could reflect improvements in auto safety between 1982 and 1988 that reduced the average fatality rate.

In contrast to the cross-sectional regression results, the estimated effect of a change in the real beer tax is negative, as predicted by economic theory. The hypothesis that the population slope coefficient is 0 is rejected at the 5% significance level. According to this estimated coefficient, an increase in the real beer tax by $1 per case reduces the traffic fatality rate by 1.04 deaths per 10,000 people. This estimated effect is very large:

FIGURE 10.2 Changes in Fatality Rates and Beer Taxes from 1982 to 1988

This is a scatterplot of the *change* in the traffic fatality rate and the *change* in the real beer tax between 1982 and 1988 for 48 states. There is a negative relationship between changes in the fatality rate and changes in the beer tax.

$FatalityRate_{1988} - FatalityRate_{1982} = -0.072 - 1.04(BeerTax_{1988} - BeerTax_{1982})$

Change in fatality rate (fatalities per 10,000)

Change in beer tax (dollars per case $1988)
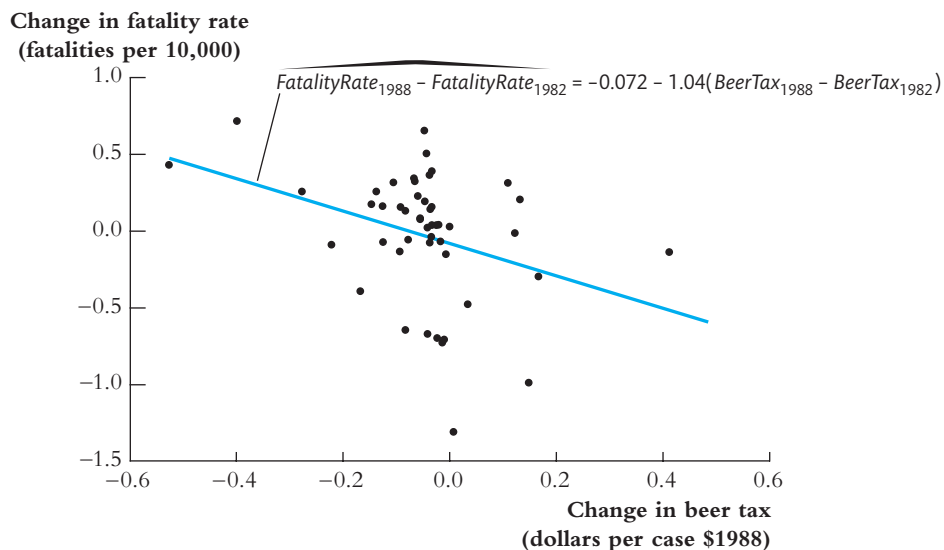
The average fatality rate is approximately 2 in these data (that is, 2 fatalities per year per 10,000 members of the population), so the estimate suggests that traffic fatalities can be cut in half merely by increasing the real tax on beer by $1 per case.

By examining changes in the fatality rate over time, the regression in Equation (10.8) controls for fixed factors such as cultural attitudes toward drinking and driving. But there are many factors that influence traffic safety, and if they change over time and are correlated with the real beer tax, then their omission will produce omitted variable bias. In Section 10.6, we undertake a more careful analysis that controls for several such factors, so for now it is best to refrain from drawing any substantive conclusions about the effect of real beer taxes on traffic fatalities.

This "before and after" or "differences" analysis works when the data are observed in two different years. Our data set, however, contains observations for seven different years, and it seems foolish to discard those potentially useful additional data. But the "before and after" method does not apply directly when $T > 2$. To analyze all the observations in our panel data set, we use the method of fixed effects regression.

10.3 Fixed Effects Regression

Fixed effects regression is a method for controlling for omitted variables in panel data when the omitted variables vary across entities (states) but do not change over time. Unlike the "before and after" comparisons of Section 10.2, fixed effects regression can be used when there are two or more time observations for each entity.

The fixed effects regression model has n different intercepts, one for each entity. These intercepts can be represented by a set of binary (or indicator) variables. These binary variables absorb the influences of all omitted variables that differ from one entity to the next but are constant over time.

The Fixed Effects Regression Model

Consider the regression model in Equation (10.4) with the dependent variable (*FatalityRate*) and observed regressor (*BeerTax*) denoted as Y_{it} and X_{it}, respectively:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}, \qquad (10.9)$$

where Z_i is an unobserved variable that varies from one state to the next but does not change over time (for example, Z_i represents cultural attitudes toward drinking and driving). We want to estimate β_1, the effect on Y of X, holding constant the unobserved state characteristics Z.

Because Z_i varies from one state to the next but is constant over time, the population regression model in Equation (10.9) can be interpreted as having n intercepts, one for each state. Specifically, let $\alpha_i = \beta_0 + \beta_2 Z_i$. Then Equation (10.9) becomes

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}. \qquad (10.10)$$

Equation (10.10) is the **fixed effects regression model**, in which $\alpha_1, \ldots, \alpha_n$ are treated as unknown intercepts to be estimated, one for each state. The interpretation of α_i as a state-specific intercept in Equation (10.10) comes from considering the population regression line for the ith state; this population regression line is $\alpha_i + \beta_1 X_{it}$. The slope coefficient of the population regression line, β_1, is the same for all states, but the intercept of the population regression line varies from one state to the next.

Because the intercept α_i in Equation (10.10) can be thought of as the "effect" of being in entity i (in the current application, entities are states), the terms $\alpha_1, \ldots, \alpha_n$ are known as **entity fixed effects**. The variation in the entity fixed effects comes from omitted variables that, like Z_i in Equation (10.9), vary across entities but not over time.

The state-specific intercepts in the fixed effects regression model also can be expressed using binary variables to denote the individual states. Section 8.3 considered the case in which the observations belong to one of two groups and the population regression line has the same slope for both groups but different intercepts (see Figure 8.8a). That population regression line was expressed mathematically using a single binary variable indicating one of the groups (case 1 in Key Concept 8.4). If we had only two states in our data set, that binary variable regression model would apply here. Because we have more than two states, however, we need additional binary variables to capture all the state-specific intercepts in Equation (10.10).

To develop the fixed effects regression model using binary variables, let $D1_i$ be a binary variable that equals 1 when $i = 1$ and equals 0 otherwise, let $D2_i$ equal 1 when $i = 2$ and equal 0 otherwise, and so on. We cannot include all n binary variables plus a common intercept, for if we do, the regressors will be perfectly multicollinear (this is the dummy variable trap of Section 6.7), so we arbitrarily omit the binary variable $D1_i$ for the first entity. Accordingly, the fixed effects regression model in Equation (10.10) can be written equivalently as

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + \cdots + \gamma_n Dn_i + u_{it}, \qquad (10.11)$$

where $\beta_0, \beta_1, \gamma_2, \ldots, \gamma_n$ are unknown coefficients to be estimated. To derive the relationship between the coefficients in Equation (10.11) and the intercepts in Equation (10.10), compare the population regression lines for each state in the two equations. In Equation (10.11), the population regression equation for the first state is $\beta_0 + \beta_1 X_{it}$, so $\alpha_1 = \beta_0$. For the second and remaining states, it is $\beta_0 + \beta_1 X_{it} + \gamma_i$, so $\alpha_i = \beta_0 + \gamma_i$ for $i \geq 2$.

Thus there are two equivalent ways to write the fixed effects regression model, Equations (10.10) and (10.11). In Equation (10.10), it is written in terms of n state-specific intercepts. In Equation (10.11), the fixed effects regression model has a common intercept and $n - 1$ binary regressors. In both formulations, the slope coefficient on X is the same from one state to the next. The state-specific intercepts in Equation (10.10) and the binary regressors in Equation (10.11) have the same source: the unobserved variable Z_i that varies across states but not over time.

The Fixed Effects Regression Model

The fixed effects regression model is

$$Y_{it} = \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \alpha_i + u_{it}, \tag{10.12}$$

where $i = 1, \ldots, n$; $t = 1, \ldots, T$; $X_{1,it}$ is the value of the first regressor for entity i in time period t, $X_{2,it}$ is the value of the second regressor, and so forth; and $\alpha_1, \ldots, \alpha_n$ are entity-specific intercepts.

Equivalently, the fixed effects regression model can be written in terms of a common intercept, the X's, and $n - 1$ binary variables representing all but one entity:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \gamma_2 D2_i$$
$$+ \gamma_3 D3_i + \cdots + \gamma_n Dn_i + u_{it}, \tag{10.13}$$

where $D2_i = 1$ if $i = 2$ and $D2_i = 0$ otherwise, and so forth.

Extension to multiple X's. If there are other observed determinants of Y that are correlated with X and that change over time, then these should also be included in the regression to avoid omitted variable bias. Doing so results in the fixed effects regression model with multiple regressors, summarized in Key Concept 10.2.

Estimation and Inference

In principle, the binary variable specification of the fixed effects regression model [Equation (10.13)] can be estimated by OLS. This regression, however, has $k + n$ regressors (the k X's, the $n - 1$ binary variables, and the intercept), so in practice this OLS regression is tedious or, in some software packages, impossible to implement if the number of entities is large. Econometric software therefore has special routines for OLS estimation of fixed effects regression models. These special routines are equivalent to using OLS on the full binary variable regression, but they are faster because they employ some mathematical simplifications that arise in the algebra of fixed effects regression.

The "entity-demeaned" OLS algorithm. Regression software typically computes the OLS fixed effects estimator in two steps. In the first step, the entity-specific average is subtracted from each variable. In the second step, the regression is estimated using "entity-demeaned" variables. Specifically, consider the case of a single regressor in the version of the fixed effects model in Equation (10.10), and take the average of both sides of Equation (10.10); then $\overline{Y}_i = \beta_1 \overline{X}_i + \alpha_i + \overline{u}_i$, where $\overline{Y}_i = (1/T)\sum_{t=1}^T Y_{it}$, and $\overline{X}_i$ and $\overline{u}_i$ are defined similarly. Thus Equation (10.10)

implies that $Y_{it} - \overline{Y}_i = \beta_1(X_{it} - \overline{X}_i) + (u_{it} - \overline{u}_i)$. Let $\widetilde{Y}_{it} = Y_{it} - \overline{Y}_i$, $\widetilde{X}_{it} = X_{it} - \overline{X}_i$ and $\widetilde{u}_{it} = u_{it} - \overline{u}_i$; accordingly,

$$\widetilde{Y}_{it} = \beta_1\widetilde{X}_{it} + \widetilde{u}_{it}. \tag{10.14}$$

Thus β_1 can be estimated by the OLS regression of the "entity-demeaned" variables $\widetilde{Y}_{it}$ on $\widetilde{X}_{it}$. In fact, this estimator is identical to the OLS estimator of β_1 obtained by estimation of the fixed effects model in Equation (10.11) using $n - 1$ binary variables (Exercise 19.6).

The "before and after" (differences) regression versus the binary variables specification. Although Equation (10.11) with its binary variables looks quite different from the "before and after" regression model in Equation (10.7), in the special case that $T = 2$ the OLS estimator of β_1 from the binary variable specification and that from the "before and after" specification are identical if the intercept is excluded from the "before and after" specification. Thus, when $T = 2$, there are three ways to estimate β_1 by OLS: the "before and after" specification in Equation (10.7) (without an intercept), the binary variable specification in Equation (10.11), and the entity-demeaned specification in Equation (10.14). These three methods are equivalent; that is, they produce identical OLS estimates of β_1 (Exercise 10.11).

The sampling distribution, standard errors, and statistical inference. In multiple regression with cross-sectional data, if the four least squares assumptions in Key Concept 6.4 hold, then the sampling distribution of the OLS estimator is normal in large samples. The variance of this sampling distribution can be estimated from the data, and the square root of this estimator of the variance—that is, the standard error—can be used to test hypotheses using a *t*-statistic and to construct confidence intervals.

Similarly, in multiple regression with panel data, if a set of assumptions—called the fixed effects regression assumptions—holds, then the sampling distribution of the fixed effects OLS estimator is normal in large samples, the variance of that distribution can be estimated from the data, the square root of that estimator is the standard error, and the standard error can be used to construct *t*-statistics and confidence intervals. Given the standard error, statistical inference—testing hypotheses (including joint hypotheses using *F*-statistics) and constructing confidence intervals—proceeds in exactly the same way as in multiple regression with cross-sectional data.

The fixed effects regression assumptions and standard errors for fixed effects regression are discussed further in Section 10.5.

Application to Traffic Deaths

The OLS estimate of the fixed effects regression line relating the real beer tax to the fatality rate, based on all 7 years of data (336 observations), is

$$\widehat{FatalityRate} = -0.66\,BeerTax + state\ fixed\ effects, \tag{10.15}$$
$$(0.29)$$

where, as is conventional, the estimated state fixed intercepts are not listed to save space and because they are not of primary interest in this application.

Like the "before and after" specification in Equation (10.8), the estimated coefficient in the fixed effects regression in Equation (10.15) is negative, so, as predicted by economic theory, higher real beer taxes are associated with fewer traffic deaths, which is the opposite of what we found in the initial cross-sectional regressions of Equations (10.2) and (10.3). The two regressions are not identical because the "before and after" regression in Equation (10.8) uses only the data for 1982 and 1988 (specifically, the difference between those two years), whereas the fixed effects regression in Equation (10.15) uses the data for all 7 years. Because of the additional observations, the standard error is smaller in Equation (10.15) than in Equation (10.8).

Including state fixed effects in the fatality rate regression lets us avoid omitted variables bias arising from omitted factors, such as cultural attitudes toward drinking and driving, that vary across states but are constant over time within a state. Still, a skeptic might suspect that other factors could lead to omitted variables bias. For example, over this period cars were getting safer, and occupants were increasingly wearing seat belts; if the real tax on beer rose, on average, during the mid-1980s, then *BeerTax* could be picking up the effect of overall automobile safety improvements. If, however, safety improvements evolved over time but were the same for all states, then we can eliminate their influence by including time fixed effects.

10.4   Regression with Time Fixed Effects

Just as fixed effects for each entity can control for variables that are constant over time but differ across entities, so time fixed effects can control for variables that are constant across entities but evolve over time.

Because safety improvements in new cars are introduced nationally, they serve to reduce traffic fatalities in all states. So it is plausible to think of automobile safety as an omitted variable that changes over time but has the same value for all states. The population regression in Equation (10.9) can be modified to make explicit the effect of automobile safety, which we will denote S_t:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}, \tag{10.16}$$

where S_t is unobserved and where the single t subscript emphasizes that safety changes over time but is constant across states. Because $\beta_3 S_t$ represents variables that determine Y_{it}, if S_t is correlated with X_{it}, then omitting S_t from the regression leads to omitted variable bias.

Time Effects Only

For the moment, suppose that the variables Z_i are not present, so that the term $\beta_2 Z_i$ can be dropped from Equation (10.16), although the term $\beta_3 S_t$ remains. Our objective is to estimate β_1, controlling for S_t.

Although S_t is unobserved, its influence can be eliminated because it varies over time but not across states, just as it is possible to eliminate the effect of Z_i, which varies across states but not over time. In the entity fixed effects model, the presence of Z_i leads to the fixed effects regression model in Equation (10.10), in which each state has its own intercept (or fixed effect). Similarly, because S_t varies over time but not over states, the presence of S_t leads to a regression model in which each time period has its own intercept.

The **time fixed effects regression model** with a single X regressor is

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}. \tag{10.17}$$

This model has a different intercept, λ_t, for each time period. The intercept λ_t in Equation (10.17) can be thought of as the "effect" on Y of year t (or, more generally, time period t), so the terms $\lambda_1, \ldots, \lambda_T$ are known as **time fixed effects**. The variation in the time fixed effects comes from omitted variables that, like S_t in Equation (10.16), vary over time but not across entities.

Just as the entity fixed effects regression model can be represented using $n - 1$ binary indicators, so, too, can the time fixed effects regression model be represented using $T - 1$ binary indicators:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_t + \cdots + \delta_T BT_t + u_{it}, \tag{10.18}$$

where $\delta_2, \ldots, \delta_T$ are unknown coefficients and where $B2_t = 1$ if $t = 2$ and $B2_t = 0$ otherwise, and so forth. As in the fixed effects regression model in Equation (10.11), in this version of the time effects model the intercept is included, and the first binary variable $(B1_t)$ is omitted to prevent perfect multicollinearity.

When there are additional observed "X" regressors, then these regressors appear in Equations (10.17) and (10.18) as well.

In the traffic fatalities regression, the time fixed effects specification allows us to eliminate bias arising from omitted variables like nationally introduced safety standards that change over time but are the same across states in a given year.

Both Entity and Time Fixed Effects

If some omitted variables are constant over time but vary across states (such as cultural norms), while others are constant across states but vary over time (such as national safety standards), then it is appropriate to include *both* entity (state) *and* time effects.

The combined **entity and time fixed effects regression model** is

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}, \tag{10.19}$$

where α_i is the entity fixed effect and λ_t is the time fixed effect. This model can equivalently be represented using $n - 1$ entity binary indicators and $T - 1$ time binary indicators, along with an intercept:

$$\begin{aligned} Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \cdots + \gamma_n Dn_i \\ + \delta_2 B2_t + \cdots + \delta_T BT_t + u_{it}, \end{aligned} \tag{10.20}$$

where $\beta_0, \beta_1, \gamma_2, \ldots, \gamma_n$, and $\delta_2, \ldots, \delta_T$ are unknown coefficients.

When there are additional observed "X" regressors, then these appear in Equations (10.19) and (10.20) as well.

The combined entity and time fixed effects regression model eliminates omitted variables bias arising both from unobserved variables that are constant over time and from unobserved variables that are constant across states.

Estimation. The time fixed effects model and the entity and time fixed effects model are both variants of the multiple regression model. Thus their coefficients can be estimated by OLS by including the additional time and entity binary variables. Alternatively, in a balanced panel the coefficients on the X's can be computed by first deviating Y and the X's from their entity *and* time-period means and then by estimating the multiple regression equation of deviated Y on the deviated X's. This algorithm, which is commonly implemented in regression software, eliminates the need to construct the full set of binary indicators that appear in Equation (10.20). An equivalent approach is to deviate Y, the X's, and the time indicators from their entity (but not time-period) means and to estimate $k + T$ coefficients by multiple regression of the deviated Y on the deviated X's and the deviated time indicators. Finally, if $T = 2$, the entity and time fixed effects regression can be estimated using the "before and after" approach of Section 10.2, including the intercept in the regression. Thus the "before and after" regression reported in Equation (10.8), in which the change in *FatalityRate* from 1982 to 1988 is regressed on the change in *BeerTax* from 1982 to 1988 including an intercept, provides the same estimate of the slope coefficient as the OLS regression of *FatalityRate* on *BeerTax*, including entity and time fixed effects, estimated using data for the two years 1982 and 1988.

Application to traffic deaths. Adding time effects to the state fixed effects regression results in the OLS estimate of the regression line:

$$\widehat{FatalityRate} = -0.64 \, BeerTax \, + \, State \ Fixed \ Effects \, + \, Time \ Fixed \ Effects. \quad (10.21)$$
$$(0.36)$$

This specification includes the beer tax, 47 state binary variables (state fixed effects), 6 single-year binary variables (time fixed effects), and an intercept, so this regression actually has $1 + 47 + 6 + 1 = 55$ right-hand variables! The coefficients on the time and state binary variables and the intercept are not reported because they are not of primary interest.

Including time effects has little impact on the coefficient on the real beer tax [compare Equations (10.15) and (10.21)]. Although this coefficient is less precisely estimated when time effects are included, it is still significant at the 10%, but not the 5%, significance level ($t = -0.64/0.36 = -1.78$).

This estimated relationship between the real beer tax and traffic fatalities is immune to omitted variable bias from variables that are constant either over time or across states. However, many important determinants of traffic deaths do not fall into this category, so this specification could still be subject to omitted variable bias.

Section 10.6 therefore undertakes a more complete empirical examination of the effect of the beer tax and of laws aimed directly at eliminating drunk driving, controlling for a variety of factors. Before turning to that study, we first discuss the assumptions underlying panel data regression and the construction of standard errors for fixed effects estimators.

10.5 The Fixed Effects Regression Assumptions and Standard Errors for Fixed Effects Regression

In panel data, the regression error can be correlated over time within an entity. Like heteroskedasticity, this correlation does not introduce bias into the fixed effects estimator, but it affects the variance of the fixed effects estimator, and therefore it affects how one computes standard errors. The standard errors for fixed effects regressions reported in this chapter are so-called clustered standard errors, which are robust both to heteroskedasticity and to correlation over time within an entity. When there are many entities (when n is large), hypothesis tests and confidence intervals can be computed using the usual large-sample normal and F critical values.

This section describes clustered standard errors. We begin with the fixed effects regression assumptions, which extend the least squares regression assumptions for causal inference to panel data; under these assumptions, the fixed effects estimator is consistent and asymptotically normally distributed when n is large. To keep the notation as simple as possible, this section focuses on the entity fixed effects regression model of Section 10.3, in which there are no time effects.

The Fixed Effects Regression Assumptions

The four fixed effects regression assumptions are summarized in Key Concept 10.3. These assumptions extend the four least squares assumptions for causal inference, stated for cross-sectional data in Key Concept 6.4, to panel data.

The first assumption is that the error term has conditional mean 0 given all T values of X for that entity. This assumption plays the same role as the first least squares assumption for cross-sectional data in Key Concept 6.4 and implies that there is no omitted variable bias. The requirement that the conditional mean of u_{it} not depend on *any* of the values of X for that entity—past, present, or future—adds an important subtlety beyond the first least squares assumption for cross-sectional data. This assumption is violated if current u_{it} is correlated with past, present, or future values of X.

The second assumption is that the variables for one entity are distributed identically to, but independently of, the variables for another entity; that is, the variables are i.i.d. across entities for $i = 1, \ldots, n$. Like the second least squares assumption in Key Concept 6.4, the second assumption for fixed effects regression holds if entities are selected by simple random sampling from the population.

The Fixed Effects Regression Assumptions

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}, i = 1, \ldots, n, t = 1, \ldots, T,$$

where β_1 is the causal effect on Y of X and

1. u_{it} has conditional mean 0: $E(u_{it} \mid X_{i1}, X_{i2}, \ldots, X_{iT}, \alpha_i) = 0$.
2. $(X_{i1}, X_{i2}, \ldots, X_{iT}, u_{i1}, u_{i2}, \ldots, u_{iT}), i = 1, \ldots, n$, are i.i.d. draws from their joint distribution.
3. Large outliers are unlikely: (X_{it}, u_{it}) have nonzero finite fourth moments.
4. There is no perfect multicollinearity.

For multiple regressors, X_{it} should be replaced by the full list $X_{1,it}, X_{2,it}, \ldots, X_{k,it}$.

The third and fourth assumptions for fixed effects regression are analogous to the third and fourth least squares assumptions for cross-sectional data in Key Concept 6.4.

Under the least squares assumptions for panel data in Key Concept 10.3, the fixed effects estimator is consistent and is normally distributed when n is large. The details are discussed in Appendix 10.2.

An important difference between the panel data assumptions in Key Concept 10.3 and the assumptions for cross-sectional data in Key Concept 6.4 is assumption 2. The cross-sectional counterpart of assumption 2 holds that each observation is independent, which arises under simple random sampling. In contrast, assumption 2 for panel data holds that the variables are independent across entities but makes no such restriction within an entity. For example, assumption 2 allows X_{it} to be correlated over time within an entity.

If X_{it} is correlated with X_{is} for different values of s and t—that is, if X_{it} is correlated over time for a given entity—then X_{it} is said to be **autocorrelated** (correlated with itself, at different dates) or **serially correlated**. Autocorrelation is a pervasive feature of time series data: What happens one year tends to be correlated with what happens the next year. In the traffic fatality example, X_{it}, the beer tax in state i in year t, is autocorrelated: Most of the time the legislature does not change the beer tax, so if it is high one year relative to its mean value for state i, it will tend to be high the next year, too. Similarly, it is possible to think of reasons why u_{it} would be autocorrelated. Recall that u_{it} consists of time-varying factors that are determinants of Y_{it} but are not included as regressors, and some of these omitted factors might be autocorrelated. For example, a downturn in the local economy might produce layoffs and diminish commuting traffic, thus reducing traffic fatalities for 2 or more years. Similarly, a major road improvement project might reduce traffic accidents not only in the year of completion but also in future years. Such omitted factors, which persist over

multiple years, produce autocorrelated regression errors. Not all omitted factors will produce autocorrelation in u_{it}; for example, severe winter driving conditions plausibly affect fatalities, but if winter weather conditions for a given state are independently distributed from one year to the next, then this component of the error term would be serially uncorrelated. In general, though, as long as some omitted factors are autocorrelated, then u_{it} will be autocorrelated.

Standard Errors for Fixed Effects Regression

If the regression errors are autocorrelated, then the usual heteroskedasticity-robust standard error formula for cross-section regression [Equations (5.3) and (5.4)] is not valid. One way to see this is to draw an analogy to heteroskedasticity. In a regression with cross-sectional data, if the errors are heteroskedastic, then (as discussed in Section 5.4) the homoskedasticity-only standard errors are not valid because they were derived under the false assumption of homoskedasticity. Similarly, if the errors are autocorrelated, then the usual standard errors will not be valid because they were derived under the false assumption of no serial correlation.

Standard errors that are valid if u_{it} is potentially heteroskedastic and potentially correlated over time within an entity are referred to as **heteroskedasticity-and autocorrelation-robust (HAR) standard errors**. The standard errors used in this chapter are one type of HAR standard errors, **clustered standard errors**. The term *clustered* arises because these standard errors allow the regression errors to have an arbitrary correlation within a cluster, or grouping, but assume that the regression errors are uncorrelated across clusters. In the context of panel data, each cluster consists of an entity. Thus clustered standard errors allow for heteroskedasticity and for arbitrary autocorrelation within an entity but treat the errors as uncorrelated across entities. That is, clustered standard errors allow for heteroskedasticity and autocorrelation in a way that is consistent with the second fixed effects regression assumption in Key Concept 10.3.

Like heteroskedasticity-robust standard errors in regression with cross-sectional data, clustered standard errors are valid whether or not there is heteroskedasticity, autocorrelation, or both. If the number of entities n is large, inference using clustered standard errors can proceed using the usual large-sample normal critical values for t-statistics and $F_{q,\infty}$ critical values for F-statistics testing q restrictions.

In practice, there can be a large difference between clustered standard errors and standard errors that do not allow for autocorrelation of u_{it}. For example, the usual (cross-sectional data) heteroskedasticity-robust standard error for the *Beer-Tax* coefficient in Equation (10.21) is 0.25, substantially smaller than the clustered standard error, 0.36, and the respective t-statistics testing $\beta_1 = 0$ are -2.51 and -1.78. The reason we report the clustered standard error is that it allows for serial correlation of u_{it} within an entity, whereas the usual heteroskedasticity-robust standard error does not. The formula for clustered standard errors is given in Appendix 10.2.

10.6 Drunk Driving Laws and Traffic Deaths

Alcohol taxes are only one way to discourage drinking and driving. States differ in their punishments for drunk driving, and a state that cracks down on drunk driving could do so by toughening driving laws as well as raising taxes. If so, omitting these laws could produce omitted variable bias in the OLS estimator of the effect of real beer taxes on traffic fatalities, even in regressions with state and time fixed effects. In addition, because vehicle use depends in part on whether drivers have jobs and because tax changes can reflect economic conditions (a state budget deficit can lead to tax hikes), omitting state economic conditions also could result in omitted variable bias. In this section, we therefore extend the preceding analysis of traffic fatalities to include other driving laws and economic conditions.

The results are summarized in Table 10.1. The format of the table is the same as that of the tables of regression results in Chapters 7 through 9: Each column reports a different regression, and each row reports a coefficient estimate and standard error, a 95% confidence interval for the coefficients on the policy variables of interest, a F-statistic and p-value, or other information about the regression.

Column (1) in Table 10.1 presents results for the OLS regression of the fatality rate on the real beer tax without state and time fixed effects. As in the cross-sectional regressions for 1982 and 1988 [Equations (10.2) and (10.3)], the coefficient on the real beer tax is *positive* (0.36): According to this estimate, increasing beer taxes *increases* traffic fatalities! However, the regression in column (2) [reported previously as Equation (10.15)], which includes state fixed effects, suggests that the positive coefficient in regression (1) is the result of omitted variable bias (the coefficient on the real beer tax is −0.66). The regression $\overline{R}^2$ jumps from 0.091 to 0.889 when fixed effects are included; evidently, the state fixed effects account for a large amount of the variation in the data.

Little changes when time effects are added, as reported in column (3) [reported previously as Equation (10.21)], except that the beer tax coefficient is now estimated less precisely. The results in columns (1) through (3) are consistent with the omitted fixed factors—historical and cultural factors, general road conditions, population density, attitudes toward drinking and driving, and so forth—being important determinants of the variation in traffic fatalities across states.

The next four regressions in Table 10.1 include additional potential determinants of fatality rates along with state and time effects. The base specification, reported in column (4), includes variables related to drunk driving laws plus variables that control for the amount of driving and overall state economic conditions. The first legal variables are the minimum legal drinking age, represented by three binary variables for a minimum legal drinking age of 18, 19, and 20 (so the omitted group is a minimum legal drinking age of 21 or older). The other legal variable is the punishment associated with the first conviction for driving under the influence of alcohol, either mandatory jail time or mandatory community service (the omitted group is less

TABLE 10.1	Regression Analysis of the Effect of Drunk Driving Laws on Traffic Deaths						
Dependent variable: traffic fatality rate (deaths per 10,000).							
Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36 (0.05) [0.26, 0.46]	−0.66 (0.29) [−1.23, −0.09]	−0.64 (0.36) [−1.35, 0.07]	−0.45 (0.30) [−1.04, 0.14]	−0.69 (0.35) [−1.38, 0.00]	−0.46 (0.31) [−1.07, 0.15]	−0.93 (0.34) [−1.60, −0.26]
Drinking age 18		0.10		0.03 (0.07) [−0.11, 0.17]	−0.01 (0.08) [−0.17, 0.15]		0.04 (0.10) [−0.16, 0.24]
Drinking age 19				−0.02 (0.05) [−0.12, 0.08]	−0.08 (0.07) [−0.21, 0.06]		−0.07 (0.10) [−0.26, 0.13]
Drinking age 20				0.03 (0.05) [−0.07, 0.13]	−0.10 (0.06) [−0.21, 0.01]		−0.11 (0.13) [−0.36, 0.14]
Drinking age						0.00 (0.02) [−0.05, 0.04]	
Mandatory jail or community service?				0.04 (0.10) [−0.17, 0.25]	0.09 (0.11) [−0.14, 0.31]	0.04 (0.10) [−0.17, 0.25]	0.09 (0.16) [−0.24, 0.42]
Average vehicle miles per driver				0.008 (0.007)	0.017 (0.011)	0.009 (0.007)	0.124 (0.049)
Unemployment rate				−0.063 (0.013)		−0.063 (0.013)	−0.091 (0.021)
Real income per capita (logarithm)				1.82 (0.64)		1.79 (0.64)	1.00 (0.68)
Years	1982–88	1982–88	1982–88	1982–88	1982–88	1982–88	1982 & 1988 only
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	yes	yes	yes	yes	yes	yes
F-Statistics and p-Values Testing Exclusion of Groups of Variables							
Time effects = 0			4.22 (0.002)	10.12 (<0.001)	3.48 (0.006)	10.28 (<0.001)	37.49 (<0.001)
Drinking age coefficients = 0				0.35 (0.786)	1.41 (0.253)		0.42 (0.738)
Unemployment rate, income per capita = 0				29.62 (<0.001)		31.96 (<0.001)	25.20 (<0.001)
$\bar{R}^2$	0.091	0.889	0.891	0.926	0.893	0.926	0.899

These regressions were estimated using panel data for 48 U.S. states. Regressions (1) through (6) use data for all years 1982 to 1988, and regression (7) uses data from 1982 and 1988 only. The data set is described in Appendix 10.1. Standard errors are given in parentheses under the coefficients, 95% confidence intervals are given in square brackets under the coefficients, and p-values are given in parentheses under the F-statistics.

severe punishment). The three measures of driving and economic conditions are average vehicle miles per driver, the unemployment rate, and the logarithm of real (1988 dollars) personal income per capita (using the logarithm of income permits the coefficient to be interpreted in terms of percentage changes of income; see Section 8.2). The final regression in Table 10.1 follows the "before and after" approach of Section 10.2 and uses only data from 1982 and 1988; thus regression (7) extends the regression in Equation (10.8) to include the additional regressors.

   The regression in column (4) has four interesting results.

1. Including the additional variables reduces the estimated effect of the beer tax from -0.64 in column (3) to -0.45 in column (4). One way to evaluate the magnitude of this coefficient is to imagine a state with an average real beer tax doubling its tax; because the average real beer tax in these data is approximately \$0.50 per case (in 1988 dollars), this entails increasing the tax by \$0.50 per case. The estimated effect of a \$0.50 increase in the beer tax is to decrease the expected fatality rate by $0.45 \times 0.50 = 0.23$ deaths per 10,000. This estimated effect is large: Because the average fatality rate is 2 deaths per 10,000, a reduction of 0.23 corresponds to reducing traffic deaths by nearly one-eighth. This said, the estimate is quite imprecise: Because the standard error on this coefficient is 0.30, the 95% confidence interval for this effect is $-0.45 \times 0.50 \pm 1.96 \times 0.30 \times 0.50 = (-0.52, 0.08)$. This wide 95% confidence interval includes 0, so the hypothesis that the beer tax has no effect cannot be rejected at the 5% significance level.

2. The minimum legal drinking age is precisely estimated to have a small effect on traffic fatalities. According to the regression in column (4), the 95% confidence interval for the increase in the fatality rate in a state with a minimum legal drinking age of 18, relative to age 21, is $(-0.11, 0.17)$. The joint hypothesis that the coefficients on the minimum legal drinking age variables are 0 cannot be rejected at the 10% significance level: The F-statistic testing the joint hypothesis that the three coefficients are 0 is 0.35, with a p-value of 0.786.

3. The coefficient on the first offense punishment variable is also estimated to be small and is not significantly different from 0 at the 10% significance level.

4. The economic variables have considerable explanatory power for traffic fatalities. High unemployment rates are associated with fewer fatalities: An increase in the unemployment rate by 1 percentage point is estimated to reduce traffic fatalities by 0.063 deaths per 10,000. Similarly, high values of real per capita income are associated with high fatalities: The coefficient is 1.82, so a 1% increase in real per capita income is associated with an increase in traffic fatalities of 0.0182 deaths per 10,000 (see case I in Key Concept 8.2 for interpretation of this coefficient). According to these estimates, good economic conditions are associated with higher fatalities, perhaps because of increased traffic density when the unemployment rate is low or greater alcohol consumption when

income is high. The two economic variables are jointly significant at the 0.1% significance level (the *F*-statistic is 29.62).

Columns (5) through (7) of Table 10.1 report regressions that check the sensitivity of these conclusions to changes in the base specification. The regression in column (5) drops the variables that control for economic conditions. The result is an increase in the estimated effect of the real beer tax, which becomes significant at the 5% level, but there is no appreciable change in the other coefficients. The sensitivity of the estimated beer tax coefficient to including the economic variables, combined with the statistical significance of the coefficients on those variables in column (4), indicates that the economic variables should remain in the base specification. The regression in column (6) shows that the results in column (4) are not sensitive to changing the functional form when the three drinking age indicator variables are replaced by the drinking age itself. When the coefficients are estimated using the changes of the variables from 1982 to 1988 [column (7)], as in Section 10.2, the findings from column (4) are largely unchanged except that the coefficient on the beer tax is larger and is significant at the 1% level.

The strength of this analysis is that including state and time fixed effects mitigates the threat of omitted variable bias arising from unobserved variables that either do not change over time (like cultural attitudes toward drinking and driving) or do not vary across states (like safety innovations). As always, however, it is important to think about possible threats to validity. One potential source of omitted variable bias is that the measure of alcohol taxes used here, the real tax on beer, could move with other alcohol taxes, which suggests interpreting the results as pertaining more broadly than just to beer. A subtler possibility is that hikes in the real beer tax could be associated with public education campaigns. If so, changes in the real beer tax could pick up the effect of a broader campaign to reduce drunk driving.

Taken together, these results present a provocative picture of measures to control drunk driving and traffic fatalities. According to these estimates, neither stiff punishments nor increases in the minimum legal drinking age have important effects on fatalities. In contrast, there is evidence that increasing alcohol taxes, as measured by the real tax on beer, does reduce traffic deaths, presumably through reduced alcohol consumption. The imprecision of the estimated beer tax coefficient means, however, that we should be cautious about drawing policy conclusions from this analysis and that additional research is warranted.[2]

---

[2]For further analysis of these data, see Ruhm (1996). A meta-analysis by Wagenaar, Salois, and Komro (2009) of 112 studies of the effect of alcohol prices and taxes on consumption found elasticities of −0.46 for beer, −0.69 for wine, and −0.80 for spirits and concluded that alcohol taxes have large effects on reducing consumption relative to other programs. Carpenter and Dobkin (2011) provide evidence that, in contrast to the findings here, raising the minimum legal drinking age substantially reduces fatalities among drivers in the affected age range, especially at night, although they do not control for the other variables in Table 10.1. To learn more about drunk driving and alcohol and about the economics of alcohol more generally, also see Cook and Moore (2000), Chaloupka, Grossman, and Saffer (2002), Young and Bielinska-Kwapisz (2006), and Dang (2008).

10.7 Conclusion

This chapter showed how multiple observations over time on the same entity can be used to control for unobserved omitted variables that differ across entities but are constant over time. The key insight is that if the unobserved variable does not change over time, then any changes in the dependent variable must be due to influences other than these fixed characteristics. If cultural attitudes toward drinking and driving do not change appreciably over 7 years within a state, then explanations for changes in the traffic fatality rate over those 7 years must lie elsewhere.

To exploit this insight, you need data in which the same entity is observed at two or more time periods; that is, you need panel data. With panel data, the multiple regression model of Part II can be extended to include a full set of entity binary variables; this is the fixed effects regression model, which can be estimated by OLS. A twist on the fixed effects regression model is to include time fixed effects, which control for unobserved variables that change over time but are constant across entities. Both entity and time fixed effects can be included in the regression to control for variables that vary across entities but are constant over time and for variables that vary over time but are constant across entities.

Despite these virtues, entity and time fixed effects regression cannot control for omitted variables that vary *both* across entities *and* over time. And, obviously, panel data methods require panel data, which often are not available. Thus there remains a need for a method that can eliminate the influence of unobserved omitted variables when panel data methods cannot do the job. A powerful and general method for doing so is instrumental variables regression, the topic of Chapter 12.

Summary

1. Panel data consist of observations on multiple (n) entities—states, firms, people, and so forth—where each entity is observed at two or more time periods (T).
2. Regression with entity fixed effects controls for unobserved variables that differ from one entity to the next but remain constant over time.
3. When there are two time periods, fixed effects regression can be estimated by a "before and after" regression of the change in Y from the first period to the second on the corresponding change in X.
4. Entity fixed effects regression can be estimated by including binary variables for $n - 1$ entities plus the observable independent variables (the X's) and an intercept.
5. Time fixed effects control for unobserved variables that are the same across entities but vary over time.
6. A regression with time and entity fixed effects can be estimated by including binary variables for $n - 1$ entities and binary variables for $T - 1$ time periods plus the X's and an intercept.

7. In panel data, variables are typically autocorrelated—that is, correlated over time within an entity. Standard errors need to allow both for this autocorrelation and for potential heteroskedasticity, and one way to do so is to use clustered standard errors.

Key Terms

panel data (320)
balanced panel (320)
unbalanced panel (320)
fixed effects regression model (326)
entity fixed effects (326)
time fixed effects regression model (330)
time fixed effects (330)

entity and time fixed effects regression model (330)
autocorrelated (333)
serially correlated (333)
heteroskedasticity-and autocorrelation-robust (HAR) standard errors (334)
clustered standard errors (334)

Review the Concepts

10.1 Why is it necessary to use two subscripts, i and t, to describe panel data? What does i refer to? What does t refer to?

10.2 A researcher is using a panel data set on $n = 1000$ workers over $T = 10$ years (from 2008 through 2017) that contains the workers' earnings, sex, education, and age. The researcher is interested in the effect of education on earnings. Give some examples of unobserved person-specific variables that are correlated with both education and earnings. Can you think of examples of time-specific variables that might be correlated with education and earnings? How would you control for these person-specific and time-specific effects in a panel data regression?

10.3 Can the regression that you suggested in response to Question 10.2 be used to estimate the effect of a worker's sex on his or her earnings? Can that regression be used to estimate the effect of the national unemployment rate on an individual's earnings? Explain.

10.4 In the context of the regression you suggested for Question 10.2, explain why the regression error for a given individual might be serially correlated.

Exercises

10.1 This exercise refers to the drunk driving panel data regressions summarized in Table 10.1.

  a. New Jersey has a population of 8.1 million people. Suppose New Jersey increased the tax on a case of beer by $1 (in 1988 dollars). Use the results in column (4) to predict the number of lives that would be saved over the next year. Construct a 95% confidence interval for your answer.

  b. The drinking age in New Jersey is 21. Suppose New Jersey lowered its drinking age to 18. Use the results in column (4) to predict the change in the number of traffic fatalities in the next year. Construct a 95% confidence interval for your answer.

  c. Should time effects be included in the regression? Why or why not?

  d. A researcher conjectures that the unemployment rate has a different effect on traffic fatalities in the western states than in the other states. How would you test this hypothesis? (Be specific about the specification of the regression and the statistical test you would use.)

10.2 Consider the binary variable version of the fixed effects model in Equation (10.11) except with an additional regressor, $D1_i$; that is, let

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_1 D1_i + \gamma_2 D2_i + \cdots + \gamma_n Dn_i + u_{it}.$$

  a. Suppose that $n = 3$. Show that the binary regressors and the "constant" regressor are perfectly multicollinear; that is, express one of the variables $D1_i$, $D2_i$, $D3_i$, and $X_{0,it}$ as a perfect linear function of the others, where $X_{0,it} = 1$ for all i, t.

  b. Show the result in (a) for general n.

  c. What will happen if you try to estimate the coefficients of the regression by OLS?

10.3 Section 9.2 gave a list of five potential threats to the internal validity of a regression study. Apply that list to the empirical analysis in Section 10.6 and thereby draw conclusions about its internal validity.

10.4 Using the regression in Equation (10.11), what are the slope and intercept for

  a. Entity 1 in time period 1?

  b. Entity 1 in time period 3?

  c. Entity 3 in time period 1?

  d. Entity 3 in time period 3?

10.5   Consider the model with a single regressor. This model also can be written as

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \delta_2 B2_t + \cdots + \delta_T BT_t + \gamma_2 D2_i + \cdots + \gamma_n Dn_i + u_{it},$$

where $B2_t = 1$ if $t = 2$ and 0 otherwise, $D2_i = 1$ if $i = 2$ and 0 otherwise, and so forth. How are the coefficients $(\beta_0, \delta_2, \ldots, \delta_T, \gamma_2, \ldots, \gamma_n)$ related to the coefficients $(\alpha_1, \ldots, \alpha_n, \lambda_1, \ldots, \lambda_T)$?

10.6   Do the fixed effects regression assumptions in Key Concept 10.3 imply that $\text{cov}(\tilde{v}_{it}, \tilde{v}_{is}) = 0$ for $t \neq s$ in Equation (10.28)? Explain.

10.7   A researcher believes that traffic fatalities increase when roads are icy and thinks that therefore states with more snow will have more fatalities than other states. Comment on the following methods designed to estimate the effect of snow on fatalities:

a. The researcher collects data on the average snowfall for each state and adds this regressor ($AverageSnow_i$) to the regressions given in Table 10.1.

b. The researcher collects data on the snowfall in each state for each year in the sample and adds this regressor to the regressions.

10.8   Consider observations (Y_{it}, X_{it}) from the linear panel data model

$$Y_{it} = X_{it}\beta_1 + \alpha_i + \lambda_i t + u_{it},$$

where $t = 1, \ldots, T; i = 1, \ldots, n;$ and $\alpha_i + \lambda_i t$ is an unobserved entity-specific time trend. How would you estimate β_1?

10.9   **a.** In the fixed effects regression model, are the fixed entity effects, α_i, consistently estimated as $n \longrightarrow \infty$ with T fixed? (*Hint:* Analyze the model with no X's: $Y_{it} = \alpha_i + u_{it}$.)

   b. If n is large (say, $n = 2000$) but T is small (say, $T = 4$), do you think that the estimated values of α_i are approximately normally distributed? Why or why not? (*Hint:* Analyze the model $Y_{it} = \alpha_i + u_{it}$.)

10.10 In a study of the effect on earnings of education using panel data on annual earnings for a large number of workers, a researcher regresses earnings in a given year on age, education, union status, and the worker's earnings in the previous year, using fixed effects regression. Will this regression give reliable estimates of the effects of the regressors (age, education, union status, and previous year's earnings) on earnings? Explain. (*Hint:* Check the fixed effects regression assumptions in Section 10.5.)

10.11 Let $\hat{\beta}_1^{DM}$ denote the entity-demeaned estimator given in Equation (10.22), and let $\hat{\beta}_1^{BA}$ denote the "before and after" estimator without an intercept, so that $\hat{\beta}_1^{BA} = [\Sigma_{i=1}^n (X_{i2} - X_{i1})(Y_{i2} - Y_{i1})]/[\Sigma_{i=1}^n (X_{i2} - X_{i1})^2]$. Show that, if

$T = 2, \hat{\beta}_1^{DM} = \hat{\beta}_1^{BA}$. [*Hint:* Use the definition of $\tilde{X}_{it}$ before Equation (10.22) to show that $\tilde{X}_{i1} = -\frac{1}{2}(X_{i2} - X_{i1})$ and $\tilde{X}_{i2} = \frac{1}{2}(X_{i2} - X_{i1})$.]

Empirical Exercises

E10.1   Some U.S. states have enacted laws that allow citizens to carry concealed weapons. These laws are known as "shall-issue" laws because they instruct local authorities to issue a concealed weapons permit to all applicants who are citizens, are mentally competent, and have not been convicted of a felony. (Some states have some additional restrictions.) Proponents argue that if more people carry concealed weapons, crime will decline because criminals will be deterred from attacking other people. Opponents argue that crime will increase because of accidental or spontaneous use of the weapons. In this exercise, you will analyze the effect of concealed weapons laws on violent crimes. On the text website, **http://www.pearsonhighered.com/stock_watson**, you will find the data file **Guns**, which contains a balanced panel of data from the 50 U.S. states plus the District of Columbia for the years 1977 through 1999.[3] A detailed description is given in **Guns_Description**, available on the website.

a. Estimate (1) a regression of ln(*vio*) against *shall* and (2) a regression of ln(*vio*) against *shall*, *incarc_rate*, *density*, *avginc*, *pop*, *pb1064*, *pw1064*, and *pm1029*.

    i. Interpret the coefficient on *shall* in regression (2). Is this estimate large or small in a real-world sense?

    ii. Does adding the control variables in regression (2) change the estimated effect of a shall-issue law in regression (1) as measured by statistical significance? As measured by the real-world significance of the estimated coefficient?

    iii. Suggest a variable that varies across states but plausibly varies little—or not at all—over time and that could cause omitted variable bias in regression (2).

b. Do the results change when you add fixed state effects? If so, which set of regression results is more credible, and why?

c. Do the results change when you add fixed time effects? If so, which set of regression results is more credible, and why?

d. Repeat the analysis using ln(*rob*) and ln(*mur*) in place of ln(*vio*).

---

[3]These data were provided by Professor John Donohue of Stanford University and were used in his paper with Ian Ayres, "Shooting Down the 'More Guns Less Crime' Hypothesis," *Stanford Law Review*, 2003, 55: 1193–1312.

     e. In your view, what are the most important remaining threats to the internal validity of this regression analysis?

     f. Based on your analysis, what conclusions would you draw about the effects of concealed weapons laws on these crime rates?

E10.2 Do citizens demand more democracy and political freedom as their incomes grow? That is, is democracy a normal good? On the text website, **http://www.pearsonhighered.com/stock_watson**, you will find the data file **Income_Democracy**, which contains a panel data set from 195 countries for the years 1960, 1965, . . . , 2000. A detailed description is given in **Income_Democracy_Description**, available on the website.[4] The data set contains an index of political freedom/democracy for each country in each year, together with data on each country's income and various demographic controls. (The income and demographic controls are lagged five years relative to the democracy index to allow time for democracy to adjust to changes in these variables.)

     a. Is the data set a balanced panel? Explain.

     b. The index of political freedom/democracy is labeled *Dem_ind*.

        i. What are the minimum and maximum values of *Dem_ind* in the data set? What are the mean and standard deviation of *Dem_ind* in the data set? What are the 10th, 25th, 50th, 75th, and 90th percentiles of its distribution?

        ii. What is the value of *Dem_ind* for the United States in 2000? Averaged over all years in the data set?

        iii. What is the value of *Dem_ind* for Libya in 2000? Averaged over all years in the data set?

        iv. List five countries with an average value of *Dem_ind* greater than 0.95; less than 0.10; and between 0.3 and 0.7.

     c. The logarithm of per capita income is labeled *Log_GDPPC*. Regress *Dem_ind* on *Log_GDPPC*. Use standard errors that are clustered by country.

        i. How large is the estimated coefficient on *Log_GDPPC*? Is the coefficient statistically significant?

        ii. If per capita income in a country increases by 20%, by how much is *Dem_ind* predicted to increase? What is a 95% confidence interval for the prediction? Is the predicted increase in *Dem_ind* large or small? (Explain what you mean by large or small.)

---

[4] These data were provided by Daron Acemoglu of M.I.T. and were used in his paper with Simon Johnson, James Robinson, and Pierre Yared, "Income and Democracy," *American Economic Review*, 2008, 98:3, 808–842.

      iii. Why is it important to use clustered standard errors for the regression? Do the results change if you do not use clustered standard errors?

d.   i. Suggest a variable that varies across countries but plausibly varies little—or not at all—over time and that could cause omitted variable bias in the regression in (c).

      ii. Estimate the regression in (c), allowing for country fixed effects. How do your answers to (c)(i) and (c)(ii) change?

     iii. Exclude the data for Azerbaijan, and rerun the regression. Do the results change? Why or why not?

     iv. Suggest a variable that varies over time but plausibly varies little—or not at all—across countries and that could cause omitted variable bias in the regression in (c).

      v. Estimate the regression in (c), allowing for time and country fixed effects. How do your answers to (c)(i) and (c)(ii) change?

     vi. There are additional demographic controls in the data set. Should these variables be included in the regression? If so, how do the results change when they are included?

e. Based on your analysis, what conclusions do you draw about the effects of income on democracy?

10.1 The State Traffic Fatality Data Set

The data are for the contiguous 48 U.S. states (excluding Alaska and Hawaii), annually for 1982 through 1988. The traffic fatality rate is the number of traffic deaths in a given state in a given year per 10,000 people living in that state in that year. Traffic fatality data were obtained from the U.S. Department of Transportation Fatal Accident Reporting System. The beer tax (the tax on a case of beer) was obtained from Beer Institute's *Brewers Almanac*. The drinking age variables in Table 10.1 are binary variables indicating whether the legal drinking age is 18, 19, or 20. The binary punishment variable in Table 10.1 describes the state's minimum sentencing requirements for an initial drunk driving conviction: This variable equals 1 if the state requires jail time or community service and equals 0 otherwise (a lesser punishment). Data on the total vehicle miles traveled annually by state were obtained from the Department of Transportation. Personal income data were obtained from the U.S. Bureau of Economic Analysis, and the unemployment rate was obtained from the U.S. Bureau of Labor Statistics.

These data were graciously provided by Professor Christopher J. Ruhm of the Department of Economics at the University of North Carolina.

10.2 Standard Errors for Fixed Effects Regression

This appendix provides formulas for clustered standard errors for fixed effects regression with a single regressor. These formulas are extended to multiple regressors in Exercise 19.15.

The Asymptotic Distribution of the Fixed Effects Estimator with Large *n*

The fixed effects estimator. The fixed effects estimator of β_1 is the OLS estimator obtained using the entity-demeaned regression of Equation (10.14), in which $\tilde{Y}_{it}$ is regressed on $\tilde{X}_{it}$, where $\tilde{Y}_{it} = Y_{it} - \overline{Y}_i$, $\tilde{X}_{it} = X_{it} - \overline{X}_i$, $\overline{Y}_i = T^{-1}\sum_{t=1}^{T}Y_{it}$, and $\overline{X}_i = T^{-1}\sum_{t=1}^{T}X_{it}$. The formula for the OLS estimator is obtained by replacing $X_i - \overline{X}$ by $\tilde{X}_{it}$ and $Y_i - \overline{Y}$ by $\tilde{Y}_{it}$ in Equation (4.5) and by replacing the single summations in Equation (4.5) by two summations, one over entities $(i = 1, \ldots, n)$ and one over time periods $(t = 1, \ldots, T)$,[5] so

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{Y}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2}. \tag{10.22}$$

The derivation of the sampling distribution of $\hat{\beta}_1$ parallels the derivation in Appendix 4.3 of the sampling distribution of the OLS estimator with cross-sectional data. First, substitute $\tilde{Y}_{it} = \beta_1\tilde{X}_{it} + \tilde{u}_{it}$ [Equation (10.14)] into the numerator of Equation (10.22) to obtain the panel data counterpart of Equation (4.28):

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{u}_{it}}{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2}. \tag{10.23}$$

Next rearrange this expression, and multiply both sides by $\sqrt{nT}$ to obtain

$$\sqrt{nT}(\hat{\beta}_1 - \beta_1) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\eta_i}}{\hat{Q}_{\tilde{X}}}, \text{ where } \eta_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{u}_{it}} \text{ and } \hat{Q}_{\tilde{x}} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2. \tag{10.24}$$

The scaling factor in Equation (10.24), nT, is the total number of observations.

Distribution and standard errors when n is large. In most panel data applications, n is much larger than T, which motivates approximating sampling distributions by letting $n \to \infty$

---

[5] The double summation is the extension to double subscripts of a single summation:

$$\sum_{i=1}^{n}\sum_{t=1}^{T}X_{it} = \sum_{i=1}^{n}\left(\sum_{t=1}^{T}X_{it}\right)$$

$$= \sum_{i=1}^{n}(X_{i1} + X_{i2} + \cdots + X_{iT})$$

$$= (X_{11} + X_{12} + \cdots + X_{1T}) + (X_{21} + X_{22} + \cdots + X_{2T}) + \cdots + (X_{n1} + X_{n2} + \cdots + X_{nT}).$$

while keeping T fixed. Under the fixed effects regression assumptions of Key Concept 10.3, $\hat{Q}_{\tilde{X}} \xrightarrow{p} Q_{\tilde{X}} = ET^{-1}\sum_{t=1}^{T}\tilde{X}_{it}^{2}$ as $n \to \infty$. Also, η_i is i.i.d. over $i = 1, \ldots, n$ (by assumption 2) with mean 0 (by assumption 1) and variance σ_η^2 (which is finite by assumption 3), so by the central limit theorem, $\sqrt{1/n}\sum_{i=1}^{n}\eta_i \xrightarrow{d} N(0, \sigma_\eta^2)$. It follows from Equation (10.24) that

$$\sqrt{nT}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N\left(0, \frac{\sigma_\eta^2}{Q_{\tilde{X}}^2}\right). \tag{10.25}$$

From Equation (10.25), the variance of the large-sample distribution of $\hat{\beta}_1$ is

$$\text{var}(\hat{\beta}_1) = \frac{1}{nT}\frac{\sigma_\eta^2}{Q_{\tilde{X}}^2}. \tag{10.26}$$

The clustered standard error formula replaces the population moments in Equation (10.26) by their sample counterparts:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{nT}\frac{s_{\hat{\eta}}^2}{\hat{Q}_{\tilde{X}}^2}},$$

$$\text{where } s_{\hat{\eta}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\eta}_i - \overline{\hat{\eta}})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\hat{\eta}_i^2, \tag{10.27}$$

where $\hat{\eta}_i = \sqrt{1/T}\sum_{t=1}^{T}\tilde{X}_{it}\hat{u}_{it}$ is the sample counterpart of η_i [$\hat{\eta}_i$ is η_i in Equation (10.24), with $\tilde{u}_{it}$ replaced by the fixed effects regression residual $\hat{u}_{it}$] and $\overline{\hat{\eta}} = (1/n)\sum_{i=1}^{n}\hat{\eta}_i$. The final equality in Equation (10.27) arises because $\overline{\hat{\eta}} = 0$, which in turn follows from the residuals and regressors being uncorrelated [Equation (4.32)]. Note that $s_{\hat{\eta}}^2$ is just the sample variance of $\hat{\eta}_i$ [see Equation (3.7)].

The estimator $s_{\hat{\eta}}^2$ is a consistent estimator of σ_η^2 as $n \to \infty$, even if there is heteroskedasticity or autocorrelation (Exercise 18.15); thus the clustered standard error in Equation (10.27) is heteroskedasticity- and autocorrelation-robust. Because the clustered standard error is consistent, the t-statistic testing $\beta_1 = \beta_{1,0}$ has a standard normal distribution under the null hypothesis as $n \to \infty$.

All the foregoing results apply if there are multiple regressors. In addition, if n is large, then the F-statistic testing q restrictions (computed using the clustered variance formula) has its usual asymptotic $F_{q,\infty}$ distribution.

Why isn't the usual heteroskedasticity-robust estimator of Chapter 5 valid for panel data? There are two reasons. The most important reason is that the heteroskedasticity-robust estimator of Chapter 5 does not allow for serial correlation within a cluster. Recall that, for two random variables U and V, $\text{var}(U + V) = \text{var}(U) + \text{var}(V) + 2\text{cov}(U, V)$. The variance η_i in Equation (10.24) therefore can be written as the sum of variances plus covariances. Let $\tilde{v}_{it} = \tilde{X}_{it}\tilde{u}_{it}$; then

$$\text{var}(\eta_i) = \text{var}\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}\tilde{v}_{it}}\right) = \frac{1}{T}\text{var}(\tilde{v}_{i1} + \tilde{v}_{i2} + \cdots + \tilde{v}_{iT})$$

$$= \frac{1}{T}\text{var}(\tilde{v}_{i1}) + \text{var}(\tilde{v}_{i2}) + \cdots + \text{var}(\tilde{v}_{iT})$$

$$+ 2\text{cov}(\tilde{v}_{i1}, \tilde{v}_{i2}) + \cdots + 2\text{cov}(\tilde{v}_{iT-1}, \tilde{v}_{iT})]. \tag{10.28}$$

The heteroskedasticity-robust variance formula of Chapter 5 misses all the covariances in the final part of Equation (10.28), so if there is serial correlation, the usual heteroskedasticity-robust variance estimator is inconsistent.

The second reason is that if T is small, the estimation of the fixed effects introduces bias into the heteroskedasticity-robust variance estimator. This problem does not arise in cross-sectional regression.

The one case in which the usual heteroskedasticity-robust standard errors *can* be used with panel data is with fixed effects regression with $T = 2$ observations. In this case, fixed effects regression is equivalent to the differences regression in Section 10.2, and heteroskedasticity-robust and clustered standard errors are equivalent.

For empirical examples showing the importance of using clustered standard errors in economic panel data, see Bertrand, Duflo, and Mullainathan (2004).

Extensions: Other applications of clustered standard errors. In some cases, u_{it} might be correlated across entities. For example, in a study of earnings, suppose the sampling scheme selects families by simple random sampling, then tracks all siblings within a family. Because the omitted factors that enter the error term could have common elements for siblings, it is not reasonable to assume that the errors are independent for siblings (even though they are independent across families).

In the siblings example, families are natural clusters, or groupings, of observations, where u_{it} is correlated within the cluster but not across clusters. The derivation leading to Equation (10.27) can be modified to allow for clusters across entities (for example, families) or across both entities and time, as long as there are many clusters.

Clustered standard errors also apply in some applications with cross-sectional data when collection schemes other than simple random sampling are used. For example, suppose cross-sectional student-level data on test scores and student characteristics are obtained by first randomly sampling classrooms, then collecting data on all students within a classroom. Because the classrooms are randomly sampled, errors would be uncorrelated for students from different classrooms. However, the errors might be correlated for students within the same classroom, so clustered standard errors would be appropriate, with the clustering done at the classroom level.

For additional discussion of clustered standard errors, see Cameron and Miller (2015).

Distribution and Standard Errors When n Is Small

If n is small and T is large, then it remains possible to use clustered standard errors; however, t-statistics need to be compared with critical values from the t_{n-1} tables, and the F-statistic testing q restrictions needs to be compared to the $F_{q,n-q}$ critical value multiplied by $(n - 1)/(n - q)$. These distributions are valid under the assumptions in Key Concept 10.3, plus some additional assumptions on the joint distribution of X_{it} and u_{it} over time within an entity. Although the validity of the t-distribution in cross-sectional regression requires normality and homoskedasticity of the regression errors (Section 5.6), neither requirement is needed to justify using the t-distribution with clustered standard errors in panel data when T is large.

To see why the clustered t-statistic has a t_{n-1} distribution when n is small and T is large, even if u_{it} is neither normally distributed nor homoskedastic, first note that if T is large, then under additional assumptions, η_i in Equation (10.24) will obey a central limit theorem, so $\eta_i \xrightarrow{d} N(0, \sigma_\eta^2)$. (The additional assumptions required for this result are substantial and technical, and we defer further discussion of them to our treatment of time series data in Chapter 15.) Thus, if T is large, then $\sqrt{nT}(\hat{\beta}_1 - \beta_1)$ in Equation (10.24) is a scaled average of the n normal random variables η_i. Moreover, the clustered formula s_η^2 in Equation (10.27) is the usual formula for the sample variance, and if it could be computed using η_i, then $(n-1)s_\eta^2/\sigma_\eta^2$ would have a χ_{n-1}^2 distribution, so the t-statistic would have a t_{n-1} distribution [see Section 3.6]. Using the residuals to compute $\hat{\eta}_i$ and s_η^2 does not change this conclusion. In the case of multiple regressors, analogous reasoning leads to the conclusion that the F-statistic testing q restrictions, computed using the cluster variance estimator, is distributed as $\left(\frac{n-1}{n-q}\right)F_{q,n-q}$. [For example, the 5% critical value for this F-statistic when $n = 10$ and $q = 4$ is $\left(\frac{10-1}{10-4}\right) \times 4.53 = 6.80$, where 4.53 is the 5% critical value from the $F_{4,6}$ distribution given in Appendix Table 5B.] Note that, as n increases, the t_{n-1} and $\left(\frac{n-1}{n-q}\right)F_{q,n-q}$ distributions approach the usual standard normal and $F_{q,\infty}$ distributions.[6]

If both n and T are small, then, in general, $\hat{\beta}_1$ will not be normally distributed, and clustered standard errors will not provide reliable inference.

---

[6] Not all software implements clustered standard errors using the t_{n-1} and $\left(\frac{n-1}{n-q}\right)F_{q,n-q}$ distributions that apply if n is small, so you should check how your software implements and treats clustered standard errors.

Regression with a Binary Dependent Variable

Two people, identical but for their race, walk into a bank and apply for a mortgage, a large loan so that each can buy an identical house. Does the bank treat them the same way? Are they both equally likely to have their mortgage application accepted? By law, they must receive identical treatment. But whether they actually do is a matter of great concern among bank regulators.

Loans are made and denied for many legitimate reasons. For example, if the proposed loan payments take up most or all of the applicant's monthly income, a loan officer might justifiably deny the loan. Also, even loan officers are human and they can make honest mistakes, so the denial of a single minority applicant does not prove anything about discrimination. Many studies of discrimination thus look for statistical evidence of discrimination, that is, evidence contained in large data sets showing that whites and minorities are treated differently.

But how, precisely, should one check for statistical evidence of discrimination in the mortgage market? A start is to compare the fraction of minority and white applicants who were denied a mortgage. In the data examined in this chapter, gathered from mortgage applications in 1990 in the Boston, Massachusetts, area, 28% of black applicants were denied mortgages but only 9% of white applicants were denied. But this comparison does not really answer the question that opened this chapter because the black applicants and the white applicants were not necessarily "identical but for their race." Instead, we need a method for comparing rates of denial, *holding other applicant characteristics constant*.

This sounds like a job for multiple regression analysis—and it is, but with a twist. The twist is that the dependent variable—whether the applicant is denied—is binary. In Part II, we regularly used binary variables as regressors, and they caused no particular problems. But when the dependent variable is binary, things are more difficult: What does it mean to fit a line to a dependent variable that can take on only two values, 0 and 1?

The answer to this question is to interpret the regression function as a conditional probability. This interpretation is discussed in Section 11.1, and it allows us to apply the multiple regression models from Part II to binary dependent variables. Section 11.1 goes over this "linear probability model." But the predicted probability interpretation also suggests that alternative, nonlinear regression models can do a better job modeling these probabilities. These methods, called "probit" and "logit" regression, are discussed in Section 11.2. Section 11.3, which is optional, discusses the method used to estimate the coefficients of the probit and logit regressions, the method of

maximum likelihood estimation. In Section 11.4, we apply these methods to the Boston mortgage application data set to see whether there is evidence of racial bias in mortgage lending.

The binary dependent variable considered in this chapter is an example of a dependent variable with a limited range; in other words, it is a **limited dependent variable**. Models for other types of limited dependent variables—for example, dependent variables that take on multiple discrete values—are surveyed in Appendix 11.3.

11.1   Binary Dependent Variables and the Linear Probability Model

Whether a mortgage application is accepted or denied is one example of a binary variable. Many other important questions also concern binary outcomes. What is the effect of a tuition subsidy on an individual's decision to go to college? What deter-mines whether a teenager takes up smoking? What determines whether a country receives foreign aid? What determines whether a job applicant is successful? In all these examples, the outcome of interest is binary: The student does or does not go to college, the teenager does or does not take up smoking, a country does or does not receive foreign aid, the applicant does or does not get a job.

This section discusses what distinguishes regression with a binary dependent variable from regression with a continuous dependent variable and then turns to the simplest model to use with binary dependent variables, the linear probability model.

Binary Dependent Variables

The application examined in this chapter is whether race is a factor in denying a mortgage application; the binary dependent variable is whether a mortgage applica-tion is denied. The data are a subset of a larger data set compiled by researchers at the Federal Reserve Bank of Boston under the Home Mortgage Disclosure Act (HMDA) and relate to mortgage applications filed in the Boston, Massachusetts, area in 1990. The Boston HMDA data are described in Appendix 11.1.

Mortgage applications are complicated. During the period covered by these data, the decision to approve a loan application typically was made by a bank loan officer. The loan officer must assess whether the applicant will make his or her loan pay-ments. One important piece of information is the size of the required loan payments relative to the applicant's income. As anyone who has borrowed money knows, it is much easier to make payments that are 10% of your income than 50%! We therefore begin by looking at the relationship between two variables: the binary dependent variable *deny*, which equals 1 if the mortgage application was denied and equals 0 if it was accepted, and the continuous variable *P/I ratio*, which is the ratio of the appli-cant's anticipated total monthly loan payments to his or her monthly income.

Scatterplot of Mortgage Application Denial and the Payment-to-Income Ratio

Mortgage applicants with a high ratio of debt payments to income (*P/I ratio*) are more likely to have their application denied (*deny = 1* if denied; *deny = 0* if approved). The linear probability model uses a straight line to model the probability of denial, conditional on the *P/I ratio*.
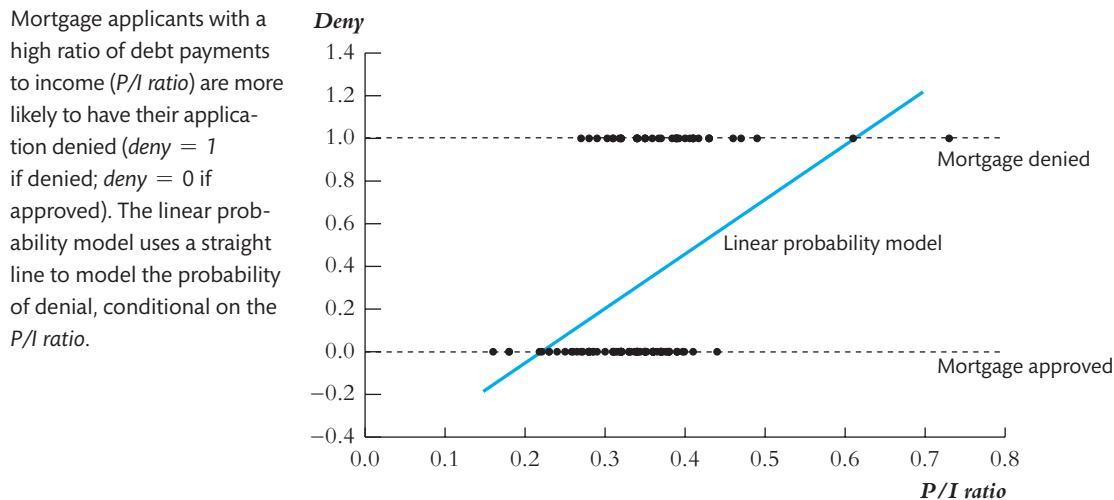
Figure 11.1 presents a scatterplot of *deny* versus *P/I ratio* for 127 of the 2380 observations in the data set. (The scatterplot is easier to read using this subset of the data.) This scatterplot looks different from the scatterplots of Part II because the variable *deny* is binary. Still, it seems to show a relationship between *deny* and *P/I ratio*: Few applicants with a payment-to-income ratio less than 0.3 have their application denied, but most applicants with a payment-to-income ratio exceeding 0.4 are denied.

This positive relationship between *P/I ratio* and *deny* (the higher the *P/I ratio*, the greater the fraction of denials) is summarized in Figure 11.1 by the OLS regression line estimated using these 127 observations. As usual, this line plots the predicted value of *deny* as a function of the regressor, the payment-to-income ratio. For example, when *P/I ratio* = 0.3, the predicted value of deny is 0.20. But what, precisely, does it mean for the predicted value of the binary variable *deny* to be 0.20?

The key to answering this question—and more generally to understanding regression with a binary dependent variable—is to interpret the regression as modeling the *probability* that the dependent variable equals 1. Thus the predicted value of 0.20 is interpreted as meaning that, when *P/I ratio* is 0.3, the probability of denial is estimated to be 20%. Said differently, if there were many applications with *P/I ratio* = 0.3, then 20% of them would be denied.

This interpretation follows from two facts. First, from Part II, the population regression function is the expected value of Y given the regressors, $E(Y|X_1, \ldots, X_k)$. Second, from Section 2.2, if Y is a 0–1 binary variable, its expected value (or mean) is the probability that Y = 1; that is, $E(Y) = 0 \times Pr(Y = 0) + 1 \times Pr(Y = 1) = Pr(Y = 1)$. In the regression context, the expected value is conditional on the value of the regressors, so the probability is conditional on X. Thus for a binary variable,

$E(Y|X_1, \ldots, X_k) = \Pr(Y = 1|X_1, \ldots, X_k)$. In short, for a binary dependent variable, the predicted value from the population regression is the probability that $Y = 1$ given X.

The linear multiple regression model applied to a binary dependent variable is called the linear probability model: *linear* because it is a straight line and *probability model* because it models the probability that the dependent variable equals 1 (in our example, the probability of loan denial).

The Linear Probability Model

The **linear probability model** is the name for the multiple regression model of Part II when the dependent variable is binary rather than continuous. Because the dependent variable Y is binary, the population regression function corresponds to the probability that the dependent variable equals 1 given X. The population coefficient β_1 on a regressor X is the *change in the probability* that $Y = 1$ associated with a *unit change* in X. Similarly, the OLS predicted value, $\hat{Y}_i$, computed using the estimated regression function, is the predicted probability that the dependent variable equals 1, and the OLS estimator $\hat{\beta}_1$ estimates the change in the probability that $Y = 1$ associated with a unit change in X.

Almost all of the tools of Part II carry over to the linear probability model. The coefficients can be estimated by OLS. Ninety-five percent confidence intervals can be formed as ± 1.96 standard errors, hypotheses concerning several coefficients can be tested using the F-statistic discussed in Chapter 7, and interactions between variables can be modeled using the methods of Section 8.3. Because the errors of the linear probability model are always heteroskedastic (Exercise 11.8), it is essential that heteroskedasticity-robust standard errors be used for inference.

One tool that does not carry over is the R^2. When the dependent variable is continuous, it is possible to imagine a situation in which the R^2 equals 1: All the data lie exactly on the regression line. This is impossible when the dependent variable is binary unless the regressors are also binary. Accordingly, the R^2 is not a particularly useful statistic here. We return to measures of fit in the next section.

The linear probability model is summarized in Key Concept 11.1.

Application to the Boston HMDA data. The OLS regression of the binary dependent variable, *deny*, against the payment-to-income ratio, *P/I ratio*, estimated using all 2380 observations in our data set is

$$\widehat{deny} = -0.080 + 0.604 \, P/I \, ratio. \tag{11.1}$$
$$(0.032) \ (0.098)$$

The estimated coefficient on *P/I ratio* is positive, and the population coefficient is statistically significantly different from 0 at the 1% level (the *t*-statistic is 6.13). Thus applicants with higher debt payments as a fraction of income are more likely to have their application denied. This coefficient can be used to compute the predicted

The Linear Probability Model

The linear probability model is the linear multiple regression model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \tag{11.2}$$

applied to a binary dependent variable Y_i. Because Y is binary, $E(Y \mid X_1, X_2, \ldots, X_k) = \Pr(Y = 1 \mid X_1, X_2, \ldots, X_k)$, so for the linear probability model,

$$\Pr(Y = 1 \mid X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$

The regression coefficient β_1 is the difference in the probability that $Y = 1$ associated with a unit difference in X_1, holding constant the other regressors, and so forth for $\beta_2, \ldots, \beta_k$. The regression coefficients can be estimated by OLS, and the usual (heteroskedasticity-robust) OLS standard errors can be used for confidence intervals and hypothesis tests.

change in the probability of denial given a change in the regressor. For example, according to Equation (11.1), if *P/I ratio* increases by 0.1, the probability of denial increases by $0.604 \times 0.1 \cong 0.060$—that is, by 6.0 percentage points.

The estimated linear probability model in Equation (11.1) can be used to compute predicted denial probabilities as a function of *P/I ratio*. For example, if projected debt payments are 30% of an applicant's income, *P/I ratio* is 0.3, and the predicted value from Equation (11.1) is $-0.080 + 0.604 \times 0.3 = 0.101$. That is, according to this linear probability model, an applicant whose projected debt payments are 30% of income has a probability of 10.1% that his or her application will be denied. [This is different from the probability of 20% based on the regression line in Figure 11.1 because that line was estimated using only 127 of the 2380 observations used to estimate Equation (11.1).]

What is the effect of race on the probability of denial, holding constant the *P/I ratio*? To keep things simple, we focus on differences between black applicants and white applicants. To estimate the effect of race, holding constant *P/I ratio*, we augment Equation (11.1) with a binary regressor that equals 1 if the applicant is black and equals 0 if the applicant is white. The estimated linear probability model is

$$\widehat{deny} = -0.091 + 0.559 \, P/I \, ratio + 0.177 \, black. \tag{11.3}$$
$$\qquad\quad (0.029) \;\; (0.089) \qquad\qquad (0.025)$$

The coefficient on *black*, 0.177, indicates that an African American applicant has a 17.7% higher probability of having a mortgage application denied than a white applicant, holding constant their payment-to-income ratio. This coefficient is significant at the 1% level (the *t*-statistic is 7.11).

Taken literally, this estimate suggests that there might be racial bias in mortgage decisions, but such a conclusion would be premature. Although the payment-to-income ratio plays a role in the loan officer's decision, so do many other factors, such as the applicant's earning potential and his or her credit history. If any of these variables is correlated with the regressors *black* given the *P/I ratio*, its omission from Equation (11.3) will cause omitted variable bias. Thus we must defer any conclusions about discrimination in mortgage lending until we complete the more thorough analysis in Section 11.3.

Shortcomings of the linear probability model.   The linearity that makes the linear probability model easy to use is also its major flaw. Because probabilities cannot exceed 1, the effect on the probability that $Y = 1$ of a given change in X must be nonlinear: Although a change in *P/I ratio* from 0.3 to 0.4 might have a large effect on the probability of denial, once *P/I ratio* is so large that the loan is very likely to be denied, increasing *P/I ratio* further will have little effect. In contrast, in the linear probability model, the effect of a given change in *P/I ratio* is constant, which leads to predicted probabilities in Figure 11.1 that drop below 0 for very low values of *P/I ratio* and exceed 1 for high values! But this is nonsense: A probability cannot be less than 0 or greater than 1. This nonsensical feature is an inevitable consequence of the linear regression. To address this problem, we introduce new nonlinear models specifically designed for binary dependent variables, the probit and logit regression models.

11.2   Probit and Logit Regression

Probit and **logit**[1] regression are nonlinear regression models specifically designed for binary dependent variables. Because a regression with a binary dependent variable Y models the probability that $Y = 1$, it makes sense to adopt a nonlinear formulation that forces the predicted values to be between 0 and 1. Because cumulative probability distribution functions (c.d.f.'s) produce probabilities between 0 and 1 (Section 2.1), they are used in logit and probit regressions. Probit regression uses the standard normal c.d.f. Logit regression, also called **logistic regression**, uses the logistic c.d.f.

Probit Regression

Probit regression with a single regressor.   The probit regression model with a single regressor X is

$$\Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X), \tag{11.4}$$

where Φ is the cumulative standard normal distribution function (tabulated in Appendix Table 1).

---

[1]Pronounced prō-bit and lō-jit.

For example, suppose that Y is the binary mortgage denial variable (*deny*), X is the payment-to-income ratio (*P/I ratio*), $\beta_0 = -2$, and $\beta_1 = 3$. What then is the probability of denial if *P/I ratio* = 0.4? According to Equation (11.4), this probability is $\Phi(\beta_0 + \beta_1 P/I\ ratio) = \Phi(-2 + 3P/I\ ratio) = \Phi(-2 + 3 \times 0.4) = \Phi(-0.8)$. According to the cumulative normal distribution table (Appendix Table 1), $\Phi(-0.8) = \Pr(Z \leq -0.8) = 21.2\%$. That is, when *P/I ratio* is 0.4, the predicted probability that the application will be denied is 21.2%, computed using the probit model with the coefficients $\beta_0 = -2$ and $\beta_1 = 3$.

In the probit model, the term $\beta_0 + \beta_1 X$ plays the role of "z" in the cumulative standard normal distribution table in Appendix Table 1. Thus the calculation in the previous paragraph can, equivalently, be done by first computing the "z-value," $z = \beta_0 + \beta_1 X = -2 + 3 \times 0.4 = -0.8$, and then looking up the probability in the tail of the normal distribution to the left of $z = -0.8$, which is 21.2%.

The probit coefficient β_1 in Equation (11.4) is the difference in the z-value associated with a unit difference in X. If β_1 is positive, a greater value for X increases the z-value and thus increases the probability that $Y = 1$; if β_1 is negative, a greater value for X decreases the probability that $Y = 1$. Although the effect of X on the z-value is linear, its effect on the probability is nonlinear. Thus in practice the easiest way to interpret the coefficients of a probit model is to compute the predicted probability, or the change in the predicted probability, for one or more values of the regressors. When there is just one regressor, the predicted probability can be plotted as a function of X.

Figure 11.2 plots the estimated regression function produced by the probit regression of *deny* on *P/I ratio* for the 127 observations in the scatterplot. The

FIGURE 11.2 **Probit Model of the Probability of Denial Given *P/I Ratio***

The probit model uses the cumulative normal distribution function to model the probability of denial given the payment-to-income ratio or, more generally, to model $Pr(Y = 1|X)$. Unlike the linear probability model, the probit conditional probabilities are always between 0 and 1.
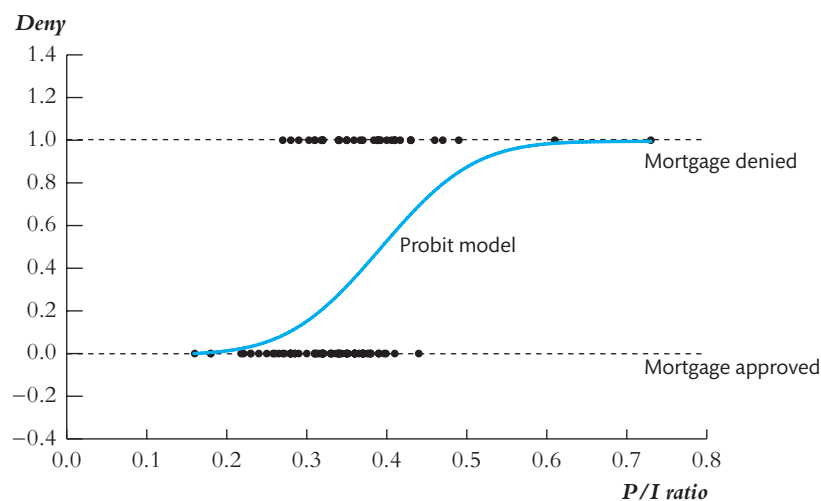
estimated probit regression function has a stretched "S" shape: It is nearly 0 and flat for small values of *P/I ratio*, it turns and increases for intermediate values, and it flattens out again and is nearly 1 for large values. For small values of the payment-to-income ratio, the probability of denial is small. For example, for *P/I ratio* $= 0.2$, the estimated probability of denial based on the estimated probit function in Figure 11.2 is $\Pr(deny = 1 | P/I\,ratio = 0.2) = 2.1\%$. When *P/I ratio* $= 0.3$, the estimated probability of denial is 16.1%. When *P/I ratio* $= 0.4$, the probability of denial increases sharply to 51.9%, and when *P/I ratio* $= 0.6$, the denial probability is 98.3%. According to this estimated probit model, for applicants with high payment-to-income ratios, the probability of denial is nearly 1.

Probit regression with multiple regressors. In all the regression problems we have studied so far, leaving out a determinant of Y that is correlated with the included regressors results in omitted variable bias. Probit regression is no exception. In linear regression, the solution is to include the additional variable as a regressor. This is also the solution to omitted variable bias in probit regression.

The probit model with multiple regressors extends the single-regressor probit model by adding regressors to compute the z-value. Accordingly, the probit population regression model with two regressors, X_1 and X_2, is

$$Pr(Y = 1 | X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2). \tag{11.5}$$

For example, suppose that $\beta_0 = -1.6$, $\beta_1 = 2$, and $\beta_2 = 0.5$. If $X_1 = 0.4$ and $X_2 = 1$, the z-value is $z = -1.6 + 2 \times 0.4 + 0.5 \times 1 = -0.3$. So the probability that $Y = 1$ given $X_1 = 0.4$ and $X_2 = 1$ is $\Pr(Y = 1 | X_1 = 0.4, X_2 = 1) = \Phi(-0.3) = 38\%$.

Effect of a change in X. In general, the regression model can be used to determine the expected change in Y arising from a change in X. When Y is binary, its conditional expectation is the conditional probability that it equals 1, so the expected change in Y arising from a change in X is the change in the probability that $Y = 1$.

Recall from Section 8.1 that, when the population regression function is a nonlinear function of X, this expected change is estimated in three steps: First, compute the predicted value at the original value of X using the estimated regression function; next, compute the predicted value at the changed value of X, $X + \Delta X$; finally, compute the difference between the two predicted values. This procedure is summarized in Key Concept 8.1. As emphasized in Section 8.1, this method *always* works for computing predicted effects of a change in X, no matter how complicated the nonlinear model. When applied to the probit model, the method of Key Concept 8.1 yields the estimated effect on the probability that $Y = 1$ of a change in X.

The probit regression model, predicted probabilities, and estimated effects are summarized in Key Concept 11.2.

The Probit Model, Predicted Probabilities, and Estimated Effects

The population probit model with multiple regressors is

$$\Pr(Y = 1 | X_1, X_2, \ldots, X_k) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k), \quad (11.6)$$

where the dependent variable Y is binary, Φ is the cumulative standard normal distribution function, and X_1, X_2, and so on are regressors. The model is best interpreted by computing predicted probabilities and the effect of a change in a regressor.

The predicted probability that $Y = 1$, given values of $X_1, X_2, \ldots, X_k$, is calculated by computing the z-value, $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$, and then looking up this z-value in the normal distribution table (Appendix Table 1).

The coefficient β_1 is the difference in the z-value arising from a unit difference in X_1, holding constant $X_2, \ldots, X_k$.

The effect on the predicted probability of a change in a regressor is computed by (1) computing the predicted probability for the initial value of the regressor, (2) computing the predicted probability for the new or changed value of the regressor, and (3) taking their difference.

Application to the mortgage data. As an illustration, we fit a probit model to the 2380 observations in our data set on mortgage denial (*deny*) and the payment-to-income ratio (*P/I ratio*):

$$\widehat{\Pr(deny = 1 | P/I\ ratio)} = \Phi(-2.19 + 2.97\, P/I\ ratio). \quad (11.7)$$
$$(0.16) \quad (0.47)$$

The estimated coefficients of -2.19 and 2.97 are difficult to interpret because they affect the probability of denial via the z-value. Indeed, the only things that can be readily concluded from the estimated probit regression in Equation (11.7) are that the payment-to-income ratio is positively related to probability of denial (the coefficient on *P/I ratio* is positive) and that this relationship is statistically significant ($t = 2.97/0.47 = 6.32$).

What is the change in the predicted probability that an application will be denied when the payment-to-income ratio increases from 0.3 to 0.4? To answer this question, we follow the procedure in Key Concept 8.1: Compute the probability of denial for *P/I ratio* = 0.3 and for *P/I ratio* = 0.4, and then compute the difference. The probability of denial when $P/I\ ratio = 0.3$ is $\Phi(-2.19 + 2.97 \times 0.3) = \Phi(-1.30) = 0.097$. The probability of denial when $P/I\ ratio = 0.4$ is $\Phi(-2.19 + 2.97 \times 0.4) = \Phi(-1.00) = 0.159$. The estimated change in the probability of denial is $0.159 - 0.097 = 0.062$. That is, an increase in the payment-to-income ratio from 0.3 to 0.4 is associated with an increase in the probability of denial of 6.2 percentage points, from 9.7% to 15.9%.

Because the probit regression function is nonlinear, the effect of a change in X depends on the starting value of X. For example, if $P/I\ ratio = 0.5$, the estimated denial probability based on Equation (11.7) is $\Phi(-2.19 + 2.97 \times 0.5) = \Phi(-0.71) = 0.239$. Thus the change in the predicted probability when $P/I\ ratio$ increases from 0.4 to 0.5 is $0.239 - 0.159$, or 8.0 percentage points, larger than the increase of 6.2 percentage points when $P/I\ ratio$ increases from 0.3 to 0.4.

What is the effect of race on the probability of mortgage denial, holding constant the payment-to-income ratio? To estimate this effect, we estimate a probit regression with both *P/I ratio* and *black* as regressors:

$$\widehat{\Pr(deny = 1 | P/I\ ratio, black)} = \Phi(-2.26 + 2.74\ P/I\ ratio + 0.71\ black). \quad (11.8)$$
$$\qquad\qquad\qquad (0.16)\ \ (0.44)\qquad\qquad (0.083)$$

Again, the values of the coefficients are difficult to interpret, but the sign and statistical significance are not. The coefficient on *black* is positive, indicating that an African American applicant has a higher probability of denial than a white applicant, holding constant their payment-to-income ratio. This coefficient is statistically significant at the 1% level (the *t*-statistic on the coefficient multiplying *black* is 8.55). For a white applicant with $P/I\ ratio = 0.3$, the predicted denial probability is 7.5%, while for a black applicant with $P/I\ ratio = 0.3$, it is 23.3%; the difference in denial probabilities between these two hypothetical applicants is 15.8 percentage points.

Estimation of the probit coefficients. The probit coefficients reported here were estimated using the method of maximum likelihood, which produces efficient (minimum variance) estimators in a wide variety of applications, including regression with a binary dependent variable. The maximum likelihood estimator is consistent and normally distributed in large samples, so *t*-statistics and confidence intervals for the coefficients can be constructed in the usual way.

Regression software for estimating probit models typically uses maximum likelihood estimation, so this is a simple method to apply in practice. Standard errors produced by such software can be used in the same way as the standard errors of regression coefficients; for example, a 95% confidence interval for the true probit coefficient can be constructed as the estimated coefficient ± 1.96 standard errors. Similarly, *F*-statistics computed using maximum likelihood estimators can be used to test joint hypotheses. Maximum likelihood estimation is discussed further in Section 11.3, with additional details given in Appendix 11.2.

Logit Regression

The logit regression model. The logit regression model is similar to the probit regression model except that the cumulative standard normal distribution function Φ in Equation (11.6) is replaced by the cumulative standard logistic distribution function, which we denote by F. Logit regression is summarized in Key Concept 11.3. The logistic

Logit Regression

The population logit model of the binary dependent variable Y with multiple regressors is

$$\Pr(Y = 1 \mid X_1, X_2, \ldots, X_k) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}. \tag{11.9}$$

Logit regression is similar to probit regression except that the cumulative distribution function is different.

cumulative distribution function has a specific functional form, defined in terms of the exponential function, which is given as the final expression in Equation (11.9).

As with probit, the logit coefficients are best interpreted by computing predicted probabilities and differences in predicted probabilities.

The coefficients of the logit model can be estimated by maximum likelihood. The maximum likelihood estimator is consistent and normally distributed in large samples, so t-statistics and confidence intervals for the coefficients can be constructed in the usual way.

The logit and probit regression functions are similar. This is illustrated in Figure 11.3, which graphs the probit and logit regression functions for the dependent variable *deny* and the single regressor *P/I ratio*, estimated by maximum likelihood using the same 127 observations as in Figures 11.1 and 11.2. The differences between the two functions are small.

FIGURE 11.3 Probit and Logit Models of the Probability of Denial Given *P/I Ratio*

These logit and probit models produce nearly identical estimates of the probability that a mortgage application will be denied, given the payment-to-income ratio.
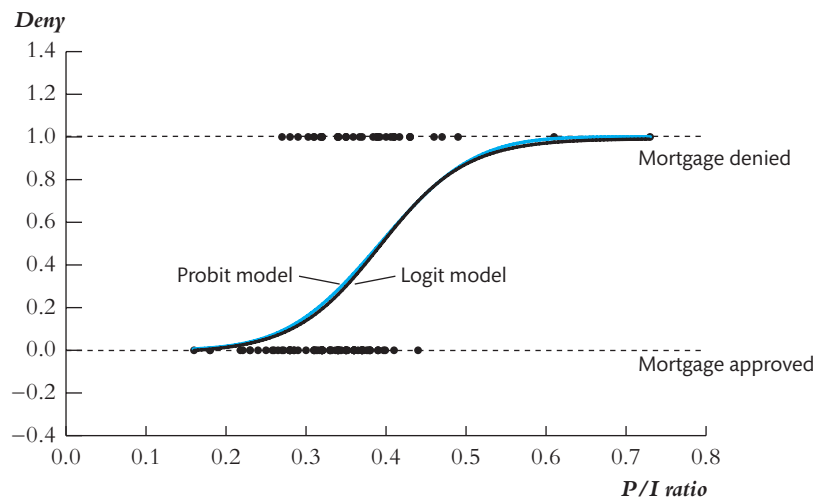
Historically, the main motivation for logit regression was that the logistic cumulative distribution function could be computed faster than the normal cumulative distribution function. With the advent of more powerful computers, this distinction is no longer important.

Application to the Boston HMDA data. A logit regression of *deny* against *P/I ratio* and *black*, using the 2380 observations in the data set, yields the estimated regression function

$$\widehat{\Pr\left(deny = 1 \mid P/I\,ratio, black\right)} = F(-4.13 + 5.37\,P/I\,ratio + 1.27\,black). \quad (11.10)$$
$$\qquad\qquad\qquad (0.35)\ \ (0.96) \qquad\qquad (0.15)$$

The coefficient on *black* is positive and statistically significant at the 1% level (the *t*-statistic is 8.47). The predicted denial probability of a white applicant with $P/I\,ratio = 0.3$ is $1/[1 + e^{-(-4.13+5.37\times0.3+1.27\times0)}] = 1/[1 + e^{2.52}] = 0.074$, or 7.4%. The predicted denial probability of an African American applicant with $P/I\,ratio = 0.3$ is $1/[1 + e^{1.25}] = 0.222$, or 22.2%, so the difference between the two probabilities is 14.8 percentage points.

Comparing the Linear Probability, Probit, and Logit Models

All three models—linear probability, probit, and logit—are just approximations to the unknown population regression function $E(Y|X) = \Pr(Y = 1|X)$. The linear probability model is easiest to use and to interpret, but it cannot capture the nonlinear nature of the true population regression function. Probit and logit regressions model this nonlinearity in the probabilities, but their regression coefficients are more difficult to interpret. So which should you use in practice?

There is no one right answer, and different researchers use different models. Probit and logit regressions frequently produce similar results. For example, according to the estimated probit model in Equation (11.8), the difference in denial probabilities between a black applicant and a white applicant with $P/I\,ratio = 0.3$ was estimated to be 15.8 percentage points, whereas the logit estimate of this gap, based on Equation (11.10), was 14.9 percentage points. For practical purposes, the two estimates are very similar. One way to choose between logit and probit is to pick the method that is easier to use in your statistical software.

The linear probability model provides the least sensible approximation to the nonlinear population regression function. Even so, in some data sets there may be few extreme values of the regressors, in which case the linear probability model still can provide an adequate approximation. In the denial probability regression in Equation (11.3), the estimated black/white gap from the linear probability model is 17.7 percentage points, larger than the probit and logit estimates but still qualitatively similar. The only way to know this, however, is to estimate both a linear and a nonlinear model and to compare their predicted probabilities.

11.3 Estimation and Inference in the Logit and Probit Models[2]

The nonlinear models studied in Sections 8.2 and 8.3 are nonlinear functions of the independent variables but are linear functions of the unknown coefficients (parameters). Consequently, the unknown coefficients of those nonlinear regression functions can be estimated by OLS. In contrast, the probit and logit regression functions are nonlinear functions of the coefficients. That is, the probit coefficients $\beta_0, \beta_1, \ldots, \beta_k$ in Equation (11.6) appear *inside* the cumulative standard normal distribution function Φ, and the logit coefficients in Equation (11.9) appear *inside* the cumulative standard logistic distribution function F. Because the population regression function is a nonlinear function of the coefficients $\beta_0, \beta_1, \ldots, \beta_k$, those coefficients cannot be estimated by OLS.

This section provides an introduction to the standard method for estimation of probit and logit coefficients, maximum likelihood; additional mathematical details are given in Appendix 11.2. Because it is built into modern statistical software, maximum likelihood estimation of the probit and logit coefficients is easy in practice. The theory of maximum likelihood estimation, however, is more complicated than the theory of least squares. We therefore first discuss another estimation method, nonlinear least squares, before turning to maximum likelihood.

Nonlinear Least Squares Estimation

Nonlinear least squares is a general method for estimating the unknown parameters of a regression function when, like the probit coefficients, those parameters enter the population regression function nonlinearly. The nonlinear least squares estimator, which was introduced in Appendix 8.1, extends the OLS estimator to regression functions that are nonlinear functions of the parameters. Like OLS, nonlinear least squares finds the values of the parameters that minimize the sum of squared prediction mistakes produced by the model.

To be concrete, consider the nonlinear least squares estimator of the parameters of the probit model. The conditional expectation of Y given the X's is $E(Y|X_1, \ldots, X_k) = \Pr(Y = 1|X_1, \ldots, X_k) = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$. Estimation by nonlinear least squares fits this conditional expectation function, which is a nonlinear function of the parameters, to the dependent variable. That is, the nonlinear least squares estimator of the probit coefficients is the values of $b_0, \ldots, b_k$ that minimize the sum of squared prediction mistakes:

$$\sum_{i=1}^{n} [Y_i - \Phi(b_0 + b_1 X_{1i} + \cdots + b_k X_{ki})]^2. \tag{11.11}$$

The nonlinear least squares estimator shares two key properties with the OLS estimator in linear regression: It is consistent (the probability that it is close to the true

---

[2]This section contains more advanced material that can be skipped without loss of continuity.

value approaches 1 as the sample size gets large), and it is normally distributed in large samples. There are, however, estimators that have a smaller variance than the nonlinear least squares estimator; that is, the nonlinear least squares estimator is inefficient. For this reason, the nonlinear least squares estimator of the probit coefficients is rarely used in practice, and instead the parameters are estimated by maximum likelihood.

Maximum Likelihood Estimation

The **likelihood function** is the joint probability distribution of the data, treated as a function of the unknown coefficients. The **maximum likelihood estimator (MLE)** of the unknown coefficients consists of the values of the coefficients that maximize the likelihood function. Because the MLE chooses the unknown coefficients to maximize the likelihood function, which is in turn the joint probability distribution, in effect the MLE chooses the values of the parameters to maximize the probability of drawing the data that are actually observed. In this sense, the MLEs are the parameter values "most likely" to have produced the data.

To illustrate maximum likelihood estimation, consider two i.i.d. observations, Y_1 and Y_2, on a binary dependent variable with no regressors. Thus Y is a Bernoulli random variable, and the only unknown parameter to estimate is the probability p that $Y = 1$, which is also the mean of Y.

To obtain the maximum likelihood estimator, we need an expression for the likelihood function, which in turn requires an expression for the joint probability distribution of the data. The joint probability distribution of the two observations Y_1 and Y_2 is $\Pr(Y_1 = y_1, Y_2 = y_2)$. Because Y_1 and Y_2 are independently distributed, the joint distribution is the product of the individual distributions [Equation (2.24)], so $\Pr(Y_1 = y_1, Y_2 = y_2) = \Pr(Y_1 = y_1)\Pr(Y_2 = y_2)$. The Bernoulli distribution can be summarized in the formula $\Pr(Y = y) = p^y(1 - p)^{1-y}$: When $y = 1$, $\Pr(Y = 1) = p^1(1 - p)^0 = p$, and when $y = 0, \Pr(Y = 0) = p^0(1 - p)^1 = 1 - p$. Thus the joint probability distribution of Y_1 and Y_2 is $\Pr(Y_1 = y_1, Y_2 = y_2) = [p^{y_1}(1 - p)^{1-y_1}] \times [p^{y_2}(1 - p)^{1-y_2}] = p^{(y_1 + y_2)}(1 - p)^{2-(y_1+y_2)}$.

The likelihood function is the joint probability distribution, treated as a function of the unknown coefficients. For $n = 2$ i.i.d. observations on Bernoulli random variables, the likelihood function is

$$f(p; Y_1, Y_2) = p^{(Y_1 + Y_2)}(1 - p)^{2-(Y_1+Y_2)}. \tag{11.12}$$

The maximum likelihood estimator of p is the value of p that maximizes the likelihood function in Equation (11.12). As with all maximization or minimization problems, this can be done by trial and error; that is, you can try different values of p and compute the likelihood $f(p; Y_1, Y_2)$ until you are satisfied that you have maximized this function. In this example, however, maximizing the likelihood function using calculus produces a simple formula for the MLE: The MLE is $\hat{p} = \frac{1}{2}(Y_1 + Y_2)$.

In other words, the MLE of p is just the sample average! In fact, for general n, the MLE $\hat{p}$ of the Bernoulli probability p is the sample average; that is, $\hat{p} = \overline{Y}$ (this is shown in Appendix 11.2). In this example, the MLE is the usual estimator of p, the fraction of times $Y_i = 1$ in the sample.

This example is similar to the problem of estimating the unknown coefficients of the probit and logit regression models. In those models, the success probability p is not constant but rather depends on X; that is, it is the success probability conditional on X, which is given in Equation (11.6) for the probit model and Equation (11.9) for the logit model. Thus the probit and logit likelihood functions are similar to the likelihood function in Equation (11.12) except that the success probability varies from one observation to the next (because it depends on X_i). Expressions for the probit and logit likelihood functions are given in Appendix 11.2.

Like the nonlinear least squares estimator, the MLE is consistent and normally distributed in large samples. Because regression software commonly computes the MLE of the probit coefficients, this estimator is easy to use in practice. All the estimated probit and logit coefficients reported in this chapter are MLEs.

Statistical inference based on the MLE. Because the MLE is normally distributed in large samples, statistical inference about the probit and logit coefficients based on the MLE proceeds in the same way as inference about the linear regression function coefficients based on the OLS estimator. That is, hypothesis tests are performed using the t-statistic, and 95% confidence intervals are formed as ± 1.96 standard errors. Tests of joint hypotheses on multiple coefficients use the F-statistic in a way similar to that discussed in Chapter 7 for the linear regression model. All of this is completely analogous to statistical inference in the linear regression model.

An important practical point is that some statistical software reports tests of joint hypotheses using the F-statistic, while other software uses the chi-squared statistic. The chi-squared statistic is $q \times F$, where q is the number of restrictions being tested. Because the F-statistic is, under the null hypothesis, distributed as χ_q^2/q in large samples, $q \times F$ is distributed as χ_q^2 in large samples. Because the two approaches differ only in whether they divide by q, they produce identical inferences, but you need to know which approach is implemented in your software so that you use the correct critical values.

Measures of Fit

In Section 11.1, it was mentioned that the R^2 is a poor measure of fit for the linear probability model. This is also true for probit and logit regression. Two measures of fit for models with binary dependent variables are the fraction correctly predicted and the pseudo-R^2. The **fraction correctly predicted** uses the following rule: If $Y_i = 1$ and the predicted probability exceeds 50% or if $Y_i = 0$ and the predicted probability is less than 50%, then Y_i is said to be correctly predicted. Otherwise, Y_i is said to be incorrectly predicted. The fraction correctly predicted is the fraction of the n observations $Y_1, \ldots, Y_n$ that are correctly predicted.

An advantage of this measure of fit is that it is easy to understand. A disadvantage is that it does not reflect the quality of the prediction: If $Y_i = 1$, the observation is treated as correctly predicted whether the predicted probability is 51% or 90%.

The **pseudo-R^2** measures the fit of the model using the likelihood function. Because the MLE maximizes the likelihood function, adding another regressor to a probit or logit model increases the value of the maximized likelihood, just like adding a regressor necessarily reduces the sum of squared residuals in linear regression by OLS. This suggests measuring the quality of fit of a probit model by comparing values of the maximized likelihood function with all the regressors to the value of the likelihood with none. This is, in fact, what the pseudo-R^2 does. A formula for the pseudo-R^2 is given in Appendix 11.2.

11.4   Application to the Boston HMDA Data

The regressions of the previous two sections indicated that denial rates were higher for black than white applicants, holding constant their payment-to-income ratio. Loan officers, however, legitimately weigh many factors when deciding on a mortgage application, and if any of those other factors differ systematically by race, the estimators considered so far have omitted variable bias.

In this section, we take a closer look at whether there is statistical evidence of discrimination in the Boston HMDA data. Specifically, our objective is to estimate the effect of race on the probability of denial, holding constant those applicant characteristics that a loan officer might legally consider when deciding on a mortgage application.

The most important variables available to loan officers through the mortgage applications in the Boston HMDA data set are listed in Table 11.1; these are the variables we will focus on in our empirical models of loan decisions. The first two variables are direct measures of the financial burden the proposed loan would place on the applicant, measured in terms of his or her income. The first of these is the *P/I ratio*; the second is the ratio of housing-related expenses to income. The next variable is the size of the loan, relative to the assessed value of the home; if the loan-to-value ratio is nearly 1, the bank might have trouble recouping the full amount of the loan if the applicant defaults on the loan and the bank forecloses. The final three financial variables summarize the applicant's credit history. If an applicant has been unreliable paying off debts in the past, the loan officer legitimately might worry about the applicant's ability or desire to make mortgage payments in the future. The three variables measure different types of credit histories, which the loan officer might weigh differently. The first concerns consumer credit, such as credit card debt; the second is previous mortgage payment history; and the third measures credit problems so severe that they appeared in a public legal record, such as filing for bankruptcy.

TABLE 11.1	Variables Included in Regression Models of Mortgage Decisions	
Variable	**Definition**	**Sample Average**
Financial Variables		
P/I ratio	Ratio of total monthly debt payments to total monthly income	0.331
housing expense-to-income ratio	Ratio of monthly housing expenses to total monthly income	0.255
loan-to-value ratio	Ratio of size of loan to assessed value of property	0.738
consumer credit score	1 if no "slow" payments or delinquencies 2 if one or two slow payments or delinquencies 3 if more than two slow payments 4 if insufficient credit history for determination 5 if delinquent credit history with payments 60 days overdue 6 if delinquent credit history with payments 90 days overdue	2.1
mortgage credit score	1 if no late mortgage payments 2 if no mortgage payment history 3 if one or two late mortgage payments 4 if more than two late mortgage payments	1.7
public bad credit record	1 if any public record of credit problems (bankruptcy, charge-offs, collection actions) 0 otherwise	0.074
Additional Applicant Characteristics		
denied mortgage insurance	1 if applicant applied for mortgage insurance and was denied, 0 otherwise	0.020
self-employed	1 if self-employed, 0 otherwise	0.116
single	1 if applicant reported being single, 0 otherwise	0.393
high school diploma	1 if applicant graduated from high school, 0 otherwise	0.984
unemployment rate	1989 Massachusetts unemployment rate in the applicant's industry	3.8
condominium	1 if unit is a condominium, 0 otherwise	0.288
black	1 if applicant is black, 0 if white	0.142
deny	1 if mortgage application denied, 0 otherwise	0.120

Table 11.1 also lists some other variables relevant to the loan officer's decision. Sometimes the applicant must apply for private mortgage insurance.[3] The loan officer knows whether that application was denied, and that denial would weigh negatively with the loan officer. The next four variables, which concern the applicant's employment status, marital status, and educational attainment, as well as the unemployment rate in the applicant's industry, relate to the prospective ability of the applicant to repay. In the event of foreclosure, characteristics of the property are relevant as well, and the next variable indicates whether the property is a condominium. The final two variables in Table 11.1 are whether the applicant is black or white and

---

[3]Mortgage insurance is an insurance policy under which the insurance company makes the monthly payment to the bank if the borrower defaults. During the period of this study, if the loan-to-value ratio exceeds 80%, the applicant typically was required to buy mortgage insurance.

whether the application was denied or accepted. In these data, 14.2% of applicants are black, and 12.0% of applications are denied.

Table 11.2 presents regression results based on these variables. The base specifications, reported in columns (1) through (3), include the financial variables in Table 11.1 plus the variables indicating whether private mortgage insurance was denied and whether the applicant is self-employed. In the 1990s, loan officers commonly used thresholds, or cutoff values, for the loan-to-value ratio, so the base specification for that variable uses binary variables for whether the loan-to-value ratio is high (≥ 0.95), medium (between 0.8 and 0.95), or low (< 0.8; this case is omitted to avoid perfect multicollinearity). The regressors in the first three columns are similar to those in the base specification considered by the Federal Reserve Bank of Boston researchers in their original analysis of these data.[4] The regressions in columns (1) through (3) differ only in how the denial probability is modeled, using a linear probability model, a logit model, and a probit model, respectively.

Because the coefficients of the logit and probit models in columns (2)–(6) are not directly interpretable, the table reports standard errors but not confidence intervals. In addition, because the aim of these regressions is to approximate the loan officers' decision rule, it is of interest to know whether individual variables—especially the applicant's race—enter that decision rule. Thus the table reports, through asterisks, whether the test that the coefficient is 0 rejects at the 5% or 1% significance level.

Because the regression in column (1) is a linear probability model, its coefficients are estimated changes in predicted probabilities arising from a unit change in the independent variable. Accordingly, an increase in *P/I ratio* of 0.1 is estimated to increase the probability of denial by 4.5 percentage points (the coefficient on *P/Iratio* in column (1) is 0.449, and 0.449 × 0.1 ≅ 0.045). Similarly, having a high loan-to-value ratio increases the probability of denial: A loan-to-value ratio exceeding 95% is associated with an 18.9 percentage point increase (the coefficient is 0.189) in the denial probability, relative to the omitted case of a loan-to-value ratio less than 80%, holding the other variables in column (1) constant. Applicants with a poor credit rating also have a more difficult time getting a loan, all else being constant, although interestingly the coefficient on consumer credit is statistically significant but the coefficient on mortgage credit is not. Applicants with a public record of credit problems, such as filing for bankruptcy, have much greater difficulty obtaining a loan: All else equal, a public bad credit record is estimated to increase the probability of denial by 0.197, or 19.7 percentage points. Being denied private mortgage insurance is estimated to be virtually decisive: The estimated coefficient of 0.702 means that being denied mortgage insurance increases your chance of being denied a mortgage by 70.2 percentage points, all else

---

[4]The difference between the regressors in columns (1) through (3) and those in Munnell et al. (1996), table 2 (1), is that Munnell et al. include additional indicators for the location of the home and the identity of the lender, data that are not publicly available; an indicator for a multifamily home, which is irrelevant here because our subset focuses on single-family homes; and net wealth, which we omit because this variable has a few very large positive and negative values and thus risks making the results sensitive to a few specific outlier observations.

TABLE 11.2 Mortgage Denial Regressions Using the Boston HMDA Data

Dependent variable: *deny* = 1 if mortgage application is denied, = 0 if accepted; 2380 observations.

Regression Model	LPM	Logit	Probit	Probit	Probit	Probit
Regressor	(1)	(2)	(3)	(4)	(5)	(6)
black	0.084** (0.023)	0.688** (0.182)	0.389** (0.098)	0.371** (0.099)	0.363** (0.100)	0.246 (0.448)
P/I ratio	0.449** (0.114)	4.76** (1.33)	2.44** (0.61)	2.46** (0.60)	2.62** (0.61)	2.57** (0.66)
housing expense-to-income ratio	−0.048 (0.110)	−0.11 (1.29)	−0.18 (0.68)	−0.30 (0.68)	−0.50 (0.70)	−0.54 (0.74)
medium loan-to-value ratio (0.80 ≤ loan-value ratio ≤ 0.95)	0.031* (0.013)	0.46** (0.16)	0.21** (0.08)	0.22** (0.08)	0.22** (0.08)	0.22** (0.08)
high loan-to-value ratio (loan-value ratio > 0.95)	0.189** (0.050)	1.49** (0.32)	0.79** (0.18)	0.79** (0.18)	0.84** (0.18)	0.79** (0.18)
consumer credit score	0.031** (0.005)	0.29** (0.04)	0.15** (0.02)	0.16** (0.02)	0.34** (0.11)	0.16** (0.02)
mortgage credit score	0.021 (0.011)	0.28* (0.14)	0.15* (0.07)	0.11 (0.08)	0.16 (0.10)	0.11 (0.08)
public bad credit record	0.197** (0.035)	1.23** (0.20)	0.70** (0.12)	0.70** (0.12)	0.72** (0.12)	0.70** (0.12)
denied mortgage insurance	0.702** (0.045)	4.55** (0.57)	2.56** (0.30)	2.59** (0.29)	2.59** (0.30)	2.59** (0.29)
self-employed	0.060** (0.021)	0.67** (0.21)	0.36** (0.11)	0.35** (0.11)	0.34** (0.11)	0.35** (0.11)
single				0.23** (0.08)	0.23** (0.08)	0.23** (0.08)
high school diploma				−0.61** (0.23)	−0.60* (0.24)	−0.62** (0.23)
unemployment rate				0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
condominium					−0.05 (0.09)	
black × P/I ratio						−0.58 (1.47)
black × housing expense-to-income ratio						1.23 (1.69)
additional credit rating indicator variables	no	no	no	no	yes	no
constant	−0.183** (0.028)	−5.71** (0.48)	−3.04** (0.23)	−2.57** (0.34)	−2.90** (0.39)	−2.54** (0.35)

(Table 11.2 continued)

F-Statistics and p-Values Testing Exclusion of Groups of Variables

	(1)	(2)	(3)	(4)	(5)	(6)
applicant single; high school diploma; industry unemployment rate				5.85 (< 0.001)	5.22 (0.001)	5.79 (< 0.001)
additional credit rating indicator variables					1.22 (0.291)	
race interactions and black						4.96 (0.002)
race interactions only						0.27 (0.766)
difference in predicted probability of denial, white vs. black (percentage points)	8.4%	6.0%	7.1%	6.6%	6.3%	6.5%

These regressions were estimated using the $n = 2380$ observations in the Boston HMDA data set described in Appendix 11.1. The linear probability model was estimated by OLS, and probit and logit regressions were estimated by maximum likelihood. Standard errors are given in parentheses under the coefficients, and p-values are given in parentheses under the F-statistics. The change in predicted probability in the final row was computed for a hypothetical applicant whose values of the regressors, other than race, equal the sample mean. Individual coefficients are statistically significant at the *5% or **1% level.

equal. Of the nine variables (other than race) in the regression, the coefficients on all but two are statistically significant at the 5% level, which is consistent with loan officers' considering many factors when they make their decisions.

The coefficient on *black* in regression (1) is 0.084, indicating that the difference in denial probabilities for black and white applicants is 8.4 percentage points, holding constant the other variables in the regression. This is statistically significant at the 1% significance level ($t = 3.65$).

The logit and probit estimates reported in columns (2) and (3) yield similar conclusions. In the logit and probit regressions, eight of the nine coefficients on variables other than race are individually statistically significantly different from 0 at the 5% level, and the coefficient on *black* is statistically significant at the 1% level. As discussed in Section 11.2, because these models are nonlinear, specific values of all the regressors must be chosen to compute the difference in predicted probabilities for white applicants and black applicants. A conventional way to make this choice is to consider an "average" applicant who has the sample average values of all the regressors other than race. The final row in Table 11.2 reports this estimated difference in probabilities, evaluated for this average applicant. The estimated racial differentials are similar to each other: 8.4 percentage points for the linear probability model [column (1)], 6.0 percentage points for the logit model [column (2)], and 7.1 percentage points for the probit model [column (3)]. These estimated race effects and the coefficients on *black* are less than in the regressions of the previous sections, in which the only regressors were *P/I ratio* and *black*, indicating that those earlier estimates had omitted variable bias.

The regressions in columns (4) through (6) investigate the sensitivity of the results in column (3) to changes in the regression specification. Column (4) modifies

column (3) by including additional applicant characteristics. These characteristics help to predict whether the loan is denied; for example, having at least a high school diploma reduces the probability of denial (the estimate is negative, and the coefficient is statistically significant at the 1% level). However, controlling for these personal characteristics does not change the estimated coefficient on *black* or the estimated difference in denial probabilities (6.6%) in an important way.

Column (5) breaks out the six consumer credit categories and four mortgage credit categories to test the null hypothesis that these two variables enter linearly; this regression also adds a variable indicating whether the property is a condominium. The null hypothesis that the credit rating variables enter the expression for the *z*-value linearly is not rejected, nor is the condominium indicator significant, at the 5% level. Most importantly, the estimated racial difference in denial probabilities (6.3%) is essentially the same as in columns (3) and (4).

Column (6) examines whether there are interactions. Are different standards applied to evaluating the payment-to-income and housing expense-to-income ratios for black applicants versus white applicants? The answer appears to be no: The interaction terms are not jointly statistically significant at the 5% level. However, race continues to have a significant effect, because the race indicator and the interaction terms are jointly statistically significant at the 1% level. Again, the estimated racial difference in denial probabilities (6.5%) is essentially the same as in the other probit regressions.

In all six specifications, the effect of race on the denial probability, holding other applicant characteristics constant, is statistically significant at the 1% level. The estimated difference in denial probabilities between black applicants and white applicants ranges from 6.0 percentage points to 8.4 percentage points.

One way to assess whether this differential is large or small is to return to a variation on the question posed at the beginning of this chapter. Suppose two individuals apply for a mortgage, one white and one black, but otherwise having the same values of the other independent variables in regression (3); specifically, aside from race, the values of the other variables in regression (3) are the sample average values in the HMDA data set. The white applicant faces a 7.4% chance of denial, but the black applicant faces a 14.5% chance of denial. The estimated racial difference in denial probabilities, 7.1 percentage points, means that the black applicant is nearly twice as likely to be denied as the white applicant.

The results in Table 11.2 (and in the original Boston Fed study) provide statistical evidence of racial patterns in mortgage denial that, by law, ought not be there. This evidence played an important role in spurring policy changes by bank regulators.[5] But economists love a good argument, and not surprisingly these results have also stimulated a vigorous debate.

Because the suggestion that there is (or was) racial discrimination in lending is charged, we briefly review some points of this debate. In so doing, it is useful to adopt the framework of Chapter 9—that is, to consider the internal and external validity of

---

[5]These policy shifts include changes in the way that fair lending examinations were done by federal bank regulators, changes in inquiries made by the U.S. Department of Justice, and enhanced education programs for banks and other home loan origination companies.

the results in Table 11.2, which are representative of previous analyses of the Boston HMDA data. A number of the criticisms made of the original Federal Reserve Bank of Boston study concern internal validity: possible errors in the data, alternative nonlinear functional forms, additional interactions, and so forth. The original data were subjected to a careful audit, some errors were found, and the results reported here (and in the final published Boston Fed study) are based on the "cleaned" data set. Estimation of other specifications—different functional forms and/or additional regressors—also produces estimates of racial differentials comparable to those in Table 11.2. A potentially more difficult issue of internal validity is whether there is relevant nonracial financial information obtained during in-person loan interviews, but not recorded on the loan application itself, that is correlated with race; if so, there still might be omitted variable bias in the Table 11.2 regressions. Finally, some have questioned external validity: Even if there was racial discrimination in Boston in 1990, it is wrong to implicate lenders elsewhere today. Moreover, racial discrimination might be less likely using modern online applications because the mortgage can be approved or denied without a face-to-face meeting. The only way to resolve the question of external validity is to consider data from other locations and years.[6]

11.5   Conclusion

When the dependent variable Y is binary, the population regression function is the probability that $Y = 1$, conditional on the regressors. Estimation of this population regression function entails finding a functional form that does justice to its probability interpretation, estimating the unknown parameters of that function, and interpreting the results. The resulting predicted values are predicted probabilities, and the estimated effect of a change in a regressor X is the estimated change in the probability that $Y = 1$ arising from the change in X.

A natural way to model the probability that $Y = 1$ given the regressors is to use a cumulative distribution function, where the argument of the c.d.f. depends on the regressors. Probit regression uses a normal c.d.f. as the regression function, and logit regression uses a logistic c.d.f. Because these models are nonlinear functions of the unknown parameters, those parameters are more complicated to estimate than linear regression coefficients. The standard estimation method is maximum likelihood. In practice, statistical inference using the maximum likelihood estimates proceeds the same way as it does in linear multiple regression; for example, 95% confidence intervals for a coefficient are constructed as the estimated coefficient ± 1.96 standard errors.

---

[6]If you are interested in further reading on this topic, a good place to start is the symposium on racial discrimination and economics in the Spring 1998 issue of the *Journal of Economic Perspectives*. The article in that symposium by Helen Ladd (1998) surveys the evidence and debate on racial discrimination in mortgage lending. A more detailed treatment is given in Goering and Wienk (1996). The U.S. mortgage market has changed dramatically since the Boston Fed study, including a relaxation of lending standards, a bubble in housing prices, the financial crisis of 2008–2009, and a return to tighter lending standards. For an introduction to changes in mortgage markets, see Green and Wachter (2008).

James Heckman and Daniel McFadden, Nobel Laureates

The 2000 Nobel Prize in Economics was awarded jointly to two econometricians, James J. Heckman of the University of Chicago and Daniel L. McFadden of the University of California at Berkeley, for fundamental contributions to the analysis of data on individuals and firms. Much of their work addressed difficulties that arise with limited dependent variables.

Heckman was awarded the prize for developing tools for handling sample selection. As discussed in Section 9.2, sample selection bias occurs when the availability of data is influenced by a selection process related to the value of the dependent variable. For example, suppose you want to estimate the relationship between earnings and some regressor, X, using a random sample from the population. If you estimate the regression using the subsample of employed workers— that is, those reporting positive earnings—the OLS estimate could be subject to selection bias. Heckman's solution was to specify a preliminary equation with a binary dependent variable indicating whether the worker is in or out of the labor force (in or out of the subsample) and to treat this equation and the earnings equation as a system of simultaneous equations. This general strategy has been extended to selection problems that arise in many fields, ranging from labor economics to industrial organization to finance.

McFadden was awarded the prize for developing models for analyzing discrete choice data (does a high school graduate join the military, go to college, or get a job?). He started by considering the problem of an individual maximizing the expected utility of each possible choice, which could depend on observable variables (such as wages, job characteristics, and family background). He then derived models for the individual choice probabilities with unknown coefficients, which in turn could be estimated by maximum likelihood. These models and their extensions have proven widely useful in analyzing discrete choice data in many fields, including labor economics, health economics, and transportation economics.

For more information on these and other Nobel laureates in economics, visit the Nobel Foundation website, http://www.nobel.se/economics.
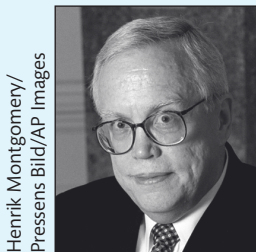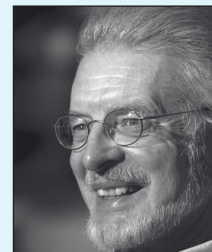
James J. Heckman   Daniel L. McFadden

Despite its intrinsic nonlinearity, sometimes the population regression function can be adequately approximated by a linear probability model—that is, by the straight line produced by linear multiple regression. The linear probability model, probit regression, and logit regression all give similar bottom-line answers when they are applied to the Boston HMDA data: All three methods estimate substantial differences in mortgage denial rates for otherwise similar black applicants and white applicants.

Binary dependent variables are the most common example of limited dependent variables, which are dependent variables with a limited range. The final quarter of the 20th century saw important advances in econometric methods for analyzing other limited dependent variables (see the box "James Heckman and Daniel McFadden, Nobel Laureates"). Some of these methods are reviewed in Appendix 11.3.

Summary

1. When Y is a binary variable, the population regression function shows the probability that $Y = 1$ given the value of the regressors, $X_1, X_2, \ldots, X_k$.

2. The linear multiple regression model is called the linear probability model when Y is a binary variable because the probability that $Y = 1$ is a linear function of the regressors.

3. Probit and logit regression models are nonlinear regression models used when Y is a binary variable. Unlike the linear probability model, probit and logit regressions ensure that the predicted probability that $Y = 1$ is between 0 and 1 for all values of X.

4. Probit regression uses the standard normal cumulative distribution function. Logit regression uses the logistic cumulative distribution function. Logit and probit coefficients are estimated by maximum likelihood.

5. The values of coefficients in probit and logit regressions are not easy to interpret. Changes in the probability that $Y = 1$ associated with changes in one or more of the X's can be calculated using the general procedure for nonlinear models outlined in Key Concept 8.1.

6. Hypothesis tests on coefficients in the linear probability, logit, and probit models are performed using the usual t- and F-statistics.

Key Terms

limited dependent variable (351)

linear probability model (353)

probit (355)

logit (355)

logistic regression (355)

likelihood function (363)

maximum likelihood estimator (MLE) (363)

fraction correctly predicted (364)

pseudo-R^2 (365)

MyLab Economics Can Help You Get a Better Grade

MyLab Economics If your exam were tomorrow, would you be ready? For each chapter, **MyLab Economics** Practice Tests and Study Plan help you prepare for your exams. You can also find the Exercises and all Review the Concepts Questions available now in **MyLab Economics**. To see how it works, turn to the **MyLab Economics** spread on the inside front cover of this text and then go to **www.pearson.com/mylab/economics**.

For additional Empirical Exercises and Data Sets, log on to the Companion Website at **www.pearsonhighered.com/stock_watson**.

Review the Concepts

11.1 Suppose a linear probability model yields a predicted value of Y that is equal to 1.3. Explain why this is nonsensical.

11.2   In Table 11.2, the estimated coefficient on *black* is 0.084 in column (1), 0.688 in column (2), and 0.389 in column (3). In spite of these large differences, all three models yield similar estimates of the marginal effect of race on the probability of mortgage denial. How can this be?

11.3   One of your friends is using data on individuals to study the determinants of smoking at your university. She asks you whether she should use a probit, logit, or linear probability model. What advice do you give her? Why?

11.4   Why are the coefficients of probit and logit models estimated by maximum likelihood instead of OLS?

Exercises

Exercises 11.1 through 11.5 are based on the following scenario: Four hundred driver's license applicants were randomly selected and asked whether they passed their driving test ($Pass_i = 1$) or failed their test ($Pass_i = 0$); data were also collected on their sex ($Male_i = 1$ if male and $= 0$ if female) and their years of driving experience ($Experience_i$, in years). The following table summarizes several estimated models.

11.1   Using the results in column (1):

   a. Does the probability of passing the test depend on *Experience*? Explain.

   b. Matthew has 10 years of driving experience. What is the probability that he will pass the test?

   c. Christopher is a new driver (0 years of experience). What is the probability that he will pass the test?

   d. The sample included values of *Experience* between 0 and 40 years, and only four people in the sample had more than 30 years of driving experience. Jed is 95 years old and has been driving since he was 15. What

Dependent Variable: *Pass*

	Probit	Logit	Linear Probability	Probit	Logit	Linear Probability	Probit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Experience	0.031 (0.009)	0.040 (0.016)	0.006 (0.002)				0.041 (0.156)
Male				−0.333 (0.161)	−0.622 (0.303)	−0.071 (0.034)	−0.174 (0.259)
Male × Experience							−0.015 (0.019)
Constant	0.712 (0.126)	1.059 (0.221)	0.774 (0.034)	1.282 (0.124)	2.197 (0.242)	0.900 (0.022)	0.806 (0.200)

is the model's prediction for the probability that Jed will pass the test? Do you think that this prediction is reliable? Why or why not?

11.2 **a.** Answer (a) through (c) from Exercise 11.1 using the results in column (2).

   b. Sketch the predicted probabilities from the probit and logit regressions in columns (1) and (2) for values of *Experience* between 0 and 60. Are the probit and logit models similar?

11.3 **a.** Answer (a) through (c) from Exercise 11.1 using the results in column (3).

   b. Sketch the predicted probabilities from the probit and linear probability model regressions in columns (1) and (3) as a function of *Experience* for values of *Experience* between 0 and 60. Do you think that the linear probability is appropriate here? Why or why not?

11.4 Using the results in columns (4) through (6):

   a. Compute the estimated probabilities of passing the test for men and for women.

   b. Are the models in (4) through (6) different? Why or why not?

11.5 Using the results in column (7):

   a. Akira is a man with 10 years of driving experience. What is the probability that he will pass the test?

   b. Jane is a woman with 2 years of driving experience. What is the probability that she will pass the test?

   c. Does the effect of experience on test performance depend on his or her sex? Explain.

11.6 Use the estimated probit model in Equation (11.8) to answer the following questions:

   a. A black mortgage applicant has a *P/I ratio* of 0.35. What is the probability that his application will be denied?

   b. Suppose the applicant reduced this ratio to 0.30. What effect would this have on his probability of being denied a mortgage?

   c. Repeat (a) and (b) for a white applicant.

   d. Does the marginal effect of the *P/I ratio* on the probability of mortgage denial depend on race? Explain.

11.7 Repeat Exercise 11.6 using the logit model in Equation (11.10). Are the logit and probit results similar? Explain.

11.8 Consider the linear probability model $Y_i = \beta_0 + \beta_1 X_i + u_i$, and assume that $E(u_i | X_i) = 0$.

   a. Show that $Pr(Y_i = 1 | X_i) = \beta_0 + \beta_1 X_i$.

   b. Show that $\text{var}(u_i|X_i) = (\beta_0 + \beta_1 X_i)[1 - (\beta_0 + \beta_1 X_i)]$. [*Hint:* Review Equation (2.7).]

   c. Is u_i heteroskedastic? Explain.

   d. (Requires Section 11.3) Derive the likelihood function.

11.9 Use the estimated linear probability model shown in column (1) of Table 11.2 to answer the following:

   a. Two applicants, one white and one black, apply for a mortgage. They have the same values for all the regressors other than race. How much more likely is the black applicant to be denied a mortgage?

   b. Construct a 95% confidence interval for your answer to (a).

   c. Think of an important omitted variable that might bias the answer in (a). What is it, and how would it bias the results?

11.10 (Requires Section 11.3 and calculus) Suppose a random variable Y has the following probability distribution: $\Pr(Y = 1) = p$, $\Pr(Y = 2) = q$, and $\Pr(Y = 3) = 1 - p - q$. A random sample of size n is drawn from this distribution, and the random variables are denoted $Y_1, Y_2, \ldots, Y_n$.

   a. Derive the likelihood function for the parameters p and q.

   b. Derive formulas for the MLE of p and q.

11.11 (Requires Appendix 11.3) Which model would you use for

   a. A study explaining the number of minutes that a person spends talking on a cell phone during the month?

   b. A study explaining grades (A through F) in a large Principles of Economics class?

   c. A study of consumers' choices for Coke, Pepsi, or generic cola?

   d. A study of the number of cell phones owned by a family?

Empirical Exercises

E11.1 In April 2008, the unemployment rate in the United States stood at 5.0%. By April 2009, it had increased to 9.0%, and it had increased further, to 10.0%, by October 2009. Were some groups of workers more likely to lose their jobs than others during the Great Recession? For example, were young workers more likely to lose their jobs than middle-aged workers? What about workers with a college degree versus those without a degree or women versus men? On the text website, **http://www.pearsonhighered.com/stock_watson**, you will find the data file **Employment_08_09**, which contains a random sample of 5440 workers who were surveyed in April 2008 and reported that they were employed full-time. A detailed description is given in **Employment_08_09_Description**,

available on the website. These workers were surveyed one year later, in April 2009, and asked about their employment status (employed, unemployed, or out of the labor force). The data set also includes various demographic measures for each individual. Use these data to answer the following questions.

a. What fraction of workers in the sample were employed in April 2009? Use your answer to compute a 95% confidence interval for the probability that a worker was employed in April 2009, conditional on being employed in April 2008.

b. Regress *Employed* on *Age* and *Age*^{2}, using a linear probability model.

   i. Based on this regression, was age a statistically significant determinant of employment in April 2009?

   ii. Is there evidence of a nonlinear effect of age on the probability of being employed?

   iii. Compute the predicted probability of employment for a 20-year-old worker, a 40-year-old worker, and a 60-year-old worker.

c. Repeat (b) using a probit regression.

d. Repeat (b) using a logit regression.

e. Are there important differences in your answers to (b)–(d)? Explain.

f. The data set includes variables measuring the workers' educational attainment, sex, race, marital status, region of the country, and weekly earnings in April 2008.

   i. Construct a table like Table 11.2 to investigate whether the conclusions on the effect of age on employment from (b)–(d) are affected by omitted variable bias.

   ii. Use the regressions in your table to discuss the characteristics of workers who were hurt most by the Great Recession.

g. The results in (a)–(f) were based on the probability of employment. Workers who are not employed can either be (i) unemployed or (ii) out the labor force. Do the conclusions you reached in (a)–(f) also hold for workers who became unemployed? (*Hint:* Use the binary variable *Unemployed* instead of *Employed*.)

h. These results have covered employment transitions during the Great Recession, but what about transitions during normal times? On the text website, you will find the data file **Employment_06_07**, which measures the same variables but for the years 2006–2007. Analyze these data and comment on the differences in employment transitions during recessions and normal times.

E11.2 Believe it or not, workers used to be able to smoke inside office buildings. Smoking bans were introduced in several areas during the 1990s. Supporters of these bans argued that in addition to eliminating the externality of secondhand

smoke, they would encourage smokers to quit by reducing their opportunities to smoke. In this assignment, you will estimate the effect of workplace smoking bans on smoking, using data on a sample of 10,000 U.S. indoor workers from 1991 to 1993, available on the text website, **http://www.pearsonhighered.com/stock_watson**, in the file **Smoking**. The data set contains information on whether individuals were or were not subject to a workplace smoking ban, whether the individuals smoked, and other individual characteristics.[7] A detailed description is given in **Smoking_Description**, available on the website.

a. Estimate the probability of smoking for (i) all workers, (ii) workers affected by workplace smoking bans, and (iii) workers not affected by workplace smoking bans.

b. What is the difference in the probability of smoking between workers affected by a workplace smoking ban and workers not affected by a workplace smoking ban? Use a linear probability model to determine whether this difference is statistically significant.

c. Estimate a linear probability model with *smoker* as the dependent variable and the following regressors: *smkban*, *female*, *age*, age^2, *hsdrop*, *hsgrad*, *colsome*, *colgrad*, *black*, and *hispanic*. Compare the estimated effect of a smoking ban from this regression with your answer from (b). Suggest an explanation, based on the substance of this regression, for the change in the estimated effect of a smoking ban between (b) and (c).

d. Test the hypothesis that the coefficient on *smkban* is 0 in the population version of the regression in (c) against the alternative that it is nonzero, at the 5% significance level.

e. Test the hypothesis that the probability of smoking does not depend on the level of education in the regression in (c). Does the probability of smoking increase or decrease with the level of education?

f. Repeat (c)–(e) using a probit model.

g. Repeat (c)–(e) using a logit model.

h. i. Mr. A is white, non-Hispanic, 20 years old, and a high school dropout. Using the probit regression and assuming that Mr. A is not subject to a workplace smoking ban, calculate the probability that Mr. A smokes. Carry out the calculation again, assuming that he is subject to a workplace smoking ban. What is the effect of the smoking ban on the probability of smoking?

    ii. Repeat (i) for Ms. B, a female, black, 40-year-old college graduate.

   iii. Repeat (i)–(ii) using the linear probability model.

---

[7]These data were provided by Professor William Evans of the University of Maryland and were used in his paper with Matthew Farrelly and Edward Montgomery, "Do Workplace Smoking Bans Reduce Smoking?" *American Economic Review*, 1999, 89(4): 728–747.

iv. Repeat (i)–(ii) using the logit model.

v. Based on your answers to (i)–(iv), do the logit, probit, and linear probability models differ? If they do, which results make most sense? Are the estimated effects large in a real-world sense?

11.1 The Boston HMDA Data Set

The Boston HMDA data set was collected by researchers at the Federal Reserve Bank of Boston. The data set combines information from mortgage applications and a follow-up survey of the banks and other lending institutions that received these mortgage applications. The data pertain to mortgage applications made in 1990 in the greater Boston metropolitan area. The full data set has 2925 observations, consisting of all mortgage applications by blacks and Hispanics plus a random sample of mortgage applications by whites.

To narrow the scope of the analysis in this chapter, we use a subset of the data for single-family residences only (thereby excluding data on multifamily homes) and for black applicants and white applicants only (thereby excluding data on applicants from other minority groups). This leaves 2380 observations. Definitions of the variables used in this chapter are given in Table 11.1.

These data were graciously provided to us by Geoffrey Tootell of the Research Department of the Federal Reserve Bank of Boston. More information about this data set, along with the conclusions reached by the Federal Reserve Bank of Boston researchers, is available in Munnell et al. (1996).

11.2 Maximum Likelihood Estimation

This appendix provides a brief introduction to maximum likelihood estimation in the context of the binary response models discussed in this chapter. We start by deriving the MLE of the success probability p for n i.i.d. observations of a Bernoulli random variable. We then turn to the probit and logit models and discuss the pseudo-R^2. We conclude with a discussion of standard errors for predicted probabilities. This appendix uses calculus at two points.

MLE for n i.i.d. Bernoulli Random Variables

The first step in computing the MLE is to derive the joint probability distribution. For n i.i.d. observations on a Bernoulli random variable, this joint probability distribution is the extension of the $n = 2$ case in Section 11.3 to general n:

$$\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$$
$$= [p^{y_1}(1-p)^{(1-y_1)}] \times [p^{y_2}(1-p)^{(1-y_2)}] \times \cdots \times [p^{y_n}(1-p)^{(1-y_n)}]$$
$$= p^{(y_1 + \cdots + y_n)}(1-p)^{n-(y_1 + \cdots + y_n)}. \tag{11.13}$$

The likelihood function is the joint probability distribution, treated as a function of the unknown coefficients. Let $S = \sum_{i=1}^{n} Y_i$; then the likelihood function is

$$f_{Bernoulli}(p; Y_1, \ldots, Y_n) = p^S (1 - p)^{n-S}. \tag{11.14}$$

The MLE of p is the value of p that maximizes the likelihood in Equation (11.14). The likelihood function can be maximized using calculus. It is convenient to maximize not the likelihood but rather its logarithm (because the logarithm is a strictly increasing function, maximizing the likelihood or its logarithm gives the same estimator). The log likelihood is $S\ln(p) + (n - S)\ln(1 - p)$, and the derivative of the log likelihood with respect to p is

$$\frac{d}{dp} \ln\left[f_{Bernoulli} \left(p; Y_1, \ldots, Y_n \right) \right] = \frac{S}{p} - \frac{n - S}{1 - p}. \tag{11.15}$$

Setting the derivative in Equation (11.15) to 0 and solving for p yields the MLE $\hat{p} = S/n = \overline{Y}$.

MLE for the Probit Model

For the probit model, the probability that $Y_i = 1$, conditional on $X_{1i}, \ldots, X_{ki}$, is $p_i = \Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$. The conditional probability distribution for the ith observation is $\Pr[Y_i = y_i | X_{1i}, \ldots, X_{ki}] = p_i^{y_i}(1 - p_i)^{1-y_i}$. Assuming that $(X_{1i}, \ldots, X_{ki}, Y_i)$ are i.i.d., $i = 1, \ldots, n$, the joint probability distribution of $Y_1, \ldots, Y_n$, conditional on the X's, is

$$\begin{aligned}
\Pr(Y_1 = y_1, \ldots, Y_n &= y_n | X_{1i}, \ldots, X_{ki}, i = 1, \ldots, n) \\
&= Pr(Y_1 = y_1 | X_{11}, \ldots, X_{k1}) \times \cdots \times Pr(Y_n = y_n | X_{1n}, \ldots, X_{kn}) \\
&= p_1^{y_1}(1 - p_1)^{1-y_1} \times \cdots \times p_n^{y_n}(1 - p_n)^{1-y_n}.
\end{aligned} \tag{11.16}$$

The likelihood function is the joint probability distribution, treated as a function of the unknown coefficients. It is conventional to consider the logarithm of the likelihood. Accordingly, the log likelihood function is

$$\begin{aligned}
\ln[f_{probit}(\beta_0, \ldots, \beta_k; Y_1, &\ldots, Y_n X_{1i}, \ldots, X_{ki}, i = 1, \ldots, n)] \\
&= \sum_{i=1}^{n} Y_i \ln[\Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})] \\
&\quad + \sum_{i=1}^{n} (1 - Y_i) \ln[1 - \Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})],
\end{aligned} \tag{11.17}$$

where this expression incorporates the probit formula for the conditional probability, $p_i = \Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$.

The MLE for the probit model maximizes the likelihood function or, equivalently, the logarithm of the likelihood function given in Equation (11.17). Because there is no simple formula for the MLE, the probit likelihood function must be maximized using a numerical algorithm on the computer.

Under general conditions, maximum likelihood estimators are consistent and have a normal sampling distribution in large samples.

MLE for the Logit Model

The likelihood for the logit model is derived in the same way as the likelihood for the probit model. The only difference is that the conditional success probability p_i for the logit model is given by Equation (11.9). Accordingly, the log likelihood of the logit model is given by Equation (11.17), with $\Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$ replaced by $[1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}]^{-1}$. As with the probit model, there is no simple formula for the MLE of the logit coefficients, so the log likelihood must be maximized numerically.

Pseudo-R^2

The pseudo-R^2 compares the value of the likelihood of the estimated model to the value of the likelihood when none of the X's are included as regressors. Specifically, the pseudo-R^2 for the probit model is

$$\text{pseudo} - R^2 = 1 - \frac{\ln(f_{probit}^{\max})}{\ln(f_{Bernoulli}^{\max})}, \tag{11.18}$$

where $f_{probit}^{\max}$ is the value of the maximized probit likelihood (which includes the X's) and $f_{Bernoulli}^{\max}$ is the value of the maximized Bernoulli likelihood (the probit model excluding all the X's).

Standard Errors for Predicted Probabilities

For simplicity, consider the case of a single regressor in the probit model. Then the predicted probability at a fixed value of that regressor, x, is $\hat{p}(x) = \Phi(\hat{\beta}_0^{MLE} + \hat{\beta}_1^{MLE} x)$, where $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$ are the MLEs of the two probit coefficients. Because this predicted probability depends on the estimators $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$, and because those estimators have a sampling distribution, the predicted probability will also have a sampling distribution.

The variance of the sampling distribution of $\hat{p}(x)$ is calculated by approximating the function $\Phi(\hat{\beta}_0^{MLE} + \hat{\beta}_1^{MLE} x)$, a nonlinear function of $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$, by a linear function of $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$. Specifically, let

$$\hat{p}(x) = \Phi(\hat{\beta}_0^{MLE} + \hat{\beta}_1^{MLE} x) \cong c + a_0(\hat{\beta}_0^{MLE} - \beta_0) + a_1(\hat{\beta}_1^{MLE} - \beta_1), \tag{11.19}$$

where the constant c and factors a_0 and a_1 depend on x and are obtained from calculus. [Equation (11.19) is a first-order Taylor series expansion; $c = \Phi(\beta_0 + \beta_1 x)$; and a_0 and a_1 are the partial derivatives, $a_0 = \partial \Phi(\beta_0 + \beta_1 x)/\partial \beta_0|_{\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}}$ and $a_1 = \partial \Phi(\beta_0 + \beta_1 x)/\partial \beta_1|_{\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}}$.] The variance of $\hat{p}(x)$ now can be calculated using the approximation in Equation (11.19) and the expression for the variance of the sum of two random variables in Equation (2.32):

$$\begin{aligned} \text{var}[\hat{p}(x)] &\cong \text{var}[c + a_0(\hat{\beta}_0^{MLE} - \beta_0) + a_1(\hat{\beta}_1^{MLE} - \beta_1)] \\ &= a_0^2 \text{var}(\hat{\beta}_0^{MLE}) + a_1^2 \text{var}(\hat{\beta}_1^{MLE}) + 2a_0 a_1 \text{cov}(\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}). \end{aligned} \tag{11.20}$$

Using Equation (11.20), the standard error of $\hat{p}(x)$ can be calculated using estimates of the variances and covariance of the MLEs.

11.3 Other Limited Dependent Variable Models

This appendix surveys some models for limited dependent variables, other than binary variables, found in econometric applications. In most cases, the OLS estimators of the parameters of limited dependent variable models are inconsistent, and estimation is routinely done using maximum likelihood. There are several advanced references available to the reader interested in further details; see, for example, Greene (2018), Ruud (2000), and Wooldridge (2010).

Censored and Truncated Regression Models

Suppose you have cross-sectional data on car purchases by individuals in a given year. Car buyers have positive expenditures, which can reasonably be treated as continuous random variables, but nonbuyers spend $0. Thus the distribution of car expenditures is a combination of a discrete distribution (at 0) and a continuous distribution.

Nobel laureate James Tobin developed a useful model for a dependent variable with a partly continuous and partly discrete distribution (Tobin, 1958). Tobin suggested modeling the ith individual in the sample as having a desired level of spending, Y_i^*, that is related to the regressors (for example, family size) according to a linear regression model. That is, when there is a single regressor, the desired level of spending is

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n. \tag{11.21}$$

If Y_i^* (what the consumer wants to spend) exceeds some cutoff, such as the minimum price of a car, the consumer buys the car and spends $Y_i = Y_i^*$, which is observed. However, if Y_i^* is less than the cutoff, spending of $Y_i = 0$ is observed instead of Y_i^*.

When Equation (11.21) is estimated using observed expenditures Y_i in place of Y_i^*, the OLS estimator is inconsistent. Tobin solved this problem by deriving the likelihood function using the additional assumption that u_i has a normal distribution, and the resulting MLE has been used by applied econometricians to analyze many problems in economics. In Tobin's honor, Equation (11.21), combined with the assumption of normal errors, is called the *tobit* regression model. The tobit model is an example of a *censored regression model*, so called because the dependent variable has been "censored" above or below a certain cutoff.

Sample Selection Models

In the censored regression model, there are data on buyers and nonbuyers, as there would be if the data were obtained via simple random sampling of the adult population. If, however, the data are collected from sales tax records, then the data would include only buyers: There would

be no data at all for nonbuyers. Data in which observations are unavailable above or below a threshold (data for buyers only) are called truncated data. The *truncated regression model* is a regression model applied to data in which observations are simply unavailable when the dependent variable is above or below a certain cutoff.

The truncated regression model is an example of a sample selection model, in which the selection mechanism (an individual is in the sample by virtue of buying a car) is related to the value of the dependent variable (expenditure on a car). As discussed in the box "James Heckman and Daniel McFadden, Nobel Laureates" in Section 11.5, one approach to estimation of sample selection models is to develop two equations, one for Y_i^* and one for whether Y_i^* is observed. The parameters of the model can then be estimated by maximum likelihood, or, in a stepwise procedure, estimating the selection equation first and then estimating the equation for Y_i^*. For additional discussion, see Ruud (2000, Chapter 28), Greene (2018, Chapter 19), or Wooldridge (2010, Chapter 17).

Count Data

Count data arise when the dependent variable is a counting number—for example, the number of restaurant meals eaten by a consumer in a week. When these numbers are large, the variable can be treated as approximately continuous, but when they are small, the continuous approximation is a poor one. The linear regression model, estimated by OLS, can be used for count data, even if the number of counts is small. Predicted values from the regression are interpreted as the expected value of the dependent variable, conditional on the regressors. So when the dependent variable is the number of restaurant meals eaten, a predicted value of 1.7 means, on average, 1.7 restaurant meals per week. As in the binary regression model, however, OLS does not take advantage of the special structure of count data and can yield nonsense predictions: for example, −0.2 restaurant meals per week. Just as probit and logit eliminate nonsense predictions when the dependent variable is binary, special models do so for count data. The two most widely used models are the Poisson and negative binomial regression models.

Ordered Responses

Ordered response data arise when mutually exclusive qualitative categories have a natural ordering, such as obtaining a high school diploma, obtaining some college education (but not graduating), or graduating from college. Like count data, ordered response data have a natural ordering, but unlike count data, they do not have natural numerical values.

Because there are no natural numerical values for ordered response data, OLS is inappropriate. Instead, ordered data are often analyzed using a generalization of probit called the *ordered probit model*, in which the probability of each outcome (e.g., a college education), conditional on the independent variables (such as parents' income), is modeled using the cumulative normal distribution.

Discrete Choice Data

A *discrete choice* or *multiple choice* variable can take on multiple unordered qualitative values. One example in economics is the mode of transport chosen by a commuter: She might take the subway, ride the bus, drive, or make her way under her own power (walk, bicycle). If we were to analyze these choices, the dependent variable would have four possible outcomes (subway, bus, car, and human-powered). These outcomes are not ordered in any natural way. Instead, the outcomes are a choice among distinct qualitative alternatives.

The econometric task is to model the probability of choosing the various options given various regressors such as individual characteristics (how far the commuter's house is from the subway station) and the characteristics of each option (the price of the subway). As discussed in the box in Section 11.5, models for analysis of discrete choice data can be developed from principles of utility maximization. Individual choice probabilities can be expressed in probit or logit form, and those models are called *multinomial probit* and *multinomial logit* regression models.

CHAPTER 12

Instrumental Variables Regression

Chapter 9 discussed several problems, including omitted variables, errors in variables, and simultaneous causality, that make the error term correlated with the regressor. Omitted variable bias can be addressed directly by including the omitted variable in a multiple regression, but this is only feasible if you have data on the omitted variable. And sometimes, such as when causality runs *both* from X to Y *and* from Y to X so that there is simultaneous causality bias, multiple regression simply cannot eliminate the bias. If a direct solution to these problems is either infeasible or unavailable, a new method is required.

Instrumental variables (IV) regression is a general way to obtain a consistent estimator of the unknown causal coefficients when the regressor, X, is correlated with the error term, u. To understand how IV regression works, think of the variation in X as having two parts: one part that, for whatever reason, is correlated with u (this is the part that causes the problems) and a second part that is uncorrelated with u. If you had information that allowed you to isolate the second part, you could focus on those variations in X that are uncorrelated with u and disregard the variations in X that bias the OLS estimates. This is, in fact, what IV regression does. The information about the movements in X that are uncorrelated with u is gleaned from one or more additional variables, called **instrumental variables** or simply **instruments**. Instrumental variables regression uses these additional variables as tools or "instruments" to isolate the movements in X that are uncorrelated with u, which in turn permits consistent estimation of the regression coefficients.

The first two sections of this chapter describe the mechanics and assumptions of IV regression: why IV regression works, what is a valid instrument, and how to implement and to interpret the most common IV regression method, two stage least squares. The key to successful empirical analysis using instrumental variables is finding valid instruments, and Section 12.3 takes up the question of how to assess whether a set of instruments is valid. As an illustration, Section 12.4 uses IV regression to estimate the elasticity of demand for cigarettes. Finally, Section 12.5 turns to the difficult question of where valid instruments come from in the first place.

12.1 The IV Estimator with a Single Regressor and a Single Instrument

We start with the case of a single regressor, X, which might be correlated with the error, u. If X and u are correlated, the OLS estimator is inconsistent; that is, it may not be close to the true value of the causal coefficient even when the sample is very large [see Equation (6.1)]. As discussed in Section 9.2, this correlation between X and u can stem from various sources, including omitted variables, errors in variables (measurement errors in the regressors), and simultaneous causality (when causality runs "backward" from Y to X as well as "forward" from X to Y). Whatever the source of the correlation between X and u, if there is a valid instrumental variable, Z, the effect on Y of a unit change in X can be estimated using the instrumental variables estimator.

The IV Model and Assumptions

Let β_1 be the causal effect of X on Y. The model relating the dependent variable Y_i and regressor X_i, without any control variables, is

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n, \tag{12.1}$$

where u_i is the error term representing omitted factors that determine Y_i. If X_i and u_i are correlated, the OLS estimator is inconsistent. Instrumental variables estimation uses an additional, "instrumental" variable Z to isolate that part of X that is uncorrelated with u.

Endogeneity and exogeneity. Instrumental variables regression has some specialized terminology to distinguish variables that are correlated with the population error term u from ones that are not. Variables correlated with the error term are called **endogenous variables**, while variables uncorrelated with the error term are called **exogenous variables**. The historical source of these terms traces to models with multiple equations, in which an "endogenous" variable is determined within the model, while an "exogenous" variable is determined outside the model. For example, Section 9.2 considered the possibility that if low test scores produced decreases in the student–teacher ratio because of political intervention and increased funding, causality would run *both* from the student–teacher ratio to test scores *and* from test scores to the student–teacher ratio. This was represented mathematically as a system of two simultaneous equations [Equations (9.3) and (9.4)], one for each causal connection. As discussed in Section 9.2, because both test scores and the student–teacher ratio are determined within the model, both are correlated with the population error term u; that is, in this example, both variables are endogenous. In contrast, an exogenous variable, which is determined outside the model, is uncorrelated with u.

The two conditions for a valid instrument. A valid instrumental variable ("instrument") Z must satisfy two conditions, known as the **instrument relevance condition** and the **instrument exogeneity condition**:

1. Instrument relevance: $\text{corr}(Z_i, X_i) \neq 0$.
2. Instrument exogeneity: $\text{corr}(Z_i, u_i) = 0$.

If an instrument is relevant, then variation in the instrument is related to variation in X_i. If in addition the instrument is exogenous, then that part of the variation of X_i captured by the instrumental variable is exogenous. Thus an instrument that is relevant and exogenous can capture movements in X_i that are exogenous. This exogenous variation can in turn be used to estimate the population coefficient β_1.

The two conditions for a valid instrument are vital for instrumental variables regression, and we return to them (and their extension to multiple regressors and multiple instruments) repeatedly throughout this chapter.

The Two Stage Least Squares Estimator

If the instrument Z satisfies the conditions of instrument relevance and exogeneity, the coefficient β_1 can be estimated using an IV estimator called **two stage least squares (TSLS)**. As the name suggests, the two stage least squares estimator is calculated in two stages. The first stage decomposes X into two components: a problematic component that may be correlated with the regression error and another, problem-free component that is uncorrelated with the error. The second stage uses the problem-free component to estimate β_1.

The first stage begins with a population regression linking X and Z:

$$X_i = \pi_0 + \pi_1 Z_i + v_i, \tag{12.2}$$

where π_0 is the intercept, π_1 is the slope, and v_i is the error term. This regression provides the needed decomposition of X_i. One component is $\pi_0 + \pi_1 Z_i$, the part of X_i that can be predicted by Z_i. Because Z_i is exogenous, this component of X_i is uncorrelated with u_i, the error term in Equation (12.1). The other component of X_i is v_i, which is the problematic component of X_i that is correlated with u_i.

The idea behind TSLS is to use the problem-free component of X_i, $\pi_0 + \pi_1 Z_i$, and to disregard v_i. The only complication is that the values of π_0 and π_1 are unknown, so $\pi_0 + \pi_1 Z_i$ cannot be calculated. Accordingly, the first stage of TSLS applies OLS to Equation (12.2) and uses the predicted value from the OLS regression, $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, where $\hat{\pi}_0$ and $\hat{\pi}_1$ are the OLS estimates.

The second stage of TSLS is easy: Regress Y_i on $\hat{X}_i$ using OLS. The resulting estimators from the second-stage regression are the TSLS estimators, $\hat{\beta}_0^{TSLS}$ and $\hat{\beta}_1^{TSLS}$.

Why Does IV Regression Work?

Two examples provide some insight into why IV regression solves the problem of correlation between X_i and u_i.

Who Invented Instrumental Variables Regression?

Instrumental variables regression was first proposed in Philip G. Wright's 1928 book, *The Tariff on Animal and Vegetable Oils*. If you yearn to know how animal and vegetable oils were produced, transported, and sold in the early 20th century, the first 285 pages of the book are for you. Econometricians, however, are more interested in Appendix B. The appendix explains the simultaneous causality problem and provides two derivations of "the method of introducing external factors"—what we now call the instrumental variables estimator. It then uses IV regression to estimate the supply and demand elasticities for butter and flaxseed oil. Philip Wright was an obscure economist who never held a major academic position, but his son Sewall went on to become a preeminent population geneticist and statistician. Many econometricians assumed that Sewall wrote the appendix anonymously, but the mystery of who really wrote the appendix, and who invented IV regression, remained unsolved—until recently.

Either father or son could have been the author. Philip Wright (1861–1934) received a master's degree in economics from Harvard University in 1887, and he taught mathematics and economics (as well as literature and physical education) at a small college in Illinois. In a book review (Wright, 1915), he used a figure like Figures 12.1a and 12.1b to show how a regression of quantity on price will not, in general, estimate a demand curve but instead estimates a combination of the supply and demand curves. In the early 1920s, Sewall Wright (1889–1988) was researching the statistical analysis of multiple equations with multiple causal variables in the context of genetics, research that in part led to his assuming a professorship in 1930 at the University of Chicago.

The first clue to the mystery came from some statistical detective work using textual data analysis.

Stylometrics is the subfield of statistics, invented by Frederick Mosteller and David Wallace (1963), that uses subtle, subconscious differences in writing styles to identify authorship of disputed texts using statistical analysis of grammatical constructions and word choice. When stylometrics is used to compare Appendix B to texts known to have been written independently by Philip and by Sewall, the results are clear: Philip was the author (see Stock and Trebbi, 2003).

Does this mean that Philip Wright invented IV regression? Not quite. Recently, correspondence between Philip and Sewall has come to light. In a series of letters between December 1925 and April 1926, the two worked through the challenge of simultaneous causality and together invented IV regression. Sewall provided one derivation, which is very difficult to follow. In fact, Philip admitted he could not follow his son's derivation, so Philip provided an entirely different, much easier derivation. Philip's derivation, from a letter of March 4, 1926, is reproduced here. Philip's derivation is the one used today, and we use it in Equations (12.5) and (12.6).

Between the statistical sleuthing and the letters, the mystery is solved: Philip Wright wrote Appendix B, but the invention of IV was a joint collaboration between father and son.

Philip G. Wright   Sewall Wright

Courtesy of Rosalind W. Harris

Philip Wright's Derivation of the IV Estimator of the Supply Elasticity in His Letter of March 4, 1926, to His Son Sewall.

Philip's notation differs from ours: O is quantity (Y in our notation), P is price (X), S is other determinants of supply (the error term in the supply equation, u), and A is the external factor (the instrument Z). All variables are in percentage deviations with mean 0 (we use logs and an intercept), and e is the supply elasticity (β). The subscripts denote observation number. The first line rearranges $Y = \beta X + u$ to be $\beta X = Y - u$. He multiplies both sides by the instrument and uses instrument exogeneity ("since A is uncorrelated with S") to set the term $\sum_{i=1}^{n} A_i S_i$ (in our notation, $\sum_{i=1}^{n} Z_i u_i$) to 0 in expectation. He then uses instrument relevance to divide by $\sum_{i=1}^{n} A_i P_i$ (in our notation, $\sum_{i=1}^{n} Z_i X_i$), thereby obtaining the formula for the IV estimator, which, in our notation, is Equation (12.4).

Example 1: Philip Wright's problem. The method of instrumental variables estimation was first published in 1928 in an appendix to a book written by Philip G. Wright (1928), although the key ideas of IV regression were developed collaboratively with his son Sewall Wright (see the box "Who Invented Instrumental Variables Regression?"). Philip Wright was concerned with an important economic problem of his day: how to set an import tariff (a tax on imported goods) on animal and vegetable oils and fats, such as butter and soy oil. In the 1920s, import tariffs were a major source of tax revenue for the United States. The key to understanding the economic effect of a tariff was having quantitative estimates of the demand and supply curves of the goods. Recall that the supply elasticity is the percentage change in the quantity supplied arising from a 1% increase in the price and that the demand elasticity is the percentage change in the quantity demanded arising from a 1% increase in the price. Philip Wright needed estimates of these elasticities of supply and demand.

To be concrete, consider the problem of estimating the elasticity of demand for butter. Recall from Key Concept 8.2 that the coefficient in a linear equation relating $\ln(Y_i)$ to $\ln(X_i)$ has the interpretation of the elasticity of Y with respect to X. In Wright's problem, this suggests the demand equation

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i, \qquad (12.3)$$

where Q_i^{butter} is the ith observation on the quantity of butter consumed, P_i^{butter} is its price, and u_i represents other factors that affect demand, such as income and consumer tastes. In Equation (12.3), a 1% increase in the price of butter yields a β_1 percent change in demand, so β_1 is the demand elasticity.

Philip Wright had data on total annual butter consumption and its average annual price in the United States for 1912 to 1922. It would have been easy to use

these data to estimate the demand elasticity by applying OLS to Equation (12.3), but he had a key insight: Because of the interactions between supply and demand, the regressor, $\ln(P_i^{butter})$, was likely to be correlated with the error term.

To see this, look at Figure 12.1a, which shows the market demand and supply curves for butter for three different years. The demand and supply curves for the first period are denoted D_1 and S_1, and the first period's equilibrium price and quantity are determined by their intersection. In year 2, demand increases from D_1 to D_2 (say, because of an increase in income), and supply decreases from S_1 to S_2 (because of an increase in the cost of producing butter); the equilibrium price and quantity are determined by the intersection of the new supply and demand curves. In year 3, the factors affecting demand and supply change again; demand increases again to D_3, supply increases to S_3, and a new equilibrium quantity and price are determined. Figure 12.1b shows the equilibrium quantity and price pairs for these three periods and for eight subsequent years, where in each year the supply and demand curves are subject to shifts associated with factors other than price that affect market supply and demand. This scatterplot is like the one that Wright would have seen when he plotted his data. As he reasoned, fitting a line to these points by OLS will estimate neither a demand curve nor a supply curve because the points have been determined by changes in both demand and supply.

Wright realized that a way to get around this problem was to find some third variable that shifted supply but did not shift demand. Figure 12.1c shows what happens when such a variable shifts the supply curve but demand remains stable. Now all of the equilibrium price and quantity pairs lie on a stable demand curve, and the slope of the demand curve is easily estimated. In the instrumental variable formulation of Wright's problem, this third variable—the instrumental variable—is correlated with price (it shifts the supply curve, which leads to a change in price) but is uncorrelated with u (the demand curve remains stable). Wright considered several potential instrumental variables; one was the weather. For example, below-average rainfall in a dairy region could impair grazing and thus reduce butter production at a given price (it would shift the supply curve to the left and increase the equilibrium price), so dairy-region rainfall satisfies the condition for instrument relevance. But dairy-region rainfall should not have a direct influence on the demand for butter, so the correlation between dairy-region rainfall and u_i would be 0; that is, dairy-region rainfall satisfies the condition for instrument exogeneity.

Example 2: Estimating the effect on test scores of class size. Despite controlling for student and district characteristics, the estimates of the effect on test scores of class size reported in Part II still might have omitted variable bias resulting from unmeasured variables such as learning opportunities outside school or the quality of the teachers. If data on these variables, or on suitable control variables, are unavailable, this omitted variable bias cannot be addressed by including the variables in the multiple regressions.

Instrumental variables regression provides an alternative approach to this problem. Consider the following hypothetical example: Some California schools

FIGURE 12.1   Equilibrium Price and Quantity Data

(a) Price and quantity are determined by the intersection of the supply and demand curves. The equilibrium in the first period is determined by the intersection of the demand curve D_1 and the supply curve S_1. Equilibrium in the second period is the intersection of D_2 and S_2, and equilibrium in the third period is the intersection of D_3 and S_3.
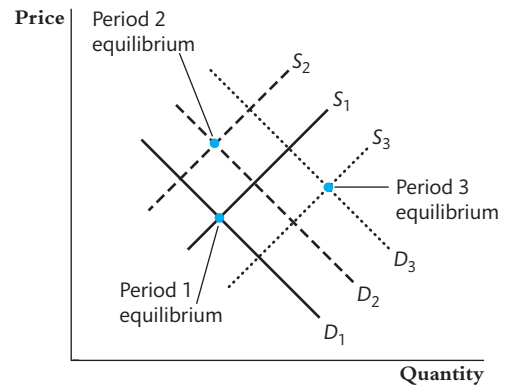
(a) Demand and supply in three time periods

(b) This scatterplot shows equilibrium price and quantity in 11 different time periods. The demand and supply curves are hidden. Can you determine the demand and supply curves from the points on the scatterplot?
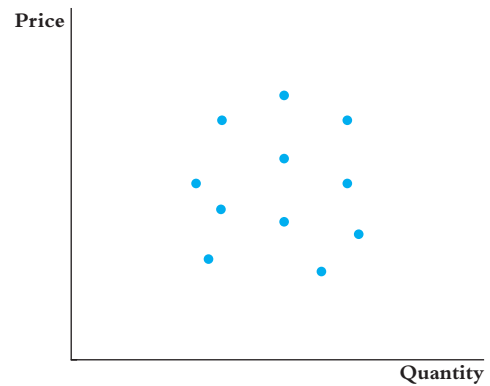
(b) Equilibrium price and quantity for 11 time periods

(c) When the supply curve shifts from S_1 to S_2 to S_3 but the demand curve remains at D_1, the equilibrium prices and quantities trace out the demand curve.
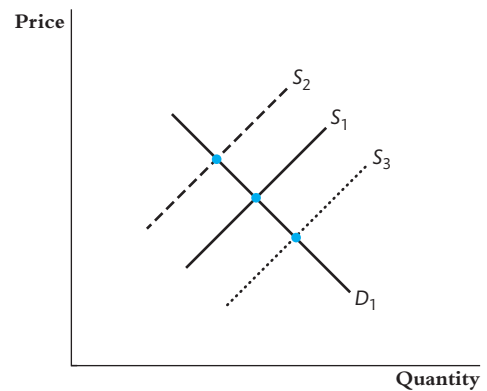
(c) Equilibrium price and quantity when only the supply curve shifts

are forced to close for repairs because of a summer earthquake. Districts closest to the epicenter are most severely affected. A district with some closed schools needs to "double up" its students, temporarily increasing class size. This means that distance from the epicenter satisfies the condition for instrument relevance because it is correlated with class size. But if distance to the epicenter is unrelated to any of the other factors affecting student performance (such as whether the students are still learning English or disruptive effects of the earthquake on student performance), then it will be exogenous because it is uncorrelated with the error term. Thus the instrumental variable, distance to the epicenter, could be used to circumvent omitted variable bias and to estimate the effect of class size on test scores.

The Sampling Distribution of the TSLS Estimator

The exact distribution of the TSLS estimator in small samples is complicated. However, like the OLS estimator, its distribution in large samples is simple: The TSLS estimator is consistent and is normally distributed.

Formula for the TSLS estimator. Although the two stages of TSLS make the estimator seem complicated, when there is a single X and a single instrument Z, as we assume in this section, there is a simple formula for the TSLS estimator. Let s_{ZY} be the sample covariance between Z and Y, and let s_{ZX} be the sample covariance between Z and X. As shown in Appendix 12.2, the TSLS estimator with a single instrument is

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}}. \tag{12.4}$$

That is, the TSLS estimator of β_1 is the ratio of the sample covariance between Z and Y to the sample covariance between Z and X.

Sampling distribution of $\hat{\beta}_1^{TSLS}$ when the sample size is large. The formula in Equation (12.4) can be used to show that $\hat{\beta}_1^{TSLS}$ is consistent and, in large samples, normally distributed. The argument is summarized here, with mathematical details given in Appendix 12.3.

The argument that $\hat{\beta}_1^{TSLS}$ is consistent combines the assumptions that Z_i is relevant and exogenous with the consistency of sample covariances for population covariances. To begin, note that because $Y_i = \beta_0 + \beta_1 X_i + u_i$ in Equation (12.1),

$$\text{cov}(Z_i, Y_i) = \text{cov}(Z_i, \beta_0 + \beta X_i + u_i) = \beta_1 \text{cov}(Z_i, X_i) + \text{cov}(Z_i, u_i), \tag{12.5}$$

where the second equality follows from the properties of covariances [Equation (2.34)]. By the instrument exogeneity assumption, $\text{cov}(Z_i, u_i) = 0$, and by the

instrument relevance assumption, $\text{cov}(Z_i, X_i) \neq 0$. Thus, if the instrument is valid, Equation (12.5) implies that

$$\beta_1 = \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, X_i)}. \tag{12.6}$$

That is, the population coefficient β_1 is the ratio of the population covariance between Z and Y to the population covariance between Z and X.

As discussed in Section 3.7, the sample covariance is a consistent estimator of the population covariance; that is, $s_{ZY} \xrightarrow{p} \text{cov}(Z_i, Y_i)$ and $s_{ZX} \xrightarrow{p} \text{cov}(Z_i, X_i)$. It follows from Equations (12.4) and (12.6) that the TSLS estimator is consistent:

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, X_i)} = \beta_1. \tag{12.7}$$

The formula in Equation (12.4) also can be used to show that the sampling distribution of $\hat{\beta}_1^{TSLS}$ is normal in large samples. The reason is the same as for every other least squares estimator we have considered: The TSLS estimator is an average of random variables, and when the sample size is large, the central limit theorem tells us that averages of random variables are normally distributed. Specifically, the numerator of the expression for $\hat{\beta}_1^{TSLS}$ in Equation (12.4) is $s_{ZY} = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})$, an average of $(Z_i - \overline{Z})(Y_i - \overline{Y})$. A bit of algebra, sketched out in Appendix 12.3, shows that because of this averaging, the central limit theorem implies that, in large samples, $\hat{\beta}_1^{TSLS}$ has a sampling distribution that is approximately $N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}})$, where

$$\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n}\frac{\text{var}\left[(Z_i - \mu_Z)u_i\right]}{\left[\text{cov}(Z_i, X_i)\right]^2}. \tag{12.8}$$

Statistical inference using the large-sample distribution. The variance $\sigma^2_{\hat{\beta}_1^{TSLS}}$ can be estimated by estimating the variance and covariance terms appearing in Equation (12.8), and the square root of the estimate of $\sigma^2_{\hat{\beta}_1^{TSLS}}$ is the standard error of the IV estimator. This is done automatically in TSLS regression commands in econometric software packages. Because $\hat{\beta}_1^{TSLS}$ is normally distributed in large samples, hypothesis tests about β_1 can be performed by computing the t-statistic, and a 95% large-sample confidence interval is given by $\hat{\beta}_1^{TSLS} \pm 1.96\,SE(\hat{\beta}_1^{TSLS})$.

Application to the Demand for Cigarettes

Philip Wright was interested in the demand elasticity of butter, but today other commodities, such as cigarettes, figure more prominently in public policy debates. One tool in the quest for reducing illnesses and deaths from smoking—and the costs, or externalities, imposed by those illnesses on the rest of society—is to tax cigarettes so heavily that current smokers cut back and potential new smokers are discouraged from taking up the habit. But precisely how big a tax hike is needed to make a dent in cigarette consumption? For example, what would the after-tax sales price of cigarettes need to be to achieve a 20% reduction in cigarette consumption?

The answer to this question depends on the elasticity of demand for cigarettes. If the elasticity is −1, then the 20% target in consumption can be achieved by a 20% increase in price. If the elasticity is −0.5, then the price must rise 40% to decrease consumption by 20%. Of course, we do not know the demand elasticity of cigarettes: We must estimate it from data on prices and sales. But, as with butter, because of the interactions between supply and demand, the elasticity of demand for cigarettes cannot be estimated consistently by an OLS regression of log quantity on log price.

We therefore use TSLS to estimate the elasticity of demand for cigarettes using annual data for the 48 contiguous U.S. states for 1985 through 1995 (the data are described in Appendix 12.1). For now, all the results are for the cross section of states in 1995; results using data for earlier years (panel data) are presented in Section 12.4.

The instrumental variable, $SalesTax_i$, is the portion of the tax on cigarettes arising from the general sales tax, measured in dollars per pack (in real dollars, deflated by the Consumer Price Index). Cigarette consumption, $Q_i^{cigarettes}$, is the number of packs of cigarettes sold per capita in the state, and the price, $P_i^{cigarettes}$, is the average real price per pack of cigarettes including all taxes.

Before using TSLS, it is essential to ask whether the two conditions for instrument validity hold. We return to this topic in detail in Section 12.3, where we provide some statistical tools that help in this assessment. Even with those statistical tools, judgment plays an important role, so it is useful to think about whether the sales tax on cigarettes plausibly satisfies the two conditions.

First consider instrument relevance. Because a high sales tax increases the after-tax sales price $P_i^{cigarettes}$, the sales tax per pack plausibly satisfies the condition for instrument relevance.

Next consider instrument exogeneity. For the sales tax to be exogenous, it must be uncorrelated with the error in the demand equation; that is, the sales tax must affect the demand for cigarettes only indirectly through the price. This seems plausible: General sales tax rates vary from state to state, but they do so mainly because different states choose different mixes of sales, income, property, and other taxes to finance public undertakings. Those choices about public finance are driven by political considerations, not by factors related to the demand for cigarettes. We discuss the credibility of this assumption more in Section 12.4, but for now we keep it as a working hypothesis.

In modern statistical software, the first stage of TSLS is estimated automatically, so you do not need to run this regression yourself to compute the TSLS estimator. Even so, it is a good idea to look at the first-stage regression. Using data for the 48 states in 1995, it is

$$\widehat{\ln(P_i^{cigarettes})} = 4.62 + 0.031 SalesTax_i. \tag{12.9}$$
$$(0.03)\ (0.005)$$

As expected, higher sales taxes mean higher after-tax prices. The R^2 of this regression is 47%, so the variation in sales tax on cigarettes explains 47% of the variance of cigarette prices across states.

In the second stage of TSLS, $\ln(Q_i^{cigarettes})$ is regressed on $\widehat{\ln(P_i^{cigarettes})}$ using OLS. The resulting estimated regression function is

$$\widehat{\ln(Q_i^{cigarettes})} = 9.72 - 1.08\widehat{\ln(P_i^{cigarettes})}. \qquad (12.10)$$

This estimated regression function is written using the regressor in the second stage, the predicted value $\widehat{\ln(P_i^{cigarettes})}$. It is, however, conventional and less cumbersome simply to report the estimated regression function with $\ln(P_i^{cigarettes})$ rather than $\widehat{\ln(P_i^{cigarettes})}$. Reported in this notation, the TSLS estimates and heteroskedasticity-robust standard errors are

$$\widehat{\ln(Q_i^{cigarettes})} = 9.72 - 1.08\ln(P_i^{cigarettes}). \qquad (12.11)$$
$$\quad\ (1.53)\ \ (0.32)$$

The TSLS estimate suggests that the demand for cigarettes is surprisingly elastic in light of their addictive nature: An increase in the price of 1% reduces consumption by 1.08%. But, recalling our discussion of instrument exogeneity, perhaps this estimate should not yet be taken too seriously. Even though the elasticity was estimated using an instrumental variable, there might still be omitted variables that are correlated with the sales tax per pack. A leading candidate is income: States with higher incomes might depend relatively less on a sales tax and more on an income tax to finance state government. Moreover, the demand for cigarettes presumably depends on income. Thus we would like to reestimate our demand equation including income as a control variable. To do so, however, we must first extend the IV regression model to include additional regressors.

12.2 The General IV Regression Model

The general IV regression model has four types of variables: the dependent variable, Y; problematic endogenous regressors, like the price of cigarettes, which are correlated with the error term and which we will label X; additional regressors W, which are either control variables or **included exogenous variables**; and instrumental variables, Z. In general, there can be multiple endogenous regressors (X's), multiple additional regressors (W's), and multiple instrumental variables (Z's).

For IV regression to be possible, there must be at least as many instrumental variables (Z's) as endogenous regressors (X's). In Section 12.1, there was a single endogenous regressor and a single instrument. Having (at least) one instrument for this single endogenous regressor was essential. Without the instrument, we could not have computed the instrumental variables estimator: there would be no first-stage regression in TSLS.

The relationship between the number of instruments and the number of endogenous regressors has its own terminology. The regression coefficients are said to be

The General Instrumental Variables Regression Model and Terminology

The general IV regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, \quad (12.12)$$

$i = 1, \ldots, n$, where

- Y_i is the dependent variable;
- $\beta_0, \beta_1, \ldots, \beta_{k+r}$ are unknown coefficients;
- $X_{1i}, \ldots, X_{ki}$ are k endogenous regressors, which are potentially correlated with u_i;
- $W_{1i}, \ldots, W_{ri}$ are r included exogenous regressors, which are uncorrelated with u_i or are control variables;
- u_i is the error term, which represents measurement error and/or omitted factors; and
- $Z_{1i}, \ldots, Z_{mi}$ are m instrumental variables.

The coefficients are overidentified if there are more instruments than endogenous regressors ($m > k$), they are underidentified if $m < k$, and they are exactly identified if $m = k$. Estimation of the IV regression model requires exact identification or overidentification.

exactly identified if the number of instruments (m) equals the number of endogenous regressors (k); that is, $m = k$. The coefficients are **overidentified** if the number of instruments exceeds the number of endogenous regressors; that is, $m > k$. They are **underidentified** if the number of instruments is less than the number of endogenous regressors; that is, $m < k$. The coefficients must be either exactly identified or overidentified if they are to be estimated by IV regression.

The general IV regression model and its terminology are summarized in Key Concept 12.1.

Included exogenous variables and control variables in IV regression. The W variables in Equation (12.12) can be either exogenous variables, in which case $E(u_i | W_i) = 0$, or they can be control variables that need not have a causal interpretation but are included to ensure that the instrument is uncorrelated with the error term. For example, Section 12.1 raised the possibility that the sales tax might be correlated with income, which economic theory tells us is a determinant of cigarette demand. If so, the sales tax would be correlated with the error term in the cigarette demand equation, $\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$, and thus

would not be an exogenous instrument. Including income in the IV regression, or including variables that control for income, would remove this source of potential correlation between the instrument and the error term. In general, if W is an effective control variable in IV regression, then including W makes the instrument uncorrelated with u, so the TSLS estimator of the coefficient on X is consistent; if W is correlated with u, however, then the TSLS coefficient on W is subject to omitted variable bias and does not have a causal interpretation. The logic of control variables in IV regression therefore parallels the logic of control variables in OLS, discussed in Section 7.5.

The mathematical condition for W to be an effective control variable in IV regression is similar to the condition on control variables in OLS discussed in Section 7.5. Specifically, including W must ensure that the conditional mean of u does not depend on Z, so conditional mean independence holds; that is, $E(u_i | Z_i, W_i) = E(u_i | W_i)$. For clarity, in the body of this chapter we focus on the case that W variables are exogenous, so that $E(u_i | W_i) = 0$. Appendix 12.6 explains how the results of this chapter extend to the case that W is a control variable, in which case the conditional mean 0 condition, $E(u_i | W_i) = 0$, is replaced by the conditional mean independence condition, $E(u_i | Z_i, W_i) = E(u_i | W_i)$.

TSLS in the General IV Model

TSLS with a single endogenous regressor. When there is a single endogenous regressor X and some additional included exogenous variables, the equation of interest is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i, \tag{12.13}$$

where, as before, X_i might be correlated with the error term, but $W_{1i}, \ldots, W_{ri}$ are not.

The population first-stage regression of TSLS relates X to the exogenous variables—that is, the W's and the instruments (Z's):

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i, \tag{12.14}$$

where $\pi_0, \pi_1, \ldots, \pi_{m+r}$ are unknown regression coefficients and v_i is an error term.

Equation (12.14) is sometimes called the **reduced form** equation for X. It relates the endogenous variable X to all the available exogenous variables, both those included in the regression of interest (W) and the instruments (Z).

In the first stage of TSLS, the unknown coefficients in Equation (12.14) are estimated by OLS, and the predicted values from this regression are $\hat{X}_1, \ldots, \hat{X}_n$.

In the second stage of TSLS, Equation (12.13) is estimated by OLS except that X_i is replaced by its predicted value from the first stage. That is, Y_i is regressed on $\hat{X}_i, W_{1i}, \ldots, W_{ri}$ using OLS. The resulting estimator of $\beta_0, \beta_1, \ldots, \beta_{1+r}$ is the TSLS estimator.

Two Stage Least Squares

The TSLS estimator in the general IV regression model in Equation (12.12) with multiple instrumental variables is computed in two stages:

1. **First-stage regression(s)**: Regress X_{1i} on the instrumental variables $(Z_{1i}, \ldots, Z_{mi})$ and the included exogenous variables and/or control variables $(W_{1i}, \ldots, W_{ri})$ using OLS, including an intercept. Compute the predicted values from this regression; call these $\hat{X}_{1i}$. Repeat this for all the endogenous regressors $X_{2i}, \ldots, X_{ki}$, thereby computing the predicted values $\hat{X}_{1i}, \ldots, \hat{X}_{ki}$.

2. **Second-stage regression**: Regress Y_i on the predicted values of the endogenous variables $(\hat{X}_{1i}, \ldots, \hat{X}_{ki})$ and the included exogenous variables and/or control variables $(W_{1i}, \ldots, W_{ri})$ using OLS, including an intercept. The TSLS estimators $\hat{\beta}_0^{TSLS}, \ldots, \hat{\beta}_{k+r}^{TSLS}$ are the estimators from the second-stage regression.

In practice, the two stages are done automatically within TSLS estimation commands in econometric software.

Extension to multiple endogenous regressors. When there are multiple endogenous regressors $X_{1i}, \ldots, X_{ki}$, the TSLS algorithm is similar except that each endogenous regressor requires its own first-stage regression. Each of these first-stage regressions has the same form as Equation (12.14); that is, the dependent variable is one of the X's, and the regressors are all the instruments (Z's) and all the included exogenous variables (W's). Together, these first-stage regressions produce predicted values of each of the endogenous regressors.

In the second stage of TSLS, Equation (12.12) is estimated by OLS except that the endogenous regressors (X's) are replaced by their respective predicted values ($\hat{X}$'s). The resulting estimator of $\beta_0, \beta_1, \ldots, \beta_{k+r}$ is the TSLS estimator.

In practice, the two stages of TSLS are done automatically within TSLS estimation commands in econometric software. The general TSLS estimator is summarized in Key Concept 12.2.

Instrument Relevance and Exogeneity in the General IV Model

The conditions of instrument relevance and exogeneity need to be modified for the general IV regression model.

When there is one included endogenous variable but multiple instruments, the condition for instrument relevance is that at least one Z is useful for predicting X given W. When there are multiple included endogenous variables, this condition is more complicated because we must rule out perfect multicollinearity in the second-stage population regression. Intuitively, when there are multiple included

The Two Conditions for Valid Instruments

A set of m instruments $Z_{1i}, \ldots, Z_{mi}$ must satisfy the following two conditions to be valid:

1. **Instrument Relevance**

   - *In general,* let $\hat{X}_{1i}^*$ be the predicted value of X_{1i} from the population regression of X_{1i} on the instruments (Z's) and the included exogenous regressors (W's), and let "1" denote the constant regressor that takes on the value 1 for all observations. Then $(\hat{X}_{1i}^*, \ldots, \hat{X}_{ki}^*, W_{1i}, \ldots, W_{ri}, 1)$ are not perfectly multicollinear.

   - *If there is only one X,* then for the previous condition to hold, at least one Z must have a nonzero coefficient in the population regression of X on the Z's and the W's.

2. **Instrument Exogeneity**

   The instruments are uncorrelated with the error term; that is, $\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$.

endogenous variables, the instruments must provide enough information about the exogenous movements in these variables to sort out their separate effects on Y.

The general statement of the instrument exogeneity condition is that each instrument must be uncorrelated with the error term u_i. The general conditions for valid instruments are given in Key Concept 12.3.

The IV Regression Assumptions and Sampling Distribution of the TSLS Estimator

Under the IV regression assumptions, the TSLS estimator is consistent and has a sampling distribution that, in large samples, is approximately normal.

The IV regression assumptions. The IV regression assumptions are modifications of the least squares assumptions for causal inference in the multiple regression model in Key Concept 6.4.

The first IV regression assumption modifies the conditional mean assumption in Key Concept 6.4 to apply only to the included exogenous variables. Just like the second least squares assumption for the multiple regression model, the second IV regression assumption is that the draws are i.i.d., as they are if the data are collected by simple random sampling. Similarly, the third IV assumption is that large outliers are unlikely.

The IV Regression Assumptions

The variables and errors in the IV regression model in Key Concept 12.1 satisfy the following:

1. $E(u_i \mid W_{1i}, \dots, W_{ri}) = 0$;

2. $(X_{1i}, \dots, X_{ki}, W_{1i}, \dots, W_{ri}, Z_{1i}, \dots, Z_{mi}, Y_i)$ are i.i.d. draws from their joint distribution;

3. Large outliers are unlikely: The X's, W's, Z's, and Y have nonzero finite fourth moments; and

4. The two conditions for a valid instrument in Key Concept 12.3 hold.

The fourth IV regression assumption is that the two conditions for instrument validity in Key Concept 12.3 hold. The instrument relevance condition in Key Concept 12.3 subsumes the fourth least squares assumption in Key Concepts 6.4 and 6.6 (no perfect multicollinearity) by assuming that the regressors in the second-stage regression are not perfectly multicollinear. The IV regression assumptions are summarized in Key Concept 12.4.

Sampling distribution of the TSLS estimator. Under the IV regression assumptions, the TSLS estimator is consistent and normally distributed in large samples. This is shown in Section 12.1 (and Appendix 12.3) for the special case of a single endogenous regressor, a single instrument, and no included exogenous variables. Conceptually, the reasoning in Section 12.1 carries over to the general case of multiple instruments and multiple included endogenous variables. The expressions in the general case are complicated, however, and are deferred to Chapter 19.

Inference Using the TSLS Estimator

Because the sampling distribution of the TSLS estimator is normal in large samples, the general procedures for statistical inference (hypothesis tests and confidence intervals) in regression models extend to TSLS regression. For example, 95% confidence intervals are constructed as the TSLS estimator ± 1.96 standard errors. Similarly, joint hypotheses about the population values of the coefficients can be tested using the F-statistic, as described in Section 7.2.

Calculation of TSLS standard errors. There are two points to bear in mind about TSLS standard errors. First, the standard errors reported by OLS estimation of the second-stage regression are incorrect because they do not recognize that it is the second stage of a two-stage process. Specifically, the second-stage OLS standard errors fail to adjust for the second-stage regression using the predicted values of the

included endogenous variables. Formulas for standard errors that make the necessary adjustment are incorporated into (and automatically used by) TSLS regression commands in econometric software. Therefore, this issue is not a concern in practice if you use a specialized TSLS regression command.

Second, as always the error u might be heteroskedastic. It is therefore important to use heteroskedasticity-robust versions of the standard errors for precisely the same reason that it is important to use heteroskedasticity-robust standard errors for the OLS estimators of the multiple regression model.

Application to the Demand for Cigarettes

In Section 12.1, we estimated the elasticity of demand for cigarettes using data on annual consumption in 48 U.S. states in 1995 using TSLS with a single regressor (the logarithm of the real price per pack) and a single instrument (the real sales tax per pack). Income also affects demand, however, so it is part of the error term of the population regression. As discussed in Section 12.1, if the state sales tax is related to state income, it is correlated with a variable in the error term of the cigarette demand equation, which violates the instrument exogeneity condition. If so, the IV estimator in Section 12.1 is inconsistent. That is, the IV regression suffers from a version of omitted variable bias. We can solve this problem by including income in the regression.

We therefore consider an alternative specification in which the logarithm of income is included in the demand equation. In the terminology of Key Concept 12.1, the dependent variable Y is the logarithm of consumption, $\ln(Q_i^{cigarettes})$; the endogenous regressor X is the logarithm of the real after-tax price, $\ln(P_i^{cigarettes})$; the included exogenous variable W is the logarithm of the real per capita state income, $\ln(Inc_i)$; and the instrument Z is the real sales tax per pack, $SalesTax_i$. The TSLS estimates and (heteroskedasticity-robust) standard errors are

$$\widehat{\ln(Q_i^{cigarettes})} = 9.43 - 1.14\ln(P_i^{cigarettes}) + 0.21\ln(Inc_i). \qquad (12.15)$$
$$\qquad\qquad (1.26) \quad (0.37) \qquad\qquad\quad (0.31)$$

This regression uses a single instrument, $SalesTax_i$, but, in fact, another candidate instrument is available. In addition to general sales taxes, states levy special taxes that apply only to cigarettes and other tobacco products. These cigarette-specific taxes ($CigTax_i$) constitute a possible second instrumental variable. The cigarette-specific tax increases the price of cigarettes paid by the consumer, so it arguably meets the condition for instrument relevance. If it is uncorrelated with the error term in the state cigarette demand equation, it is an exogenous instrument.

With this additional instrument in hand, we now have two instrumental variables, the real sales tax per pack and the real state cigarette-specific tax per pack. With two instruments and a single endogenous regressor, the demand elasticity is overidentified; that is, the number of instruments ($SalesTax_i$ and $CigTax_i$, so $m = 2$) exceeds

the number of included endogenous variables ($P_i^{cigarettes}$, so $k = 1$). We can estimate the demand elasticity using TSLS, where the regressors in the first-stage regression are the included exogenous variable, $\ln(Inc_i)$, and both instruments.

The resulting TSLS estimate of the regression function using the two instruments $SalesTax_i$ and $CigTax_i$ is

$$\widehat{\ln(Q_i^{cigarettes})} = 9.89 - 1.28\ln(P_i^{cigarettes}) + 0.28\ln(Inc_i). \qquad (12.16)$$
$$\quad\;\; (0.96) \;\; (0.25) \qquad\qquad\quad (0.25)$$

Compare Equations (12.15) and (12.16): The standard error of the estimated price elasticity is smaller by one-third in Equation (12.16) [0.25 in Equation (12.16) versus 0.37 in Equation (12.15)]. The reason the standard error is smaller in Equation (12.16) is that this estimate uses more information than Equation (12.15): In Equation (12.15), only one instrument (the sales tax) is used, but in Equation (12.16), two instruments (the sales tax and the cigarette-specific tax) are used. Using two instruments explains more of the variation in cigarette prices than using just one, and this is reflected in smaller standard errors on the estimated demand elasticity.

Are these estimates credible? Ultimately, credibility depends on whether the set of instrumental variables—here, the two taxes—plausibly satisfies the two conditions for valid instruments. It is therefore vital that we assess whether these instruments are valid, and it is to this topic that we now turn.

12.3 Checking Instrument Validity

Whether instrumental variables regression is useful in a given application hinges on whether the instruments are valid: Invalid instruments produce meaningless results. It therefore is essential to assess whether a given set of instruments is valid in a particular application.

Assumption 1: Instrument Relevance

The role of the instrument relevance condition in IV regression is subtle. One way to think of instrument relevance is that it plays a role akin to the sample size: The more relevant are the instruments—that is, the more the variation in X is explained by the instruments—the more information is available for use in IV regression. A more relevant instrument produces a more accurate estimator, just as a larger sample size produces a more accurate estimator. Moreover, statistical inference using TSLS is predicated on the TSLS estimator having a normal sampling distribution, but according to the central limit theorem, the normal distribution is a good approximation in large—but not necessarily small—samples. If having a more relevant instrument is like having a larger sample size, this suggests, correctly, that the more relevant is the instrument, the better is the normal approximation to the sampling distribution of the TSLS estimator and its t-statistic.

Instruments that explain little of the variation in X are called **weak instruments**. In the cigarette example, the distance of the state from cigarette manufacturing plants arguably would be a weak instrument: Although a greater distance increases shipping costs (thus shifting the supply curve in and raising the equilibrium price), cigarettes are lightweight, so shipping costs are a small component of the price of cigarettes. Thus the amount of price variation explained by shipping costs, and thus distance to manufacturing plants, probably is quite small.

This section discusses why weak instruments are a problem, how to check for weak instruments, and what to do if you have weak instruments. It is assumed throughout that the instruments are exogenous.

Why weak instruments are a problem. If the instruments are weak, then the normal distribution provides a poor approximation to the sampling distribution of the TSLS estimator, even if the sample size is large. Thus there is no theoretical justification for the usual methods for performing statistical inference, even in large samples. In fact, if instruments are weak, then the TSLS estimator can be badly biased in the direction of the OLS estimator. In addition, 95% confidence intervals constructed as the TSLS estimator ± 1.96 standard errors can contain the true value of the coefficient far less than 95% of the time. In short, if instruments are weak, TSLS is no longer reliable.

To see that there is a problem with the large-sample normal approximation to the sampling distribution of the TSLS estimator, consider the special case, introduced in Section 12.1, of a single included endogenous variable, a single instrument, and no included exogenous regressor. If the instrument is valid, then $\hat{\beta}_1^{TSLS}$ is consistent because the sample covariances s_{ZY} and s_{ZX} are consistent; that is, $\hat{\beta}_1^{TSLS} = s_{ZY}/s_{ZX} \xrightarrow{p} \mathrm{cov}(Z_i, Y_i)/\mathrm{cov}(Z_i, X_i) = \beta_1$ [Equation (12.7)]. But now suppose that the instrument is not just weak but in fact is irrelevant, so that $\mathrm{cov}(Z_i, X_i) = 0$. Then $s_{ZX} \xrightarrow{p} \mathrm{cov}(Z_i, X_i) = 0$, so, taken literally, the denominator on the right-hand side of the limit $\mathrm{cov}(Z_i, Y_i)/\mathrm{cov}(Z_i, X_i)$ is 0! Clearly, the argument that $\hat{\beta}_1^{TSLS}$ is consistent breaks down when the instrument relevance condition fails. As shown in Appendix 12.4, this breakdown results in the TSLS estimator having a nonnormal sampling distribution, even if the sample size is very large. In fact, when the instrument is irrelevant, the large-sample distribution of $\hat{\beta}_1^{TSLS}$ is not the distribution of a normal random variable but rather the distribution of a *ratio* of two normal random variables! As discussed in Appendix 12.4, this ratio-of-normals distribution is centered at the large-sample value of the OLS estimator.

While this circumstance of totally irrelevant instruments might not be encountered in practice, it raises a question: How relevant must the instruments be for the normal distribution to provide a good approximation in practice? The answer to this question in the general IV model is complicated. Fortunately, however, there is a simple rule of thumb available for the most common situation in practice, the case of a single endogenous regressor.

A Rule of Thumb for Checking for Weak Instruments

The first-stage F-statistic is the F-statistic testing the hypothesis that the coefficients on the instruments $Z_{1i}, \ldots, Z_{mi}$ equal 0 in the first stage of two stage least squares. When there is a single endogenous regressor, a first-stage F-statistic less than 10 indicates that the instruments are weak, in which case the TSLS estimator is biased (even in large samples) and TSLS t-statistics and confidence intervals are unreliable.

Checking for weak instruments when there is a single endogenous regressor. One way to check for weak instruments when there is a single endogenous regressor is to compute the F-statistic testing the hypothesis that the coefficients on the instruments are all 0 in the first-stage regression of TSLS. This **first-stage F-statistic** provides a measure of the information content contained in the instruments: The more information content, the larger the expected value of the F-statistic. One simple rule of thumb is that you do not need to worry about weak instruments if the first-stage F-statistic exceeds 10. (Why 10? See Appendix 12.5.) This is summarized in Key Concept 12.5.

What do I do if I have weak instruments? If you have many instruments, some of those instruments are probably weaker than others. If you have a small number of strong instruments and many weak ones, you will be better off discarding the weakest instruments and using the most relevant subset for your TSLS analysis. Your TSLS standard errors might increase when you drop weak instruments, but keep in mind that your original standard errors were not meaningful anyway!

If, however, the coefficients are exactly identified, you cannot discard the weak instruments. Even if the coefficients are overidentified, you might not have enough strong instruments to achieve identification, so discarding some weak instruments will not help. In this case, you have two options. The first option is to find additional, stronger instruments. This is easier said than done: It requires an intimate knowledge of the problem at hand and can entail redesigning the data set and the nature of the empirical study. The second option is to proceed with your empirical analysis using the weak instruments, but employing methods other than TSLS. Although this chapter has focused on TSLS, some other methods for instrumental variable analysis are less sensitive to weak instruments than TSLS, and some of these methods are discussed in Appendix 12.5.

Assumption 2: Instrument Exogeneity

If the instruments are not exogenous, then TSLS is inconsistent: The TSLS estimator converges in probability to something other than the causal coefficient. After all, the idea of instrumental variables regression is that the instrument contains information

The First IV Regression

After he and his son Sewall derived the IV estimator (see the box "Who Invented Instrumental Variables Regression?"), Philip Wright set out to see how it worked in practice. In a letter to Sewall of March 15, 1926, Philip wrote out a table (reproduced here in part) of annual data on variables relating to U.S. production of flaxseed from 1903 through 1925. Flaxseed was grown for its oil, also called linseed oil, which was used in oil-based paint for buildings. Philip wanted to estimate the elasticity of supply. To get a percent–percent relationship, he first transformed the data to be percentage deviations from a long-term trend.

Philip then needed to make a key decision: What instrument should he use? He chose building permits on the East Coast. He reasoned that if there were more new buildings, there would be more demand for oil-based paint and thus for flaxseed, so the instrument would be relevant. He further reasoned that fluctuations in building permits on the East Coast were largely driven by broader economic conditions that had nothing to do with disturbances to flaxseed supply in a given year, so that building permits would be exogenous. Said differently, fluctuations in building permits on the East Coast were a determinant of demand but not of supply.

After laborious computations—by hand, of course—Philip obtained the IV estimate of the supply elasticity, -0.88. This elasticity has the wrong sign: It suggests that the supply curve slopes *down*. In the March 15 letter, Philip called this result "obviously absurd."

So what went wrong? Although Philip did not know it, his IV regression had a first-stage F-statistic of 1.75, far less than the rule-of-thumb cutoff of 10. As explained in the text and in Appendix 12.4, when the instrument is irrelevant, its distribution centers on the OLS estimate, which in Wright's data is -0.66. This first IV regression had a very weak instrument, and the result was biased toward OLS.

But Philip persevered. For estimating the demand elasticity, he had as an instrument rainfall in the Upper Midwest, where flaxseed was grown. More rain makes for a better harvest, so rainfall is plausibly relevant; because rainfall in the Midwest does not affect the demand for oil paint, it is plausibly exogenous. Rainfall, it turns out, has a first-stage F of 12.8 and yields an IV estimate of the demand elasticity of -0.48. This estimate indicates that the demand curve slopes down (as it should) and that demand is inelastic, which is consistent with there being no good substitute for linseed oil for paints during this period.

The First Five Observations of the First IV Regression Data Set, from Philip Wright's Letter to Sewall Wright of March 15, 1926.

The first two data columns are the real price and quantity ("output") of flaxseed. The "B" variables— acreage planted, yield, rainfall in the Upper Midwest, and the ratio of flaxseed yield that year to spring wheat yield the previous year—shift supply but not demand, so they are potential instruments for the demand elasticity. The "A" variable—building permits on the East Coast—shifts demand but not supply, so it is a potential instrument for the supply elasticity.

about variation in X_i that is unrelated to the error term u_i. If, in fact, the instrument is not exogenous, it cannot pinpoint this exogenous variation in X_i, and it stands to reason that IV regression fails to provide a consistent estimator. The math behind this argument is summarized in Appendix 12.4.

Can you statistically test the assumption that the instruments are exogenous? Yes and no. On the one hand, it is not possible to test the hypothesis that the instruments are exogenous when the coefficients are exactly identified. On the other hand, if the coefficients are overidentified, it is possible to test the overidentifying restrictions—that is, to test the hypothesis that the "extra" instruments are exogenous under the maintained assumption that there are enough valid instruments to identify the coefficients of interest.

First consider the case that the coefficients are exactly identified, so you have as many instruments as endogenous regressors. Then it is impossible to develop a statistical test of the hypothesis that the instruments are, in fact, exogenous. That is, empirical evidence cannot be brought to bear on the question of whether these instruments satisfy the exogeneity restriction. In this case, the only way to assess whether the instruments are exogenous is to draw on expert opinion and your personal knowledge of the empirical problem at hand. For example, Philip Wright's knowledge of agricultural supply and demand led him to suggest that below-average rainfall would plausibly shift the supply curve for fats and oils but would not directly shift the demand curve.

Assessing whether the instruments are exogenous *necessarily* requires making an expert judgment based on personal knowledge of the application. If, however, there are more instruments than endogenous regressors, then there is a statistical tool that can be helpful in this process: the so-called test of overidentifying restrictions.

The overidentifying restrictions test. Suppose you have a single endogenous regressor and two instruments. Then you could compute two different TSLS estimators: one using the first instrument and the other using the second. These two estimators will not be the same because of sampling variation, but if both instruments are exogenous, then they will tend to be close to each other. But what if these two instruments produce very different estimates? You might sensibly conclude that there is something wrong with one or the other of the instruments or with both. That is, it would be reasonable to conclude that one or the other or both of the instruments are not exogenous.

The **test of overidentifying restrictions** implicitly makes this comparison. We say implicitly because the test is carried out without actually computing all of the different possible IV estimates. Here is the idea. Exogeneity of the instruments means that they are uncorrelated with u_i. This suggests that the instruments should be approximately uncorrelated with $\hat{u}_i^{TSLS}$, where $\hat{u}_i^{TSLS} = Y_i - (\hat{\beta}_0^{TSLS} + \hat{\beta}_1^{TSLS} X_{1i} + \cdots + \hat{\beta}_{k+r}^{TSLS} W_{ri})$

The Overidentifying Restrictions Test (The *J*-Statistic)

Let $\hat{u}_i^{TSLS}$ be the residuals from TSLS estimation of Equation (12.12). Use OLS to estimate the regression coefficients in

$$\hat{u}_i^{TSLS} = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \cdots + \delta_{m+r} W_{ri} + e_i, \quad (12.17)$$

where e_i is the regression error term. Let F denote the homoskedasticity-only F-statistic testing the hypothesis that $\delta_1 = \cdots = \delta_m = 0$. The overidentifying restrictions test statistic is $J = mF$. Under the null hypothesis that all the instruments are exogenous, if e_i is homoskedastic, in large samples J is distributed χ^2_{m-k}, where $m - k$ is the *degree of overidentification*—that is, the number of instruments minus the number of endogenous regressors.

is the residual from the estimated TSLS regression using all the instruments (approximately rather than exactly because of sampling variation). (Note that these residuals are constructed using the true X's rather than their first-stage predicted values.) Accordingly, if the instruments are, in fact, exogenous, then the coefficients on the instruments in a regression of $\hat{u}_i^{TSLS}$ on the instruments and the included exogenous variables should all be 0, and this hypothesis can be tested.

This method for computing the overidentifying restrictions test is summarized in Key Concept 12.6. This statistic is computed using the homoskedasticity-only F-statistic. The test statistic is commonly called the *J*-statistic and is computed as $J = mF$.

In large samples, if the instruments are not weak and the errors are homoskedastic, then, under the null hypothesis that the instruments are exogenous, the *J*-statistic has a chi-squared distribution with $m - k$ degrees of freedom (χ^2_{m-k}). It is important to remember that even though the number of restrictions being tested is m, the degrees of freedom of the asymptotic distribution of the *J*-statistic is $m - k$. The reason is that it is possible to test only the *over*identifying restrictions, of which there are $m - k$. The modification of the *J*-statistic for heteroskedastic errors is given in Section 19.7.

The easiest way to see that you cannot test the exogeneity of the regressors when the coefficients are exactly identified $(m = k)$ is to consider the case of a single included endogenous variable $(k = 1)$. If there are two instruments, then you can compute two TSLS estimators, one for each instrument, and you can compare them to see if they are close. But if you have only one instrument, then you can compute only one TSLS estimator, and you have nothing to which to compare it. In fact, if the coefficients are exactly identified, so that $m = k$, then the overidentifying test statistic J is exactly 0.

12.4 Application to the Demand for Cigarettes[1]

Our attempt to estimate the elasticity of demand for cigarettes left off with the TSLS estimates summarized in Equation (12.16), in which income was an included exogenous variable and there were two instruments, the general sales tax and the cigarette-specific tax. We can now undertake a more thorough evaluation of these instruments.

As in Section 12.1, it makes sense that the two instruments are relevant because taxes are a big part of the after-tax price of cigarettes, and shortly we will look at this empirically. First, however, we focus on the difficult question of whether the two tax variables are plausibly exogenous.

The first step in assessing whether an instrument is exogenous is to think through the arguments for why it may or may not be. This requires thinking about which factors account for the error term in the cigarette demand equation and whether these factors are plausibly related to the instruments.

Why do some states have higher per capita cigarette consumption than others? One reason might be variation in incomes across states, but state income is included in Equation (12.16), so this is not part of the error term. Another reason is that there are historical factors influencing demand. For example, states that grow tobacco have higher rates of smoking than most other states. Could this factor be related to taxes? Quite possibly: If tobacco farming and cigarette production are important industries in a state, then these industries could exert influence to keep cigarette-specific taxes low. This suggests that an omitted factor in cigarette demand—whether the state grows tobacco and produces cigarettes—could be correlated with cigarette-specific taxes.

One solution to this possible correlation between the error term and the instrument would be to include information on the size of the tobacco and cigarette industry in the state; this is the approach we took when we included income as a regressor in the demand equation. But because we have panel data on cigarette consumption, a different approach is available that does not require this information. As discussed in Chapter 10, panel data make it possible to eliminate the influence of variables that vary across entities (states) but do not change over time, such as the historical circumstances that lead to a large tobacco and cigarette industry in a state. Two methods for doing this were given in Chapter 10: constructing data on *changes* in the variables between two different time periods and using fixed effects regression. To keep the analysis here as simple as possible, we adopt the former approach and perform regressions of the type described in Section 10.2, based on the changes in the variables between two different years.

The time span between the two different years influences how the estimated elasticities are to be interpreted. Because cigarettes are addictive, changes in price will take some time to alter behavior. At first, an increase in the price of cigarettes might have little effect on demand. Over time, however, the price increase might contribute

---

[1]This section assumes knowledge of the material in Sections 10.1 and 10.2 on panel data with $T = 2$ time periods.

The Externalities of Smoking

Smoking imposes costs that are not fully borne by the smoker; that is, it generates externalities. One economic justification for taxing cigarettes therefore is to "internalize" these externalities. In theory, the tax on a pack of cigarettes should equal the dollar value of the externalities created by smoking that pack. But what, precisely, are the externalities of smoking, measured in dollars per pack?

Several studies have used econometric methods to estimate the externalities of smoking. The negative externalities—costs—borne by others include medical costs paid by the government to care for ill smokers, health care costs of nonsmokers associated with secondhand smoke, and fires caused by cigarettes.

But, from a purely economic point of view, smoking also has *positive* externalities, or benefits. The biggest economic benefit of smoking is that smokers tend to pay much more in Social Security (public pension) taxes than they ever get back. There are also large savings in nursing home expenditures on the very old—smokers tend not to live that long. Because the negative externalities of smoking occur while the smoker is alive but the positive ones accrue after

death, the net present value of the per-pack externalities (the value of the net costs per pack, discounted to the present) depends on the discount rate.

The studies do not agree on a specific dollar value of the net externalities. Some suggest that the net externalities, properly discounted, are quite small, less than current taxes. In fact, the most extreme estimates suggest that the net externalities are *positive*, so smoking should be subsidized! Other studies, which incorporate costs that are probably important but difficult to quantify (such as caring for babies who are unhealthy because their mothers smoke), suggest that externalities might be $1 per pack, possibly even more. But all the studies agree that, by tending to die in late middle age, smokers pay far more in taxes than they ever get back in their brief retirement.[1]

_____

[1] An early calculation of the externalities of smoking was reported by Willard G. Manning et al. (1989). A calculation suggesting that health care costs would go *up* if everyone stopped smoking is presented in Barendregt, Bonneux, and van der Maas (1997). Other studies of the externalities of smoking are reviewed by Chaloupka and Warner (2000).

to some smokers' desire to quit, and, importantly, it could discourage nonsmokers from taking up the habit. Thus the response of demand to a price increase could be small in the short run but large in the long run. Said differently, for an addictive product like cigarettes, demand might be inelastic in the short run—that is, it might have a short-run elasticity near 0—but it might be more elastic in the long run.

In this analysis, we focus on estimating the long-run price elasticity. We do this by considering quantity and price changes that occur over 10-year periods. Specifically, in the regressions considered here, the 10-year change in log quantity, $\ln(Q_{i,1995}^{cigarettes}) - \ln(Q_{i,1985}^{cigarettes})$, is regressed against the 10-year change in log price, $\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})$, and the 10-year change in log income, $\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$. Two instruments are used: the change in the sales tax over 10 years, $SalesTax_{i,1995} - SalesTax_{i,1985}$, and the change in the cigarette-specific tax over 10 years, $CigTax_{i,1995} - CigTax_{i,1985}$.

TABLE 12.1	**Two Stage Least Squares Estimates of the Demand for Cigarettes Using Panel Data for 48 U.S. States**		

Dependent variable: $\ln(Q_{i,1995}^{cigarettes}) - \ln(Q_{i,1985}^{cigarettes})$

Regressor	**(1)**	**(2)**	**(3)**
$\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})$	−0.94 (0.21) [−1.36, −0.52]	−1.34 (0.23) [−1.80, −0.88]	−1.20 (0.20) [−1.60, −0.81]
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0.53 (0.34) [−0.16, 1.21]	0.43 (0.30) [−0.16, 1.02]	0.46 (0.31) [−0.16, 1.09]
Intercept	−0.12 (0.07)	−0.02 (0.07)	−0.05 (0.06)
Instrumental variable(s)	Sales tax	Cigarette-specific tax	Both sales tax and cigarette-specific tax
First-stage F-statistic	33.7	107.2	88.6
Overidentifying restrictions J-test and p-value	—	—	4.93 (0.026)

These regressions were estimated using data for 48 U.S. states (48 observations on the 10-year differences). The data are described in Appendix 12.1. The J-test of overidentifying restrictions is described in Key Concept 12.6 (its p-value is given in parentheses), and the first-stage F-statistic is described in Key Concept 12.5. Heteroskedasticity-robust standard errors are given in parentheses beneath coefficients, and 95% confidence intervals are given in brackets.

The results are presented in Table 12.1. As usual, each column in the table presents the results of a different regression. All regressions have the same regressors, and all coefficients are estimated using TSLS; the only difference among the three regressions is the set of instruments used. In column (1), the only instrument is the sales tax; in column (2), the only instrument is the cigarette-specific tax; and in column (3), both taxes are used as instruments.

In IV regression, the reliability of the coefficient estimates hinges on the validity of the instruments, so the first things to look at in Table 12.1 are the diagnostic statistics assessing the validity of the instruments.

First, are the instruments relevant? We need to look at the first-stage F-statistics. The first-stage regression in column (1) is

$$\widehat{\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})} = 0.53 - 0.22[\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})]$$
$$(0.03) \ \ (0.22)$$

$$+ \ 0.0255(SalesTax_{i,1995} - SalesTax_{i,1985}). \quad (12.18)$$
$$(0.0044)$$

Because there is only one instrument in this regression, the first-stage F-statistic is the square of the t-statistic testing that the coefficient on the instrumental variable, $SalesTax_{i,1995} - SalesTax_{i,1985}$, is 0; this is $F = t^2 = (0.0255/0.0044)^2 = 33.7$. For the

regressions in columns (2) and (3), the first-stage F-statistics are 107.2 and 88.6, so in all three cases the first-stage F-statistics exceed 10. We conclude that the instruments are not weak, so we can rely on the standard methods for statistical inference (hypothesis tests and confidence intervals) using the TSLS coefficients and standard errors.

Second, are the instruments exogenous? Because the regressions in columns (1) and (2) each have a single instrument and a single included endogenous regressor, the coefficients in those regressions are exactly identified. Thus we cannot deploy the J-test in either of those regressions. The regression in column (3), however, is overidentified because there are two instruments and a single included endogenous regressor, so there is one $(m - k = 2 - 1 = 1)$ overidentifying restriction. The J-statistic is 4.93; this has a χ_1^2 distribution, so the 5% critical value is 3.84 (Appendix Table 3) and the null hypothesis that both the instruments are exogenous is rejected at the 5% significance level (this deduction also can be made directly from the p-value of 0.026, reported in the table).

The reason the J-statistic rejects the null hypothesis that both instruments are exogenous is that the two instruments produce rather different estimated coefficients. When the only instrument is the sales tax [column (1)], the estimated price elasticity is -0.94, but when the only instrument is the cigarette-specific tax, the estimated price elasticity is -1.34. Recall the basic idea of the J-statistic: If both instruments are exogenous, then the two TSLS estimators using the individual instruments are consistent and differ from each other only because of random sampling variation. If, however, one of the instruments is exogenous and one is not, then the estimator based on the endogenous instrument is inconsistent, which is detected by the J-statistic. In this application, the difference between the two estimated price elasticities is sufficiently large that it is unlikely to be the result of pure sampling variation, so the J-statistic rejects the null hypothesis that both the instruments are exogenous.

The J-statistic rejection means that the regression in column (3) is based on invalid instruments (the instrument exogeneity condition fails). What does this imply about the estimates in columns (1) and (2)? The J-statistic rejection says that at least one of the instruments is endogenous, so there are three logical possibilities: The sales tax is exogenous but the cigarette-specific tax is not, in which case the column (1) regression is reliable; the cigarette-specific tax is exogenous but the sales tax is not, so the column (2) regression is reliable; or neither tax is exogenous, so neither regression is reliable. The statistical evidence cannot tell us which possibility is correct, so we must use our judgment.

We think that the case for the exogeneity of the general sales tax is stronger than that for the cigarette-specific tax because the political process can link changes in the cigarette-specific tax to changes in the cigarette market and smoking policy. For example, if smoking decreases in a state because it falls out of fashion, there will be fewer smokers and a weakened lobby against cigarette-specific tax increases, which in turn could lead to higher cigarette-specific taxes. Thus changes in tastes (which are part of u) could be correlated with changes in cigarette-specific taxes (the instrument). This suggests discounting the IV estimates that use the cigarette-only tax as

an instrument and adopting the price elasticity estimated using the general sales tax as an instrument, -0.94.

The estimate of -0.94 indicates that cigarette consumption is somewhat elastic: An increase in price of 1% leads to a decrease in consumption of 0.94%. This may seem surprising for an addictive product like cigarettes. But remember that this elasticity is computed using changes over a 10-year period, so it is a long-run elasticity. This estimate suggests that increased taxes can make a substantial dent in cigarette consumption, at least in the long run.

When the elasticity is estimated using 5-year changes from 1985 to 1990 rather than the 10-year changes reported in Table 12.1, the elasticity (estimated with the general sales tax as the instrument) is -0.79; for changes from 1990 to 1995, the elasticity is -0.68. These estimates suggest that demand is less elastic over horizons of 5 years than over 10 years. This finding of greater price elasticity at longer horizons is consistent with the large body of research on cigarette demand. Demand elasticity estimates in that literature typically fall in the range -0.3 to -0.5, but these are mainly short-run elasticities; some studies suggest that the long-run elasticity could be perhaps twice the short-run elasticity.[2]

12.5   Where Do Valid Instruments Come From?

In practice, the most difficult aspect of IV estimation is finding instruments that are both relevant and exogenous. There are two main approaches, which reflect two different perspectives on econometric and statistical modeling.

The first approach is to use economic theory to suggest instruments. For example, Philip Wright's understanding of the economics of agricultural markets led him to look for an instrument that shifted the supply curve but not the demand curve; this in turn led him to consider weather conditions in agricultural regions. One area where this approach has been particularly successful is the field of financial economics. Some economic models of investor behavior involve statements about how investors forecast, which then imply sets of variables that are uncorrelated with the error term. Those models sometimes are nonlinear in the data and in the parameters, in which case the IV estimators discussed in this chapter cannot be used. An extension of IV methods to nonlinear models, called generalized method of moments estimation, is used instead. Economic theories are, however, abstractions that often do not take into account the nuances and details necessary for analyzing a particular data set. Thus this approach does not always work.

The second approach to constructing instruments is to look for some exogenous source of variation in X arising from what is, in effect, a random phenomenon that

---

[2]A sobering economic study by Adda and Cornaglia (2006) suggests that smokers compensate for higher taxes by smoking more intensively, thus extracting more nicotine per cigarette. If you are interested in learning more about the economics of smoking, see Chaloupka and Warner (2000), Gruber (2001), and Carpenter and Cook (2008).

induces shifts in the endogenous regressor. For example, in our hypothetical example in Section 12.1, earthquake damage increased average class size in some school districts, and this variation in class size was unrelated to potential omitted variables that affect student achievement. This approach typically requires knowledge of the problem being studied and careful attention to the details of the data, and it is best explained through examples.

Three Examples

We now turn to three empirical applications of IV regression that illustrate how different researchers used their expert knowledge of their empirical problem to find instrumental variables.

Does putting criminals in jail reduce crime? This is a question only an economist would ask. After all, a criminal cannot commit a crime outside jail while in prison, and that some criminals are caught and jailed serves to deter others. But the magnitude of the combined effect—the change in the crime rate associated with a 1% increase in the prison population—is an empirical question.

One strategy for estimating this effect is to regress crime rates (crimes per 100,000 members of the general population) against incarceration rates (prisoners per 100,000 members of the general population), using annual data at a suitable level of jurisdiction (for example, U.S. states). This regression could include some control variables measuring economic conditions (crime increases when general economic conditions worsen), demographics (youths commit more crimes than the elderly), and so forth. There is, however, a serious potential for simultaneous causality bias that undermines such an analysis: If the crime rate goes up and the police do their job, there will be more prisoners. On the one hand, increased incarceration reduces the crime rate; on the other hand, an increased crime rate increases incarceration. As in the butter example in Figure 12.1, because of this simultaneous causality, an OLS regression of the crime rate on the incarceration rate will estimate some complicated combination of these two effects. This problem cannot be solved by finding better control variables.

This simultaneous causality bias, however, can be eliminated by finding a suitable instrumental variable and using TSLS. The instrument must be correlated with the incarceration rate (it must be relevant), but it must also be uncorrelated with the error term in the crime rate equation of interest (it must be exogenous). That is, it must affect the incarceration rate but be unrelated to any of the unobserved factors that determine the crime rate.

Where does one find something that affects incarceration but has no direct effect on the crime rate? One place is exogenous variation in the capacity of existing prisons. Because it takes time to build a prison, short-term capacity restrictions can force states to release prisoners prematurely or otherwise reduce incarceration rates. Using this reasoning, Levitt (1996) suggested that lawsuits aimed at reducing prison

overcrowding could serve as an instrumental variable, and he implemented this idea using panel data for the U.S. states from 1972 to 1993.

Are variables measuring overcrowding litigation valid instruments? Although Levitt did not report first-stage *F*-statistics, the prison overcrowding litigation slowed the growth of prisoner incarcerations in his data, suggesting that this instrument is relevant. To the extent that overcrowding litigation is induced by prison conditions but not by the crime rate or its determinants, this instrument is exogenous. Because Levitt breaks down overcrowding legislation into several types and thus has several instruments, he is able to test the overidentifying restrictions and fails to reject them using the *J*-statistic, which bolsters the case that his instruments are valid.

Using these instruments and TSLS, Levitt estimated the effect on the crime rate of incarceration to be substantial. This estimated effect was three times larger than the effect estimated using OLS, suggesting that OLS suffered from large simultaneous causality bias.

Does cutting class sizes increase test scores? As we saw in the empirical analysis of Part II, schools with small classes tend to be wealthier, and their students have access to enhanced learning opportunities both in and out of the classroom. In Part II, we used multiple regression to tackle the threat of omitted variables bias by controlling for various measures of student affluence, ability to speak English, and so forth. Still, a skeptic could wonder whether we did enough: If we left out something important, our estimates of the class size effect would still be biased.

This potential omitted variables bias could be addressed by including the right control variables, but if these data are unavailable (some, like outside learning opportunities, are hard to measure), then an alternative approach is to use IV regression. This regression requires an instrumental variable correlated with class size (relevance) but uncorrelated with the omitted determinants of test performance that make up the error term, such as parental interest in learning, learning opportunities outside the classroom, quality of the teachers and school facilities, and so forth (exogeneity).

Where does one look for an instrument that induces random, exogenous variation in class size, but is unrelated to the other determinants of test performance? Hoxby (2000) suggested biology. Because of random fluctuations in timings of births, the size of the incoming kindergarten class varies from one year to the next. Although the actual number of children entering kindergarten might be endogenous (recent news about the school might influence whether parents send a child to a private school), she argued that the *potential* number of children entering kindergarten—the number of 4-year-olds in the district—is mainly a matter of random fluctuations in the birth dates of children.

Is potential enrollment a valid instrument? Whether it is exogenous depends on whether it is correlated with unobserved determinants of test performance. Surely biological fluctuations in potential enrollment are exogenous, but potential enrollment also fluctuates because parents with young children choose to move into an

improving school district and out of one in trouble. If so, an increase in potential enrollment could be correlated with unobserved factors such as the quality of school management, rendering this instrument invalid. Hoxby addressed this problem by reasoning that growth or decline in the potential student pool for this reason would occur smoothly over several years, whereas random fluctuations in birth dates would produce short-term "spikes" in potential enrollment. Thus she used as her instrument not potential enrollment, but the deviation of potential enrollment from its long-term trend. These deviations satisfy the criterion for instrument relevance (the first-stage F-statistics all exceed 100). She makes a good case that this instrument is exogenous, but, as in all IV analysis, the credibility of this assumption is ultimately a matter of judgment.

Hoxby implemented this strategy using detailed panel data on elementary schools in Connecticut in the 1980s and 1990s. The panel data set permitted her to include school fixed effects, which, in addition to the instrumental variables strategy, attack the problem of omitted variables bias at the school level. Her TSLS estimates suggested that the effect on test scores of class size is small; most of her estimates were statistically insignificantly different from 0.

Does aggressive treatment of heart attacks prolong lives? Aggressive treatments for victims of heart attacks (technically, acute myocardial infarctions, or AMIs) hold the potential for saving lives. Before a new medical procedure—in this example, cardiac catheterization[3]—is approved for general use, it goes through clinical trials, a series of randomized controlled experiments designed to measure its effects and side effects. But strong performance in a clinical trial is one thing; actual performance in the real world is another.

A natural starting point for estimating the real-world effect of cardiac catheterization is to compare patients who received the treatment to those who did not. This leads to regressing the length of survival of the patient against the binary treatment variable (whether the patient received cardiac catheterization) and other control variables that affect mortality (age, weight, other measured health conditions, and so forth). The population coefficient on the indicator variable is the increment to the patient's life expectancy provided by the treatment. Unfortunately, the OLS estimator is subject to bias: Cardiac catheterization does not "just happen" to a patient randomly; rather, it is performed because the doctor and patient decide that it might be effective. If their decision is based in part on unobserved factors relevant to health outcomes not in the data set, the treatment decision will be correlated with the regression error term. If the healthiest patients are the ones who receive the treatment, the OLS estimator will be biased (treatment is correlated with an omitted variable), and the treatment will appear more effective than it really is.

---

[3]Cardiac catheterization is a procedure in which a catheter, or tube, is inserted into a blood vessel and guided all the way to the heart to obtain information about the heart and coronary arteries.

This potential bias can be eliminated by IV regression using a valid instrumental variable. The instrument must be correlated with treatment (must be relevant) but must be uncorrelated with the omitted health factors that affect survival (must be exogenous).

Where does one look for something that affects treatment but does not affect the health outcome other than through its effect on treatment? McClellan, McNeil, and Newhouse (1994) suggested geography. Most hospitals in their data set did not offer cardiac catheterization, so many patients were closer to "regular" hospitals that did not offer this treatment than to cardiac catheterization hospitals. McClellan, McNeil, and Newhouse therefore used as an instrumental variable the difference between the distance from the AMI patient's home to the nearest cardiac catheterization hospital and the distance to the nearest hospital of any sort; this distance is 0 if the nearest hospital is a cardiac catheterization hospital, and otherwise it is positive. If this relative distance affects the probability of receiving this treatment, then it is relevant. If it is distributed randomly across AMI victims, then it is exogenous.

Is relative distance to the nearest cardiac catheterization hospital a valid instrument? McClellan, McNeil, and Newhouse do not report first-stage F-statistics, but they do provide other empirical evidence that it is not weak. Is this distance measure exogenous? They make two arguments. First, they draw on their medical expertise and knowledge of the health care system to argue that distance to a hospital is plausibly uncorrelated with any of the unobservable variables that determine AMI outcomes. Second, they have data on some of the additional variables that affect AMI outcomes, such as the weight of the patient, and in their sample, distance is uncorrelated with these *observable* determinants of survival; this, they argue, makes it more credible that distance is uncorrelated with the *unobservable* determinants in the error term as well.

Using 205,021 observations on Americans aged at least 64 who had an AMI in 1987, McClellan, McNeil, and Newhouse reached a striking conclusion: Their TSLS estimates suggest that cardiac catheterization has a small, possibly 0, effect on health outcomes; that is, cardiac catheterization does not substantially prolong life. In contrast, the OLS estimates suggest a large positive effect. They interpret this difference as evidence of bias in the OLS estimates.

McClellan, McNeil, and Newhouse's IV method has an interesting interpretation. The OLS analysis used actual treatment as the regressor, but because actual treatment is itself the outcome of a decision by patient and doctor, they argue that the actual treatment is correlated with the error term. Instead, TSLS uses *predicted* treatment, where the variation in predicted treatment arises because of variation in the instrumental variable: Patients closer to a cardiac catheterization hospital are more likely to receive this treatment.

This interpretation has two implications. First, the IV regression actually estimates the effect of the treatment not on a "typical" randomly selected patient but rather on patients for whom distance is an important consideration in the treatment

decision. The effect on those patients might differ from the effect on a typical patient, which provides one explanation of the greater estimated effectiveness of the treatment in clinical trials than in McClellan, McNeil, and Newhouse's IV study. Second, it suggests a general strategy for finding instruments in this type of setting: Find an instrument that affects the probability of treatment, but does so for reasons that are unrelated to the outcome except through their effect on the likelihood of treatment. Both these implications have applicability to experimental and "quasi-experimental" studies, the topic of Chapter 13.

12.6 Conclusion

From the humble start of estimating how much less butter people will buy if its price rises, IV methods have evolved into a general approach for estimating regressions when one or more variables are correlated with the error term. Instrumental variables regression uses the instruments to isolate variation in the endogenous regressors that is uncorrelated with the error in the regression of interest; this is the first stage of two stage least squares. This in turn permits estimation of the effect of interest in the second stage of two stage least squares.

Successful IV regression requires valid instruments—that is, instruments that are both relevant (not weak) and exogenous. If the instruments are weak, then the TSLS estimator can be biased, even in large samples, and statistical inferences based on TSLS t-statistics and confidence intervals can be misleading. Fortunately, when there is a single endogenous regressor, it is possible to check for weak instruments simply by checking the first-stage F-statistic.

If the instruments are not exogenous—that is, if one or more instruments are correlated with the error term—the TSLS estimator is inconsistent. If there are more instruments than endogenous regressors, instrument exogeneity can be examined by using the J-statistic to test the overidentifying restrictions. However, the core assumption—that there are at least as many exogenous instruments as there are endogenous regressors—cannot be tested. It is therefore incumbent on both the empirical analyst and the critical reader to use their own understanding of the empirical application to evaluate whether this assumption is reasonable.

The interpretation of IV regression as a way to exploit known exogenous variation in the endogenous regressor can be used to guide the search for potential instrumental variables in a particular application. This interpretation underlies much of the empirical analysis in the area that goes under the broad heading of program evaluation, in which experiments or quasi-experiments are used to estimate the effect of programs, policies, or other interventions on some outcome measure. A variety of additional issues arises in those applications—for example, the interpretation of IV results when, as in the cardiac catheterization example, different "patients" might have different responses to the same "treatment." These and other aspects of empirical program evaluation are taken up in Chapter 13.

Summary

1. Instrumental variables regression is a way to estimate causal coefficients when one or more regressors are correlated with the error term.
2. Endogenous variables are correlated with the error term in the equation of interest; exogenous variables are uncorrelated with this error term.
3. For an instrument to be valid, it must be (1) correlated with the included endogenous variable and (2) exogenous.
4. IV regression requires at least as many instruments as included endogenous variables.
5. The TSLS estimator has two stages. First, the included endogenous variables are regressed against the included exogenous variables and the instruments. Second, the dependent variable is regressed against the included exogenous variables and the predicted values of the included endogenous variables from the first-stage regression(s).
6. Weak instruments (instruments that are nearly uncorrelated with the included endogenous variables) make the TSLS estimator biased and TSLS confidence intervals and hypothesis tests unreliable.
7. If an instrument is not exogenous, the TSLS estimator is inconsistent.

Key Terms

instrumental variables (IV) regression (385)

instrumental variable (instrument) (385)

endogenous variable (386)

exogenous variable (386)

instrument relevance condition (387)

instrument exogeneity condition (387)

two stage least squares (387)

included exogenous variables (395)

exactly identified (396)

overidentified (396)

underidentified (396)

reduced form (397)

first-stage regression (398)

second-stage regression (398)

weak instruments (403)

first-stage F-statistic (404)

test of overidentifying restrictions (406)

MyLab Economics Can Help You Get a Better Grade

MyLab Economics   If your exam were tomorrow, would you be ready? For each chapter, **MyLab Economics** Practice Tests and Study Plan help you prepare for your exams. You can also find the Exercises and all Review the Concepts Questions available now in **MyLab Economics**. To see how it works, turn to the **MyLab Economics** spread on the inside front cover of this text and then go to **www.pearson.com/mylab/economics**.

For additional Empirical Exercises and Data Sets, log on to the Companion Website at **www.pearsonhighered.com/stock_watson**.

Review the Concepts

12.1 In the demand curve model of Equation (12.3), is $\ln(P_i^{butter})$ positively or negatively correlated with the error, u_i? If β_1 is estimated by OLS, would you expect the estimated value to be larger or smaller than the true value of β_1? Explain.

12.2 In the study of cigarette demand in this chapter, suppose we used as an instrument the number of trees per capita in the state. Is this instrument relevant? Is it exogenous? Is it a valid instrument?

12.3 In his study of the effect of incarceration on crime rates, suppose Levitt had used the number of lawyers per capita as an instrument. Would this instrument be relevant? Would it be exogenous? Would it be a valid instrument?

12.4 In their study of the effectiveness of cardiac catheterization, McClellan, McNeil, and Newhouse (1994) used as an instrument the difference in distances to cardiac catheterization and regular hospitals. How could you determine whether this instrument is relevant? How could you determine whether this instrument is exogenous?

Exercises

12.1 This question refers to the panel data IV regressions summarized in Table 12.1.

   a. Suppose the federal government is considering a new tax on cigarettes that is estimated to increase the retail price by $0.50 per pack. If the current price per pack is $7.50, use the IV regression in column (1) to predict the change in demand. Construct a 95% confidence interval for the change in demand.

   b. Suppose the United States enters a recession and income falls by 2%. Use the IV regression in column (1) to predict the change in demand.

   c. Suppose the recession lasts less than one year. Do you think that the IV regression in column (1) will provide a reliable answer to the question in (b)? Why or why not?

   d. Suppose the F-statistic in column (1) were 3.7 instead of 33.7. Would the IV regression provide a reliable answer to the question posed in (a)? Why or why not?

12.2 Consider the regression model with a single regressor: $Y_i = \beta_0 + \beta_1 X_i + u_i$. Suppose the least squares assumptions in Key Concept 4.3 are satisfied.

   a. Show that X_i is a valid instrument. That is, show that Key Concept 12.3 is satisfied with $Z_i = X_i$.

b. Show that the IV regression assumptions in Key Concept 12.4 are satisfied with this choice of Z_i.

c. Show that the IV estimator constructed using $Z_i = X_i$ is identical to the OLS estimator.

12.3 A classmate is interested in estimating the variance of the error term in Equation (12.1).

a. Suppose she uses the estimator from the second-stage regression of TSLS: $\hat{\sigma}_a^2 = \frac{1}{n-2}\Sigma_{i=1}^{n}(Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS}\hat{X}_i)^2$, where $\hat{X}_i$ is the fitted value from the first-stage regression. Is this estimator consistent? (For the purposes of this question, suppose that the sample is very large and the TSLS estimators are essentially identical to β_0 and β_1.)

b. Is $\hat{\sigma}_b^2 = \frac{1}{n-2}\Sigma_{i=1}^{n}(Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS}X_i)^2$ consistent?

12.4 Consider TSLS estimation with a single included endogenous variable and a single instrument. Then the predicted value from the first-stage regression is $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$. Use the definition of the sample variance and covariance to show that $s_{\hat{X}Y} = \hat{\pi}_1 s_{ZY}$ and $s_{\hat{X}}^2 = \hat{\pi}_1^2 s_Z^2$. Use this result to fill in the steps of the derivation of Equation (12.4) in Appendix 12.2.

12.5 Consider the IV regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i,$$

where X_i is correlated with u_i and Z_i is an instrument. Suppose that the first three assumptions in Key Concept 12.4 are satisfied. Which IV assumption is not satisfied when

a. Z_i is independent of (Y_i, X_i, W_i)?

b. $Z_i = W_i$?

c. $W_i = 1$ for all i?

d. $Z_i = X_i$?

12.6 In an IV regression model with one regressor, X_i, and one instrument, Z_i, the regression of X_i onto Z_i has $R^2 = 0.05$ and $n = 100$. Is Z_i a strong instrument? [*Hint:* See Equation (7.14).] Would your answer change if $R^2 = 0.05$ and $n = 500$?

12.7 In an IV regression model with one regressor, X_i, and two instruments, Z_{1i} and Z_{2i}, the value of the *J*-statistic is $J = 18.2$.

a. Does this suggest that $E(u_i | Z_{1i}, Z_{2i}) \neq 0$? Explain.

b. Does this suggest that $E(u_i | Z_{1i}) \neq 0$? Explain.

12.8 Consider a product market with a supply function $Q_i^s = \beta_0 + \beta_1 P_i + u_i^s$, a demand function $Q_i^d = \gamma_0 + u_i^d$, and a market equilibrium condition

$Q_i^s = Q_i^d$, where u_i^s and u_i^s are mutually independent i.i.d. random variables, both with a mean of 0.

    a. Show that P_i and u_i^s are correlated.

    b. Show that the OLS estimator of β_1 is inconsistent.

    c. How would you estimate β_0, β_1, and γ_0?

12.9 A researcher is interested in the effect of military service on human capital. He collects data from a random sample of 4000 workers aged 40 and runs the OLS regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, where Y_i is a worker's annual earnings and X_i is a binary variable that is equal to 1 if the person served in the military and is equal to 0 otherwise.

    a. Explain why the OLS estimates are likely to be unreliable. (*Hint:* Which variables are omitted from the regression? Are they correlated with military service?)

    b. During the Vietnam War, there was a draft in which priority for the draft was determined by a national lottery. (The days of the year were randomly reordered 1 through 365. Those with birth dates ordered first were drafted before those with birth dates ordered second, and so forth.) Explain how the lottery might be used as an instrument to estimate the effect of military service on earnings. (For more about this issue, see Joshua D. Angrist (1990).)

12.10 Consider the IV regression model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$, where Z_i is an instrument. Suppose data on W_i are not available and the model is estimated omitting W_i from the regression.

    a. Suppose Z_i and W_i are uncorrelated. Is the IV estimator consistent?

    b. Suppose Z_i and W_i are correlated. Is the IV estimator consistent?

Empirical Exercises

E12.1 How does fertility affect labor supply? That is, how much does a woman's labor supply fall when she has an additional child? In this exercise, you will estimate this effect using data for married women from the 1980 U.S. Census.[4] The data are available on the text website, **http://www.pearsonhighered.com/ stock_watson**, in the file **Fertility** and described in the file **Fertility_Description**. The data set contains information on married women aged 21–35 with two or more children.

---

[4]These data were provided by Professor William Evans of the University of Maryland and were used in his paper with Joshua Angrist, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 1998, 88(3): 450–477.

  a. Regress *weeksworked* on the indicator variable *morekids*, using OLS. On average, do women with more than two children work less than women with two children? How much less?

  b. Explain why the OLS regression estimated in (a) is inappropriate for estimating the causal effect of fertility (*morekids*) on labor supply (*weeksworked*).

  c. The data set contains the variable *samesex*, which is equal to 1 if the first two children are of the same sex (boy–boy or girl–girl) and equal to 0 otherwise. Are couples whose first two children are of the same sex more likely to have a third child? Is the effect large? Is it statistically significant?

  d. Explain why *samesex* is a valid instrument for the IV regression of *weeksworked* on *morekids*.

  e. Is *samesex* a weak instrument?

  f. Estimate the IV regression of *weeksworked* on *morekids*, using *samesex* as an instrument. How large is the fertility effect on labor supply?

  g. Do the results change when you include the variables *agem1*, *black*, *hispan*, and *othrace* in the labor supply regression (treating these variable as exogenous)? Explain why or why not.

E12.2 Does viewing a violent movie lead to violent behavior? If so, the incidence of violent crimes, such as assaults, should rise following the release of a violent movie that attracts many viewers. Alternatively, movie viewing may substitute for other activities (such as alcohol consumption) that lead to violent behavior, so that assaults should fall when more viewers are attracted to the cinema. On the text website, **http://www.pearsonhighered.com/stock_watson**, you will find the data file **Movies**, which contains data on the number of assaults and movie attendance for 516 weekends from 1995 through 2004.[5] A detailed description is given in **Movies_Description**, available on the website. The data set includes weekend U.S. attendance for strongly violent movies (such as *Hannibal*), mildly violent movies (such as *Spider-Man*), and nonviolent movies (such as *Finding Nemo*). The data set also includes a count of the number of assaults for the same weekend in a subset of counties in the United States. Finally, the data set includes indicators for year, month, whether the weekend is a holiday, and various measures of the weather.

  a. i. Regress the logarithm of the number of assaults [$ln\_assaults =$ $\ln(assaults)$] on the year and month indicators. Is there evidence of seasonality in assaults? That is, do there tend to be more assaults in some months than others? Explain.

  ii. Regress total movie attendance ($attend = attend\_v + attend\_m +$ $attend\_n$) on the year and month indicators. Is there evidence of seasonality in movie attendance? Explain.

---

[5]These are aggregated versions of data provided by Gordon Dahl of University of California–San Diego and Stefano DellaVigna of University of California–Berkeley and were used in their paper "Does Movie Violence Increase Violent Crime?" *Quarterly Journal of Economics*, 2009, 124(2): 677–734.

b. Regress *ln_assaults* on *attend_v*, *attend_m*, *attend_n*, the year and month indicators, and the weather and holiday control variables available in the data set.

  i. Based on the regression, does viewing a strongly violent movie increase or decrease assaults? By how much? Is the estimated effect statistically significant?

  ii. Does attendance at strongly violent movies affect assaults differently than attendance at moderately violent movies? Differently than attendance at nonviolent movies?

  iii. A strongly violent blockbuster movie is released, and the weekend's attendance at strongly violent movies increases by 6 million; meanwhile, attendance falls by 2 million for moderately violent movies and by 1 million for nonviolent movies. What is the predicted effect on assaults? Construct a 95% confidence interval for the change in assaults. [*Hint:* Review Section 7.3 and material surrounding Equations (8.7) and (8.8).]

c. It is difficult to control for all the variables that affect assaults and that might be correlated with movie attendance. For example, the effect of the weather on assaults and movie attendance is only crudely approximated by the weather variables in the data set. However, the data set does include a set of instruments—*pr_attend_v*, *pr_attend_m*, and *pr_attend_n*—that are correlated with attendance but are (arguably) uncorrelated with weekend-specific factors (such as the weather) that affect both assaults and movie attendance. These instruments use historical attendance patterns, not information on a particular weekend, to predict a film's attendance in a given weekend. For example, if a film's attendance is high in the second week of its release, then this can be used to predict that its attendance was also high in the first week of its release. (The details of the construction of these instruments are available in the Dahl and DellaVigna paper referenced in footnote 5.) Run the regression from (b) (including year, month, holiday, and weather controls) but now using *pr_attend_v*, *pr_attend_m*, and *pr_attend_n* as instruments for *attend_v*, *attend_m*, and *attend_n*. Use this IV regression to answer (b)(i)–(b)(iii).

d. The intuition underlying the instruments in (c) is that attendance in a given week is correlated with attendance in surrounding weeks. For each movie category, the data set includes attendance in surrounding weeks. Run the regression using the instruments *attend_v_f*, *attend_m_f*, *attend_n_f*, *attend_v_b, attend_m_b*, and *attend_n_b* instead of the instruments used in (c). Use this IV regression to answer (b)(i)–(b)(iii).

e. There are nine instruments listed in (c) and (d), but only three are needed for identification. Carry out the test for overidentification summarized in Key Concept 12.6. What do you conclude about the validity of the instruments?

   f. Based on your analysis, what do you conclude about the effect of violent movies on (short-run) violent behavior?

E12.3 (This requires Appendix 12.5) On the text website, **http://www.pearson highered.com/stock_watson**, you will find the data set **WeakInstrument**, which contains 200 observations on (Y_i, X_i, Z_i) for the instrumental regression $Y_i = \beta_0 + \beta_1 X_i + u_i$.

   a. Construct $\hat{\beta}_1^{TSLS}$, its standard error, and the usual 95% confidence interval for β_1.

   b. Compute the F-statistic for the regression of X_i on Z_i. Is there evidence of a "weak instrument" problem?

   c. Compute a 95% confidence interval for β_1, using the Anderson–Rubin procedure. (To implement the procedure, assume that $-5 \leq \beta_1 \leq 5$.)

   d. Comment on the differences in the confidence intervals in (a) and (c). Which is more reliable?

12.1 The Cigarette Consumption Panel Data Set

The data set consists of annual data for the 48 contiguous U.S. states from 1985 to 1995. Quantity consumed is measured by annual per capita cigarette sales in packs per fiscal year, as derived from state tax collection data. The price is the real (that is, inflation-adjusted) average retail cigarette price per pack during the fiscal year, including taxes. Income is real per capita income. The general sales tax is the average tax, in cents per pack, due to the broad-based state sales tax applied to all consumption goods. The cigarette-specific tax is the tax applied to cigarettes only. All prices, income, and taxes used in the regressions in this chapter are deflated by the Consumer Price Index and thus are in constant (real) dollars. We are grateful to Professor Jonathan Gruber of MIT for providing us with these data.

12.2 Derivation of the Formula for the TSLS Estimator in Equation (12.4)

The first stage of TSLS is to regress X_i on the instrument Z_i by OLS and then compute the OLS predicted value $\hat{X}_i$; the second stage is to regress Y_i on $\hat{X}_i$ by OLS. Accordingly, the formula for the TSLS estimator, expressed in terms of the predicted value $\hat{X}_i$, is the formula for the OLS estimator in Key Concept 4.2, with $\hat{X}_i$ replacing X_i. That is, $\hat{\beta}_1^{TSLS} = s_{\hat{X}Y}/s_{\hat{X}}^2$, where $s_{\hat{X}}^2$ is the sample variance of $\hat{X}_i$ and $s_{\hat{X}Y}$ is the sample covariance between Y_i and $\hat{X}_i$.

Because $\hat{X}_i$ is the predicted value of X_i from the first-stage regression, $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, the definitions of sample variances and covariances imply that $s_{\hat{X}Y} = \hat{\pi}_1 s_{ZY}$ and $s_{\hat{X}}^2 = \hat{\pi}_1^2 s_Z^2$ (Exercise 12.4). Thus, the TSLS estimator can be written as $\hat{\beta}_1^{TSLS} = s_{\hat{X}Y}/s_{\hat{X}}^2 = s_{ZY}/(\hat{\pi}_1 s_Z^2)$. Finally, $\hat{\pi}_1$ is the OLS slope coefficient from the first stage of TSLS, so $\hat{\pi}_1 = s_{ZX}/s_Z^2$. Substitution of this formula for $\hat{\pi}_1$ into the formula $\hat{\beta}_1^{TSLS} = s_{ZY}/(\hat{\pi}_1 s_Z^2)$ yields the formula for the TSLS estimator in Equation (12.4).

12.3 Large-Sample Distribution of the TSLS Estimator

This appendix studies the large-sample distribution of the TSLS estimator in the case considered in Section 12.1—that is, with a single instrument, a single included endogenous variable, and no included exogenous variables.

To start, we derive a formula for the TSLS estimator in terms of the errors; this formula forms the basis for the remaining discussion, similar to the expression for the OLS estimator in Equation (4.28) in Appendix 4.3.

From Equation (12.1), $Y_i - \overline{Y} = \beta_1(X_i - \overline{X}) + (u_i - \overline{u})$. Accordingly, the sample covariance between Z and Y can be expressed as

$$
\begin{aligned}
s_{ZY} &= \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z})(Y_i - \overline{Y}) \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z})[\beta_1(X_i - \overline{X}) + (u_i - \overline{u})] \\
&= \beta_1 s_{ZX} + \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z})(u_i - \overline{u}) \\
&= \beta_1 s_{ZX} + \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z}) u_i,
\end{aligned}
\tag{12.19}
$$

where $s_{ZX} = [1/(n-1)]\sum_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})$ and where the final equality follows because $\sum_{i=1}^{n}(Z_i - \overline{Z}) = 0$. Substituting the definition of s_{ZX} and the final expression in Equation (12.19) into the definition of $\hat{\beta}_1^{TSLS}$ and multiplying the numerator and denominator by $(n-1)/n$ yields

$$
\hat{\beta}_1^{TSLS} = \beta_1 + \frac{\dfrac{1}{n} \sum_{i=1}^{n} (Z_i - \overline{Z}) u_i}{\dfrac{1}{n} \sum_{i=1}^{n} (Z_i - \overline{Z})(X_i - \overline{X})}.
\tag{12.20}
$$

Large-Sample Distribution of $\hat{\beta}_1^{TSLS}$ When the IV Regression Assumptions in Key Concept 12.4 Hold

Equation (12.20) for the TSLS estimator is similar to Equation (4.28) in Appendix 4.3 for the OLS estimator except that Z rather than X appears in the numerator and that the

denominator is the covariance between Z and X rather than the variance of X. Because of these similarities and because Z is exogenous, the argument in Appendix 4.3 that the OLS estimator is normally distributed in large samples extends to $\hat{\beta}_1^{TSLS}$.

Specifically, when the sample is large, $\overline{Z} \cong \mu_Z$, so the numerator is approximately $\overline{q} = (\frac{1}{n})\sum_{i=1}^n q_i$, where $q_i = (Z_i - \mu_Z)u_i$. Because the instrument is exogenous, $E(q_i) = 0$. By the IV regression assumptions in Key Concept 12.4, q_i is i.i.d. with variance $\sigma_q^2 = \text{var}[(Z_i - \mu_Z)u_i]$. It follows that $\text{var}(\overline{q}) = \sigma_{\overline{q}}^2 = \sigma_q^2/n$, and, by the central limit theorem, $\overline{q}/\sigma_{\overline{q}}$ is, in large samples, distributed $N(0, 1)$.

Because the sample covariance is consistent for the population covariance, $s_{ZX} \xrightarrow{p} \text{cov}(Z_i, X_i)$, which, because the instrument is relevant, is nonzero. Thus, by Equation (12.20), $\hat{\beta}_1^{TSLS} \cong \beta_1 + \overline{q}/\text{cov}(Z_i, X_i)$, so in large samples $\hat{\beta}_1^{TSLS}$ is approximately distributed $N(\beta_1, \sigma_{\hat{\beta}_1^{TSLS}}^2)$, where $\sigma_{\hat{\beta}_1^{TSLS}}^2 = \sigma_{\overline{q}}^2/[\text{cov}(Z_i, X_i)]^2 = (1/n)\text{var}[(Z_i - \mu_Z)u_i]/[\text{cov}(Z_i, X_i)]^2$, which is the expression given in Equation (12.8).

12.4 Large-Sample Distribution of the TSLS Estimator When the Instrument Is Not Valid

This appendix considers the large-sample distribution of the TSLS estimator in the setup of Section 12.1 (one X, one Z) when one or the other of the conditions for instrument validity fails. If the instrument relevance condition fails, the large-sample distribution of the TSLS estimator is not normal; in fact, its distribution is that of a ratio of two normal random variables. If the instrument exogeneity condition fails, the TSLS estimator is inconsistent.

Large-Sample Distribution of $\hat{\beta}_1^{TSLS}$ When the Instrument Is Weak

First consider the case that the instrument is irrelevant, so that $\text{cov}(Z_i, X_i) = 0$. Then the argument in Appendix 12.3 entails division by 0. To avoid this problem, we need to take a closer look at the behavior of the term in the denominator of Equation (12.20) when the population covariance is 0.

We start by rewriting Equation (12.20). Because of the consistency of the sample average, in large samples $\overline{Z}$ is close to μ_Z, and $\overline{X}$ is close to μ_X. Thus the term in the denominator of Equation (12.20) is approximately $(\frac{1}{n})\sum_{i=1}^n (Z_i - \mu_Z)(X_i - \mu_X) = (\frac{1}{n})\sum_{i=1}^n r_i = \overline{r}$, where $r_i = (Z_i - \mu_Z)(X_i - \mu_X)$. Let $\sigma_r^2 = \text{var}[(Z_i - \mu_Z)(X_i - \mu_X)]$, let $\sigma_{\overline{r}}^2 = \sigma_r^2/n$, and let $\overline{q}, \sigma_q^2$, and $\sigma_{\overline{q}}^2$ be as defined in Appendix 12.3. Then Equation (12.20) implies that, in large samples,

$$\hat{\beta}_1^{TSLS} \cong \beta_1 + \frac{\overline{q}}{\overline{r}} = \beta_1 + \left(\frac{\sigma_{\overline{q}}}{\sigma_{\overline{r}}}\right)\left(\frac{\overline{q}/\sigma_{\overline{q}}}{\overline{r}/\sigma_{\overline{r}}}\right) = \beta_1 + \left(\frac{\sigma_q}{\sigma_r}\right)\left(\frac{\overline{q}/\sigma_{\overline{q}}}{\overline{r}/\sigma_{\overline{r}}}\right). \qquad (12.21)$$

If the instrument is irrelevant, then $E(r_i) = \text{cov}(Z_i, X_i) = 0$. Thus $\bar{r}$ is the sample average of the random variables $r_i, i = 1, \ldots, n$, which are i.i.d. (by the second least squares assumption), have variance $\sigma_r^2 = \text{var}[(Z_i - \mu_Z)(X_i - \mu_X)]$ (which is finite by the third IV regression assumption), and have a mean of 0 (because the instruments are irrelevant). It follows that the central limit theorem applies to $\bar{r}$; specifically, $\bar{r}/\sigma_{\bar{r}}$ is approximately distributed $N(0, 1)$. Therefore, the final expression of Equation (12.21) implies that, in large samples, the distribution of $\hat{\beta}_1^{TSLS} - \beta_1$ is the distribution of aS, where $a = \sigma_q/\sigma_r$ and S is the ratio of two random variables, each of which has a standard normal distribution (these two standard normal random variables are correlated).

In other words, when the instrument is irrelevant, the central limit theorem applies to the denominator as well as the numerator of the TSLS estimator, so in large samples the distribution of the TSLS estimator is the distribution of the ratio of two normal random variables. Because X_i and u_i are correlated, these normal random variables are correlated, and the large-sample distribution of the TSLS estimator when the instrument is irrelevant is complicated. In fact, the large-sample distribution of the TSLS estimator with irrelevant instruments is centered on the probability limit of the OLS estimator. Thus when the instrument is irrelevant, TSLS does not eliminate the bias in OLS and, moreover, has a nonnormal distribution, even in large samples.

A weak instrument represents an intermediate case between an irrelevant instrument and the normal distribution derived in Appendix 12.3. When the instrument is weak but not irrelevant, the distribution of the TSLS estimator continues to be nonnormal, so the general lesson here about the extreme case of an irrelevant instrument carries over to weak instruments.

Large-Sample Distribution of $\hat{\beta}_1^{TSLS}$ When the Instrument Is Endogenous

The numerator in the final expression in Equation (12.20) converges in probability to $\text{cov}(Z_i, u_i)$. If the instrument is exogenous, this is 0, and the TSLS estimator is consistent (assuming that the instrument is not weak). If, however, the instrument is not exogenous, then, if the instrument is not weak, $\hat{\beta}_1^{TSLS} \xrightarrow{p} \beta_1 + \text{cov}(Z_i, u_i)/\text{cov}(Z_i, X_i) \neq \beta_1$. That is, if the instrument is not exogenous, the TSLS estimator is inconsistent.

APPENDIX

12.5 Instrumental Variables Analysis with Weak Instruments

This appendix discusses some methods for instrumental variables analysis in the presence of potentially weak instruments. The appendix focuses on the case of a single included endogenous regressor [Equations (12.13) and (12.14)].

Testing for Weak Instruments

The rule of thumb in Key Concept 12.5 is that a first-stage F-statistic less than 10 indicates that the instruments are weak. One motivation for this rule of thumb arises from an approximate expression for the bias of the TSLS estimator. Let β_1^{OLS} denote the probability limit of the OLS estimator β_1, and let $\beta_1^{OLS} - \beta_1$ denote the asymptotic bias of the OLS estimator (if the regressor is endogenous, then $\hat{\beta}_1 \xrightarrow{p} \beta_1^{OLS} \neq \beta_1$). It is possible to show that, when there are many instruments, the bias of the TSLS estimator is approximately $E(\hat{\beta}_1^{TSLS}) - \beta_1 \approx (\beta_1^{OLS} - \beta_1)/[E(F) - 1]$, where $E(F)$ is the expectation of the first-stage F-statistic. If $E(F) = 10$, then the bias of TSLS relative to the bias of OLS is approximately 1/9, or just over 10%, which is small enough to be acceptable in many applications. Replacing $E(F) > 10$ with $F > 10$ yields the rule of thumb in Key Concept 12.5.

The motivation in the previous paragraph involved an approximate formula for the bias of the TSLS estimator when there are many instruments. In most applications, however, the number of instruments, m, is small. Stock and Yogo (2005) provide a formal test for weak instruments that avoids the approximation that m is large. In the Stock–Yogo test, the null hypothesis is that the instruments are weak, and the alternative hypothesis is that the instruments are strong, where strong instruments are defined to be instruments for which the bias of the TSLS estimator is at most 10% of the bias of the OLS estimator. The test entails comparing the first-stage F-statistic (for technical reasons, the homoskedasticity-only version) to a critical value that depends on the number of instruments. As it happens, for a test with a 5% significance level, this critical value ranges between 9.08 and 11.52, so the rule of thumb of comparing F to 10 is a good approximation to the Stock–Yogo test.

Hypothesis Tests and Confidence Sets for β

If the instruments are weak, the TSLS estimator is biased and has a nonnormal distribution. Thus the TSLS t-test of $\beta_1 = \beta_{1,0}$ is unreliable, as is the TSLS confidence interval for β_1. There are, however, other tests of $\beta_1 = \beta_{1,0}$, along with confidence intervals based on those tests, that are valid whether instruments are strong, weak, or even irrelevant. When there is a single endogenous regressor, the preferred test is Moreira's (2003) conditional likelihood ratio (CLR) test. An older test, which works for any number of endogenous regressors, is based on the Anderson–Rubin (1949) statistic. Because the Anderson–Rubin statistic is conceptually less complicated, we describe it first.

The Anderson–Rubin test of $\beta_1 = \beta_{1,0}$ proceeds in two steps. In the first step, compute a new variable, $Y_i^* = Y_i - \beta_{1,0}X_i$. In the second step, regress Y_i^* against the included exogenous regressors (W's) and the instruments (Z's). The Anderson–Rubin statistic is the F-statistic testing the hypothesis that the coefficients on the Z's are all 0. Under the null hypothesis that $\beta_1 = \beta_{1,0}$, if the instruments satisfy the exogeneity condition (condition 2 in Key Concept 12.3), they will be uncorrelated with the error term in this regression, and the null hypothesis will be rejected in 5% of all samples.

As discussed in Sections 3.3 and 7.4, a confidence set can be constructed as the set of values of the parameters that are not rejected by a hypothesis test. Accordingly, the set of values of β_1 that are not rejected by a 5% Anderson–Rubin test constitutes a 95% confidence

set for β_1. When the Anderson–Rubin F-statistic is computed using the homoskedasticity-only formula, the Anderson–Rubin confidence set can be constructed by solving a quadratic equation (see Empirical Exercise 12.3). The logic behind the Anderson–Rubin statistic never assumes instrument relevance, and the Anderson–Rubin confidence set will have a coverage probability of 95% in large samples, whether the instruments are strong, weak, or even irrelevant.

The CLR statistic also tests the hypothesis that $\beta_1 = \beta_{1,0}$. Likelihood ratio statistics compare the value of the likelihood (see Appendix 11.2) under the null hypothesis to its value under the alternative and reject it if the likelihood under the alternative is sufficiently greater than under the null. Familiar test statistics in this text, such as the homoskedasticity-only F-statistic in multiple regression, can be derived as likelihood ratio statistics under the assumption of homoskedastic normally distributed errors. Unlike any of the other tests discussed in this text, however, the critical value of the CLR test depends on the data—specifically, on a statistic that measures the strength of the instruments. By using the right critical value, the CLR test is valid whether instruments are strong, weak, or irrelevant. CLR confidence intervals can be computed as the set of values of β_1 that are not rejected by the CLR test.

The CLR test is equivalent to the TSLS t-test when instruments are strong and has very good power when instruments are weak. With suitable software, the CLR test is easy to use. The disadvantage of the CLR test is that it does not generalize readily to more than one endogenous regressor. In that case, the Anderson–Rubin test (and confidence set) is recommended; however, when instruments are strong (so TSLS is valid) and the coefficients are overidentified, the Anderson–Rubin test is inefficient in the sense that it is less powerful than the TSLS t-test.

Estimation of β

If the instruments are irrelevant, then without further restrictions it is not possible to obtain an unbiased estimator of β_1, even in large samples. With weak instruments, CLR or Anderson–Rubin confidence intervals for the coefficients are preferable to point estimation.

The problems of estimation, testing, and confidence intervals in IV regression with weak instruments constitute an area of ongoing research. To learn more about this topic, visit the website for this text.

12.6 TSLS with Control Variables

In Key Concept 12.4, the W variables are assumed to be exogenous. This appendix considers the case in which W is not exogenous but instead is a control variable included to make Z exogenous. The logic of control variables in TSLS parallels the logic in OLS: If a control variable effectively controls for an omitted factor, then the instrument is uncorrelated with the error term. Because the control variable is correlated with the error term, the coefficient on a

control variable does not have a causal interpretation. The mathematics of control variables in TSLS also parallels the mathematics of control variables in OLS and entails relaxing the assumption that the error has conditional mean 0 given Z and W to be that the conditional mean of the error does not depend on Z. This appendix draws on Appendix 6.5 (OLS with control variables), which should be reviewed first.

Consider the IV regression model in Equation (12.12) with a single X and a single W:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i. \tag{12.22}$$

We replace IV regression assumption 1 in Key Concept 12.4 [which states that $E(u_i \mid W_i) = 0$] with the assumption that, conditional on W_i, the mean of u_i does not depend on Z_i:

$$E(u_i \mid W_i, Z_i) = E(u_i \mid W_i). \tag{12.23}$$

The next steps in the argument parallel those for regression with control variables in Equations (6.23)–(6.25) in Appendix 6.5. Assume that $E(u_i \mid W_i)$ is linear in W_i, so $E(u_i \mid W_i) = \gamma_0 + \gamma_1 W_i$, where γ_0 and γ_1 are coefficients. Then

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i - E(u_i \mid W_i, Z_i) + E(u_i \mid W_i, Z_i) \\
&= \beta_0 + \beta_1 X_i + \beta_2 W_i + \varepsilon_i + \gamma_0 + \gamma_1 W_i
\end{aligned}
\tag{12.24}
$$

where the first line adds and subtracts $E(u_i \mid W_i, Z_i)$ to the right hand side of Equation (12.22), and the second line and defines $\varepsilon_i = u_i - E(u_i \mid W_i, Z_i)$ and uses the conditional mean independence assumption plus linearity to write $E(u_i \mid W_i, Z_i) = E(u_i \mid W_i) = \gamma_0 + \gamma_1 W_i$. We thus have that,

$$Y_i = \delta_0 + \beta_1 X_i + \delta_1 W_i + \varepsilon_i, \tag{12.25}$$

where $\delta_0 = \beta_0 + \gamma_0$ and $\delta_1 = \beta_2 + \gamma_1$. Now $E(\varepsilon_i \mid W_i, Z_i) = E[u_i - E(u_i \mid W_i, Z_i) \mid W_i, Z_i] = E(u_i \mid W_i, Z_i) - E(u_i \mid W_i, Z_i) = 0$, which in turn implies $\text{corr}(Z_i, \varepsilon_i) = 0$. Thus IV regression assumption 1 and the instrument exogeneity requirement (condition 2 in Key Concept 12.3) both hold for Equation (12.24) with error term ε_i. Thus, if IV regression assumption 1 is replaced by conditional mean independence in Equation (12.23), the original IV regression assumptions in Key Concept 12.4 apply to the modified regression in Equation (12.25).

Because the IV regression assumptions of Key Concept 12.4 hold for Equation (12.25), all the methods of inference (for both weak and strong instruments) discussed in this chapter apply to Equation (12.25). In particular, if the instruments are strong, the coefficients in Equation (12.25) will be estimated consistently by TSLS and TSLS tests, and confidence intervals will be valid.

Just as in OLS with control variables, in general the TSLS coefficient on the control variable W does not have a causal interpretation. TSLS consistently estimates δ_1 in Equation (12.25), but δ_1 is the sum of β_2, the direct causal effect of W, and γ_1, which reflects the correlation between W and the omitted factors in u_i for which W controls.

In the cigarette consumption regressions in Table 12.1, it is tempting to interpret the coefficient on the 10-year change in log income as the income elasticity of demand. If, however, income growth is correlated with increases in education and if more education reduces smoking, income growth would have its own causal effect (β_2, the income elasticity) plus an effect arising from its correlation with education (γ_1). If the latter effect is negative ($\gamma_1 < 0$), the income coefficients in Table 12.1 (which estimate $\delta_1 = \beta_2 + \gamma_1$) would underestimate the income elasticity. As long as the conditional mean independence assumption in Equation (12.23) holds, however, the TSLS estimator of the price elasticity is consistent, even if the estimate of the income elasticity is not.

13

Experiments and Quasi-Experiments

In many fields, such as psychology and medicine, causal effects are commonly estimated using experiments. Before being approved for widespread medical use, for example, a new drug must be subjected to experimental trials in which some patients are randomly selected to receive the drug while others are given a harmless ineffective substitute (a placebo); the drug is approved only if this randomized controlled experiment provides convincing statistical evidence that the drug is safe and effective.

There are three reasons to study randomized controlled experiments in an econometrics course. First, an ideal randomized controlled experiment provides a conceptual benchmark against which to judge estimates of causal effects made with observational data. Second, the results of randomized controlled experiments, when conducted, can be very influential, so it is important to understand the limitations and threats to validity of actual experiments, as well as their strengths. Third, external circumstances sometimes produce what appears to be randomization; that is, because of external events, the treatment of some individual occurs "as if" it is random, possibly conditional on some control variables. This "as if" randomness produces a *quasi-experiment* or *natural experiment*, and many of the methods developed for analyzing randomized experiments can be applied (with some modifications) to quasi-experiments.

This chapter examines experiments and quasi-experiments in economics. The statistical tools used in this chapter are multiple regression analysis, regression analysis of panel data, and instrumental variables (IV) regression. What distinguishes the discussion in this chapter is not the tools used but rather the type of data analyzed and the special opportunities and challenges posed when analyzing experiments and quasi-experiments.

The methods developed in this chapter are often used for evaluating social or economic programs. **Program evaluation** is the field of study that concerns estimating the effect of a program, policy, or some other intervention or "treatment." What is the effect on earnings of going through a job training program? What is the effect on employment of low-skilled workers of an increase in the minimum wage? What is the effect on college attendance of making low-cost student aid loans available to middle-class students? This chapter discusses how such programs or policies can be evaluated using experiments or quasi-experiments.

We begin in Section 13.1 by elaborating on the discussions in Chapters 1, 3, and 4 of the estimation of causal effects using randomized controlled experiments. In reality, actual experiments with human subjects encounter practical problems that constitute threats to their internal and external validity; these threats and some econometric

tools for addressing them are discussed in Section 13.2. Section 13.3 analyzes an important randomized controlled experiment in which elementary students were randomly assigned to different-sized classes in the state of Tennessee in the late 1980s.

Section 13.4 turns to the estimation of causal effects using quasi-experiments. Threats to the validity of quasi-experiments are discussed in Section 13.5. One issue that arises in both experiments and quasi-experiments is that treatment effects can differ from one member of the population to the next, and the matter of interpreting the resulting estimates of causal effects when the population is heterogeneous is taken up in Section 13.6.

13.1   Potential Outcomes, Causal Effects, and Idealized Experiments

This section explains how the population mean of individual-level causal effects can be estimated using a randomized controlled experiment and how data from such an experiment can be analyzed using multiple regression analysis.

Potential Outcomes and the Average Causal Effect

Suppose that you are considering taking a drug for a medical condition, enrolling in a job training program, or doing an optional econometrics problem set. It is reasonable to ask, What are the benefits of doing so—receiving the treatment—for *me*? You can imagine two hypothetical situations, one in which you receive the treatment and one in which you do not. Under each hypothetical situation, there would be a measurable outcome (the progress of the medical condition, getting a job, your econometrics grade). The difference in these two potential outcomes would be the causal effect, for you, of the treatment.

More generally, a **potential outcome** is the outcome for an individual under a potential treatment. The causal effect for that individual is the difference in the potential outcome if the treatment is received and the potential outcome if it is not. In general, the causal effect can differ from one individual to the next. For example, the effect of a drug could depend on your age, whether you smoke, or other health conditions. The problem is that there is no way to measure the causal effect for a single individual: Because the individual either receives the treatment or does not, one of the potential outcomes can be observed—but not both.

Although the causal effect cannot be measured for a single individual, in many applications it suffices to know the mean causal effect in a population. For example, a job training program evaluation might trade off the average expenditure per trainee against average trainee success in finding a job. The mean of the individual causal effects in the population under study is called the **average causal effect** or the **average treatment effect**.

The average causal effect for a given population can be estimated, at least in theory, using an ideal randomized controlled experiment. To see how, first suppose that the subjects are selected at random from the population of interest. Because the

subjects are selected by simple random sampling, their potential outcomes, and thus their causal effects, are drawn from the same distribution, so the expected value of the causal effect in the sample is the average causal effect in the population. Next suppose that subjects are randomly assigned to the treatment or the control group. Because an individual's treatment status is randomly assigned, it is distributed independently of his or her potential outcomes. Thus the expected value of the outcome for those treated minus the expected value of the outcome for those not treated equals the expected value of the causal effect. Thus when the concept of potential outcomes is combined with (1) random selection of individuals from a population and (2) random experimental assignment of treatment to those individuals, the expected value of the difference in outcomes between the treatment and control groups is the average causal effect in the population. That is, as was stated in Section 3.5, the average causal effect on Y_i of treatment ($X_i = 1$) versus no treatment ($X_i = 0$) is the difference in the conditional expectations, $E(Y_i|X_i = 1) - E(Y_i|X_i = 0)$, where $E(Y_i|X_i = 1)$ and $E(Y_i|X_i = 0)$ are, respectively, the expected values of Y for the treatment and control groups in an ideal randomized controlled experiment. Appendix 13.3 provides a mathematical treatment of the foregoing reasoning.

In general, an individual causal effect can be thought of as depending both on observable variables and on unobservable variables. We have already encountered the idea that a causal effect can depend on observable variables; for example, Chapter 8 examined the possibility that the effect of a class size reduction might depend on whether a student is an English learner. Through Section 13.5, we consider the case that variation in causal effects depends only on observable variables. Section 13.6 takes up the case that causal effects depend on unobserved variables.

Econometric Methods for Analyzing Experimental Data

Data from a randomized controlled experiment can be analyzed by comparing differences in means or by a regression that includes the treatment indicator and additional control variables. This latter specification, the differences estimator with additional regressors, can also be used in more complicated randomization schemes, in which the randomization probabilities depend on observable covariates.

The differences estimator. The **differences estimator** is the difference in the sample averages for the treatment and control groups (Section 3.5), which can be computed by regressing the outcome variable Y on a binary treatment indicator X:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \ i = 1, \ldots, n. \tag{13.1}$$

As discussed in Section 4.4, if X is randomly assigned, then $E(u_i|X_i) = 0$, and the OLS estimator of the causal effect β_1 in Equation (13.1) is an unbiased and consistent estimator of the causal effect.

The differences estimator with additional regressors. The efficiency of the difference estimator often can be improved by including some control variables W in the regression; doing so leads to the **differences estimator with additional regressors**:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i, \ i = 1, \ldots, n. \qquad (13.2)$$

If W helps to explain the variation in Y, then including W reduces the standard error of the regression and, typically, the standard error of $\hat{\beta}_1$. As discussed in Section 7.5 and Appendix 6.5, for the estimator $\hat{\beta}_1$ of the causal effect β_1 in Equation (13.2) to be unbiased, the control variables W must be such that u_i satisfies conditional mean independence; that is, $E(u_i|X_i,W_i) = E(u_i|W_i)$. This condition is satisfied if W_i are pretreatment individual characteristics, such as sex: If W_i is a pretreatment characteristic and X_i is randomly assigned, then X_i is independent of u_i and W_i, so $E(u_i|X_i,W_i) = E(u_i|W_i)$. The W regressors in Equation (13.2) should not include experimental outcomes (X_i is not randomly assigned, given an experimental outcome). As usual with control variables under conditional mean independence, the coefficients on the control variables do not have a causal interpretation.

Estimating causal effects that depend on observables. As discussed in Chapter 8, variation in causal effects that depends on observables can be estimated by including suitable nonlinear functions of, or interactions with, X_i. For example, if W_{1i} is a binary indicator denoting sex, then distinct causal effects for men and women can be estimated by including the interaction variable $W_{1i} \times X_i$ in the regression in Equation (13.2).

Randomization based on covariates. Randomization in which the probability of assignment to the treatment group depends on one or more observable variables W is called **randomization based on covariates**. If randomization is based on covariates, then in general the differences estimator based on Equation (13.1) suffers from omitted variable bias. For example, consider a hypothetical experiment to estimate the causal effect of mandatory versus optional homework in an econometrics course. Suppose that there is random assignment, but economics majors ($W_i = 1$) are assigned to the treatment group (mandatory homework, $X_i = 1$) with higher probability than nonmajors ($W_i = 0$). If majors tend to do better in the course than nonmajors anyway, then there is omitted variable bias because being in the treatment group is correlated with the omitted variable, being a major.

  Because X_i is randomly assigned given W_i, this omitted variable bias can be eliminated by using the differences estimator with the additional control variable W_i. The random assignment of X_i given W_i implies that, given W_i, the mean of u_i does not depend on X_i; that is, $E(u_i|X_i,W_i) = E(u_i|W_i)$. Thus if the treatment effect is the same for majors and nonmajors, the first least squares assumption for causal inference with control variables (Key Concept 6.6) is satisfied, and the OLS estimator $\hat{\beta}_1$ in Equation (13.2) is an unbiased estimator of the causal effect when X_i is assigned randomly based on W_i. If the treatment effect is different for majors and nonmajors, then the interaction term $X_i \times W_i$ needs to be added to Equation (13.2), and with this addition, the first least squares assumption for causal inference with control variables is satisfied.

13.2 Threats to Validity of Experiments

Recall from Key Concept 9.1 that a statistical study is *internally valid* if the statistical inferences about causal effects are valid for the population being studied; it is *externally valid* if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings. Various real-world problems pose threats to the internal and external validity of the statistical analysis of actual experiments with human subjects.

Threats to Internal Validity

Threats to the internal validity of randomized controlled experiments include failure to randomize, failure to follow the treatment protocol, attrition, experimental effects, and small sample sizes.

Failure to randomize. If the treatment is not assigned randomly but instead is based in part on the characteristics or preferences of the subject, then experimental outcomes will reflect both the effect of the treatment and the effect of the nonrandom assignment. For example, suppose that participants in a job training program experiment are assigned to the treatment group depending on whether their last name falls in the first or second half of the alphabet. Because of ethnic differences in last names, ethnicity could differ systematically between the treatment and control groups. To the extent that work experience, education, and other labor market characteristics differ by ethnicity, there could be systematic differences between the treatment and control groups in these omitted factors that affect outcomes. In general, nonrandom assignment can lead to correlation between X_i and u_i in Equations (13.1) and (13.2), which in turn leads to bias in the estimator of the treatment effect.

It is possible to test for randomization. If treatment is randomly received, then X_i will be uncorrelated with observable pretreatment individual characteristics W. Thus a **test for random receipt of treatment** entails testing the hypothesis that the coefficients on $W_{1i}, \ldots, W_{ri}$ are 0 in a regression of X_i on $W_{1i}, \ldots, W_{ri}$. In the job training program example, regressing receipt of job training (X_i) on sex, race, and prior education (W's) and then computing the F-statistic testing whether the coefficients on the W's are 0 provides a test of the null hypothesis that treatment was randomly received against the alternative hypothesis that receipt of treatment depended on sex, race, or prior education. If the experimental design performs randomization conditional on covariates, then those covariates would be included in the regression, and the F-test would test the coefficients on the remaining W's.[1]

---

[1]In this example, X_i is binary, so, as discussed in Chapter 11, the regression of X_i on $W_{1i}, \ldots, W_{ri}$ is a linear probability model, and heteroskedasticity-robust standard errors are essential. Another way to test the hypothesis that $E(X_i | W_{1i}, \ldots, W_{ri})$ does not depend on $W_{1i}, \ldots, W_{ri}$ when X_i is binary is to use a probit or logit model (see Section 11.2).

Failure to follow the treatment protocol. In an actual experiment, people do not always do what they are told. In a job training program experiment, for example, some of the subjects assigned to the treatment group might not show up for the training sessions and thus not receive the treatment. Similarly, subjects assigned to the control group might somehow receive the training anyway, perhaps by making a special request to an instructor or administrator.

The failure of individuals to follow completely the randomized treatment protocol is called **partial compliance** with the treatment protocol. Suppose that the experimenter knows whether the treatment was actually received (for example, whether the trainee attended class), and the treatment actually received is recorded as X_i. With partial compliance, there is an element of choice in whether the subject receives the treatment, so X_i can be correlated with u_i even if initially there is random assignment. Thus failure to follow the treatment protocol leads to bias in the OLS estimator.

If there are data on both treatment actually received (X_i) and the initial random assignment, then the treatment effect can be estimated by instrumental variables regression. **Instrumental variables estimation of the treatment effect** entails the estimation of Equation (13.1)—or Equation (13.2) if there are control variables—using the initial random assignment (Z_i) as an instrument for the treatment actually received (X_i). Recall that a variable must satisfy the two conditions of instrument relevance and instrument exogeneity (Key Concept 12.3) to be a valid instrumental variable. As long as the protocol is partially followed, then the actual treatment level is partially determined by the assigned treatment level, so the instrumental variable Z_i is relevant. If initial assignment is random, then Z_i is distributed independently of u_i (conditional on W_i, if randomization is conditional on covariates), so the instrument is exogenous. Thus in an experiment with randomly assigned treatment, partial compliance, and data on actual treatment, the original random assignment is a valid instrumental variable.

Attrition. **Attrition** refers to subjects dropping out of the study after being randomly assigned to the treatment or the control group. Sometimes attrition occurs for reasons unrelated to the treatment program; for example, a participant in a job training study might need to leave town to care for a sick relative. But if the reason for attrition is related to the treatment itself, then the attrition can result in bias in the OLS estimator of the causal effect. For example, suppose that the most able trainees drop out of the job training program experiment because they get out-of-town jobs acquired using the job training skills, so at the end of the experiment only the least able members of the treatment group remain. Then the distribution of unmeasured characteristics (ability) will differ between the control and treatment groups (the treatment enabled the ablest trainees to leave town). In other words, the treatment X_i will be correlated with u_i (which includes ability) for those who remain in the sample at the end of the experiment, so the differences estimator will be biased. Because attrition results in a nonrandomly selected sample, attrition that is related to the treatment leads to selection bias (Key Concept 9.4).

The Hawthorne Effect

During the 1920s and 1930s, the General Electric Company conducted a series of studies of worker productivity at its Hawthorne plant. In one set of experiments, the researchers varied lightbulb wattage to see how lighting affected the productivity of women assembling electrical parts. In other experiments, they increased or decreased rest periods, changed the workroom layout, and shortened workdays. Influential early reports on these studies concluded that productivity continued to rise whether the lights were dimmer or brighter, whether workdays were longer or shorter, or whether conditions improved or worsened. Researchers concluded that the productivity improvements were not the consequence of changes in the workplace but instead came about because their special role in the experiment made the workers feel noticed and valued, so they worked harder and harder. Over the years, the idea that being in an experiment influences subject behavior has come to be known as the Hawthorne effect.

But there is a glitch to this story: Careful examination of the actual Hawthorne data reveals no Hawthorne effect (Gillespie, 1991; Jones, 1992)! Still, in some experiments, especially ones in which the subjects have a stake in the outcome, merely being in an experiment could affect behavior. The Hawthorne effect and experimental effects more generally can pose threats to internal validity—even though the Hawthorne effect is not evident in the original Hawthorne data.

Experimental effects. In experiments with human subjects, merely because the subjects are in an experiment can change their behavior, a phenomenon sometimes called the **Hawthorne effect** (see the box "The Hawthorne Effect").

In some experiments, a "double-blind" protocol can mitigate the effect of being in an experiment: Although subjects and experimenters both know that they are in an experiment, neither knows whether a subject is in the treatment group or the control group. In a medical drug experiment, for example, sometimes the drug and the placebo can be made to look the same so that neither the medical professional dispensing the drug nor the patient knows whether the administered drug is the real thing or the placebo. If the experiment is double-blind, then both the treatment and control groups should experience the same experimental effects, so different outcomes between the two groups can be attributed to the drug.

Double-blind experiments are often infeasible in real-world experiments in economics: Both the experimental subject and the instructor know whether the subject is attending the job training program. In a poorly designed experiment, this experimental effect could be substantial. For example, teachers in an experimental program might try especially hard to make the program a success if they think their future employment depends on the outcome of the experiment. Deciding whether experimental results are biased because of the experimental effects requires making judgments based on details of how the experiment was conducted.

Small sample sizes. Because experiments with human subjects can be expensive, sometimes the sample size is small. A small sample size does not bias estimators of the causal effect, but it does mean that the causal effect is estimated imprecisely. A small sample also raises threats to the validity of confidence intervals and hypothesis tests. Because inference based on normal critical values and heteroskedasticity-robust standard errors is justified using large-sample approximations, experimental data with small samples are sometimes analyzed under the assumption that the errors are normally distributed (Sections 3.6 and 5.6); however, the assumption of normality is typically as dubious for experimental data as it is for observational data.

Threats to External Validity

Threats to external validity compromise the ability to generalize the results of the study to other populations and settings.

Nonrepresentative sample. The population studied and the population of interest must be sufficiently similar to justify generalizing the experimental results. If a job training program is evaluated in an experiment with former prison inmates, then it might be possible to generalize the study results to other former prison inmates. Because a criminal record weighs heavily on the minds of potential employers, however, the results might not generalize to workers who have never committed a crime.

Nonrepresentative program or policy. The policy or program of interest must be sufficiently similar to the program studied to permit generalizing the results. A program studied in a small-scale, tightly monitored experiment could be quite different from the program actually implemented. If the program actually implemented is widely available, then the scaled-up program might not provide the same quality control as the experimental version or might be funded at a lower level; either possibility could result in the full-scale program being less effective than the smaller experimental program. Another difference between an experimental program and an actual program might be its duration: The experimental program lasts only for the length of the experiment, whereas the actual program under consideration might be available for longer periods of time.

General equilibrium effects. An issue related to scale and duration concerns what economists call general equilibrium effects. Turning a small, temporary experimental program into a widespread, permanent program might change the economic environment sufficiently that the results from the experiment cannot be generalized. A small, experimental job training program, for example, might supplement training by employers, but if the program were made widely available, it could displace employer-provided training, thereby reducing the net benefits of the program. An internally valid small experiment might correctly measure a causal effect, holding constant the market or policy environment, but general equilibrium effects mean that these other factors are not, in fact, held constant when the program is implemented broadly.

13.3 Experimental Estimates of the Effect of Class Size Reductions

In this section, we return to a question addressed in Part II: What is the effect on test scores of reducing class size in the early grades? In the late 1980s, Tennessee conducted a large, multimillion-dollar randomized controlled experiment to ascertain whether class size reduction was an effective way to improve elementary education. The results of this experiment have strongly influenced our understanding of the effect of class size reductions.

Experimental Design

The Tennessee class size reduction experiment, known as Project STAR (Student–Teacher Achievement Ratio), was a 4-year experiment designed to evaluate the effect on learning of small class sizes. Funded by the Tennessee state legislature, the experiment cost approximately $12 million. The study compared three different class arrangements for kindergarten through third grade: a regular-sized class, with 22 to 25 students per class, a single teacher, and no teacher's aide; a small class, with 13 to 17 students per class and no teacher's aide; and a regular-sized class with a teacher's aide.

Each school participating in the experiment had at least one class of each type, and students entering kindergarten in a participating school were randomly assigned to one of these three groups at the beginning of the 1985–1986 academic year. Teachers were also assigned randomly to one of the three types of classes.

According to the original experimental protocol, students would stay in their initially assigned class type for the 4 years of the experiment (kindergarten through third grade). However, because of parent complaints, students initially assigned to a regular class (with or without an aide) were randomly reassigned at the beginning of first grade to a regular class with an aide or to a regular class without an aide; students initially assigned to a small class remained in a small class. Students entering school in first grade (kindergarten was optional), in the second year of the experiment, were randomly assigned to one of the three groups. Each year students in the experiment were given standardized tests (the Stanford Achievement Test) in reading and math.

The project paid for the additional teachers and aides necessary to achieve the target class sizes. During the first year of the study, approximately 6400 students participated in 108 small classes, 101 regular-sized classes, and 99 regular-sized classes with an aide. Over all 4 years of the study, a total of approximately 11,600 students at 80 schools participated in the study.

Deviations from the experimental design. The experimental protocol specified that the students should not switch between class groups except through the re-randomization at the beginning of first grade. However, approximately 10% of the students switched in subsequent years for reasons including incompatible children and behavioral problems. These switches represent a departure from the randomization scheme and,

depending on the true nature of the switches, have the potential to introduce bias into the results. Switches made purely to avoid personality conflicts might be sufficiently unrelated to the experiment that they would not introduce bias. If, however, the switches arose because the parents most concerned with their children's education pressured the school into switching a child into a small class, then this failure to follow the experimental protocol could bias the results toward overstating the effectiveness of small classes. Another deviation from the experimental protocol was that the class sizes changed over time because students switched between classes and moved in and out of the school district.

Analysis of the STAR Data

Because there are two treatment groups—small class and regular-sized class with an aide—the regression version of the differences estimator needs to be modified to handle the two treatment groups and the control group. This modification is done by introducing two binary variables, one indicating whether the student is in a small class and another indicating whether the student is in a regular-sized class with an aide, which leads to the population regression model

$$Y_i = \beta_0 + \beta_1 SmallClass_i + \beta_2 RegAide_i + u_i, \tag{13.3}$$

where Y_i is a test score, $SmallClass_i = 1$ if the ith student is in a small class and $= 0$ otherwise, and $RegAide_i = 1$ if the ith student is in a regular class with an aide and $= 0$ otherwise. The effect on the test score of a small class relative to a regular class is β_1, and the effect of a regular class with an aide relative to a regular class is β_2. The differences estimator for the experiment can be computed by estimating β_1 and β_2 in Equation (13.3) by OLS.

Because of the design of the experiment, the observations are not plausibly i.i.d. In particular, once a school is chosen, all students at the school participate. Because students at a given school typically come from the same area, they can share similar unobserved characteristics, such as parental education. Thus, the error term u_i in Equation (13.3) could be correlated across students in the same school. While this correlation does not lead to bias, the standard errors need to be computed in a way that allows for this correlation. Because clustered standard errors allow for correlation within entities (schools) but not across entities (see Section 10.5 and Appendix 10.2), we compute standard errors clustered at the school level.

Table 13.1 presents the differences estimates of the effect on test scores of being in a small class or in a regular-sized class with an aide. The dependent variable Y_i in the regressions in Table 13.1 is the student's total score on the combined math and reading portions of the Stanford Achievement Test. According to the estimates in Table 13.1, for students in kindergarten, the effect of being in a small class is an increase of 13.9 points on the test, relative to being in a regular class; the estimated effect of being in a regular class with an aide is only 0.31 points on

TABLE 13.1	Project STAR: Differences Estimates of Effect on Standardized Test Scores of Class Size Treatment Group			
	Grade			
Regressor	**K**	**1**	**2**	**3**
Small class	13.90 (4.23) [5.48, 22.32]	29.78 (4.79) [20.24, 39.32]	19.39 (5.12) [9.18, 29.61]	15.59 (4.21) [7.21, 23.97]
Regular-sized class with aide	0.31 (3.77) [−7.19, 7.82]	11.96 (4.87) [2.27, 21.65]	3.48 (4.91) [−6.31, 13.27]	−0.29 (4.04) [−8.35, 7.77]
Intercept	918.04 (4.82)	1039.39 (5.82)	1157.81 (5.29)	1228.51 (4.66)
Number of observations	5786	6379	6049	5967

The regressions were estimated using the Project STAR public access data set described in Appendix 13.1. The dependent variable is the student's combined score on the math and reading portions of the Stanford Achievement Test. Standard errors, clustered at the school level, appear in parentheses, and 95% confidence intervals appear in brackets.

the test. For each grade, the null hypothesis that small classes provide no improvement is rejected at the 0.5% (two-sided) significance level. However, it is not possible to reject the null hypothesis that having an aide in a regular class provides no improvement, relative to not having an aide, except in first grade, even at the 10% significance level. The estimated magnitudes of the improvements in small classes are broadly similar in grades K, 2, and 3, although the estimate is larger for first grade.

The differences estimates in Table 13.1 suggest that reducing class size has an effect on test performance, but that adding an aide to a regular-sized class has a much smaller effect, possibly 0. As discussed in Section 13.1, augmenting the regressions in Table 13.1 with additional regressors—the W regressors in Equation (13.2)—can provide more efficient estimates of the causal effects. Moreover, if the treatment received is not random because of failures to follow the treatment protocol, then the estimates of the experimental effects based on regressions with additional regressors could differ from the difference estimates reported in Table 13.1. For these two reasons, estimates of the experimental effects in which additional regressors are included in Equation (13.3) are reported for kindergarten in Table 13.2; the first column of Table 13.2 repeats the results of the first column of Table 13.1, and the remaining three columns include additional regressors that measure teacher, school, and student characteristics.

The main conclusion from Table 13.2 is that the multiple regression estimates of the causal effects of the two treatments (small class and regular-sized class with aide) in the final three columns of Table 13.2 are similar to the differences estimates reported in the first column. That adding these observable regressors does not change the estimated causal effects of the different treatments makes it more plausible that

Regressor	(1)	(2)	(3)	(4)
TABLE 13.2 Project STAR: Differences Estimates with Additional Regressors for Kindergarten				
Small class	13.90	14.00	15.93	15.89
	(4.23)	(4.25)	(4.08)	(3.95)
	[5.48, 22.32]	[5.55, 22.46]	[7.81, 24.06]	[8.03, 23.74]
Regular-sized class with aide	0.31	−0.60	1.22	1.79
	(3.77)	(3.84)	(3.64)	(3.60)
	[−7.19, 7.82]	[−8.25, 7.05]	[−6.04, 8.47]	[−5.38, 8.95]
Teacher's years of experience		1.47	0.74	0.66
		(0.44)	(0.35)	(0.36)
		[0.60, 2.34]	[0.04, 1.45]	[−0.05, 1.37]
Boy				−12.09
				(1.54)
Free lunch eligible				−34.70
				(2.47)
Black				−25.43
				(4.52)
Race other than black or white				−8.50
				(12.64)
School indicator variables?	no	no	yes	yes
$\overline{R}^2$	0.01	0.02	0.22	0.28
Number of observations	5786	5766	5766	5748

The regressions were estimated using the Project STAR public access data set described in Appendix 13.1. The dependent variable is the student's combined test score on the math and reading portions of the Stanford Achievement Test. All regressions include an intercept (not reported). The number of observations differs in the different regressions because of some missing data. Standard errors, clustered at the school level, appear in parentheses, and 95% confidence intervals appear in brackets.

the random assignment to the smaller classes also does not depend on unobserved variables. As expected, these additional regressors increase the $\overline{R}^2$ of the regression, and the standard error of the estimated class size effect decreases from 4.23 in column (1) to 3.95 in column (4).

Because teachers were randomly assigned to class types within a school, the experiment also provides an opportunity to estimate the effect on test scores of teacher experience. In the terminology of Section 13.1, randomization is conditional on the covariates W, where W denotes a full set of binary variables indicating each school; that is, W denotes a full set of school fixed effects. Thus, conditional on W, years of experience is randomly assigned, which in turn implies that u_i in Equation (13.2) satisfies conditional mean independence, where the X variables are the class size treatments and the teacher's years of experience and W is the full set of school fixed effects. Because teachers were not reassigned randomly across schools, without school fixed effects in the regression [Table 13.2, column (2)] years of experience will, in general, be correlated with the error term; for example, wealthier districts might have teachers with more years of experience. When school effects are included, the estimated coefficient on experience is cut in half, from 1.47 in column (2) of Table 13.2

to 0.74 in column (3). Because teachers were randomly assigned within a school, column (3) produces an unbiased estimator of the effect on test scores of an additional year of experience. The estimate, 0.74, is moderately large, although imprecisely estimated: Ten years of experience corresponds to a predicted increase in test scores of 7.4 points, with a 95% confidence interval of (0.4, 14.5).

It is tempting to interpret some of the other coefficients in Table 13.2 but, like coefficients on control variables generally, those coefficients do not have a causal interpretation.

Interpreting the estimated effects of class size. Are the estimated effects of class size reported in Tables 13.1 and 13.2 large or small in a practical sense? There are two ways to answer this: first, by translating the estimated changes in raw test scores into units of standard deviations of test scores, so that the estimates in Table 13.1 are comparable across grades; and, second, by comparing the estimated class size effect to the other coefficients in Table 13.2.

Because the distribution of test scores is not the same for each grade, the estimated effects in Table 13.1 are not directly comparable across grades. We faced this problem in Section 9.4, when we wanted to compare the effect on test scores of a reduction in the student–teacher ratio estimated using data from California to the effect estimated using data from Massachusetts. Because the two tests differed, the coefficients could not be compared directly. The solution in Section 9.4 was to translate the estimated effects into units of standard deviations of the test, so that a unit decrease in the student–teacher ratio corresponds to a change of an estimated fraction of a standard deviation of test scores. We adopt this approach here so that the estimated effects in Table 13.1 can be compared across grades. For example, the standard deviation of test scores for children in kindergarten is 73.75, so the effect of being in a small class in kindergarten, based on the estimate in Table 13.1, is $13.9/73.75 = 0.19$, with a standard error of $4.23/73.75 = 0.06$.

The estimated effects of class size from Table 13.1, converted into units of the standard deviation of test scores across students, are summarized in Table 13.3. Expressed in standard deviation units, the estimated effect of being in a small class is similar for grades K, 2, and 3 and is approximately one-fifth of a standard deviation of test scores. Similarly, the result of being in a regular-sized class with an aide is approximately 0 for grades K, 2, and 3. The estimated treatment effects are larger for first grade; however, the estimated difference between the small class and the regular-sized class with an aide is 0.20 for first grade, the same as for the other grades. Thus one interpretation of the first-grade results is that the students in the control group—the regular-sized class without an aide—happened to do poorly on the test that year for some unusual reason, perhaps simply random sampling variation.

Another way to gauge the magnitude of the estimated effect of being in a small class is to compare the estimated treatment effects with the other coefficients in Table 13.2. In kindergarten, the estimated effect of being in a small class is 13.9 points on the test (first row of Table 13.2). Holding constant race, teacher's years of experience,

TABLE 13.3	Estimated Class Size Effects in Units of Standard Deviations of the Test Score Across Students			
	Grade			
Treatment Group	**K**	**1**	**2**	**3**
Small class	0.19 (0.06)	0.33 (0.05)	0.23 (0.06)	0.21 (0.06)
Regular-sized class with aide	0.00 (0.05)	0.13 (0.05)	0.04 (0.06)	0.00 (0.06)
Sample standard deviation of test scores (s_Y)	73.75	91.25	84.08	73.27

The estimates and standard errors in the first two rows are the estimated effects in Table 13.1, divided by the sample standard deviation of the Stanford Achievement Test for that grade (the final row in this table), computed using data on the students in the experiment. Standard errors, clustered at the school level, appear in parentheses.

eligibility for free lunch, and the treatment group, boys score lower on the standardized test than girls by approximately 12 points, according to the estimates in column (4) of Table 13.2. Thus the estimated effect of being in a small class is somewhat larger than the performance gap between girls and boys. As another comparison, the estimated coefficient on the teacher's years of experience in column (4) is 0.66, so having a teacher with 20 years of experience is estimated to improve test performance by 13 points. Thus the estimated effect of being in a small class is approximately the same as the effect of having a 20-year veteran as a teacher relative to having a new teacher. These comparisons suggest that the estimated effect of being in a small class is meaningfully large.

Additional results.   Econometricians, statisticians, and specialists in elementary education have studied this experiment extensively, and we briefly summarize some of their findings here. One is that the effect of a small class is concentrated in the earliest grades, as can be seen in Table 13.3; except for the anomalous first-grade results, the test score gap between regular-sized and small classes reported in Table 13.3 is essentially constant across grades (0.19 standard deviation units in kindergarten, 0.23 in second grade, and 0.21 in third grade). Because the children initially assigned to a small class stayed in that small class, staying in a small class did not result in additional gains; rather, the gains made upon initial assignment were retained in the higher grades, but the gap between the treatment and control groups did not increase. Another finding is that, as indicated in the second row of Table 13.3, this experiment shows little benefit of having an aide in a regular-sized classroom. One potential concern about interpreting the results of the experiment is the failure to follow the treatment protocol for some students (some students switched from the small classes). If initial placement in a kindergarten classroom is random and has no direct effect on test scores, then initial placement can be used as an instrumental variable that partially, but not entirely, influences placement. This strategy was pursued by Krueger (1999), who used two stage least squares (TSLS) to estimate the effect on test scores of class size using initial classroom placement as the instrumental variable;

he found that the TSLS and OLS estimates were similar, leading him to conclude that deviations from the experimental protocol did not introduce substantial bias into the OLS estimates. An external validity concern about all these results is that they pertain to a narrow measure, test scores at young ages. Chetty et al. (2011) used tax data to examine long-term outcomes for the students in the STAR experiment. Strikingly, they found that students randomly assigned to the small class in kindergarten had higher rates of college attendance than their peers randomly assigned to a regular-sized class.[2]

Comparison of the Observational and Experimental Estimates of Class Size Effects

The Project STAR experiment provides an opportunity that is rare in economics to compare an experimental estimate of a causal effect to estimates made using observational data. Part II presented multiple regression estimates of the class size effect based on observational data for California and Massachusetts school districts. In those data, class size was *not* randomly assigned but instead was determined by local school officials trying to balance educational objectives against budgetary realities. How do those observational estimates compare with the experimental estimates from Project STAR?

To compare the California and Massachusetts estimates with those in Table 13.3, it is necessary to consider the same class size reduction and to express the predicted effect in comparable units, such as standard deviations of test scores. Over the four years of the STAR experiment, the small classes had, on average, approximately 7.5 fewer students than the regular-sized classes, so we use the observational estimates to predict the effect on test scores of a reduction of 7.5 students per class. Based on the OLS estimates for the linear specifications summarized in the first column of Table 9.3, the California estimates predict an increase of 5.5 points on the test for a 7.5 student reduction in the student–teacher ratio ($0.73 \times 7.5 \cong 5.5$ points). The standard deviation of the test across students in California is approximately 38 points, so the estimated effect of the reduction of 7.5 students, expressed in units of standard deviations across students, is $5.5/38 \cong 0.14$ standard deviations.[3] The standard error of the estimated slope coefficient for California is 0.26 (Table 9.3), so the standard error of the estimated effect of a 7.5 student reduction in standard deviation units is

---

[2]For further reading about Project STAR, see Mosteller (1995), Mosteller, Light, and Sachs (1996), and Krueger (1999). Ehrenberg et al. (2001a, 2001b) discuss Project STAR and place it in the context of the policy debate on class size and related research on the topic. For some criticisms of Project STAR, see Hanushek (1999a), and for a critical view of the relationship between class size and performance more generally, see Hanushek (1999b).

[3]In Table 9.3, the estimated effects are presented in terms of the standard deviation of test scores across *districts*; in Table 13.3, the estimated effects are presented in terms of the standard deviation of test scores across *students*. The standard deviation across students is greater than the standard deviation across districts. For California, the standard deviation across students is 38, but the standard deviation across districts is 19.1.

| | | Change in | Standard Deviation | | |
| | | Student–Teacher | of Test Scores Across | Estimated | 95% Confidence |
Study	$\hat{\beta}_1$	Ratio	Students	Effect	Interval
STAR (grade K)	−13.90	Small class vs.	73.8	0.19	[0.08, 0.30]
	(4.23)	regular-sized class		(0.06)	
California	−0.73	−7.5	38.0	0.14	[0.04, 0.24]
	(0.26)			(0.05)	
Massachusetts	−0.64	−7.5	39.0	0.12	[0.02, 0.22]
	(0.27)			(0.05)	

TABLE 13.4   Estimated Effects of Reducing the Student–Teacher Ratio by 7.5 Based on the STAR Data and the California and Massachusetts Observational Data

The estimated coefficient $\hat{\beta}_1$ for the STAR study is taken from column (1) of Table 13.2. The estimated coefficients for the California and Massachusetts studies are taken from the first column of Table 9.3. The estimated effect is the effect of being in a small class versus a regular-sized class (for STAR) or the effect of reducing the student–teacher ratio by 7.5 (for the California and Massachusetts studies). The 95% confidence interval for the reduction in the student–teacher ratio is this estimated effect ±1.96 standard errors. Standard errors are given in parentheses under estimated effects.

$0.26 \times 7.5/38 \cong 0.05$. Thus, based on the California data, the estimated effect of reducing classes by 7.5 students, expressed in units of standard deviations of test scores across students, is 0.14 standard deviations, with a standard error of 0.05. These calculations and similar calculations for Massachusetts are summarized in Table 13.4, along with the STAR estimates for kindergarten taken from column (1) of Table 13.2.

The estimated effects from the California and Massachusetts observational studies are somewhat smaller than the STAR estimates. One reason that estimates from different studies differ, however, is random sampling variability, so it makes sense to compare confidence intervals for the estimated effects from the three studies. Based on the STAR data for kindergarten, the 95% confidence interval for the effect of being in a small class (reported in the final column of Table 13.4) is 0.08 to 0.30. The comparable 95% confidence interval based on the California observational data is 0.04 to 0.24, and for Massachusetts, it is 0.02 to 0.22. Thus the 95% confidence intervals from the California and Massachusetts studies contain most of the 95% confidence interval from the STAR kindergarten data. Viewed in this way, the three studies give strikingly similar ranges of estimates.

There are many reasons the experimental and observational estimates might differ. One reason is that, as discussed in Section 9.4, there are remaining threats to the internal validity of the observational studies. For example, because children move into and out of districts, the district student–teacher ratio might not reflect the student–teacher ratio actually experienced by the students, so the coefficient on the student–teacher ratio in the Massachusetts and California studies could be biased toward 0 because of errors-in-variables bias. In addition, the district average student–teacher ratio used in the observational studies is not the same thing as the actual number of children actually in a class, the STAR experimental variable. Other reasons concern external validity. Project STAR was conducted in a southern state in

the 1980s, potentially different from California and Massachusetts in the late 1990s, and the grades being compared differ (K through 3 in STAR, fourth grade in Massachusetts, and fifth grade in California). In light of all these reasons to expect different estimates, the findings of the three studies are remarkably similar. That the estimates from the observational studies are similar to the Project STAR estimates suggests that the remaining threats to the internal validity of the observational estimates are minor.

13.4   Quasi-Experiments

The statistical insights and methods of randomized controlled experiments can carry over to nonexperimental settings. In a **quasi-experiment**, also called a **natural experiment**, randomness is introduced by variations in individual circumstances that make it appear *as if* the treatment is randomly assigned. These variations in individual circumstances might arise because of vagaries in legal institutions, location, timing of policy or program implementation, natural randomness such as birth dates, rainfall, or other factors that are unrelated to the causal effect under study.

We consider two types of quasi-experiments. In the first, whether an individual (more generally, an entity) receives treatment is viewed as if it is randomly determined. In this case, the causal effect can be estimated by OLS using the treatment, X_i, as a regressor. In the second type of quasi-experiment, the as-if random variation only partially determines the treatment. In this case, the causal effect is estimated by instrumental variables regression, where the as-if random source of variation provides the instrumental variable.

After providing some examples, this section presents some extensions of the econometric methods in Sections 13.1 and 13.2 that can be useful for analyzing data from quasi-experiments.

Examples

We illustrate these two types of quasi-experiments by examples. The first example is a quasi-experiment in which the treatment is as-if randomly determined. The second and third examples illustrate quasi-experiments in which the as-if random variation influences, but does not entirely determine, the level of the treatment.

Example 1: Labor market effects of immigration.   Does immigration reduce wages? Economic theory suggests that if the supply of labor increases because of an influx of immigrants, the "price" of labor—the wage—should fall. However, all else being equal, immigrants are attracted to cities with high labor demand, so the OLS estimator of the effect on wages of immigration will be biased. An ideal randomized controlled experiment for estimating the effect on wages of immigration would randomly assign different numbers of immigrants (different "treatments") to different labor

markets ("subjects") and measure the effect on wages (the "outcome"). Such an experiment, however, faces severe practical, financial, and ethical problems.

The labor economist David Card (1990) therefore used a quasi-experiment in which a large number of Cuban immigrants entered the Miami, Florida, labor market in the Mariel boatlift, which resulted from a temporary lifting of restrictions on emigration from Cuba in 1980. Half of the immigrants settled in Miami, in part because it had a large preexisting Cuban community. Card estimated the causal effect on wages of an increase in immigration by comparing the change in wages of low-skilled workers in Miami to the change in wages of similar workers in comparable U.S. cities over the same period. He concluded that this influx of immigrants had a negligible effect on wages of less-skilled workers.

Example 2: Effects on civilian earnings of military service.   Does serving in the military improve your subsequent labor market prospects? The military provides training that future employers might find attractive. However, an OLS regression of individual civilian earnings against prior military service could produce a biased estimator of the effect on civilian earnings of military service because military service is determined, at least in part, by individual choices and characteristics.

To circumvent this selection bias, Joshua Angrist (1990) used a quasi-experimental design in which he examined labor market histories of those who served in the U.S. military during the Vietnam War. During this period, whether a young man was drafted into the military was determined in part by a national lottery system based on birthdays: Men randomly assigned low lottery numbers were eligible to be drafted, whereas those with high numbers were not. Actual entry into the military was determined by complicated rules, including physical screening and certain exemptions, and some young men volunteered for service, so serving in the military was only partially influenced by whether a man was draft-eligible. Thus being draft-eligible serves as an instrumental variable that partially determines military service but is randomly assigned. In this case, there was true random assignment of draft eligibility via the lottery, but because this randomization was not done as part of an experiment to evaluate the effect of military service, it is a quasi-experiment. Angrist concluded that the long-term effect of military service was to reduce earnings of white, but not nonwhite, veterans.

Example 3: The effect of cardiac catheterization.   Section 12.5 described the study by McClellan, McNeil, and Newhouse (1994), in which they used the distance from a heart attack patient's home to a cardiac catheterization hospital, relative to the distance to a hospital lacking catheterization facilities, as an instrumental variable for actual treatment by cardiac catheterization. This study is a quasi-experiment with a variable that partially determines the treatment. The treatment itself, cardiac catheterization, is determined by personal characteristics of the patient and by the decision of the patient and doctor; however, it is also influenced by whether a nearby hospital is capable of performing this procedure. If the location of the patient is as-if

randomly assigned and has no direct effect on health outcomes other than through its effect on the probability of catheterization, then the relative distance to a catheterization hospital is a valid instrumental variable.

The Differences-in-Differences Estimator

If the treatment in a quasi-experiment is as-if randomly assigned, conditional on some observed variables W, then the treatment effect can be estimated using the differences regression in Equation (13.2). Because the researcher does not have control over the randomization, however, some differences might remain between the treatment and control groups even after controlling for W. One way to adjust for those remaining differences between the two groups is to compare not the outcomes Y but the *change* in the outcomes pre- and posttreatment, thereby adjusting for differences in pretreatment values of Y in the two groups. Because this estimator is the difference across groups in the change, or difference over time, it is called the differences-in-differences estimator. For example, in his study of the effect of immigration on low-skilled workers' wages, Card (1990) used a differences-in-differences estimator to compare the *change* in wages in Miami with the *change* in wages in other U.S. cities.

The differences-in-differences estimator. Let $\overline{Y}^{treatment,\,before}$ be the sample average of Y for those in the treatment group before the experiment, and let $\overline{Y}^{treatment,\,after}$ be the sample average for the treatment group after the experiment. Let $\overline{Y}^{control,\,before}$ and $\overline{Y}^{control,\,after}$ be the corresponding pretreatment and post-treatment sample averages for the control group. The average change in Y over the course of the experiment for those in the treatment group is $\overline{Y}^{treatment,\,after} - \overline{Y}^{treatment,\,before}$, and the average change in Y over this period for those in the control group is $\overline{Y}^{control,\,after} - \overline{Y}^{control,\,before}$. The **differences-in-differences estimator** is the average change in Y for those in the treatment group minus the average change in Y for those in the control group:

$$
\begin{aligned}
\hat{\beta}_1^{diffs\text{-}in\text{-}diffs} &= \left(\overline{Y}^{treatment,\,after} - \overline{Y}^{treatment,\,before}\right) - \left(\overline{Y}^{control,\,after} - \overline{Y}^{control,\,before}\right) \\
&= \Delta\overline{Y}^{treatment} - \Delta\overline{Y}^{control},
\end{aligned}
\tag{13.4}
$$

where $\Delta\overline{Y}^{treatment}$ is the average change in Y in the treatment group and $\Delta\overline{Y}^{control}$ is the average change in Y in the control group. If the treatment is randomly assigned, then $\hat{\beta}_1^{diffs\text{-}in\text{-}diffs}$ is an unbiased and consistent estimator of the causal effect.

The differences-in-differences estimator can be written in regression notation. Let ΔY_i be the postexperimental value of Y for the ith individual minus the preexperimental value. The differences-in-differences estimator is the OLS estimator of β_1 in the regression

$$
\Delta Y_i = \beta_0 + \beta_1 X_i + u_i.
\tag{13.5}
$$

FIGURE 13.1   The Differences-in-Differences Estimator

The posttreatment difference between the treatment and control groups is $80 - 30 = 50$, but this overstates the treatment effect because before the treatment $\bar{Y}$ was higher for the treatment group than the control group by $40 - 20 = 20$. The differences-in-differences estimator is the difference between the final and initial gaps, so $\hat{\beta}_1^{diffs\text{-}in\text{-}diffs} = (80 - 30) - (40 - 20) = 50 - 20 = 30$. Equivalently, the differences-in-differences estimator is the average change for the treatment group minus the average change for the control group; that is, $\hat{\beta}_1^{diffs\text{-}in\text{-}diffs} = \Delta\bar{Y}^{treatment} - \Delta\bar{Y}^{control} = (80 - 40) - (30 - 20) = 30$.
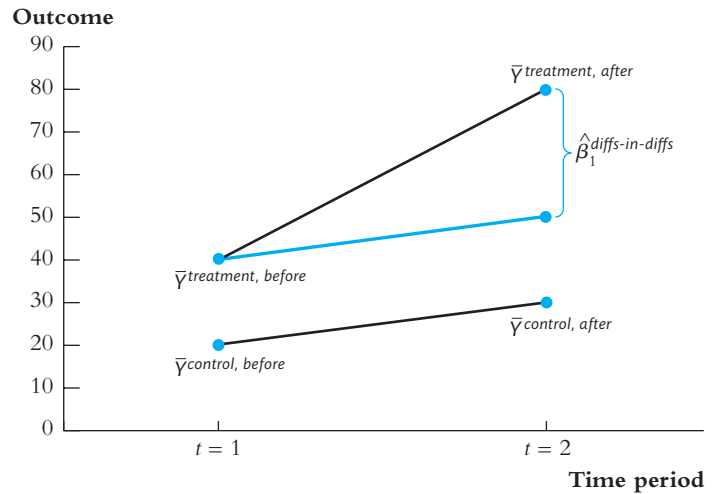
The differences-in-differences estimator is illustrated in Figure 13.1. In that figure, the sample average of Y for the treatment group is 40 before the experiment, whereas the pretreatment sample average of Y for the control group is 20. Over the course of the experiment, the sample average of Y increases in the control group to 30, whereas it increases to 80 for the treatment group. Thus the mean difference of the posttreatment sample averages is $80 - 30 = 50$. However, some of this difference arises because the treatment and control groups had different pretreatment means: The treatment group started out ahead of the control group. The differences-in-differences estimator measures the gains of the treatment group relative to the control group, which in this example is $(80 - 40) - (30 - 20) = 30$. By focusing on the change in Y over the course of the experiment, the differences-in-differences estimator removes the influence of initial values of Y that vary between the treatment and control groups.

The differences-in-differences estimator with additional regressors. The differences-in-differences estimator can be extended to include additional regressors $W_{1i}, \ldots, W_{ri}$. These variables can be individual characteristics prior to the experiment, or they can be control variables. These additional regressors can be incorporated using the multiple regression model

$$\Delta Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i, \, i = 1, \ldots, n. \quad (13.6)$$

The OLS estimator of β_1 in Equation (13.6) is the **differences-in-differences estimator with additional regressors**. If X_i is as-if randomly assigned, conditional on $W_{1i}, \ldots, W_{ri}$,

then u_i satisfies conditional mean independence, and the OLS estimator of $\hat{\beta}_1$ in Equation (13.6) is unbiased.

The differences-in-differences estimator described here considers two time periods, before and after the experiment. In some settings, there are panel data with multiple time periods. The differences-in-differences estimator can be extended to multiple time periods using the panel data regression methods of Chapter 10.

Differences-in-differences using repeated cross-sectional data. A **repeated cross-sectional data** set is a collection of cross-sectional data sets, where each cross-sectional data set corresponds to a different time period. For example, the data set might contain observations on 400 individuals in the year 2004 and on 500 different individuals in 2005, for a total of 900 different individuals. One example of repeated cross-sectional data is political polling data, in which political preferences are measured by a series of surveys of randomly selected potential voters, where the surveys are taken at different dates and each survey has different respondents.

The premise of using repeated cross-sectional data is that if the individuals (more generally, entities) are randomly drawn from the same population, then the individuals in the earlier cross section can be used as surrogates for the individuals in the treatment and control groups in the later cross section.

When there are two time periods, the regression model for repeated cross-sectional data is

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 G_i + \beta_3 D_t + \beta_4 W_{1it} + \cdots + \beta_{3+r} W_{rit} + u_{it}, \quad (13.7)$$

where X_{it} is the actual treatment of the ith individual (entity) in the cross section in period t $(t = 1, 2)$, G_i is a binary variable indicating whether the individual is in the treatment group (or in the surrogate treatment group if the observation is in the pretreatment period), and D_t is the binary indicator that equals 0 in the first period and equals 1 in the second period. The ith individual receives treatment if he or she is in the treatment group in the second period, so in Equation (13.7), $X_{it} = G_i \times D_t$; that is, X_{it} is the interaction between G_i and D_t.

If the quasi-experiment makes X_{it} as-if randomly received, conditional on the W's, then the causal effect can be estimated by the OLS estimator of β_1 in Equation (13.7). If there are more than two time periods, then Equation (13.7) is modified to contain $T - 1$ binary variables indicating the different time periods (see Section 10.4).

Instrumental Variables Estimators

If the quasi-experiment yields a variable Z_i that influences receipt of treatment, if data are available both on Z_i and on the treatment actually received (X_i), and if Z_i is as-if randomly assigned (perhaps after controlling for some additional variables W_i), then Z_i is a valid instrument for X_i, and the coefficients of Equation (13.2) can be estimated using two stage least squares. Any control variables appearing in Equation (13.2) also appear as control variables in the first stage of the two stage least squares estimator of β_1.

FIGURE 13.2 A Hypothetical Regression Discontinuity Design Scatterplot

Suppose that the binary treatment X is required if W is less than the threshold value $w_0 = 2$. As long as the only role of the threshold w_0 is to mandate treatment, the treatment effect is given by the magnitude of the jump, or discontinuity, in the regression function at $W = 2$.
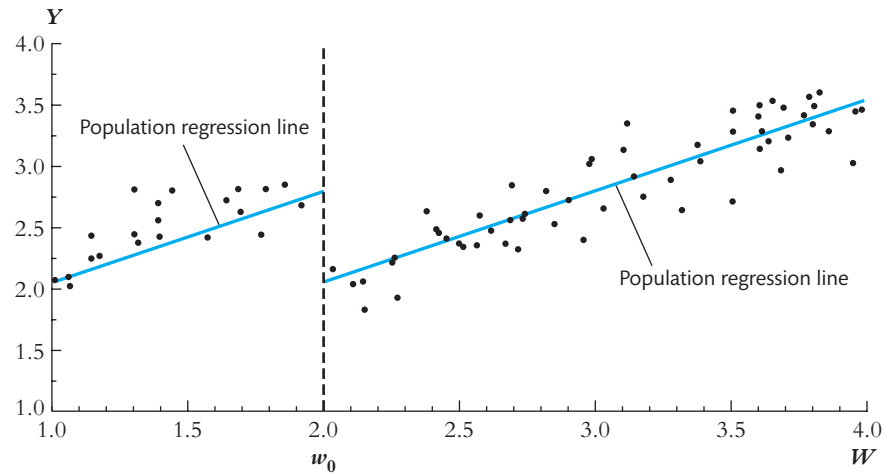
Regression Discontinuity Estimators

One situation that gives rise to a quasi-experiment is when receipt of the treatment depends in whole or in part on whether an observable variable W crosses a threshold value. For example, suppose that students are required to attend summer school if their end-of-year grade point average (GPA) falls below a threshold.[4] Then one way to estimate the effect of mandatory summer school is to compare outcomes for students whose GPA was just below the threshold (and thus were required to attend) to outcomes for students whose GPA was just above the threshold (so they escaped summer school). The outcome Y could be next year's GPA, whether the student drops out, or future earnings. As long as there is nothing special about the threshold value other than its use in mandating summer school, it is reasonable to attribute any jump in outcomes at that threshold to summer school. Figure 13.2 illustrates a hypothetical scatterplot of a data set in which the treatment (summer school, X) is required if GPA (W) is less than a threshold value ($w_0 = 2.0$). The scatterplot shows next year's GPA (Y) for a hypothetical sample of students as a function of this year's GPA, along with the population regression function. If the only role of the threshold w_0 is to mandate summer school, then the jump in next year's GPA at w_0 is an estimate of the effect of summer school on next year's GPA.

Because of the jump, or discontinuity, in treatment at the threshold, studies that exploit a discontinuity in the probability of receiving treatment at a threshold value are called **regression discontinuity** designs. There are two types of regression discontinuity designs, sharp and fuzzy.

---

[4]This example is a simplified version of the regression discontinuity study of the effect of summer school for elementary and middle school students by Jordan Matsudaira (2008), in which summer school attendance was based in part on end-of-year tests.

Sharp regression discontinuity design. In a sharp regression discontinuity design, receipt of treatment is entirely determined by whether W exceeds the threshold: All students with $W < w_0$ attend summer school, and no students with $W \geq w_0$ attend; that is, $X_i = 1$ if $W < w_0$, and $X_i = 0$ if $W \geq w_0$. In this case, the jump in Y at the threshold equals the average treatment effect for the subpopulation with $W = w_0$, which might be a useful approximation to the average treatment effect in the larger population of interest. If the regression function is linear in W, other than for the treatment-induced discontinuity, the treatment effect can be estimated by β_1 in the regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i. \tag{13.8}$$

If the regression function is nonlinear, then a suitable nonlinear function of W can be used (Section 8.2).

Fuzzy regression discontinuity design. In a fuzzy regression discontinuity design, crossing the threshold influences receipt of the treatment but is not the sole determinant. For example, suppose that some students whose GPA falls below the threshold are exempted from summer school while some whose GPA exceeds the threshold nevertheless attend. This situation could arise if the threshold rule is part of a more complicated process for determining treatment. In a fuzzy design, X_i will, in general, be correlated with u_i in Equation (13.8). If, however, any special effect of crossing the threshold operates solely by increasing the probability of treatment—that is, the direct effect of crossing the threshold is captured by the linear term in W—then an instrumental variables approach is available. Specifically, let the binary variable Z_i indicate crossing the threshold (so $Z_i = 1$ if $W_i < w_0$ and $Z_i = 0$ if $W_i \geq w_0$). Then Z_t influences receipt of treatment but is uncorrelated with u_i, so it is a valid instrument for X_i. Thus, in a fuzzy regression discontinuity design, β_1 can be estimated by instrumental variables estimation of Equation (13.8), using as an instrument the binary variable indicating that $W_i < w_0$.

13.5 Potential Problems with Quasi-Experiments

Like all empirical studies, quasi-experiments face threats to internal and external validity. A particularly important potential threat to internal validity is whether the as-if randomization, in fact, can be treated reliably as true randomization.

Threats to Internal Validity

The threats to the internal validity of true randomized controlled experiments listed in Section 13.2 also apply to quasi-experiments but with some modifications.

Failure of randomization. Quasi-experiments rely on differences in individual circumstances—legal changes, sudden unrelated events, and so forth—to provide the as-if randomization in the treatment level. If this as-if randomization fails to produce

a treatment level X (or an instrumental variable Z) that is random, then, in general, the OLS estimator is biased (or the instrumental variable estimator is not consistent).

As in a true experiment, one way to test for failure of randomization is to check for systematic differences between the treatment and control groups, for example by regressing X (or Z) on the individual characteristics (the W's) and testing the hypothesis that the coefficients on the W's are 0. If differences exist that are not readily explained by the nature of the quasi-experiment, then that is evidence that the quasi-experiment did not produce true randomization. Even if there is no relationship between X (or Z) and the W's, the possibility remains that X (or Z) could be related to some of the unobserved factors in the error term u. Because these factors are unobserved, this possibility cannot be tested, and the validity of the assumption of as-if randomization must be evaluated using expert knowledge and judgment applied to the application at hand.

Failure to follow the treatment protocol. In a true experiment, failure to follow the treatment protocol arises when members of the treatment group fail to receive treatment, members of the control group actually receive treatment, or both; in consequence, the OLS estimator of the causal effect has selection bias. The counterpart to failing to follow the treatment protocol in a quasi-experiment is when the as-if randomization influences, but does not determine, the treatment level. In this case, the instrumental variables estimator based on the quasi-experimental influence Z can be consistent even though the OLS estimator is not.

Attrition. Attrition in a quasi-experiment is similar to attrition in a true experiment in the sense that if attrition arises because of personal choices or characteristics, then it can induce correlation between the treatment level and the error term. The result is sample selection bias, so the OLS estimator of the causal effect is biased and inconsistent.

Experimental effects. An advantage of quasi-experiments is that because they are not true experiments, there typically is no reason for individuals to think that they are experimental subjects. Thus experimental effects such as the Hawthorne effect generally are not germane in quasi-experiments.

Instrument validity in quasi-experiments. An important step in evaluating a study that uses instrumental variables regression is careful consideration of whether the instrument is in fact valid. This general statement remains true in quasi-experimental studies in which the instrument is as-if randomly determined. As discussed in Chapter 12, instrument validity requires both instrument relevance and instrument exogeneity. Because instrument relevance can be checked using the statistical methods summarized in Key Concept 12.5, here we focus on the second, more judgmental requirement of instrument exogeneity.

Although it might seem that a randomly assigned instrumental variable is necessarily exogenous, that is not so. Consider the examples of Section 13.4. In Angrist's (1990) use of draft lottery numbers as an instrumental variable in studying the effect

on civilian earnings of military service, the lottery numbers were, in fact, randomly assigned. But as Angrist (1990) points out and discusses, if a low draft number results in behavior aimed at avoiding the draft and that avoidance behavior subsequently affects civilian earnings, then a low lottery number (Z_i) could be related to unobserved factors that determine civilian earnings (u_i); that is, Z_i and u_i are correlated even though Z_i is randomly assigned. As a second example, McClellan, McNeil, and Newhouse's (1994) study of the effect on heart attack patients of cardiac catheterization treated the relative distance to a catheterization hospital as if it were randomly assigned. But as the authors highlight and examine, if patients who live close to a catheterization hospital are healthier than those who live far away (perhaps because of better access to medical care generally), then the relative distance to a catheterization hospital would be correlated with omitted variables in the error term of the health outcome equation. In short, just because an instrument is randomly determined or as-if randomly determined does not necessarily mean it is exogenous in the sense that $\text{corr}(Z_i, u_i) = 0$. Thus the case for exogeneity must be scrutinized closely even if the instrument arises from a quasi-experiment.

Threats to External Validity

Quasi-experimental studies use observational data, and the threats to the external validity of a study based on a quasi-experiment are generally similar to the threats discussed in Section 9.1 for conventional regression studies using observational data.

One important consideration is that the special events that create the as-if randomness at the core of a quasi-experimental study can result in other special features that threaten external validity. For example, Card's (1990) study of labor market effects of immigration discussed in Section 13.4 used the as-if randomness induced by the influx of Cuban immigrants in the Mariel boatlift. There were, however, special features of the Cuban immigrants, Miami, and its Cuban community that might make it difficult to generalize these findings to immigrants from other countries or to other destinations. Similarly, Angrist's (1990) study of the labor market effects of serving in the U.S. military during the Vietnam War presumably would not generalize to peacetime military service. As usual, whether a study generalizes to a specific population and setting of interest depends on the details of the study and must be assessed on a case-by-case basis.

13.6 Experimental and Quasi-Experimental Estimates in Heterogeneous Populations

As discussed in Section 13.1, the causal effect can vary from one member of the population to the next. Section 13.1 discusses estimating causal effects that vary depending on observable variables, such as sex. In this section, we consider the consequences of *unobserved* variation in the causal effect. We refer to unobserved variation in the causal effect as having a heterogeneous population. To keep things simple

and to focus on the role of unobserved heterogeneity, in this section we omit control variables W; the conclusions of this section carry over to regressions including control variables.

If the population is heterogeneous, then the ith individual now has his or her own causal effect, β_{1i}, which (in the terminology of Section 13.1) is the difference in the ith individual's potential outcomes if the treatment is or is not received. For example, β_{1i} might be 0 for a resume-writing training program if the ith individual already knows how to write a resume. With this notation, the population regression equation can be written

$$Y_i = \beta_0 + \beta_{1i}X_i + u_i. \tag{13.9}$$

Appendix 13.3 derives Equation (13.9) from the potential outcomes framework for a heterogeneous population. Because β_{1i} varies from one individual to the next in the population and the individuals are selected from the population at random, β_{1i} is a random variable that, just like u_i, reflects unobserved variation across individuals (for example, variation in preexisting resume-writing skills). The average causal effect is the population mean value of the causal effect, $E(\beta_{1i})$; that is, it is the expected causal effect of a randomly selected member of the population under study.

What do the estimators of Sections 13.1, 13.2, and 13.4 estimate if there is population heterogeneity of the form in Equation (13.9)? We first consider the OLS estimator when X_i is as-if randomly determined; in this case, the OLS estimator is a consistent estimator of the average causal effect. That is generally not true for the IV estimator, however. Instead, if X_i is partially influenced by Z_i, then the IV estimator using the instrument Z estimates a weighted average of the causal effects, where those for whom the instrument is most influential receive the most weight.

OLS with Heterogeneous Causal Effects

If there is heterogeneity in the causal effect and if X_i is randomly assigned, then the differences estimator is a consistent estimator of the average causal effect. This result follows from the discussion in Section 13.1 and Appendix 13.3, which make use of the potential outcome framework; here it is shown without reference to potential outcomes by applying concepts from Chapters 3 and 4 directly to the random coefficients regression model in Equation (13.9).

The OLS estimator of β_1 in Equation (13.1) is $\hat{\beta}_1 = s_{XY}/s_X^2$ [Equation (4.5)]. If the observations are i.i.d., then the sample covariance and variance are consistent estimators of the population covariance and variance, so $\hat{\beta}_1 \xrightarrow{P} \sigma_{XY}/\sigma_X^2$. If X_i is randomly assigned, then X_i is distributed independently of other individual characteristics, both observed and unobserved, and in particular is distributed independently of β_{1i}. Accordingly, the OLS estimator $\hat{\beta}_1$ has the limit

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} \xrightarrow{P} \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\text{cov}(\beta_0 + \beta_{1i}X_i + u_i, X_i)}{\sigma_X^2}$$
$$= \frac{\text{cov}(\beta_{1i}X_i, X_i)}{\sigma_X^2} = E(\beta_{1i}), \tag{13.10}$$

where the third equality uses the facts about covariances in Key Concept 2.3 and $\text{cov}(u_i, X_i) = 0$, which is implied by $E(u_i|X_i) = 0$ [Equation (2.28)], and where the final equality follows from β_{1i} being distributed independently of X_i, which it is if X_i is randomly assigned (Exercise 13.9). Thus, if X_i is randomly assigned, $\hat{\beta}_1$ is a consistent estimator of the average causal effect $E(\beta_{1i})$.

IV Regression with Heterogeneous Causal Effects

Suppose that the causal effect is estimated by instrumental variables regression of Y_i on X_i (treatment actually received) using Z_i (initial randomly or as-if randomly assigned treatment) as an instrument. Suppose that Z_i is a valid instrument (relevant and exogenous) and that there is heterogeneity in the effect on X_i of Z_i. Specifically, suppose that X_i is related to Z_i by the linear model

$$X_i = \pi_0 + \pi_{1i}Z_i + v_i, \tag{13.11}$$

where the coefficient π_{1i} varies from one individual to the next. Equation (13.11) is the first-stage equation of TSLS with the modification that the effect on X_i of a change in Z_i is allowed to vary from one individual to the next.

The TSLS estimator is $\hat{\beta}_1^{TSLS} = s_{ZY}/s_{ZX}$ [Equation (12.4)], the ratio of the sample covariance between Z and Y to the sample covariance between Z and X. If the observations are i.i.d., then these sample covariances are consistent estimators of the population covariances, so $\hat{\beta}_1^{TSLS} \xrightarrow{P} \sigma_{ZY}/\sigma_{ZX}$. Suppose that the instrument Z_i is randomly assigned or as-if randomly assigned, so that Z_i is distributed independently of $(u_i, v_i, \pi_{1i}, \beta_{1i})$, and that $E(\pi_{1i}) \neq 0$ (instrument relevance). It is shown in Appendix 13.2 that, under these assumptions,

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{P} \frac{\sigma_{ZY}}{\sigma_{ZX}} = \frac{E(\beta_{1i}\pi_{1i})}{E(\pi_{1i})}. \tag{13.12}$$

That is, the TSLS estimator converges in probability to the ratio of the expected value of the product of β_{1i} and π_{1i} to the expected value of π_{1i}.

The final ratio in Equation (13.12) is a weighted average of the individual causal effects β_{1i}. The weights are $\pi_{1i}/E(\pi_{1i})$, which measure the relative degree to which the instrument influences whether the ith individual receives treatment. Thus the TSLS estimator is a consistent estimator of a weighted average of the individual causal effects, where the individuals who receive the most weight are those for whom the instrument is most influential. The weighted average causal effect that is estimated by TSLS is called the **local average treatment effect** (LATE). The term *local* emphasizes that it is the weighted average that places the most weight on those individuals (more generally, entities) whose treatment probability is most influenced by the instrumental variable.

There are three special cases in which the LATE equals the average treatment effect:

1. The treatment effect is the same for all individuals. This case corresponds to $\beta_{1i} = \beta_1$ for all i. Then the final expression in Equation (13.12) simplifies to $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = \beta_1 E(\pi_{1i})/E(\pi_{1i}) = \beta_1$.

2. The instrument affects each individual equally. This case corresponds to $\pi_{1i} = \pi_1$ for all i. In this case, the final expression in Equation (13.12) simplifies to $E(\beta_{1i}\pi_{1i}) / E(\pi_{1i}) = E(\beta_{1i})\pi_1 / \pi_1 = E(\beta_{1i})$.

3. The heterogeneity in the treatment effect and heterogeneity in the effect of the instrument are uncorrelated. This case corresponds to β_{1i} and π_{1i} being random but $\text{cov}(\beta_{1i}, \pi_{1i}) = 0$. Because $E(\beta_{1i}\pi_{1i}) = \text{cov}(\beta_{1i}, \pi_{1i}) + E(\beta_{1i})E(\pi_{1i})$ [Equation (2.35)], if $\text{cov}(\beta_{1i}, \pi_{1i}) = 0$, then $E(\beta_{1i}\pi_{1i}) = E(\beta_{1i})E(\pi_{1i})$, and the final expression in Equation (13.12) simplifies to $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = E(\beta_{1i})E(\pi_{1i})/E(\pi_{1i}) = E(\beta_{1i})$.

In each of these three cases, there is population heterogeneity in the effect of the instrument, in the effect of the treatment, or in both, but the LATE equals the average treatment effect. That is, in all three cases, TSLS is a consistent estimator of the average treatment effect.

Aside from these three special cases, in general, the LATE differs from the average treatment effect. For example, suppose that Z_i has no influence on the treatment decision for half the population (for them, $\pi_{1i} = 0$), while for the other half, Z_i has a common, nonzero influence on the treatment decision (for them, π_{1i} takes on the same nonzero value). Then TSLS is a consistent estimator of the average treatment effect in the half of the population for which the instrument influences the treatment decision. To be concrete, suppose workers are eligible for a job training program and are randomly assigned a priority number Z, which influences how likely they are to be admitted to the program. Half the workers know they will benefit from the program and thus may decide to enroll in the program; for them, $\beta_{1i} = \beta_1^+ > 0$ and $\pi_{1i} = \pi_1^+ > 0$. The other half know that, for them, the program is ineffective, so they would not enroll even if admitted; that is, for them $\beta_{1i} = \beta_1^-$ and $\pi_{1i} = 0$. The average treatment effect is $E(\beta_{1i}) = \frac{1}{2}(\beta_1^+ + \beta_1^-)$. The local average treatment effect is $E(\beta_{1i}\pi_{1i})/E(\pi_{1i})$. Now $E(\pi_{1i}) = \frac{1}{2}\pi_1^+$ and $E(\beta_{1i}\pi_{1i}) = \frac{1}{2}(\beta_1^- \times 0 + \beta_1^+\pi_1^+) = \frac{1}{2}\beta_1^+\pi_1^+$, so $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = \beta_1^+$. Thus in this example the LATE is the causal effect for those workers who might enroll in the program, and it gives no weight to those who will not enroll under any circumstances. In contrast, the average treatment effect places equal weight on all individuals, regardless of whether they would enroll. Because individuals decide to enroll based in part on their knowledge of how effective the program will be for them, in this example the LATE exceeds the average treatment effect.

Implications. If an individual's decision to receive treatment depends on the effectiveness of the treatment for that individual, then the TSLS estimator, in general, is not a consistent estimator of the average causal effect. Instead, TSLS estimates a LATE, where the causal effects of the individuals who are most influenced by the instrument receive the greatest weight.

This conclusion leads to a disconcerting situation in which two researchers, armed with different instrumental variables that are both valid in the sense that both

are relevant and exogenous, would obtain different estimates of "the" causal effect, even in large samples. The difference arises because each researcher is implicitly estimating a different weighted average of the individual causal effects in the population. In fact, a *J*-test of overidentifying restrictions can reject if the two instruments estimate different LATEs, even if both instruments are valid. Although both estimators provide some insight into the distribution of the causal effects via their respective weighted averages of the form in Equation (13.12), in general, neither estimator is a consistent estimator of the average causal effect.[5]

Example: The cardiac catheterization study. Sections 12.5 and 13.4 discuss McClellan, McNeil, and Newhouse's (1991) study of the effect on mortality of cardiac catheterization of heart attack patients. The authors used instrumental variables regression, with the relative distance to a cardiac catheterization hospital as the instrumental variable. Based on their TSLS estimates, they found that cardiac catheterization had little or no effect on health outcomes. This result is surprising: Medical procedures such as cardiac catheterization are subjected to rigorous clinical trials prior to approval for widespread use. Moreover, cardiac catheterization allows surgeons to perform medical interventions that would have required major surgery a decade earlier, making these interventions safer and, presumably, better for long-term patient health. How could this econometric study fail to find beneficial effects of cardiac catheterization?

One possible answer is that there is heterogeneity in the treatment effect of cardiac catheterization. For some patients, this procedure is an effective intervention, but for others, perhaps those who are healthier, it is less effective or, given the risks involved with any surgery, perhaps on the whole ineffective. Thus the average causal effect in the population of heart attack patients could be, and presumably is, positive. The IV estimator, however, measures a marginal effect, not an average effect, where the marginal effect is the effect of the procedure on those patients for whom relative distance to a cardiac catheterization hospital is an important factor in whether they receive treatment. But those patients could be just the relatively healthy patients for whom, on the margin, cardiac catheterization is a relatively ineffective procedure. If so, McClellan, McNeil, and Newhouse's TSLS estimator measures the effect of the procedure for the marginal patient (for whom it is relatively ineffective), not for the average patient (for whom it might be effective).

---

[5]There are several good (but advanced) discussions of the effect of population heterogeneity on program evaluation estimators. They include the survey by Heckman, LaLonde, and Smith (1999, Section 7) and James Heckman's lecture delivered when he received the Nobel Prize in Economics (Heckman, 2001, Section 7). The latter reference and Angrist, Graddy, and Imbens (2000) provide detailed discussion of the random effects model (which treats β_{1i} as varying across individuals) and provide more general versions of the result in Equation (13.12). The concept of the LATE was introduced by Imbens and Angrist (1994), who showed that, in general, it does not equal the average treatment effect. Imbens and Wooldridge (2009) provide an advanced survey of methods for program evaluation with treatment effect heterogeneity, including those discussed in this chapter.

13.7 Conclusion

In Chapter 1, we defined the causal effect in terms of the expected outcome of an ideal randomized controlled experiment. If a randomized controlled experiment is available or can be performed, it can provide compelling evidence on the causal effect under study, although even randomized controlled experiments are subject to potentially important threats to internal and external validity.

Despite their advantages, randomized controlled experiments in economics face considerable hurdles, including ethical concerns and cost. The insights of experimental methods can, however, be applied to quasi-experiments, in which special circumstances make it seem as if randomization has occurred. In quasi-experiments, the causal effect can be estimated using a differences-in-differences estimator, possibly augmented with additional regressors; if the as-if randomization only partly influences the treatment, then instrumental variables regression can be used instead. An important advantage of quasi-experiments is that the source of the as-if randomness in the data is usually transparent and thus can be evaluated in a concrete way. An important threat confronting quasi-experiments is that sometimes the as-if randomization is not really random, so the treatment (or the instrumental variable) is correlated with omitted variables and the resulting estimator of the causal effect is biased.

Quasi-experiments provide a bridge between observational data sets and true randomized controlled experiments. The econometric methods used in this chapter for analyzing quasi-experiments are familiar ones developed in different contexts in earlier chapters: OLS, panel data estimation methods, and instrumental variables regression. What differentiates quasi-experiments from the applications examined in Part II and the earlier chapters in Part III are the way in which these methods are interpreted and the data sets to which they are applied. Quasi-experiments provide econometricians with a way to think about how to acquire new data sets, how to think of instrumental variables, and how to evaluate the plausibility of the exogeneity assumptions that underlie OLS and instrumental variables estimation.[6]

Summary

1. The average causal effect in the population under study is the expected difference in the average outcomes for the treatment and control groups in an ideal randomized controlled experiment. Actual experiments with human subjects deviate from an ideal experiment for various practical reasons, including the failure of people to comply with the experimental protocol.

---

[6]Shadish, Cook, and Campbell (2002) provide a comprehensive treatment of experiments and quasi-experiments in the social sciences and in psychology. An important line of research in development economics focuses on experimental evaluations of health and education programs in developing countries. For examples, see Kremer, Miguel, and Thornton (2009) and the website of MIT's Poverty Action Laboratory (http://www.povertyactionlab.org). Deaton (2010) provides a thoughtful critique of this research.

2. If the *actual* treatment level X_i is random, then the treatment effect can be estimated by regressing the outcome on the treatment. If the *assigned* treatment Z_i is random but the actual treatment X_i is partly determined by individual choice, then the causal effect can be estimated by instrumental variables regression, using Z_i as an instrument. If the treatment (or assigned treatment) is random, conditional on some variables W, those control variables need to be included in the regressions.

3. In a quasi-experiment, variations in laws or circumstances or accidents of nature are treated as if they induce random assignment to treatment and control groups. If the actual treatment is as-if random, then the causal effect can be estimated by regression (possibly with additional pretreatment characteristics as regressors); if the assigned treatment is as-if random, then the causal effect can be estimated by instrumental variables regression.

4. Regression discontinuity estimators are based on quasi-experiments in which treatment depends on whether an observable variable crosses a threshold value.

5. A key threat to the internal validity of a quasi-experimental study is whether the as-if randomization actually results in exogeneity. Because of behavioral responses, the regression error may change in response to the treatment induced by the quasi-experiment, so the treatment is not exogenous.

6. When the treatment effect varies from one individual to the next, the OLS estimator is a consistent estimator of the average causal effect if the actual treatment is randomly assigned or as-if randomly assigned. However, the instrumental variables estimator is a weighted average of the individual treatment effects, where the individuals for whom the instrument is most influential receive the greatest weight.

Key Terms

program evaluation (432)

potential outcome (433)

average causal effect (433)

average treatment effect (433)

differences estimator (434)

differences estimator with additional regressors (435)

randomization based on covariates (435)

test for random receipt of treatment (436)

partial compliance (437)

instrumental variables estimation of the treatment effect (437)

attrition (437)

Hawthorne effect (438)

quasi-experiment (448)

natural experiment (448)

differences-in-differences estimator (450)

differences-in-differences estimator with additional regressors (451)

repeated cross-sectional data (452)

regression discontinuity (453)

local average treatment effect (458)

Review the Concepts

13.1 A researcher studying the effects of a new fertilizer on crop yields plans to carry out an experiment in which different amounts of the fertilizer are applied to 100 different one-acre parcels of land. There will be four treatment levels. Treatment level 1 is no fertilizer, treatment level 2 is 50% of the manufacturer's recommended amount of fertilizer, treatment level 3 is 100%, and treatment level 4 is 150%. The researcher plans to apply treatment level 1 to the first 25 parcels of land, treatment level 2 to the second 25 parcels, and so forth. Can you suggest a better way to assign treatment levels? Why is your proposal better than the researcher's method?

13.2 A clinical trial is carried out for a new cholesterol-lowering drug. The drug is given to 500 patients, and a placebo is given to another 500 patients, using random assignment of the patients. How would you estimate the treatment effect of the drug? Suppose you had data on the weight, age, and sex of each patient. Could you use these data to improve your estimate? Explain. Suppose you had data on the cholesterol level of each patient before he or she entered the experiment. Could you use these data to improve your estimate? Explain.

13.3 Researchers studying the STAR data report anecdotal evidence that school principals were pressured by some parents to place their children in the small classes. Suppose some principals succumbed to this pressure and transferred some children into the small classes. How would such transfers compromise the internal validity of the study? Suppose you had data on the original random assignment of each student before the principal's intervention. How could you use this information to restore the internal validity of the study?

13.4 Explain whether experimental effects (like the Hawthorne effect) might be important in each of the experiments in the previous three questions.

13.5 Consider the quasi-experiment described in Section 13.4 involving the draft lottery, military service, and civilian earnings. Explain why there might be heterogeneous effects of military service on civilian earnings; that is, explain why β_{1i} in Equation (13.9) depends on i. Explain why there might be hetero-

geneous effects of the lottery outcome on the probability of military service; that is, explain why π_{1i} in Equation (13.11) depends on i. If there are heterogeneous responses of the sort you described, what behavioral parameter is being estimated by the TSLS estimator?

Exercises

13.1 Using the results in Table 13.1, calculate the following for each grade: an estimate of the small class treatment effect relative to the regular-sized class; its standard error; and its 90% confidence interval. (For this exercise, ignore the results for regular classes with aides.)

13.2 For the following calculations, use the results in column (4) of Table 13.2. Consider two classrooms, A and B, which have identical values of the regressors in column (4) of Table 13.2 but differ as noted:

   a. Classroom A is a small class, and classroom B is a regular-sized class. Construct a 90% confidence interval for the expected difference in average test scores.

   b. Classroom A has a teacher with 5 years of experience, and classroom B has a teacher with 10 years of experience. Construct a 90% confidence interval for the expected difference in average test scores.

   c. Classroom A is a small class with a teacher with 5 years of experience, and classroom B is a regular-sized class with a teacher with 10 years of experience. Construct a 90% confidence interval for the expected difference in average test scores. (*Hint:* In STAR, the teachers were randomly assigned to the different types of classrooms.)

   d. Why is the intercept missing from column (4)?

13.3 Suppose that, in a randomized controlled experiment of the effect of an SAT preparatory course on SAT scores, the following results are reported:

	Treatment Group	**Control Group**
Average SAT score ($\overline{X}$)	1241	1201
Standard deviation of SAT score (s_X)	93.2	97.1
Number of men	55	45
Number of women	45	55

   a. Estimate the average treatment effect on test scores.

   b. Is there evidence of nonrandom assignment? Explain.

13.4 A new law will increase minimum wages in City A next year but not in City B, a city much like City A. You collect employment data from a random selected

sample of restaurants in cities A and B this year, and you plan to return and collect data at restaurants next year. Let Y_{it} denote the employment level at restaurant i in year t.

a. Suppose you design your analysis so you sample the *same* restaurants this year and next year. Explain how you will use the data to estimate the average causal effect of the minimum wage increase on restaurant employment.

b. Suppose you design your analysis so you sample *different*, independently selected restaurants this year and next year. Explain how you will use the data to estimate the average causal effect of the minimum wage increase on restaurant employment.

c. Which sampling design, using the same restaurants in (a) or using different restaurants in (b), is likely to yield a more precise estimate of the average causal effect? (*Hint:* You might find it useful to solve Exercise 13.6 first.)

13.5 Consider a study to evaluate the effect on college student grades of dorm room Internet connections. In a large dorm, half the rooms are randomly wired for high-speed Internet connections (the treatment group), and final course grades are collected for all residents. Which of the following pose threats to internal validity, and why?

a. Midway through the year all the male athletes move into a fraternity and drop out of the study. (Their final grades are not observed.)

b. Engineering students assigned to the control group put together a local area network so that they can share a private wireless Internet connection that they pay for jointly.

c. The art majors in the treatment group never learn how to access their Internet accounts.

d. The economics majors in the treatment group provide access to their Internet connection to those in the control group, for a fee.

13.6 Suppose there are panel data for $T = 2$ time periods for a randomized controlled experiment, where the first observation ($t = 1$) is taken before the experiment and the second observation ($t = 2$) is for the posttreatment period. Suppose the treatment is binary; that is, suppose $X_{it} = 1$ if the ith individual is in the treatment group and $t = 2$, and $X_{it} = 0$ otherwise. Further suppose the treatment effect can be modeled using the specification

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it},$$

where α_i are individual-specific effects with a mean of 0 and a variance of σ_α^2 and u_{it} is an error term, where u_{it} is homoskedastic, $\text{cov}(u_{i1}, u_{i2}) = 0$, and $\text{cov}(u_{it}, \alpha_i) = 0$ for all i. Let $\hat{\beta}_1^{differences}$ denote the differences estimator— that is, the OLS estimator in a regression of Y_{i2} on X_{i2} with an intercept— and let $\hat{\beta}_1^{diffs-in-diffs}$ denote the differences-in-differences estimator—that is,

the estimator of β_1 based on the OLS regression of $\Delta Y_i = Y_{i2} - Y_{i1}$ against $\Delta X_i = X_{i2} - X_{i1}$ and an intercept.

a. Show that $n \operatorname{var}(\hat{\beta}_1^{differences}) \longrightarrow (\sigma_u^2 + \sigma_\alpha^2)/\operatorname{var}(X_{i2})$. (*Hint:* Use the homoskedasticity-only formulas for the variance of the OLS estimator in Appendix 5.1.)

b. Show that $n \operatorname{var}(\hat{\beta}_1^{diffs - in - diffs}) \longrightarrow 2\sigma_u^2/\operatorname{var}(X_{i2})$. (*Hint:* Note that $X_{i2} - X_{i1} = X_{i2}$. Why?)

c. Based on your answers to (a) and (b), when would you prefer the differences-in-differences estimator over the differences estimator, based purely on efficiency considerations?

13.7 Suppose you have panel data from an experiment with $T = 2$ periods (so $t = 1, 2$). Consider the panel data regression model with fixed individual and time effects and individual characteristics W_i that do not change over time. Let the treatment be binary, so that $X_{it} = 1$ for $t = 2$ for the individuals in the treatment group and $X_{it} = 0$ otherwise. Consider the population regression model

$$Y_{it} = \alpha_i + \beta_1 X_{it} + \beta_2(D_t \times W_i) + \beta_0 D_t + v_{it},$$

where α_i are individual fixed effects, D_t is the binary variable that equals 1 if $t = 2$ and equals 0 if $t = 1$, $D_t \times W_i$ is the product of D_t and W_i, and the α's and β's are unknown coefficients. Let $\Delta Y_i = Y_{i2} - Y_{i1}$. Derive Equation (13.6) (in the case of a single W regressor, so $r = 1$) from this population regression model.

13.8 Suppose you have the same data as in Exercise 13.7 (panel data with two periods, n observations), but ignore the W regressor. Consider the alternative regression model

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 G_i + \beta_3 D_t + u_{it},$$

where $G_i = 1$ if the individual is in the treatment group and $G_i = 0$ if the individual is in the control group. Show that the OLS estimator of β_1 is the differences-in-differences estimator in Equation (13.4). (*Hint:* See Section 8.3.)

13.9 Derive the final equality in Equation (13.10). (*Hint:* Use the definition of the covariance, and remember that, because the actual treatment X_i is random, β_{1i} and X_i are independently distributed.)

13.10 Consider the regression model with heterogeneous regression coefficients

$$Y_i = \beta_0 + \beta_{1i} X_i + v_i,$$

where (v_i, X_i, β_{1i}) are i.i.d. random variables with $\beta_1 = E(\beta_{1i})$.

a. Show that the model can be written as $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $u_i = (\beta_{1i} - \beta_1)X_i + v_i$.

b. Suppose X_i is randomly assigned, so that $E[\beta_{1i}|X_i] = \beta_1$ and $E[v_i|X_i] = 0$. Show that $E[u_i|X_i] = 0$.

c. Show that assumption 1 and assumption 2 of Key Concept 4.3 are satisfied.

d. Suppose outliers are rare, so that (u_i, X_i) have finite fourth moments. Is it appropriate to use OLS and the methods of Chapters 4 and 5 to estimate and carry out inference about the average values of β_{0i} and β_{1i}?

e. Now suppose X_i is not randomly assigned, that $E[v_i|X_i] = 0$, but that β_{1i} and X_i are positively correlated, so that observations with larger-than-average values of X_i tend to have larger-than-average values of β_{1i}. Are the assumptions in Key Concept 4.3 satisfied? If not, which assumption(s) is (are) violated? Will the OLS estimator of β_1 be unbiased for $E(\beta_{1i})$?

13.11 In Chapter 12, state-level panel data were used to estimate the price elasticity of demand for cigarettes, using the state sales tax as an instrumental variable. Consider in particular regression (1) in Table 12.1. In this case, in your judgment does the local average treatment effect differ from the average treatment effect? Explain.

13.12 Consider the potential outcomes framework from Appendix 13.3. Suppose X_i is a binary treatment that is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$. Let $TE_i = Y_i(1) - Y_i(0)$ denote the treatment effect for individual i.

a. Can you consistently estimate $E[Y_i(1)]$ and $E[Y_i(0)]$? If yes, explain how; if not, explain why not.

b. Can you consistently estimate $E(TE_i)$? If yes, explain how; if not, explain why not.

c. Can you consistently estimate $\text{var}[Y_i(1)]$ and $\text{var}[Y_i(0)]$? If yes, explain how; if not, explain why not.

d. Can you consistently estimate $\text{var}(TE_i)$? If yes, explain how; if not, explain why not.

e. Do you think you can consistently estimate the median treatment effect in the population? Explain.

Empirical Exercises

E13.1 A prospective employer receives two resumes: a resume from a white job applicant and a similar resume from an African American applicant. Is the employer more likely to call back the white applicant to arrange an interview? Marianne Bertrand and Sendhil Mullainathan carried out a randomized controlled experiment to answer this question. Because race is not typically included on a resume, they differentiated resumes on the basis of "white-sounding names"

(such as Emily Walsh or Gregory Baker) and "African American–sounding names" (such as Lakisha Washington or Jamal Jones). A large collection of fictitious resumes was created, and the presupposed "race" (based on the "sound" of the name) was randomly assigned to each resume. These resumes were sent to prospective employers to see which resumes generated a phone call (a callback) from the prospective employer. Data from the experiment and a detailed data description are on the text website, **http://www.pearsonhighered.com/stock_watson/**, in the files **Names** and **Names_Description**.[7]

a. Define the *callback rate* as the fraction of resumes that generate a phone call from the prospective employer. What was the callback rate for whites? For African Americans? Construct a 95% confidence interval for the difference in the callback rates. Is the difference statistically significant? Is it large in a real-world sense?

b. Is the African American/white callback rate differential different for men than for women?

c. What is the difference in callback rates for high-quality versus low-quality resumes? What is the high-quality/low-quality difference for white applicants? For African American applicants? Is there a significant difference in this high-quality/low-quality difference for whites versus African Americans?

d. The authors of the study claim that race was assigned randomly to the resumes. Is there any evidence of nonrandom assignment?

13.1 The Project STAR Data Set

The Project STAR public access data set contains data on test scores, treatment groups, and student and teacher characteristics for the 4 years of the experiment, from academic year 1985–1986 to academic year 1988–1989. The test score data analyzed in this chapter are the sum of the scores on the math and reading portions of the Stanford Achievement Test. The binary variable "Boy" in Table 13.2 indicates whether the student is a boy ($=1$) or girl ($=0$); the binary variables "Black" and "Race other than black or white" indicate the student's race. The binary variable "Free lunch eligible" indicates whether the student is eligible for a free lunch during that school year. The "Teacher's years of experience" is the total years of experience of the teacher whom the student had in the grade for which the test data apply. The data set also indicates which school the student attended in a given year, making it possible to construct binary school-specific indicator variables.

---

[7]These data were provided by Professor Marianne Bertrand of the University of Chicago and were used in her paper with Sendhil Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 2004, 94(4): 991–1013.

13.2 IV Estimation When the Causal Effect Varies Across Individuals

This appendix derives the probability limit of the TSLS estimator in Equation (13.12) when there is population heterogeneity in the treatment effect and in the influence of the instrument on the receipt of treatment. Specifically, we assume that the IV regression assumptions in Key Concept 12.4 hold except that treatment effects are heterogeneous, as in Equations (13.9) and (13.11). Further assume that Z_i is randomly assigned or as-if randomly assigned, so $(u_i, v_i, \pi_{1i}, \beta_{1i})$ are distributed independently of Z_i; also assume that $E(\pi_{1i}) \neq 0$ (so the instrument is relevant on average).

Because $(X_i, Y_i, Z_i), i = 1, \ldots, n,$ are i.i.d. with four moments, the law of large numbers in Key Concept 2.6 applies and

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{\sigma_{ZY}}{\sigma_{ZX}}. \tag{13.13}$$

(See Appendix 3.3 and Exercise 18.2.) The task thus is to obtain expressions for σ_{ZY} and σ_{ZX} in terms of the moments of π_{1i} and β_{1i}. Now $\sigma_{ZX} = E[(Z_i - \mu_Z)(X_i - \mu_X)] = E[(Z_i - \mu_Z)X_i]$. Substituting Equation (13.11) into this expression for σ_{ZX} yields

$$\begin{aligned}
\sigma_{ZX} &= E(Z_i - \mu_Z)(\pi_0 + \pi_{1i}Z_i + v_i) \\
&= \pi_0 E(Z_i - \mu_Z) + E[\pi_{1i}Z_i(Z_i - \mu_Z)] + \text{cov}(Z_i, v_i) \\
&= \sigma_Z^2 E(\pi_{1i}),
\end{aligned} \tag{13.14}$$

where the third equality follows because $E(Z_i - \mu_Z) = 0$; because Z_i and v_i are independent, so that $\text{cov}(Z_i, v_i) = 0$; and because π_{1i} and Z_i are independent, so that $E[\pi_{1i}Z_i(Z_i - \mu_Z)] = E(\pi_{1i})E[Z_i(Z_i - \mu_Z)] = \sigma_Z^2 E(\pi_{1i})$.

Next consider σ_{ZY}. Substituting Equation (13.11) into Equation (13.9) yields $Y_i = \beta_0 + \beta_{1i}(\pi_0 + \pi_{1i}Z_i + v_i) + u_i$, so

$$\begin{aligned}
\sigma_{ZY} &= E[(Z_i - \mu_Z)Y_i] \\
&= E[(Z_i - \mu_Z)(\beta_0 + \beta_{1i}\pi_0 + \beta_{1i}\pi_{1i}Z_i + \beta_{1i}v_i + u_i)] \\
&= \beta_0 E(Z_i - \mu_Z) + \pi_0 E[\beta_{1i}(Z_i - \mu_Z)] + E[\beta_{1i}\pi_{1i}Z_i(Z_i - \mu_Z)] \\
&\quad + E[\beta_{1i}v_i(Z_i - \mu_Z)] + \text{cov}(Z_i, u_i).
\end{aligned} \tag{13.15}$$

The assumption that $(u_i, v_i, \beta_{1i}, \pi_{1i})$ is independent of Z_i, along with the fact that $E(Z_i - \mu_Z) = 0$, implies the following simplifications for the five terms after the final equality in Equation (13.15): $\beta_0 E(Z_i - \mu_Z) = 0$, $\pi_0 E[\beta_{1i}(Z_i - \mu_Z)] = \pi_0 E(\beta_{1i})E(Z_i - \mu_Z) = 0$, $E[\beta_{1i}\pi_{1i}Z_i(Z_i - \mu_Z)] = E(\beta_{1i}\pi_{1i})E[Z_i(Z_i - \mu_Z)] = E(\beta_{1i}\pi_{1i})\sigma_Z^2$, $E[\beta_{1i}v_i(Z_i - \mu_Z)] = E(\beta_{1i}v_i)E(Z_i - \mu_Z) = 0$, and $\text{cov}(Z_i, u_i) = 0$. Thus the final expression in Equation (13.15) simplifies to

$$\sigma_{ZY} = \sigma_Z^2 E(\beta_{1i}\pi_{1i}). \tag{13.16}$$

Substituting Equations (13.14) and (13.16) into Equation (13.13) yields $\hat{\beta}_1^{TSLS} \xrightarrow{p} \sigma_Z^2 E(\beta_{1i}\pi_{1i})/\sigma_Z^2 E(\pi_{1i}) = E(\beta_{1i}\pi_{1i})/E(\pi_{1i})$, which is the result stated in Equation (13.12).

13.3 The Potential Outcomes Framework for Analyzing Data from Experiments

This appendix provides a mathematical treatment of the potential outcomes framework discussed in Section 13.1. The potential outcomes framework, combined with a constant treatment effect, implies the regression model in Equation (13.1). If assignment is random, conditional on covariates, the potential outcomes framework leads to Equation (13.2) and conditional mean independence. We consider a binary treatment with $X_i = 1$ indicating receipt of treatment.

Let $Y_i(1)$ denote individual i's potential outcome if treatment is received, and let $Y_i(0)$ denote the potential outcome if treatment is not received, so individual i's treatment effect is $Y_i(1) - Y_i(0)$. The average treatment effect in the population is $E[Y_i(1) - Y_i(0)]$. Because the individual is either treated or not, only one of the two potential outcomes is observed. The observed outcome, Y_i, is related to the potential outcomes by

$$Y_i = Y_i(1)X_i + Y_i(0)(1 - X_i). \tag{13.17}$$

If some individuals receive the treatment and some do not, the expected difference in observed outcomes between the two groups is $E(Y_i|X_i = 1) - E(Y_i|X_i = 0) = E[Y_i(1)|X_i = 1] - E[Y_i(0)|X_i = 0]$. This is true no matter how treatment is determined and simply says that the expected difference is the mean treatment outcome for the treated minus the mean no-treatment outcome for the untreated.

If the individuals are randomly assigned to the treatment and control groups, then X_i is distributed independently of all personal attributes and in particular is independent of $[Y_i(1), Y_i(0)]$. With random assignment, the mean difference between the treatment and control groups is

$$\begin{aligned} E(Y_i|X_i = 1) - E(Y_i|X_i = 0) &= E[Y_i(1)|X_i = 1] - E[Y_i(0)|X_i = 0] \\ &= E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)], \end{aligned} \tag{13.18}$$

where the second equality uses the fact that $[Y_i(1), Y_i(0)]$ are independent of X_i by random assignment and the third equality uses the linearity of expectations [Equation (2.29)]. Thus if X_i is randomly assigned, the mean difference in the experimental outcomes between the two groups is the average treatment effect in the population from which the subjects were drawn.

The potential outcome framework translates directly into the regression notation used throughout this text. Let $u_i = Y_i(0) - E[Y_i(0)]$, and denote $E[Y_i(0)] = \beta_0$. Also denote

$Y_i(1) - Y_i(0) = \beta_{1i}$, so that β_{1i} is the treatment effect for individual i. Starting with Equation (13.17), we have

$$
\begin{aligned}
Y_i &= Y_i(1)X_i + Y_i(0)(1 - X_i) \\
&= Y_i(0) + [Y_i(1) - Y_i(0)]X_i \\
&= E[Y_i(0)] + [Y_i(1) - Y_i(0)]X_i + \{Y_i(0) - E[Y_i(0)]\} \\
&= \beta_0 + \beta_{1i}X_i + u_i.
\end{aligned}
\tag{13.19}
$$

Thus, starting with the relationship between observed and potential outcomes in Equation (13.17) and simply changing notation, we obtain the random coefficients regression model in Equation (13.9). If X_i is randomly assigned, then X_i is independent of $[Y_i(1), Y_i(0)]$ and thus is independent of β_{1i} and u_i. If the treatment effect is constant, then $\beta_{1i} = \beta_1$ and Equation (13.9) becomes Equation (13.1). If the outcome Y_i is measured with error, then the first line of Equation (13.19) would include a measurement error term, which would be subsumed in u_i in the final line.

As discussed in Section 13.1, in some designs X_i is randomly assigned based on the value of a third variable, W_i. If W_i and the potential outcomes are not independent, then, in general, the mean difference between groups does not equal the average treatment effect; that is, Equation (13.18) does not hold. However, random assignment of X_i given W_i implies that, conditional on W_i, X_i and $[Y_i(1), Y_i(0)]$ are independent. This condition—that $[Y_i(1), Y_i(0)]$ is independent of X_i, conditional on W_i—is sometimes called *unconfoundedness*.

If the treatment effect does not vary across individuals and if $E(Y|X_i, W_i)$ is linear, then unconfoundedness implies conditional mean independence of the regression error in Equation (13.2). It follows from Appendix 6.5 that, under these conditions, the OLS estimator of β_1 in Equation (13.2) is unbiased, although, in general, the OLS estimator of γ is biased because $E(u_i|W_i) \neq 0$. To show conditional mean independence under these conditions, let $Y_i(0) = \beta_0 + \gamma W_i + u_i$, where γ is the causal effect (if any) on $Y_i(0)$ of W_i, and let $Y_i(1) - Y_i(0) = \beta_1$ (constant treatment effect). Then the logic leading to Equation (13.19) yields $Y_i = \beta_0 + \beta_1 X_i + \gamma W_i + u_i$, which is Equation (13.2). Thus $E(u_i|X_i, W_i) = E[Y_i(0) - \beta_0 - \gamma W_i|X_i, W_i] = E[Y_i(0) - \beta_0 - \gamma W_i|W_i] = E(u_i|W_i)$, where the second equality follows from unconfoundedness, which implies that $E[Y_i(0)|X_i, W_i] = E[Y_i(0)|W_i]$.

Prediction with Many Regressors and Big Data

Chapter 4 began with two different questions about student performance at elementary schools. A superintendent wanted to know whether test scores would improve if she reduced the student–teacher ratio in her schools—and if they would, by how much. A father, trying to decide where to live, wanted to predict which schools had the highest-performing students. Answering the superintendent's question requires you to estimate the causal effect on test scores of the student–teacher ratio, and estimating causal effects is the focus of Chapters 4–13. In contrast, answering the father's question requires you to predict school test scores given one or more relevant variables—in Chapter 4, the student-teacher ratio, extended in Chapter 6 to include additional information on school and community characteristics.

Statistical prediction entails using data to estimate a prediction model and then applying that model to new, out-of-sample observations. The goal is accurate out-of-sample prediction. In a prediction problem, there are neither specific regressors of interest nor control variables; there are only predictors and the variable to be predicted.

If there are only a handful of predictors, ordinary least squares (OLS) works well if the least squares assumptions for prediction in Appendix 6.4 hold. But modern data sets often have many predictors. For example, the empirical application in this chapter is the prediction of school-level test scores using data on school and community characteristics. We use data on 3932 elementary schools in California; half of these observations are used to estimate prediction models, while the other half are reserved to test their performance.[1] For most of the chapter, we consider a data set with 817 predictors, which is expanded in Section 14.6 to 2065 predictors. This problem of predicting school test scores is typical of many prediction applications using cross-sectional data, such as forecasting sales for a business, predicting patient-level outcomes of medical procedures, or predicting demand for services by state and local government. In such applications, the number of predictors can be nearly as large as, or even larger than, the number of observations.

With so many predictors, OLS overfits the data and makes poor out-of-sample predictions. Fortunately, it is possible to improve upon OLS by using estimators that are broadly referred to as shrinkage estimators. These estimators are biased (they "shrink" the estimator), and the coefficients, in general, do not have a causal interpretation. Remarkably, however, when there are many predictors, introducing bias can reduce the variance of the estimator sufficiently that the overall out-of-sample prediction accuracy is improved.

---

[1] In California, a school district typically contains multiple individual schools. The test score data set used in Chapters 4–9 contains district-level data, while the data used here are for individual schools.

This chapter considers prediction using cross-sectional data sampled from a larger population (shoppers, patients, schools) to predict outcomes for members of the population not in the estimation sample. A related problem is prediction of future events, such as the number of jobs the economy will add next month. Predictions about the future are typically referred to as forecasts, and we adopt that terminology. Forecasting uses time series data, which introduce additional notation and technicalities. Forecasting is taken up in Part IV.

The availability of many predictors is one of the opportunities provided by very large data sets. The field of analyzing big data sets goes by multiple names, including machine learning, data science, and the term we shall use, *big data*.

14.1   What Is "Big Data"?

Data sets can be big in the sense of having many observations, or having many predictors relative to the number of observations, or both. Big data sets can be nonstandard—for example, containing text or images.

Big data sets make available new families of applications. One such family, which is the focus of this chapter, is prediction when the number of predictors k is large compared to the number of observations n. The prediction methods considered in this chapter start with linear regression, so having many predictors corresponds to having many regressors. This situation can arise if one has many distinct primitive predictors, or it can arise if one is considering predictions that are nonlinear functions of the primitive predictors. Even if one starts with only a few dozen primitive predictors, including squares, cubes, and interactions very quickly expands the number of regressors into the hundreds or thousands.

A second family of applications that arises with big data is categorization. We have encountered this problem before, in the context of regression with a binary dependent variable. The logit and probit models of Chapter 11 predict the probability that the dependent variable is 1—in the empirical application, the probability that a loan application is denied. An alternative framing of this problem is to divide the data set into two groups, or categories: those applications that are likely to be denied and those that are likely to be accepted. From a prediction perspective, the aim is to develop a model of loan applications that mimics the decision-making process of a loan officer. Said differently, by fitting that model, a machine (computer) would have learned (estimated) the decision process made by a loan officer. Using that machine learning model, the computer then can make the accept/deny decision itself for future applications. Indeed, the online home loan application industry relies heavily on machine learning, applied to very large data sets on loan applications, to assess the eligibility of an applicant for a mortgage.

A third family of applications concerns testing multiple hypotheses. In the regression context, for example, there might be a potentially large set of coefficients representing different treatments, and the econometrician might be interested in

ascertaining which, if any, of these treatments is effective. Because the *F*-statistic tests a joint hypothesis on a group of coefficients, it is not well suited for the problem of testing many treatments to find out *which* of the treatments is effective. Testing many individual hypotheses with the aim of determining which treatment effect is nonzero requires specialized methods that have been developed for big data applications.

A fourth family of applications concerns handling nonstandard data, such as text and images. The key step is turning these nonstandard data into numerical data, which can then be handled using techniques for high-dimensional data sets. Section 14.7 discusses methods for handling text data.

A fifth, related family of applications is pattern recognition, such as facial recognition or translating text from one language to another. This area has seen great progress using procedures such as "deep learning," which are in essence highly nonlinear models estimated ("trained") using very many observations.

A common feature of all of these problems is that handling large data sets creates computational challenges. Those challenges include storing and accessing large data sets efficiently and developing fast algorithms for estimating models. These computational issues are important; however, we do not address them in this chapter and instead leave them to computer science curricula.

The results of machine learning applied to large data sets are increasingly part of our everyday world. Examples range from software that helps doctors make diagnoses to techniques that target online advertisements to facial recognition algorithms that are used by law enforcement officials. In economics, applications include estimating local incomes based on satellite data, predicting sales for a firm using detailed customer data, interpreting network data on social media sites, searching for patterns in high-frequency asset price databases to use in computerized trading algorithms, and forecasting macroeconomic growth using up-to-the-minute data. Increasingly, computerized analysis of nonstandard data, especially text data, is playing a role in econometric applications.

This chapter cannot cover all these uses of big data, so it focuses on one of the most important for economic applications: the many-predictor problem. Although the nomenclature of this growing field—machine learning, data science, and so forth—makes it seem difficult and new, the methods discussed in this chapter are, at their core, extensions of linear regression analysis that are tailored to the opportunities and challenges of large data sets.

14.2 The Many-Predictor Problem and OLS

This chapter considers the problem of predicting test scores for a school using variables describing the school, its students, and its community. The full data set consists of data gathered on 3932 elementary schools in the state of California in 2013. The task is to use these data to develop a prediction model that will provide good out-of-sample predictions—that is, predictions for schools not in the data set. To simulate

the out-of-sample prediction problem, for most of the chapter we use half the observations ($n = 1966$) for estimating prediction models. The remaining half of the observations are reserved as a test data set to assess how the models perform and are not used until Section 14.6.

The variable to be predicted is the average fifth-grade test score at the school. The primary data set contains 817 distinct variables relating to school and community characteristics; these variables are summarized in Table 14.1. For comparison, smaller and larger data sets are used in Section 14.6. The data are described in more detail in Appendix 14.1.

If only the main variables in Table 14.1 were used, there would be 38 regressors. The analysis of the district test score data in Section 8.4, however, revealed several interesting nonlinearities and interactions in the test score regressions. For example, the regressions in Table 8.3 indicate that there is a nonlinear relationship between test scores and the student–teacher ratio and, in addition, that this relationship differs depending on whether there are a large number of English learners in the district. In Section 8.4, these nonlinearities were handled by including third-degree polynomials of the student–teacher ratio and interaction terms. As laid out in Table 14.1, including interactions, squares, and cubes increases the number of predictors to 817. In Section 14.6, we consider an even larger data set with 2065 predictors, which exceeds the 1966 observations in the estimation sample! Regression with 817 regressors, not to mention 2065 regressors, goes well beyond anything attempted so far in this text.

A natural starting point is OLS. Unfortunately, OLS can produce quite poor predictions when the number of predictors is large relative to the sample size. Fortunately, there are estimators other than OLS that can produce more reliable predictions

---

TABLE 14.1   Variables in the 817-Predictor School Test Score Data Set

Main variables (38)

Fraction of students eligible for free or reduced-price lunch	Ethnicity variables (8): fraction of students who are American Indian, Asian, Black, Filipino, Hispanic, Hawaiian, two or more, none reported
Fraction of students eligible for free lunch	Number of teachers
Fraction of English learners	Fraction of first-year teachers
Teachers' average years of experience	Fraction of second-year teachers
Instructional expenditures per student	Part-time ratio (number of teachers divided by teacher full-time equivalents)
Median income of the local population	
Student–teacher ratio	Per-student expenditure by category, district level (7)
Number of enrolled students	
Fraction of English-language proficient students	Per-student expenditure by type, district level (5)
Ethnic diversity index	Per-student revenues by revenue source, district level (4)

+ Squares of main variables (38)

+ Cubes of main variables (38)

+ All interactions of main variables ($38 \times 37/2 = 703$)

Total number of predictors $= k = 38 + 38 + 38 + 703 = 817$

when the number of predictors relative to the sample size is large. This fact might seem surprising in light of the Gauss–Markov theorem, which says that the OLS estimator has the lowest variance of all unbiased estimators as long as the Gauss–Markov conditions hold (Appendix 5.2). The reason for this surprising result, and the reason it does not violate the Gauss–Markov theorem, is that these alternative estimators are biased. Although the estimators are biased, their variance is sufficiently smaller than the variance of the OLS estimator for them to produce better predictions.

The Mean Squared Prediction Error

To compare prediction models, we need a quantitative measure of predictive accuracy. As we have throughout this text, we will use the square of the error—in this case, the error from out-of-sample predictions. Using the squared prediction error means that small errors receive little weight but large errors receive great weight. This makes sense in many prediction problems, where small errors have negligible impact but very large errors can undercut the usefulness and credibility of the prediction.

The **mean squared prediction error (MSPE)** is the expected value of the square of the prediction error that arises when the model is used to make a prediction for an observation not in the data set.

Stated mathematically, the MSPE is

$$MSPE = E[Y^{oos} - \hat{Y}(X^{oos})]^2, \tag{14.1}$$

where X^{oos} and Y^{oos} are out-of-sample ("oos") observations on X and Y and $\hat{Y}(x)$ is the predicted value of Y for a value x of the predictors. As usual, X is shorthand for the k separate predictors. The notation of Equation (14.1) is taken from Appendix 6.4 (the least squares assumptions for prediction). The notation distinguishes between the n observations $(X_i, Y_i), i = 1, \ldots, n$, used to estimate the prediction model that produces $\hat{Y}(x)$ and the out-of-sample observation for which the prediction is made. The out-of-sample observation is not used to estimate the prediction model.

From the perspective of minimizing the MSPE, the best possible prediction is the conditional mean—that is, $E(Y^{oos}|X^{oos})$ (Appendix 2.2 and Exercise 14.8). This best-possible prediction, $E(Y^{oos}|X^{oos})$, is sometimes called the **oracle prediction**. Because the conditional mean is unknown, the oracle prediction cannot be used in practice (it is infeasible); however, it is the benchmark against which to judge all feasible predictions. In the regression model, the oracle prediction corresponds to the prediction that would be made using the true (unknown) population regression coefficients.

The MSPE embodies two sources of prediction errors. First, even if the conditional mean were known, the prediction would be imperfect: The oracle prediction makes the prediction error, $Y^{oos} - E(Y^{oos}|X^{oos})$. Second, $E(Y^{oos}|X^{oos})$ is unknown, and estimating its parameters—that is, estimating the coefficients of the prediction model $\hat{Y}(x)$—introduces an additional source of error.

The First Least Squares Assumption for Prediction

The school test score application uses data on some (but not all) California schools to estimate the prediction model. We can have confidence that this prediction model will generalize to other California schools; however, we have much less confidence that it will apply to schools in Europe and even less confidence that it will apply to schools in India.

The first least squares assumption for prediction makes this intuition precise. This assumption, which was introduced in Appendix 6.4, states that the out-of-sample observation is drawn from the same distribution as the in-sample observations used to estimate the model:

> **First least squares assumption for prediction**: (X^{oos}, Y^{oos}) are randomly drawn from the same population distribution as the estimation sample $(X_i, Y_i), i = 1, \ldots, n$.

Because the in- and out-of-sample observations are drawn from the same distribution, the conditional mean, $E(Y|X)$, is the oracle prediction for both in- and out-of-sample observations.

The first least squares assumption for prediction is a statement about external validity: The in-sample model can be generalized to the out-of-sample observation of interest.

Although we refer to this assumption as the first least squares assumption for prediction, the requirement applies for estimation methods other than least squares. This condition is assumed to hold for the remainder of this chapter.

The Predictive Regression Model with Standardized Regressors

This chapter uses a modified version of the linear regression model in which the regressors are all standardized; that is, they are transformed to have mean 0 and variance 1. In addition, the dependent variable is transformed to have mean 0. By using standardized regressors, all the regression coefficients have the same units, a property used in the methods of Sections 14.3–14.5.

Let $(X^*_{1i}, \ldots, X^*_{ki}, Y^*_i), i = 1, \ldots, n$, denote the data as originally collected, where X^*_{ji} is the ith observation on the jth original regressor. The standardized regressors are $X_{ji} = (X^*_{ji} - \mu_{X^*_j})/\sigma_{X^*_j}$, where $\mu_{X^*_j}$ and $\sigma_{X^*_j}$ are, respectively, the population mean and standard deviation of $X^*_{j1}, \ldots, X^*_{jn}$. The transformed (demeaned) dependent variable is $Y_i = Y^*_i - \mu_{Y^*}$, where μ_{Y^*} is the population mean of $Y^*_1, \ldots, Y^*_n$.

With this notation, the **standardized predictive regression model** is the regression of Y, which has mean 0, on the k standardized X's:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i. \tag{14.2}$$

The intercept is excluded from Equation (14.2) because all the variables have mean 0.

Because the regressors are standardized, the regression coefficients have the same units: β_j is the difference in the predicted value of Y associated with a one standard deviation difference in X_j^*, holding constant the other X's.

Because the focus of this chapter is prediction, we adopt throughout the prediction interpretation of the regression model in Appendix 6.4; that is, $E(Y|X) = \sum_{j=1}^{k} \beta_j X_j$ and $E(u|X) = 0$.

As usual, the linear structure in Equation (14.2) means that the predictions are linear in the coefficients; however, the regression function can be nonlinear in the predictors because X can include nonlinear terms such as squares or interactions.

The MSPE in the standardized predictive regression model. In the standardized regression model in Equation (14.2), the prediction for the out-of-sample value of the predictors is $\hat{Y}(X^{oos}) = \hat{\beta}_1 X_1^{oos} + \ldots + \hat{\beta}_k X_k^{oos}$. The prediction error is $Y^{oos} - (\hat{\beta}_1 X_1^{oos} + \ldots + \hat{\beta}_k X_k^{oos}) = u^{oos} - [(\hat{\beta}_1 - \beta_1)X_1^{oos} + \ldots + (\hat{\beta}_k - \beta_k)X_k^{oos}]$, where the final expression obtains using Equation (14.2), and u^{oos} is the value of the error u for the out-of-sample observation. Because u^{oos} is independent of the data used to estimate the coefficients and is uncorrelated with X^{oos}, the MSPE in Equation (14.1) for the standardized predictive regression model can be written as the sum of two components:

$$\text{MSPE} = \sigma_u^2 + E[(\hat{\beta}_1 - \beta_1)X_1^{oos} + \ldots + (\hat{\beta}_k - \beta_k)X_k^{oos}]^2. \qquad (14.3)$$

The first term in Equation (14.3), σ_u^2, is the variance of the oracle prediction error—that is, of the prediction error made using the true (unknown) conditional mean, $E(Y|X)$.

The second term in Equation (14.3) is the contribution to the prediction error arising from the estimated regression coefficients. This second term represents the cost, measured in terms of increased mean squared prediction error, of needing to estimate the coefficients instead of using the oracle prediction.

Because the mean square is the sum of the variance and the square of the bias (Equation (2.33)), the second term in Equation (14.3) is the sum of the variance of the prediction arising from estimating β and the squared bias of the prediction. When it comes to determining which estimator to use, the goal is to make this second term in Equation (14.3) as small as possible. As we shall see, when there are many predictors, this entails trading off the bias of the estimated coefficients against their variance.

Standardization using the sample means and variances. In practice, the population means and standard deviations of the original variables are not known. Accordingly, the in-sample means and variances are used to standardize the regressors, and the in-sample mean is subtracted from the dependent variable.

Because the regressors are standardized and the dependent variable is demeaned, an additional step is needed to produce the prediction for an out-of-sample observation. Specifically, the out-of-sample observation on the predictors must be standardized using the in-sample mean and standard deviation, and the in-sample mean of

the dependent variable must be added back into the prediction. Formulas are given in Appendix 14.5.

The MSPE of OLS and the Principle of Shrinkage

In the special case that the regression error u in Equation (14.2) is homoskedastic, the MSPE of OLS is given by

$$\text{MSPE}_{\text{OLS}} \cong \left(1 + \frac{k}{n} \right) \sigma_u^2. \tag{14.4}$$

The approximation in Equation (14.4) holds exactly in some special cases (Exercise 14.12), and it holds more generally as an approximation when n is large and k/n is small. In the case of a single regressor, Equation (14.4) is derived in Appendix 14.2. The derivation of Equation (14.4) for general k uses matrix algebra and is given in Appendix 19.7.

This expression has a simple interpretation. As discussed following Equation (14.3), the MSPE of the oracle prediction—that is, the prediction using the true value of β—is σ_u^2. When the k regression coefficients are estimated by OLS, the MSPE increases by the factor $(1 + k/n)$ relative to the best-possible MSPE. Thus the cost, as measured by the MSPE, of using OLS depends on the ratio of the number of regressors to the sample size.

For example, in the school test score application, suppose the 38 main regressors in Table 14.1 are used to predict test scores. Although 38 regressors sounds like a lot, $k/n = 38/1966 \approx 0.02$, so using OLS entails only a 2% loss in MSPE relative to the oracle prediction. In many applications, a loss of 2% might not be important. In the data set with 817 regressors, however, $k/n = 817/1966 \cong 0.40$, and a 40% deterioration is large enough that it is worth investigating estimators that have a lower MSPE than OLS.

Because OLS is unbiased under the prediction interpretation of Equation (14.2), the inflation factor $(1 + k/n)$ arises solely from the variance of the OLS estimator. Under the Gauss–Markov conditions, the OLS estimator has the smallest variance of all linear unbiased estimators. As a result, one might naturally be discouraged about making much headway when k/n is large. But a major conceptual breakthrough in the many-predictor problem, dating to the early 1960s, was the discovery that if one allows for biased estimators, the estimator variance can be reduced by so much that the MSPE can be less than that of OLS.

The principle of shrinkage. A **shrinkage estimator** introduces bias by "shrinking" the OLS estimator toward a specific number and thereby reducing the variance of the estimator. Because the mean squared error is the sum of the variance and the squared bias (Equation (2.33)), if the estimator variance is reduced by enough, then the decrease in the variance can more than compensate for the increase in the squared bias. The result is an estimator with a lower mean squared error than OLS.

James and Stein (1961) developed the first estimator that achieved this goal of reducing the estimator mean squared error by introducing bias. When the regressors are uncorrelated, the James–Stein estimator can be written as $\widetilde{\beta}^{JS} = c\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator and c is a factor that is less than 1 and depends on the data. Because c is less than 1, the James–Stein estimator shrinks the OLS estimator toward 0 and thus is biased toward 0. It is not surprising that the James–Stein estimator has a lower mean squared error than the OLS estimator when the true β's are small. What James and Stein showed, however, is that if the errors are normally distributed, their estimator has a lower mean squared error than the OLS estimator, *regardless* of the true value of β, as long as $k \geq 3$.

James and Stein's remarkable result is the foundation of many-predictor methods used with big data. Their result leads to the family of shrinkage estimators, which includes ridge regression and the Lasso estimator, the topics of Sections 14.3 and 14.4, respectively.

Estimation of the MSPE

The MSPE is a population expectation and thus is unknown. However, it can be estimated from a sample of data. Here, we discuss two ways to estimate the MSPE. The first, split-sample estimation, draws directly on the definition of the MSPE and entails dividing the sample into two subsamples, one for estimation and one for prediction. The second, called m-fold cross validation, extends this idea but uses the data symmetrically and more efficiently by dividing the sample into m subsamples.

Estimating the MSPE using a split sample. Recall that the MSPE is the variance of the prediction error for a randomly drawn X, where the observation is not used to estimate β. This definition suggests estimating the MSPE by dividing the data set into two parts: an estimation subsample and a "test" subsample used to simulate out-of-sample prediction. The estimation subsample is used to estimate β, yielding the estimate $\widetilde{\beta}$, which could be obtained by OLS or some other estimator. This estimate is then used to make a prediction $\hat{Y}$ for each of the n_{test} observations in the test subsample. The MSPE is then estimated using the resulting n_{test} prediction errors:

$$\widehat{\text{MSPE}}_{split\text{-}sample} = \frac{1}{n_{test}} \sum_{\substack{observations\ in \\ test\ subsample}} (Y_i - \hat{Y}_i)^2. \tag{14.5}$$

Estimating the MSPE by m-fold cross validation. The split-sample procedure treats the data asymmetrically by arbitrarily splitting the observations into two subsamples that are then used for different purposes. This estimator can be improved by treating the data symmetrically. Specifically, the two subsamples can be used to produce two different estimators of the MSPE by swapping which subsample is used to estimate β and which is used to estimate the MSPE.

This idea extends to m different, randomly chosen subsamples. The resulting procedure is called m-fold cross validation. In **m-fold cross validation**, there are m separate estimates of the MSPE, each produced by sequentially leaving out one of

m-fold Cross Validation

The m-fold cross-validation estimator of the MSPE is determined according to the following six steps.

1. Divide the test sample into m randomly chosen subsets of approximately equal size.

2. Use the combined subsamples 2, 3, . . . , m to compute $\tilde{\beta}$, an estimate of β.

3. Use $\tilde{\beta}$ and Equation (14.12) to compute predicted values $\hat{Y}$ and prediction errors $Y - \hat{Y}$ for the observations in subsample 1.

4. Using subsample 1 as the test sample, estimate the MSPE with the predicted values in subsample 1 and Equation (14.5); call this estimate $\widehat{MSPE}_1$.

5. Repeat steps 2–4 using subsample 2 as the left-out test sample, then subsample 3, and so forth, yielding a total of m estimates $\widehat{MSPE}_i$, $i = 1, \ldots, m$.

6. The m-fold cross-validation estimator of the MSPE is then estimated by averaging these m subsample estimates of the MSPE:

$$\widehat{MSPE}_{m\text{-}fold\ cross\ validation} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{n_i}{n/m}\right)\widehat{MSPE}_i, \qquad (14.6)$$

where n_i is the number of observations in subsample i and the factor in parentheses allows for different numbers of observations in the different subsamples.

the m subsamples when estimating β and using that reserved subsample to estimate the MSPE. The m-fold cross-validation estimator of the MSPE is the average of the m subset estimators of the MSPE. The m-fold cross-validation estimator of the MSPE is summarized in Key Concept 14.1.

A loose end in m-fold cross validation is how to choose m. This involves a trade-off. A larger value of m produces more efficient estimators of β because more observations are used each time β is estimated. From this perspective, ideally one would use the so-called leave-one-out cross-validation estimator, for which $m = n - 1$. But a larger value of m means that β must be estimated m times. In applications in which k is large (in the hundreds or more), this can take considerable computer time, and leave-one-out cross validation takes too long computationally. As a result, the choice of m must be made taking into account practical constraints on your and your computer's time. In the school test score application in this chapter, we settle on $m = 10$ as a practical compromise given the computer we used, so that each subsample estimator of β uses 90% of the sample.

The m-fold cross-validation estimator can be used to estimate the MSPE in very general settings, regardless of how β is estimated. It even works for models that can be expressed only as algorithms, not in terms of parameters. This general applicability makes it widely used in empirical work with big data.

14.3 Ridge Regression

Sections 14.3 and 14.4 describe two shrinkage estimators that are designed for use with many predictors. The method discussed in this section, ridge regression, shrinks the estimated parameter to 0 by adding to the sum of squared residuals a penalty that increases with the square of the estimated parameter. By minimizing the sum of these two terms, which is called the penalized sum of squared residuals, ridge regression introduces bias into the estimator but reduces its variance. In some applications, ridge regression can result in large improvements in MSPE compared to OLS.

Shrinkage via Penalization and Ridge Regression

One way to shrink the estimated coefficients toward 0 is to penalize large values of the estimate. The ridge regression estimator is based on this idea. Specifically, the **ridge regression** estimator minimizes the penalized sum of squares, which is the sum of squared residuals plus a penalty factor that increases with the sum of the squared coefficients:

$$S^{Ridge}(b; \lambda_{Ridge}) = \sum_{i=1}^{n}(Y_i - b_1 X_{1i} - \ldots - b_k X_{ki})^2 + \lambda_{Ridge}\sum_{j=1}^{k} b_j^2, \quad (14.7)$$

where $\lambda_{Ridge} \geq 0$. The parameter λ_{Ridge} is called the ridge shrinkage parameter. The ridge regression estimator is the value of b that minimizes $S^{Ridge}(b; \lambda_{Ridge})$.

The first term on the right-hand side of Equation (14.7) is the usual sum of squared residuals for a trial coefficient value b. If this were the only term, then the ridge and OLS estimators would be the same. The second term, however, is new. This second term increases with the sum of the squared coefficients. This second term in Equation (14.7) is called a **penalty term** because it penalizes the estimator for choosing a large estimate of the coefficient. When the penalty term is scaled by the shrinkage parameter and added to the sum of squared residuals, as it is in Equation (14.7), the result is called the **penalized sum of squared residuals**.

The penalty term shrinks the ridge regression estimator toward 0. Figure 14.1 shows how ridge penalization works when there is only one regressor. Without the penalty, one would minimize the sum of squared residuals, which yields the OLS estimator. Adding in the penalty shifts the minimum of the penalized function toward 0. Thus the estimated ridge coefficient will be closer to 0 than the OLS estimator is; that is, the ridge regression estimator is shrunk toward 0.

The magnitude of the shrinkage depends on the shrinkage parameter λ_{Ridge}. If $\lambda_{Ridge} = 0$, there is no shrinkage, and the ridge regression estimator equals the OLS estimator. The larger λ_{Ridge}, the greater the penalty for a given value of b, and the greater the shrinkage of the estimator toward 0. Because we are using the standardized predictive regression model, all the coefficients have the same units, so a single shrinkage parameter λ_{Ridge} can be used for all the coefficients.

FIGURE 14.1 Components of the Ridge Regression Penalty Function

The ridge regression estimator minimizes $S^{Ridge}(b)$, which is the sum of squared residuals, $SSR(b)$, plus a penalty that increases with the square of the estimated parameter. The SSR is minimized at the OLS estimator, $\hat{\beta}$. Including the penalty shrinks the ridge estimator, $\hat{\beta}^{Ridge}$, toward 0.
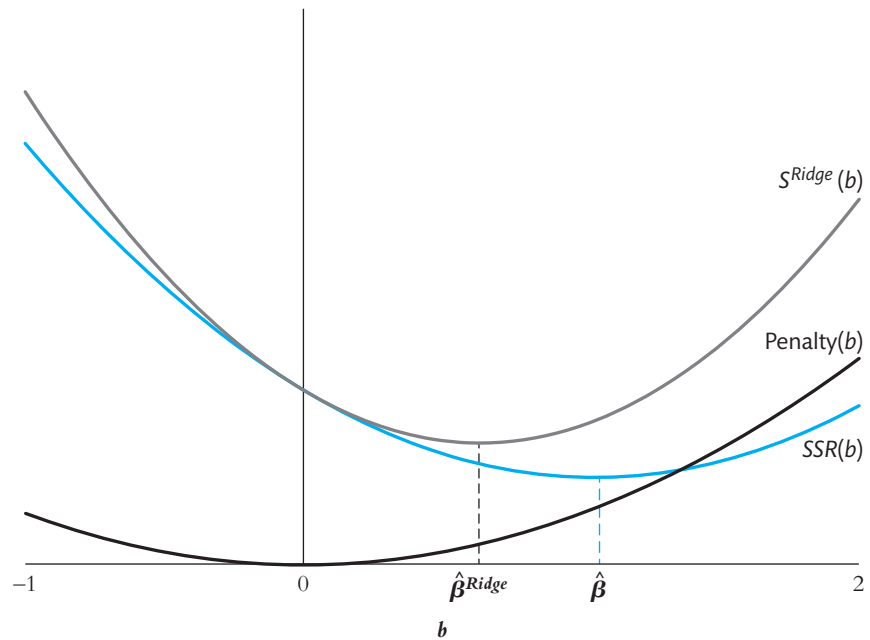
The penalized sum of squared residuals in Equation (14.7) can be minimized using calculus to give a simple expression for the ridge regression estimator. This formula is derived in Appendix 14.3 for the case of a single regressor. When $k > 2$, the formula is best expressed using matrix notation, and it is given in Appendix 19.7.

In the special case that the regressors are uncorrelated, the ridge regression estimator is

$$\hat{\beta}_j^{Ridge} = \left(\frac{1}{1 + \lambda_{Ridge}/\sum_{i=1}^{n} X_{ji}^2} \right) \hat{\beta}_j, \tag{14.8}$$

where $\hat{\beta}_j$ is the OLS estimator of β_j. In this case, the ridge regression estimator shrinks the OLS estimator toward 0, like the James–Stein estimator. When the regressors are correlated, the ridge regression estimate can sometimes be greater than the OLS estimate although overall the ridge regression estimates are shrunk towards zero.

When there is perfect multicollinearity, such as when $k > n$, the OLS estimator can no longer be computed, but the ridge estimator can.

Estimation of the Ridge Shrinkage Parameter by Cross Validation

The ridge regression estimator depends on the shrinkage parameter λ_{Ridge}. While the value of λ_{Ridge} could be chosen arbitrarily, a better strategy is to pick λ_{Ridge} so that the ridge regression estimator works well for the data at hand.

One might initially think that the shrinkage parameter λ_{Ridge} could be estimated by minimizing $S^{Ridge}(b; \lambda_{Ridge})$ in Equation (14.7). However, for any trial value of b, minimizing $S^{Ridge}(b; \lambda_{Ridge})$ with respect to λ_{Ridge} simply leads to setting λ_{Ridge} to 0; but when $\lambda_{Ridge} = 0$, the ridge regression estimator is just the OLS estimator! The reason that this approach yields the OLS estimator is that it provides the best *in-sample* fit—which is given by OLS. In contrast, the goal of prediction is to have a good *out-of-sample* fit—that is, a low MSPE.

That insight suggests choosing λ_{Ridge} to minimize the estimated MSPE. This strategy can be implemented using the m-fold cross-validation estimator of the MSPE (Key Concept 14.1). Specifically, suppose you have two candidate values of λ_{Ridge}—for example, 0.1 and 0.2—and choose some value of m. Let $\tilde{\beta}$ in Key Concept 14.1 denote the ridge regression estimator using $\lambda_{Ridge} = 0.1$. Given $\tilde{\beta}$, compute the predictions in the test sample, and use those predictions to compute $\widehat{MSPE}$ for that estimator. Now repeat, but use $\lambda_{Ridge} = 0.2$. You now have two estimates of the MSPE, one for $\lambda_{Ridge} = 0.1$ and one for $\lambda_{Ridge} = 0.2$, so choose the value of λ_{Ridge} that provides the lowest estimated MSPE. Repeating these steps for multiple values of λ_{Ridge} yields an estimator of λ_{Ridge} that minimizes the m-fold cross-validation MSPE. Although this estimator could potentially be 0—so that the best ridge estimator is the OLS estimator—typically the best shrinkage parameter will not be 0 and the ridge estimator will differ from the OLS estimator.

Application to School Test Scores

We illustrate the use of ridge regression by fitting a predictive model for school test scores using the 817 predictors in Table 14.1 with 1966 observations.

Figure 14.2 plots the square root of the 10-fold cross-validation estimator of the MSPE as a function of the ridge shrinkage parameter λ_{Ridge}. The square root of the MSPE is plotted so that it provides an estimate of the magnitude of a typical out-of-sample prediction error. For a given value of λ_{Ridge}, the MSPE was computed as described in Key Concept 14.1. The choice of $m = 10$ represents a practical balance between the desire to use as many observations as possible to estimate the parameters and the computational burden of repeating that estimation m times for each value of λ_{Ridge}.

As Figure 14.2 shows, the MSPE has a U shape. It is minimized at $\lambda_{Ridge} = 2233$, so the 10-fold cross-validation estimate of the ridge shrinkage parameter is $\hat{\lambda}_{Ridge} = 2233$.

The square root of the MSPE, evaluated at $\hat{\lambda}_{Ridge}$, is 39.5. In contrast, the root MSPE for OLS, estimated using the same 817 predictors and 1966 observations, is much larger, 78.2. Because the OLS estimator is the ridge estimator with $\lambda_{Ridge} = 0$, in principle the root MSPE of the OLS estimator could also be shown in Figure 14.2 as the point ($\lambda_{Ridge} = 0$, root MSPE = 78.2); however, the root MSPE for OLS is so large that it is off the scale of the figure.

The MSPE is estimated using 10-fold cross validation for the school test score data set with $k = 817$ predictors and $n = 1966$ observations.

FIGURE 14.2 Square Root of the MSPE for the Ridge Regression Prediction as a Function of the Shrinkage Parameter (Log Scale for λ_{Ridge})
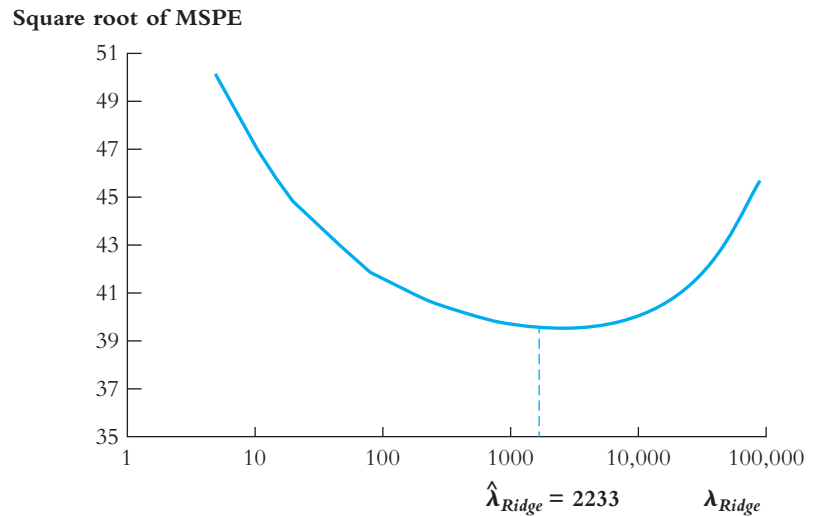
The fact that the OLS MSPE is much larger than the ridge MSPE provides an empirical demonstration of the main theoretical point discussed in Section 14.2: When there are many predictors, introducing bias into the parameter estimates via shrinkage can reduce the variance of the prediction by more than enough to compensate for the bias and therefore produce much more accurate predictions.

Because $\hat{\lambda}_{Ridge}$ is chosen to minimize the cross-validated MSPE, the cross-validated MSPE evaluated at $\hat{\lambda}_{Ridge}$ is no longer an unbiased estimator of the MSPE. In Section 14.6, we use the remaining 1966 observations (not used so far) to obtain an unbiased estimator of the MSPE for ridge regression using $\hat{\lambda}_{Ridge}$.

It is also of interest to compare the ridge regression coefficients to the OLS coefficients. That comparison is made in Section 14.6, where these coefficients are also compared to the methods discussed in Sections 14.4 and 14.5, the Lasso and principal components, respectively.

14.4 The Lasso

In OLS and ridge regression, none of the estimated coefficients is exactly 0 so all the regressors are used to make the prediction. In some applications, however, only a few predictors might be useful, with the rest irrelevant. For example, among the predictors in Table 14.1, all but 38 are constructed as squares, cubes, or interactions of the 38 main variables; if the true conditional expectation is, in fact, linear in the 38 main variables, then $817 - 38 = 779$ of the variables would have a coefficient of 0.

A regression model in which the coefficients are nonzero for only a small fraction of the predictors is called a **sparse model**. If the model is sparse, predictions can be improved by estimating many of the coefficients to be *exactly* 0.

The estimator examined in this section, the Lasso (least absolute shrinkage and selection operator), is designed for sparse models. Like ridge regression, the Lasso shrinks estimated coefficients to 0. Unlike ridge regression, it sets many of the estimated coefficients exactly to 0, thereby dropping those regressors from the model. Moreover, the regressors it keeps are subject to less shrinkage than with ridge regression. Thus, the Lasso provides a way to select a subset of the regressors and then estimate their coefficients with a modest amount of shrinkage.

Like ridge regression, the Lasso can be used when $k > n$. Also like ridge regression, the Lasso has a shrinkage parameter that can be estimated by minimizing the cross-validated MSPE.

Shrinkage Using the Lasso

The **Lasso** estimator minimizes a penalized sum of squares, where the penalty increases with the sum of the absolute values of the coefficients:

$$S^{Lasso}(b; \lambda_{Lasso}) = \sum_{i=1}^{n}(Y_i - b_1 X_{1i} - \ldots - b_k X_{ki})^2 + \lambda_{Lasso}\sum_{j=1}^{k}|b_j|, \quad (14.9)$$

where λ_{Lasso} is called the Lasso shrinkage parameter. The Lasso estimator is the value of b that minimizes $S^{Lasso}(b; \lambda_{Lasso})$. As with ridge regression, if the shrinkage parameter $\lambda_{Lasso} = 0$, the Lasso estimator minimizes the sum of squared residuals in which case the Lasso is just OLS. The second term in Equation (14.9) penalizes large values of b and thus shrinks the Lasso estimate toward 0.[2]

The first part of the Lasso name—least absolute shrinkage—reflects the nature of the penalty term in Equation (14.9). Whereas the ridge regression penalty increases with the square of b, the Lasso penalty increases with its absolute value.

The second part of the Lasso name—selection operator—arises because the Lasso estimates many coefficients to be exactly 0, thereby dropping some of the predictors. Thus the Lasso, in effect, selects a subset of the predictors to be used in the model.

The reason that the Lasso estimates some coefficients to be exactly 0 is illustrated in Figure 14.3 for $k = 1$. This figure shows the sum of squared residuals, the Lasso penalty, and the combined Lasso minimization function in Equation (14.9). Parts a and b of Figure 14.3 differ only in the value of the OLS estimate, which minimizes the first term in Equation (14.9). In Figure 14.3a, the OLS estimate is far from

---

[2]The ridge and Lasso penalty terms can both be written as $\lambda \sum_{j=1}^{k}|b_j|^p$, where $p = 2$ for ridge and $p = 1$ for Lasso. The expression $(\sum_{j=1}^{k}|b_j|^p)^{1/p}$ is called the L_p length of b, where $p = 2$ corresponds to the usual Euclidean distance. As a result, the ridge is sometimes called L_2 penalization, and the Lasso is sometimes called L_1 penalization.

For a single regressor,
(a) when the OLS esti-
mator is far from zero,
the Lasso estimator
shrinks it toward 0; (b)
when the OLS estimator
is close to 0, the Lasso
estimator becomes
exactly 0.
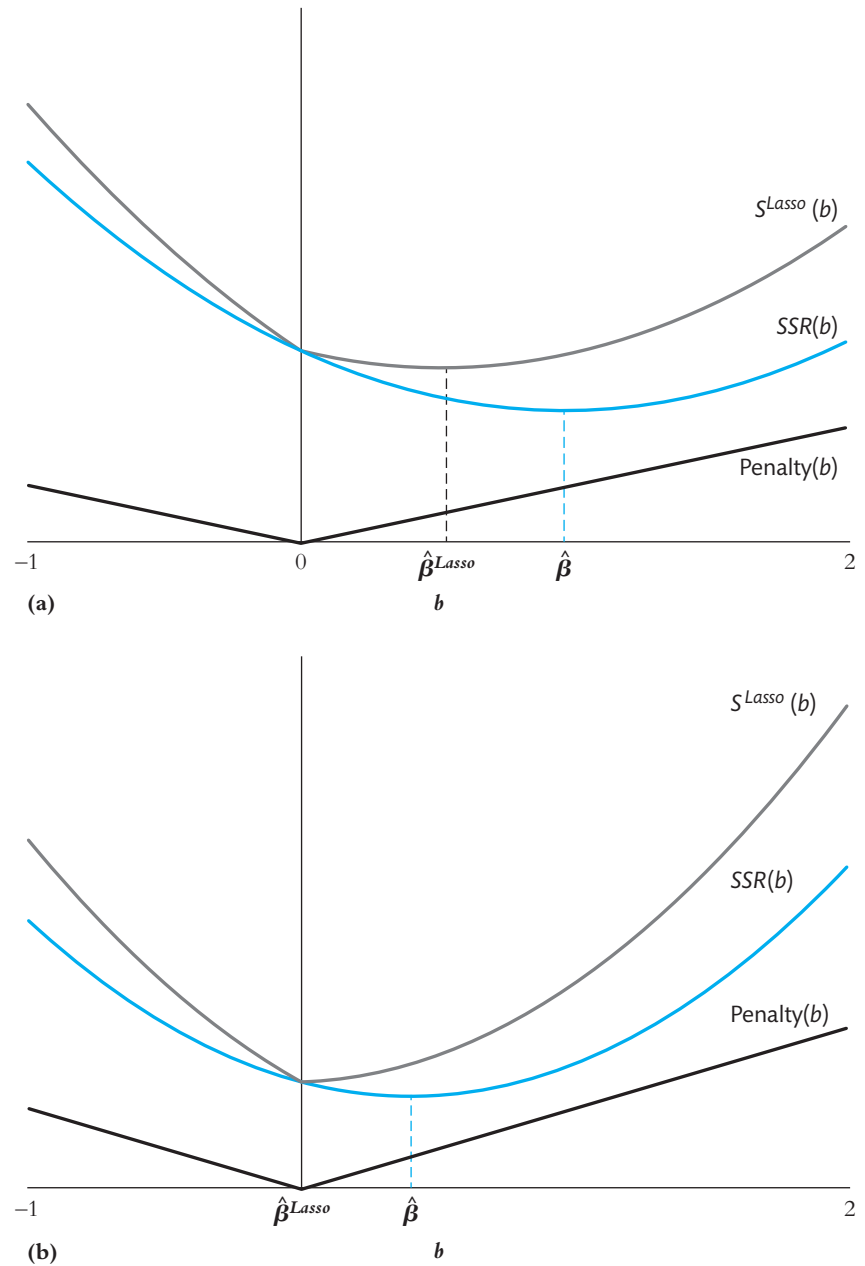
0 ($\hat{\beta} = 1.0$), and the Lasso shrinks it to a smaller value ($\hat{\beta}^{Lasso} = 0.5$). In Figure 14.3b, the curve representing the sum of squared residuals is shifted to the left, so the OLS estimate is smaller ($\hat{\beta} = 0.4$), and the Lasso estimate is exactly 0 ($\hat{\beta}^{Lasso} = 0$). This estimate of exactly 0 arises because the sum of squared residuals function in Figure 14.3b is so flat near 0 that the penalty term takes over from the sum of squared residuals and drives the estimate to 0.

Appendix 14.4 provides a formula for the Lasso estimator when $k = 1$. The formula shows mathematically that for sufficiently small values of the OLS estimator, the Lasso estimator is exactly 0.

The ridge and Lasso estimators also behave differently when the OLS estimate is large. For large values of b, the ridge penalty exceeds the Lasso penalty. Thus, when the OLS estimate is large, the Lasso shrinks it less than ridge, but when the OLS estimate is small, the Lasso shrinks it more than ridge—in some cases, all the way to 0.

Figure 14.3 considers the case of a single regressor, for which the Lasso always shrinks the OLS estimator toward 0. If there are multiple predictors, then the Lasso generally shrinks the OLS estimates toward 0; however, it is possible that the Lasso estimate of some of the coefficients could be larger than the OLS estimate.

Computation of the Lasso estimator.   Unlike OLS and ridge regression, there is no simple expression for the Lasso estimator when $k > 1$, so the Lasso minimization problem must be done numerically using a computer. One of the many computational advances in machine learning is the development of specialized algorithms to compute the Lasso estimator. Some econometric software packages incorporate these algorithms and make it straightforward to use the Lasso estimator.

Estimation of the shrinkage parameter by cross validation.   As in ridge regression, the Lasso tuning parameter can be estimated by minimizing an estimate of the MSPE. The algorithm for estimating λ_{Lasso} is the same as that laid out in Section 14.3 for estimating λ_{Ridge}.

A word of warning about the ridge and Lasso estimators.   The ridge and Lasso estimators differ from all the other estimators used in this text in an important way. In OLS, the fit of the regression model is the same whether one uses the k original regressors or k linear combinations of the regressors as long as one avoids perfect multicollinearity. For example, one can use an intercept and a dummy variable for *male*, or an intercept and a dummy variable for *female*, or both a *male* dummy and a *female* dummy and no intercept; all yield identical fits of the OLS regression and identical predictions. Moreover, which of these three specifications is used makes no difference for the other estimated coefficients in the model.

In contrast, with ridge and Lasso the regression fit, the estimated coefficients, and the predictions in general depend on the specific choice of the linear combination of regressors used. This is easiest to see for the Lasso because the population values of the coefficients change as you change linear combinations. For example, the

coefficient on male in the (intercept, *male*) specification differs from that in the (*female*, *male*) specification. Thus the Lasso might drop *male* from the (intercept, *male*) specification but not from the (*female*, *male*) specification. If so, the (intercept, *male*) and (*female*, *male*) specifications would have different selected predictors and thus would make different predictions.

The reason that the choice of linear combinations matters for ridge is more subtle and stems from the fact that different linear combinations will have different correlations with each other. An explanation of this result for ridge regression is given in Appendix 19.7.

The dependence of the ridge and Lasso estimators on the choice of linear combination of regressors implies that one needs to put thought into choosing the regressors when using these estimators—a decision that does not matter for OLS or for the principal components method of Section 14.5 (or, for that matter, for logit, probit, or IV regression).

Application to School Test Scores

We now turn to estimation of a Lasso prediction model using the same 817 regressors and 1966 observations as in Section 14.3.

Figure 14.4 plots the square root of the 10-fold cross-validation estimate of the MSPE as a function of the Lasso shrinkage parameter λ_{Lasso}. The MSPE is minimized when the shrinkage parameter is 4527, so $\hat{\lambda}_{Lasso} = 4527$. At this estimated value of λ_{Lasso}, the MSPE is 39.7. This MSPE is much less than the MSPE of OLS, 78.2, which is equivalent to the Lasso estimator for $\lambda_{Lasso} = 0$. The Lasso MSPE is close to, but slightly greater than, the minimized ridge MSPE of 39.5 (from Section 14.3).

FIGURE 14.4   Square Root of the MSPE for the Lasso Prediction as a Function of the Lasso Shrinkage Parameter (Log Scale for λ_{Lasso})

The MSPE is estimated by 10-fold cross validation using the school test score data set with $k = 817$ predictors and $n = 1966$ observations.

$\hat{\lambda}_{Lasso} = 4527$
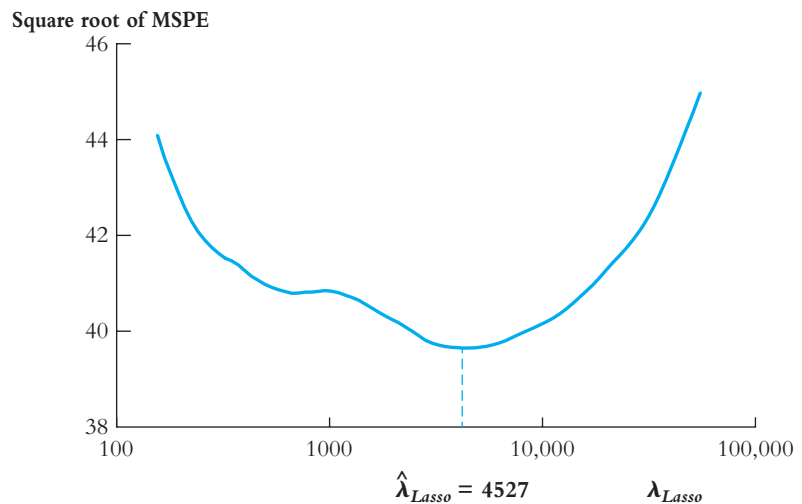
The Lasso estimates nonzero coefficients on only 56 of the 817 predictors; thus the Lasso estimator excludes 761, or 93%, of the candidate predictors in Table 14.1. Of the retained predictors, all but 4 are interactions among the 38 main predictors in Table 14.1.

14.5   Principal Components

When the regressors are perfectly collinear, at least one of them can be dropped from the data set without any loss of information because the dropped regressor can be perfectly reconstructed from the retained regressors. This observation suggests that there might be little loss of information from dropping a variable that is highly, but imperfectly, correlated with the other regressors. This insight forms the basis for an alternative strategy for handling many predictors: Exploit the correlations among the regressors to reduce the number of regressors while retaining as much of the information in the original regressors as possible. Principal components analysis implements this strategy and can reduce sharply the number of regressors so that estimation and prediction can proceed using OLS.

This section begins by showing how principal components analysis works when there are two regressors. We then turn to the more relevant case when the number of regressors is large.

Principal Components with Two Variables

The **principal components** of a set of standardized variables X are linear combinations of those variables, where the linear combinations are chosen so that the principal components are mutually uncorrelated and sequentially contain as much of the information in the original variables as possible. Specifically, the linear combination weights for the first principal component are chosen to maximize its variance, in this sense capturing as much of the variation of the X's as possible. The linear combination weights for the second principal component are chosen so that it is uncorrelated with the first principal component and captures as much of the variance of the X's as possible, controlling for the first principal component. The third principal component is uncorrelated with the first two and captures as much of the variance of the X's as possible, controlling for the first two principal components, and so forth. If $k \leq n$ and there is no perfect multicollinearity, then the total number of principal components is k. If $k > n$, then the total number of principal components is n.

It is easiest to see how this procedure works when there are two X's. Figure 14.5 illustrates this case when X_1 and X_2 are standard normal random variables with a correlation of 0.7. The first principal component is the weighted average, $PC_1 = w_1 X_1 + w_2 X_2$, with the maximum variance, where w_1 and w_2 are the principal component weights. Choosing the weights corresponds to choosing a direction in which to add the variables or, equivalently, choosing a direction in which the spread

The first principal component (PC_1) maximizes the variance of the linear combination of these variables, which is done by adding X_1 and X_2. The second principal component (PC_2) is uncorrelated with the first and is obtained by subtracting the two variables. The principal component weights are normalized so that the sum of squared weights adds to 1.

$$PC_1 = (X_1 + X_2)/\sqrt{2}$$

$$PC_2 = (X_1 - X_2)/\sqrt{2}$$
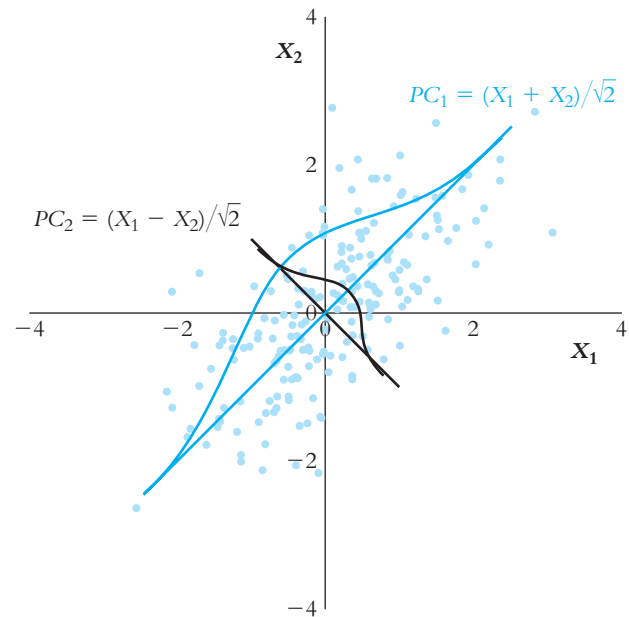
of the variables is greatest. As Figure 14.5 illustrates, the spread of the variables is greatest in the direction of the 45° line. Along this direction, the variables are added together with equal weights.

Without further restrictions, the variance of the linear combination can always be increased simply by increasing both w_1 and w_2. Thus, for the principal components problem to have a solution, it is necessary to restrict the weights. This is done by requiring the sum of squared weights to equal 1; that is, $w_1^2 + w_2^2 = 1$. Along the 45° line, the weights are equal, so $w_1 = w_2 = 1/\sqrt{2}$ and $PC_1 = (X_1 + X_2)/\sqrt{2}$, a result derived mathematically in Exercise 14.11.

The second principal component is chosen to be uncorrelated with the first principal component, and the sum of its squared weights also equals 1. When there are two variables, these two requirements imply that $PC_2 = (X_1 - X_2)/\sqrt{2}$. This corresponds to adding the variables along the downward-sloping 45° line in Figure 14.5. As illustrated in the figure, the spread of the variables is minimized in this direction. Thus, when there are only two variables, the first principal component maximizes the variance of the linear combination, while the second principal component minimizes the variance of the linear combination.

The variances of the two principal components are $var(PC_1) = 1 + |\rho|$ and $var(PC_2) = 1 - |\rho|$, where $\rho = corr(X_1, X_2)$ (Exercise 14.11). These expressions confirm that if the variables are correlated, PC_1 has a greater variance than PC_2.

These expressions for the variances of PC_1 and PC_2 have another, more subtle feature: $\text{var}(PC_1) + \text{var}(PC_2) = \text{var}(X_1) + \text{var}(X_2)$.[3] This provides an R^2 interpretation of principal components: The fraction of the total variance explained by the first principal component is $\text{var}(PC_1)/[\text{var}(X_1) + \text{var}(X_2)]$, and the fraction explained by the second is $\text{var}(PC_2)/[\text{var}(X_1) + \text{var}(X_2)]$. Together, the two principal components explain all the variance of X. For the two variables in Figure 14.5, the correlation is 0.7, so the first principal component explains $(1 + \rho)/2 = 85\%$ of the variance of X, while the second principal component explains the remaining 15% of the variance of X.

If there are only two variables, there is little reason to reduce their number using principal components. The utility of principal components arises when there are many correlated variables, in which case much or most of the variation in those variables can be captured by a smaller number of principal components.

Principal Components with k Variables

The principal components of the k variables $X_1, \ldots, X_k$ are the linear combinations of those variables that are mutually uncorrelated, have squared weights that sum to 1, and maximize the variance of the linear combination controlling for the previous principal components. Assuming there is no perfect multicollinearity among the variables, the number of principal components of X is the minimum of n and k.

Expressions for the principal component weights for $k > 2$ are more complicated than when $k = 2$. Fortunately, there is a fast method for computing the principal components and their weights. Because this method entails matrix calculations, it is deferred to Appendix 19.7. This procedure for computing principal components is widely available in standard statistical software.

Principal components with k variables is summarized in Key Concept 14.2.

The scree plot. The equality in Equation (14.10) leads to a useful graph, known as a scree plot, for visualizing the amount of variation in X that is captured by the jth principal component.

A **scree plot** is the plot of the sample variance of the jth principal component relative to the total sample variance in the X's (that is, the sample value of $\text{var}(PC_j)/\sum_{j=1}^{k} \text{var}(X_j)$) against the number of the principal component, j. Because this ratio has the interpretation of the R^2 of the jth principal component, the scree plot makes it possible to read off the fraction of the sample variance of the Xs explained by any particular principal component. Because the principal components are mutually uncorrelated, the cumulative sum of these ratios through the pth principal component is the fraction of the total sample variance of X explained by the first p principal components.

Figure 14.6 is the scree plot for the first 50 principal components of the 817-variable data set in Table 14.1. The first principal component explains 18% of the

---

[3]For $k = 2$, this can be verified by adding the two expressions for the variances of the principal components: $\text{var}(PC_1) + \text{var}(PC_2) = (1 + |\rho|) + (1 - |\rho|) = 2 = \text{var}(X_1) + \text{var}(X_2)$, where the final equality follows because X_1 and X_2 are standardized and thus have unit variance.

The Principal Components of X

The principal components of the k variables $X_1, \ldots, X_k$ are the linear combinations of X that have the following properties:

(i) The squared weights of the linear combinations sum to 1;

(ii) The first principal component maximizes the variance of its linear combination;

(iii) The second principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first principal component; and

(iv) More generally, the jth principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first $j - 1$ principal components.

- Assuming there is no perfect multicollinearity in X, the number of principal components is the minimum of n and k.

- The sum of the sample variances of the principal components equals the sum of the sample variances of the X's:

$$\sum_{j=1}^{\min(n,k)} \text{var}(PC_j) = \sum_{j=1}^{k} \text{var}(X_j). \tag{14.10}$$

- The ratio $\text{var}(PC_j)/\sum_{j=1}^{k} \text{var}(X_j)$ is the fraction of the total variance of the X's explained by the jth principal component. This measure is like an R^2 for the total variance of the X's.

total sample variance of the 817 X's, and the second principal component explains 11% of the total variance. Thus 29%, or more than one-fourth, of the total variance of the 817 variables is explained by just these two principal components. The first 10 principal components explain 63% of the total variance of the 817 X's, and the first 40 principal components explain 92% of the total variance.

The flattening in Figure 14.6 after the first few principal components is typical of many data sets in which the variables are highly correlated, as they are in the 817-variable school test score data set. This feature gives the scree plot its name: It looks like a cliff, with boulders, or scree, cascading into a valley.

Prediction using principal components. The fact that so much of the variation in the 817 predictors is captured by the first 10, or 50, principal components suggests that one could replace the 817 predictors with far fewer principal components and use

**Fraction of total variance
of X explained**

Principal component number
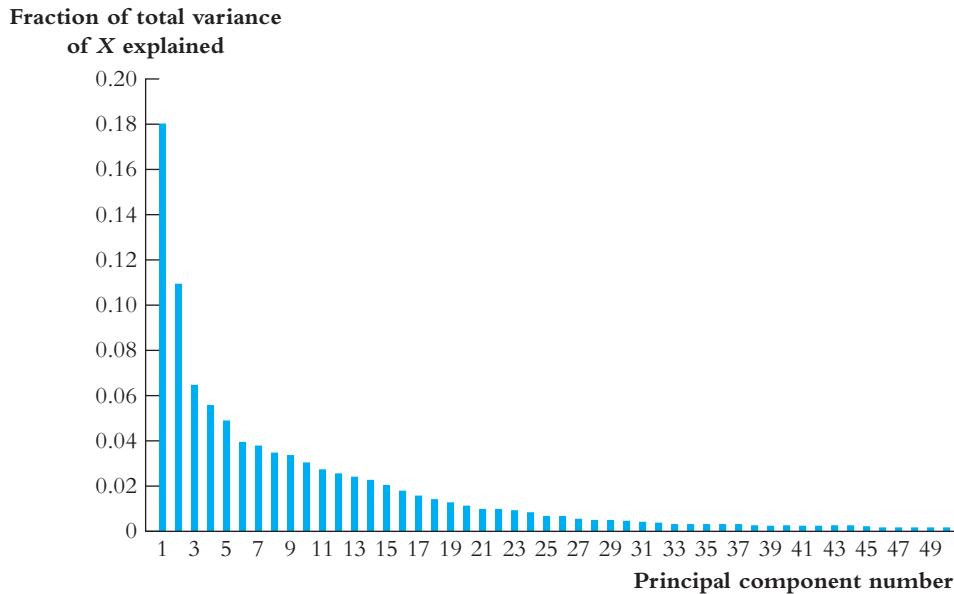
Plotted values are the fraction of the total variance of the 817 regressors explained by the indicated principal component. The first principal component explains 18% of the total variance of the 817 X's, and the first 10 principal components together explain 63% of the total variance.

those principal components as regressors. With many fewer regressors, the coefficients can be estimated using OLS.

A key question is how many principal components p to include in the regression. Like the ridge and Lasso shrinkage parameters, the number of principal components p can be estimated by minimizing the MSPE, where the MSPE is estimated by m-fold cross validation.

As discussed following Equation (14.3), computing the predicted value for an out-of-sample observation requires standardizing the observation using the in-sample mean and variance of each predictor. In the case of principal components regression, the out-of-sample values of the principal components must, in addition, be computed by applying the weights (the w's) estimated using in-sample data values to the standardized X's. The details are discussed in Appendix 14.5.

Application to School Test Scores

Figure 14.7 plots the square root of the 10-fold cross-validation estimate of the MSPE of the principal components predictor of school test scores as a function of the number p of principle components used as regressors; the principle components were computed using the same 817 predictors and 1966 observations as in Sections 14.3

Square Root of the MSPE for the Principal Components Prediction as a Function of the Number of Principal Components p Used as Predictors

The MSPE is estimated using 10-fold cross validation for the school test score data set with $k = 817$ predictors and $n = 1966$ observations.
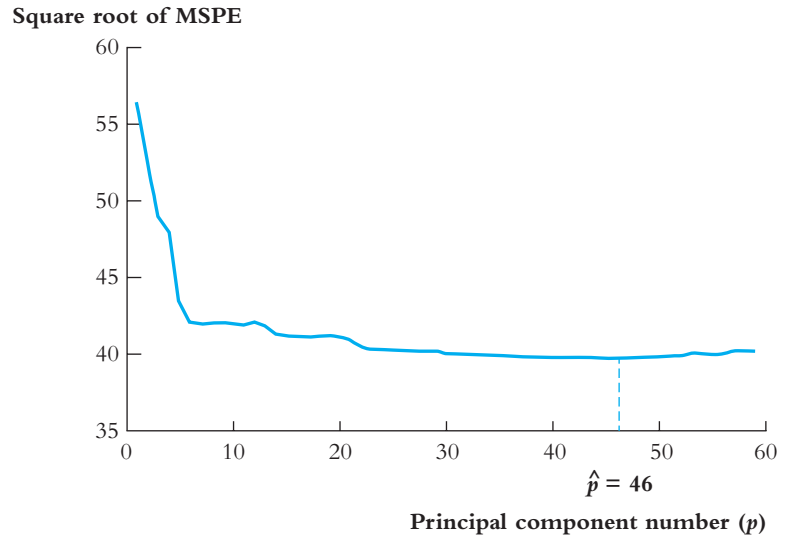
and 14.4. Initially, increasing the number of principal components used as predictors results in a sharp decline in the MSPE. After $p = 5$ principal components, the improvement slows down, and after $p = 23$ principal components, the MSPE is essentially flat in the number of predictors. The MSPE is minimized at 46 predictors, so this is the cross-validation estimate of p; that is, $\hat{p} = 46$. Using 46 principal components, the MSPE is 39.7, the same as for Lasso and just slightly more than for ridge.

14.6 Predicting School Test Scores with Many Predictors

Do the many-predictor methods improve upon test score predictions made using OLS with a small data set and, if so, how do the many-predictor methods compare? To find out, we predict school test scores using small ($k = 4$), large ($k = 817$), and very large ($k = 2065$) data sets. For the small data set, the predictions are made using OLS. For the other data sets, they are made using OLS, ridge regression, the Lasso, and principal components.

As was stressed in Section 14.2, the predictive performance that matters is performance out of sample. Because the m-fold MSPE is used to estimate the ridge and Lasso shrinkage parameters and the number of included principal components p, the MSPE no longer provides a true out-of-sample comparison among the prediction methods. We therefore have reserved half the observations for assessing the performance of the estimated models; we call these remaining observations the reserved test sample.

Specifically, we use the following procedure, explained here for ridge regression, to assess predictive performance. Using the 1966 observations in the estimation sample, we estimate the shrinkage parameter λ_{Ridge} by 10-fold cross validation; for the 817-predictor data set, this yields the estimate $\hat{\lambda}_{Ridge}$ reported in Section 14.3. Using this estimated shrinkage parameter, the ridge regression coefficients are reestimated using all 1966 observations in the estimation sample. Those estimated coefficients are then used to predict the out-of-sample values Y^{*oos} for all the observations in the reserved test sample. Analogous procedures are used for the Lasso and principal components.

Table 14.2 lists the three sets of predictors. The 4 predictors in the small set are similar to some regressors in Chapters 5–9 for the district-level test score regressions. The 817 predictors are those in Table 14.1. The very large set augments the 38 main variables in Table 14.1 with demographic data on residents in the neighborhood of the school (age distribution, sex, marital status, education, and immigrant status), as well as some binary descriptors of the school and district, for a total of 65 main variables. For the very large data set, these 65 main variables are further augmented by all interactions, squares, and cubes, for a total of 2065 predictors—more than the number of observations (1966) in the estimation sample!

TABLE 14.2   The Three Sets of Predictors, School Test Score Data Set

Small ($k = 4$)
School-level data on Student–teacher ratio
Median income of the local population
Teachers' average years of experience
Instructional expenditures per student

Large ($k = 817$)
The full data set in Table 14.1

Very Large ($k = 2065$)
The main variables are those in Table 14.1, augmented with the 27 variables below, for a total of 65 main variables, 5 of which are binary:

Population	Immigration status variables (4)
Age distribution variables in local population (8)	Charter school (binary)
Fraction of local population that is male	School has full-year calendar (binary)
Local population marital status variables (3)	School is in a unified school district (large city) (binary)
Local population educational level variables (4)	School is in Los Angeles (binary)
Fraction of local housing that is owner occupied	School is in San Diego (binary)

+ Squares and cubes of the 60 nonbinary variables ($60 + 60$)
+ All interactions of the nonbinary variables ($60 \times 59/2 = 1770$)
+ All interactions between the binary variables and the nonbinary demographic variables ($5 \times 22 = 110$)
Total number of variables $= 65 + 60 + 60 + 1770 + 110 = 2065$

TABLE 14.3	Out-of-Sample Performance of Predictive Models for School Test Scores			
Predictor Set	OLS	Ridge Regression	Lasso	Principal Components
Small ($k = 4$)				
Estimated λ or p	—	—	—	—
In-sample root MSPE	53.6	—	—	—
Out-of-sample root MSPE	52.9	—	—	—
Large ($k = 817$)				
Estimated λ or p	—	2233	4527	46
In-sample root MSPE	78.2	39.5	39.7	39.7
Out-of-sample root MSPE	64.4	38.9	39.1	39.5
Very large ($k = 2065$)				
Estimated λ or p	—	3362	4221	69
In-sample root MSPE	—	39.2	39.2	39.6
Out-of-sample root MSPE	—	39.0	39.1	39.6

Notes: The in-sample MSPE is the 10-fold cross-validation estimate computed using the 1966 observations in the estimation sample. For the many-predictor methods, the shrinkage parameter or p was estimated by minimizing this in-sample MSPE. The out-of-sample MSPE is a split-sample estimate, computed with the 1966 observations in the reserved test sample and using the model estimated from the full estimation sample.

The results of this comparison are summarized in Table 14.3. Four features stand out. First, the MSPE of OLS is much less using the small data set than using the large data set (OLS cannot be computed in the very large data set because $k > n$). When there are many regressors, OLS is unable to use the additional information to improve out-of-sample prediction.

Second, for the many-predictor methods, there are substantial gains from increasing the number of predictors from 4 to 817, with the square root of the MSPE falling by roughly one-fourth. There are no further gains, however, from going to the very large set of regressors.

Third, the in-sample estimates of MSPE (the 10-fold cross-validation estimates) are similar to the out-of-sample estimates. In fact, the out-of-sample MSPEs are slightly less than the in-sample MSPEs. There are two reasons for this surprising result. First, the 10-fold MSPE uses only 90% of the data for estimating the coefficients at any one time (that is, $0.9 \times 1966 = 1769$ observations), whereas the coefficients used for the out-of-sample estimate of the MSPE are estimated using all

TABLE 14.4 Coefficients on Selected Standardized Regressors, 4- and 817-Variable Data Sets					
Predictor	**k = 4**	**k = 817**			
	OLS	**OLS**	**Ridge Regression**	**Lasso**	**Principal Components**
Student–teacher ratio	4.51	118.03	0.31	0	0.25
Median income of the local population	34.46	−21.73	0.38	0	0.30
Teachers' average years of experience	1.00	−79.59	−0.11	0	−0.17
Instructional expenditures per student	0.54	−1020.77	0.11	0	0.19
Student–teacher ratio × Instruction expenditures per student		−89.79	0.72	2.31	0.84
Student–teacher ratio × Fraction of English learners		−81.66	−0.87	−5.09	−0.55
Free or reduced-price lunch × Index of part-time teachers		29.42	−0.92	−8.17	−0.95

Notes: The index of part-time teachers measures the fraction of teachers who work part-time. For OLS, ridge, and Lasso, the coefficients in Table 14.4 are produced directly by the estimation algorithms. For principal components, the coefficients in Table 14.4 are computed from the principal component regression coefficients (the γ's in Equation ((14.13)), combined with the principal component weights. The formula for the β coefficients for principal components is presented using matrix algebra in Appendix 19.7.

1966 observations in the estimation sample. As a result, those latter coefficient estimates are more precise. Second, there is random sampling variation in both estimates. The more general point is that the in-sample 10-fold MSPEs provide a good guide to the out-of-sample MSPE.

Fourth, the MSPE in the reserved test sample is generally similar for all the many-predictor methods. This is not always the case; it just happens to be so in this application. For these data, ridge regression has a slight edge, and the lowest out-of-sample MSPE is obtained using ridge in the large data set.

Table 14.4 lists the coefficients on 7 of the variables in the 817-predictor data set; 4 of the 7 are those in the small data set. Although none of these coefficients has a causal interpretation, comparing them across the different methods and data sets gives insights into how the various methods work. Because the regressors are standardized, all the coefficients have the same units, points on the test per standard deviation of the original predictor.[4]

Table 14.4 has several striking features. For the small model, the magnitudes of the coefficients accord with the findings of Chapter 9 using the district-level data; for example, a one-standard-deviation greater value of median income predicts a

---

[4]The coefficients on the X's in the principal components column are obtained by combining the two steps of prediction using principal components. Specifically, the principal components are linear combinations of the X's, and the principal components regression model is a linear combination of the principal components. Thus the prediction can be written as a linear combination of the X's, where the weights involve both the principal components weights and the regression weights. The relevant formulas are given in Appendix 19.7.

34-point higher score on the test (the standard deviation of the test scores across schools is 64 points). In the large data set, however, many of the OLS coefficients are extremely large, and the pattern is erratic. With many regressors, OLS can fit individual observations by estimating large coefficients on specific variables, and this seems to be what is happening. This overfitting is why the predictive performance of OLS deteriorates moving from the small to the large data set. In contrast, the estimated coefficients for the many-predictor methods are substantially smaller and do not exhibit wild values. For the seven predictors in the table, the ridge and principal components coefficients are numerically similar. The Lasso coefficients, however, differ substantially from the ridge and principal components coefficients. Most notably, many of the Lasso coefficients (92% in all) are 0, including the coefficients on the four variables in the small data set. For the three coefficients in the table that are nonzero, they have the same sign as the ridge and principal components coefficients but are much larger, an empirical illustration of the tendency of Lasso to shrink more than ridge for small coefficients but to shrink less than ridge for large ones.

Another way to compare predictive models is to look at their predictions. Figure 14.8 shows scatterplots of the four sets of predictions for the 817-variable model, where the predictions are for the 1966 observations in the reserved test set.

FIGURE 14.8   Scatterplots for Out-of-Sample Predictions Using the 817-Predictor Data Set
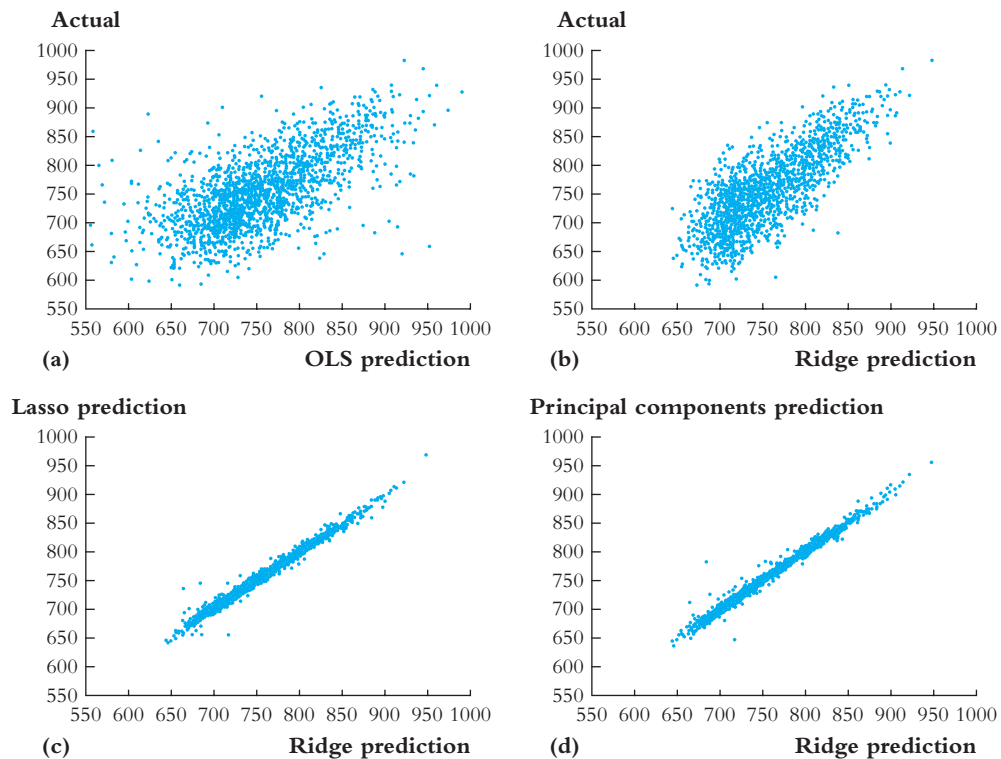
(a) Actual versus OLS, (b) actual versus ridge, (c) Lasso versus ridge, and (d) principal components versus ridge.

Specifically, Figure 14.8a shows a scatterplot of the actual test scores versus the OLS predictions, and Figure 14.8b is the scatterplot of the actual test scores versus the ridge predictions. Figure 14.8c and Figure 14.8d are scatterplots of the Lasso versus the ridge predictions and the principal components versus the ridge predictions, respectively.

In Figure 14.8a and Figure 14.8b, the tighter the spread of the scatter along the 45° line, the better the prediction. Ridge has a tighter scatter than OLS, and it makes better out-of-sample predictions. (These scatterplots understate the improvement of ridge over OLS because some of the OLS predictions are outside the vertical scale of the plot.)

The clustering of the points along the 45° line in Figure 14.8c and Figure 14.8d indicate that the ridge, Lasso, and principal components predictions are generally quite similar. Still, one can see quite a few schools for which the predictions differ by at least 15 points, a substantial amount. Thus, while the three models have quite similar performance as measured by the MSPE (Table 14.3), for any given school the predictions can differ meaningfully.

The most important conclusion from this application is that for the large data set the many-predictor methods succeed where OLS fails. The reason for this success is that the many-predictor methods allow the coefficient estimates to be biased in a way that reduces their variance by enough to compensate for the increased bias. Another important conclusion is that the m-fold MSPE is close to the MSPE computed using the reserved test sample. One finding that does not generalize, however, is that the three methods happen to perform equally well in these data.

14.7 Conclusion

The coefficients in the predictive regression model do not have a causal interpretation. This does not matter, however, when the goal is prediction; the aim simply is to make out-of-sample predictions that are as accurate as possible, where accuracy is measured by the MSPE.

This chapter presented three methods for making predictions with many predictors. These methods provide different ways to overcome the poor performance of OLS predictions when the number of regressors is large relative to the sample size. The methods covered in this chapter—ridge regression, the Lasso, and principal components regression—all introduce bias into the estimator of the β's. However, this bias is introduced in a way that reduces the variance of the prediction by enough to yield a smaller MSPE.

Although ridge regression, the Lasso, and principal components regression all reduce variance by introducing bias, they do so in quite different ways. The Lasso sets many of the coefficients exactly to 0, in effect discarding those predictors. This approach works well when the oracle prediction model is sparse or approximately so. Principal components regression is most appropriate when the predictors, or groups of predictors, are highly correlated, in which case most of the variation in the regressors can be captured by a relatively small number of linear combinations of the

Text as Data

Text contains a lot of information! That is why you read the newspaper or posts on social media. That information keeps you abreast of political developments and helps you decide what to do tonight. By reading these sources, you use textual information—textual data—to make predictions about outcomes that are relevant to you.

A major accomplishment of statistics and machine learning is figuring out how to use computers to read text and to make predictions using textual data. At a conceptual level, it is a big leap to go from analyzing numbers to analyzing texts. The key step in doing so is turning text data into numerical data.

One way to turn text data into numerical data is to develop a list of words or phrases and then count the number of times that these words or phrases occur in a given text excerpt (for example, a newspaper article or blog post). These counts of words or phrases are numerical data that summarize the text. The unit of observation is the text excerpt, and the number of observations is the number of excerpts analyzed. This method of distilling a set of texts into occurrence counts of words or phrases was developed by Frederick Mosteller and David Wallace (1963) and is the basis of the field of stylometrics (see the box titled "Who Invented Instrumental Variables Regression?" in Chapter 12).

The approach of distilling text into counts of words or phrases has its own jargon. The list of words in a text is called a *bag of words*. The list of words and phrases of interest is called the *dictionary*. The dictionary may include only the words or phrases that are relevant to the prediction problem at hand, or it may contain all the words in the bag of words, excluding (for example) articles, pronouns, and conjunctions.

The word counts now can be used as predictors (X's) to predict a variable Y of interest. Thus, this bag-of-words approach has turned a seemingly intractable problem of combining text and numerical data into a regression problem.

Because the dictionary typically contains many words, the number of predictors can be large relative to the number of texts (n). If so, OLS would tend to produce poor predictions, but the methods in this chapter can be applied directly. For example, principal components analysis can be a useful tool in this setting because often words appear in groups (think of the words used in an article about a baseball game compared to an article about macroeconomic conditions). Putting all these pieces together results in predictive models that take text as the input and yield a prediction of Y.

variables—specifically, by their first few principal components. Because these principal components are relatively few in number, they can be the regressors used in a multiple regression model estimated by OLS. Ridge regression shrinks the OLS estimates toward 0 but does not rely on there being sparsity or on the regressors being highly correlated; thus it provides a useful approach when the regressors are not highly correlated and there is no a priori reason to assume sparsity. As it happens, in the school test score data, the three methods perform similarly, but this coincidence does not occur in general.

As discussed in Section 14.1, making predictions using many predictors that take on numerical values is only one of the opportunities provided by the methods of machine learning. For example, the box "Text as Data" describes how the

tools of this section can be used to analyze text data. Similarly, principal components analysis and its extensions can be used to turn images into numerical data, which then can be analyzed by the many-predictor methods described in this chapter. While many of the procedures in machine learning are new and the computational algorithms and tools are sophisticated, at their core are the key ideas of regression analysis, estimation, and testing that are at the heart of Parts I–III of this text.

The use of machine learning in economics is young, and many exciting applications await. For some examples and further reading, see Jean et. al. (2016) (predicting poverty using satellite imagery), Davis and Heller (2017) (examining treatment heterogeneity for a summer jobs program), and Kleinberg et. al. (2018) (application of machine learning to criminal sentencing).[5]

Summary

1. The goal of prediction is to make accurate predictions for out-of-sample observations—that is, for observations not used to estimate the prediction model.

2. The coefficients in prediction models do not have a causal interpretation.

3. One of the opportunities provided by big data sets is making predictions using many predictors. However, OLS works poorly for prediction when the number of regressors is large relative to the sample size.

4. The shortcomings of OLS can be overcome by using prediction methods that have lower variance at the cost of introducing estimator bias. These many-predictor methods can produce predictions with substantially better predictive performance than OLS, as measured by the MSPE.

5. Ridge regression and the Lasso are shrinkage estimators that minimize a penalized sum of squared residuals. The penalty introduces a cost to estimating large values of the regression coefficient. The weight on the penalty (the shrinkage parameter) can be estimated by minimizing the m-fold cross-validation estimator of the MSPE.

6. The principal components of a set of correlated variables capture most of the variation in those variables in a reduced number of linear combinations. Those principal components can be used in a predictive regression, and the number of principal components included can be estimated by minimizing the m-fold cross-validation MSPE.

---

[5]The field of machine learning is growing rapidly. A textbook introduction to this area, which is accessible to students after completing Parts I–III of this text, is Gareth James et al., *An Introduction to Statistical Learning* (2013).

Key Terms

mean squared prediction error (476)

oracle prediction (476)

first least squares assumption for
     prediction (477)

standardized predictive regression
     model (477)

shrinkage estimator (479)

m-fold cross validation (480)

ridge regression (482)

penalty term (482)

penalized sum of squared
     residuals (482)

sparse model (486)

Lasso (486)

principal components (490)

scree plot (492)

MyLab Economics Can Help You Get a Better Grade

MyLab Economics If your exam were tomorrow, would you be ready? For each chapter, **MyLab Economics** Practice Tests and Study Plan help you prepare for your exams. You can also find the Exercises and all Review the Concepts Questions available now in **MyLab Economics**. To see how it works, turn to the **MyLab Economics** spread on the inside front cover of this text and then go to **www.pearson.com/mylab/economics**.

For additional Empirical Exercises and Data Sets, log on to the Companion Website at **www.pearsonhighered.com/stock_watson**.

Review the Concepts

14.1 Using data from a random sample of elementary schools, a researcher regresses average test scores on the fraction of students who qualify for reduced-price meals. The regression indicates a negative coefficient that is highly statistically significant and yields a high R^2. Is this regression useful for determining the causal effect of school meals on student test scores? Why or why not? Is this regression useful for predicting test scores? Why or why not?

14.2 Cross-validation uses in-sample observations. How does it estimate the MSPE for out-of-sample observations, even though the econometrician does not have those observations?

14.3 Regression coefficients estimated using shrinkage estimators are biased. Why might these biased estimators yield more accurate predictions than an unbiased estimator?

14.4 Ridge regression and Lasso are two regression estimators based on penalization. Explain how they are similar and how they differ.

14.5 Suppose a data set with 10 variables produces a scree plot that is flat. What does this tell you about the correlation of the variables? What does this suggest about the usefulness of using the first few principal components of these variables in a predictive regression?

Exercises

14.1 A researcher is interested in predicting average test scores for elementary schools in Arizona. She collects data on three variables from 200 randomly chosen Arizona elementary schools: average test scores (*TestScore*) on a standardized test, the fraction of students who qualify for reduced-priced meals (*RPM*), and the average years of teaching experience for the school's teachers (*TExp*). The table below shows the sample means and standard deviations from her sample.

Variable	Sample Mean	Sample Standard Deviation
TestScore	750.1	65.9
RPM	0.60	0.28
TExp	13.2	3.8

After standardizing *RPM* and *TEXP* and subtracting the sample mean from *TestScore*, she estimates the following regression:

$$\widehat{TestScore} = -48.7 \times RPM + 8.7 \times TExp, SER = 44.0$$

a. You are interested in using the estimated regression to predict average test scores for an out-of-sample school with $RPM = 0.52$ and $TEXP = 11.1$.

   i. Compute the transformed (standardized) values of *RPM* and *TEXP* for this school; that is, compute the X^{oos} values from the X^{*oos} values, as discussed preceding Equation (14.2).

   ii. Compute the predicted value of average test scores for this school.

b. The actual average test score for the school is 775.3. Compute the error for your prediction.

c. The regression shown above was estimated using the standardized regressors and the demeaned value of *TestScore*. Suppose the regression had been estimated using the raw data for *TestScore*, *RMP*, and *TExp*. Calculate the values of the regression intercept and slope coefficients for this regression.

d. Use the regression coefficients that you computed in (c) to predict average test scores for an out-of-sample school with $RPM = 0.52$ and $TExp = 11.1$. Verify that the prediction is identical to the prediction you computed in (a.ii).

14.2 A school principal is trying to raise funds so that all her students will receive reduced-price meals; currently, only 40% qualify for reduced-priced meals. Can she use the regression in Exercise 14.1 to estimate the effect of the new policy on test scores? Explain why or why not.

14.3   Describe the relationship, if any, between the standard error of a regression and the square root of the MSPE of the regression's out-of-sample predictions.

14.4   A large online retailer sells thousands of products. The retailer has detailed data on the products purchased by each of its customers. Explain how you would use these data to predict the next product purchased by a randomly selected customer.

14.5 Y is a random variable with mean $\mu = 2$ and variance $\sigma^2 = 25$.

   a.   Suppose you know the value of μ.

      i.   What is the best (lowest MSPE) prediction of the value of Y? That is, what is the oracle prediction of Y?

      ii.   What is the MSPE of this prediction?

   b.   Suppose you don't know the value of μ but you have access to a random sample of size $n = 10$ from the same population. Let $\overline{Y}$ denote the sample mean from this random sample. You predict the value of Y using $\overline{Y}$.

      i.   Show that the prediction error can be decomposed as $Y - \overline{Y} = (Y - \mu) - (\overline{Y} - \mu)$, where $(Y - \mu)$ is the prediction error of the oracle predictor and $(\mu - \overline{Y})$ is the error associated with using $\overline{Y}$ as an estimate of μ.

      ii.   Show that $(Y - \mu)$ has a mean of 0, that $(\overline{Y} - \mu)$ has a mean of 0, and that $Y - \overline{Y}$ has a mean of 0.

      iii.   Show that $(Y - \mu)$ and $(\overline{Y} - \mu)$ are uncorrelated.

      iv.   Show that the MSPE of $\overline{Y}$ is MSPE $= E(Y - \mu)^2 + E(\overline{Y} - \mu)^2 = \text{var}(Y) + \text{var}(\overline{Y})$.

      v.   Show that MSPE $= 25(1 + 1/10) = 27.5$.

14.6   In Exercise 14.5(b), suppose you predict Y using $\overline{Y}/2$ instead of $\overline{Y}$.

   a.   Compute the bias of the prediction.

   b.   Compute the mean of the prediction error.

   c.   Compute the variance of the prediction error.

   d.   Compute the MSPE of the prediction.

   e.   Does $\overline{Y}/2$ produce a prediction with a lower MSPE than the $\overline{Y}$ prediction?

   f.   Suppose $\mu = 10$ (instead of $\mu = 2$). Does $\overline{Y}/2$ produce a prediction with a lower MSPE than the $\overline{Y}$ prediction?

   g.   In a realistic setting, the value of μ is unknown. What advice would you give someone who is deciding between using $\overline{Y}$ and $\overline{Y}/2$?

14.7   In Exercise 14.5(b), suppose you predict Y using $\overline{Y} - 1$ instead of $\overline{Y}$.

   a.   Compute the bias of the prediction.

   b.   Compute the mean of the prediction error.

   c.   Compute the variance of the prediction error.

   d.   Compute the MSPE of the prediction.

e. Does $\overline{Y} - 1$ produce a prediction with a lower MSPE than the $\overline{Y}$ prediction?

f. Does $\overline{Y} - 1$ produce a prediction with a lower MSPE than the $\overline{Y}/2$ prediction from Exercise 14.6?

14.8 Let X and Y be two random variables. Denote the mean of Y given $X = x$ by $\mu(x)$ and the variance of Y by $\sigma^2(x)$.

a. Show that the best (minimum MSPE) prediction of Y given $X = x$ is $\mu(x)$ and the resulting MSPE is $\sigma^2(x)$. (*Hint:* Review Appendix 2.2.)

b. Suppose X is chosen at random. Use the result in (a) to show that the best prediction of Y is $\mu(X)$ and the resulting MSPE is $E[Y - \mu(X)]^2 = E[\sigma^2(X)]$.

14.9 You have a sample of size $n = 1$ with data $y_1 = 2$ and $x_1 = 1$. You are interested in the value of β in the regression $Y = X\beta + u$. (Note there is no intercept.)

a. Plot the sum of squared residuals $(y_1 - bx_1)^2$ as function of b.

b. Show that the least squares estimate of β is $\hat{\beta}^{OLS} = 2$.

c. Using $\lambda_{Ridge} = 1$, plot the ridge penalty term $\lambda_{Ridge}b^2$ as a function of b.

d. Using $\lambda_{Ridge} = 1$, plot the ridge penalized sum of squared residuals $(y_1 - bx_1)^2 + \lambda_{Ridge}b^2$.

e. Find the value of $\hat{\beta}^{Ridge}$.

f. Using $\lambda_{Ridge} = 0.5$, repeat (c) and (d). Find the value of $\hat{\beta}^{Ridge}$.

g. Using $\lambda_{Ridge} = 3$, repeat (c) and (d). Find the value of $\hat{\beta}^{Ridge}$.

h. Use the graphs that you produced in (a)–(d) for the various values of λ_{Ridge} to explain why a larger value of λ_{Ridge} results in more shrinkage of the OLS estimate.

14.10 You have a sample of size $n = 1$ with data $y_1 = 2$ and $x_1 = 1$. You are interested in the value of β in the regression $Y = X\beta + u$. (Note there is no intercept.)

a. Plot the sum of squared residuals $(y_1 - bx_1)^2$ as function of b.

b. Show that the least squares estimate of β is $\hat{\beta}^{OLS} = 2$.

c. Using $\lambda_{Lasso} = 1$, plot the Lasso penalty term $\lambda_{Lasso}|b|$ as a function of b.

d. Using $\lambda_{Lasso} = 1$, plot the Lasso penalized sum of squared residuals $(y_1 - bx_1)^2 + \lambda_{Lasso}|b|$.

e. Find the value of $\hat{\beta}^{Lasso}$.

f. Using $\lambda_{Lasso} = 0.5$, repeat (c) and (d). Find the value of $\hat{\beta}^{Lasso}$.

g. Using $\lambda_{Lasso} = 5$, repeat (c) and (d). Find the value of $\hat{\beta}^{Lasso}$.

h. Use the graphs that you produced in (a)–(d) for the various values of λ_{Lasso} to explain why a larger value of λ_{Lasso} results in more shrinkage of the OLS estimate.

14.11 Let X_1 and X_2 be two positively correlated random variables, both with variance 1.

    a. (Requires calculus) The first principal component, PC_1, is the linear combination of X_1 and X_2 that maximizes $\text{var}(w_1 X_1 + w_2 X_2)$, where $w_1^2 + w_2^2 = 1$. Show that $PC_1 = (X_1 + X_2)/\sqrt{2}$. (*Hint:* First derive an expression for $\text{var}(w_1 X_1 + w_2 X_2)$ as a function of w_1 and w_2.)

    b. The second principal component is $PC_2 = (X_1 - X_2)/\sqrt{2}$. Show that $\text{cov}(PC_1, PC_2) = 0$.

    c. Show that $\text{var}(PC_1) = 1 + \rho$ and $\text{var}(PC_2) = 1 - \rho$, where $\rho = \text{cor}(x_1, x_2)$.

14.12 Consider the fixed-effects panel data model $Y_{jt} = \alpha_j + u_{jt}$ for $j = 1, \ldots, k$ and $t = 1, \ldots, T$. Assume that u_{jt} is i.i.d. across entities j and over time t with $E(u_{jt}) = 0$ and $\text{var}(u_{jt}) = \sigma_u^2$.

    a. The OLS estimator of α_j is the value of a_j that makes the sum of squared residuals $\sum_{j=1}^{k}\sum_{t=1}^{T}(Y_{jt} - a_j)^2$ as small as possible. Show that the OLS estimator is $\hat{\alpha}_j = \overline{Y}_j = \dfrac{1}{T}\sum_{t=1}^{T} Y_{jt}$.

    b. Show that

        i. $\hat{\alpha}_j$ is an unbiased estimator of α_j.

        ii. $\text{var}(\hat{\alpha}_j) = \sigma_u^2/T$.

        iii. $\text{cov}(\hat{\alpha}_i, \hat{\alpha}_j) = 0$ for $i \neq j$.

    c. You are interested in predicting an out-of-sample value for entity j—that is, for $Y_{j,T+1}$—and use $\hat{\alpha}_j$ as the predictor. Show that $\text{MSPE} = \sigma_u^2 + \sigma_u^2/T$.

    d. You are interested in predicting an out-of-sample value for a randomly selected entity—that is, for $Y_{j,T+1}$, where j is selected at random. You again use $\hat{\alpha}_j$ as the predictor. Show the $\text{MSPE} = \sigma_u^2 + \sigma_u^2/T$.

    e. The total number of in-sample observations is $n = kT$. Show that in both (c) and (d) $\text{MSPE} = \sigma_u^2(1 + k/n)$.

Empirical Exercises

E14.1 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find a data set **CASchools_EE14_InSample** that contains a subset of $n = 500$ schools from the data set used in this chapter. Included are data on test scores and 20 of the primitive predictor variables; see **CASchools_EE141_Description** for a description of the variables. In this exercise, you will construct prediction models like those described in the text and use these models to predict test scores for 500 out-of-sample schools. (Please read **EE141_SoftwareNotes** on the text website before solving the exercise.)

    a. From the 20 primitive predictors, construct squares of all the predictors, along with all of the interactions (that is, the cross products $X_{ji}X_{ki}$ for all j and k). Collect the 20 primitive predictors, their squares,

and all interactions into a set of k predictors. Verify that you have $20 + 20 + (20 \times 19)/2 = 230$ predictors. One of the primitive predictors is the binary variable *charter_s*. Drop the predictor $(charter\_s)^2$ from the list of 230 predictors, leaving 229 predictors for the analysis. Why should $(charter\_s)^2$ be dropped from the original list of predictors?

b. Compute the sample mean and standard deviation of each of the predictors, and use these to compute the standardized regressors. Compute the sample mean of *TestScore*, and subtract the sample mean from *TestScore* to compute its demeaned value.

c. Using OLS, regress the demeaned value of *TestScore* on the standardized regressors.

   i. Did you include an intercept in the regression? Why or why not?

   ii. Compute the standard error of the regression.

d. Using ridge regression with $\lambda_{Ridge} = 300$, regress the demeaned value of *TestScore* on the standardized regressors. Compare the OLS and ridge estimates of the standardized regression coefficients.

e. Using Lasso with $\lambda_{Lasso} = 1000$, regress the demeaned value of *TestScore* on the standardized regressors. How many of the estimated Lasso coefficients are different from 0? Which predictors have a nonzero coefficient.

f. Compute the scree plot for the 229 predictors. How much of the variance in the standardized regressors is captured by the first principal component? By the first two principal components? By the first 15 principal components?

g. Compute 15 principal components from the 229 predictors. Regress the demeaned value of *TestScore* on the 15 principal components.

h. On the text website, you will find a data set **CASchools_EE14_OutOfSample** that contains data from another $n = 500$ schools.

   i. Predict the average test score for each of these 500 schools using the OLS, ridge, Lasso, and principal components prediction models that you estimated in (c), (d), (e), and (g). Compute the root mean square prediction error for each of the methods.

   ii. Construct four scatter plots like those in Figure 14.8. What do you learn from the plots?

i. Estimate λ_{Ridge}, λ_{Lasso}, and the number of principal components using 10-fold cross validation from the in-sample data set.

j. Use the estimated values of λ_{Ridge}, λ_{Lasso}, and the number of principal components from (i) to construct predictions of test scores for the out-of-sample schools. Are these predictions more accurate than the predictions you computed in (h)? Is the difference in line with what you expected from the cross-validation calculations in (i)?

14.1 The California School Test Score Data Set

The test scores used in this chapter are from the California Standards Tests (part of California's Standardized Testing and Reporting program) given to fifth-grade students in the spring of 2013. The average test score for each of California's schools is available from the California Department of Education, where you can also find much of the other school and district data used in the chapter. The remaining school and district data were obtained from ED-Data (www.ed-data.org). All school and district data are for the 2012–13 academic year. In addition to school and district data, demographic data for 2013 are constructed from the census tracts making up the zip code for each school. These data are available from the American Community Survey (see factfinder.census.gov). More detail is available in the replication files for the chapter at **http://www.pearsonhighered.com/stock_watson/.**

14.2 Derivation of Equation (14.4) for $k = 1$

With a single regressor, the OLS prediction in the standardized predictive regression model (Equation (14.2)) for a given value $X = x$ is $\hat{Y}(x) = \hat{\beta}x$. The second term in Equation (14.3) is $E[(\hat{\beta} - \beta)X^{OOS}]^2 = E(\hat{\beta} - \beta)^2E(X^{OOS})^2 = E(\hat{\beta} - \beta)^2$, where the first equality uses the independence of $\hat{\beta}$ and X^{oos} ($\hat{\beta}$ is estimated using the in-sample data) and the second equality uses the fact that the regressors are standardized, so $E(X^{OOS})^2 = \text{var}(X^{OOS}) = 1$. Because the OLS estimator is unbiased in the prediction model, $E(\hat{\beta} - \beta)^2 = \text{var}(\hat{\beta}) = \sigma_u^2/(n\sigma_X^2) = \sigma_u^2/n$, where the second equality uses the large-n formula for the variance of the OLS estimator under homoskedasticity in Equation (5.27) and the final equality uses the fact that $\sigma_X^2 = 1$ because the regressors in Equation (14.2) are standardized using the population mean and variance. It follows from Equation (14.3) that, with $k = 1$ under homoskedasticity, the MSPE of OLS $\cong (1 + 1/n)\sigma_u^2$ for large n, which is Equation (14.4) with $k = 1$.

14.3 The Ridge Regression Estimator When $k = 1$

When $k = 1$, the ridge estimator minimizes the penalized sum of squares, $S^{Ridge}(b; \lambda_{Ridge}) = \sum_{i=1}^{n}(Y_i - bX_i)^2 + \lambda_{Ridge}b^2$. Taking the derivative of $S^{Ridge}(b; \lambda_{Ridge})$ with respect to b and setting the derivative equal to 0 yields $-\sum_{i=1}^{n}X_i(Y_i - \hat{\beta}^{Ridge}X_i) + \lambda_{Ridge}\hat{\beta}^{Ridge} = 0$. Solving for $\hat{\beta}^{Ridge}$ yields $\hat{\beta}^{Ridge} = \sum_{i=1}^{n}X_iY_i/(\sum_{i=1}^{n}X_i^2 + \lambda_{Ridge}) = (1 + \lambda_{Ridge}/\sum_{i=1}^{n}X_i^2)^{-1}\hat{\beta}$, where $\hat{\beta} = \sum_{i=1}^{n}X_iY_i/\sum_{i=1}^{n}X_i^2$ is the OLS estimator.

14.4 The Lasso Estimator When $k = 1$

When $k = 1$, the Lasso minimizes the penalized sum of squared residuals, $S^{Lasso}(b; \lambda_{Lasso}) = \sum_{i=1}^{n}(Y_i - bX_i)^2 + \lambda_{Lasso}|b|$. Inspection of Figure 14.3 shows that $\hat{\beta}$ and $\hat{\beta}^{Lasso}$ must have the same sign when $k = 1$. Suppose $\hat{\beta}$ is positive. Then, over the relevant range $b \geq 0$, the Lasso minimizes $\sum_{i=1}^{n}(Y_i - bX_i)^2 + \lambda_{Lasso}b$, and its derivative with respect to b is $-2\sum_{i=1}^{n}X_i(Y_i - bX_i) + \lambda_{Lasso}$. For $\hat{\beta}^{Lasso} > 0$, setting this derivative equal to 0 implies $-2\sum_{i=1}^{n}X_i(Y_i - \hat{\beta}^{Lasso}X_i) + \lambda_{Lasso} = 0$; otherwise, $\hat{\beta}^{Lasso} = 0$. Solving for $\hat{\beta}^{Lasso}$ yields

$$\hat{\beta}^{Lasso} = \max\left(\hat{\beta} - \tfrac{1}{2}\lambda_{Lasso}/\sum_{i=1}^{n}X_i^2, 0\right) \text{ when } \hat{\beta} \geq 0. \tag{14.11}$$

Similar reasoning shows that $\hat{\beta}^{Lasso} = \min(\hat{\beta} + \tfrac{1}{2}\lambda_{Lasso}/\sum_{i=1}^{n}X_i^2, 0)$ when $\hat{\beta} < 0$.

14.5 Computing Out-of-Sample Predictions in the Standardized Regression Model

The estimators of this chapter are all computed using the standardized predictive regression model in Equation (14.2). Computing the prediction for an out-of-sample observation entails first standardizing the out-of-sample predictors, then computing the demeaned out-of-sample prediction, then adding back in the in-sample mean of Y. These transformations must all be done using the same means, variances, and weights for the out-of-sample data as for the in-sample data. Details are provided first for ridge regression and the Lasso, and then for principal components regression.

Out-of-Sample Predictions Using the Standardized Regression Model of Equation (14.2) (Ridge and Lasso)

Following Section 14.2, let $X_1^{*oos}, \ldots, X_k^{*oos}$ denote an out-of-sample observation on the original, untransformed values of the k predictors, and let Y^{*oos} denote the out-of-sample observation on the variable to be predicted. The transformed out-of-sample value of the jth predictor is $X_j^{oos} = (X_j^{*oos} - \overline{X}_j^*)/s_{X_j^*}$, where $\overline{X}_j^*$ and $s_{X_j^*}$ are the in-sample mean and standard deviation of the jth predictor. Let $\widetilde{\beta}_j$ be some estimator of β_j, e.g., the ridge regression or Lasso estimator. Then the predicted value of the original dependent variable in terms of the original predictors is

$$\hat{Y}^{*oos} = \overline{Y}^* + \sum_{j=1}^{k}\widetilde{\beta}_j\left(\frac{X_j^{*oos} - \overline{X}_j^*}{s_{X_j^*}}\right), \tag{14.12}$$

where $\overline{Y}^*, \overline{X}_j^*, s_{X_j^*},$ and $\widetilde{\beta}_j$ $(j = 1, \ldots, k)$ are all computed using the estimation sample.

Out-of-Sample Predictions Using Principal Components Regression

To compute the predicted value for an out-of-sample observation using principal components regression, it is necessary, in addition, to compute the out-of-sample values of the principal components using the in-sample weights. Let γ denote the coefficients in the regression of Y on the first p principal components:

$$Y_i = \gamma_1 PC_{1i} + \gamma_2 PC_{2i} + \ldots + \gamma_p PC_{pi} + v_i, \tag{14.13}$$

where v_i is an error term. The prediction of Y^{*oos} is computed in the following steps:

1. Compute the principal components in the estimation sample:

   a. Compute the demeaned Y and standardized X for the in-sample observations on Y^* and X^* as described preceding Equation (14.2).

   b. Compute the in-sample principal components of X; call these $PC_1, \ldots, PC_{min(n,k)}$.

2. Given p, estimate the regression coefficients in Equation (14.13); call these estimates $\hat{\gamma}_1^{PC}, \ldots, \hat{\gamma}_p^{PC}$.

3. Compute the out-of-sample values of the principal components:

   a. Standardize the out-of-sample predictors X^{*oos} using the in-sample mean and standard deviation from step 1(a). Denote this transformed observation as X^{oos}.

   b. Compute the principal components for the out-of-sample observation using the in-sample weights from step 1(b); call these $PC_1^{oos}, \ldots, PC_p^{oos}$.

4. Compute the predicted value for the out-of-sample observation as $\hat{Y}^{*oos} = \overline{Y}^* + \sum_{j=1}^{p} \hat{\gamma}_j^{PC} PC_j^{oos}$.

Introduction to Time Series Regression and Forecasting

Time series data—data collected for a single entity at multiple points in time—can be used to answer quantitative questions for which cross-sectional data are inadequate. One such question is, what is the causal effect on a variable of interest, Y, of a change in another variable, X, over time? In other words, what is the *dynamic* causal effect on Y of a change in X? For example, what is the effect on traffic fatalities of a law requiring passengers to wear seatbelts, both initially and subsequently, as drivers adjust to the law? Another such question is, what is your best forecast of the value of some variable at a future date? For example, what is your best forecast of next month's unemployment rate, interest rates, or stock prices? Both of these questions—one about dynamic causal effects, the other about economic forecasting—can be answered using time series data.

This chapter and Chapters 16 and 17 introduce techniques for econometric analysis of time series data and apply those techniques to the problems of forecasting and estimating dynamic causal effects. This chapter introduces the basic concepts and tools of regression using time series data and applies them to economic forecasting. Chapter 16 applies these tools to the estimation of dynamic causal effects. Chapter 17 takes up some more advanced topics in time series econometrics, including forecasting multiple time series, forecasting with many predictors, and modeling changes in volatility over time.

Economic forecasting is the prediction of future values of economic variables. Firms use economic forecasts when they plan production levels. Governments use revenue forecasts when they develop their budgets for the upcoming year. Economists at central banks, like the U.S. Federal Reserve System, forecast economic variables including the inflation rate and the growth of Gross Domestic Product (GDP) as part of setting monetary policy. Wall Street investors rely on forecasts of profits when deciding whether to invest in a company.

Forecasting is an application of the more general prediction problem in statistics, in which a given set of data is used to predict observations not in the data set. Forecasting refers to the prediction of *future* values of time series data. As with prediction more generally, forecasting models need not and generally do not have a causal interpretation.

Section 15.1 presents some examples of economic time series data and introduces basic concepts of time series analysis. Section 15.2 sets out the forecasting problem and introduces a measure of forecast accuracy, the mean squared forecast error. It also introduces the concept of stationarity, which implies that historical relationships among variables hold in the future, so that past data can reliably be used to make forecasts. Section 15.3 introduces autoregressions, time series regression models in

which the regressors are past values of the dependent variable, and Section 15.4 explains how to include additional regressors. For example, we find that including the term spread (the difference between long- and short-term interest rates) improves forecasts of the growth of U.S. GDP relative to using only lagged values of GDP growth. Section 15.5 discusses how to estimate the mean squared forecast error and how to compute forecast intervals—that is, ranges that are likely to contain the actual value of the variable being forecasted. Section 15.6 describes methods for choosing the number of lags in forecasting models. Sections 15.7 and 15.8 take up two common departures from the assumption of stationarity, trends and breaks, and show how to modify forecasting regressions if they are present.

15.1   Introduction to Time Series Data and Serial Correlation

A good place to start any empirical analysis is plotting the data, so that is where we begin.

Real GDP in the United States

Gross Domestic Product (GDP) measures the value of goods and services produced in an economy over a given time period. Figure 15.1a plots the value of "real" GDP per year in the United States from 1960 through 2017, where "real" indicates that the values have been adjusted for inflation. The values of GDP are expressed in $2009, which means that the price level is held fixed at its 2009 value. Because U.S. GDP grows at approximately an exponential rate, Figure 15.1a plots GDP on a logarithmic scale. GDP increased dramatically over a recent 58-year period, from approximately $3 trillion in 1960 to over $17 trillion in 2017. Measured on a logarithmic scale, this greater-than-five-fold increase corresponds to an increase of 1.7 log points. The rate of growth was not constant, however, and the figure shows declines in GDP during the recessions of 1960–1961, 1970, 1974–1975, 1980, 1981–1982, 1990–1991, 2001, and 2007–2009, episodes denoted by shading in Figure 15.1.

Lags, First Differences, Logarithms, and Growth Rates

The observation on the time series variable Y made at date t is denoted Y_t, and the total number of observations is denoted T. The interval between observations—that is, the period of time between observation t and observation $t + 1$—is some unit of time such as weeks, months, quarters (three-month units), or years. A set of T observations on a time series variable Y is denoted $Y_1, \ldots, Y_T$, or $\{Y_t\}, t = 1, \ldots, T$. This notation parallels the notation for cross-sectional data, in which the observations are denoted by $i = 1, \ldots, n$. In a given data set, the date $t = 1$ corresponds to the first date in the data set, and $t = T$ corresponds to the final date in the data set. For example, the GDP data studied in this chapter are quarterly, so the unit of time

GDP increased from $3 trillion per year in 1960 to over $17 trillion per year in 2017 when measured in inflation-adjusted 2009 dollars. This greater-than-five-fold increase corresponds to an increase of 1.7 log points. The growth rate of GDP was not constant, and it varied considerably from quarter to quarter.

Logarithm
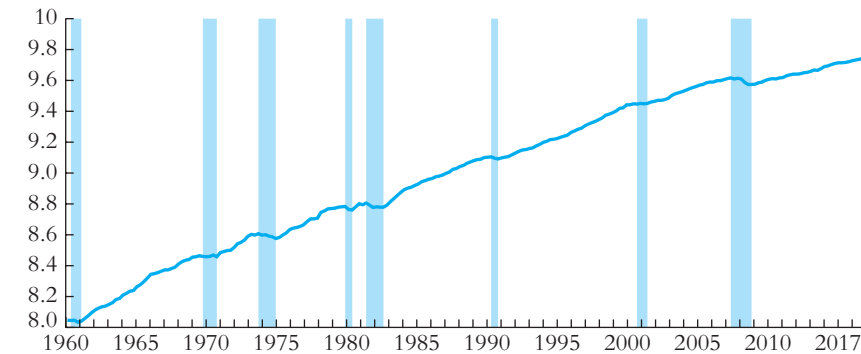
(a) U.S. GDP ($2009, billions)

Percentage at an Annual Rate
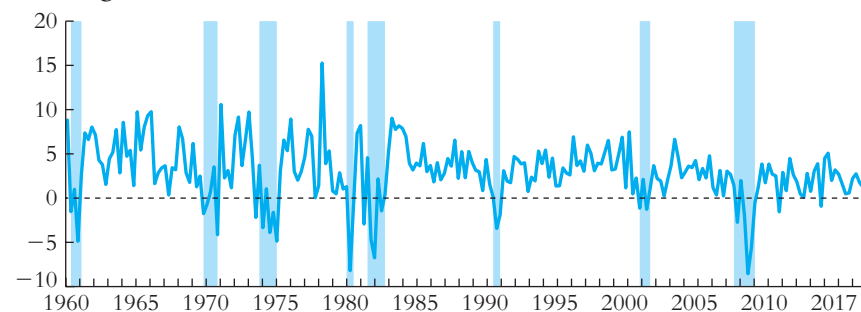
(b) Growth rate in U.S. GDP

(a period) is a quarter of a year. The data plotted in Figure 15.1b are quarterly growth rates of GDP from the first quarter of 1960, or 1960:Q1, through the fourth quarter of 2017, or 2017:Q4, for a total of $T = 232$ observations.

The change in the value of Y between period $t - 1$ and period t is $Y_t - Y_{t-1}$; this change is called the **first difference** in the variable Y_t. In time series data, "Δ" is used to represent the first difference, so $\Delta Y_t = Y_t - Y_{t-1}$.

Special terminology and notation are used to indicate future and past values of Y. The value of Y in the previous period (relative to the current period, t) is called its *first lagged value* (or, more simply, its **first lag**) and is denoted Y_{t-1}. Its jth *lagged value* (or, more simply, its *j^{th} lag*) is its value j periods ago, which is Y_{t-j}. Similarly, Y_{t+1} denotes the value of Y one period into the future.

Economic time series are often analyzed after computing their logarithms or the changes in their logarithms. One reason for this is that many economic series exhibit growth that is approximately exponential; that is, over the long run, the series tends to grow by a certain percentage per year on average. This implies that the logarithm of the series grows approximately linearly and is why Figure 15.1a plots the logarithm

Lags, First Differences, Logarithms, and Growth Rates

- The first lag of a time series Y_t is Y_{t-1}; its jth lag is Y_{t-j}.
- The first difference of a series, ΔY_t, is its change between periods $t-1$ and t; that is, $\Delta Y_t = Y_t - Y_{t-1}$.
- The first difference of the logarithm of Y_t is $\Delta\ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$.
- The percentage change of a time series Y_t between periods $t-1$ and t is approximately $100\Delta\ln(Y_t)$, where the approximation is most accurate when the percentage change is small.

of U.S. GDP. Another reason is that the standard deviation of many economic time series is approximately proportional to its level; that is, the standard deviation is well expressed as a percentage of the level of the series. This implies that the standard deviation of the logarithm of the series is approximately constant. In either case, it is useful to transform the series so that changes in the transformed series are proportional (or percentage) changes in the original series, and this is achieved by taking the logarithm of the series.[1]

Lags, first differences, and growth rates are summarized in Key Concept 15.1.

Lags, changes, and percentage changes are illustrated using the U.S. GDP data in Table 15.1. The first column shows the date, or period, where the fourth quarter of 2016 is denoted 2016:Q4, the first quarter of 2017 is denoted 2017:Q1, and so forth. The second column shows the value of GDP in that quarter, the third column shows the logarithm of GDP, and the fourth column shows the growth rate of GDP (in percent at an annual rate). For example, from the fourth quarter of 2016 to the first quarter of 2017, GDP increased from \$16,851 to \$16,903 billion, which is a percentage increase of $100 \times (16{,}903 - 16{,}851)/16{,}851 = 0.31\%$. This is the percentage increase from one quarter to the next. It is conventional to report rates of growth in quarterly macroeconomic time series on an annual basis, which is the percentage increase in GDP that would occur over a year if the series were to continue to increase at the same rate. Because there are four quarters in a year, the annualized rate of GDP growth in 2017:Q1 is $0.31 \times 4 = 1.24$, or 1.24%.

---

[1]The change of the logarithm of a variable is approximately equal to the proportional change of that variable; that is, $\ln(X + a) - \ln(X) \cong a/X$, where the approximation works best when a/X is small [see Equation (8.16) and the surrounding discussion]. Now, replace X with Y_{t-1} and a with ΔY_t, and note that $Y_t = Y_{t-1} + \Delta Y_t$. This means that the proportional change in the series Y_t between periods $t-1$ and t is approximately $\ln(Y_t) - \ln(Y_{t-1}) = \ln(Y_{t-1} + \Delta Y_t) - \ln(Y_{t-1}) \cong \Delta Y_t/Y_{t-1}$ (see Equation 18.16). The expression $\ln(Y_t) - \ln(Y_{t-1})$ is the first difference of $\ln(Y_t)$—that is, $\Delta\ln(Y_t)$. Thus $\Delta\ln(Y_t) \cong \Delta Y_t/Y_{t-1}$. The percentage change is 100 times the fractional change, so the percentage change in the series Y_t is approximately $100\Delta\ln(Y_t)$.

	TABLE 15.1 GDP in the United States in the Last Quarter of 2016 and in 2017				
Quarter	**U.S. GDP (billions of $2009), GDP_t**	**Logarithm of GDP, $\ln(GDP_t)$**	**Growth Rate of GDP at an Annual Rate, $GDPGR_t = 400 \times \Delta\ln(GDP_t)$**	**First Lag, $GDPGR_{t-1}$**
2016:Q4	16,851	9.732	1.74	2.74
2017:Q1	16,903	9.735	1.23	1.74
2017:Q2	17,031	9.743	3.01	1.23
2017:Q3	17,164	9.751	3.11	3.01
2017:Q4	17,272	9.757	2.50	3.11

Note: The quarterly rate of GDP growth is the first difference of the logarithm. This is converted into percentages at an annual rate by multiplying by 400. The first lag is its value in the previous quarter. All entries are rounded to the nearest decimal.

In the table, this percentage change is computed using the differences-of-logarithms approximation in Key Concept 15.1. The difference in the logarithm of GDP from 2016:Q4 to 2017:Q1 is $\ln(16{,}903) - \ln(16{,}851) = 0.00308$, yielding the approximate quarterly percentage difference $100 \times 0.00308 = 0.308\%$. On an annualized basis, this is $0.308 \times 4 = 1.23$, or 1.23%, essentially the same as the change obtained by directly computing the percentage growth. These calculations can be summarized as

$$\text{Annualized rate of GDP growth} = GDPGR_t \cong 400\left[\ln(GDP_t) - \ln(GDP_{t-1})\right]$$

$$= 400\Delta\ln(GDP_t), \tag{15.1}$$

where GDP_t is the value of GDP at date t. The factor of 400 arises from converting the decimal change to a percentage (multiplying by 100) and then converting the quarterly percentage change to an equivalent annual rate (multiplying by 4).

The final column of Table 15.1 illustrates lags. The first lag of $GDPGR$ in 2017:Q1 is 1.74%, the value of $GDPGR$ in 2016:Q4.

Figure 15.1b plots $GDPGR_t$ from 1960:Q1 through 2017:Q4. It shows substantial variability in the growth rate of GDP. For example, GDP grew at an annual rate of over 15% in 1978:Q2 and fell at an annual rate of over 8% in 2008:Q4. Over the entire period, the growth rate averaged 3.0% (which is responsible for the increase of GDP from $3.1 trillion in 1960 to $17.3 trillion in 2017), and the sample standard deviation was 3.3%.

Autocorrelation

In time series data, the value of Y in one period typically is correlated with its value in the next period. The correlation of a series with its own lagged values is called **autocorrelation** or **serial correlation**. The first autocorrelation (or **autocorrelation coefficient**) is the correlation between Y_t and Y_{t-1}—that is, the correlation between

Autocorrelation (Serial Correlation) and Autocovariance

The jth autocovariance of a series Y_t is the covariance between Y_t and its jth lag, Y_{t-j}, and the jth autocorrelation coefficient is the correlation between Y_t and Y_{t-j}. That is,

$$j^{th} \text{ autocovariance} = \text{cov}(Y_t, Y_{t-j}) \qquad (15.2)$$

$$j^{th} \text{ autocorrelation} = \rho_j = \text{corr}(Y_t, Y_{t-j}) = \frac{\text{cov}(Y_t, Y_{t-j})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t-j})}}. \qquad (15.3)$$

The jth autocorrelation coefficient is sometimes called the jth serial correlation coefficient.

values of Y at two adjacent dates. The second autocorrelation is the correlation between Y_t and Y_{t-2}, and the jth autocorrelation is the correlation between Y_t and Y_{t-j}. Similarly, the **j^{th} autocovariance** is the covariance between Y_t and Y_{t-j}. Autocorrelation and autocovariance are summarized in Key Concept 15.2.

The jth population autocovariances and autocorrelations in Key Concept 15.2 can be estimated by the jth sample autocovariances and autocorrelations, $\widehat{\text{cov}(Y_t, Y_{t-j})}$ and $\hat{\rho}_j$:

$$\widehat{\text{cov}(Y_t, Y_{t-j})} = \frac{1}{T}\sum_{t=j+1}^{T}(Y_t - \overline{Y}_{j+1:T})(Y_{t-j} - \overline{Y}_{1:T-j}) \qquad (15.4)$$

$$\hat{\rho}_j = \frac{\widehat{\text{cov}(Y_t, Y_{t-j})}}{\widehat{\text{var}(Y_t)}}, \qquad (15.5)$$

where $\overline{Y}_{j+1:T}$ denotes the sample average of Y_t computed using the observations $t = j + 1, \ldots, T$ and where $\widehat{\text{var}(Y_t)}$ is the sample variance of Y.[2]

The first four sample autocorrelations of *GDPGR*, the growth rate of GDP, are $\hat{\rho}_1 = 0.33$, $\hat{\rho}_2 = 0.26$, $\hat{\rho}_3 = 0.10$, and $\hat{\rho}_4 = 0.11$. These values suggest that GDP growth rates are mildly positively autocorrelated: If GDP grows faster than average in one period, it tends to also grow faster than average in the following period.

---

[2]The summation in Equation (15.4) is divided by T, whereas in the usual formula for the sample covariance [see Equation (3.24)], the summation is divided by the number of observations in the summation minus a degrees-of-freedom adjustment. The formula in Equation (15.4) is conventional for the purpose of computing the autocovariance. Equation (15.5) uses the assumption that $\text{var}(Y_t)$ and $\text{var}(Y_{t-j})$ are the same—an implication of the assumption that Y is stationary, a concept introduced in Section 15.3.

FIGURE 15.2 Four Economic Time Series

(a) U.S. unemployment rate

(b) U.S. dollar/British pound exchange rate

(c) Logarithm of Japan Index of Industrial Production

(d) Percentage change in daily value of the Wilshire 5000 Total Market Index
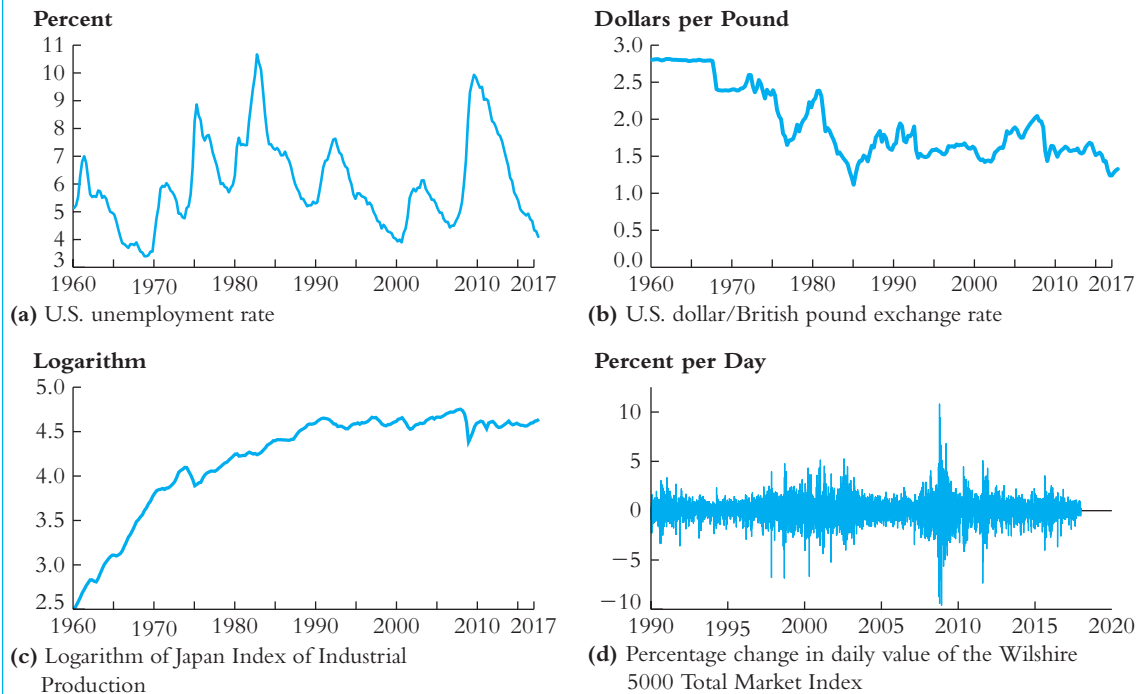
The four time series have markedly different patterns. The unemployment rate (Figure 15.2a) increases during recessions and declines during expansions. The exchange rate between the U.S. dollar and the British pound (Figure 15.2b) shows a discrete change after the 1972 collapse of the Bretton Woods system of fixed exchange rates. The logarithm of the Japan Index of Industrial Production (Figure 15.2c) shows decreasing growth. The daily percentage changes in the Wilshire 5000 Total Market Index, a stock price index (Figure 15.2d), are essentially unpredictable, but the variance changes: This series shows *volatility clustering*.

Other Examples of Economic Time Series

Economic time series differ greatly. Four examples of economic time series are plotted in Figure 15.2: the U.S. unemployment rate; the rate of exchange between the U.S. dollar and the British pound; the logarithm of the Japan Index of Industrial Production; and the percentage change in daily values of the Wilshire 5000 Total Market Index, a stock price index.

The U.S. unemployment rate (Figure 15.2a) is the fraction of the labor force out of work, as measured in the Current Population Survey (see Appendix 3.1). Figure 15.2a shows that the unemployment rate increases by large amounts during recessions (the shaded areas in Figure 15.1) and falls during expansions.

The dollar/pound exchange rate (Figure 15.2b) is the price of a British pound (£) in U.S. dollars. Before 1972, the developed economies ran a system of fixed exchange rates—called the Bretton Woods system—under which governments kept exchange

rates from fluctuating. In 1972, inflationary pressures led to the breakdown of this system; thereafter, the major currencies were allowed to "float"; that is, their values were determined by the supply and demand for currencies in the market for foreign exchange. Prior to 1972, the exchange rate was approximately constant, with the exception of a single devaluation in 1968, in which the official value of the pound relative to the dollar was decreased to $2.40. Since 1972, the exchange rate has fluctuated over a very wide range.

The Japan Index of Industrial Production (Figure 15.2c) measures Japan's output of industrial commodities. The logarithm of the series is plotted in Figure 15.2c, and changes in this series can be interpreted as (fractional) growth rates. During the 1960s and early 1970s, Japanese industrial production grew quickly, but this growth slowed in the late 1970s and 1980s, and industrial production has grown little since the early 1990s.

The Wilshire 5000 Total Market Index is an index of the share prices of all firms traded on exchanges in the United States. Figure 15.2d plots the daily percentage change in this index for trading days from January 2, 1990, to December 29, 2017 (a total of 7305 observations). Unlike the other series in Figure 15.2, there is very little serial correlation in these daily percentage changes; if there were, then you could predict them using past daily changes and make money by buying when you expect the market to rise and selling when you expect it to fall. Although the changes are essentially unpredictable, inspection of Figure 15.2d reveals patterns in their volatility. For example, the standard deviation of daily percentage changes was relatively large in 1998–2003 and 2007–2012, and it was relatively small in 1994, 2004, and 2017. This *volatility clustering* is found in many financial time series, and econometric models for modeling this special type of heteroskedasticity are taken up in Section 17.5.

15.2 Stationarity and the Mean Squared Forecast Error

Stationarity

Time series forecasts use data on the past to forecast the future. Doing so presumes that the future is similar to the past in the sense that the correlations, and more generally the distributions, of the data in the future will be like they were in the past. If the future differs fundamentally from the past, then historical relationships might not be reliable guides to the future.

In the context of regression with time series data, the idea that historical relationships can be generalized to the future is formalized by the concept of **stationarity**. The precise definition of stationarity, given in Key Concept 15.3, is that the probability distribution of the time series variable does not change over time. Under the assumption of stationarity, regression models estimated using past data can be used to forecast future values.

Stationarity

A time series Y_t is *stationary* if its probability distribution does not change over time—that is, if the joint distribution of $(Y_{s+1}, Y_{s+2}, \ldots, Y_{s+T})$ does not depend on s, regardless of the value of T; otherwise, Y_t is said to be *nonstationary*. A pair of time series, X_t and Y_t, are said to be *jointly stationary* if the joint distribution of $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2}, \ldots, X_{s+T}, Y_{s+T})$ does not depend on s, regardless of the value of T. Stationarity requires the future to be like the past, at least in a probabilistic sense.

Stationarity can fail to hold for multiple reasons, in which case the time series is said to be **nonstationary**. One reason is that the unconditional mean might have a trend. For example, the logarithm of U.S. GDP plotted in Figure 15.1a has a persistent upward trend, reflecting long-term economic growth. Another type of nonstationarity arises when the population regression coefficients change at a given point in time. Ways to detect and to address these two types of nonstationarity are taken up in Sections 15.6 and 15.7. Until then, we assume that the time series is stationary.

Forecasts and Forecast Errors

This chapter considers the problem of forecasting the value of a time series variable Y in the period immediately following the end of the available data—that is, of forecasting Y_{T+1} using data through date T. This forecast answers questions such as, Given data through the current quarter, what is my forecast of GDP growth for the next quarter? Because the forecast is for the next time period, this forecast is called a **one-step ahead forecast**. A more ambitious question is, Given data through the current quarter, what is my forecast of GDP growth for *each* of the next eight quarters? Answering that question entails making a forecast over a longer horizon, called a **multi-step ahead forecast**. Multi-step ahead forecasts are taken up in Chapter 17.

We let $\hat{Y}_{T+1|T}$ denote a candidate one-step ahead forecast of Y_{T+1}. In this notation, the subscript $T + 1|T$ indicates that the forecast is of the value of Y at time $T + 1$, made using data through time T, and the caret (ˆ) indicates that the forecast is based on an estimated model. For example, suppose you have quarterly observations on GDP growth (Y) from 1960:Q1 to 2017:Q3. The one-step ahead forecasting problem is to use these data to forecast GDP growth in 2017:Q4, and the forecast is denoted $\hat{Y}_{2017:Q4|2017:Q3}$.

Because the future is unknown, errors in forecasting are inevitable. The **forecast error** is the difference between the actual value of Y_{T+1} and its forecast:

$$\text{Forecast error} = Y_{T+1} - \hat{Y}_{T+1|T}. \tag{15.6}$$

A forecast refers to a prediction made for a future date that is not in the data set used to make the forecast—that is, the forecast is for an out-of-sample future observation. The forecast error is the mistake made by the forecast, which is realized only after time has elapsed and the actual value of Y_{T+1} is observed.

The Mean Squared Forecast Error

Because forecast errors are inevitable, the aim of the forecaster is not to eliminate errors but rather to make them as small as possible—that is, to make the forecasts as accurate as possible. To make this goal precise, we need a quantitative measure of what it means for a forecast error to be small. The most commonly used measure, which we adopt in this text, is the **mean squared forecast error (MSFE)**, which is the expected value of the square of the forecast error:

$$\text{MSFE} = E[(Y_{T+1} - \hat{Y}_{T+1|T})^2]. \tag{15.7}$$

The MSFE is the time series counterpart of the mean squared prediction error introduced in Section 14.2 for out-of-sample prediction with cross-sectional data.

In practice, large forecast errors can be much more costly than small ones. A series of small forecast errors often causes only minor problems for the user, but a single very large forecast error can call the entire forecasting activity into question. The MSFE captures this idea by using the square of the forecast error, so that large errors receive a much greater penalty than small ones.

The **root mean squared forecast error (RMSFE)** is the square root of the MSFE. The RMSFE is easily interpreted because it has the same units as Y. If the forecast is unbiased, forecast errors have mean zero and the RMSFE is the standard deviation of the out-of-sample forecast made using a given model.

The MSFE incorporates two sources of randomness. The first is the randomness of the future value, Y_{T+1}. The second is the randomness arising from estimating a forecasting model. For example, suppose a forecaster uses a very simple model, in which the value of Y_{T+1} is forecasted to be its historical mean value, μ_Y. (This simple model is a plausible starting point for forecasting stock returns, as discussed in the box "Can You Beat the Market?" later in this section.) Because the mean is unknown, it must be estimated—say, by $\hat{\mu}_Y$. In this example, the forecast is $\hat{Y}_{T+1|T} = \hat{\mu}_Y$, the forecast error is $Y_{T+1} - \hat{Y}_{T+1|T} = Y_{T+1} - \hat{\mu}_Y$, and the MSFE is $\text{MSFE} = E[(Y_{T+1} - \hat{\mu}_Y)^2]$. By adding and subtracting μ_Y, if Y_{T+1} is uncorrelated with $\hat{\mu}_Y$, the MSFE can be written as $\text{MSFE} = E[(Y_{T+1} - \mu_Y)^2] + E[(\hat{\mu}_Y - \mu_Y)^2]$. The first term in this expression is the error the forecaster would make if the population mean were known: This term captures the random future (out-of-sample) fluctuations in Y_{T+1} around the population mean. The second term in this expression is the additional error made because the population mean is unknown, so the forecaster must estimate it.

From the perspective of the MSFE, the best-possible prediction is the conditional mean given the in-sample observations on Y—that is, $E(Y_{T+1}|Y_1, \ldots, Y_T)$

Can You Beat the Market?

Have you ever dreamed of getting rich quickly by beating the stock market? If you think that the market will be going up, you should buy stocks today and sell them later, before the market turns down. If you are good at forecasting swings in stock prices, then this active trading strategy will produce better returns than a passive "buy and hold" strategy, in which you purchase stocks and just hang onto them. The trick, of course, is having a reliable forecast of future stock returns.

Forecasts based on past values of stock returns are sometimes called momentum forecasts: If the value of a stock rose this month, perhaps it has momentum and will also rise next month. If so, then returns will be autocorrelated, and the autoregressive model will provide useful forecasts. You can implement a momentum-based strategy for a specific stock or for a stock index that measures the overall value of the market.

Table 15.2 presents autoregressive models of the excess return on a broad-based index of stock prices, called the CRSP value-weighted index, using monthly data from 1960:M1 to 2002:M12, where M1 denotes the first month of the year (January), M2 denotes the second month, and so forth. The monthly excess return is what you earn, in percentage terms, by purchasing a stock at the end of the previous month and selling it at the end of this month minus

TABLE 15.2	**Autoregressive Models of Monthly Excess Stock Returns, 1960:M1–2002:M12**		

Dependent variable: excess returns on the CRSP value-weighted index

	(1)	(2)	(3)
Specification	AR(1)	AR(2)	AR(4)
Regressors			
excess return$_{t-1}$	0.050 (0.051)	0.053 (0.051)	0.054 (0.051)
excess return$_{t-2}$		−0.053 (0.048)	−0.054 (0.048)
excess return$_{t-3}$			0.009 (0.050)
excess return$_{t-4}$			−0.016 (0.047)
Intercept	0.312 (0.197)	0.328 (0.199)	0.331 (0.202)
F-statistic for coefficients on lags of excess return (p-value)	0.968 (0.325)	1.342 (0.261)	0.707 (0.587)
$\overline{R}^2$	0.0006	0.0014	−0.0022

Note: Excess returns are measured in percentage points per month. The data are described in Appendix 15.1. All regressions are estimated over 1960:M1–2002:M12 ($T = 516$ observations), with earlier observations used for initial values of lagged variables. Entries in the regressor rows are coefficients, with standard errors in parentheses. The final two rows report the F-statistic testing the hypothesis that the coefficients on lags of *excess return* in the regression are 0, with its p-value in parentheses, and the adjusted R^2, or $\overline{R}^2$.

what you would have earned had you purchased a safe asset (a U.S. Treasury bill). The return on the stock includes the capital gain (or loss) from the change in price plus any dividends you receive during the month. The data are described further in Appendix 15.1.

Sadly, the results in Table 15.2 are negative. The coefficient on lagged returns in the AR(1) model is not statistically significant, and we cannot reject the null hypothesis that the coefficients on lagged returns are all 0 in the AR(2) or AR(4) model. In fact, the adjusted R^2, or $\overline{R}^2$, of one of the models is negative, and those of the other two are only slightly positive, suggesting that none of these models is useful for forecasting.

These negative results are consistent with the theory of efficient capital markets, which holds that excess returns should be unpredictable because stock prices already embody all currently available information. The reasoning is simple: If market participants think that a stock will have a positive excess return next month, then they will buy that stock now, but doing so will drive up the price of the stock to exactly the point at which there is no expected excess return. As a result, we should not be able to forecast future excess returns by using past publicly available information, and we cannot do it, at least using the regressions in Table 15.2.

(Appendix 2.2). This best-possible forecast, $E(Y_{T+1}|Y_1, \ldots, Y_T)$, is called the **oracle forecast**. The oracle forecast is infeasible because the conditional mean is unknown in practice. Because it minimizes the MSFE, the oracle forecast is a conceptual benchmark against which to assess an actual forecast.

The MSFE is an unknown population expectation, so to use it in practice it must be estimated using data. We discuss estimation of the RMSFE in Section 15.4.

15.3   Autoregressions

If you want to predict the future, a good place to start is the immediate past. For example, if you want to forecast the rate of GDP growth in the next quarter, you might use data on how fast GDP grew in the current quarter or perhaps over the past several quarters as well. To do so, a forecaster would fit an autoregression.

The First-Order Autoregressive Model

An **autoregression** expresses the conditional mean of a time series variable Y_t as a linear function of its own lagged values. A **first-order autoregression** uses only one lag of Y in this conditional expectation. That is, in a first-order autoregression, $E(Y_t|Y_{t-1}, Y_{t-2}, \ldots) = \beta_0 + \beta_1 Y_{t-1}$. The first-order autoregression [AR(1)] model can be written in the familiar form of a regression model as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t, \tag{15.8}$$

where u_t is the error term. The first-order autoregression in Equation (15.8) is a population autoregression with two unknown coefficients, β_0 and β_1.

The unknown population coefficients β_0 and β_1 in Equation (15.8) can be estimated by ordinary least square (OLS). How to estimate β_0 and β_1 might initially seem puzzling: Unlike a cross-sectional regression with X on the right-hand side, Equation (15.8) has Y on both the right- *and* the left-hand sides! The solution to this puzzle is to realize that the variable Y_{t-1} on the right-hand side differs from the dependent variable Y_t because the regressor is the first lag of Y. That is, Equation (15.8) has the form of a standard regression model, with X being the first lag of Y. Thus, to estimate β_0 and β_1, you must create a new variable—the first lag of Y—and then use that as the regressor. Doing so yields the OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$.

To make this concrete, consider estimating a first-order autoregression for GDP growth. Observations on the dependent variable, $Y_t = GDPGR_t$, are given in the fourth column of Table 15.1 for 2016:Q4-2017:Q4. Data on the regressor, $Y_{t-1} = GDPGR_{t-1}$ for those dates are given in the final column of Table 15.1. Thus the OLS estimator is obtained by regressing the data in the fourth column of Table 15.1 (extended back to the start of the sample) against the data in the final column, including an intercept. To estimate this AR(1) model, we use data starting in 1962:Q1 and reserve the final observation, 2017:Q4, to illustrate computing the forecast and forecast error. The resulting first-order autoregression, estimated using data from 1962:Q1–2017:Q3, is

$$\widehat{GDPGR_t} = 1.950 + 0.341 \, GDPGR_{t-1}. \tag{15.9}$$
$$(0.322) \quad (0.073)$$

As usual, standard errors are given in parentheses under the estimated coefficients, and $\widehat{GDPGR}$ is the predicted value of $GDPGR$ based on the estimated regression line.

Forecasts and forecast errors. If the population coefficients in Equation (15.8) were known, then the one-step ahead forecast of Y_{T+1}, made using data through date T, would be $\beta_0 + \beta_1 Y_T$. Although β_0 and β_1 are unknown, the forecaster can use their OLS estimates instead. Accordingly, the forecast based on the AR(1) model in Equation (15.8) is

$$\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T, \tag{15.10}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated using historical data through time T. The forecast error is $Y_{T+1} - \hat{Y}_{T+1|T}$.

Application to GDP growth. What is the forecast of the growth rate of GDP in the fourth quarter of 2017 (2017:Q4) that a forecaster would have made in 2017:Q3, based on the estimated AR(1) model in Equation (15.9) (which was estimated using data through 2017:Q3)? According to Table 15.1, the growth rate of GDP in 2017:Q3 was 3.11% (so $GDPGR_{2017:Q3} = 3.11$). Plugging this value

into Equation (15.8), the forecast of the growth rate of GDP in 2017:Q4 is
$\widehat{GDPGR}_{2017:Q4|2017:Q3} = 1.950 + 0.341 \times GDPGR_{2017:Q3} = 1.950 + 0.341 \times 3.11 = 3.0$
(rounded to the nearest tenth). Thus, the AR(1) model forecasts that the growth rate
of GDP will be 3.0% in 2017:Q4. Because data for 2017:Q4 are available, we can
evaluate the forecast error for this forecast. Table 15.1 shows that the actual growth
rate of GDP in 2017:Q4 was 2.5%, so the AR(1) forecast is high by 0.5 percentage
points; that is, the forecast error is −0.5.[3]

The $\overline{R}^2$ of the AR(1) model in Equation (15.9) is only 0.11, so the lagged value
of GDP growth explains only a small fraction of the variation in GDP growth in the
sample used to fit the autoregression. It is therefore of interest to see whether includ-
ing additional variables, beyond the first lag, could improve the fit of the forecasting
model.

The pth-Order Autoregressive Model

The AR(1) model uses Y_{t-1} to forecast Y_t, but doing so ignores potentially useful
information in the more distant past. One way to incorporate this information is to
include additional lags in the AR(1) model; this yields the pth-order autoregressive
model.

The **p^{th}-order autoregressive [AR(p)] model** represents Y_t as a linear function of
p of its lagged values; that is, in the AR(p) model, the regressors are $Y_{t-1},\ Y_{t-2},\ \ldots,\ Y_{t-p}$,
plus an intercept. The number of lags, p, included in an AR(p) model is called the
order, or lag length, of the autoregression.

For example, an AR(2) model of GDP growth uses two lags of GDP growth as
regressors. Estimated by OLS over the period 1962:Q1–2017:Q3, the AR(2) model is

$$\widehat{GDPGR}_t = \underset{(0.37)}{1.60} + \underset{(0.08)}{0.28}\,GDPGR_{t-1} + \underset{(0.08)}{0.18}\,GDPGR_{t-2}. \qquad (15.11)$$

The coefficient on the additional lag in (Equation (15.11)) is significantly different
from 0 at the 5% significance level: The t-statistic is 2.30 (p-value $= 0.02$). This is
reflected in an improvement in the $\overline{R}^2$ from 0.11 for the AR(1) model in Equation (15.8)
to 0.14 for the AR(2) model.

The AR(p) model is summarized in Key Concept 15.4.

Properties of the forecast and error term in the AR(p) model.   The assumption that
the conditional expectation of u_t is 0 given past values of Y_t—that is,
$E(u_t\,|\,Y_{t-1}, Y_{t-2}, \ldots) = 0]$—has two important implications.

The first implication is that the best forecast of Y_{T+1} based on its entire
history depends on only the most recent p past values. Specifically, let
$Y_{T+1|T} = E(Y_{T+1}\,|\,Y_T, Y_{T-1}, \ldots)$ denote the conditional mean of Y_{T+1} given its

---

[3]The units of the arithmetic difference between two percentages is percentage points. For example, if an
interest rate is 3.5% at an annual rate and it rises to 3.8%, then it has risen by 0.3 percentage points.

Autoregressions

The pth-order autoregressive [AR(p)] model represents the conditional expectation of Y_t as a linear function of p of its lagged values:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + u_t, \qquad (15.12)$$

where $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$. The number of lags p is called the order, or the lag length, of the autoregression.

entire history. Then $Y_{T+1|T}$ is the oracle forecast and has the smallest MSFE of any forecast, based on the history of Y (Exercise 15.5). That is, if Y_t follows an AR(p), then the oracle forecast of Y_{T+1} based on $Y_T, Y_{T-1}, \ldots$ is

$$Y_{T+1|T} = \beta_0 + \beta_1 Y_T + \beta_2 Y_{T-1} + \cdots + \beta_p Y_{T-p+1}. \qquad (15.13)$$

In practice, the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are unknown, so actual forecasts from an AR(p) use Equation (15.13) with estimated coefficients.

The second implication is that the errors u_t are serially uncorrelated. This result follows from Equation (2.28) (Exercise 15.5).

Application to GDP growth. What is the forecast of the growth rate of GDP in 2017:Q4, using data through 2017:Q3, based on the AR(2) model of GDP growth in Equation (15.11)? To compute this forecast, substitute the values of GDP growth in 2017:Q2 and 2017:Q3 into Equation (15.11): $GDPGR_{2017:Q4|2017:Q3} = 160 + 0.28 \, GDPGR_{2017:Q3} + 0.18 \, GDPGR_{2017:Q2} = 1.60 + 0.28 \times 3.11 + 0.18 \times 3.01 \approx 3.0$, where the 2017:Q3 and 2017:Q2 values for $GDPGR$ are taken from the fourth column of Table 15.1. The forecast error is the actual value, 2.5%, minus the forecast, or $2.5\% - 3.0\% = -0.5$ percentage points, essentially the same as the AR(1) forecast error.

15.4 Time Series Regression with Additional Predictors and the Autoregressive Distributed Lag Model

Economic theory often suggests other variables that could help forecast a variable of interest. These other variables, or predictors, can be added to an autoregression to produce a time series regression model with multiple predictors. When other

Interest Rates and the Term Spread, 1960–2017

Long-term and short-term interest rates on bonds move together but not one-for-one. The difference between long-term rates and short-term rates is called the term spread. The term spread has fallen sharply before U.S. recessions, which are shown as shaded regions in the figures.

Percent per Annum

10-Year Interest Rate

3-Month Interest Rate
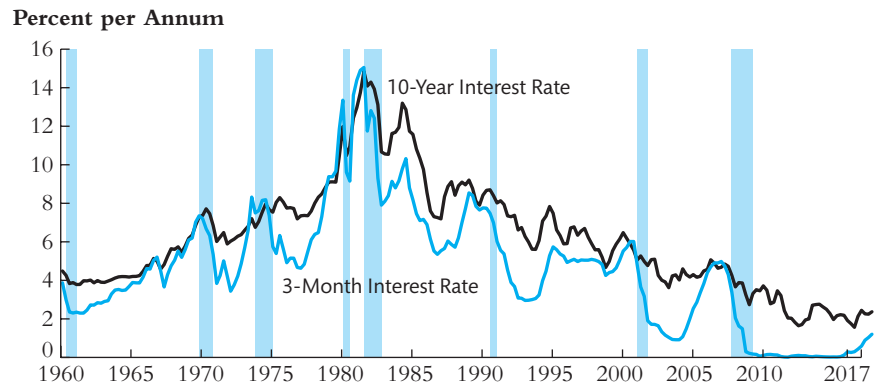
(a) 10–year interest rate and 3–month interest rate

Percentage points
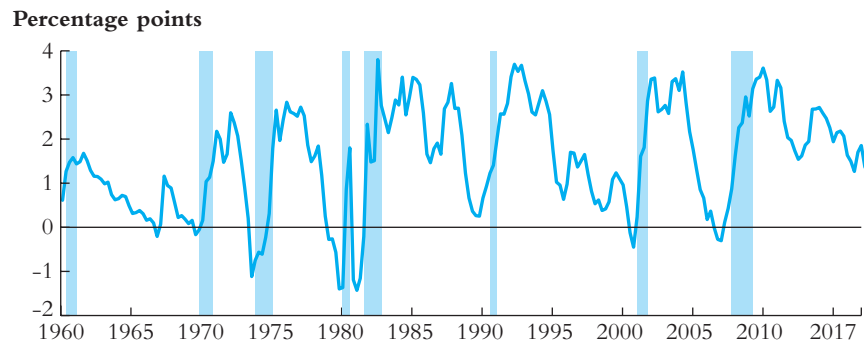
(b) Term spread

variables and their lags are added to an autoregression, the result is an autoregressive distributed lag model.

Forecasting GDP Growth Using the Term Spread

Interest rates on long-term and short-term bonds move together but not one for one. Figure 15.3a plots interest rates on 10-year U.S. Treasury bonds and 3-month Treasury bills from 1960 through 2017. These interest rates show the same long-run tendencies: Both were low in the 1960s, both rose through the 1970s and peaked in the early 1980s, and both fell subsequently. But the gap, or difference, between the two interest rates has not been constant: While short-term rates are generally below long-term rates, the gap between them narrows and even disappears shortly before the start of a recession; recessions are shown as the shaded bars in the figure. This difference between long-term and short-term interest rates is called the **term spread** and is plotted in Figure 15.3b. The term spread is generally positive, but it falls toward or below 0 before recessions.

Figure 15.3 suggests that the term spread might contain information about the future GDP growth that is not already contained in past values of GDP growth. This conjecture is readily checked by augmenting the AR(2) model in Equation (15.11) to include the first lag of the term spread:

$$\widehat{GDPGR}_t = 0.94 + 0.27\,GDPGR_{t-1} + 0.19\,GDPGR_{t-2} + 0.42\,TSpread_{t-1}.$$
$$(0.47)\ \ (0.08) \qquad\qquad (0.08) \qquad\qquad\quad (0.18) \qquad (15.14)$$

The t-statistic on $TSpread_{t-1}$ is -2.34, so this coefficient is significant at the 1% level. The $\overline{R}^2$ of this regression is 0.16, an improvement over the AR(2) $\overline{R}^2$ of 0.14.

The forecast of the rate of GDP growth in 2017:Q4 is obtained by substituting the 2017:Q2 and 2017:Q3 values of GDP growth into Equation (15.14), along with the value of the term spread in 2017:Q3 (which is 1.21); the resulting forecast is $\widehat{GDPGR}_{2017:Q4|2017:Q3} = 2.9\%$, and the forecast error is -0.4%.

If one lag of the term spread is helpful for forecasting GDP growth, more lags might be even more helpful; adding an additional lag of the term spread yields

$$\widehat{GDPGR}_t = 0.94 + 0.25\,GDPGR_{t-1} + 0.18\,GDPGR_{t-2}$$
$$(0.46)\ \ (0.08) \qquad\qquad (0.08)$$
$$(15.15)$$
$$-\,0.13\,TSpread_{t-1} + 0.62\,TSpread_{t-2}.$$
$$(0.42) \qquad\qquad (0.43)$$

The t-statistic testing the significance of the second lag of the term spread is 1.46 (p-value $= 0.14$), so it falls just short of statistical significance at the 10% level. The $\overline{R}^2$ of the regression in Equation (15.15) is 0.16, essentially the same as that in Equation (15.14).

The forecasted rate of GDP growth in 2017:Q4 is computed by substituting the values of the variables into Equation (15.15). The term spread was 1.37 in 2017:Q2 and 1.21 in 2017:Q3. The forecasted value of the rate of GDP growth in 2017:Q4, based on Equation (15.15), is

$$\widehat{GDPGR}_{2017:Q4|2017:Q3} = 0.94 + 0.25 \times 3.11 + 0.18 \times 3.01$$
$$-\,0.13 \times 1.21 + 0.62 \times 1.37 \approx 2.9. \qquad (15.16)$$

The forecast error is -0.4 percentage points.

The Autoregressive Distributed Lag Model

Each model in Equations (15.14) and (15.15) is an **autoregressive distributed lag (ADL) model**: *autoregressive* because lagged values of the dependent variable are included as regressors, as in an autoregression, and *distributed lag* because the regression also includes multiple lags (a "distributed lag") of an additional predictor. In general, an ADL model with p lags of the dependent variable Y_t and q lags of an

The Autoregressive Distributed Lag Model

The autoregressive distributed lag model with p lags of Y_t and q lags of X_t, denoted ADL(p, q), is

$$
\begin{aligned}
Y_t = {} & \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \\
& + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \cdots + \delta_q X_{t-q} + u_t,
\end{aligned}
\tag{15.17}
$$

where $\beta_0, \beta_1, \ldots, \beta_p, \delta_1, \ldots, \delta_q$, are unknown coefficients and u_t is the error term with $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots) = 0$.

additional predictor X_t is called an **ADL(p, q)** model. In this notation, the model in Equation (15.14) is an ADL(2, 1) model, and the model in Equation (15.15) is an ADL(2, 2) model.

The ADL model is summarized in Key Concept 15.5. The notation in Equation (15.17) is somewhat cumbersome, and alternative optional notation, based on the so-called lag operator, is presented in Appendix 15.3.

The assumption that the errors in the ADL model have a conditional mean of 0 given all past values of Y and X—that is, that $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots) = 0$ — implies that no additional lags of either Y or X belong in the ADL model. In other words, the lag lengths p and q are the true lag lengths, and the coefficients on additional lags are 0.

The Least Squares Assumptions for Forecasting with Multiple Predictors

The general time series regression model with multiple predictors extends the ADL model to include multiple predictors and their lags. The model is summarized in Key Concept 15.6. The presence of multiple predictors and their lags leads to double subscripting of the regression coefficients and regressors.

The assumptions in Key Concept 15.6 are the time series counterparts of the four least squares assumptions for prediction with multiple regression using cross-sectional data (Appendix 6.4).

The first assumption is that u_t has conditional mean 0 given the history of all the regressors. This assumption extends the assumption used in the AR and ADL models and implies that the oracle forecast of Y_t using all past values of Y and the X's is given by the regression in Equation (15.18).

The second least squares assumption for cross-sectional data is that $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, are independently and identically distributed (i.i.d.). The second assumption for time series regression replaces the i.i.d. assumption by a more appropriate one with two parts. Part (a) is that the data are drawn

The Least Squares Assumptions for Forecasting with Time Series Data

The general time series regression model allows for k additional predictors, where q_1 lags of the first predictor are included, q_2 lags of the second predictor are included, and so forth:

$$
\begin{aligned}
Y_t = \beta_0 &+ \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \\
&+ \delta_{11} X_{1t-1} + \delta_{12} X_{1t-2} + \cdots + \delta_{1q_1} X_{1t-q_1} \\
&+ \cdots + \delta_{k1} X_{kt-1} + \delta_{k2} X_{kt-2} + \cdots + \delta_{kq_k} X_{kt-q_k} + u_t,
\end{aligned} \tag{15.18}
$$

where

1. $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots, X_{1t-1}, X_{1t-2}, \ldots, X_{kt-1}, X_{kt-2}, \ldots) = 0$;

2. (a) The random variables $(Y_t, X_{1t}, \ldots, X_{kt})$ have a stationary distribution, and
   (b) $(Y_t, X_{1t}, \ldots, X_{kt})$ and $(Y_{t-j}, X_{1t-j}, \ldots, X_{kt-j})$ become independent as j gets large;

3. Large outliers are unlikely: $X_{1t}, \ldots, X_{kt}$ and Y_t have nonzero, finite fourth moments; and

4. There is no perfect multicollinearity.

from a stationary distribution, so that the distribution of the time series today is the same as its distribution in the past. This assumption is a time series version of the *identically distributed* part of the i.i.d. assumption: The cross-sectional requirement of each draw being identically distributed is replaced by the time series requirement that the joint distribution of the variables, *including lags*, not change over time. If the time series variables are nonstationary, then one or more problems can arise in time series regression, including biased forecasts.

The assumption of stationarity implies that the conditional mean for the data used to estimate the model is also the conditional mean for the out-of-sample observation of interest. Thus the assumption of stationarity is also an assumption about external validity, and it plays the role of the first least squares assumption for prediction in Appendix 6.4.

Part (b) of the second assumption requires that the random variables become independently distributed when the amount of time separating them becomes large. This replaces the cross-sectional requirement that the variables be independently distributed from one observation to the next with the time series requirement that they be independently distributed when they are separated by long periods of time. This assumption is sometimes referred to as **weak dependence**, and it ensures that in large samples there is sufficient randomness in the data for the law of large numbers and the central limit theorem to hold. For a precise mathematical statement of the weak dependence condition, see Hayashi (2000, Chapter 2).

The third assumption (no outliers) and fourth assumption (no perfect multicollinearity) are the same as for cross-sectional data.

Under the assumptions of Key Concept 15.6, inference on the regression coefficients using OLS proceeds in the same way as it usually does using cross-sectional data.

15.5 Estimation of the MSFE and Forecast Intervals

An estimate of the MSFE can be used to summarize forecast uncertainty and to construct forecast intervals.

Estimation of the MSFE

The MSFE, defined in Equation (15.7), is an expected value that depends on the distribution of Y and on the forecasting model. Because it is an expectation, its value is not known and must be estimated from the data.

A natural instinct would be to estimate the MSFE by replacing the expectation in Equation (15.7) with an average over out-of-sample observations. The out-of-sample data, however, are not observed, so this approach is not feasible. Instead, there are three commonly used methods, with increasing complexity, for estimation of the MSFE. All three methods necessarily rely on the in-sample data. The simplest estimator focuses only on future uncertainty and ignores uncertainty associated with estimation of the regression coefficients. The second estimator incorporates future uncertainty and estimation error, under the assumption of stationarity so that the conditional expectation estimated by the model applies to the out-of-sample forecast. The third incorporates uncertainty and estimation error and in addition allows for the possibility that the conditional expectation might change over the course of the sample.

The first two methods are based on an expression for the MSFE derived from Equation (15.7) and the assumption of stationarity. We provide this expression here for an AR(p); it extends directly to the models with additional predictors in Key Concept 15.6. Under the assumption of stationarity,

$$\text{MSFE} = \sigma_u^2 + \text{var}(\hat{\beta}_0 + \hat{\beta}_1 Y_T + \cdots + \hat{\beta}_p Y_{T-p+1}). \tag{15.19}$$

This result is shown for an AR(1) in Exercise 15.12.

The first term in Equation (15.19) is the variance of Y_{T+1} around its conditional mean. This is the variance of the oracle forecast. The second term in Equation (15.19) arises because the coefficients of the autoregression are unknown and must be estimated.

Method 1: Estimating the MSFE by the standard error of the regression. Because the variance of the OLS estimator is proportional to $1/T$, the second term in

Equation (15.19) is proportional to $1/T$. Consequently, if the number of observations T is large relative to the number of autoregressive lags p, then the contribution of the second term is small relative to the first term. That is, if T is large relative to p, Equation (15.19) simplifies to the approximation MSFE $\approx \sigma_u^2$. This simplification in turn suggests estimating the MSFE by

$$\widehat{\text{MSFE}}_{SER} = s_{\hat{u}}^2, \text{ where } s_{\hat{u}}^2 = \frac{SSR}{T - p - 1}, \tag{15.20}$$

where SSR is the sum of squared residuals of the autoregression. The statistic $s_{\hat{u}}^2$ is the square of the standard error of the regression (SER), originally defined in Equation (6.13) and restated in Equation (15.20) using the notation of autoregressions.

Method 2: Estimating the MSFE by the final prediction error. If T is not large relative to p, the sampling error of the estimated autoregression coefficients can be sufficiently large that the second term in Equation (15.19) should not be ignored. The **final prediction error (FPE)** is an estimate of the MSFE that incorporates both terms in Equation (15.19), under the additional assumption that the errors are homoskedastic. With homoskedastic errors, $\text{var}(\hat{\beta}_0 + \hat{\beta}_1 Y_T + \cdots + \hat{\beta}_p Y_{T-p+1}) \approx \sigma_u^2[(p+1)/T]$ (shown in Appendix 19.7); substitution of this expression into Equation (15.19) yields, MSFE $= \sigma_u^2 + \sigma_u^2 \frac{p+1}{T} = \sigma_u^2(1 + \frac{p+1}{T})$. The FPE uses this final expression, along with the estimator $s_{\hat{u}}^2$, to estimate the MSFE:

$$\widehat{MSFE}_{FPE} = \left(\frac{T + p + 1}{T}\right)s_{\hat{u}}^2 = \left(\frac{T + p + 1}{T - p - 1}\right)\frac{SSR}{T}. \tag{15.21}$$

The FPE estimator improves upon the squared SER in Equation (15.20) by adjusting for the sampling uncertainty in estimating the autoregression coefficients.

Method 3: Estimating the MSFE by pseudo out-of-sample forecasting. The third estimate of the MSFE uses the data to simulate out-of-sample forecasting. This method proceeds by first dividing the data set into two parts: an initial estimation sample (the first T-P observations) and a reserved sample (the final P observations). The initial estimation sample is used to estimate the forecasting model, which is then used to forecast the first observation in the reserved sample. Next the estimation sample is augmented by the first observation in the reserved sample, and the model is reestimated and is used to forecast the second observation in the reserved sample. This procedure is repeated until the forecast is made for the final observation in the reserved sample and produces P forecasts and thus P forecast errors. Those P forecast errors can then be used to estimate the MSFE.[4]

---

[4]Readers of Chapter 14 will recognize that this method for estimating the MSFE is related to estimation of the mean squared prediction error by cross validation.

Pseudo Out-of-Sample Forecasts

Pseudo out-of-sample forecasts are computed using the following steps:

1. Choose a number of observations, P, for which you will generate pseudo out-of-sample forecasts; for example, P might be 10% or 20% of the sample size. Let $s = T - P$.

2. Estimate the forecasting regression using the estimation sample—that is, using observations $t = 1, \ldots, s$.

3. Compute the forecast for the first period beyond this shortened sample, $s + 1$; call this $\tilde{Y}_{s+1|s}$.

4. Compute the forecast error, $\tilde{u}_{s+1} = Y_{s+1} - \tilde{Y}_{s+1|s}$.

5. Repeat steps 2 through 4 for the remaining periods, $s = T - P + 1$ to $T - 1$ (reestimate the regression for each period). The pseudo out-of-sample forecasts are $\tilde{Y}_{s+1|s}$, $s = T - P, \ldots, T - 1$, and the pseudo out-of-sample forecast errors are $\tilde{u}_{s+1}$, $s = T - P, \ldots, T - 1$.

This method of estimating a model on a subsample of the data and then using that model to forecast on a reserved sample is called **pseudo out-of-sample forecasting**: *out-of-sample* because the observations being forecasted were not used for model estimation but *pseudo* because the reserved data are not truly out-of-sample observations. The construction of pseudo out-of-sample forecasts is summarized in Key Concept 15.7.

With the resulting pseudo out-of-sample forecast errors $\tilde{u}_s$, $s = T - P + 1, \ldots, T$ in hand, the pseudo out-of-sample estimate of the MSFE is

$$\widehat{MSFE}_{POOS} = \frac{1}{P} \sum_{s=T-P+1}^{T} \tilde{u}_s^2. \tag{15.22}$$

Compared to the squared SER estimate in Equation (15.20) and the final prediction error estimate in Equation (15.21), the pseudo out-of-sample estimate in Equation (15.22) has both advantages and disadvantages. The main advantage is that it does not rely on the assumption of stationarity, so that the conditional mean might differ between the estimation and the reserved samples. For example, the coefficients of the autoregression need not be the same in the two samples, and indeed the pseudo out-of-sample forecast error need not have mean 0. Thus any bias in the forecast arising because of a change in coefficients will be captured by $\widehat{MSFE}_{POOS}$ but not by the other two estimators [which rely on Equation (15.19), which was derived under the assumption of stationarity]. Three disadvantages of the pseudo out-of-sample estimate are that it is more difficult to compute, that the estimate of the MSFE will have greater sampling variability than the other two estimates if Y is, in fact,

stationary (because $\widehat{MSFE}_{POOS}$ uses only P forecast errors), and that it requires choosing P.

The choice of P entails a trade-off between the precision of the coefficient estimates and the number of observations available for estimating the MSFE. In practice, choosing P to be 10% or 20% of the total number of observations can provide a reasonable balance between these two considerations.

Application to GDP growth. For the AR(1) in Equation (15.9), $\widehat{RMSFE}_{SER} = 3.05$, $\widehat{RMSFE}_{FPE} = 3.07$, and $\widehat{RMSFE}_{POOS} = 2.60$ (computed over the final 44 quarters or 20% of the sample). For the AR(2) in Equation (15.11), $\widehat{RMSFE}_{SER} = 3.01$, $\widehat{RMSFE}_{FPE} = 3.03$, and $\widehat{RMSFE}_{POOS} = 2.52$. The FPE estimates are larger than the SER estimates because of the additional factor that estimates the variance from estimating the coefficients. The pseudo out-of-sample estimates of the RMSFE are smaller than the in-sample estimates. In part, this reflects the reduction in the variability of GDP growth that occurred in the early 1980s that is evident in Figure 15.1b, a phenomenon known as the Great Moderation.

Forecast Uncertainty and Forecast Intervals

In any estimation problem, it is good practice to report a measure of the uncertainty of that estimate, and forecasting is no exception. One measure of the uncertainty of a forecast is its root mean squared forecast error (RMSFE). Under the additional assumption that the errors u_t are normally distributed, the estimates of the RMSFE introduced in Section 15.3 can be used to construct a forecast interval—that is, an interval that contains the future value of the variable with a certain probability.

Forecast intervals. A forecast interval is like a confidence interval except that it pertains to a forecast. For example, a 95% **forecast interval** is an interval that contains the future value of the variable being forecasted in 95% of repeated applications.

One important difference between a forecast interval and a confidence interval is that the usual formula for a 95% confidence interval (the estimator ± 1.96 standard errors) is justified by the central limit theorem and therefore holds for a wide range of distributions of the error term. In contrast, because the forecast error in Equation (15.15) includes the future value of the error u_{T+1}, computing a forecast interval requires either estimating the distribution of the error term or making some assumption about that distribution.

In practice, it is convenient to assume that u_{T+1} is normally distributed. Under the assumption of stationarity, the forecast error is the sum of u_{T+1} and a term reflecting the estimation error of the regression coefficients. In large samples, this second term is approximately normally distributed (by the central limit theorem) and is uncorrelated with u_{T+1}. Thus, if u_{T+1} is normally distributed, the forecast error is approximately normally distributed and has a variance equal to the MSFE (Exercise 15.12).

The River of Blood

A s part of its efforts to inform the public about monetary policy decisions, the Bank of England regularly publishes forecasts of inflation. These forecasts combine output from econometric models maintained by professional econometricians at the bank with the expert judgment of the members of the bank's senior staff and Monetary Policy Committee. The forecasts are presented as a set of forecast intervals designed to reflect what these economists consider to be the range of probable paths that inflation might take. In its *Inflation Report*, the bank prints these ranges in red, with the darkest red reserved for the central band. Although the bank prosaically refers to this as the "fan chart," the press has called these spreading shades of red the "river of blood."

The river of blood for February 2017 is shown in Figure 15.4. (In this figure, the blood is blue, not red, so you will need to use your imagination.) This chart shows that, as of February 2017, the bank's economists expected the rate of inflation to rise from below its 2.0% target in early 2017 to 2.7% in

the first quarter of 2018. The economists cited an expected strengthening of demand and a depreciation in the British pound as reasons for the increase in the inflation rate. As it turned out, inflation rose over the next year by more than they had forecasted, to 3.0% in early 2018.

The Bank of England has been a pioneer in the movement toward greater openness by central banks, and other central banks now also publish inflation forecasts. The decisions made by monetary policy makers are difficult ones and affect the lives—and wallets—of many of their fellow citizens. In a democracy in the information age, reasoned the economists at the Bank of England, it is particularly important for citizens to understand the bank's economic outlook and the reasoning behind its difficult decisions.

To see the river of blood in its original red hue, visit the Bank of England's website, at **http://www .bankofengland.co.uk.** To learn more about the performance of the Bank of England inflation forecasts, see Clements (2004).

FIGURE 15.4 The River of Blood

The Bank of England's fan chart for February 2017 shows forecast ranges for inflation.

Source: Reprinted with permission from the Bank of England.
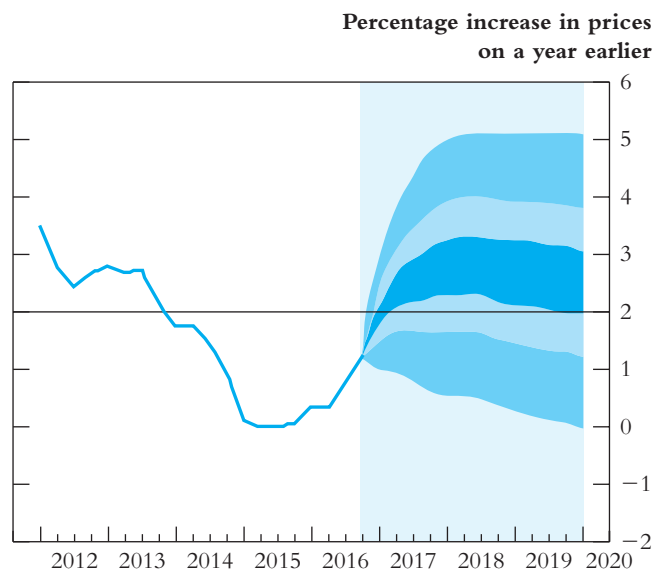
The second two of the estimators of the MSFE, $\widehat{MSE}_{FPE}$ and $\widehat{MSE}_{POOS}$, incorporate estimation error, and either one can be used to construct forecast intervals. That is, if u_{T+1} is normally distributed, a 95% forecast interval is given by $\hat{Y}_{T+1|T} \pm 1.96 \, \widehat{RMSE}$, where $\widehat{RMSE}$ is either $\widehat{RMSE}_{FPE}$ in Equation (15.21) or $\widehat{RMSE}_{POOS}$ in Equation (15.22).

This discussion has focused on the case that u_t is homoskedastic. If instead it is heteroskedastic, then one needs to develop a model of the heteroskedasticity so that the term σ_u^2 in Equation (15.19) can be estimated given the most recent values of Y and X. Methods for modeling this conditional heteroskedasticity are presented in Chapter 17.

Fan charts.   To convey the full range of uncertainty about future values of a variable, professional forecasters sometimes report multiple forecast intervals. Taken together, multiple forecast intervals summarize the full distribution of future values of the variable. A forecast of the distribution of future values of a variable provides a great deal more information to consumers of forecasts than does a forecast of just its mean.

Forecast distributions are frequently conveyed graphically in what is known as a **fan chart**. Fan charts portray the distribution at a future date by shaded overlaid forecast intervals, connected over an expanding forecast horizon. The Bank of England was one of the early users of fan charts as a way to convey forecast paths and uncertainty to the public and to monetary policy makers (see the box "The River of Blood").

15.6 Estimating the Lag Length Using Information Criteria

The estimated GDP growth regressions in Sections 15.3 and 15.4 have either one or two lags of the predictors. Why not more lags? How many lags should you include in a time series regression? This section discusses statistical methods for choosing the number of lags, first in an autoregression and then in a time series regression model with multiple predictors.

Determining the Order of an Autoregression

In practice, choosing the order p of an autoregression requires balancing the marginal benefit of including more lags against the marginal cost of additional estimation error. On the one hand, if the order of an estimated autoregression is too low, you will omit potentially valuable information contained in the more distant lagged values. On the other hand, if it is too high, you will be estimating more coefficients than necessary, which in turn introduces additional estimation error into your forecasts.

The F-statistic approach.   One approach to choosing p is to start with a model with many lags and to perform hypothesis tests on the final lag. For example, you might

start by estimating an AR(6) and test whether the coefficient on the sixth lag is significant at the 5% level; if not, drop it and estimate an AR(5), test the coefficient on the fifth lag, and so forth. The drawback to this method is that it will tend to produce large models: Even if the true AR order is five, so the sixth coefficient is 0, a 5% test using the t-statistic will incorrectly reject this null hypothesis 5% of the time just by chance. Thus, if the true value of p is five, this method will estimate p to be six 5% of the time.

The BIC. One way around this problem is to estimate p by minimizing an information criterion. One such information criterion is the **Bayes information criterion (BIC)**, also called the *Schwarz information criterion (SIC)*, which is

$$\mathrm{BIC}(p) = \ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{\ln(T)}{T}, \tag{15.23}$$

where $SSR(p)$ is the sum of squared residuals of the estimated AR(p). The BIC estimator of p, $\hat{p}$, is the value that minimizes BIC(p) among the possible choices $p = 0, 1, \ldots, p_{max}$, where p_{max} is the largest value of p considered and $p = 0$ corresponds to the model that contains only an intercept.

The formula for the BIC might look a bit mysterious at first, but it has an intuitive interpretation. Consider the first term in Equation (15.23). Because the regression coefficients are estimated by OLS, the sum of squared residuals necessarily decreases (or at least does not increase) when you add a lag. In contrast, the second term is the number of estimated regression coefficients (the number of lags, p, plus one for the intercept) times the factor $\ln(T)/T$. This second term increases when you add a lag and thus provides a penalty for including another lag. The BIC trades off these two forces so that the number of lags that minimizes the BIC is a consistent estimator of the true lag length. Appendix 15.5 provides the mathematics of this argument.

As an example, consider estimating the AR order for an autoregression of the growth rate of GDP. The various steps in the calculation of the BIC are carried out in Table 15.3 for autoregressions of maximum order six $(p_{max} = 6)$. For example, for the AR(1) model in Equation (15.8), $[SSR(1)/T] = 9.247$, so $\ln[SSR(1)/T] = 2.224$. Because $T = 223$ (1962:Q1–2017:Q3), $\ln(T)/T = 0.024$, and $(p+1)\ln(T)/T = 2 \times 0.024 = 0.048$. Thus BIC(1) $= 2.224 + 0.048 = 2.273$.

The BIC is smallest when $p = 2$ in Table 15.3. Thus the BIC estimate of the lag length is 2. As can be seen in Table 15.3, as the number of lags increases, the R^2 increases, and the SSR decreases. The increase in the R^2 is large from zero to one lag, smaller for one to two lags, and smaller yet for other lags. The BIC helps decide precisely how large the increase in the R^2 must be to justify including the additional lag.

The AIC. Another information criterion is the **Akaike information criterion (AIC)**:

$$\mathrm{AIC}(p) = \ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{2}{T}. \tag{15.24}$$

TABLE 15.3	**The Bayes Information Criterion (BIC) and the R^2 for Autoregressive Models of U.S. GDP Growth Rates, 1962:Q1–2017:Q3**				
p	**SSR(p)/T**	**ln[SSR(p)/T]**	**(p + 1) ln(T)/T**	**BIC(p)**	**R²**
0	10.477	2.349	0.024	2.373	0.000
1	9.247	2.224	0.048	2.273	0.117
2	8.954	2.192	0.073	2.265	0.145
3	8.954	2.192	0.097	2.289	0.145
4	8.920	2.188	0.121	2.310	0.149
5	8.788	2.173	0.145	2.319	0.161
6	8.779	2.172	0.170	2.342	0.162

The difference between the AIC and the BIC is that the term $\ln(T)$ in the BIC is replaced by 2 in the AIC, so the second term in the AIC is smaller. For example, for the 223 observations used to estimate the GDP autoregressions, $\ln(T) = \ln(223) = 5.41$, so the second term for the BIC is more than twice as large as the term in the AIC. Thus a smaller decrease in the SSR is needed in the AIC to justify including another lag.

The AIC has an appealing motivation: In large samples, it corresponds to choosing p to minimize the MSFE as estimated by the final prediction error; that is, it minimizes $\widehat{MSFE}_{FPE}$ in Equation (15.21).[5] However, as a matter of theory, the second term in the AIC is not large enough to ensure that the correct lag length is chosen, even in large samples, so the AIC estimator of p is not consistent. As is discussed in Appendix 15.5, in large samples the AIC will overestimate p with nonzero probability.

Both the AIC and the BIC are widely used in practice. If you are concerned that the BIC might yield a model with too few lags, the AIC provides a reasonable alternative.[6]

---

[5]Start with Equation (15.21) to write $\widehat{MSFE}_{FPE} = \left[\dfrac{T + p + 1}{T - (p + 1)}\right]\dfrac{SSR}{T} = \left[\dfrac{1 + (p + 1)/T}{1 - (p + 1)/T}\right]\dfrac{SSR}{T}$.

Taking logarithms of the final expression yields $\ln(\widehat{MSFE}_{FPE}) = \ln\left(1 + \dfrac{p + 1}{T}\right) - \ln\left(1 - \dfrac{p + 1}{T}\right) +$

$\ln\left(\dfrac{SSR}{T}\right) \cong 2\left(\dfrac{p + 1}{T}\right) + \ln\left(\dfrac{SSR}{T}\right)$, where the final expression uses the approximation that

$\ln(1 + x) \cong x$ when x is small [Equation (8.16)]. The final expression is the AIC in Equation (15.24). The approximation $\widehat{MSFE}_{FPE} \approx$ AIC holds when $(p + 1)/T$ is small.

[6]The BIC and the AIC tackle the same problem—restricting the number of parameters to estimate—as the penalized least squares methods of ridge regression and the LASSO discussed in Sections 14.3 and 14.4. One difference between the variable selection problem discussed in Chapter 14 and the lag selection problem discussed here is that, in the general prediction problem with cross-sectional data, there is no natural ordering of the potential regressors. In contrast, in the lag selection problem, it is natural to think that the first lag will be the most useful predictor, followed by the second lag, and so forth, so the predictors have a natural ordering. The AIC and the BIC exploit that natural ordering.

A note on calculating information criteria. For the AIC and BIC to decide between competing regressions with different numbers of lags, those regressions must be estimated using the same observations. For example, in Table 15.3 all the regressions were estimated using data from 1962:Q1 to 2017:Q3, for a total of 223 observations. Because the autoregressions involve lags of the GDP growth rate, this means that the regression uses earlier values of GDP growth (values before 1962:Q1) for initial observations. Said differently, each of the regressions examined in Table 15.3 includes observations on $GDPGR_t$, $GDPGR_{t-1}, \ldots, GDPGR_{t-p}$ for $t = 1962$:Q1, $\ldots$, 2017:Q3 corresponding to 223 observations on the dependent variable and regressors, so $T = 223$ in Equations (15.23) and (15.24).

Lag Length Selection in Time Series Regression with Multiple Predictors

The trade-off involved with lag length choice in the general time series regression model with multiple predictors [Equation (15.18)] is similar to that in an autoregression: Using too few lags can decrease forecast accuracy because valuable information is lost, but adding lags increases estimation error. The choice of lags must balance the benefit of using additional information against the cost of estimating the additional coefficients.

The F-statistic approach. As in the univariate autoregression, one way to determine the number of lags is to use F-statistics to test joint hypotheses that sets of coefficients are equal to 0. For example, in the discussion of Equation (15.15), we tested the hypothesis that the coefficient on the second lag of the term spread was equal to 0 against the alternative that it was nonzero; this hypothesis was not rejected at the 10% significance level, suggesting that the second lag of the term spread could be dropped from the regression. If the number of models being compared is small, then this F-statistic method is easy to use. In general, however, the F-statistic method can produce models that are large and thus have considerable estimation error.

Information criteria. As in an autoregression, the BIC and the AIC can be used to estimate the number of lags and variables in the time series regression model with multiple predictors. If the regression model has K coefficients (including the intercept), the BIC is

$$\mathrm{BIC}(K) = \ln\left[\frac{SSR(K)}{T}\right] + K\frac{\ln(T)}{T}. \tag{15.25}$$

The AIC is defined in the same way, but with 2 replacing $\ln(T)$ in Equation (15.25). For each candidate model, the BIC (or the AIC) can be evaluated, and the model with the lowest value of the BIC (or the AIC) is the preferred model, based on the information criterion.

There are two important practical considerations when using an information criterion to estimate the lag lengths. First, as is the case for the autoregression, all the candidate models must be estimated over the same sample; in the notation of Equation (15.25), the number of observations used to estimate the model, T, must be the same for all models. Second, when there are multiple predictors, this approach is computationally demanding because it requires computing many different models (many combinations of the lag parameters). In practice, a convenient shortcut is to require all the regressors to have the same number of lags—that is, to require that $p = q_1 = \cdots = q_k$, so that only $p_{max} + 1$ models need to be compared (corresponding to $p = 0, 1, \ldots, p_{max}$. Applying this lag-length selection method to the ADL for GDP growth and the term spread results in the ADL(2, 2) model in Equation (15.15).

15.7 Nonstationarity I: Trends

In Key Concept 15.6, it was assumed that the dependent variable and the regressors are stationary. If this is not the case—that is, if the dependent variable and/or the regressors are nonstationary—then conventional hypothesis tests, confidence intervals, and forecasts can be unreliable. The precise problem created by nonstationarity, and the solution to that problem, depends on the nature of that nonstationarity.

In this and the next section, we examine two types of nonstationarity that are frequently encountered in economic time series: trends and breaks. In each section, we first describe the nature of the nonstationarity and then discuss the consequences for time series regression if this type of nonstationarity is present but ignored. We next present tests for nonstationarity and discuss remedies for, or solutions to, the problems caused by that particular type of nonstationarity. We begin by discussing trends.

What Is a Trend?

A **trend** is a persistent long-term movement of a variable over time. A time series variable fluctuates around its trend.

Inspection of Figure 15.1a suggests that the logarithm of U.S. GDP has a clear upwardly increasing trend. The series in Figures 15.2a, 15.2b, and 15.2c also have trends, but their trends are quite different. The trend in the unemployment rate is increasing from the late 1960s through the early 1980s, then decreasing until the early 2000s, and then increasing again. The \$/£ exchange rate clearly had a prolonged downward trend after the collapse of the fixed exchange rate system in 1972. The logarithm of the Japan Industrial Production Index has a complicated trend: fast growth at first, then moderate growth, and finally no growth.

Deterministic and stochastic trends. There are two types of trends in time series data: deterministic and stochastic. A **deterministic trend** is a nonrandom function of

time. For example, a deterministic trend might be linear in time; if the logarithm of U.S. GDP had a deterministic linear trend, so that it increased by 0.75 percentage points per quarter, this trend could be written as $0.75t$, where t is measured in quarters. In contrast, a **stochastic trend** is random and varies over time. For example, a stochastic trend might exhibit a prolonged period of increase followed by a prolonged period of decrease, like the unemployment rate trend in Figure 15.2a. But stochastic trends can be more subtle. For example, if you look carefully at Figure 15.1a, you will notice that the trend growth rate of GDP is not constant; for example, GDP grew faster in the 1960s than in the 1970s (the plot is steeper in the 1960s than in the 1970s), and it grew faster in the 1990s than in the 2000s.

Like many econometricians, we think it is more appropriate to model economic time series as having stochastic rather than deterministic trends. It is hard to reconcile the predictability implied by a deterministic trend with the complications and surprises faced year after year by workers, businesses, and governments. For example, although the U.S. unemployment rate rose through the 1970s, it was neither destined to rise forever nor destined to fall again. Rather, the slow rise of unemployment rates is now understood to have occurred because of a combination of demographic changes (including an influx of younger workers), bad luck (such as oil price shocks and a productivity slowdown), and monetary policy mistakes. Similarly, the \$/£ exchange rate trended down from 1972 to 1985 and subsequently drifted up, but these movements, too, were the consequences of complex economic forces; because these forces change unpredictably, these trends are usefully thought of as having a large unpredictable, or random, component.

For these reasons, our treatment of trends in economic time series focuses on stochastic rather than deterministic trends, and when we refer to "trends" in time series data, we mean stochastic trends unless we explicitly say otherwise.

The random walk model of a trend. The simplest model of a variable with a stochastic trend is the random walk. A time series Y_t is said to follow a **random walk** if the change in Y_t is i.i.d.—that is, if

$$Y_t = Y_{t-1} + u_t, \tag{15.26}$$

where u_t is i.i.d. We will, however, use the term *random walk* more generally to refer to a time series that follows Equation (15.26), where u_t has conditional mean 0; that is, $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$. Another term for a time series for which $E(\Delta Y_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$ is a *martingale*.

The basic idea of a random walk is that the value of the series tomorrow is its value today plus an unpredictable change: Because the path followed by Y_t consists of random "steps" u_t, that path is a "random walk." The conditional mean of Y_t based on data through time $t - 1$ is Y_{t-1}; that is, because $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$, $E(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots) = Y_{t-1}$. In other words, if Y_t follows a random walk, then the best forecast of tomorrow's value is its value today.

If Y_t follows a random walk, its variance increases over time. Because it does not have a constant variance, a random walk is nonstationary (Exercise 15.13).

Some series, such as the logarithm of U.S. GDP in Figure 15.1a, have an obvious upward tendency, in which case the best forecast of the series must include an adjustment for the tendency of the series to increase. This adjustment leads to an extension of the random walk model to include a tendency to move, or drift, in one direction or the other. This extension is referred to as a **random walk with drift**:

$$Y_t = \beta_0 + Y_{t-1} + u_t, \tag{15.27}$$

where $E(u_t \mid Y_{-1}, Y_{-2}, \ldots) = 0$ and β_0 is the drift in the random walk. If β_0 is positive, then Y_t increases on average. In the random walk with drift model, the best forecast of the series tomorrow is the value of the series today plus the drift β_0.

The random walk model (with drift, as appropriate) is simple yet versatile, and it is the primary model for trends used in this book.

Stochastic trends, autoregressive models, and a unit root. The random walk model is a special case of the AR(1) model [Equation (15.8)] in which $\beta_1 = 1$. In other words, if Y_t follows an AR(1) with $\beta_1 = 1$, then Y_t contains a stochastic trend and is nonstationary. If, however, $|\beta_1| < 1$ and u_t is stationary, then the joint distribution of Y_t and its lags does not depend on t (a result shown in Appendix 15.2), so Y_t is stationary.

The analogous condition for an AR(p) to be stationary is more complicated than the condition $|\beta_1| < 1$ for an AR(1). Its formal statement involves the roots of the polynomial, $1 - \beta_1 z - \beta_2 z^2 - \beta_3 z^3 - \cdots - \beta_p z^p$. (The roots of this polynomial are the values of z that satisfy $1 - \beta_1 z - \beta_2 z^2 - \beta_3 z^3 - \cdots - \beta_p z^p = 0$.) For an AR($p$) to be stationary, the roots of this polynomial must all be greater than 1 in absolute value. In the special case of an AR(1), the root is the value of z that solves $1 - \beta_1 z = 0$, so its root is $z = 1/\beta_1$. Thus the statement that the root must be greater than 1 in absolute value is equivalent to $|\beta_1| < 1$.

If an AR(p) has a root that equals 1, the series is said to have a *unit autoregressive root* or, more simply, a **unit root**. If Y_t has a unit root, then it contains a stochastic trend. If Y_t is stationary (and thus does not have a unit root), it does not contain a stochastic trend. For this reason, we will use the terms *stochastic trend* and *unit root* interchangeably.

Problems Caused by Stochastic Trends

If a regressor has a stochastic trend (that is, has a unit root), then inferences made using the OLS estimator of the autoregressive coefficient can be misleading. Moreover, two series that are independent but have stochastic trends will, with high probability, misleadingly appear to be related, a situation known as spurious regression.

Downward bias and nonnormal distributions of the OLS estimator and t-statistic. If
a regressor has a stochastic trend, then its usual OLS *t*-statistic can have a nonnormal
distribution under the null hypothesis, even in large samples, and the estimate of the
autoregressive coefficient is biased toward 0. This nonnormal distribution means that
conventional confidence intervals are not valid and hypothesis tests cannot be con-
ducted as usual.

The downward bias of the OLS estimator poses a problem for forecasts. Recall
that the oracle forecast is the conditional mean. If the coefficient in an AR(1) model
of the conditional mean is 1 (a unit root), then the OLS estimator will tend to take
on a value less than 1, and its sampling distribution has a mean that is less than 1. In
a forecasting application, this can lead to systematic bias in the forecast. Moreover,
because the distribution of the *t*-statistic testing that coefficient is not normal, even
in large samples, standard inferences based on that *t*-statistic will not detect this mis-
take of downward-biased forecasts. Fortunately, as is discussed later in this section,
there are ways to detect whether a series has a unit root and thus to avoid these
problems.

Spurious regression. Stochastic trends can lead two time series to appear related
when they are not, a problem called **spurious regression**.

For example, the U.S. unemployment rate was steadily rising from the mid-1960s
through the early 1980s, and at the same time, Japanese industrial production (plotted
in logarithms in Figure 15.2c) was steadily rising. These two trends conspire to
produce a regression that appears to be "significant" using conventional measures.
Estimated by OLS using data from 1962 through 1985, this regression is

$$\widehat{U.\,S.\ Unemployment\ Rate}_t = -2.37 + 2.22 \times \ln(Japanese\ IP_t),\ \overline{R}^2 = 0.34.$$
$$(1.19)\ \ (0.32) \tag{15.28}$$

The *t*-statistic on the slope coefficient is 7, which by usual standards indicates a strong
positive relationship between the two series, and the $\overline{R}^2$ is moderately high. However,
running this regression using data from 1986 through 2017 yields

$$\widehat{U.\,S.\ Unemployment\ Rate}_t = 42.37 - 7.92 \times \ln(Japanese\ IP_t),\ \overline{R}^2 = 0.14.$$
$$(7.74)\ \ (1.69) \tag{15.29}$$

The regressions in Equations (15.28) and (15.29) could hardly be more different.
Interpreted literally, Equation (15.28) indicates a strong positive relationship, while
Equation (15.29) indicates a negative relationship.

The source of these conflicting results is that both series have stochastic trends.
These trends happened to align from 1962 through 1985 but were reversed from 1986
through 2017. There is, in fact, no compelling economic or political reason to think that
the trends in these two series are related. In short, these regressions are spurious.

The regressions in Equations (15.28) and (15.29) illustrate empirically the
theoretical point that OLS can be misleading when the series contain stochastic
trends. (See Exercise 15.6 for a computer simulation that demonstrates this result.)

One special case in which certain regression-based methods *are* reliable is when the trend component of the two series is the same—that is, when the series contain a *common* stochastic trend; in such a case, the series are said to be cointegrated. Econometric methods for detecting and analyzing cointegrated economic time series are discussed in Chapter 17.

Detecting Stochastic Trends: Testing for a Unit AR Root

The starting point for detecting a trend in a time series is inspecting its time series plot. If the series looks like it might have a trend, the hypothesis that it has a stochastic trend can be tested using a Dickey–Fuller test.

The Dickey–Fuller test in the AR(1) model. The random walk in Equation (15.27) is a special case of the AR(1) model with $\beta_1 = 1$. Thus, when Y_t follows an AR(1), the hypothesis that Y_t has a stochastic trend corresponds to

$$H_0: \beta_1 = 1 \text{ vs. } H_1: \beta_1 < 1, \text{ where } Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t. \quad (15.30)$$

The null hypothesis in Equation (15.30) is that the AR(1) has a unit root, and the one-sided alternative is that it is stationary.

This test is most easily implemented by estimating a modified version of Equation (15.30), obtained by subtracting Y_{t-1} from both sides. Let $\delta = \beta_1 - 1$; then Equation (15.30) becomes

$$H_0: \delta = 0 \text{ vs. } H_1: \delta < 0, \text{ where } \Delta Y_t = \beta_0 + \delta Y_{t-1} + u_t. \quad (15.31)$$

The OLS t-statistic testing $\delta = 0$ in Equation (15.31) is called the **Dickey–Fuller statistic** [Dickey and Fuller (1979)]. The Dickey–Fuller statistic is computed using nonrobust standard errors—that is, the homoskedasticity-only standard errors, presented in Appendix 5.1.[7]

Critical values for the ADF statistic. Under the null hypothesis of a unit root, the Dickey–Fuller statistic does not have a normal distribution, even in large samples. Because its distribution is nonnormal, a different set of critical values is required.

The critical values for the ADF test of the null and alternative hypotheses in Equation (15.31) are given in the first row of Table 15.4. Because the alternative hypothesis of stationarity implies that $\delta < 0$ in Equation (15.31), the ADF test is one-sided. For example, if the regression does not include a time trend, then the hypothesis of a unit root is rejected at the 5% significance level if the ADF statistic is less than -2.86.

---

[7]Under the null hypothesis of a unit root, the usual nonrobust standard errors produce a t-statistic that is, in fact, robust to heteroskedasticity, a surprising and special result.

TABLE 15.4	**Large-Sample Critical Values of the Augmented Dickey–Fuller Statistic**		
Deterministic Regressors	**10%**	**5%**	**1%**
Intercept only	−2.57	−2.86	−3.43
Intercept and time trend	−3.12	−3.41	−3.96

The critical values in Table 15.4 are substantially larger (more negative) than the one-sided critical values of −1.28 (at the 10% level) and –1.64 (at the 5% level) from the standard normal distribution. The nonstandard distribution of the ADF statistic is an example of how OLS *t*-statistics for regressors with stochastic trends can have nonnormal distributions.

The Dickey–Fuller test in the AR(p) model.   The Dickey–Fuller statistic in Equation (15.31) applies to first-order autoregression. The extension of the Dickey–Fuller test to the AR(p) model entails including $p - 1$ lags of ΔY_t as additional regressors, so that Equation (15.31) becomes

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \cdots + \gamma_{p-1} \Delta Y_{t-p+1} + u_t. \quad (15.32)$$

Under the null hypothesis that $\delta = 0$, Y_t has a stochastic trend; under the alternative hypothesis that $\delta < 0$, Y_t is stationary. The *t*-statistic testing the hypothesis that $\Delta = 0$ in Equation (15.32) is called the **augmented Dickey–Fuller (ADF) statistic**. In general, the lag length p is unknown, but it can be estimated using an information criterion applied to regressions of the form in Equation (15.32) for various values of p. Studies of the ADF statistic suggest that it is better to have too many lags than too few, so it is recommended to use the AIC instead of the BIC to estimate p for the ADF statistic.[8]

Testing against the alternative of stationarity around a linear deterministic time trend.   The discussion so far has considered the null hypothesis that a series has a unit root and the alternative hypothesis that it is stationary. This alternative hypothesis of stationarity is appropriate for series such as the unemployment rate that do not exhibit growth over the long run. But for series such as U.S. GDP, the alternative of stationarity around a constant mean is inappropriate, and it makes more sense to test for stationarity around a deterministic trend. One specific formulation of this alternative hypothesis is that the trend is a linear function of t. Thus the null hypothesis is that the series has a unit root, and the alternative is that it does not have a unit root but does have a deterministic time trend.

---

[8]See Stock (1994) and Haldrup and Jansson (2006) for reviews of simulation studies of the finite-sample properties of the Dickey–Fuller and other unit root test statistics.

If the alternative hypothesis is that Y_t is stationary around a deterministic linear time trend, then this trend, t (the observation number), must be added as an additional regressor, in which case the Dickey–Fuller regression becomes

$$\Delta Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \cdots + \gamma_{p-1} \Delta Y_{t-p+1} + u_t, \quad (15.33)$$

where α is an unknown coefficient. The ADF statistic now is the OLS t-statistic testing $\delta = 0$ in Equation (15.33), and the one-sided critical values are given in the second row of Table 15.4.[9]

Does U.S. GDP have a stochastic trend? The null hypothesis that the logarithm of U.S. GDP has a stochastic trend can be tested against the alternative that it is stationary by performing the ADF test for a unit autoregressive root. The ADF regression with two lags of $\Delta \ln(GDP_t)$ is

$$\widehat{\Delta \ln(GDP_t)} = 0.162 + 0.0001t - 0.019 \ln(GDP_{t-1})$$
$$(0.080) \quad (0.0001) \quad (0.010)$$
$$\quad (15.34)$$
$$+ 0.261 \Delta \ln(GDP_{t-1}) + 0.165 \Delta \ln(GDP_{t-2}).$$
$$(0.066) \quad\quad\quad\quad (0.066)$$

The ADF t-statistic is the t-statistic testing the hypothesis that the coefficient on $\ln(GDP_{t-1})$ is 0; this is, $t = -1.95$. From Table 15.4, the 10% critical value is -3.12. Because the ADF statistic of -1.95 is less negative than -3.12, the test does not reject the null hypothesis at the 10% significance level. Based on the regression in Equation (15.34), we therefore cannot reject (at the 10% significance level) the null hypothesis that the logarithm of GDP has a unit autoregressive root—that is, that $\ln(GDP)$ has a stochastic trend—against the alternative that it is stationary around a linear trend.

Avoiding the Problems Caused by Stochastic Trends

The most reliable way to handle a trend in a series is to transform the series so that it does not have the trend. If the series has a stochastic trend, then its difference does not. For example, if Y_t follows a random walk, so that $Y_t = \beta_0 + Y_{t-1} + u_t$, then $\Delta Y_t = \beta_0 + u_t$ is stationary. Thus using first differences eliminates random walk trends in a series.

In practice, you can rarely be sure whether a series has a stochastic trend. Recall that, as a general point, failure to reject the null hypothesis does not necessarily mean that the null hypothesis is true; rather, it simply means that you have insufficient evidence to conclude that it is false. Thus failure to reject the null hypothesis of a unit root using the ADF statistic does not mean that the series actually *has* a unit root. Even though failure to reject the null hypothesis of a unit root does not mean the series has

---

[9]For extensions of the Dickey–Fuller test to nonlinear time trends, see Maddala and Kim (1998).

a unit root, it still can be reasonable to approximate the true autoregressive root as equaling 1 and therefore to use differences of the series rather than its levels.[10]

15.8 Nonstationarity II: Breaks

A second type of nonstationarity arises when the population regression function changes over the course of the sample. In economics, this can occur for a variety of reasons, such as changes in economic policy, changes in the structure of the economy, or changes in a specific industry due to an invention. If such changes, or **breaks**, occur, then a regression model that neglects those changes can provide a misleading basis for inference and forecasting. It is therefore important to check a forecasting model for breaks and to adjust the model if one is found.

What Is a Break?

Breaks can arise either from a discrete change in the population regression coefficients at a distinct date or from a gradual evolution of the coefficients over a longer period of time.

One source of discrete breaks in macroeconomic data is a major change in macroeconomic policy. For example, the breakdown of the Bretton Woods system of fixed exchange rates in 1972 produced the break in the time series behavior of the $/£ exchange rate that is evident in Figure 15.2b. Prior to 1972, the exchange rate was essentially constant, with the exception of a single devaluation in 1968, when the official value of the pound relative to the dollar was decreased. In contrast, since 1972 the exchange rate has fluctuated over a very wide range.

Breaks also can occur more slowly, as the population regression evolves over time. For example, such changes can arise because of slow evolution of economic policy and ongoing changes in the structure of the economy. The methods for detecting breaks described in this section can detect both types of breaks: distinct changes and slow evolution.

Problems caused by breaks. If a break occurs in the population regression function during the sample, then the OLS regression estimates over the full sample will estimate a relationship that holds on average in the sense that the estimate combines the two different periods. Depending on the location and the size of the break, the "average" regression function can be quite different from the true regression function at the end of the sample, and this leads to poor forecasts.

Testing for Breaks

One way to detect breaks is to test for discrete changes, or breaks, in the regression coefficients. How this is done depends on whether the **break date** (the date of the suspected break) is known.

---

[10]For additional discussion of stochastic trends in economic time series variables and of the problems they pose for regression analysis, see Stock and Watson (1988).

Testing for a break at a known date. In some applications, you might suspect that there is a break at a known date. For example, if you are studying international trade relationships using data from the 1970s, you might hypothesize that there is a break in the population regression function of interest in 1972, when the Bretton Woods system of fixed exchange rates was abandoned in favor of floating exchange rates.

If the date of the hypothesized break in the coefficients is known, then the null hypothesis of no break can be tested using a binary variable interaction regression (Key Concept 8.4). To keep things simple, consider an ADL(1, 1) model, so there is an intercept, a single lag of Y_t, and a single lag of X_t. Let τ denote the hypothesized break date, and let $D_t(\tau)$ be a binary variable that equals 0 before the break date and 1 after, so $D_t(\tau) = 0$ if $t \leq \tau$ and $D_t(\tau) = 1$ if $t > \tau$. Then the regression including the binary break indicator and all interaction terms is

$$
\begin{aligned}
Y_t = {} & \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \gamma_0 D_t(\tau) + \gamma_1 [D_t(\tau) \times Y_{t-1}] \\
& + \gamma_2 [D_t(\tau) \times X_{t-1}] + u_t.
\end{aligned} \tag{15.35}
$$

If there is not a break, then the population regression function is the same over both parts of the sample, so the terms involving the break binary variable $D_t(\tau)$ do not enter Equation (15.35). That is, under the null hypothesis of no break, $\gamma_0 = \gamma_1 = \gamma_2 = 0$. Under the alternative hypothesis that there is a break, the population regression function is different before and after the break date τ, in which case at least one of the γ's is nonzero. Thus the hypothesis of a break can be tested using the F-statistic that tests the hypothesis that $\gamma_0 = \gamma_1 = \gamma_2 = 0$ against the hypothesis that at least one of the γ's is nonzero. This is often called a Chow test for a break at a known break date, named for its inventor, Gregory Chow (1960).

If there are multiple predictors or more lags, then this test can be extended by constructing binary variable interaction variables for all the regressors and testing the hypothesis that all the coefficients on terms involving $D_t(\tau)$ are 0.

This approach can be modified to check for a break in a subset of the coefficients by including only the binary variable interactions for that subset of regressors of interest.

Testing for a break at an unknown date. Often the date of a possible break is unknown or known only within a range. Suppose, for example, you suspect that a break occurred sometime between two dates, τ_0 and τ_1. The Chow test can be extended to handle this situation by testing for breaks at all possible dates τ between τ_0 and τ_1 and then using the largest of the resulting F-statistics to test for a break at an unknown date. This modified Chow test is variously called the **Quandt likelihood ratio (QLR) statistic** (Quandt 1960) (the term we shall use) or, more obscurely, the sup-Wald statistic.

Because the QLR statistic is the largest of many F-statistics, its distribution is not the same as an individual F-statistic. Instead, the critical values for the QLR statistic must be obtained from a special distribution. Like the F-statistic, this distribution depends on the number of restrictions being tested, q—that is, the number of coefficients (including the intercept) that are being allowed to break, or change, under the alternative hypothesis. The distribution of the QLR statistic also depends on

τ_0/T and τ_1/T—that is, on the endpoints, τ_0 and τ_1, of the subsample over which the F-statistics are computed, expressed as a fraction of the total sample size.

For the large-sample approximation to the distribution of the QLR statistic to be a good one, the subsample endpoints, τ_0 and τ_1, cannot be too close to the beginning or the end of the sample. For this reason, in practice the QLR statistic is computed over a "trimmed" range, or subset, of the sample. A common choice is to use 15% trimming—that is, to set $\tau_0 = 0.15T$ and $\tau_1 = 0.85T$ (rounded to the nearest integer). With 15% trimming, the F-statistic is computed for break dates in the central 70% of the sample.

The critical values for the QLR statistic, computed with 15% trimming, are given in Table 15.5. Comparing these critical values with those of the $F_{q,\infty}$ distribution (Appendix Table 4) shows that the critical values for the QLR statistics are larger.

TABLE 15.5	Critical Values of the QLR Statistic with 15% Trimming		
Number of Restrictions (q)	**10%**	**5%**	**1%**
1	7.12	8.68	12.16
2	5.00	5.86	7.78
3	4.09	4.71	6.02
4	3.59	4.09	5.12
5	3.26	3.66	4.53
6	3.02	3.37	4.12
7	2.84	3.15	3.82
8	2.69	2.98	3.57
9	2.58	2.84	3.38
10	2.48	2.71	3.23
11	2.40	2.62	3.09
12	2.33	2.54	2.97
13	2.27	2.46	2.87
14	2.21	2.40	2.78
15	2.16	2.34	2.71
16	2.12	2.29	2.64
17	2.08	2.25	2.58
18	2.05	2.20	2.53
19	2.01	2.17	2.48
20	1.99	2.13	2.43

Note: These critical values apply when $\tau_0 = 0.15T$ and $\tau_1 = 0.85T$ (rounded to the nearest integer), so the F-statistic is computed for all potential break dates in the central 70% of the sample. The number of restrictions q is the number of restrictions tested by each individual F-statistic. Critical values for other trimming percentages are given in Andrews (2003).

The QLR Test for Coefficient Stability

Let $F(\tau)$ denote the F-statistic testing the hypothesis of a break in the regression coefficients at date τ; in the regression in Equation (15.35), for example, this is the F-statistic testing the null hypothesis that $\gamma_0 = \gamma_1 = \gamma_2 = 0$. The QLR (or sup-Wald) test statistic is the largest of the F-statistics in the range $\tau_0 \le \tau \le \tau_1$:

$$\text{QLR} = \max[F(\tau_0), F(\tau_0 + 1), \dots, F(\tau_1)]. \qquad (15.36)$$

1. Like the F-statistic, the QLR statistic can be used to test for a break in all or just some of the regression coefficients.

2. In large samples, the distribution of the QLR statistic under the null hypothesis depends on the number of restrictions being tested, q, and on the endpoints τ_0 and τ_1 as a fraction of T. Critical values are given in Table 15.5 for 15% trimming ($\tau_0 = 0.15T$ and $\tau_1 = 0.85T$, rounded to the nearest integer).

3. The QLR test can detect a single discrete break, multiple discrete breaks, and/or slow evolution of the regression function.

4. If there is a distinct break in the regression function, the date at which the largest Chow statistic occurs is an estimator of the break date.

This reflects the fact that the QLR statistic looks at the largest of many individual F-statistics. By examining F-statistics at many possible break dates, the QLR statistic has many opportunities to reject the null hypothesis, leading to QLR critical values that are larger than the individual F-statistic critical values.

The QLR test can be used to test for a break in only some of the regression coefficients by using interactions between the date binary indicators and only the variables in question, and then computing the largest of the resulting F-statistics. The critical values for this version of the QLR test are also taken from Table 15.5, where the number of restrictions (q) is the number of restrictions tested.

If there is a discrete break at a date within the range tested, the date at which the constituent F-statistic is at its maximum, $\hat{\tau}$, is an estimate of the break date τ.

The QLR statistic also rejects the null hypothesis with high probability in large samples when there are multiple discrete breaks or when the break comes in the form of a slow evolution of the regression function. This means that the QLR statistic detects forms of instability other than a single discrete break. As a result, if the QLR statistic rejects the null hypothesis, it can mean that there is a single discrete break, that there are multiple discrete breaks, or that there is slow evolution of the regression function.

The QLR statistic is summarized in Key Concept 15.8.

***Warning: You probably don't know the break date even if you think you
do.*** Sometimes an expert might believe that he or she knows the date of a possible
break, so that the Chow test can be used instead of the QLR test. But if this knowl-
edge is based on the expert's knowledge of the series being analyzed, then, in fact,
this date was estimated using the data, albeit in an informal way. Preliminary estima-
tion of the break date means that the usual F critical values cannot be used for the
Chow test for a break at that date. Thus it remains appropriate to use the QLR sta-
tistic in this circumstance.

Application: Has the predictive power of the term spread been stable? The QLR test
provides a way to check whether the GDP–term spread relation has been stable from
1962 to 2017. Specifically, we focus on whether there have been changes in the
coefficients on the lagged values of the term spread and the intercept in the ADL(2, 2)
specification in Equation (15.15), containing two lags each of $GDPGR_t$ and $TSpread_t$.

The Chow F-statistics testing the hypothesis that the intercept and the coeffi-
cients on $TSpread_{t-1}$, $TSpread_{t-2}$, and the intercept in Equation (15.15) are constant
against the alternative that they break at a given date are plotted in Figure 15.5 for
breaks in the central 70% of the sample. For example, the F-statistic testing for a
break in 1975:Q1 is 2.07, the value plotted at that date in the figure. Each F-statistic
tests three restrictions (no change in the intercept and in the two coefficients on lags
of the term spread), so $q = 3$. The largest of these F-statistics is 6.47, which occurs in
1980:Q4; this is the QLR statistic. Comparing 6.47 to the critical values for $q = 3$ in
Table 15.5 indicates that the hypothesis that these coefficients are stable is rejected
at the 1% significance level. (The 1% critical value is 6.02.) Thus, there is statistically
significant evidence that at least one of these coefficients changed over the sample.

FIGURE 15.5   *F-Statistics Testing for a Break in Equation (15.15) at Different Dates*
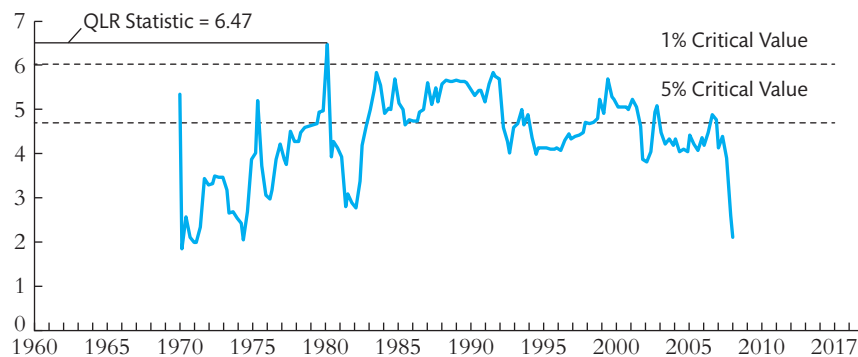
At a given break date, the *F*-statistic plotted here tests the null hypothesis of a break in at least one of the coefficients
on *TSpread*$_{t-1}$, *TSpread*$_{t-2}$, or the intercept in Equation (15.15). For example, the *F*-statistic testing for a break in 1975:Q1
is 2.07. The QLR statistic, 6.47, is the largest of these *F*-statistics and exceeds the 1% critical value of 6.02.

Detecting Breaks Using Pseudo Out-of-Sample Forecasts

The ultimate test of a forecasting model is its out-of-sample performance—that is, its forecasting performance in "real time," after the model has been estimated. Pseudo out-of-sample forecasting, introduced in Key Concept 15.7 for the purpose of estimating the MSFE, simulates the real-time performance of a forecasting model and can be used to detect breaks near the end of the sample.

The most direct and often most useful way to do so is via a time series plot of the in-sample predicted values, the pseudo out-of-sample forecasts, and the actual values of the series. A visible deterioration of the forecasts in the pseudo out-of-sample period is a red flag warning of a possible breakdown of the forecasting model. Another check is to compare $\widehat{MSFE}_{POOS}$ with $\widehat{MSFE}_{FPE}$, where $\widehat{MSFE}_{FPE}$ is computed on the same estimation sample as used for $\widehat{MSFE}_{POOS}$ (the first $T - P$ observations). If the series is stationary, these two estimates of the MSFE should be numerically close. A value of $\widehat{MSFE}_{POOS}$ that is much larger than $\widehat{MSFE}_{FPE}$ suggests some violation of stationarity, possibly a breakdown of the forecasting equation.

Application: Did the predictive power of the term spread change during the 2000s? Using the QLR statistic, we rejected the null hypothesis that the predictive power of the term spread has been stable against the alternative of a break at the 1% significance level, with a break occurring in the early 1980s. Does the ADL(2, 2) model provide a stable forecasting model subsequent to the 1980:Q4 break?

If the coefficients of the ADL(2, 2) model changed toward the end of the 1981:Q1–2017:Q3 period, then pseudo out-of-sample forecasts computed using an estimation sample starting in 1981:Q1 should deteriorate. The pseudo out-of-sample forecasts of the growth rate of GDP for the period 2003:Q1–2017:Q3, computed using the estimation sample of 1981:Q1–2002:Q4 and the method of Key Concept 5.7, are plotted in Figure 15.6, along with the actual values of the growth rate of GDP. The pseudo out-of-sample forecast errors are the differences between the actual growth rate of GDP and its pseudo out-of-sample forecast—that is, the differences between the two lines in Figure 15.6. For example, in 2006:Q4, the growth rate of GDP was 3.1 percentage points (at an annual rate), but the pseudo out-of-sample forecast of $GDPGR_{2006:Q4}$ was 1.6 percentage points, so the pseudo out-of-sample forecast error was $GDPGR_{2006:Q4} - \widehat{GDPGR}_{2006:Q4|2006:Q3} = 1.5$ percentage points. In other words, a forecaster using the ADL(2, 2) estimated through 2006:Q3 would have forecasted GDP growth of 1.6 percentage points in 2006:Q4, whereas in reality GDP grew by 3.1 percentage points.

How do the mean and standard deviation of the pseudo out-of-sample forecast errors compare with the in-sample fit of the model? If the forecasting model is stable, the pseudo out-of-sample forecast errors should have mean 0. However, over the 2003:Q1–2017:Q4 pseudo out-of-sample forecast period, the average forecast error is −0.57, and the t-statistic testing the hypothesis that the mean forecast error equals 0 is −2.00; thus the hypothesis that the forecasts have mean 0 is rejected

FIGURE 15.6   U.S. GDP Growth Rates and Pseudo Out-of-Sample Forecasts

The pseudo out-of-sample forecasts made using the ADL(2, 2) model of the form in Equation (15.15) generally track the actual growth rate of GDP from 2003 to 2017 but fail to forecast the sharp decline in GDP following the financial crisis of 2008.
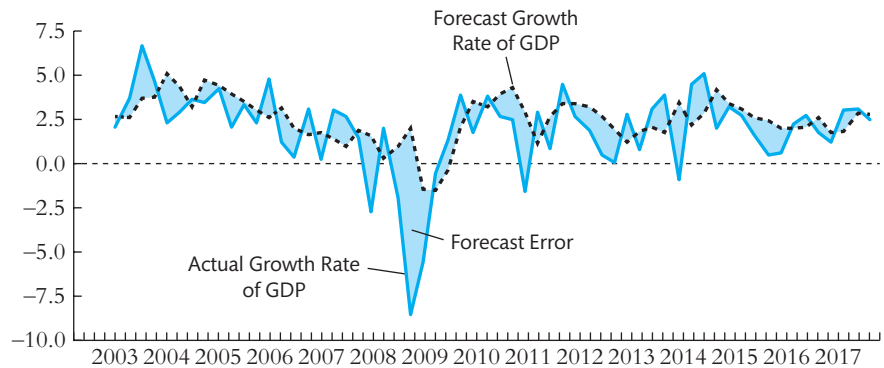
at the 5% significance level. That said, $\widehat{RMSFE}_{FPE} = 2.45$ (1981:Q1–2002:Q4) and $\widehat{RMSFE}_{POOS} = 2.29$ (2003:Q1–2017:Q4), indicating a slight improvement of the forecast in the out-of-sample period. Figure 15.6 shows that the pseudo out-of-sample forecasts track actual GDP growth reasonably well except during late 2008 and early 2009, the period of steepest decline of GDP during the financial crisis and its immediate aftermath. Excluding the single quarter 2008:Q4 lowers $\widehat{RMSFE}_{POOS}$ from 2.29 to 1.85.

According to the pseudo out-of-sample forecasting exercise, the performance of the ADL(2, 2) forecasting model during the pseudo out-of-sample period 2003:Q1–2017:Q4 was, with the exception of the sharp decline in GDP in late 2008, better than its performance during the in-sample period of 1981:Q1–2002:Q4.[11]

Avoiding the Problems Caused by Breaks

How best to adjust for a break in the population regression function depends on the source of that break. If a distinct break occurs at a specific date, that break will be detected with high probability by the QLR statistic, and the break date can be estimated. The regression function can then be reestimated using a binary variable indicating the two subsamples associated with this break and including interactions with the other regressors as appropriate. If all the coefficients break, then this simplifies to reestimating the regression using the post-break data. If there is, in fact, a distinct break, then subsequent inference on the regression coefficients can proceed as usual—for example, using normal critical values for hypothesis tests based on

---

[11]The ADL(2, 2) was not alone in failing to forecast GDP growth in 2008:Q4. Researchers at the Federal Reserve Bank of Philadelphia surveyed 47 professional forecasters in the third quarter of 2008 and asked for their forecasts of the growth rate of GDP in the fourth quarter. The median of the 47 forecasts was 0.7%, lower than the ADL(2, 2) forecast of 2.0%. The actual growth rate of GDP in 2008:Q4 was −8.5%.

t-statistics. In addition, forecasts can be produced using the regression function estimated using the post-break model.

If the break is not distinct but rather arises from a slow, ongoing change in the parameters, the remedy is more difficult and goes beyond the scope of this book.[12]

15.9   Conclusion

In time series data, a variable generally is correlated from one observation, or date, to the next. A consequence of this correlation is that linear regression can be used to forecast future values of a time series based on its current and past values. The starting point for time series regression is an autoregression, in which the regressors are lagged values of the dependent variable. If additional predictors are available, then their lags can be added to the regression. This chapter has described methods for specifying and estimating forecasting regressions, for selecting among competing forecasting regressions, for handling trends in the data, and for assessing the stability of forecasting models.

The time series regressions in this chapter were developed for forecasting, and in general, the coefficients do not have a causal interpretation. In some applications, however, the task is not to develop a forecasting model but rather to estimate causal relationships among time series variables—that is, to estimate the *dynamic* causal effect on *Y over time* of a change in *X*. Under the right conditions, the methods of this chapter, or closely related methods, can be used to estimate dynamic causal effects, and that is the topic of the next chapter.

Summary

1. Regression models used for forecasting need not have a causal interpretation.
2. A time series variable generally is correlated with one or more of its lagged values; that is, it is serially correlated.
3. The accuracy of a forecast is measured by its mean squared forecast error.
4. An autoregression of order p is a linear multiple regression model in which the regressors are the first p lags of the dependent variable. The coefficients of an AR(p) can be estimated by OLS, and the estimated regression function can be used for forecasting. The lag order p can be estimated using an information criterion such as the BIC or the AIC.
5. Adding other variables and their lags to an autoregression can improve forecasting performance. Under the least squares assumptions for prediction with time series regression (Key Concept 15.6), the OLS estimators have normal distributions in large samples, and statistical inference proceeds the same way as for cross-sectional data.

---

[12]For additional discussion of estimation and testing in the presence of discrete breaks, see Hansen (2001). For an advanced discussion of estimation and forecasting when there are slowly evolving coefficients, see Hamilton (1994, Chapter 13).

6. Forecast intervals quantify forecast uncertainty. If the errors are normally distributed, an approximate 68% forecast interval can be constructed as the forecast plus or minus an estimate of the root mean squared forecast error.
7. A series that contains a stochastic trend is nonstationary. A random walk stochastic trend can be detected using the ADF statistic and can be eliminated by using the first difference of the series.
8. If the population regression function changes over time, then OLS estimates neglecting this instability produce unreliable forecasts. The QLR statistic can be used to test for a break, and if a discrete break is found, the regression function can be reestimated allowing for the break.
9. Pseudo out-of-sample forecasts can be used to estimate the root mean squared forecast error, to compare different forecasting models, and to assess model stability toward the end of the sample.

Key Terms

gross domestic product (GDP) (513)
first difference (514)
first lag (514)
j^{th} lag (514)
autocorrelation (516)
serial correlation (516)
autocorrelation coefficient (516)
j^{th} autocovariance (517)
stationarity (519)
nonstationarity (520)
one-step ahead forecast (520)
multi-step ahead forecast (520)
forecast error (520)
mean squared forecast error (MSFE) (521)
root mean squared forecast error (RMSFE) (521)
oracle forecast (523)
autoregression (523)
first-order autoregression (523)
p^{th}-order autoregressive [AR(p)] model (525)
term spread (527)
autoregressive distributed lag (ADL) model (528)
ADL(p, q) (529)

weak dependence (530)
final prediction error (FPE) (532)
pseudo out-of-sample forecasting (533)
forecast interval (534)
fan chart (536)
Bayes information criterion (BIC) (537)
Akaike information criterion (AIC) (537)
trend (540)
deterministic trend (540)
stochastic trend (541)
random walk (541)
random walk with drift (542)
unit root (542)
spurious regression (543)
Dickey–Fuller statistic (544)
augmented Dickey–Fuller (ADF) statistic (545)
break (547)
break date (547)
Quandt likelihood ratio (QLR) statistic (548)
lag operator (564)
lag polynomial (564)
autoregressive–moving average (ARMA) model (565)

Review the Concepts

15.1 Look at the plot of the logarithm of the Japan Index of Industrial Production in Figure 15.2c. Does this time series appear to be stationary? Explain. Suppose you calculated the first difference of this series. Would it appear to be stationary? Explain.

15.2 Many financial economists believe that the random walk model is a good description of the logarithm of stock prices. It implies that the percentage changes in stock prices are unforecastable. A financial analyst claims to have a new model that makes better predictions than the random walk model. Explain how you would examine the analyst's claim that his model is superior.

15.3 A researcher estimates an AR(1) with an intercept and finds that the OLS estimate of β_1 is 0.95, with a standard error of 0.02. Does a 95% confidence interval include $\beta_1 = 1$? Explain.

15.4 Suppose you suspected that the intercept in Equation (15.15) changed in 1992:Q1. How would you modify the equation to incorporate this change? How would you test for a change in the intercept? How would you test for a change in the intercept if you did not know the date of the change?

Exercises

15.1 Consider the AR(1) model $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$. Suppose the process is stationary.

    a. Show that $E(Y_t) = E(Y_{t-1})$. (*Hint:* Read Key Concept 15.3.)

    b. Show that $E(Y_t) = \beta_0/(1 - \beta_1)$.

15.2 The Index of Industrial Production (IP_t) is a monthly time series that measures the quantity of industrial commodities produced in a given month. This problem uses data on this index for the United States. All regressions are estimated over the sample period 1986:M1–2017:M12 (that is, January 1986 through December 2017). Let $Y_t = 1200 \times \ln(IP_t/IP_{t-1})$.

a. A forecaster states that Y_t shows the monthly percentage change in *IP*, measured in percentage points per annum. Is this correct? Why?

b. Suppose she estimates the following AR(4) model for Y_t:

$$\hat{Y}_t = 0.749 + 0.071Y_{t-1} + 0.170Y_{t-2} + 0.216Y_{t-3} + 0.167Y_{t-4}.$$
$$\quad\ (0.488)\ \ (0.088)\qquad (0.053)\qquad (0.078)\qquad (0.064)$$

Use this AR(4) to forecast the value of Y_t in January 2018, using the following values of *IP* for July 2017 through December 2017:

Date	2017:M7	2017:M8	2017:M9	2017:M10	2017:M11	2017:M12
IP	105.01	104.56	104.82	106.58	106.86	107.30

c. Worried about potential seasonal fluctuations in production, she adds Y_{t-12} to the autoregression. The estimated coefficient on Y_{t-12} is −0.061, with a standard error of 0.043. Is this coefficient statistically significant?

d. Worried about a potential break, she computes a QLR test (with 15% trimming) on the constant and AR coefficients in the AR(4) model. The resulting QLR statistic is 1.80. Is there evidence of a break? Explain.

e. Worried that she might have included too few or too many lags in the model, the forecaster estimates AR(*p*) models for $p = 0, 1, \ldots, 6$ over the same sample period. The sum of squared residuals from each of these estimated models is shown in the table. Use the BIC to estimate the number of lags that should be included in the autoregression. Do the results differ if you use the AIC?

AR Order	0	1	2	3	4	5	6
SSR	21,045	20,043	18,870	17,838	17,344	17,337	17,306

15.3. Using the same data as in Exercise 15.2, a researcher tests for a stochastic trend in $\ln(IP_t)$, using the following regression:

$$\widehat{\Delta \ln(IP_t)} = 0.026 + 0.000097t - 0.0070\ln(IP_{t-1}) + 0.068\Delta\ln(IP_{t-1})$$
$$\qquad\quad (0.013)\ \ (0.000067)\quad (0.0037)\qquad\qquad (0.050)$$

$$\qquad + 0.169\Delta\ln(IP_{t-2}) + 0.219\Delta\ln(IP_{t-3}) + 0.173\Delta\ln(IP_{t-4}),$$
$$\qquad\quad (0.049)\qquad\qquad (0.050)\qquad\qquad (0.051)$$

where the standard errors shown in parentheses are computed using the homoskedasticity-only formula and the regressor *t* is a linear time trend.

a. Use the ADF statistic to test for a stochastic trend (unit root) in $\ln(IP)$.

b. Do these results support the specification used in Exercise 15.2? Explain.

15.4   The forecaster in Exercise 15.2 augments her AR(4) model for *IP* growth to include four lagged values of ΔR_t, where R_t is the interest rate on three-month U.S. Treasury bills (measured in percentage points at an annual rate).

    a.   The *F*-statistic on the four lags of ΔR_t is 3.91. Do interest rates help predict *IP* growth? Explain.

    b.   The researcher also regresses ΔR_t on a constant, four lags of ΔR_t, and four lags of *IP* growth. The resulting *F*-statistic on the four lags of *IP* growth is 1.48. Does *IP* growth help to predict interest rates? Explain.

15.5   Prove the following results about conditional means, forecasts, and forecast errors:

    a.   Let *W* be a random variable with mean μ_W and variance σ_w^2, and let *c* be a constant. Show that $E[(W - c)^2] = \sigma_w^2 + (\mu_W - c)^2$.

    b.   Consider the problem of forecasting Y_t, using data on $Y_{t-1}, Y_{t-2}, \ldots$. Let f_{t-1} denote some forecast of Y_t, where the subscript $t - 1$ on f_{t-1} indicates that the forecast is a function of data through date $t - 1$. Let $E[(Y_t - f_{t-1})^2 \mid Y_{t-1}, Y_{t-2}, \ldots]$ be the conditional mean squared error of the forecast f_{t-1}, conditional on values of *Y* observed through date $t - 1$. Show that the conditional mean squared forecast error is minimized when $f_{t-1} = Y_{t|t-1}$, where $Y_{t|t-1} = E(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots)$. (*Hint:* Review Appendix 2.2.)

    c.   Let u_t denote the error in Equation (15.12). Show that $\mathrm{cov}(u_t, u_{t-j}) = 0$ for $j \neq 0$. [*Hint:* Use Equation (2.28).]

15.6   In this exercise, you will conduct a Monte Carlo experiment to study the phenomenon of spurious regression discussed in Section 15.7. In a Monte Carlo study, artificial data are generated using a computer, and then those artificial data are used to calculate the statistics being studied. This makes it possible to compute the distribution of statistics for known models when mathematical expressions for those distributions are complicated (as they are here) or even unknown. In this exercise, you will generate data so that two series, Y_t and X_t, are independently distributed random walks. The specific steps are as follows:

    i.   Use your computer to generate a sequence of $T = 100$ i.i.d. standard normal random variables. Call these variables $e_1, e_2, \ldots, e_{100}$. Set $Y_1 = e_1$ and $Y_t = Y_{t-1} + e_t$ for $t = 2, 3, \ldots, 100$.

    ii.   Use your computer to generate a new sequence, $a_1, a_2, \ldots, a_{100}$, of $T = 100$ i.i.d. standard normal random variables. Set $X_1 = a_1$ and $X_t = X_{t-1} + a_t$ for $t = 2, 3, \ldots, 100$.

    iii.   Regress Y_t onto a constant and X_t. Compute the OLS estimator, the regression R^2, and the (homoskedasticity-only) *t*-statistic testing the null hypothesis that β_1 (the coefficient on X_t) is 0.

Use this algorithm to answer the following questions:

a. Run the algorithm (i) through (iii) once. Use the t-statistic from (iii) to test the null hypothesis that $\beta_1 = 0$, using the usual 5% critical value of 1.96. What is the R^2 of your regression?

b. Repeat (a) 1000 times, saving each value of R^2 and the t-statistic. Construct a histogram of the R^2 and t-statistic. What are the 5%, 50%, and 95% percentiles of the distributions of the R^2 and the t-statistic? In what fraction of your 1000 simulated data sets does the t-statistic exceed 1.96 in absolute value?

c. Repeat (b) for different numbers of observations, such as $T = 50$ and $T = 200$. As the sample size increases, does the fraction of times that you reject the null hypothesis approach 5%, as it should because you have generated Y and X to be independently distributed? Does this fraction seem to approach some other limit as T gets large? What is that limit?

15.7 Suppose Y_t follows the stationary AR(1) model $Y_t = 2.5 + 0.7Y_{t-1} + u_t$, where u_t is i.i.d. with $E(u_t) = 0$ and $\text{var}(u_t) = 9$.

a. Compute the mean and variance of Y_t. (*Hint:* See Exercise 15.1.)

b. Compute the first two autocovariances of Y_t. (*Hint:* Read Appendix 15.2.)

c. Compute the first two autocorrelations of Y_t.

d. Suppose $Y_T = 102.3$. Compute $Y_{T+1|T} = E(Y_{T+1} | Y_T, Y_{t-1}, \ldots)$.

15.8 Suppose Y_t is the monthly value of the number of new home construction projects started in the United States. Because of the weather, Y_t has a pronounced seasonal pattern; for example, housing starts are low in January and high in June. Let μ_{Jan} denote the average value of housing starts in January, and let $\mu_{Feb}, \mu_{Mar}, \ldots, \mu_{Dec}$ denote the average values in the other months. Show that the values of $\mu_{Jan}, \mu_{Feb}, \ldots, \mu_{Dec}$ can be estimated from the OLS regression $Y_t = \beta_0 + \beta_1 Feb_t + \beta_2 Mar_t + \cdots + \beta_{11} Dec_t + u_t$, where Feb_t is a binary variable equal to 1 if t is February, Mar_t is a binary variable equal to 1 if t is March, and so forth. (*Hint:* Show that $\beta_0 + \beta_2 = \mu_{Mar}$ and so forth.)

15.9 The moving average model of order q has the form

$$Y_t = \beta_0 + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \cdots + b_q e_{t-q},$$

where e_t is a serially uncorrelated random variable with mean 0 and variance σ_e^2.

a. Show that $E(Y_t) = \beta_0$.

b. Show that the variance of Y_t is $\text{var}(Y_t) = \sigma_e^2(1 + b_1^2 + b_2^2 + \cdots + b_q^2)$.

c. Show that $\rho_j = 0$ for $j > q$.

d. Suppose $q = 1$. Derive the autocovariances for Y.

15.10 A researcher carries out a QLR test using 25% trimming, and there are $q = 5$ restrictions. Answer the following questions, using the values in Table 15.5 and the $F_{m,\infty}$ critical values in Appendix Table 4.

    a. The QLR F-statistic is 4.2. Should the researcher reject the null hypothesis at the 5% level?

    b. The QLR F-statistic is 2.1. Should the researcher reject the null hypothesis at the 5% level?

    c. The QLR F-statistic is 3.5. Should the researcher reject the null hypothesis at the 5% level?

15.11 Suppose ΔY_t follows the AR(1) model $\Delta Y_t = \beta_0 + \beta_1 \Delta Y_{t-1} + u_t$.

    a. Show that Y_t follows an AR(2) model.

    b. Derive the AR(2) coefficients for Y_t as a function of β_0 and β_1.

15.12 Consider the stationary AR(1) model $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$, where u_t is i.i.d. with mean 0 and variance σ_u^2. The model is estimated using data from time periods $t = 1$ through $t = T$, yielding the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. You are interested in forecasting the value of Y at time $T + 1$—that is, Y_{T+1}. Denote the forecast by $\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T$.

    a. Show that the forecast error is $Y_{T+1} - \hat{Y}_{T+1|T} = u_{T+1} - [(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)Y_T]$.

    b. Show that u_{T+1} is independent of Y_T.

    c. Show that u_{T+1} is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$

    d. Show that $\text{var}(Y_{T+1|T} - \hat{Y}_{T+1|T}) = \sigma_u^2 + \text{var}[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)Y_T]$.

15.13 Suppose Y_t follows a random walk, $Y_t = Y_{t-1} + u_t$, for $t = 1, \ldots, T$, where $Y_0 = 0$ and u_t is i.i.d. with mean 0 and variance σ_u^2.

    a. Compute the mean and variance of Y_t.

    b. Compute the covariance between Y_t and Y_{t-k}.

    c. Use the results in (a) and (b) to show that Y_t is nonstationary.

Empirical Exercises

E15.1 On the text website, **http://www.pearsonhighered.com/stock_watson**, you will find the data file **USMacro_Quarterly**, which contains quarterly data on several macroeconomic series for the United States; the data are described in the file **USMacro_Description**. The variable *PCEP* is the price index for personal consumption expenditures from the U.S. National Income and Product Accounts. In this exercise, you will construct forecasting models for the rate of inflation based on *PCEP*. For this analysis, use the sample period

1963:Q1–2017:Q4 (where data before 1963 may be used, as necessary, as initial values for lags in regressions).

a. i. Compute the inflation rate, $Infl = 400 \times [\ln(PCEP_t) - \ln(PCEP_{t-1})]$. What are the units of $Infl$? (Is $Infl$ measured in dollars, percentage points, percentage per quarter, percentage per year, or something else? Explain.)

   ii. Plot the value of $Infl$ from 1963:Q1 through 2017:Q4. Based on the plot, do you think that $Infl$ has a stochastic trend? Explain.

b. i. Compute the first four autocorrelations of $\Delta Infl$.

   ii. Plot the value of $\Delta Infl$ from 1963:Q1 through 2017:Q4. The plot should look choppy or jagged. Explain why this behavior is consistent with the first autocorrelation that you computed in (i).

c. i. Run an OLS regression of $\Delta Infl_t$ on $\Delta Infl_{t-1}$. Does knowing the change in inflation over the current quarter help predict the change in inflation over the next quarter? Explain.

   ii. Estimate an AR(2) model for $\Delta Infl$. Is the AR(2) model better than an AR(1) model? Explain.

   iii. Estimate an AR(p) model for $p = 0, \ldots, 8$. What lag length is chosen by the BIC? What lag length is chosen by the AIC?

   iv. Use the AR(2) model to predict the change in inflation from 2017:Q4 to 2018:Q1—that is, to predict the value of $\Delta Infl_{2018:Q1}$.

   v. Use the AR(2) model to predict the level of the inflation rate in 2018:Q1—that is, $Infl_{2018:Q1}$.

d. i. Use the ADF test for the regression in Equation (15.32) with two lags of $\Delta Infl$ (so that $p = 3$ in Equation (15.32)) to test for a stochastic trend in $Infl$.

   ii. Is the ADF test based on Equation (15.32) preferred to the test based on Equation (15.33) for testing for a stochastic trend in $Infl$? Explain.

   iii. In (i), you used two lags of $\Delta Infl$. Should you use more lags? Fewer lags? Explain.

   iv. Based on the test you carried out in (i), does the AR model for $Infl$ contain a unit root? Explain carefully. (*Hint:* Does the failure to reject a null hypothesis mean that the null hypothesis is true?)

e. Use the QLR test with 15% trimming to test the stability of the coefficients in the AR(2) model for $\Delta Infl$. Is the AR(2) model stable? Explain.

f. i. Using the AR(2) model for $\Delta Infl$ with a sample period that begins in 1963:Q1, compute pseudo out-of-sample forecasts for the change in inflation beginning in 2003:Q1 and going through 2017:Q4. (That is, compute $\widehat{\Delta Infl}_{2003:Q1|2002:Q4}, \widehat{\Delta Infl}_{2003:Q2|2003:Q1}, \ldots, \widehat{\Delta Infl}_{2017:Q4|2017:Q3}.$)

      ii. Are the pseudo out-of-sample forecasts biased? That is, do the forecast errors have a nonzero mean?

      iii. How large is the RMSFE of the pseudo out-of-sample forecasts? Is this consistent with the AR(2) model for $\Delta Infl$ estimated over the 1963:Q1–2002:Q4 sample period?

      iv. There is a large outlier in 2008:Q4. Why did inflation fall so much in 2008:Q4? (*Hint:* Collect some data on oil prices. What happened to oil prices during 2008?)

E15.2 Read the box "Can You Beat the Market?" Next go to the course website, where you will find an extended version of the data set described in the box; the data are in the file **Stock_Returns_1931_2002** and are described in the file **Stock_Returns_1931_2002_Description**.

    a. Repeat the calculations reported in Table 15.2 using regressions estimated over the 1932:M1–2002:M12 sample period.

    b. Construct pseudo out-of-sample forecasts of excess returns over the 1983:M1–2002:M12 period using regressions that begin in 1932:M1.

    c. Do the results in (a)–(b) suggest any important changes to the conclusions reached in the box? Explain.

15.1 Time Series Data Used in Chapter 15

Macroeconomic time series data for the United States are collected and published by various government agencies. The Bureau of Economic Analysis in the Department of Commerce publishes the National Income and Product Accounts, which include the GDP data used in this chapter. The unemployment rate is computed from the Bureau of Labor Statistics' Current Population Survey (see Appendix 3.1). The quarterly data used here were computed by averaging the monthly values. The 10-year Treasury bond rate, 3-month Treasury bill rate, and the $/£ exchange rate data are quarterly averages of daily rates, as reported by the Federal Reserve System. The Japan Index of Industrial Production is published by the Organisation for Economic Co-operation and Development (OECD). The daily percentage change in the Wilshire 5000 Total Market Index, a stock price index, was computed as $100\Delta\ln(W5000_t)$, where $W5000_t$ is the daily value of the index; because the stock exchange is not open on weekends and holidays, the time period of analysis is a business day. We obtained all these data series from the Federal Reserve Economic Data (FRED) website at the Federal Reserve Bank of St. Louis. There you can find times series data on thousands of macroeconomic variables.

    The regressions in Table 15.2 use monthly financial data for the United States. Stock prices (P_t) are measured by the broad-based (NYSE and AMEX), value-weighted index of stock prices constructed by the Center for Research in Security Prices (CRSP). The monthly

percentage excess return is $100 \times \{\ln[(P_t + Div_t)/P_{t-1}] - \ln(TBill_t)\}$, where Div_t is the dividends paid on the stocks in the CRSP index and $TBill_t$ is the gross return (1 plus the interest rate) on a 30-day Treasury bill during month t. We thank Motohiro Yogo for providing both his help and these data.

15.2 Stationarity in the AR(1) Model

This appendix shows that if $|\beta_1| < 1$ and u_t is stationary, then Y_t is stationary. Recall from Key Concept 15.3 that the time series variable Y_t is stationary if the joint distribution of $(Y_{s+1}, \ldots, Y_{s+T})$ does not depend on s, regardless of the value of T. To streamline the argument, we show this for $T = 2$ under the simplifying assumptions that $\beta_0 = 0$ and $\{u_t\}$ are i.i.d. $N(0, \sigma_u^2)$.

The first step is deriving an expression for Y_t in terms of the u_t's. Because $\beta_0 = 0$, Equation (15.8) implies that $Y_t = \beta_1 Y_{t-1} + u_t$. Substituting $Y_{t-1} = \beta_1 Y_{t-2} + u_{t-1}$ into this expression yields $Y_t = \beta_1(\beta_1 Y_{t-2} + u_{t-1}) + u_t = \beta_1^2 Y_{t-2} + \beta_1 u_{t-1} + u_t$. Continuing this substitution another step yields $Y_t = \beta_1^3 Y_{t-3} + \beta_1^2 u_{t-2} + \beta_1 u_{t-1} + u_t$, and continuing indefinitely yields

$$Y_t = u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \beta_1^3 u_{t-3} + \cdots = \sum_{i=0}^{\infty} \beta_1^i u_{t-i}. \tag{15.37}$$

Thus Y_t is a weighted average of current and past u_t's. Because the u_t's are normally distributed and because the weighted average of normal random variables is normal (Section 2.4), Y_{s+1} and Y_{s+2} have a bivariate normal distribution. Recall from Section 2.4 that the bivariate normal distribution is completely determined by the means of the two variables, their variances, and their covariance. Thus, to show that Y_t is stationary, we need to show that the means, variances, and covariance of (Y_{s+1}, Y_{s+2}) do not depend on s. An extension of the argument used below can be used to show that the distribution of $(Y_{s+1}, Y_{s+2}, \ldots, Y_{s+T})$ does not depend on s.

The means and variances of Y_{s+1} and Y_{s+2} can be computed using Equation (15.37), with the subscript $s + 1$ or $s + 2$ replacing t. First, because $E(u_t) = 0$ for all t, $E(Y_t) = E(\sum_{i=0}^{\infty} \beta_1^i u_{t-i}) = \sum_{i=0}^{\infty} \beta_1^i E(u_{t-i}) = 0$, so the means of Y_{s+1} and Y_{s+2} are both 0 and in particular do not depend on s. Second, $\text{var}(Y_t) = \text{var}(\sum_{i=0}^{\infty} \beta_1^i u_{t-i}) = \sum_{i=0}^{\infty}(\beta_1^i)^2 \text{var}(u_{t-i}) = \sigma_u^2 \sum_{i=0}^{\infty}(\beta_1^2)^i = \sigma_u^2/(1 - \beta_1^2)$, where the final equality follows from the fact that if $|a| < 1, \sum_{i=0}^{\infty} a^i = 1/(1 - a)$; thus $\text{var}(Y_{s+1}) = \text{var}(Y_{s+2}) = \sigma_u^2/(1 - \beta_1^2)$. Finally, because $Y_{s+2} = \beta_1 Y_{s+1} + u_{s+2}$, $\text{cov}(Y_{s+1}, Y_{s+2}) = E(Y_{s+1}, Y_{s+2}) = E[Y_{s+1}(\beta_1 Y_{s+1} + u_{s+2})] = \beta_1 \text{var}(Y_{s+1}) + \text{cov}(Y_{s+1}, u_{s+2}) = \beta_1 \text{var}(Y_{s+1}) = \beta_1 \sigma_u^2/(1 - \beta_1^2)$.

The covariance does not depend on s, so Y_{s+1} and Y_{s+2} have a joint probability distribution that does not depend on s; that is, their joint distribution is stationary. If $|\beta_1| \geq 1$, this calculation breaks down because the infinite sum in Equation (15.37) does not converge, and the variance of Y_t is infinite. Thus Y_t is stationary if $|\beta_1| < 1$ but not if $|\beta_1| \geq 1$.

The preceding argument was made under the assumptions that $\beta_0 = 0$ and u_t is normally distributed. If $\beta_0 \neq 0$, the argument is similar except that the means of Y_{s+1} and Y_{s+2} are $\beta_0/(1 - \beta_1)$ and Equation (15.37) must be modified for this nonzero mean. The assumption that u_t is i.i.d. normal can be replaced with the assumption that u_t is stationary with a finite variance because, by Equation (15.37), Y_t can still be expressed as a function of current and past u_t's, so the distribution of Y_t is stationary as long as the distribution of u_t is stationary and the infinite sum expression in Equation (15.37) is meaningful in the sense that it converges, which requires that $|\beta_1| < 1$.

15.3 Lag Operator Notation

The notation in this and the next two chapters is streamlined considerably by adopting what is known as lag operator notation. Let L denote the **lag operator**, which has the property that it transforms a variable into its lag. That is, the lag operator L has the property $LY_t = Y_{t-1}$. By applying the lag operator twice, one obtains the second lag: $L^2Y_t = L(LY_t) = LY_{t-1} = Y_{t-2}$. More generally, by applying the lag operator j times, one obtains the jth lag. In summary, the lag operator has the property that

$$LY_t = Y_{t-1}, L^2Y_t = Y_{t-2}, \text{ and } L^jY_t = Y_{t-j}. \tag{15.38}$$

The lag operator notation permits us to define the **lag polynomial**, which is a polynomial in the lag operator:

$$a(L) = a_0 + a_1L + a_2L^2 + \cdots + a_pL^p = \sum_{j=0}^{p} a_jL^j, \tag{15.39}$$

where $a_0, \ldots, a_p$ are the coefficients of the lag polynomial and $L^0 = 1$. The degree of the lag polynomial $a(L)$ in Equation (15.39) is p. Multiplying Y_t by $a(L)$ yields

$$a(L)Y_t = \left(\sum_{j=0}^{p} a_jL^j\right)Y_t = \sum_{j=0}^{p} a_j(L^jY_t) = \sum_{j=0}^{p} a_jY_{t-j} = a_0Y_t + a_1Y_{t-1} + \cdots + a_pY_{t-p}. \tag{15.40}$$

The expression in Equation (15.40) implies that the AR(p) model in Equation (15.12) can be written compactly as

$$a(L)Y_t = \beta_0 + u_t, \tag{15.41}$$

where $a_0 = 1$ and $a_j = -\beta_j$, for $j = 1, \ldots, p$. Similarly, an ADL(p, q) model can be written

$$a(L)Y_t = \beta_0 + c(L)X_{t-1} + u_t, \tag{15.42}$$

where $a(L)$ is a lag polynomial of degree p with $a_0 = 1$ and $c(L)$ is a lag polynomial of degree $q - 1$.

15.4 ARMA Models

The **autoregressive–moving average (ARMA) model** extends the autoregressive model by modeling u_t as serially correlated—specifically, as being a distributed lag (or moving average) of another unobserved error term. In the lag operator notation of Appendix 15.3, let $u_t = b(L)e_t$, where $b(L)$ is a lag polynomial of degree q with $b_0 = 1$ and e_t is a serially uncorrelated, unobserved random variable. Then the ARMA(p, q) model is

$$a(L)Y_t = \beta_0 + b(L)e_t, \tag{15.43}$$

where $a(L)$ is a lag polynomial of degree p with $a_0 = 1$.

Both the AR and ARMA models can be thought of as ways to approximate the autocovariances of Y_t. The reason for this is that any stationary time series Y_t with a finite variance can be written either as an AR or as a MA with a serially uncorrelated error term, although the AR or MA model might need to have an infinite order. The second of these results, that a stationary process can be written in moving average form, is known as the Wold decomposition theorem and is one of the fundamental results underlying the theory of stationary time series analysis.

The families of AR, MA, and ARMA models are equally rich as long as the lag polynomials have a sufficiently high degree. In some cases, the autocovariances can be better approximated by an ARMA(p, q) model with small p and q than by a pure AR model with only a few lags. That said, ARMA models are more difficult to extend to additional regressors than are AR models.

15.5 Consistency of the BIC Lag Length Estimator

This appendix summarizes the argument that the BIC estimator of the lag length, $\hat{p}$, in an autoregression is correct in large samples; that is, $\Pr(\hat{p} = p) \rightarrow 1$. This is not true for the AIC estimator, which can overestimate p even in large samples.

BIC

First consider the special case in which the BIC is used to choose among autoregressions with zero, one, or two lags, when the true lag length is one. It is shown below that (i) $\Pr(\hat{p} = 0) \rightarrow 0$ and (ii) $\Pr(\hat{p} = 2) \rightarrow 0$, from which it follows that $\Pr(\hat{p} = 1) \rightarrow 1$. The extension of this argument to the general case of searching over $0 \leq p \leq p_{max}$ entails showing that $\Pr(\hat{p} < p) \rightarrow 0$ and $\Pr(\hat{p} > p) \rightarrow 0$; the strategy for showing these is the same as used in (i) and (ii) below.

Proof of (i) and (ii)

Proof of (i). To choose $\hat{p} = 0$, it must be the case that $BIC(0) < BIC(1)$; that is, $BIC(0) - BIC(1) < 0$. Now $BIC(0) - BIC(1) = [\ln(SSR(0)/T) + (\ln T)/T] - [\ln(SSR(1)/T) + 2(\ln T)/T] = \ln(SSR(0)/T) - \ln(SSR(1)/T) - (\ln T)/T$. Now $SSR(0)/T = [(T-1)/T]s_Y^2 \xrightarrow{p} \sigma_Y^2$, $SSR(1)/T \xrightarrow{p} \sigma_u^2$, and $(\ln T)/T \longrightarrow 0$; putting these pieces together, $BIC(0) - BIC(1) \xrightarrow{p} \ln \sigma_Y^2 - \ln \sigma_u^2 > 0$ because $\sigma_Y^2 > \sigma_u^2$. It follows that $\Pr[BIC(0) < BIC(1)] \rightarrow 0$, so $\Pr(\hat{p} = 0) \longrightarrow 0$.

Proof of (ii). To choose $\hat{p} = 2$, it must be the case that $BIC(2) < BIC(1)$ or $BIC(2) - BIC(1) < 0$. Now $T[BIC(2) - BIC(1)] = T\{[\ln(SSR(2)/T) + 3(\ln T)/T] - [\ln(SSR(1)/T) + 2(\ln T)/T]\} = T\ln[SSR(2)/SSR(1)] + \ln T = -T\ln[1 + F/(T-2)] + \ln T$, where $F = [SSR(1) - SSR(2)]/[SSR(2)/(T-2)]$ is the homoskedasticity-only F-statistic [Equation (7.13)] testing the null hypothesis that $\beta_2 = 0$ in the AR(2). If u_t is homoskedastic, then F has a χ_1^2 asymptotic distribution; if not, it has some other asymptotic distribution. Thus $\Pr[BIC(2) - BIC(1) < 0] = \Pr\{T[BIC(2) - BIC(1)] < 0\} = \Pr\{-T\ln[1 + F/(T-2)] + (\ln T) < 0\} = \Pr\{T\ln[1 + F/(T-2)] > \ln T\}$. As T increases, $T\ln[1 + F/(T-2)] - F \xrightarrow{p} 0$ [a consequence of the logarithmic approximation $\ln(1 + a) \cong a$, which becomes exact as $a \longrightarrow 0$]. Thus $\Pr[BIC(2) - BIC(1) < 0] \longrightarrow \Pr(F > \ln T) \longrightarrow 0$, so $\Pr(\hat{p} = 2) \longrightarrow 0$.

AIC

In the special case of an AR(1) when zero, one, or two lags are considered, the proof of (i) for the BIC applies to the AIC where the term $\ln T$ is replaced by 2, so $\Pr(\hat{p} = 0) \longrightarrow 0$. All the steps in the proof of (ii) for the BIC also apply to the AIC, with the modification that $\ln T$ is replaced by 2; thus $\Pr[AIC(2) - AIC(1) < 0] \longrightarrow \Pr(F > 2) > 0$. If u_t is homoskedastic, then $\Pr(F > 2) \longrightarrow \Pr(\chi_1^2 > 2) = 0.16$, so $\Pr(\hat{p} = 2) \longrightarrow 0.16$. In general, when $\hat{p}$ is chosen using the AIC, $\Pr(\hat{p} < p) \longrightarrow 0$, but $\Pr(\hat{p} > p)$ tends to a positive number, so $\Pr(\hat{p} = p)$ does not tend to 1.

Estimation of Dynamic Causal Effects

In the 1983 movie *Trading Places*, the characters played by Dan Aykroyd and Eddie Murphy used inside information on how well Florida oranges had fared over the winter to make millions in the orange juice concentrate futures market, a market for contracts to buy or sell large quantities of orange juice concentrate at a specified price on a future date. In real life, traders in orange juice futures, in fact, do pay close attention to the weather in Florida: Freezes in Florida kill Florida oranges, the source of almost all frozen orange juice concentrate made in the United States, so its supply falls and the price rises. But precisely how much does the price rise when the weather in Florida turns sour? Does the price rise all at once, or are there delays; if so, for how long? These are questions that real-life traders in orange juice futures need to answer if they want to succeed.

This chapter takes up the problem of estimating the effect on Y now and in the future of a change in X—that is, the **dynamic causal effect** on Y of a change in X. What, for example, is the effect on the path of orange juice prices over time of a freezing spell in Florida? The starting point for modeling and estimating dynamic causal effects is the so-called distributed lag regression model, in which Y_t is expressed as a function of current and past values of X_t. Section 16.1 introduces the distributed lag model in the context of estimating the effect of cold weather in Florida on the price of orange juice concentrate over time. Section 16.2 takes a closer look at what, precisely, is meant by a dynamic causal effect.

One way to estimate dynamic causal effects is to estimate the coefficients of the distributed lag regression model using ordinary least squares (OLS). As discussed in Section 16.3, this estimator is consistent if the regression error has a conditional mean of 0 given current and past values of X, a condition that is referred to as exogeneity (as in Chapter 12). Because the omitted determinants of Y_t are correlated over time—that is, because they are serially correlated—the error term in the distributed lag model can be serially correlated. This possibility in turn requires heteroskedasticity- and autocorrelation-consistent (HAC) standard errors, the topic of Section 16.4.

A second way to estimate dynamic causal effects, discussed in Section 16.5, is to model the serial correlation in the error term as an autoregression and then to use this autoregressive model to derive an autoregressive distributed lag (ADL) model. Alternatively, the coefficients of the original distributed lag model can be estimated by generalized least squares (GLS). Both the ADL and the GLS methods, however, require a stronger version of exogeneity than we have used so far: *strict* exogeneity, under which the regression errors have a conditional mean of 0 given past, present, *and future* values of X.

Section 16.6 provides a more complete analysis of the relationship between orange juice prices and the weather. In this application, the weather is exogenous

(although, as discussed in Section 16.6, economic theory suggests that it is not necessarily strictly exogenous). Because exogeneity is necessary for estimating dynamic causal effects, Section 16.7 examines this assumption in several applications taken from macroeconomics and finance.

This chapter builds on the material in Sections 15.1 through 15.4 but, with the exception of a subsection (that can be skipped) of the empirical analysis in Section 16.6, does not require the material in Sections 15.5 through 15.7.

16.1 An Initial Taste of the Orange Juice Data

Orlando, the historical center of Florida's orange-growing region, is normally sunny and warm. But now and then there is a cold snap, and if temperatures drop below freezing for too long, the trees drop many of their oranges. If the cold snap is severe, the trees freeze. Following a freeze, the supply of orange juice concentrate falls, and its price rises. The timing of the price increases is rather complicated, however. Orange juice concentrate is a "durable," or storable, commodity; that is, it can be stored in its frozen state, albeit at some cost (to run the freezer). Thus the price of orange juice concentrate depends not only on current supply but also on expectations of future supply. A freeze today means that future supplies of concentrate will be low, but because concentrate currently in storage can be used to meet either current or future demand, the price of existing concentrate rises today. But precisely how much does the price of concentrate rise when there is a freeze? The answer to this question is of interest not just to orange juice traders but more generally to economists interested in studying the operations of commodity markets. To learn how the price of orange juice changes in response to weather conditions, we must analyze data on orange juice prices and the weather.

Monthly data on the price of frozen orange juice concentrate, its monthly percentage change, and temperatures in the orange-growing region of Florida from January 1950 to December 2000 are plotted in Figure 16.1. The price, plotted in Figure 16.1a, is a measure of the average real price of frozen orange juice concentrate paid by wholesalers. This price was deflated by the overall producer price index for finished goods to eliminate the effects of overall price inflation. The percentage price change plotted in Figure 16.1b is the percentage change in the price over the month. The temperature data plotted in Figure 16.1c are the number of *freezing degree days* at the Orlando, Florida, airport, calculated as the sum of the number of degrees Fahrenheit that the minimum temperature falls below freezing in a given day over all days in the month; for example, in November 1950 the airport temperature dropped below freezing twice, on the 25th ($31°F$) and on the 29th ($29°F$), for a total of 4 freezing degree days $[(32 - 31) + (32 - 29) = 4]$. (The data are described in more detail in Appendix 16.1.) As you can see by comparing the panels in Figure 16.1, the price of orange juice concentrate has large swings, some of which appear to be associated with cold weather in Florida.

FIGURE 16.1 **Orange Juice Prices and Florida Weather, 1950–2000**

(a) Price index for frozen concentrated orange juice

(b) Percentage change in the price of frozen concentrated orange juice

(c) Monthly freezing degree days in Orlando, Florida
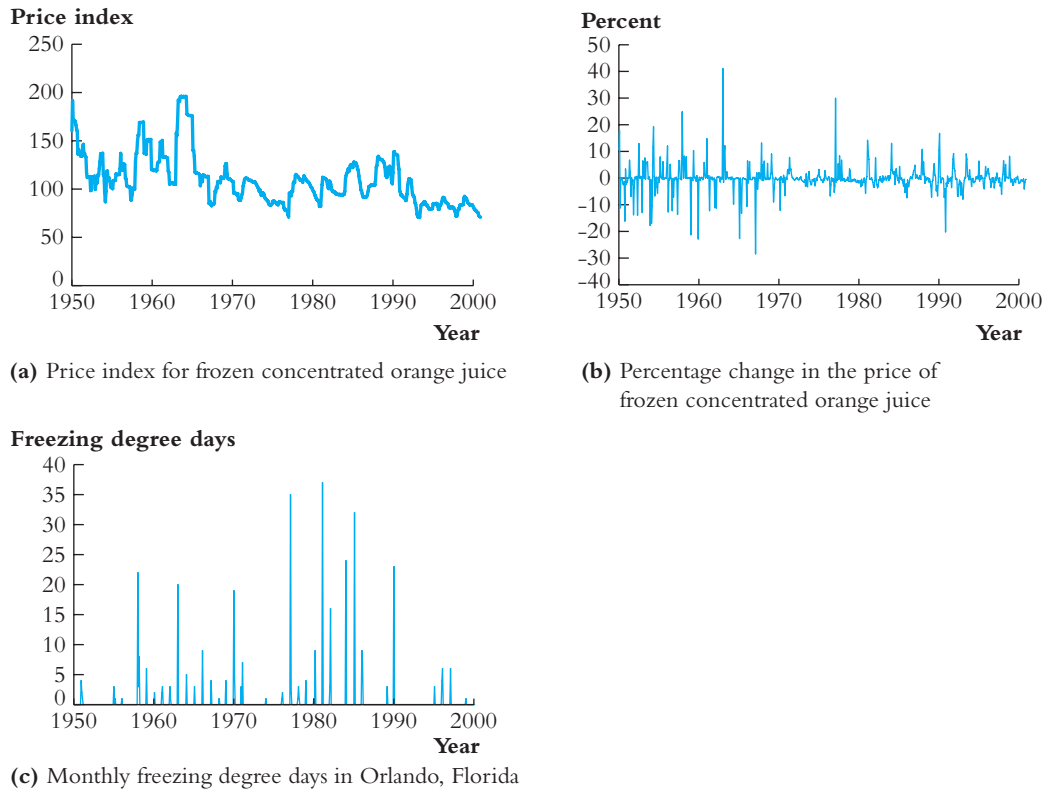
There have been large month-to-month changes in the price of frozen concentrated orange juice. Many of the large movements coincide with freezing weather in Orlando, home of many orange groves.

We begin our quantitative analysis of the relationship between orange juice price and the weather by using a regression to estimate the amount by which orange juice prices rise when the weather turns cold. The dependent variable is the percentage change in the price over that month [$\%Chg\,P_t$, where $\%Chg\,P_t = 100 \times \Delta \ln(P_t^{OJ})$ and P_t^{OJ} is the real price of orange juice]. The regressor is the number of freezing degree days during that month (FDD_t). This regression is estimated using monthly data from January 1950 to December 2000 (as are all regressions in this chapter), for a total of $T = 612$ observations:

$$\widehat{\%Chg\,P_t} = -0.40 + 0.47\,FDD_t.$$
$$\qquad\qquad (0.22)\ \ (0.13) \qquad\qquad\qquad (16.1)$$

The standard errors reported in this section are not the usual OLS standard errors but rather are HAC standard errors that are appropriate when the error term and regressors are autocorrelated. HAC standard errors are discussed in Section 16.4, and for now, they are used without further explanation.

According to this regression, an additional freezing degree day during a month increases the price of orange juice concentrate over that month by 0.47%. In a month with 4 freezing degree days, such as November 1950, the price of orange juice concentrate is estimated to have increased by 1.88% ($4 \times 0.47\% = 1.88\%$) relative to a month with no days below freezing.

Because the regression in Equation (16.1) includes only a contemporaneous measure of the weather, it does not capture any lingering effects of the cold snap on the orange juice price over the coming months. To capture these we need to consider the effect on prices of both contemporaneous and lagged values of FDD, which in turn can be done by augmenting the regression in Equation (16.1) with, for example, lagged values of FDD over the previous six months:

$$\widehat{\%ChgP_t} = -0.65 + 0.47\,FDD_t + 0.14\,FDD_{t-1} + 0.06\,FDD_{t-2}$$
$$\phantom{\widehat{\%ChgP_t} =} (0.23)\quad (0.14)\qquad\quad (0.08)\qquad\qquad (0.06)$$

$$+\ 0.07\ FDD_{t-3} + 0.03\,FDD_{t-4} + 0.05\,FDD_{t-5} + 0.05\,FDD_{t-6}. \quad (16.2)$$
$$(0.05)\qquad\qquad (0.03)\qquad\qquad (0.03)\qquad\qquad (0.04)$$

Equation (16.2) is a distributed lag regression. The coefficient on FDD_t in Equation (16.2) estimates the percentage increase in prices over the course of the month in which the freeze occurs; an additional freezing degree day is estimated to increase prices that month by 0.47%. The coefficient on the first lag of FDD_t, FDD_{t-1}, estimates the percentage increase in prices arising from a freezing degree day in the preceding month, the coefficient on the second lag estimates the effect of a freezing degree day two months ago, and so forth. Equivalently, the coefficient on the first lag of FDD estimates the effect of a unit increase in FDD one month after the freeze occurs. Thus the estimated coefficients in Equation (16.2) are estimates of the effect of a unit increase in FDD_t on current and future values of $\%ChgP$; that is, they are estimates of the dynamic effect of FDD_t on $\%ChgP_t$. For example, the 4 freezing degree days in November 1950 are estimated to have increased orange juice prices by 1.88% during November 1950, by an additional 0.56% ($= 4 \times 0.14$) in December 1950, by an additional 0.24% ($= 4 \times 0.06$) in January 1951, and so forth.

16.2 Dynamic Causal Effects

Before learning more about the tools for estimating dynamic causal effects, we should spend a moment thinking about what, precisely, is meant by a dynamic causal effect. Having a clear idea about what a dynamic causal effect is leads to a clearer understanding of the conditions under which it can be estimated.

Causal Effects and Time Series Data

Section 1.2 defined a causal effect as the outcome of an ideal randomized controlled experiment: When a horticulturalist randomly applies fertilizer to some tomato plots

but not others and then measures the yield, the expected difference in yield between the fertilized and unfertilized plots is the causal effect on tomato yield of the fertilizer. This concept of an experiment, however, is one in which there are multiple subjects (multiple tomato plots or multiple people), so the data are either cross-sectional (the tomato yield at the end of the harvest) or panel data (individual incomes before and after an experimental job training program). By having multiple subjects, it is possible to have both treatment and control groups and thereby to estimate the causal effect of the treatment.

In time series applications, this definition of causal effects in terms of an ideal randomized controlled experiment needs to be modified. To be concrete, consider an important problem of macroeconomics: estimating the effect of the central bank making an unanticipated change in the short-term interest rate on the current and future economic activity in a given country, as measured by gross domestic product (GDP). Taken literally, the randomized controlled experiment of Section 1.2 would entail randomly assigning different economies to treatment and control groups. The central banks in the treatment group would apply the treatment of a random interest rate change, while those in the control group would apply no such random changes; for both groups, economic activity (for example, GDP) would be measured over the next few years. But what if we are interested in estimating this effect for a specific country—say, the United States? Then this experiment would entail having different "clones" of the United States as subjects and assigning some clone economies to the treatment group and some to the control group. Obviously, this "parallel universes" experiment is infeasible.

Instead, in time series data it is useful to think of a randomized controlled experiment as consisting of the same subject (e.g., the U.S. economy) being given different treatments (randomly chosen changes in interest rates) at different points in time (the 1970s, the 1980s, and so forth). In this framework, the single subject at different times plays the role of both treatment and control group: Sometimes the Fed changes the interest rate, while at other times it does not. Because data are collected over time, it is possible to estimate the dynamic causal effect—that is, the time path of the effect on the outcome of interest of the treatment. For example, a surprise increase in the short-term interest rate of 2 percentage points, sustained for one quarter, might initially have a negligible effect on output; after two quarters, GDP growth might slow, with the greatest slowdown after six quarters; then over the next 2 years, GDP growth might return to normal. This time path of causal effects is the dynamic causal effect on GDP growth of a surprise change in the interest rate.

As a second example, consider the causal effect on orange juice price changes of a freezing degree day. It is possible to imagine a variety of hypothetical experiments, each yielding a different causal effect. One experiment would be to change the weather in the Florida orange groves, holding weather constant elsewhere—for example, holding weather constant in the Texas grapefruit groves and in other citrus fruit regions. This experiment would measure a partial effect, holding other weather constant. A second experiment might change the weather in all the regions, where the "treatment" is application of overall weather patterns. If weather is correlated across regions for competing

crops, then these two dynamic causal effects differ. In this chapter, we consider the causal effect in the latter experiment—that is, the causal effect of applying general weather patterns. This corresponds to measuring the dynamic effect on prices of a change in Florida weather, *not* holding weather constant in other agricultural regions.

Dynamic effects and the distributed lag model.   Because dynamic effects necessarily occur over time, the econometric model used to estimate dynamic causal effects needs to incorporate lags. To do so, Y_t can be expressed as a distributed lag of current and r past values of X_t:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \cdots + \beta_{r+1} X_{t-r} + u_t, \qquad (16.3)$$

where u_t is an error term that includes the measurement error in Y_t and the effect of omitted determinants of Y_t. The model in Equation (16.3) is called the **distributed lag model** relating X_t, and r of its lags, to Y_t.

As an illustration of Equation (16.3), consider a modified version of the tomato/fertilizer experiment: Because fertilizer applied today might remain in the ground in future years, the horticulturalist wants to determine the effect on tomato yield *over time* of applying fertilizer. Accordingly, she designs a three-year experiment and randomly divides her plots into four groups: The first is fertilized in only the first year; the second is fertilized in only the second year; the third is fertilized in only the third year; and the fourth, the control group, is never fertilized. Tomatoes are grown annually in each plot, and the third-year harvest is weighed. The three treatment groups are denoted by the binary variables X_{t-2}, X_{t-1}, and X_t, where t represents the third year (the year in which the harvest is weighed), $X_{t-2} = 1$ if the plot is in the first group (fertilized two years earlier), $X_{t-1} = 1$ if the plot was fertilized one year earlier, and $X_t = 1$ if the plot was fertilized in the final year. In the context of Equation (16.3) (which applies to a single plot), the effect of being fertilized in the final year is β_1, the effect of being fertilized one year earlier is β_2, and the effect of being fertilized two years earlier is β_3. If the effect of fertilizer is greatest in the year it is applied, then β_1 will be larger than β_2 and β_3.

More generally, the coefficient on the contemporaneous value of X_t, β_1, is the contemporaneous or immediate effect of a unit change in X_t on Y_t. The coefficient on X_{t-1}, β_2, is the effect on Y_t of a unit change in X_{t-1} or, equivalently, the effect on Y_{t+1} of a unit change in X_t; that is, β_2 is the effect of a unit change in X on Y one period later. In general, the coefficient on X_{t-h} is the effect of a unit change in X on Y after h periods. The dynamic causal effect is the effect of a change in X_t on Y_t, Y_{t+1}, Y_{t+2}, and so forth; that is, it is the sequence of causal effects on current and future values of Y. Thus, in the context of the distributed lag model in Equation (16.3), the dynamic causal effect is the sequence of coefficients $\beta_1, \beta_2, \ldots, \beta_{r+1}$.

Implications for empirical time series analysis.   This formulation of dynamic causal effects in time series data as the expected outcome of an experiment in which different treatment levels are repeatedly applied to the same subject has two implications for empirical attempts to measure the dynamic causal effect with observational time

series data. The first implication is that the dynamic causal effect should not change over the sample on which we have data. This in turn is implied by the data being jointly stationary (Key Concept 15.3). As discussed in Section 15.7, the hypothesis that a population regression function is stable over time can be tested using the Quandt likelihood ratio (QLR) test for a break, and it is possible to estimate the dynamic causal effect in different subsamples. The second implication is that X must be uncorrelated with the error term, and it is to this implication that we now turn.

Two Types of Exogeneity

Section 12.1 defined an *exogenous* variable as a variable that is uncorrelated with the regression error term and an *endogenous* variable as a variable that is correlated with the error term. This terminology traces to models with multiple equations, in which an endogenous variable is determined within the model, while an exogenous variable is determined outside the model. Loosely speaking, if we are to estimate dynamic causal effects using the distributed lag model in Equation (16.3), the regressors (the X's) must be uncorrelated with the error term. Thus X must be exogenous. Because we are working with time series data, however, we need to refine the definitions of exogeneity. In fact, there are two different concepts of exogeneity that we use here.

The first concept of exogeneity is that the error term has a conditional mean of 0 given current and all past values of X_t—that is, that $E(u_t \mid X_t, X_{t-1}, X_{t-2}, \dots) = 0$. This modifies the standard conditional mean assumption for multiple regression with cross-sectional data (assumption 1 in Key Concept 6.4), which requires only that u_t have a conditional mean of 0 given the included regressors—that is, $E(u_t \mid X_t, X_{t-1}, \dots, X_{t-r}) = 0$. Including all lagged values of X_t in the conditional expectation implies that all the more distant causal effects—all the causal effects beyond lag r—are 0. Thus, under this assumption, the r distributed lag coefficients in Equation (16.3) constitute all the nonzero dynamic causal effects. We can refer to this assumption—that $E(u_t \mid X_t, X_{t-1}, \dots) = 0$—as *past and present exogeneity*, but because of the similarity of this definition and the definition of exogeneity in Chapter 12 , we just use the term **exogeneity**.

The second concept of exogeneity is that the error term has mean 0 given all past, present, and *future* values of X_t—that is, that $E(u_t \mid \dots, X_{t+2}, X_{t+1}, X_t, X_{t-1}, X_{t-2}, \dots) = 0$. This is called **strict exogeneity**; for clarity, we also call it *past, present, and future exogeneity*. The reason for introducing the concept of strict exogeneity is that, when X is strictly exogenous, there are more efficient estimators of dynamic causal effects than the OLS estimators of the coefficients of the distributed lag regression in Equation (16.3).

The difference between exogeneity (past and present) and strict exogeneity (past, present, and future) is that strict exogeneity includes future values of X in the conditional expectation. Thus strict exogeneity implies exogeneity but not the reverse. One way to understand the difference between the two concepts is to consider the implications of these definitions for correlations between X and u. If X is (past and present) exogenous, then u_t is uncorrelated with current and past values of X_t.

The Distributed Lag Model and Exogeneity

In the distributed lag model

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \cdots + \beta_{r+1} X_{t-r} + u_t, \quad (16.4)$$

there are two different types of exogeneity—that is, two different exogeneity conditions:

- Past and present exogeneity (exogeneity):

$$E(u_t \mid X_t, X_{t-1}, X_{t-2}, \ldots) = 0; \text{ and} \quad (16.5)$$

- Past, present, and future exogeneity (strict exogeneity):

$$E(u_t \mid \ldots, X_{t+2}, X_{t+1}, X_t, X_{t-1}, X_{t-2}, \ldots) = 0. \quad (16.6)$$

If X is strictly exogenous, it is exogenous, but exogeneity does not imply strict exogeneity.

If X is strictly exogenous, then in addition u_t is uncorrelated with *future* values of X_t. For example, if a change in Y_t causes *future* values of X_t to change, then X_t is not strictly exogenous even though it might be (past and present) exogenous.

As an illustration, consider the hypothetical multiyear tomato/fertilizer experiment described following Equation (16.3). Because the fertilizer is randomly applied in the hypothetical experiment, it is exogenous. Because tomato yield today does not depend on the amount of fertilizer applied in the future, the fertilizer time series is also strictly exogenous.

As a second illustration, consider the orange juice price example, in which Y_t is the monthly percentage change in orange juice prices and X_t is the number of freezing degree days in that month. From the perspective of orange juice markets, we can think of the weather—the number of freezing degree days—as if it were randomly assigned in the sense that the weather is outside human control. If the effect of *FDD* is linear and if it has no effect on prices after r months, then it follows that the weather is exogenous. But is the weather *strictly* exogenous? If the conditional mean of u_t given future *FDD* is nonzero, then *FDD* is not strictly exogenous. Answering this question requires thinking carefully about what, precisely, is contained in u_t. In particular, if orange juice market participants use forecasts of *FDD* when they decide how much they will buy or sell at a given price, then orange juice prices, and thus the error term u_t, could incorporate information about future *FDD* that would make u_t a useful predictor of *FDD*. This means that u_t will be correlated with future values of FDD_t. According to this logic, because u_t includes forecasts of future Florida weather, *FDD* would be (past and present) exogenous but not *strictly* exogenous. The difference between this and the tomato/fertilizer example is that, while tomato plants are unaffected by future fertilization, orange juice market participants *are* influenced by forecasts of future Florida weather. We return to the question of whether *FDD* is strictly exogenous when we analyze the orange juice price data in more detail in Section 16.6.

The two definitions of exogeneity are summarized in Key Concept 16.1.

16.3 Estimation of Dynamic Causal Effects with Exogenous Regressors

If X is exogenous, then its dynamic causal effect on Y can be estimated by OLS estimation of the distributed lag regression in Equation (16.4). This section summarizes the conditions under which these OLS estimators lead to valid statistical inferences and introduces dynamic multipliers and cumulative dynamic multipliers.

The Distributed Lag Model Assumptions

The four assumptions of the distributed lag regression model are similar to the four assumptions for the cross-sectional multiple regression model (Key Concept 6.4), but they have been modified for time series data.

The first assumption is that X is exogenous, which extends the 0 conditional mean assumption for cross-sectional data to include all lagged values of X. As discussed in Section 16.2, this assumption implies that the r distributed lag coefficients in Equation (16.3) constitute all the nonzero dynamic causal effects. In this sense, the population regression function summarizes the entire dynamic effect on Y of a change in X.

The second assumption has two parts: Part (a) requires that the variables have a stationary distribution, and part (b) requires that they become independently distributed when the amount of time separating them becomes large. This assumption is the same as the corresponding assumption for the ADL model (the second assumption in Key Concept 15.6), and the discussion of that assumption in Section 15.4 applies here as well.

The third assumption is that large outliers are unlikely, made mathematically precise by assuming that the variables have more than eight nonzero finite moments. This is stronger than the assumption of four finite moments that is used elsewhere in this text. As discussed in Section 16.4, this stronger assumption is used in the mathematics behind the HAC variance estimator.

The fourth assumption, which is the same as that in the cross-sectional multiple regression model, is that there is no perfect multicollinearity.

The distributed lag regression model assumptions are summarized in Key Concept 16.2.

Extension to additional X's. The distributed lag model extends directly to multiple X's: The additional X's and their lags are simply included as regressors in the distributed lag regression, and the assumptions in Key Concept 16.2 are modified to include these additional regressors. Although the extension to multiple X's is conceptually straightforward, it complicates the notation, obscuring the main ideas of estimation and inference in the distributed lag model. For this reason, the case of multiple X's is not treated explicitly in this chapter but is left as a straightforward extension of the distributed lag model with a single X.

The Distributed Lag Model Assumptions

The distributed lag model is given in Key Concept 16.1 [Equation (16.4)], where $\beta_1, \beta_2, \ldots, \beta_{r+1}$ are dynamic causal effects and

1. X is exogenous; that is, $E(u_t \mid X_t, X_{t-1}, X_{t-2}, \ldots) = 0$;

2. (a) The random variables Y_t and X_t have a stationary distribution, and
   (b) (Y_t, X_t) and (Y_{t-j}, X_{t-j}) become independent as j gets large;

3. Large outliers are unlikely: Y_t and X_t have more than eight nonzero finite moments; and

4. There is no perfect multicollinearity.

Autocorrelated u_t, Standard Errors, and Inference

In the distributed lag regression model, the error term u_t can be autocorrelated; that is, u_t can be correlated with its lagged values. This autocorrelation arises because, in time series data, the omitted factors included in u_t can themselves be serially correlated. For example, suppose that the demand for orange juice also depends on income, so one factor that influences the price of orange juice is income—specifically, the aggregate income of potential orange juice consumers. Then aggregate income is an omitted variable in the distributed lag regression of orange juice price changes against freezing degree days. Aggregate income, however, is serially correlated: Income tends to fall in recessions and rise in expansions. Thus income is serially correlated, and because it is part of the error term, u_t will be serially correlated. This example is typical: Because omitted determinants of Y are themselves serially correlated, in general u_t in the distributed lag model will be serially correlated.

The autocorrelation of u_t does not affect the consistency of OLS, nor does it introduce bias. If, however, the errors are autocorrelated, then, in general, the usual OLS standard errors are inconsistent, and a different formula must be used. Thus serial correlation of the errors is analogous to heteroskedasticity: The homoskedasticity-only standard errors are "wrong" when the errors are, in fact, heteroskedastic in the sense that using homoskedasticity-only standard errors results in misleading statistical inferences when the errors are heteroskedastic. Similarly, when the errors are serially correlated, standard errors predicated on independently and identically distributed (i.i.d.) errors are "wrong" in the sense that they result in misleading statistical inferences. The solution to this problem is to use HAC standard errors, the topic of Section 16.4.

Dynamic Multipliers and Cumulative Dynamic Multipliers

Another name for the dynamic causal effect is the dynamic multiplier. The cumulative dynamic multipliers are the cumulative causal effects, up to a given lag; thus the cumulative dynamic multipliers measure the cumulative effect on Y of a change in X.

Dynamic multipliers. The effect of a unit change in X on Y after h periods, which is β_{h+1} in Equation (16.4), is called the h-period **dynamic multiplier**. Thus the dynamic multipliers relating X to Y are the coefficients on X_t and its lags in Equation (16.4). For example, β_2 is the one-period dynamic multiplier, β_3 is the two-period dynamic multiplier, and so forth. In this terminology, the zero-period (or contemporaneous) dynamic multiplier, or **impact effect**, is β_1, the effect on Y of a change in X in the same period.

Because the dynamic multipliers are estimated by the OLS regression coefficients, their standard errors are the HAC standard errors of the OLS regression coefficients.

Cumulative dynamic multipliers. The h-period **cumulative dynamic multiplier** is the cumulative effect of a unit change in X on Y over the next h periods. Thus the cumulative dynamic multipliers are the cumulative sum of the dynamic multipliers. In terms of the coefficients of the distributed lag regression in Equation (16.4), the zero-period cumulative multiplier is β_1, the one-period cumulative multiplier is $\beta_1 + \beta_2$, and the h-period cumulative dynamic multiplier is $\beta_1 + \beta_2 + \cdots + \beta_{h+1}$. The sum of all the individual dynamic multipliers, $\beta_1 + \beta_2 + \cdots + \beta_{r+1}$, is the cumulative long-run effect on Y of a change in X and is called the **long-run cumulative dynamic multiplier**.

For example, consider the regression in Equation (16.2). The immediate effect of an additional freezing degree day is that the price of orange juice concentrate rises by 0.47%. The cumulative effect of a price change over the next month is the sum of the impact effect and the dynamic effect one month ahead; thus the cumulative effect on prices is the initial increase of 0.47% plus the subsequent smaller increase of 0.14%, for a total of 0.61%. Similarly, the cumulative dynamic multiplier over two months is $0.47\% + 0.14\% + 0.06\% = 0.67\%$.

The cumulative dynamic multipliers can be estimated directly using a modification of the distributed lag regression in Equation (16.4). This modified regression is

$$Y_t = \delta_0 + \delta_1 \Delta X_t + \delta_2 \Delta X_{t-1} + \delta_3 \Delta X_{t-2} + \cdots + \delta_r \Delta X_{t-r+1} + \delta_{r+1} X_{t-r} + u_t.$$
(16.7)

The coefficients in Equation (16.7), $\delta_1, \delta_2, \ldots, \delta_{r+1}$, are, in fact, the cumulative dynamic multipliers. This can be shown by a bit of algebra (Exercise 16.5), which demonstrates that the population regressions in Equations (16.7) and (16.4) are equivalent, where $\delta_0 = \beta_0, \delta_1 = \beta_1, \delta_2 = \beta_1 + \beta_2, \delta_3 = \beta_1 + \beta_2 + \beta_3$, and so forth. The coefficient on X_{t-r}, δ_{r+1}, is the long-run cumulative dynamic multiplier; that is, $\delta_{r+1} = \beta_1 + \beta_2 + \beta_3 + \cdots + \beta_{r+1}$. Moreover, the OLS estimators of the coefficients in Equation (16.7) are the same as the corresponding cumulative sum of the OLS estimators in Equation (16.4). For example, $\hat{\delta}_2 = \hat{\beta}_1 + \hat{\beta}_2$. The main benefit of estimating the cumulative dynamic multipliers using the specification in Equation (16.7) is that, because the OLS estimators of the regression coefficients are estimators of the cumulative dynamic multipliers, the HAC standard errors of the coefficients in Equation (16.7) are the HAC standard errors of the cumulative dynamic multipliers.

16.4 Heteroskedasticity- and Autocorrelation-Consistent Standard Errors

If the error term u_t is autocorrelated, then OLS coefficient estimators are consistent, but in general the usual OLS standard errors for cross-sectional data are not. This means that conventional statistical inferences—hypothesis tests and confidence intervals—based on the usual OLS standard errors will, in general, be misleading. For example, confidence intervals constructed as the OLS estimator ± 1.96 conventional standard errors need not contain the true value in 95% of repeated samples, even if the sample size is large. This section begins with a derivation of the correct formula for the variance of the OLS estimator with autocorrelated errors and then turns to HAC standard errors.

This section covers HAC standard errors for regression with time series data. Chapter 10 introduced a type of HAC standard errors, clustered standard errors, that are appropriate for panel data. Although clustered standard errors for panel data and HAC standard errors for time series data have the same goal, the different data structures lead to different formulas. This section is self-contained, and Chapter 10 is not a prerequisite.

Distribution of the OLS Estimator with Autocorrelated Errors

To keep things simple, consider the OLS estimator $\hat{\beta}_1$ in the distributed lag regression model with no lags—that is, the linear regression model with a single regressor X_t:

$$Y_t = \beta_0 + \beta_1 X_t + u_t, \tag{16.8}$$

where the assumptions of Key Concept 16.2 are satisfied. This section shows that the variance of $\hat{\beta}_1$ can be written as the product of two terms: the expression for $\mathrm{var}(\hat{\beta}_1)$, applicable if u_t is not serially correlated, multiplied by a correction factor that arises from the autocorrelation in u_t or, more precisely, the autocorrelation in $(X_t - \mu_X)u_t$.

As shown in Appendix 4.3, the formula for the OLS estimator $\hat{\beta}_1$ in Key Concept 4.2 can be rewritten as

$$\hat{\beta}_1 = \beta_1 + \frac{\dfrac{1}{T}\displaystyle\sum_{t=1}^{T}(X_t - \overline{X})u_t}{\dfrac{1}{T}\displaystyle\sum_{t=1}^{T}(X_t - \overline{X})^2}, \tag{16.9}$$

where Equation (16.9) is Equation (4.28) with a change of notation so that i and n are replaced by t and T. Because $\overline{X} \xrightarrow{P} \mu_X$ and $\frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})^2 \xrightarrow{P} \sigma_X^2$, in large samples $\hat{\beta}_1 - \beta_1$ is approximately given by

$$\hat{\beta}_1 - \beta_1 \cong \frac{\dfrac{1}{T}\displaystyle\sum_{t=1}^{T}(X_t - \mu_X)u_t}{\sigma_X^2} = \frac{\dfrac{1}{T}\displaystyle\sum_{t=1}^{T}v_t}{\sigma_X^2} = \frac{\overline{v}}{\sigma_X^2}, \tag{16.10}$$

where $v_t = (X_t - \mu_X)u_t$ and $\bar{v} = \frac{1}{T}\sum_{t=1}^{T} v_t$. Thus

$$\text{var}(\hat{\beta}_1) = \text{var}\left(\frac{\bar{v}}{\sigma_X^2}\right) = \frac{\text{var}(\bar{v})}{(\sigma_X^2)^2}. \tag{16.11}$$

If v_t is i.i.d.—as assumed for cross-sectional data in Key Concept 4.3—then $\text{var}(\bar{v}) = \text{var}(v_t)/T$, and the formula for the variance of $\hat{\beta}_1$ from Key Concept 4.4 applies. If, however, u_t and X_t are not independently distributed over time, then, in general, v_t will be serially correlated, so $\text{var}(\bar{v}) \neq \text{var}(v_t)/T$ and Key Concept 4.4 does not apply. Instead, if v_t is serially correlated, the variance of $\bar{v}$ is given by

$$
\begin{aligned}
\text{var}(\bar{v}) &= \text{var}[(v_1 + v_2 + \cdots + v_T)/T] \\
&= [\text{var}(v_1) + \text{cov}(v_1, v_2) + \cdots + \text{cov}(v_1, v_T) \\
&\quad + \text{cov}(v_2, v_1) + \text{var}(v_2) + \cdots + \text{var}(v_T)]/T^2 \\
&= [T\text{var}(v_t) + 2(T-1)\text{cov}(v_t, v_{t-1}) \\
&\quad + 2(T-2)\text{cov}(v_t, v_{t-2}) + \cdots + 2\text{cov}(v_t, v_{t-T+1})]/T^2 \\
&= \frac{\sigma_v^2}{T}f_T,
\end{aligned}
\tag{16.12}
$$

where

$$f_T = 1 + 2\sum_{j=1}^{T-1}\left(\frac{T-j}{T}\right)\rho_j, \tag{16.13}$$

where $\rho_j = \text{corr}(v_t, v_{t-j})$. In large samples, f_T tends to the limit, $f_T \longrightarrow f_\infty = 1 + 2\sum_{j=1}^{\infty}\rho_j$.

Combining the expressions in Equation (16.10) for $\hat{\beta}_1$ and Equation (16.12) for $\text{var}(\bar{v})$ gives the formula for the variance of $\hat{\beta}_1$ when v_t is autocorrelated:

$$\text{var}(\hat{\beta}_1) = \left[\frac{1}{T}\frac{\sigma_v^2}{(\sigma_X^2)^2}\right]f_T, \tag{16.14}$$

where f_T is given in Equation (16.13).

Equation (16.14) expresses the variance of $\hat{\beta}_1$ as the product of two terms. The first, in square brackets, is the formula for the variance of $\hat{\beta}_1$ given in Key Concept 4.4, which applies in the absence of serial correlation. The second is the factor f_T, which adjusts this formula for serial correlation. Because of this additional factor f_T in Equation (16.14), the usual OLS standard error computed using Equation (5.4) is incorrect if the errors are serially correlated: If $v_t = (X_t - \mu_X)u_t$ is serially correlated, the estimator of the variance is off by the factor f_T.

HAC Standard Errors

If the factor f_T, defined in Equation (16.13), were known, then the variance of $\hat{\beta}_1$ could be estimated by multiplying the usual cross-sectional estimator of the variance by f_T. This factor, however, depends on the unknown autocorrelations of v_t, so it must

be estimated. The estimator of the variance of $\hat{\beta}_1$ that incorporates this adjustment is consistent whether or not there is heteroskedasticity and whether or not v_t is autocorrelated. Accordingly, this estimator is called the **heteroskedasticity- and autocorrelation-consistent (HAC)** estimator of the variance of $\hat{\beta}_1$, and the square root of the HAC variance estimator is the **HAC standard error** of $\hat{\beta}_1$.

The HAC variance formula. The HAC estimator of the variance of $\hat{\beta}_1$ is

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \hat{\sigma}^2_{\hat{\beta}_1} \hat{f}_T, \tag{16.15}$$

where $\hat{\sigma}^2_{\hat{\beta}_1}$ is the estimator of the variance of $\hat{\beta}_1$ in the absence of serial correlation, given in Equation (5.4), and where $\hat{f}_T$ is an estimator of the factor f_T in Equation (16.13).

The task of constructing a consistent estimator $\hat{f}_T$ is challenging. To see why, consider two extremes. At one extreme, given the formula in Equation (16.13), it might seem natural to replace the population autocorrelations ρ_j with the sample autocorrelations $\hat{\rho}_j$ [defined in Equation (15.5)], yielding the estimator $1 + 2\sum_{j=1}^{T-1}\left(\frac{T-j}{T}\right)\hat{\rho}_j$. But this estimator contains so many estimated autocorrelations that it is inconsistent. Intuitively, because each of the estimated autocorrelations contains an estimation error, by estimating so many autocorrelations the estimation error in this estimator of f_T remains large even in large samples. At the other extreme, one could imagine using only a few sample autocorrelations—for example, using only the first sample autocorrelation and ignoring all the higher autocorrelations. Although this estimator eliminates the problem of estimating too many autocorrelations, it has a different problem: It is inconsistent because it ignores the additional autocorrelations that appear in Equation (16.13). In short, using too many sample autocorrelations makes the estimator have a large variance, but using too few autocorrelations ignores the autocorrelations at higher lags, so in either of these extreme cases the estimator is inconsistent.

Estimators of f_T used in practice strike a balance between these two extreme cases by choosing the number of autocorrelations to include in a way that depends on the sample size T. If the sample size is small, only a few autocorrelations are used, but if the sample size is large, more autocorrelations are included (but still far fewer than T). Specifically, let $\hat{f}_T$ be given by

$$\hat{f}_T = 1 + 2\sum_{j=1}^{m-1}\left(\frac{m-j}{m}\right)\tilde{\rho}_j, \tag{16.16}$$

where $\tilde{\rho}_j = \sum_{t=j+1}^{T}\hat{v}_t\hat{v}_{t-j}/\sum_{t=1}^{T}\hat{v}_t^2$, where $\hat{v}_t = (X_t - \overline{X})\hat{u}_t$ (as in the definition of $\hat{\sigma}^2_{\hat{\beta}_1}$). The parameter m in Equation (16.16) is called the **truncation parameter** of the HAC estimator because the sum of autocorrelations is shortened, or truncated, to include only $m - 1$ autocorrelations instead of the $T - 1$ autocorrelations appearing in the population formula in Equation (16.13).

For $\hat{f}_T$ to be consistent, m must be chosen so that it is large in large samples, although still much less than T. One guideline for choosing m in practice is to use the formula

$$m = 0.75T^{1/3}, \tag{16.17}$$

rounded to an integer. This formula, which is based on the assumption that there is at most a moderate amount of autocorrelation in v_t, gives a benchmark rule for determining m as a function of the number of observations in the regression.[1]

The value of the truncation parameter m resulting from Equation (16.17) can be modified using your knowledge of the series at hand. On the one hand, if there is a great deal of serial correlation in v_t, then you should increase m beyond the value from Equation (16.17). On the other hand, if v_t has little serial correlation, you could decrease m. Because of the ambiguity associated with the choice of m, it is good practice to try one or two alternative values of m for at least one specification to make sure your results are not sensitive to m.

The HAC estimator in Equation (16.15), with $\hat{f}_T$ given in Equation (16.16), is called the **Newey–West variance estimator**, after the econometricians Whitney Newey and Kenneth West, who proposed it. They showed that, when used along with a rule like that in Equation (16.17), under general assumptions this estimator is a consistent estimator of the variance of $\hat{\beta}_1$ (Newey and West 1987). Their proofs (and those in Andrews 1991) assume that v_t has more than four moments, which in turn is implied by X_t and u_t having more than eight moments, and this is the reason that the third assumption in Key Concept 16.2 is that X_t and u_t have more than eight moments.

Other HAC estimators.   The Newey–West variance estimator is not the only HAC estimator. For example, the weights $(m - j)/m$ in Equation (16.16) can be replaced by different weights. If different weights are used, then the rule for choosing the truncation parameter in Equation (16.17) no longer applies, and a different rule, developed for those weights, should be used instead. Discussion of HAC estimators using other weights goes beyond the scope of this text. For more information on this topic, see Hayashi (2000, Section 6.6).

Extension to multiple regression.   All the issues discussed in this section generalize to the distributed lag regression model in Key Concept 16.1 with multiple lags and, more generally, to the multiple regression model with serially correlated errors. In particular, if the error term is serially correlated, then the usual OLS standard errors are an unreliable basis for inference, and HAC standard errors should be used instead. If the HAC variance estimator used is the Newey–West estimator [the HAC variance estimator based on the weights $(m - j)/m$], then the truncation parameter m can be

---

[1]Equation (16.17) gives the value of m that minimizes $E(\tilde{\sigma}^2_{\hat{\beta}_1} - \sigma^2_{\hat{\beta}_1})^2$ when u_t and X_t are first-order autoregressive processes with first autocorrelation coefficient 0.5. Equation (16.17) is based on a more general formula derived by Andrews [1991, Equation (5.3)].

HAC Standard Errors

The problem: The error term u_t in the distributed lag regression model in Key Concept 16.1 can be serially correlated. If so, the OLS coefficient estimators are consistent, but, in general, the usual OLS standard errors are not, resulting in misleading hypothesis tests and confidence intervals.

The solution: Standard errors should be computed using a HAC estimator of the variance. The HAC estimator involves estimates of $m - 1$ autocorrelations as well as the variance; in the case of a single regressor, the relevant formulas are given in Equations (16.15) and (16.16).

In practice, using HAC standard errors entails choosing the truncation parameter m. To do so, use the formula in Equation (16.17) as a benchmark and then increase or decrease m, depending on whether your regressors and errors have high or low serial correlation.

chosen according to the rule in Equation (16.17) whether there is a single regressor or multiple regressors. The formula for HAC standard errors in multiple regression is incorporated into modern regression software designed for use with time series data. Because this formula involves matrix algebra, we omit it here and instead refer the reader to Hayashi (2000, Section 6.6) for the mathematical details.

HAC standard errors are summarized in Key Concept 16.3.

16.5 Estimation of Dynamic Causal Effects with Strictly Exogenous Regressors

When X_t is strictly exogenous, two alternative estimators of dynamic causal effects are available. The first such estimator involves estimating an ADL model instead of a distributed lag model and calculating the dynamic multipliers from the estimated ADL coefficients. This method can entail estimating fewer coefficients than OLS estimation of the distributed lag model, thus potentially reducing estimation error. The second method is to estimate the coefficients of the distributed lag model, using generalized least squares (GLS) instead of OLS. Although GLS estimates the same number of coefficients in the distributed lag model as OLS, the GLS estimator has a smaller variance. To keep the exposition simple, these two estimation methods are laid out and discussed in the context of a distributed lag model with a single lag and AR(1) errors. Appendix 16.2 extends these estimators to the general distributed lag model with higher-order autoregressive errors.

The Distributed Lag Model with AR(1) Errors

Suppose that the causal effect on Y of a change in X lasts for only two periods; that is, it has an initial impact effect β_1 and an effect in the next period of β_2 but no effect thereafter. Then the appropriate distributed lag regression model is the distributed lag model with only current and past values of X_{t-1}:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + u_t. \tag{16.18}$$

As discussed in Section 16.2, in general the error term u_t in Equation (16.18) is serially correlated. One consequence of this serial correlation is that, if the distributed lag coefficients are estimated by OLS, then inference based on the usual OLS standard errors can be misleading. For this reason, Sections 16.3 and 16.4 emphasized the use of HAC standard errors when β_1 and β_2 in Equation (16.18) are estimated by OLS.

In this section, we take a different approach toward the serial correlation in u_t. This approach, which is possible if X_t is strictly exogenous, involves adopting an autoregressive model for the serial correlation in u_t and then using this AR model to derive estimators that can be more efficient than OLS.

Specifically, suppose that u_t follows the AR(1) model

$$u_t = \phi_1 u_{t-1} + \tilde{u}_t, \tag{16.19}$$

where ϕ_1 is the autoregressive parameter, $\tilde{u}_t$ is serially uncorrelated, and no intercept is needed because $E(u_t) = 0$. Equations (16.18) and (16.19) imply that the distributed lag model with a serially correlated error can be rewritten as an autoregressive distributed lag model with a serially uncorrelated error. To do so, lag each side of Equation (16.18), and subtract ϕ_1 multiplied by this lag from each side:

$$Y_t - \phi_1 Y_{t-1} = (\beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + u_t) - \phi_1(\beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + u_{t-1})$$
$$= \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} - \phi_1 \beta_0 - \phi_1 \beta_1 X_{t-1} - \phi_1 \beta_2 X_{t-2} + \tilde{u}_t, \tag{16.20}$$

where the second equality uses $\tilde{u}_t = u_t - \phi_1 u_{t-1}$. Collecting terms in Equation (16.20), we have that

$$Y_t = \alpha_0 + \phi_1 Y_{t-1} + \delta_0 X_t + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \tilde{u}_t, \tag{16.21}$$

where

$$\alpha_0 = \beta_0(1 - \phi_1), \delta_0 = \beta_1, \delta_1 = \beta_2 - \phi_1 \beta_1, \text{ and } \delta_2 = -\phi_1 \beta_2, \tag{16.22}$$

where $\beta_0, \beta_1,$ and β_2 are the coefficients in Equation (16.18) and ϕ_1 is the autocorrelation coefficient in Equation (16.19).

Equation (16.21) is an ADL model that includes a contemporaneous value of X and two of its lags. We will refer to Equation (16.21) as the ADL representation of the distributed lag model with autoregressive errors given in Equations (16.18) and (16.19).

The terms in Equation (16.20) can be reorganized differently to obtain an expression that is equivalent to Equations (16.21) and (16.22). Let $\widetilde{Y}_t = Y_t - \phi_1 Y_{t-1}$ be the **quasi-difference** of Y_t (*quasi* because it is not the first difference, the difference between Y_t and Y_{t-1}; rather, it is the difference between Y_t and $\phi_1 Y_{t-1}$). Similarly, let $\widetilde{X}_t = X_t - \phi_1 X_{t-1}$ be the quasi-difference of X_t. Then Equation (16.20) can be written

$$\widetilde{Y}_t = \alpha_0 + \beta_1 \widetilde{X}_t + \beta_2 \widetilde{X}_{t-1} + \widetilde{u}_t. \tag{16.23}$$

We will refer to Equation (16.23) as the quasi-difference representation of the distributed lag model with autoregressive errors given in Equations (16.18) and (16.19).

The ADL model in Equation (16.21) [with the parameter restrictions in Equation (16.22)] and the quasi-difference model in Equation (16.23) are equivalent. In both models, the error term, $\widetilde{u}_t$, is serially uncorrelated. The two representations, however, suggest different estimation strategies. But before discussing those strategies, we turn to the assumptions under which they yield consistent estimators of the dynamic multipliers, β_1 and β_2.

The conditional mean 0 assumption in the ADL and quasi-difference models. Because Equations (16.21) [with the restrictions in Equation (16.22)] and (16.23) are equivalent, the conditions for their estimation are the same, so for convenience we consider Equation (16.23).

The quasi-difference model in Equation (16.23) is a distributed lag model involving the quasi-differenced variables with a serially uncorrelated error. Accordingly, the conditions for OLS estimation of the coefficients in Equation (16.23) are the least squares assumptions for the distributed lag model in Key Concept 16.2, expressed in terms of $\widetilde{u}_t$ and $\widetilde{X}_t$. The critical assumption here is the first assumption, which, applied to Equation (16.23), is that $\widetilde{X}_t$ is exogenous; that is,

$$E(\widetilde{u}_t | \widetilde{X}_t, \widetilde{X}_{t-1}, \dots) = 0, \tag{16.24}$$

where letting the conditional expectation depend on distant lags of $\widetilde{X}_t$ ensures that no additional lags of $\widetilde{X}_t$, other than those appearing in Equation (16.23), enter the population regression function.

Because $\widetilde{X}_t = X_t - \phi_1 X_{t-1}$, so $X_t = \widetilde{X}_t + \phi_1 X_{t-1}$, conditioning on $\widetilde{X}_t$ and all of its lags is equivalent to conditioning on X_t and all of its lags. Thus the conditional expectation condition in Equation (16.24) is equivalent to the condition that $E(\widetilde{u}_t | X_t, X_{t-1}, \dots) = 0$. Furthermore, because $\widetilde{u}_t = u_t - \phi_1 u_{t-1}$, this condition in turn implies that

$$
\begin{aligned}
0 &= E(\widetilde{u}_t \,|\, X_t, X_{t-1}, \dots) \\
&= E(u_t - \phi_1 u_{t-1} \,|\, X_t, X_{t-1}, \dots) \\
&= E(u_t \,|\, X_t, X_{t-1}, \dots) - \phi_1 E(u_{t-1} \,|\, X_t, X_{t-1}, \dots).
\end{aligned}
\tag{16.25}
$$

For the equality in Equation (16.25) to hold for general values of ϕ_1, it must be the case that both $E(u_t \mid X_t, X_{t-1}, \dots) = 0$ and $E(u_{t-1} \mid X_t, X_{t-1}, \dots) = 0$. By shifting the time subscripts forward one time period, the condition that $E(u_{t-1} \mid X_t, X_{t-1}, \dots) = 0$ can be rewritten as

$$E(u_t \mid X_{t+1}, X_t, X_{t-1}, \dots) = 0, \tag{16.26}$$

which (by the law of iterated expectations) implies that $E(u_t \mid X_t, X_{t-1}, \dots) = 0$. In summary, having the 0 conditional mean assumption in Equation (16.24) hold for general values of ϕ_1 is equivalent to having the condition in Equation (16.26) hold.

The condition in Equation (16.26) is implied by X_t being strictly exogenous, but it is *not* implied by X_t being (past and present) exogenous. Thus the least squares assumptions for estimation of the distributed lag model in Equation (16.23) hold if X_t is strictly exogenous, but it is not enough that X_t be (past and present) exogenous.

Because the ADL representation [Equations (16.21) and (16.22)] is equivalent to the quasi-differenced representation [Equation (16.23)], the conditional mean assumption needed to estimate the coefficients of the quasi-differenced representation [that $E(u_t \mid X_{t+1}, X_t, X_{t-1}, \dots) = 0$] is also the conditional mean assumption for consistent estimation of the coefficients of the ADL representation.

We now turn to the two estimation strategies suggested by these two representations: estimation of the ADL coefficients and estimation of the coefficients of the quasi-difference model.

OLS Estimation of the ADL Model

The first strategy is to use OLS to estimate the coefficients in the ADL model in Equation (16.21). As the derivation leading to Equation (16.21) shows, including the lag of Y and the extra lag of X as regressors makes the error term serially uncorrelated (under the assumption that the error follows a first-order autoregression). Thus the usual OLS standard errors can be used; that is, HAC standard errors are not needed when the ADL model coefficients in Equation (16.21) are estimated by OLS.

The estimated ADL coefficients are not themselves estimates of the dynamic multipliers, but the dynamic multipliers can be computed from the ADL coefficients. A general way to compute the dynamic multipliers is to express the estimated regression function as a function of current and past values of X_t—that is, to eliminate Y_t from the estimated regression function. To do so, repeatedly substitute expressions for lagged values of Y_t into the estimated regression function. Specifically, consider the estimated regression function

$$\hat{Y}_t = \hat{\phi}_1 Y_{t-1} + \hat{\delta}_0 X_t + \hat{\delta}_1 X_{t-1} + \hat{\delta}_2 X_{t-2}, \tag{16.27}$$

where the estimated intercept has been omitted because it does not enter any expression for the dynamic multipliers. Lagging both sides of Equation (16.27) yields

$\hat{Y}_{t-1} = \hat{\phi}_1 Y_{t-2} + \hat{\delta}_0 X_{t-1} + \hat{\delta}_1 X_{t-2} + \hat{\delta}_2 X_{t-3}$, so replacing $\hat{Y}_{t-1}$ in Equation (16.27) by this expression for $\hat{Y}_{t-1}$ and collecting terms yields

$$\hat{Y}_t = \hat{\phi}_1(\hat{\phi}_1 Y_{t-2} + \hat{\delta}_0 X_{t-1} + \hat{\delta}_1 X_{t-2} + \hat{\delta}_2 X_{t-3}) + \hat{\delta}_0 X_t + \hat{\delta}_1 X_{t-1} + \hat{\delta}_2 X_{t-2}$$
$$= \hat{\delta}_0 X_t + (\hat{\delta}_1 + \hat{\phi}_1 \hat{\delta}_0) X_{t-1} + (\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1) X_{t-2} + \hat{\phi}_1 \hat{\delta}_2 X_{t-3} + \hat{\phi}_1^2 Y_{t-2}. \quad (16.28)$$

Repeating this process by repeatedly substituting expressions for Y_{t-2}, Y_{t-3}, and so forth yields

$$\hat{Y}_t = \hat{\delta}_0 X_t + (\hat{\delta}_1 + \hat{\phi}_1 \hat{\delta}_0) X_{t-1} + (\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1 + \hat{\phi}_1^2 \hat{\delta}_0) X_{t-2}$$
$$+ \hat{\phi}_1(\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1 + \hat{\phi}_1^2 \hat{\delta}_0) X_{t-3} + \hat{\phi}_1^2(\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1 + \hat{\phi}_1^2 \hat{\delta}_0) X_{t-4} + \cdots. \quad (16.29)$$

The coefficients in Equation (16.29) are the estimators of the dynamic multipliers, computed from the OLS estimators of the coefficients in the ADL model in Equation (16.21). If the restrictions on the coefficients in Equation (16.22) were to hold exactly for the *estimated* coefficients, then the dynamic multipliers beyond the second (that is, the coefficients on X_{t-2}, X_{t-3}, and so forth) would all be 0.[2] However, under this estimation strategy those restrictions will not hold exactly, so the estimated multipliers beyond the second in Equation (16.29) will generally be nonzero.

GLS Estimation

The second strategy for estimating the dynamic multipliers when X_t is strictly exogenous is to use **generalized least squares (GLS)**, which entails estimating Equation (16.23). To describe the GLS estimator, we initially assume that ϕ_1 is known. Because in practice it is unknown, this estimator is infeasible, so it is called the infeasible GLS estimator. The infeasible GLS estimator, however, can be modified using an estimator of ϕ_1, which yields a feasible version of the GLS estimator.

Infeasible GLS. If ϕ_1 is known, then the quasi-differenced variables $\widetilde{X}_t$ and $\widetilde{Y}_t$ can be computed directly. As discussed in the context of Equations (16.24) and (16.26), if X_t is strictly exogenous, then $E(\widetilde{u}_t \mid \widetilde{X}_t, \widetilde{X}_{t-1}, \dots) = 0$. Thus, if X_t is strictly exogenous and if ϕ_1 is known, the coefficients α_0, β_1, and β_2 in Equation (16.23) can be estimated by the OLS regression of $\widetilde{Y}_t$ on $\widetilde{X}_t$ and $\widetilde{X}_{t-1}$ (including an intercept). The resulting estimator of β_1 and β_2—that is, the OLS estimator of the slope coefficients in Equation (16.23) when ϕ_1 is known—is the **infeasible GLS estimator**. This estimator is infeasible because in reality ϕ_1 is unknown, so $\widetilde{X}_t$ and $\widetilde{Y}_t$ cannot be computed and thus these OLS estimators cannot actually be computed.

---

[2]Substitute the equalities in Equation (16.22) to show that, if those equalities hold, then $\delta_2 + \phi_1 \delta_1 + \phi_1^2 \delta_0 = 0$.

Feasible GLS.   The **feasible GLS estimator** modifies the infeasible GLS estimator by using a preliminary estimator of ϕ_1, $\hat{\phi}_1$, to compute the estimated quasi-differences. Specifically, the feasible GLS estimators of β_1 and β_2 are the OLS estimators of β_1 and β_2 in Equation (16.23), computed by regressing $\hat{\tilde{Y}}_t$ on $\hat{\tilde{X}}_t$ and $\hat{\tilde{X}}_{t-1}$ (with an intercept), where $\hat{\tilde{X}}_t = X_t - \hat{\phi}_1 X_{t-1}$ and $\hat{\tilde{Y}}_t = Y_t - \hat{\phi}_1 Y_{t-1}$.

The preliminary estimator, $\hat{\phi}_1$, can be computed by first estimating the distributed lag regression in Equation (16.18) by OLS and then using OLS to estimate ϕ_1 in Equation (16.19) with the OLS residuals $\hat{u}_t$ replacing the unobserved regression errors u_t. This version of the GLS estimator is called the Cochrane–Orcutt (1949) estimator.

An extension of the Cochrane–Orcutt method is to continue this process iteratively: Use the GLS estimate of β_1 and β_2 to compute revised estimates of u_t; use these new residuals to reestimate ϕ_1; use this revised estimate of ϕ_1 to compute revised estimated quasi-differences; use these revised estimated quasi-differences to reestimate β_1 and β_2; and continue this process until the estimates of β_1 and β_2 converge. This is referred to as the iterated Cochrane–Orcutt estimator.

Efficiency of GLS.   The virtue of the GLS estimator is that when X is strictly exogenous and the transformed errors $\tilde{u}_t$ are homoskedastic, it is efficient among linear estimators, at least in large samples. To see this, first consider the infeasible GLS estimator. If $\tilde{u}_t$ is homoskedastic, if ϕ_1 is known (so that $\tilde{X}_t$ and $\tilde{Y}_t$ can be treated as if they are observed), and if X_t is strictly exogenous, then the Gauss–Markov theorem implies that the OLS estimator of α_0, β_1, and β_2 in Equation (16.23) is efficient among all linear conditionally unbiased estimators based on $\tilde{X}_t$ and $\tilde{Y}_t$, for $t = 2, \ldots, T$, where the first observation ($t = 1$) is lost because of quasi-differencing. That is, the OLS estimator of the coefficients in Equation (16.23) is the best linear unbiased estimator, or BLUE (Section 5.5). Because the OLS estimator of Equation (16.23) is the infeasible GLS estimator, this means that the infeasible GLS estimator is BLUE. The feasible GLS estimator is similar to the infeasible GLS estimator except that ϕ_1 is estimated. Because the estimator of ϕ_1 is consistent and its variance is inversely proportional to T, the feasible and infeasible GLS estimators have the same variances in large samples, and the loss of information from the first observation ($t = 1$) is negligible when T is large. In this sense, if X is strictly exogenous, then the feasible GLS estimator is BLUE in large samples. In particular, if X is strictly exogenous, then GLS is more efficient than the OLS estimator of the distributed lag coefficients discussed in Section 16.3.

The Cochrane–Orcutt and iterated Cochrane–Orcutt estimators presented here are special cases of GLS estimation. In general, GLS estimation involves transforming the regression model so that the errors are homoskedastic and serially uncorrelated and then estimating the coefficients of the transformed regression model by OLS. In general, the GLS estimator is consistent and BLUE in large samples if X is strictly exogenous, but it is not consistent if X is only (past and present) exogenous. The mathematics of GLS involves matrix algebra, so it is postponed to Section 19.6.

16.6 Orange Juice Prices and Cold Weather

This section uses the tools of time series regression to squeeze additional insights from our data on Florida temperatures and orange juice prices. First, how long lasting is the effect of a freeze on the price? Second, has this dynamic effect been stable, or has it changed over the 51 years spanned by the data and, if so, how?

We begin this analysis by estimating the dynamic causal effects using the method of Section 16.3—that is, by OLS estimation of the coefficients of a distributed lag regression of the percentage change in prices ($\%ChgP_t$) on the number of freezing degree days in that month (FDD_t) and its lagged values. For the distributed lag estimator to be consistent, FDD must be (past and present) exogenous. As discussed in Section 16.2, this assumption is reasonable here. Humans cannot influence the weather, so treating the weather as if it were randomly assigned experimentally is appropriate as a working hypothesis (we return to this below). If FDD is exogenous, we can estimate the dynamic causal effects by OLS estimation of the coefficients in the distributed lag model of Equation (16.4) in Key Concept 16.1.

As discussed in Sections 16.3 and 16.4, the error term can be serially correlated in distributed lag regressions, so it is important to use HAC standard errors, which adjust for this serial correlation. For the initial results, the truncation parameter for the Newey–West standard errors (m in the notation of Section 16.4) was chosen using the rule in Equation (16.17): Because there are 612 monthly observations, according to that rule $m = 0.75 \ T^{1/3} = 0.75 \times 612^{1/3} = 6.37$, but because m must be an integer, this was rounded up to $m = 7$. The sensitivity of the standard errors to this choice of truncation parameter is investigated below.

The results of OLS estimation of the distributed lag regression of $\%ChgP_t$ on FDD_t, $FDD_{t-1}, \ldots, FDD_{t-18}$ are summarized in column (1) of Table 16.1. The coefficients of this regression (only some of which are reported in the table) are estimates of the dynamic causal effect on orange juice price changes (in percent) for the first 18 months following a unit increase in the number of freezing degree days in a month. For example, a single freezing degree day is estimated to increase prices by 0.50% over the month in which the freezing degree day occurs. The subsequent effect on price in later months of a freezing degree day is less: After one month, the estimated effect is to increase the price by a further 0.17%, and after two months, the estimated effect is to increase the price by an additional 0.07%. The R^2 from this regression is 0.12, indicating that much of the monthly variation in orange juice prices is not explained by current and past values of FDD.

Plots of dynamic multipliers can convey information more effectively than tables such as Table 16.1. The dynamic multipliers from column (1) of Table 16.1 are plotted in Figure 16.2a along with their 95% confidence intervals, computed as the estimated coefficient ± 1.96 HAC standard errors. After the initial sharp price rise, subsequent price rises are less, although prices are estimated to rise slightly in each of the first six months after the freeze. As can be seen from Figure 16.2a, for months other than the first, the dynamic multipliers are not statistically significantly different from 0 at the 5% significance level, although they are estimated to be positive through the seventh month.

TABLE 16.1 The Dynamic Effect of a Freezing Degree Day (*FDD*) on the Price of Orange Juice: Selected Estimated Dynamic Multipliers and Cumulative Dynamic Multipliers

Lag Number	(1) Dynamic Multipliers	(2) Cumulative Multipliers	(3) Cumulative Multipliers	(4) Cumulative Multipliers
0	0.50 (0.14)	0.50 (0.14)	0.50 (0.14)	0.51 (0.15)
1	0.17 (0.09)	0.67 (0.14)	0.67 (0.13)	0.70 (0.15)
2	0.07 (0.06)	0.74 (0.17)	0.74 (0.16)	0.76 (0.18)
3	0.07 (0.04)	0.81 (0.18)	0.81 (0.18)	0.84 (0.19)
4	0.02 (0.03)	0.84 (0.19)	0.84 (0.19)	0.87 (0.20)
5	0.03 (0.03)	0.87 (0.19)	0.87 (0.19)	0.89 (0.20)
6 . . .	0.03 (0.05)	0.90 (0.20)	0.90 (0.21)	0.91 (0.21)
12 . . .	−0.14 (0.08)	0.54 (0.27)	0.54 (0.28)	0.54 (0.28)
18	0.00 (0.02)	0.37 (0.30)	0.37 (0.31)	0.37 (0.30)
Monthly indicators?	No	No	No	Yes $F = 1.01$ ($p = 0.43$)
HAC standard error truncation parameter (*m*)	7	7	14	7

All regressions were estimated by OLS using monthly data (described in Appendix 16.1) from January 1950 to December 2000, for a total of $T = 612$ monthly observations. The dependent variable is the monthly percentage change in the price of orange juice ($\%ChgP_t$). Regression (1) is the distributed lag regression with the monthly number of freezing degree days and 18 of its lagged values—that is, $FDD_t, FDD_{t-1}, \ldots, FDD_{t-18}$—and the reported coefficients are the OLS estimates of the dynamic multipliers. The cumulative multipliers are the cumulative sum of the estimated dynamic multipliers. All regressions include an intercept, which is not reported. Newey–West HAC standard errors, computed using the truncation number given in the final row, are reported in parentheses.

Column (2) of Table 16.1 contains the cumulative dynamic multipliers for this specification—that is, the cumulative sum of the dynamic multipliers reported in column (1). These cumulative dynamic multipliers are plotted in Figure 16.2b along with their 95% confidence intervals. After 1 month, the cumulative effect of the freezing degree day is to increase prices by 0.67%; after 2 months, the price is estimated to have risen by 0.74%; and after 6 months, the price is estimated to have risen by 0.90%. As can be seen in Figure 16.2b, these cumulative multipliers increase through the seventh month because the individual dynamic multipliers are positive for the first 7 months. In the 8th month, the dynamic multiplier is negative, so the price of orange juice begins to fall slowly from its peak. After 18 months, the cumulative increase in prices is only 0.37%; that is, the long-run cumulative dynamic multiplier is

The Dynamic Effect of a Freezing Degree Day (*FDD*) on the Price of Orange Juice

The estimated dynamic multipliers show that a freeze leads to an immediate increase in prices. Subsequent price rises are much smaller than the initial impact. The cumulative multiplier shows that freezes have a persistent effect on the level of orange juice prices, with prices peaking seven months after the freeze.
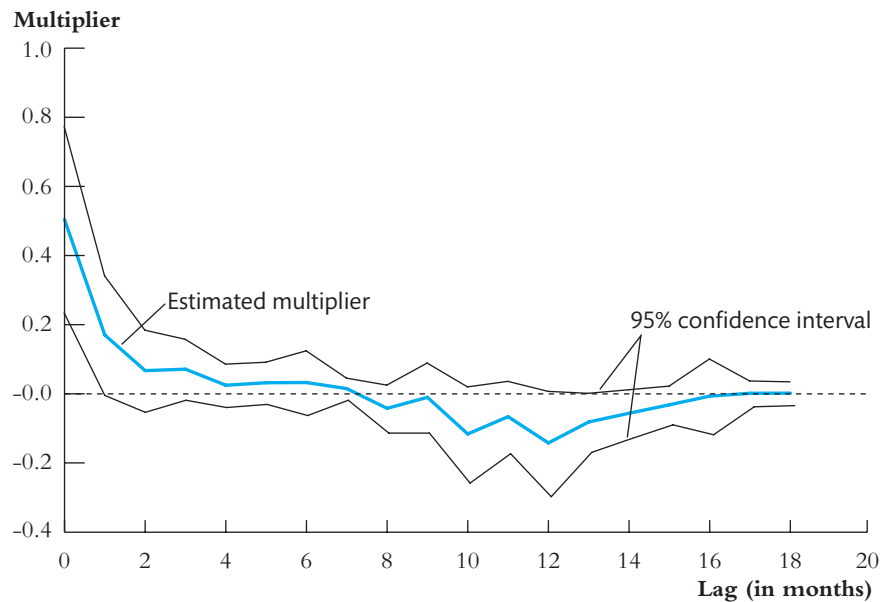
(a) Estimated dynamic multipliers and 95% confidence interval
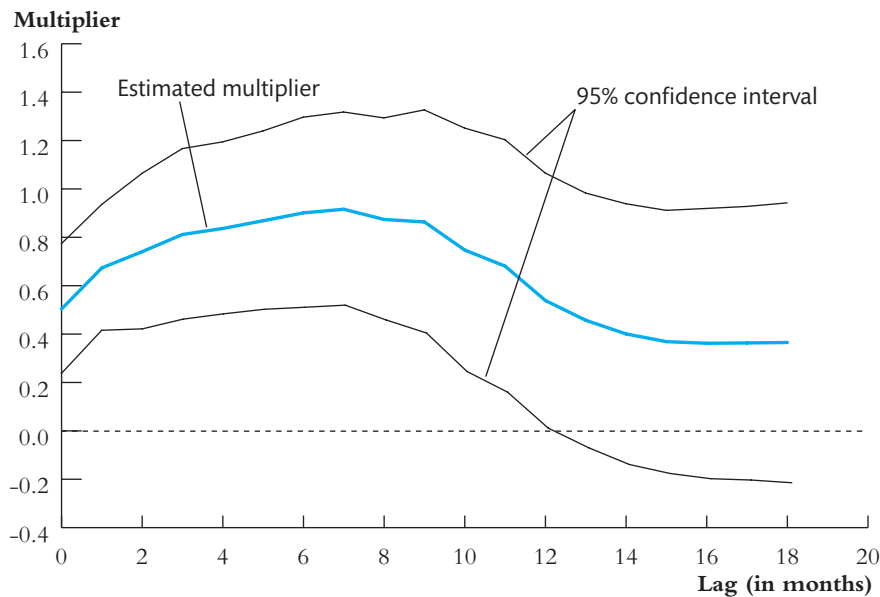
(b) Estimated cumulative dynamic multipliers and 95% confidence interval

only 0.37%. This long-run cumulative dynamic multiplier is not statistically significantly different from 0 at the 10% significance level ($t = 0.37/0.30 = 1.23$).

Sensitivity analysis. As in any empirical analysis, it is important to check whether these results are sensitive to changes in the details of the empirical analysis. We

therefore examine three aspects of this analysis: sensitivity to the computation of the HAC standard errors, an alternative specification that investigates potential omitted variable bias, and an analysis of the stability over time of the estimated multipliers.

First, we investigate whether the standard errors reported in the second column of Table 16.1 are sensitive to different choices of the HAC truncation parameter m. In column (3), results are reported for $m = 14$, twice the value used in column (2). The regression specification is the same as in column (2), so the estimated coefficients and dynamic multipliers are identical; only the standard errors differ but, as it happens, not by much. We conclude that the results are insensitive to changes in the HAC truncation parameter.

Second, we investigate a possible source of omitted variable bias. Freezes in Florida are not randomly assigned throughout the year but rather occur in the winter (of course). If demand for orange juice is seasonal (is demand for orange juice greater in the winter than in the summer?), then the seasonal patterns in orange juice demand could be correlated with FDD, resulting in omitted variable bias. The quantity of oranges sold for juice is endogenous: Prices and quantities are simultaneously determined by the forces of supply and demand. Thus, as discussed in Section 9.2, including quantity would lead to simultaneity bias. Nevertheless, the seasonal component of demand can be captured by including seasonal variables as regressors. The specification in column (4) of Table 16.1 therefore includes 11 monthly binary variables, one indicating whether the month is January, one indicating whether the month is February, and so forth (as usual, one binary variable must be omitted to prevent perfect multicollinearity with the intercept). These monthly indicator variables are not jointly statistically significant at the 10% level ($p = 0.43$), and the estimated cumulative dynamic multipliers are essentially the same as for the specifications excluding the monthly indicators. In summary, seasonal fluctuations in demand are not an important source of omitted variable bias.

Have the dynamic multipliers been stable over time? [3]   To assess the stability of the dynamic multipliers, we need to check whether the distributed lag regression coefficients have been stable over time. Because we do not have a specific break date in mind, we test for instability in the regression coefficients using the Quandt likelihood ratio (QLR) statistic (Key Concept 15.9). The QLR statistic (with 15% trimming and HAC variance estimator) testing the stability of all the coefficients in the regression of column (1) has a value of 21.19, with $q = 20$ degrees of freedom (the coefficients on FDD_t, its 18 lags, and the intercept). The 1% critical value in Table 15.5 is 2.43, so the QLR statistic rejects at the 1% significance level. These QLR regressions have 40 regressors, a large number; recomputing them for 6 lags only (so that there are 16 regressors and $q = 8$) also results in rejection at the 1% level. Thus the hypothesis that the dynamic multipliers are stable is rejected at the 1% significance level.

---

[3]The discussion of stability in this subsection draws on material from Section 15.7 and can be skipped if that material has not been covered.

The dynamic effect on orange juice prices of freezes changed significantly over the second half of the 20th century. A freeze had a larger impact on prices during 1950–1966 than later, and the effect of a freeze was less persistent during 1984–2000 than earlier.
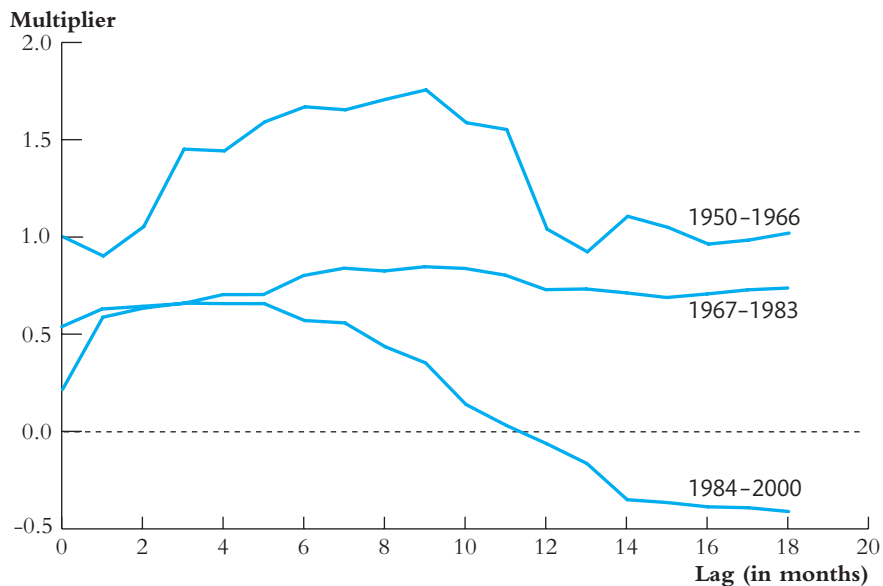
One way to see how the dynamic multipliers have changed over time is to compute them for different parts of the sample. Figure 16.3 plots the estimated cumulative dynamic multipliers for the first third (1950–1966), middle third (1967–1983), and final third (1984–2000) of the sample, computed by running separate regressions on each subsample. These estimates show an interesting and noticeable pattern. In the 1950s and early 1960s, a freezing degree day had a large and persistent effect on the price. The magnitude of the effect on price of a freezing degree day diminished in the 1970s, although it remained highly persistent. In the late 1980s and 1990s, the short-run effect of a freezing degree day was the same as in the 1970s, but it became much less persistent and was essentially eliminated after a year. These estimates suggest that the dynamic causal effect on orange juice prices of a Florida freeze became smaller and less persistent over the second half of the 20th century. The box "Orange Trees on the March" discusses one possible explanation for the instability of the dynamic causal effects.

ADL and GLS estimates. As discussed in Section 16.5, if the error term in the distributed lag regression is serially correlated and *FDD* is strictly exogenous, it is possible to estimate the dynamic multipliers more efficiently than by OLS estimation of the distributed lag coefficients. Before using either the GLS estimator or the estimator based on the ADL model, however, we need to consider whether *FDD* is, in fact, strictly exogenous. True, humans cannot affect the daily weather, but does that mean that the weather is *strictly* exogenous? Does the error term u_t in the distributed lag regression have conditional mean 0 given past, present, and *future* values of *FDD*?

Orange Trees on the March

Why do the dynamic multipliers in Figure 16.3 vary over time? One possible explanation is changes in markets, but another is that the trees moved south.

According to the Florida Department of Citrus, the severe freezes in the 1980s, which are visible in Figure 16.1c, spurred citrus growers to seek a warmer climate. As shown in Figure 16.4, the number of acres of orange trees in the more frost-prone northern and western counties fell from 232,000 acres in 1981 to 53,000 acres in 1985, and orange acreage in southern and central counties subsequently increased from 413,000 in 1985 to 588,000 in 1993. With the groves farther south, northern frosts damage a smaller fraction of the crop, and—as indicated by the dynamic multipliers in Figure 16.3—price becomes less sensitive to temperatures in the more northern city of Orlando.

OK, the orange trees themselves might not have been on the march—that can be left to *Macbeth*—but southern migration of the orange groves does give new meaning to the term *nonstationarity*.[4]

---

[4]The Florida orange juice industry has experienced many other changes since the end of this data set in 2000. Demand for orange juice has declined, and imports from Brazil have increased. Perhaps most important has been the spread of a bacterial disease, citrus greening, that prevents oranges from maturing and kills citrus trees. Between 2000 and 2015, total Florida orange production fell by approximately 60%. We are grateful to Professor James Cobbe of Florida State University for telling us about the southern movement of the orange groves.

FIGURE 16.4   Orange Grove Acreage in Regions of Florida
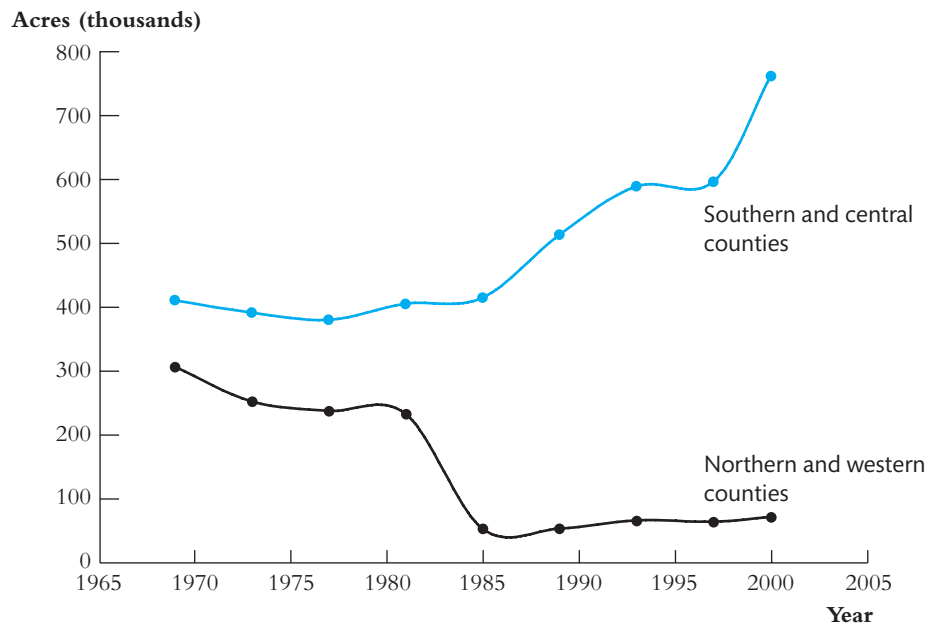
The error term in the population counterpart of the distributed lag regression in column (1) of Table 16.1 is the discrepancy between the price and its population prediction based on the past 18 months of weather. This discrepancy might arise for many reasons, one of which is that traders use forecasts of the weather in Orlando.

NEWS FLASH: Commodity Traders Send Shivers Through Disney World

Although the weather at Disney World in Orlando, Florida, is usually pleasant, now and then a cold spell can settle in. If you are visiting Disney World on a winter evening, should you bring a warm coat? Some people might check the weather forecast on TV, but those in the know can do better: They can check that day's closing price on the New York orange juice futures market!

The financial economist Richard Roll (1984) undertook a detailed study of the relationship between orange juice prices and the weather. He examined the effect on prices of cold weather in Orlando, but he also studied the "effect" of changes in the price of an orange juice futures contract (a contract to buy frozen orange juice concentrate at a specified date in the future) on the weather. Roll used daily data from 1975 to 1981 on the prices of orange juice futures contracts traded at the New York Cotton Exchange and on daily and overnight temperatures in Orlando. He found that a rise in the price of the futures contract during the trading day in New York predicted cold weather—in particular, a freezing spell—in Orlando over the following night. In fact, the market was so effective in predicting cold weather in Florida that a price rise during the trading day actually predicted forecast errors in the official U.S. government weather forecasts for that night.

Roll's study is also interesting for what he did *not* find: Although his detailed weather data explained some of the variation in daily orange juice futures prices, most of the daily movements in orange juice prices remained unexplained. He therefore suggested that the orange juice futures market exhibits "excess volatility"—that is, more volatility than can be attributed to movements in fundamentals. Understanding why (and if) there is excess volatility in financial markets is now an important area of research in financial economics.

Roll's finding also illustrates the difference between forecasting and estimating dynamic causal effects. Price changes on the orange juice futures market are a useful predictor of cold weather, but that does not mean that commodity traders are so powerful that they can *cause* the temperature to fall. Visitors to Disney World might shiver after an orange juice futures contract price rise, but they are not shivering *because* of the price rise—unless, of course, they went short in the orange juice futures market.

For example, if an especially cold winter is forecasted, then traders would incorporate this into the price, so the price would be above its predicted value based on the population regression; that is, the error term would be positive. If this forecast is accurate, then, in fact, future weather would turn out to be cold. Thus future freezing degree days would be positive ($X_{t+1} > 0$) when the current price is unusually high ($u_t > 0$), so corr(X_{t+1}, u_t) is positive. Stated more simply, although orange juice traders cannot influence the weather, they can—and do—predict it (see the box, "NEWS FLASH: Commodity Traders Send Shivers Through Disney World"). Consequently, the error term in the price/weather regression is correlated with future weather. In other words, *FDD* is exogenous, but if this reasoning is true, it is not strictly exogenous, and the GLS and ADL estimators will not be consistent estimators of the dynamic multipliers. These estimators therefore are not used in this application.

16.7   Is Exogeneity Plausible? Some Examples

As in regression with cross-sectional data, the interpretation of the coefficients in a distributed lag regression as causal dynamic effects hinges on the assumption that X is exogenous. If X_t or its lagged values are correlated with u_t, then the conditional mean of u_t will depend on X_t or its lags, in which case X is not (past and present) exogenous. Regressors can be correlated with the error term for several reasons, but with economic time series data, a particularly important concern is that there could be simultaneous causality, which (as discussed in Sections 9.2 and 12.1) results in endogenous regressors. In Section 16.6, we discussed the assumptions of exogeneity and strict exogeneity of freezing degree days in detail. In this section, we examine the assumption of exogeneity in four other economic applications.

U.S. Income and Australian Exports

The United States is an important source of demand for Australian exports. Precisely how sensitive Australian exports are to fluctuations in U.S. aggregate income could be investigated by regressing Australian exports to the United States against a measure of U.S. income. Strictly speaking, because the world economy is integrated, there is simultaneous causality in this relationship: A decline in Australian exports reduces Australian income, which reduces demand for imports from the United States, which reduces U.S. income. As a practical matter, however, this effect is very small because the Australian economy is much smaller than the U.S. economy. Thus U.S. income plausibly can be treated as exogenous in this regression.

In contrast, in a regression of European Union exports to the United States against U.S. income, the argument for treating U.S. income as exogenous is less convincing because demand by residents of the European Union for U.S. exports constitutes a substantial fraction of the total demand for U.S. exports. Thus a decline in U.S. demand for EU exports would decrease EU income, which in turn would decrease demand for U.S. exports and thus decrease U.S. income. Because of these linkages through international trade, EU exports to the United States and U.S. income are simultaneously determined, so in this regression U.S. income arguably is not exogenous. This example illustrates a more general point that whether a variable is exogenous depends on the context: U.S. income is plausibly exogenous in a regression explaining Australian exports but not in a regression explaining EU exports.

Oil Prices and Inflation

Ever since the oil price increases of the 1970s, macroeconomists have been interested in estimating the dynamic effect of an increase in the international price of crude oil

on the U.S. rate of inflation. Because oil prices are set in world markets in large part by foreign oil-producing countries, initially one might think that oil prices are exogenous. But oil prices are not like the weather: Members of the Organization of Petroleum Exporting Countries set oil production levels strategically, taking many factors, including the state of the world economy, into account. To the extent that oil prices (or quantities) are set based on an assessment of current and future world economic conditions, including inflation in the United States, oil prices are endogenous.

Monetary Policy and Inflation

The central bankers in charge of monetary policy need to know the effect on inflation of monetary policy. Because an important tool of monetary policy is the short-term interest rate (the *short rate*), they need to know the dynamic causal effect on inflation of a change in the short rate. Although the short rate is determined by the central bank, it is not set by the central bankers at random (as it would be in an ideal randomized experiment); rather, it is set endogenously: The central bank determines the short rate based on an assessment of the current and future states of the economy, especially including the current and future rates of inflation. The rate of inflation in turn depends on the interest rate (higher interest rates reduce aggregate demand), but the interest rate depends on the rate of inflation, its past value, and its (expected) future value. Thus the short rate is endogenous, and the dynamic causal effect of a change in the short rate on future inflation cannot be consistently estimated by an OLS regression of the rate of inflation on current and past interest rates.

The Growth Rate of GDP and the Term Spread

In Chapter 15, lagged values of the term spread were used to forecast future values of the growth rate of GDP. Because lags of the term spread happened in the past, one might initially think that there cannot be feedback from current growth rates of GDP to past values of the term spread, so past values of the term spread can be treated as exogenous. But past values of the term spread were not randomly assigned in an experiment; instead, the past term spread was simultaneously determined with past values of the growth rate of GDP. Because GDP and the interest rates making up the term spread are simultaneously determined, the other factors that determine the growth rate of GDP contained in u_t are correlated with past values of the term spread; that is, the term spread is not exogenous. It follows that the term spread is not strictly exogenous, so the dynamic multipliers computed using an ADL model [for example, the ADL model in Equation (15.20)] are not consistent estimates of the dynamic causal effect on the growth rate of GDP of a change in the term spread.

16.8 Conclusion

Time series data provide the opportunity to estimate the time path of the effect on Y of a change in X—that is, the dynamic causal effect on Y of a change in X. To estimate dynamic causal effects using a distributed lag regression, however, X must be exogenous, as it would be if it were set randomly in an ideal randomized experiment. If X is not just exogenous but is *strictly* exogenous, then the dynamic causal effects can be estimated using an autoregressive distributed lag model or by GLS.

In some applications, such as estimating the dynamic causal effect on the price of orange juice of freezing weather in Florida, a convincing case can be made that the regressor (freezing degree days) is exogenous; thus the dynamic causal effect can be estimated by OLS estimation of the distributed lag coefficients. Even in this application, however, economic theory suggests that the weather is not strictly exogenous, so the ADL and GLS methods are inappropriate. Moreover, in many relations of interest to econometricians, there is simultaneous causality, so the regressor in these specifications is not exogenous, strictly or otherwise. Ascertaining whether the regressor is exogenous (or strictly exogenous) ultimately requires combining economic theory, institutional knowledge, and careful judgment.

Summary

1. Dynamic causal effects in time series are defined in the context of a randomized experiment, where the same subject (entity) receives different randomly assigned treatments at different times. The coefficients in a distributed lag regression of Y on X and its lags can be interpreted as the dynamic causal effects when the time path of X is determined randomly and independently of other factors that influence Y.

2. The variable X is (past and present) exogenous if the conditional mean of the error u_t in the distributed lag regression of Y on current and past values of X does not depend on current and past values of X. If, in addition, the conditional mean of u_t does not depend on future values of X, then X is strictly exogenous.

3. If X is exogenous, then the OLS estimators of the coefficients in a distributed lag regression of Y on current and past values of X are consistent estimators of the dynamic causal effects. In general, the error u_t in this regression is serially correlated, so conventional standard errors are misleading and HAC standard errors must be used instead.

4. If X is strictly exogenous, then the dynamic multipliers can be estimated using either OLS estimation of an ADL model or GLS.

5. Exogeneity is a strong assumption that often fails to hold in economic time series data because of simultaneous causality, and the assumption of strict exogeneity is even stronger.

Key Terms

dynamic causal effect (567)

distributed lag model (572)

exogeneity (573)

strict exogeneity (573)

dynamic multiplier (576)

impact effect (577)

cumulative dynamic multiplier (577)

long-run cumulative dynamic
    multiplier (577)

heteroskedasticity- and autocorrelation-
    consistent (HAC) standard error
    (580)

truncation parameter (580)

Newey–West variance estimator (581)

quasi-difference (584)

generalized least squares (GLS) (586)

infeasible GLS estimator (586)

feasible GLS estimator (587)

MyLab Economics Can Help You Get a Better Grade

MyLab Economics   If your exam were tomorrow, would you be ready? For each
    chapter, **MyLab Economics** Practice Tests and Study Plan
help you prepare for your exams. You can also find the Exercises and all Review the Concepts Questions
available now in **MyLab Economics**. To see how it works, turn to the **MyLab Economics** spread on the
inside front cover of this text and then go to **www.pearson.com/mylab/economics**.

For additional Empirical Exercises and Data Sets, log on to the Companion Website at
www.pearsonhighered.com/stock_watson.

Review the Concepts

16.1   In the 1970s, a common practice was to estimate a distributed lag model relat-
ing changes in nominal GDP (Y) to current and past changes in the money
supply (X). Under what assumptions will this regression estimate the causal
effects of money on nominal GDP? Are these assumptions likely to be satis-
fied in a modern economy like that of the United States?

16.2   Suppose that X is strictly exogenous. A researcher estimates an ADL(1, 1)
model, calculates the regression residual, and finds the residual to be highly
serially correlated. Should the researcher estimate a new ADL model with
additional lags or simply use HAC standard errors for the ADL(1, 1) esti-
mated coefficients?

16.3   Suppose that a distributed lag regression is estimated, where the dependent
variable is ΔY_t instead of Y_t. Explain how you would compute the dynamic
multipliers of X_t on Y_t.

16.4   Suppose that you added FDD_{t+1} as an additional regressor in Equation (16.2).
If FDD is strictly exogenous, would you expect the coefficient on FDD_{t+1} to
be 0 or nonzero? Would your answer change if FDD is exogenous but not
strictly exogenous?

Exercises

16.1 Increases in oil prices have been blamed for several recessions in developed countries. To quantify the effect of oil prices on real economic activity, researchers have run regressions like those discussed in this chapter. Let GDP_t denote the value of quarterly real GDP in the United States, and let $Y_t = 100\ln(GDP_t/GDP_{t-1})$ be the quarterly percentage change in GDP. James Hamilton, an econometrician and macroeconomist, has suggested that oil prices adversely affect that economy only when they jump above their values in the recent past. Specifically, let O_t equal the greater of 0 or the percentage point difference between oil prices at date t and their maximum value during the past three years. A distributed lag regression relating Y_t and O_t, estimated over 1960:Q1–2017:Q4, is

$$\hat{Y}_t = 1.0 - 0.006O_t - 0.014O_{t-1} - 0.020O_{t-2} - 0.024O_{t-3} - 0.036O_{t-4}$$
$$\quad (0.1)\ \ (0.013) \qquad (0.011) \qquad (0.010) \qquad (0.009) \qquad (0.012)$$

$$\quad - 0.013O_{t-5} + 0.005O_{t-6} - 0.007O_{t-7} + 0.005O_{t-8}.$$
$$\qquad (0.007) \qquad\quad (0.010) \qquad\quad (0.008) \qquad\quad (0.008)$$

a. Suppose that oil prices jump 25% above their previous peak value and stay at this new higher level (so that $O_t = 25$ and $O_{t+1} = O_{t+2} = \cdots = 0$). What is the predicted effect on output growth for each quarter over the next two years?

b. Construct a 95% confidence interval for your answers to (a).

c. What is the predicted cumulative change in GDP growth over eight quarters?

d. The HAC F-statistic testing whether the coefficients on O_t and its lags are 0 is 5.45. Are the coefficients significantly different from 0?

16.2 Macroeconomists have also noticed that interest rates change following oil price jumps. Let R_t denote the interest rate on three-month Treasury bills (in percentage points at an annual rate). The distributed lag regression relating the change in R_t (ΔR_t) to O_t estimated over 1960:Q1–2017:Q4 is

$$\widehat{\Delta R_t} = 0.03 + 0.013O_t + 0.013O_{t-1} - 0.004O_{t-2} - 0.024O_{t-3} - 0.000O_{t-4}$$
$$\qquad (0.05)\ \ (0.010) \qquad (0.010) \qquad (0.008) \qquad (0.015) \qquad (0.010)$$

$$\quad + 0.006O_{t-5} - 0.005O_{t-6} - 0.018O_{t-7} - 0.004O_{t-8}.$$
$$\qquad (0.015) \qquad\quad (0.015) \qquad\quad (0.010) \qquad\quad (0.006)$$

a. Suppose that oil prices jump 25% above their previous peak value and stay at this new higher level (so that $O_t = 25$ and $O_{t+1} = O_{t+2} = \cdots = 0$). What is the predicted change in interest rates for each quarter over the next two years?

b. Construct 95% confidence intervals for your answers to (a).

    c. What is the effect of this change in oil prices on the level of interest rates in period $t + 8$? How is your answer related to the cumulative multiplier?

    d. The HAC F-statistic testing whether the coefficients on O_t and its lags are 0 is 1.92. Are the coefficients significantly different from 0?

16.3 Consider two different randomized experiments. In experiment A, oil prices are set randomly, and the central bank reacts according to its usual policy rules in response to economic conditions, including changes in the oil price. In experiment B, oil prices are set randomly, and the central bank holds interest rates constant and in particular does not respond to the oil price changes. In both experiments, GDP growth is observed. Now suppose that oil prices are exogenous in the regression in Exercise 16.1. To which experiment, A or B, does the dynamic causal effect estimated in Exercise 16.1 correspond?

16.4 Suppose that oil prices are strictly exogenous. Discuss how you could improve on the estimates of the dynamic multipliers in Exercise 16.1.

16.5 Derive Equation (16.7) from Equation (16.4), and show that $\delta_0 = \beta_0$, $\delta_1 = \beta_1$, $\delta_2 = \beta_1 + \beta_2$, $\delta_3 = \beta_1 + \beta_2 + \beta_3$ (etc.). (*Hint:* Note that $X_t = \Delta X_t + \Delta X_{t-1} + \cdots + \Delta X_{t-p+1} + X_{t-p}$.)

16.6 Consider the regression model $Y_t = \beta_0 + \beta_1 X_t + u_t$, where u_t follows the stationary AR(1) model $u_t = \phi_1 u_{t-1} + \tilde{u}_t$ with $\tilde{u}_t$ i.i.d. with mean 0 and variance $\sigma_{\tilde{u}}^2$ and $|\phi_1| < 1$; the regressor X_t follows the stationary AR(1) model $X_t = \gamma_1 X_{t-1} + e_t$ with e_t i.i.d. with mean 0 and variance σ_e^2 and $|\gamma| < 1$; and e_t is independent of $\tilde{u}_i$ for all t and i.

    a. Show that $\mathrm{var}(u_t) = \dfrac{\sigma_{\tilde{u}}^2}{1 - \phi_1^2}$ and $\mathrm{var}(X_t) = \dfrac{\sigma_e^2}{1 - \gamma_1^2}$.

    b. Show that $\mathrm{cov}(u_t, u_{t-j}) = \phi_1^j \mathrm{var}(u_t)$ and $\mathrm{cov}(X_t, X_{t-j}) = \gamma_1^j \mathrm{var}(X_t)$.

    c. Show that $\mathrm{corr}(u_t, u_{t-j}) = \phi_1^j$ and $\mathrm{corr}(X_t, X_{t-j}) = \gamma_1^j$.

    d. Consider the terms σ_v^2 and f_T in Equation (16.14).

        i. Show that $\sigma_v^2 = \sigma_X^2 \sigma_u^2$, where σ_X^2 is the variance of X and σ_u^2 is the variance of u.

        ii. Derive an expression for f_∞.

16.7 Consider the regression model $Y_t = \beta_0 + \beta_1 X_t + u_t$, where u_t follows the stationary AR(1) model $u_t = \phi_1 u_{t-1} + \tilde{u}_t$ with $\tilde{u}_t$ i.i.d. with mean 0 and variance $\sigma_{\tilde{u}}^2$ and $|\phi_1| < 1$.

    a. Suppose that X_t is independent of $\tilde{u}_j$ for all t and j. Is X_t exogenous (past and present)? Is X_t strictly exogenous (past, present, and future)?

    b. Suppose that $X_t = \tilde{u}_{t+1}$. Is X_t exogenous? Is X_t strictly exogenous?

16.8 Consider the model in Exercise 16.7 with $X_t = \tilde{u}_{t+1}$.

    a. Is the OLS estimator of β_1 consistent? Explain.

   b. Explain why the GLS estimator of β_1 is not consistent.

   c. Show that the infeasible GLS estimator $\hat{\beta}_1^{GLS} \xrightarrow{p} \beta_1 - \frac{\phi_1}{1 + \phi_1^2}$.

   [*Hint:* Apply the omitted variable formula in Equation (6.1) to the quasi-differenced regression in Equation (16.23).]

16.9 Consider the constant-term-only regression model $Y_t = \beta_0 + u_t$, where u_t follows the stationary AR(1) model $u_t = \phi_1 u_{t-1} + \tilde{u}_t$ with $\tilde{u}_t$ i.i.d. with mean 0 and variance $\sigma_{\tilde{u}}^2$ and $|\phi_1| < 1$.

   a. Show that the OLS estimator is $\hat{\beta}_0 = T^{-1} \sum_{t=1}^{T} Y_t$.

   b. Show that the (infeasible) GLS estimator is $\hat{\beta}_0^{GLS} = (1 - \phi_1)^{-1}(T - 1)^{-1}$ $\sum_{t=2}^{T}(Y_t - \phi_1 Y_{t-1})$. [*Hint:* The GLS estimator of β_0 is $(1 - \phi_1)^{-1}$ multiplied by the OLS estimator of α_0 in Equation (16.23). Why?]

   c. Show that $\hat{\beta}_0^{GLS}$ can be written as $\hat{\beta}_0^{GLS} = (T - 1)^{-1} \sum_{t=2}^{T-1} Y_t + (1 - \phi_1)^{-1}$ $(T - 1)^{-1}(Y_T - \phi_1 Y_1)$. [*Hint:* Rearrange the formula in (b).]

   d. Derive the difference $\hat{\beta}_0 - \hat{\beta}_0^{GLS}$, and discuss why it is likely to be small when T is large.

16.10 Consider the ADL model $Y_t = 3.1 + 0.4Y_{t-1} + 2.0X_t - 0.8X_{t-1} + \tilde{u}_t$, where X_t is strictly exogenous.

   a. Derive the impact effect of X on Y.

   b. Derive the first five dynamic multipliers.

   c. Derive the first five cumulative multipliers.

   d. Derive the long-run cumulative dynamic multiplier.

16.11 Suppose that $a(L) = (1 - \phi L)$, with $|\phi_1| < 1$, and $b(L) = 1 + \phi L + \phi^2 L^2 + \phi^3 L^3 \dots$.

   a. Show that the product $b(L)a(L) = 1$, so that $b(L) = a(L)^{-1}$.

   b. Why is the restriction $|\phi_1| < 1$ important?

16.12 Suppose $Y_t = \beta_0 + u_t$, where u_t follows a stationary stationary AR(1) $u_t = \phi_1 u_{t-1} + \tilde{u}_t$ with $\tilde{u}_t$ i.i.d. with mean 0 and variance $\sigma_{\tilde{u}}^2$ and $|\phi_1| < 1$.

   a. Show that $\beta_0 = \mu_Y = E(Y_t)$.

   b. Let $\overline{Y}_{1:T} = \frac{1}{T}\sum_{t=1}^{T} Y_t$ denote the sample mean of Y_t using observations from $t = 1$ through $t = T$. Show that the OLS estimator of β_0 is $\hat{\beta}_0 = \overline{Y}_{1:T}$.

   c. Show that $\text{var}[\sqrt{T}(\overline{Y}_{1:T} - \mu_Y)] \rightarrow \sigma_{\tilde{u}}^2/(1 - \phi_1)^2$.

   d. Assume that $\overline{Y}_{1:T}$ is approximately normally distributed with mean μ_Y and variance $\sigma_{\tilde{u}}^2/[T(1 - \phi_1)^2]$. Suppose $T = 200$, $\sigma_{\tilde{u}}^2 = 7.9$, $\phi_1 = 0.3$, and the sample mean of Y_t is $\overline{Y}_{1:T} = 2.8$. Construct a 95% confidence interval for μ_Y.

   e. Suppose you are interested in the average value of Y_t from $t = T + 1$ through $T + h$; that is, $\overline{Y}_{T+1:T+h} = \frac{1}{h}\sum_{t=T+1}^{T+h} Y_t$, where h is a large number. Show that $\overline{Y}_{T+1:T+h}$ has mean μ_Y and variance $\sigma_{\tilde{u}}^2/[h(1 - \phi_1)^2]$.

f. Assume that $\overline{Y}_{T+1:T+h}$ is approximately normally distributed. Suppose $h = 100, \sigma_{\tilde{u}}^2 = 7.9, \phi_1 = 0.3$, and $\mu_Y = 2.9$. Construct a 95% forecast interval for $\overline{Y}_{T+1:T+h}$.

g. Let $r = h/T$. Show that $\text{var}\left[\sqrt{T}(\overline{Y}_{T+1:T+h} - \overline{Y}_{1:T})\right] \rightarrow (1 + r^{-1})\frac{\sigma_{\tilde{u}}^2}{(1 - \phi_1)^2}$, where r is held fixed as $T \rightarrow \infty$.

h. Show that $\overline{Y}_{T+1:T+h} - \overline{Y}_{1:T}$ has mean 0 and variance $\left(\frac{1}{T} + \frac{1}{h}\right)\frac{\sigma_{\tilde{u}}^2}{(1 - \phi_1)^2}$.

i. Use the result in (i) to show that the forecast interval $\overline{Y}_{1:T} \pm 1.96\sqrt{\left(\frac{1}{T} + \frac{1}{h}\right)\frac{\sigma_{\tilde{u}}^2}{(1 - \phi_1)^2}}$ will contain the value of $\overline{Y}_{T+1:T+h}$ with probability 95%, approximately, when T and h are large. (Assume that $\overline{Y}_{T+1:T+h} - \overline{Y}_{1:T}$ is approximately normally distributed.)

j. Suppose $T = 200, h = 100, \sigma_{\tilde{u}}^2 = 7.9, \phi_1 = 0.3$, and $\overline{Y}_{1:T} = 2.8$. Construct a 95% forecast interval for $\overline{Y}_{T+1:T+h}$.

Empirical Exercises

E16.1 In this exercise, you will estimate the effect of oil prices on macroeconomic activity using monthly data on the Index of Industrial Production (IP) and the monthly measure of O_t described in Exercise 16.1. The data can be found on the text website, **http://www.pearsonhighered.com/stock_watson**, in the file **USMacro_Monthly**.

a. Compute the monthly growth rate in IP, expressed in percentage points, $ip\_growth_t = 100 \times \ln(IP_t/IP_{t-1})$. What are the mean and standard deviation of ip_growth over the 1960:M1–2017:M12 sample period? What are the units for ip_growth (percent, percent per annum, percent per month, or something else)?

b. Plot the value of O_t. Why are so many values of O_t equal to 0? Why aren't some values of O_t negative?

c. Estimate a distributed lag model by regressing ip_growth onto the current value and 18 lagged values of O_t, including an intercept. What value of the HAC standard error truncation parameter m did you choose? Why?

d. Taken as a group, are the coefficients on O_t statistically significantly different from 0?

e. Construct graphs like those in Figure 16.2, showing the estimated dynamic multipliers, cumulative multipliers, and 95% confidence intervals. Comment on the real-world size of the multipliers.

f. Suppose that high demand in the United States (evidenced by large values of ip_growth) leads to increases in oil prices. Is O_t exogenous? Are the estimated multipliers shown in the graphs in (e) reliable? Explain.

E16.2 In the data file **USMacro_Quarterly**, you will find data on two aggregate price series for the United States: the price index for personal consumption

expenditures (PCEP), which you used in Empirical Exercise 15.1, and the Consumer Price Index (CPI). These series are alternative measures of consumer prices in the United States. The CPI prices a basket of goods whose composition is updated every 5–10 years. The PCEP uses chain weighting to price a basket of goods whose composition changes from month to month. Economists have argued that the CPI will overstate inflation because it does not take into account the substitution that occurs when relative prices change. If this substitution bias is important, then average CPI inflation should be systematically higher than PCEP inflation. Let $\pi_t^{CPI} = 400 \times [\ln(CPI_t) - \ln(CPI_{t-1})]$, $\pi_t^{PCEP} = 400 \times [\ln(PCEP_t) - \ln(PCEP_{t-1})]$, and $Y_t = \pi_t^{CPI} - \pi_t^{PCEP}$, so π_t^{CPI} is the quarterly rate of price inflation (measured in percentage points at an annual rate) based on the CPI, π_t^{PCEP} is the quarterly rate of price inflation from the PCEP, and Y_t is their difference. Using data from 1963:Q1 through 2017:Q4, carry out the following exercises.

   a. Compute the sample means of π_t^{CPI} and π_t^{PCEP}. Are these point estimates consistent with the presence of economically significant substitution bias in the CPI?

   b. Compute the sample mean of Y_t. Explain why it is numerically equal to the difference in the means computed in (a).

   c. Show that the population mean of Y is equal to the difference of the population means of the two inflation rates.

   d. Consider the constant-term-only regression $Y_t = \beta_0 + u_t$. Show that $\beta_0 = E(Y)$. Do you think that u_t is serially correlated? Explain.

   e. Construct a 95% confidence interval for β_0. What value of the HAC standard error truncation parameter m did you choose? Why?

   f. Is there statistically significant evidence that the mean inflation rate for the CPI is greater than the rate for the PCEP?

   g. Is there evidence of instability in β_0? Carry out a QLR test. (*Hint:* Make sure you use HAC standard errors for the regressions in the QLR procedure.)

E16.3 In the data file **USMacro_Quarterly**, you will find the data on U.S. real GDP (GDPC1) that was analyzed in Chapter 15. In this exercise, you will construct a 95% confidence interval for the mean growth rate of real GDP in the United States; in addition, you will construct a 95% forecast interval for the average growth rate of real GDP for 2018:Q1–2067:Q4. Before attempting this empirical exercise, you should answer Exercise 16.12.

   a. Compute the growth rate of real GDP: $Y_t = 400 \times [\ln(GDPC1_t) - \ln(GDPC1_{t-1})]$. Plot the series from 1960 through 2017, and verify that the data are the same as plotted in Figure 15.1b.

   b. Using the data from 1960:Q1 through 2017:Q4:

    i. Estimate an AR(1) model for Y_t. In the notation of Exercise 16.12, denote the estimated AR(1) coefficient by $\hat{\phi}_1$ and the standard error of the regression as $\hat{\sigma}_{\tilde{u}}$.

    ii. Compute the sample mean of Y_t.

  c. Assuming that Y_t follows an AR(1), use the results you derived in Exercise 16.12, the estimated values of ϕ_1 and $\sigma_{\tilde{u}}^2$ from (b.i), and the sample mean from (b.ii) to

    i. Construct a 95% confidence interval for μ_Y, the mean growth rate of real GDP.

    ii. Construct a 95% forecast interval for the average growth rate of real GDP over the period 2018:Q1–2067:Q4—that is, for $\overline{Y}_{2018Q1:2067Q4}$.

  d. Using the data from 1960:Q1 through 2017:Q4:

    i. Regress Y_t on a constant (with no other regressors). Construct the standard error for the estimated constant using the Newey–West HAC estimator with four lags.

    ii. Use the results from this regression to construct a 95% confidence interval for μ_Y, the mean growth rate of real GDP.

    iii. Use the results from this regression to construct a 95% forecast interval for the average growth rate of real GDP over the period 2018:Q1–2067:Q4—that is, for $\overline{Y}_{2018Q1:2067Q4}$.

  e. Are the intervals constructed in (d.ii) and (d.iii) similar to the intervals constructed in (c.i) and (c.ii)? Should they be? Explain.

APPENDIX

16.1 The Orange Juice Data Set

The orange juice price data are the frozen orange juice component of the processed foods and feeds group of the Producer Price Index (PPI), collected by the U.S. Bureau of Labor Statistics (BLS Series wpu02420301). The orange juice price series was divided by the overall PPI for finished goods to adjust for general price inflation. The freezing degree days series was constructed from daily minimum temperatures recorded at Orlando-area airports, obtained from the National Oceanic and Atmospheric Administration (NOAA) of the U.S. Department of Commerce. The *FDD* series was constructed so that its timing and the timing of the orange juice price data were approximately aligned. Specifically, the frozen orange juice price data are collected by surveying a sample of producers in the middle of every month, although the exact date varies from month to month. Accordingly, the *FDD* series was constructed to be the number of freezing degree days from the 11th of one month to the 10th of the next month; that is, *FDD* is the maximum of 0 and 32 minus the minimum daily temperature, summed over all days from the 11th to the 10th. Thus *%ChgP_t* for February is the percentage change in real orange juice prices from mid-January to mid-February, and *FDD_t* for February is the number of freezing degree days from January 11 to February 10.

16.2 The ADL Model and Generalized Least Squares in Lag Operator Notation

Section 16.5 introduced the autoregressive distributed lag model for the case that the error term in the distributed lag model is AR(1). This appendix extends the ADL model to the case of AR(p) errors, using the lag operator notation introduced in Appendix 15.3.

The Distributed Lag, ADL, and Quasi-Difference Models in Lag Operator Notation

As defined in Appendix 15.3, the lag operator, L, has the property that $L^j X_t = X_{t-j}$, and the distributed lag $\beta_1 X_t + \beta_2 X_{t-1} + \cdots + \beta_{r+1} X_{t-r}$ can be expressed as $\beta(L)X_t$, where $\beta(L) = \sum_{j=0}^{r} \beta_{j+1} L^j$, where $L^0 = 1$. Thus the distributed lag model in Key Concept 16.1 [Equation (16.4)] can be written in lag operator notation as

$$Y_t = \beta_0 + \beta(L)X_t + u_t. \tag{16.30}$$

In addition, if the error term u_t follows an AR(p), then it can be written as

$$\phi(L)u_t = \tilde{u}_t, \tag{16.31}$$

where $\phi(L) = \sum_{j=0}^{p} \phi_j L^j$, where $\phi_0 = 1$, and $\tilde{u}_t$ is serially uncorrelated [note that, in the case $p = 1, \phi_1$ as defined here is the negative of ϕ_1 in the notation of Equation (16.19)].

To derive the ADL model, premultiply each side of Equation (16.30) by $\phi(L)$ so that

$$\phi(L)Y_t = \phi(L)[\beta_0 + \beta(L)X_t + u_t] = \alpha_0 + \delta(L)X_t + \tilde{u}_t, \tag{16.32}$$

where

$$\alpha_0 = \phi(1)\beta_0 \text{ and } \delta(L) = \phi(L)\beta(L), \text{ where } \phi(1) = \sum_{j=0}^{p} \phi_j. \tag{16.33}$$

The model in Equation (16.32) is the ADL(p, q) model including the contemporaneous value of X, where p is the number of lags of Y and q is the number of lags of X.

To derive the quasi-differenced model, note that $\phi(L)\beta(L)X_t = \beta(L)\phi(L)X_t = \beta(L)\tilde{X}_t$, where $\tilde{X}_t = \phi(L)X_t$. Thus rearranging Equation (16.32) yields

$$\tilde{Y}_t = \alpha_0 + \beta(L)\tilde{X}_t + \tilde{u}_t, \tag{16.34}$$

where $\tilde{Y}_t$ is the quasi-difference of Y_t; that is, $\tilde{Y}_t = \phi(L)Y_t$.

The Inverse of a Lag Polynomial

Let $a(x) = \sum_{j=0}^{p} a_j x^j$ denote a polynomial of order p. The inverse of $a(x)$—say, $b(x)$—is a function that satisfies $b(x)a(x) = 1$. If the roots of the polynomial $a(x)$ are greater than 1 in absolute value, then $b(x)$ can be written as a polynomial in nonnegative powers of x: $b(x) = \sum_{j=0}^{\infty} b_j x^j$. Because $b(x)$ is the inverse of $a(x)$, it is denoted as $a(x)^{-1}$ or as $1/a(x)$.

The inverse of a lag polynomial $a(L)$ is defined analogously: $a(L)^{-1} = 1/a(L) = b(L) = \sum_{j=0}^{\infty} b_j L^j$, where $b(L)a(L) = 1$. For example, if $a(L) = (1 - \phi L)$, with $|\phi| < 1$, you can verify that $a(L)^{-1} = 1 + \phi L + \phi^2 L^2 + \phi^3 L^3 \ldots = \sum_{j=0}^{\infty} \phi^j L^j$. (See Exercise 16.11.)

The OLS and GLS Estimators

The OLS estimator of the ADL coefficients is obtained by OLS estimation of Equation (16.32). The original distributed lag coefficients are $\beta(L)$, which, in terms of the estimated coefficients, are $\beta(L) = \phi(L)^{-1}\delta(L)$; that is, the coefficients in $\beta(L)$ satisfy the restrictions implied by $\phi(L)\beta(L) = \delta(L)$. Thus the estimator of the dynamic multipliers based on the OLS estimators of the coefficients of the ADL model, $\hat{\delta}(L)$ and $\hat{\phi}(L)$, is

$$\hat{\beta}^{ADL}(L) = \hat{\phi}(L)^{-1}\hat{\delta}(L). \tag{16.35}$$

The expressions for the coefficients in Equation (16.29) in the text are obtained as a special case of Equation (16.35) when $p = 1$ and $q = 2$.

The feasible GLS estimator is computed by obtaining a preliminary estimator of $\phi(L)$, computing estimated quasi-differences, estimating $\beta(L)$ in Equation (16.34) using these estimated quasi-differences, and (if desired) iterating until convergence. The iterated feasible GLS estimator is the nonlinear least squares estimator of the ADL model in Equation (16.32), subject to the nonlinear restrictions on the parameters contained in Equation (16.33).

Conditions for estimation of the ADL coefficients. The discussion in Section 16.5 of the conditions for consistent estimation of the ADL coefficients in the AR(1) case extends to the general model with AR(p) errors. The conditional mean 0 assumption for Equation (16.34) is that

$$E(\tilde{u}_t \,|\, \tilde{X}_t, \tilde{X}_{t-1}, \ldots) = 0. \tag{16.36}$$

Because $\tilde{u}_t = \phi(L)u_t$ and $\tilde{X}_t = \phi(L)X_t$, this condition is equivalent to

$$\begin{aligned} E(u_t | X_t, X_{t-1}, \ldots) &+ \phi_1 E(u_{t-1} | X_t, X_{t-1}, \ldots) \\ &+ \cdots + \phi_p E(u_{t-p} | X_t, X_{t-1}, \ldots) = 0. \end{aligned} \tag{16.37}$$

For Equation (16.37) to hold for general values of $\phi_1, \ldots, \phi_p$, it must be the case that each of the conditional expectations in Equation (16.37) is 0; equivalently, it must be the case that

$$E(u_t | X_{t+p}, X_{t+p-1}, X_{t+p-2}, \ldots) = 0. \tag{16.38}$$

This condition is not implied by X_t being (past and present) exogenous, but it is implied by X_t being strictly exogenous. In fact, in the limit when p is infinite (so that the error term in the distributed lag model follows an infinite-order autoregression), the condition in Equation (16.38) becomes the condition in Key Concept 16.1 for strict exogeneity.

Additional Topics in Time Series Regression

This chapter takes up some further topics in time series regression, starting with forecasting. Chapter 15 considered forecasting a single variable. In practice, however, you might want to forecast two or more variables, such as the growth rate of gross domestic product (GDP) and the rate of inflation. Section 17.1 introduces a model for forecasting multiple variables, vector autoregressions (VARs), in which lagged values of two or more variables are used to forecast future values of those variables. Chapter 15 focused on making forecasts one period (e.g., one quarter) into the future, but making forecasts two, three, or more periods into the future is important as well. Methods for making multi-period forecasts are discussed in Section 17.2.

Sections 17.3 and 17.4 return to the topic of Section 15.6, stochastic trends. Section 17.3 introduces additional models of stochastic trends. Section 17.4 introduces the concept of cointegration, which arises when two variables share a common stochastic trend—that is, when each variable contains a stochastic trend but a weighted difference of the two variables does not.

In some time series data, especially financial data, the variance changes over time: Sometimes the series exhibits high volatility, while at other times the volatility is low, so the data exhibit clusters of volatility. Section 17.5 discusses volatility clustering and introduces models in which the variance of the forecast error changes over time—that is, models in which the forecast error is conditionally heteroskedastic. Models of conditional heteroskedasticity have several applications. One application is computing forecast intervals, where the width of the interval changes over time to reflect periods of high or low uncertainty. Another application is forecasting the uncertainty of returns on an asset, such as a stock, which in turn can be useful in assessing the risk of owning that asset or forecasting the price of derivative assets that depend on this risk.

Section 17.6 takes up the challenge of forecasting when there are many predictors, as is the case for macroeconomic data in developed economies. This section draws on material introduced in Chapter 14 and focuses on one commonly used method for forecasting with large data sets, which uses principal components analysis to reduce the information in a large time series data set to a small number of time series. The framework for doing so is the dynamic factor model, which also can be used for purposes other than forecasting.

17.1 Vector Autoregressions

Chapter 15 focused on forecasting the growth rate of GDP, but in reality, economic forecasters are in the business of forecasting other key macroeconomic variables as well, such as the rate of inflation, the unemployment rate, and interest rates. One

Vector Autoregressions

A vector autoregression (VAR) is a set of k time series regressions, in which the regressors are lagged values of all k series. A VAR extends the univariate autoregression to a list, or "vector," of time series variables. When the number of lags in each of the equations is the same and is equal to p, the system of equations is called a VAR(p).

In the case of two time series variables, Y_t and X_t, the VAR(p) consists of the two equations,

$$Y_t = \beta_{10} + \beta_{11}Y_{t-1} + \cdots + \beta_{1p}Y_{t-p} + \gamma_{11}X_{t-1} + \cdots + \gamma_{1p}X_{t-p} + u_{1t} \quad (17.1)$$

$$X_t = \beta_{20} + \beta_{21}Y_{t-1} + \cdots + \beta_{2p}Y_{t-p} + \gamma_{21}X_{t-1} + \cdots + \gamma_{2p}X_{t-p} + u_{2t}, \quad (17.2)$$

where the β's and the γ's are unknown coefficients and u_{1t} and u_{2t} are error terms.

The VAR assumptions are the time series regression assumptions of Key Concept 15.6 applied to each equation. The coefficients of a VAR are estimated by estimating each equation by OLS.

approach is to develop a separate forecasting model for each variable, using the methods of Section 15.4. Another approach is to develop a single model that can forecast all the variables, which can help to make the forecasts mutually consistent. One way to forecast several variables with a single model is to use a vector autoregression (VAR). A VAR extends the univariate autoregression to multiple time series variables; that is, it extends the univariate autoregression to a "vector" of time series variables.

The VAR Model

A **vector autoregression (VAR)** with two time series variables, Y_t and X_t, consists of two equations: In one, the dependent variable is Y_t; in the other, the dependent variable is X_t. The regressors in both equations are lagged values of both variables. More generally, a VAR with k time series variables consists of k equations, one for each of the variables, where the regressors in all equations are lagged values of all the variables. The coefficients of the VAR are estimated by estimating each of the equations by ordinary least squares (OLS).

VARs are summarized in Key Concept 17.1.

Inference in VARs. Under the VAR assumptions, the OLS estimators are consistent and have a joint normal distribution in large samples. Accordingly, statistical

inference proceeds in the usual manner; for example, 95% confidence intervals on coefficients can be constructed as the estimated coefficient ± 1.96 standard errors.

One new aspect of hypothesis testing arises in VARs because a VAR with k variables is a collection, or system, of k equations. Thus it is possible to test joint hypotheses that involve restrictions across multiple equations.

For example, in the two-variable VAR(p) in Equations (17.1) and (17.2), you could ask whether the correct lag length is p or $p - 1$; that is, you could ask whether the coefficients on Y_{t-p} and X_{t-p} are 0 in these two equations. The null hypothesis that these coefficients are 0 is

$$H_0: \beta_{1p} = 0, \ \beta_{2p} = 0, \ \gamma_{1p} = 0, \text{ and } \gamma_{2p} = 0. \tag{17.3}$$

The alternative hypothesis is that at least one of these four coefficients is nonzero. Thus the null hypothesis involves coefficients from *both* of the equations, two from each equation.

Because the estimated coefficients have a jointly normal distribution in large samples, it is possible to test restrictions on these coefficients by computing an F-statistic. The precise formula for this statistic is complicated because the notation must handle multiple equations, so we omit it. In practice, most modern software packages have automated procedures for testing hypotheses on coefficients in systems of multiple equations.

How many variables should be included in a VAR? The number of coefficients in each equation of a VAR is proportional to the number of variables in the VAR. For example, a VAR with 5 variables and 4 lags will have 21 coefficients (4 lags each of 5 variables, plus the intercept) in each of the 5 equations, for a total of 105 coefficients! As discussed in Section 14.2, estimating all these coefficients by OLS increases the amount of estimation error entering a forecast, which can result in deterioration of the accuracy of the forecast as measured by the mean squared forecast error (MSFE). If the VAR coefficients are estimated by OLS, the number of coefficients therefore should be small relative to the sample size, so the number of VAR variables should be few.

In this section, we consider small VARs with coefficients estimated by OLS. Because a small VAR has only a handful of variables, those variables should be chosen with care. One guideline is to make sure the variables are plausibly related to each other so that they will be useful for forecasting one another. For example, we know from a combination of empirical evidence (such as that discussed in Chapter 15) and economic theory that the growth rate of GDP, the term spread, and the rate of inflation are related to one another, suggesting that these variables could help forecast one another in a VAR. Including an unrelated variable in a VAR, however, introduces estimation error without adding predictive content, thereby reducing forecast accuracy.

An alternative approach is to use many variables but to use methods other than OLS. We take up forecasting with many predictors in Section 17.6.

Determining lag lengths in VARs. Lag lengths in a VAR can be determined using either *F*-tests or information criteria.

The information criterion for a system of equations extends the single-equation information criterion in Section 15.5. To define this information criterion, we need to adopt matrix notation (reviewed in Appendix 19.1). Let Σ_u be the $k \times k$ covariance matrix of the VAR errors, and let $\hat{\Sigma}_u$ be the estimate of the covariance matrix, where the i,j element of $\hat{\Sigma}_u$ is $\frac{1}{T}\sum_{t=1}^{T} \hat{u}_{it}\hat{u}_{jt}$, where $\hat{u}_{it}$ is the OLS residual from the ith equation and $\hat{u}_{jt}$ is the OLS residual from the jth equation. The Bayes information criterion (BIC) for the VAR is

$$\text{BIC}(p) = \ln[\det(\hat{\Sigma}_u)] + k(kp + 1)\frac{\ln(T)}{T}, \tag{17.4}$$

where $\det(\hat{\Sigma}_u)$ is the determinant of the matrix $\hat{\Sigma}_u$. The Akaike information criterion (AIC) is computed using Equation (17.4), modified by replacing the term $\ln(T)$ with 2.

The expression for the BIC for the k equations in the VAR in Equation (17.4) extends the expression for a single equation given in Section 15.5. When there is a single equation, the first term simplifies to $\ln[SSR(p)/T]$. The second term in Equation (17.4) is the penalty for adding additional regressors; $k(kp + 1)$ is the total number of regression coefficients in the VAR. (There are k equations, each of which has an intercept and p lags of each of the k time series variables.)

Lag length estimation in a VAR using the BIC proceeds analogously to the single-equation case: Among a set of candidate values of p, the estimated lag length $\hat{p}$ is the value of p that minimizes $\text{BIC}(p)$.

Using VARs for causal analysis. The discussion so far has focused on using VARs for forecasting. Another use of VAR models is for analyzing causal relationships among economic time series variables; indeed, it was for this purpose that VARs were first introduced to economics by the econometrician and macroeconomist Christopher Sims (1980). (See the box "Nobel Laureates in Time Series Econometrics.") The use of VARs for causal inference is known as *structural VAR modeling—structural* because in this application VARs are used to model the underlying structure of the economy. Structural VAR analysis uses the techniques introduced in this section in the context of forecasting plus some additional tools. The biggest conceptual difference between using VARs for forecasting and using them for structural modeling, however, is that structural modeling requires very specific assumptions, derived from economic theory and institutional knowledge, of what is exogenous and what is not. The discussion of structural VARs is best undertaken in the context of estimation of systems of simultaneous equations, which goes beyond the scope of this book. For an introduction to using VARs for forecasting and policy analysis, see Stock and Watson (2001). For a graduate textbook treatment of structural VAR modeling, see Kilian and Lütkepohl (2017).

A VAR Model of the Growth Rate of GDP and the Term Spread

As an illustration, consider a two-variable VAR for the growth rate of GDP, $GDPGR_t$, and the term spread, $TSpread_t$. The VAR for $GDPGR_t$ and $TSpread_t$ consists of two equations: one in which $GDPGR_t$ is the dependent variable and one in which $TSpread_t$ is the dependent variable. The regressors in both equations are lagged values of $GDPGR_t$ and $TSpread_t$. Because of the apparent break in the relation in the early 1980s found in Section 15.7 using the Quandt likelihood ratio (QLR) test, the VAR is estimated using data from 1981:Q1 to 2017:Q3.

The first equation of the VAR is the GDP growth rate equation:

$$\widehat{GDPGR_t} = 0.54 + 0.29\,GDPGR_{t-1} + 0.20\,GDPGR_{t-2}$$
$$\quad\;\; (0.50)\;\; (0.11) \qquad\qquad (0.08)$$
$$\qquad\;\; -0.86\,TSpread_{t-1} + 1.18\,TSpread_{t-2}.$$
$$\qquad\quad (0.35) \qquad\qquad (0.39)$$

(17.5)

The adjusted R^2 is $\overline{R}^2 = 0.27$.

The second equation of the VAR is the term spread equation, in which the regressors are the same as in the $GDPGR$ equation but the dependent variable is the term spread:

$$\widehat{TSpread_t} = 0.44 + 0.01\,GDPGR_{t-1} - 0.05\,GDPGR_{t-2}$$
$$\quad\;\; (0.12)\;\; (0.02) \qquad\qquad (0.03)$$
$$\qquad\;\; + 1.06\,TSpread_{t-1} - 0.22\,TSpread_{t-2}.$$
$$\qquad\quad (0.10) \qquad\qquad (0.11)$$

(17.6)

The adjusted R^2 is $\overline{R}^2 = 0.82$.

Equations (17.5) and (17.6), taken together, are a VAR(2) model of the growth rate of GDP, $GDPGR_t$, and the term spread, $TSpread_t$.

These VAR equations can be used to perform tests of predictability. The F-statistic testing the null hypothesis that the coefficients on $TSpread_{t-1}$ and $TSpread_{t-2}$ are 0 in the GDP growth rate equation [Equation (17.5)] is 5.60, which has a p-value less than 0.001. Thus the null hypothesis is rejected, so we can conclude that the term spread is a useful predictor of the growth rate of GDP, given lags in the growth rate of GDP. The F-statistic testing the hypothesis that the coefficients on the two lags of $GDPGR_t$ are zero in the term spread equation [Equation (17.6)] is 3.22, which has a p-value of 0.04. Thus the growth rate of GDP helps predict the term spread at the 5% significance level.

Forecasts of the growth rate of GDP and the term spread one period ahead are obtained exactly as discussed in Section 15.4. The forecast of the growth rate of GDP for 2017:Q4, based on Equation (17.5), is $\widehat{GDP}_{2017:Q4|2017:Q3} = 2.8\%$. A similar calculation using Equation (17.6) gives a forecast of the term spread for 2017:Q4, based on data through 2017:Q3, of $\widehat{TSpread}_{2017:Q4|2017:Q3} = 1.3$ percentage points. The actual values for 2017:Q4 are $GDPGR_{2017:Q4} = 2.5\%$ and $TSpread_{2017:Q4} = 1.2$ percentage points.

17.2 Multi-period Forecasts

The discussion of forecasting so far has focused on making forecasts one period in advance. Often, however, forecasters are called upon to make forecasts further into the future. This section describes two methods for making multi-period forecasts, which are also called multi-step forecasts. The first method is to construct iterated forecasts, in which a one-period ahead model is iterated forward one period at a time in a way that is made precise in this section. The second method is to make direct forecasts by using a regression in which the dependent variable is the multi-period variable that one wants to forecast. For reasons discussed at the end of this section, in most applications the iterated method is recommended over the direct method.

Iterated Multi-period Forecasts

The essential idea of an iterated forecast is that a forecasting model is used to make a forecast one period ahead, for period $T + 1$, using data through period T. The model then is used to make a forecast for date $T + 2$, given the data through date T, where the forecasted value for date $T + 1$ is treated as data for the purpose of making the forecast for period $T + 2$. Thus the one-period ahead forecast (which is also referred to as a one-step ahead forecast) is used as an intermediate step to make the two-period ahead forecast. This process repeats, or iterates, until the forecast is made for the desired forecast horizon h.

The iterated AR forecast method: AR(1). An iterated $AR(1)$ forecast uses an $AR(1)$ for the one-period ahead model. For example, consider the first-order autoregression for $GDPGR$ [Equation (15.9)]:

$$\widehat{GDPGR}_t = 1.95 + 0.34\ GDPGR_{t-1}.$$
$$(0.32)\quad(0.07) \tag{17.7}$$

The first step in computing the two-quarter ahead forecast of $GDPGR_{2018:Q1}$ based on Equation (17.7) and using data through 2017:Q3 is to compute the one-quarter ahead forecast of $GDPGR_{2017:Q4}$ based on data through 2017:Q3: $\widehat{GDPGR}_{2017:Q4|2017:Q3} = 1.95 + 0.34\,GDPGR_{2017:Q3} = 1.95 + 0.34 \times 3.11 = 3.0$. The second step is to substitute this forecast into Equation (17.7), so that $\widehat{GDPGR}_{2018:Q1|2017:Q3} = 1.95 + 0.34\ GDPGR_{2017:Q4|2017:Q3} = 1.95 + 0.34 \times 3.0 = 3.0$. Thus, based on information through the third quarter of 2017, this forecast states that the growth rate of GDP will be 3.0% in the first quarter of 2018.

The iterated AR forecast method: AR(p). The iterated $AR(1)$ strategy is extended to an $AR(p)$ by replacing Y_{T+1} with its forecast, $\hat{Y}_{T+1|T}$, and then treating that forecast as data for the $AR(p)$ forecast of Y_{T+2}. For example, consider the iterated two-period

ahead forecast of the growth rate of GDP based on the AR(2) model from Section 15.3 [Equation (15.11)]:

$$\widehat{GDPGR}_t = 1.60 + 0.28\,GDPGR_{t-1} + 0.18\,GDPGR_{t-2}.$$
$$\phantom{\widehat{GDPGR}_t = 1.60}\ (0.37)\ \ (0.08)\phantom{\,GDPGR_{t-1} + }\ (0.08) \tag{17.8}$$

The forecast of $GDPGR_{2017:Q4}$ based on data through 2017:Q3 using this AR(2), computed in Section 15.3, is $\widehat{GDPGR}_{2017:Q4|2017:Q3} = 3.0$. Thus the two-quarter ahead iterated forecast based on the AR(2) is $\widehat{GDPGR}_{2018:Q1|2017:Q3} = 1.60 + 0.28$ $\widehat{GDPGR}_{2017:Q4|2017:Q3} + 0.18\ GDPGR_{2017:Q3} = 1.60 + 0.28 \times 3.0 + 0.18 \times 3.1 = 3.0$. According to this iterated AR(2) forecast, based on data through the third quarter of 2017, the growth rate of GDP is predicted to be 3.0% in the first quarter of 2018.

Iterated multivariate forecasts using an iterated VAR. Iterated multivariate forecasts can be computed using a VAR in much the same way as iterated univariate forecasts are computed using an autoregression. The main new feature of an iterated multivariate forecast is that the two-step ahead (period $T + 2$) forecast of one variable depends on the forecasts of all variables in the VAR in period $T + 1$. For example, to compute the forecast of the growth rate of GDP in period $T + 2$ using a VAR with the variables $GDPGR_t$ and $TSpread_t$, one must forecast both $GDPGR_{T+1}$ and $TSpread_{T+1}$, using data through period T as an intermediate step in forecasting $GDPGR_{T+2}$. More generally, to compute multi-period iterated VAR forecasts h periods ahead, it is necessary to compute forecasts of all variables for all intervening periods between T and $T + h$.

As an example, we will compute the iterated VAR forecast of $GDPGR_{2018:Q1}$ based on data through 2017:Q3, using the VAR(2) for $GDPGR_t$ and $TSpread_t$ in Section 17.1 [Equations (17.5) and (17.6)]. The first step is to compute the one-quarter ahead forecasts $\widehat{GDPGR}_{2017:Q4|2017:Q3}$ and $\widehat{TSpread}_{2017:Q4|2017:Q3}$ from that VAR. These one-period ahead forecasts were computed in Section 17.1 based on Equations (17.5) and (17.6). The forecasts were $\widehat{GDPGR}_{2017:Q4|2017:Q3} = 2.8$ and $\widehat{TSpread}_{2017:Q4|2017:Q3} = 1.3$. In the second step, these forecasts are substituted into Equations (17.5) and (17.6) to produce the two-quarter ahead forecast:

$$\begin{aligned}
\widehat{GDPGR}_{2018:Q1|2017:Q3} = {} & 0.54 + 0.29\ \widehat{GDPGR}_{2017:Q4|2017:Q3} + 0.20\,GDPGR_{2017:Q3} \\
& -0.86\ \widehat{TSpread}_{2017:Q4|2017:Q3} + 1.28\,TSpread_{2017:Q3} \\
= {} & 0.54 + 0.29 \times 2.8 + 0.20 \times 3.1 \\
& -0.86 \times 1.3 + 1.28 \times 1.2 = 2.4.
\end{aligned} \tag{17.9}$$

Thus the iterated VAR(2) forecast, based on data through the third quarter of 2017, is that the growth rate of GDP will be 2.4% in the first quarter of 2018.

Iterated multi-period forecasts are summarized in Key Concept 17.2.

Iterated Multi-period Forecasts

The **iterated multi-period AR forecast** is computed in steps: First compute the one-period ahead forecast, and then use that to compute the two-period ahead forecast, and so forth. The two- and three-period ahead iterated forecasts based on an $AR(p)$ are

$$\hat{Y}_{T+2|T} = \hat{\beta}_0 + \hat{\beta}_1\hat{Y}_{T+1|T} + \hat{\beta}_2 Y_T + \hat{\beta}_3 Y_{T+1} + \cdots + \hat{\beta}_p Y_{T-p+2} \quad (17.10)$$

$$\hat{Y}_{T+3|T} = \hat{\beta}_0 + \hat{\beta}_1\hat{Y}_{T+2|T} + \hat{\beta}_2\hat{Y}_{T+1|T} + \hat{\beta}_3 Y_T + \cdots +\hat{\beta}_p Y_{T-p+3}, \quad (17.11)$$

where the $\hat{\beta}$'s are the OLS estimates of the $AR(p)$ coefficients. Continuing this process (iterating) produces forecasts further into the future.

The **iterated multi-period VAR forecast** is also computed in steps: First compute the one-period ahead forecast of all the variables in the VAR, then use those forecasts to compute the two-period ahead forecasts, and continue this process iteratively to the desired forecast horizon. The two-period ahead iterated forecast of Y_{T+2}, based on the two-variable $VAR(p)$ in Key Concept 17.1, is

$$\begin{aligned} \hat{Y}_{T+2|T} = \; & \hat{\beta}_{10} + \hat{\beta}_{11}\hat{Y}_{T+1|T} + \hat{\beta}_{12} Y_T + \hat{\beta}_{13} Y_{T-1} + \cdots + \hat{\beta}_{1p} Y_{T-p+2} \\ & + \hat{\gamma}_{11}\hat{X}_{T+1|T} + \hat{\gamma}_{12} X_T + \hat{\gamma}_{13} X_{T-1} + \cdots + \hat{\gamma}_{1p} X_{T-p+2}, \end{aligned} \quad (17.12)$$

where the coefficients in Equation (17.12) are the OLS estimates of the VAR coefficients. Iterating produces forecasts further into the future.

Direct Multi-period Forecasts

Direct multi-period forecasts are computed without iterating by using a single regression, in which the dependent variable is the multi-period ahead variable to be forecasted and the regressors are the predictor variables. Forecasts computed this way are called direct forecasts because the regression coefficients can be used directly to make the multi-period forecast.

The direct multi-period forecasting method. Suppose that you want to make a forecast of Y_{T+2} using data through time T. The direct multivariate method takes the ADL model as its starting point but lags the predictor variables by an additional time period. For example, if two lags of the predictors are used, then the dependent variable is Y_t, and the regressors are Y_{t-2}, Y_{t-3}, X_{t-2}, and X_{t-3}. The coefficients from this regression can be used directly to compute the forecast of Y_{T+2} using data on Y_T, Y_{T-1}, X_T, and X_{T-1}, without the need for any iteration. More generally, in a direct h-period ahead forecasting regression, all predictors are lagged h periods to produce the h-period ahead forecast.

For example, the forecast of $GDPGR_t$ two quarters ahead using two lags each of $GDPGR_{t-2}$ and $TSpread_{t-2}$ is computed by first estimating the regression:

$$
\begin{aligned}
\widehat{GDPGR}_{t|t-2} = {}& 0.56 + 0.31 GDPGR_{t-2} + 0.04 GDPGR_{t-3} \\
& (0.63) \ (0.07) \qquad\qquad (0.09) \\
& + 0.56 TSpread_{t-2} + 0.04 TSpread_{t-3}. \\
& \ \ (0.46) \qquad\qquad (0.45)
\end{aligned}
\tag{17.13}
$$

The two-quarter ahead forecast of the growth rate of GDP in 2018:Q1 based on data through 2017:Q3 is computed by substituting the values of $GDPGR_{2017:Q3}$, $GDPGR_{2017:Q2}$, $TSpread_{2017:Q3}$, and $TSpread_{2017:Q2}$ into Equation (17.13); this yields

$$
\begin{aligned}
\widehat{GDPGR}_{2018:Q1|2017:Q3} = {}& 0.56 + 0.31 GDPGR_{2017:Q3} + 0.04 GDPGR_{2017:Q2} \\
& + 0.56 TSpread_{2017:Q3} + 0.04 TSpread_{2017:Q2} = 2.4.
\end{aligned}
\tag{17.14}
$$

The three-quarter ahead direct forecast of $GDPGR_{T+3}$ is computed by lagging all the regressors in Equation (17.13) by one additional quarter, estimating that regression, and then computing the forecast. The h-quarter ahead direct forecast of $GDPGR_{T+h}$ is computed by using $GPDGR_t$ as the dependent variable and the regressors $GPDGR_{t-h}$ and $TSpread_{t-h}$ plus additional lags of $GPDGR_{t-h}$ and $TSpread_{t-h}$, as desired.

Standard errors in direct multi-period regressions.   Because the dependent variable in a multi-period regression occurs two or more periods into the future, the error term in a multi-period regression is serially correlated. To see this, consider the two-period ahead forecast of the GDP growth rate, and suppose that a surprise jump in oil prices occurs in the next quarter. Today's two-period ahead forecast of the growth rate of GDP will be too high because it does not incorporate this unexpected negative event. Because the oil price rise was also unknown in the previous quarter, the two-period ahead forecast made last quarter will also be too high. Thus the surprise oil price jump next quarter means that *both* last quarter's and this quarter's two-period ahead forecasts are too high. Because of such intervening events, the error term in a multi-period regression is serially correlated.

As discussed in Section 16.4, if the error term is serially correlated, the usual OLS standard errors are incorrect, or, more precisely, they are not a reliable basis for inference. Therefore, heteroskedasticity- and autocorrelation-consistent (HAC) standard errors must be used with direct multi-period regressions. The standard errors reported in Equation (17.13) for direct multi-period regressions therefore are Newey–West HAC standard errors, where the truncation parameter m is set according to Equation (16.17); for these data (for which $T = 147$), Equation (16.17) yields $m = 4$. For longer forecast horizons, the amount of overlap—and thus the degree of serial correlation in the error—increases: In general, the first $h - 1$ autocorrelation coefficients of the errors in an h-period ahead regression are nonzero. Thus larger values of m than indicated by Equation (16.17) are appropriate for multi-period regressions with long forecast horizons.

Direct multi-period forecasts are summarized in Key Concept 17.3.

Direct Multi-period Forecasts

The **direct multi-period forecast** h periods into the future based on p lags each of Y_t and an additional predictor X_t is computed by first estimating the regression

$$Y_t = \delta_0 + \delta_1 Y_{t-h} + \cdots + \delta_p Y_{t-p-h+1} + \delta_{p+1} X_{t-h} \\ + \cdots + \delta_{2p} X_{t-p-h+1} + u_t \tag{17.15}$$

and then using the estimated coefficients directly to make the forecast of Y_{T+h} using data through period T.

Which Method Should You Use?

In most applications, the iterated method is the recommended procedure for multi-period forecasting for two reasons. First, from a theoretical perspective, if the underlying one-period ahead model (the AR or VAR that is used to compute the iterated forecast) is specified correctly, then the coefficients are estimated more efficiently if they are estimated by a one-period ahead regression (and then iterated) than by a multi-period ahead regression. Second, from a practical perspective, forecasters are usually interested in forecasts not just at a single horizon but at multiple horizons. Because they are produced using the same model, iterated forecasts tend to have time paths that are less erratic across horizons than do direct forecasts. Because a different model is used at every horizon for direct forecasts, sampling error in the estimated coefficients can add random fluctuations to the time paths of a sequence of direct multi-period forecasts.

Under some circumstances, however, direct forecasts are preferable to iterated forecasts. One such circumstance is when you have reason to believe that the one-period ahead model (the AR or VAR) is not specified correctly. For example, you might believe that the equation for the variable you are trying to forecast in a VAR is specified correctly but that one or more of the other equations in the VAR are specified incorrectly, perhaps because of neglected nonlinear terms. If the one-step ahead model is specified incorrectly, then, in general, the iterated multi-period forecast will be biased, and the MSFE of the iterated forecast can exceed the MSFE of the direct forecast, even though the direct forecast has a larger variance.

17.3 Orders of Integration and the Nonnormality of Unit Root Test Statistics

This section extends the treatment of stochastic trends in Section 15.6 by addressing two further topics. First, the trends of some time series are not well described by the random walk model, so we introduce an extension of that model and discuss its

implications for regression modeling of such series. Next we discuss the reason for the nonnormal distribution of the ADF test for a unit root.

Other Models of Trends and Orders of Integration

Recall that the random walk model for a trend, introduced in Section 15.6, specifies that the trend at date t equals the trend at date $t - 1$ plus a random error term. If Y_t follows a random walk with drift β_0, then

$$Y_t = \beta_0 + Y_{t-1} + u_t, \tag{17.16}$$

where u_t is serially uncorrelated. Also recall from Section 15.6 that, if a series has a random walk trend, then it has an autoregressive root that equals 1.

Although the random walk model of a trend describes the long-run movements of many economic time series, some economic time series have trends that are smoother—that is, that vary less from one period to the next—than is implied by Equation (17.16). A different model is needed to describe the trends of such series.

One model of a smooth trend makes the first difference of the trend follow a random walk; that is,

$$\Delta Y_t = \beta_0 + \Delta Y_{t-1} + u_t, \tag{17.17}$$

where u_t is serially uncorrelated. Thus, if Y_t follows Equation (17.17), ΔY_t follows a random walk, so $\Delta Y_t - \Delta Y_{t-1}$ is stationary. The difference of the first differences, $\Delta Y_t - \Delta Y_{t-1}$, is called the **second difference** of Y_t and is denoted $\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1}$. In this terminology, if Y_t follows Equation (17.17), then its second difference is stationary. If a series has a trend of the form in Equation (17.17), then the first difference of the series has an autoregressive root that equals 1.

Orders of integration terminology.   Some additional terminology is useful for distinguishing between these two models of trends. A series that has a random walk trend is said to be **integrated of order one**, or $I(1)$. A series that has a trend of the form in Equation (17.17) is said to be **integrated of order two**, or $I(2)$. A series that does not have a stochastic trend and is stationary is said to be **integrated of order zero**, or $I(0)$.

The **order of integration** in the $I(1)$ and $I(2)$ terminology is the number of times that the series needs to be differenced for it to be stationary: If Y_t is $I(1)$, then the first difference of Y_t, ΔY_t, is stationary, and if Y_t is $I(2)$, then the second difference of Y_t, $\Delta^2 Y_t$, is stationary. If Y_t is $I(0)$, then Y_t is stationary.

Orders of integration are summarized in Key Concept 17.4.

KEY CONCEPT

17.4

Orders of Integration, Differencing, and Stationarity

- If Y_t is integrated of order one—that is, if Y_t is $I(1)$—then Y_t has a unit autoregressive root, and its first difference, ΔY_t, is stationary.

- If Y_t is integrated of order two—that is, if Y_t is $I(2)$—then ΔY_t has a unit autoregressive root, and its second difference, $\Delta^2 Y_t$, is stationary.

- If Y_t is **integrated of order d**—that is, if Y_t is $\boldsymbol{I(d)}$—then Y_t must be differenced d times to eliminate its stochastic trend; that is, $\Delta^d Y_t$ is stationary.

How to test whether a series is I(2) or I(1). If Y_t is $I(2)$, then ΔY_t is $I(1)$, so ΔY_t has an autoregressive root that equals 1. If, however, Y_t is $I(1)$, then ΔY_t is stationary. Thus the null hypothesis that Y_t is $I(2)$ can be tested against the alternative hypothesis that Y_t is $I(1)$ by testing whether ΔY_t has a unit autoregressive root. If the hypothesis that ΔY_t has a unit autoregressive root is rejected, then the hypothesis that Y_t is $I(2)$ is rejected in favor of the alternative that Y_t is $I(1)$.

Examples of I(2) and I(1) series: The price level and the rate of inflation. The rate of inflation is the growth rate of the price level. Recall from Section 15.2 that the growth rate of a time series X_t can be computed as the first difference of the logarithm of X_t; that is, $\Delta \ln(X_t)$ is the growth rate of X_t (expressed as fraction). If P_t is a time series for the price level measured quarterly, then $\Delta \ln(P_t)$ is its growth rate, and $Infl_t = 400 \times \Delta \ln(P_t)$ is the quarterly rate of inflation, measured in percentage points at an annual rate. As in the expression for the growth of GDP, $GDPGR$ in Equation (15.1), the factor 400 arises from converting fractional changes to percentage changes (multiplying by 100) and converting quarterly percentages to an annual rate (multiplying by 4).

In Empirical Exercise 15.1, you analyzed the inflation rate, $Infl_t$, computed using the price index for personal consumption expenditures in the United States as P_t. In that exercise, you concluded that the rate of inflation in the United States plausibly has a random walk stochastic trend—that is, that the rate of inflation is $I(1)$. If inflation is $I(1)$, then its stochastic trend is removed by first differencing, so $\Delta Infl_t$ is stationary. But treating inflation as $I(1)$ is equivalent to treating $\Delta \ln(P_t)$ as $I(1)$, and this in turn is equivalent to treating the logarithm of the price level, $\ln(P_t)$, as $I(2)$.

The logarithm of the price level and the rate of inflation are plotted in Figure 17.1. The long-run trend of the logarithm of the price level (Figure 17.1a) varies more smoothly than the long-run trend in the rate of inflation (Figure 17.1b). The smooth trend in the logarithm of the price level is typical of $I(2)$ series.

(a) Logarithm of the PCE price index
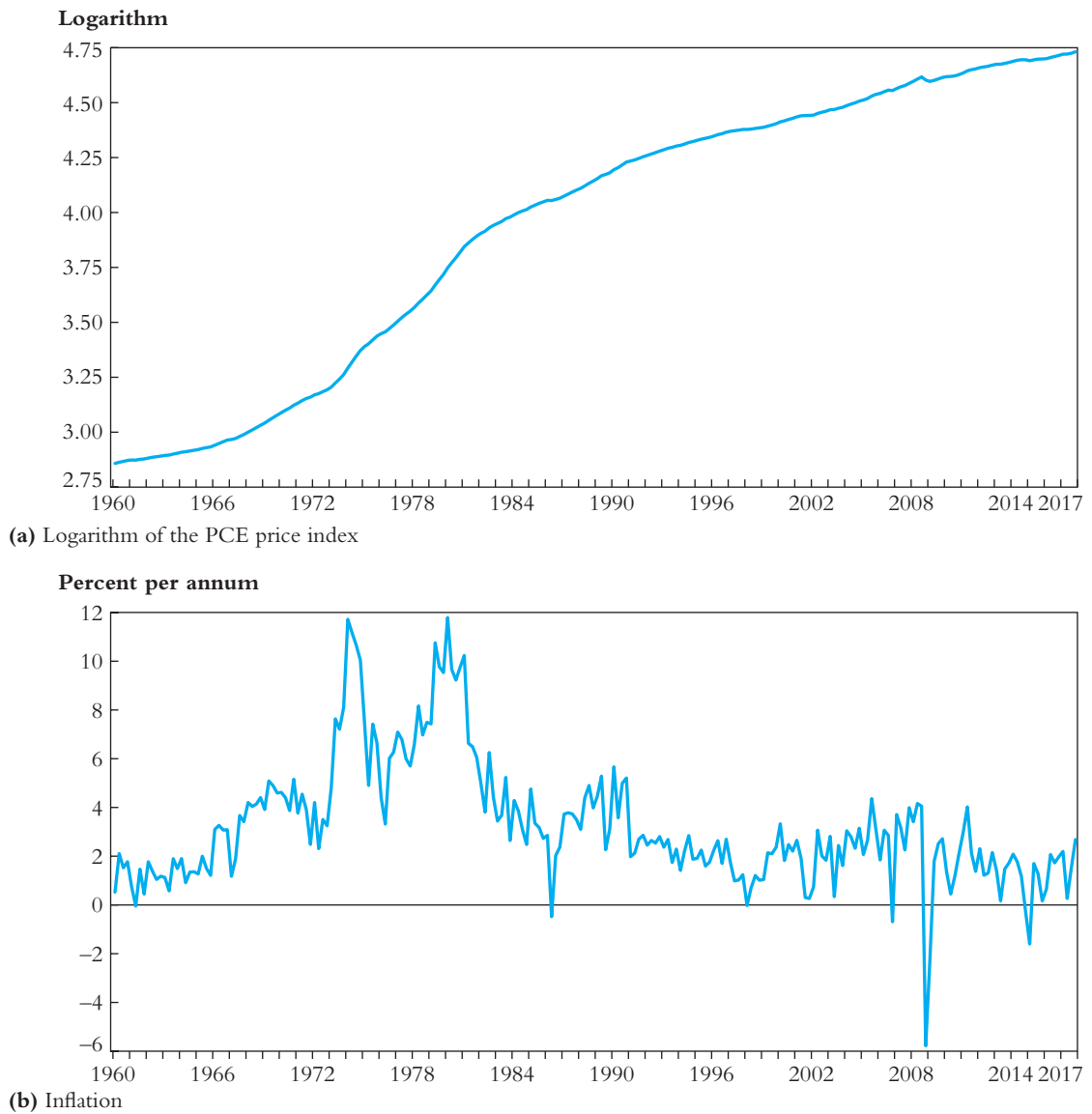
(b) Inflation

The trend in the logarithm of prices (Figure 17.1a) is much smoother than the trend in inflation (Figure 17.1b).

Why Do Unit Root Tests Have Nonnormal Distributions?

In Section 15.7, it was stressed that the large-sample normal distribution on which regression analysis relies so heavily does not apply if the regressors are nonstationary. Under the null hypothesis that the regression contains a unit root, the regressor Y_{t-1} in the Dickey–Fuller regression is nonstationary. The nonnormal distribution of the unit root test statistics is a consequence of this nonstationarity.

To gain some mathematical insight into this nonnormality, consider the simplest possible Dickey–Fuller regression, in which ΔY_t is regressed against the single regressor Y_{t-1} and the intercept is excluded. In the notation of Equation (15.32), the OLS estimator in this regression is $\hat{\delta} = \sum_{t=1}^{T} Y_{t-1} \Delta Y_t / \sum_{t=1}^{T} Y_{t-1}^2$, so

$$T\hat{\delta} = \frac{\dfrac{1}{T}\displaystyle\sum_{t=1}^{T} Y_{t-1}\Delta Y_t}{\dfrac{1}{T^2}\displaystyle\sum_{t=1}^{T} Y_{t-1}^2}. \tag{17.19}$$

Consider the numerator in Equation (17.19). Under the additional assumption that $Y_0 = 0$, a bit of algebra (Exercise 17.5) shows that

$$\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\Delta Y_t = \frac{1}{2}\left[\left(\frac{Y_T}{\sqrt{T}}\right)^2 - \frac{1}{T}\sum_{t=1}^{T}(\Delta Y_t)^2\right]. \tag{17.20}$$

Under the null hypothesis, $\Delta Y_t = u_t$, which is serially uncorrelated and has a finite variance, so the second term in Equation (17.20) has the probability limit $\frac{1}{T}\sum_{t=1}^{T}(\Delta Y_t)^2 \xrightarrow{p} \sigma_u^2$. Under the assumption that $Y_0 = 0$, the first term in Equation (17.20) can be written $Y_T/\sqrt{T} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\Delta Y_t} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}u_t}$, which in turn obeys the central limit theorem; that is, $Y_T/\sqrt{T} \xrightarrow{d} N(0, \sigma_u^2)$. Thus $(Y_T/\sqrt{T})^2 - \frac{1}{T}\sum_{t=1}^{T}(\Delta Y_t)^2 \xrightarrow{d} \sigma_u^2(Z^2 - 1)$, where Z is a standard normal random variable. Recall, however, that the square of a standard normal distribution has a chi-squared distribution with 1 degree of freedom. It therefore follows from Equation (17.20) that, under the null hypothesis, the numerator in Equation (17.19) has the limiting distribution

$$\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\Delta Y_t \xrightarrow{d} \frac{\sigma_u^2}{2}\left(\chi_1^2 - 1\right). \tag{17.21}$$

The large-sample distribution in Equation (17.21) is different than the usual large-sample normal distribution when the regressor is stationary. Instead, the numerator of the OLS estimator of the coefficient on Y_t in this Dickey–Fuller regression has a distribution that is proportional to a chi-squared distribution with 1 degree of freedom minus 1.

This discussion has considered only the numerator of $T\hat{\delta}$. The denominator also behaves unusually under the null hypothesis: Because Y_t follows a random walk under the null hypothesis, $\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}^2$ does not converge in probability to a constant. Instead, the denominator in Equation (17.19) is a random variable, even in large samples: Under the null hypothesis, $\frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}^2$ converges in distribution jointly with the numerator. The unusual joint distribution of the numerator and denominator in Equation (17.19) are the source of the nonstandard distribution of the Dickey–Fuller test statistic and the reason that the ADF statistic has its own special table of critical values.

17.4 Cointegration

Sometimes two or more series share the same stochastic trend. In this special case, referred to as cointegration, regression analysis can reveal long-run relationships among time series variables, but some new methods are needed.

Cointegration and Error Correction

Two or more time series with stochastic trends can move together so closely over the long run that they appear to have the same trend component; that is, they appear to have a **common trend**. For example, the 90-day and 10-year U.S. Treasury interest rates in Figure 15.3 exhibit the same long-run tendencies or trends: Both were low in the 1960s, both rose through the 1970s to peaks in the early 1980s, and then both fell through the 1990s. However, the difference between the long-term and short-term interest rates, the term spread shown in Figure 15.3b, does not appear to have a trend. That is, subtracting the short-term rate from the long-term rate appears to elimi-nate the trends in both of the individual rates. Said differently, although the two interest rates differ, they appear to share a common stochastic trend: Because the trend in each individual series is eliminated by subtracting one series from the other, the two series must have the same trend; that is, they must have a common stochastic trend.

Two or more series that have a common stochastic trend are said to be cointe-grated. The formal definition of **cointegration** (due to the econometrician Clive Granger; see the box "Nobel Laureates in Time Series Econometrics") is given in Key Concept 17.5. In this section, we introduce a test for whether cointegration is present, discuss estimation of the coefficients of regressions relating cointegrated variables, and illustrate the use of the cointegrating relationship for forecasting. The discussion initially focuses on the case that there are only two variables, X_t and Y_t.

Vector error correction model. If X_t and Y_t are cointegrated, the first differences of X_t and Y_t can be modeled using a VAR, augmented by including $Y_{t-1} - \theta X_{t-1}$ as an additional regressor:

$$\Delta Y_t = \beta_{10} + \beta_{11}\Delta Y_{t-1} + \cdots + \beta_{1p}\Delta Y_{t-p} + \gamma_{11}\Delta X_{t-1} \\ + \cdots + \gamma_{1p}\Delta X_{t-p} + \alpha_1(Y_{t-1} - \theta X_{t-1}) + u_{1t} \tag{17.22}$$

$$\Delta X_t = \beta_{20} + \beta_{21}\Delta Y_{t-1} + \cdots + \beta_{2p}\Delta Y_{t-p} + \gamma_{21}\Delta X_{t-1} \\ + \cdots + \gamma_{2p}\Delta X_{t-p} + \alpha_2(Y_{t-1} - \theta X_{t-1}) + u_{2t}. \tag{17.23}$$

The term $Y_t - \theta X_t$ is called the **error correction term**: if the two variables are far apart, by virtue of their sharing a trend, one would expect the variables to get closer together over time, so that the "error" $Y_t - \theta X_t$ will be "corrected."

The combined model in Equations (17.22) and (17.23) is called a **vector error correction model (VECM)**. In a VECM, past values of $Y_t - \theta X_t$ help to predict future values of ΔY_t and/or ΔX_t.

Cointegration

Suppose that X_t and Y_t are integrated of order one. If, for some coefficient θ, $Y_t - \theta X_t$ is integrated of order zero, then X_t and Y_t are said to be *cointegrated*. The coefficient θ is called the **cointegrating coefficient**.

If X_t and Y_t are cointegrated, then they have the same, or a common, stochastic trend. Computing the difference $Y_t - \theta X_t$ eliminates this common stochastic trend.

How Can You Tell Whether Two Variables Are Cointegrated?

There are three ways to determine whether two variables can plausibly be modeled as cointegrated: You can use expert knowledge and economic theory, graph the series and see whether they appear to have a common stochastic trend, and perform statistical tests for cointegration. In practice, you should use all three methods.

For example, the two interest rates in Figure 15.3 are linked together by the so-called expectations theory of the term structure of interest rates, which holds that the 10-year Treasury bond rate is the average of the sequence of expected interest rates on 3-month Treasury bills over the 10-year life of the bond. Thus, if the 3-month interest rate has a random walk stochastic trend, this theory implies that this stochastic trend is inherited by the 10-year interest rate (Exercise 17.2). Moreover, the plot of the two interest rates in Figure 15.3 shows that each of the series appears to be $I(1)$ but that the term spread appears to be $I(0)$, so it is plausible that the two series are cointegrated.

The unit root testing procedures introduced so far can be extended to tests for cointegration. The insight on which these tests are based is that if Y_t and X_t are cointegrated with cointegrating coefficient θ, then $Y_t - \theta X_t$ is stationary; otherwise, $Y_t - \theta X_t$ is nonstationary—that is, $I(1)$. The hypothesis that Y_t and X_t are not cointegrated—that is, that $Y_t - \theta X_t$ is $I(1)$—therefore can be tested by testing the null hypothesis that $Y_t - \theta X_t$ has a unit root; if this hypothesis is rejected, then Y_t and X_t can be modeled as cointegrated. The details of this test depend on whether the cointegrating coefficient θ is known.

Testing for cointegration when θ is known. In many cases, expert knowledge or economic theory suggests a value for θ. When θ is known, the ADF unit root tests can be used to test for cointegration by first constructing the series $z_t = Y_t - \theta X_t$ and then testing the null hypothesis that z_t has a unit autoregressive root.

As an illustration, applying the ADF test to the term spread (the difference between the 10-year and 90-day Treasury rates) from 1962 to 2017, with an intercept and (AIC-determined) six lags, yields an ADF statistic of -4.13. This value is less

TABLE 17.1 Critical Values for the Engle–Granger ADF Statistic			
Number of X's in Equation (17.24)	**10%**	**5%**	**1%**
1	−3.12	−3.41	−3.96
2	−3.52	−3.80	−4.36
3	−3.84	−4.16	−4.73
4	−4.20	−4.49	−5.07

than −3.43 from Table 15.4, so the null hypothesis of no cointegration (a unit root in the term spread) is rejected at the 1% significance level.

Testing for cointegration when θ is unknown.   If the cointegrating coefficient θ is unknown, then it must be estimated prior to testing for a unit root in the error correction term. This preliminary step makes it necessary to use different critical values for the subsequent unit root test.

Specifically, in the first step the cointegrating coefficient θ is estimated by OLS estimation of the regression

$$Y_t = \alpha + \theta X_t + z_t. \tag{17.24}$$

In the second step, a Dickey–Fuller t-test (with an intercept but no time trend) is used to test for a unit root in the residual from this regression, $\hat{z}_t$. This two-step procedure is called the Engle–Granger Augmented Dickey–Fuller test for cointegration, or **EG-ADF test** (Engle and Granger 1987).

Critical values of the EG-ADF statistic are given in Table 17.1.[1] The critical values in the first row apply when there is a single regressor in Equation (17.26), so there are two cointegrated variables (X_t and Y_t). The subsequent rows apply to the case of multiple cointegrated variables, which is discussed at the end of this section.

Estimation of Cointegrating Coefficients

If X_t and Y_t are cointegrated, then the OLS estimator of the coefficient in the cointegrating regression in Equation (17.24) is consistent. However, in general, the OLS estimator (like the ADF test statistic, for similar reasons) has a nonnormal distribution, and inferences based on its t-statistics can be misleading whether or not those t-statistics are computed using HAC standard errors. Because of these drawbacks of the OLS estimator of θ, econometricians have developed a number of other estimators of the cointegrating coefficient.

One such estimator of θ that is simple to use in practice is the **dynamic OLS (DOLS) estimator** (Stock and Watson 1993). The DOLS estimator is based on a

---

[1]The critical values in Table 17.1 are taken from Fuller (1976) and Phillips and Ouliaris (1990). Following a suggestion by Hansen (1992), the critical values in Table 17.1 are chosen so that they apply whether or not X_t and Y_t have drift components.

modified version of Equation (17.24) that includes past, present, and future values of the change in X_t:

$$Y_t = \beta_0 + \theta X_t + \sum_{j=-p}^{p} \delta_j \Delta X_{t-j} + u_t. \qquad (17.25)$$

Thus, in Equation (17.25), the regressors are $X_t, \Delta X_{t+p}, \ldots, \Delta X_{t-p}$. The DOLS estimator of θ is the OLS estimator of θ in the regression of Equation (17.25).

If X_t and Y_t are cointegrated, then the DOLS estimator is efficient in large samples. Moreover, statistical inferences about θ and the δ's in Equation (17.25) based on HAC standard errors are valid. For example, the t-statistic constructed using the DOLS estimator with HAC standard errors has a standard normal distribution in large samples.

As an illustration, for a DOLS regression of the 90-day Treasury rate on the 10-year Treasury rate, using the data in Figure 15.3 and $p = 4$ leads and lags, the DOLS estimate of the cointegrating coefficient is 1.02. The HAC standard error, computed using a truncation parameter of $m = 5$, is 0.05. Thus the null hypothesis that $\theta = 1$ cannot be rejected at the 10% significance level. This result, along with the finding that the term spread is stationary, is consistent with the theory of the term structure of interest rates.

Extension to Multiple Cointegrated Variables

The concepts, tests, and estimators discussed here extend to more than two variables. For example, if there are three variables, Y_t, X_{1t}, and X_{2t}, each of which is $I(1)$, then they are cointegrated with cointegrating coefficients θ_1 and θ_2 if $Y_t - \theta_1 X_{1t} - \theta_2 X_{2t}$ is stationary. When there are three or more variables, there can be multiple cointegrating relationships. For example, consider modeling the relationship among three interest rates: the three-month rate ($R3m$), the one-year ($R1y$) rate, and the ten-year rate ($R10y$). If they are $I(1)$, then the expectations theory of the term structure of interest rates suggests that they will all be cointegrated. One cointegrating relationship suggested by the theory is $R10y_t - R3m_t$, and a second relationship is $R1y_t - R3m_t$. (The relationship $R10y_t - R1y_t$ is also a cointegrating relationship, but it contains no additional information beyond that in the other relationships because it is perfectly multicollinear with the other two cointegrating relationships.)

The EG-ADF procedure for testing for a single cointegrating relationship among multiple variables is the same as for the case of two variables except that the regression in Equation (17.24) is modified so that both X_{1t} and X_{2t} are regressors; the critical values for the EG-ADF test are given in Table 17.1, where the appropriate row depends on the number of regressors in the first-stage OLS cointegrating regression. The DOLS estimator of a single cointegrating relationship among multiple X's involves including the level of each X along with leads and lags of the first difference of each X. For additional discussion of cointegration methods for multiple variables, see Hamilton (1994).

Even if economic theory does not suggest a specific value of the cointegrating coefficient, it is important to check whether the estimated cointegrating relationship

makes sense in practice. Because cointegration tests can be misleading (they can improperly reject the null hypothesis of no cointegration more frequently than they should, and frequently they improperly fail to reject the null hypothesis), it is especially important to rely on economic theory, institutional knowledge, and common sense when estimating and using cointegrating relationships.

17.5   Volatility Clustering and Autoregressive Conditional Heteroskedasticity

The phenomenon that some times are tranquil, while others are not—that is, that volatility comes in clusters—shows up in many economic time series. This section presents a pair of models for quantifying volatility clustering or, as it is also known, conditional heteroskedasticity.

Volatility Clustering

The volatility of many financial and macroeconomic variables changes over time. For example, daily percentage changes in the Wilshire 5000 Total Market Index, shown in Figure 17.2, exhibit periods of high volatility, such as in 2001 and 2008, and other periods of low volatility, such as in 2004 and 2017. A series with some periods of low volatility and some periods of high volatility is said to exhibit **volatility clustering**. Because the volatility appears in clusters, the variance of the daily percentage price

FIGURE 17.2   Daily Percentage Changes in the Wilshire 5000 Total Market Index, 1990–2017
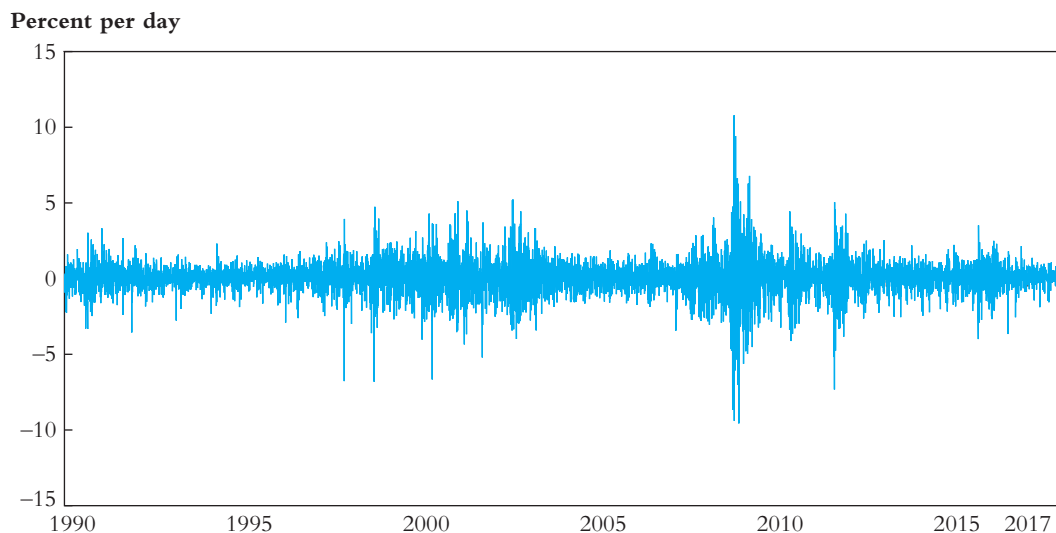
Daily percentage price changes in the Wilshire 5000 Total Market Index exhibit volatility clustering, in which there are some periods of high volatility, such as in 2008, and other periods of relative tranquility, such as in 2004.

change in the Wilshire 5000 can be forecasted, even though the daily price change itself is very difficult to forecast.

Forecasting the variance of a series is of interest for several reasons. First, the variance of price changes for some asset is a measure of the risk of owning that asset: The larger the variance of daily stock price changes, the more a stock market participant stands to gain— or lose—on a typical day. An investor who is worried about risk would be less tolerant of participating in the stock market during a period of high—rather than low—volatility.

Second, the value of some financial derivatives, such as options, depends on the variance of the underlying asset. An options trader wants the best available forecasts of future volatility to help him or her know the price at which to buy or sell options.

Third, forecasting variances can improve the accuracy of forecast intervals. Suppose that you are forecasting the rate of inflation. If the variance of the forecast error is constant, then an approximate forecast confidence interval can be constructed using the standard error of the regression or final prediction error as discussed in Section 15.5. If, however, the variance of the forecast error changes over time, then the width of the forecast interval should change over time: At periods when inflation is subject to particularly large disturbances or shocks, the interval should be wide; during periods of relative tranquility, the interval should be tighter. If the forecast error changes slowly, then the pseudo out-of-sample forecast error estimate of the MSFE in Equation (15.22) can be used, but to capture more rapid changes in volatility, such as those observed in Figure 17.2, other methods must be used.

Volatility clustering can be thought of as clustering of the variance of the error term over time: If the regression error has a small variance in one period, its variance tends to be small in the next period, too. In other words, volatility clustering implies that the error exhibits time-varying heteroskedasticity.

When data are observed at a high frequency, it is possible to measure volatility directly using a measure called realized volatility. When data are observed less frequently, it is possible to estimate a model of the volatility and use that to estimate current volatility. We address these two approaches in turn.

Realized Volatility

Suppose you have daily data on asset returns, like that shown in Figure 17.2. One way to estimate the volatility in a given month is to compute the sample variance of asset returns in that month. For asset returns measured at high frequency, the mean return is typically very small compared with the variation in the return, as is evident in Figure 17.2. For that reason, for asset returns, and more generally for series that can be measured at a high frequency, the volatility of the return is measured not by the sample variance but simply by its mean square. Accordingly, the h-period **realized volatility** of a variable X_t is the sample root mean square of X computed over h consecutive periods:

$$RV_t^h = \sqrt{\frac{1}{h}\sum_{s=t-h+1}^{t} X_s^2}.$$ (17.26)

Daily Percentage Changes in the Wilshire 5000 Total Market Index, 20-day Realized Volatility Bands, and GARCH(1, 1) Bands, 2015–2017

Percent per day
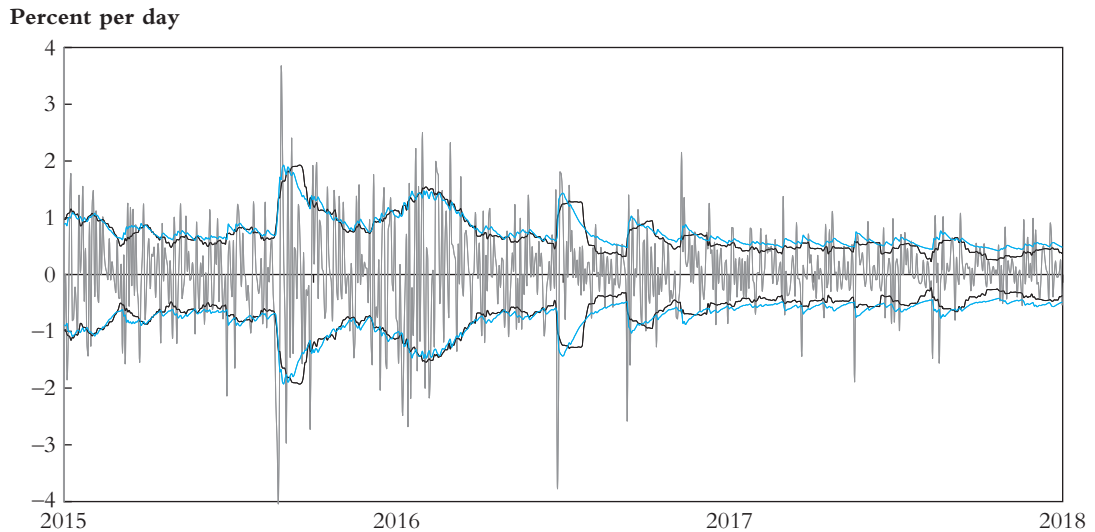
The volatility of stock price changes varies considerably over the 2015–2017 period. The volatility bands are narrow when volatility is low and wide when it is high. The 20-day realized volatility bands (black) and GARCH(1, 1) bands (dark blue) are similar to each other.

The 20-day realized volatility bands of the data in Figure 17.2 for 2015–2017 is plotted in Figure 17.3. As can be seen from the figure, the realized volatility bands provides a smooth measure of the volatility clustering evident in that figure.

In practice, realized volatility is typically computed using higher-frequency data than just daily. For example, the stock of a major company might be traded sufficiently frequently that its price can be measured at five-minute intervals. If so, these five-minute intervals can be used to compute realized volatility for a day, or even for a period of hours within a day. High-frequency realized volatility is one of the tools used in high-frequency trading.

Autoregressive Conditional Heteroskedasticity

When data are observed less frequently, an alternative is to estimate a model of the evolution of the variance over time. Two models of volatility clustering are the **autoregressive conditional heteroskedasticity (ARCH)** model and its extension, the **generalized ARCH (GARCH)** model.

ARCH.   Consider the ADL(1, 1) regression

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \gamma_1 X_{t-1} + u_t. \tag{17.27}$$

In the ARCH model, which was developed by the econometrician Robert Engle (1982; see the box "Nobel Laureates in Time Series Econometrics"), the error u_t is modeled

as being normally distributed with mean 0 and variance σ_t^2, where σ_t^2 depends on past squared values of u_t. Specifically, the ARCH model of order p, denoted ARCH(p), is

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_p u_{t-p}^2, \tag{17.28}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_p$ are unknown coefficients. If these coefficients are positive, then if recent squared errors are large, the ARCH model predicts that the current squared error will be large in magnitude in the sense that its variance, σ_t^2, is large.

Although it is described here for the ADL(1, 1) model in Equation (17.27), the ARCH model can be applied to the error variance of any time series regression model with an error that has a conditional mean of 0, including higher-order ADL models, autoregressions, and time series regressions with multiple predictors.

GARCH. The GARCH model, developed by the econometrician Tim Bollerslev (1986), extends the ARCH model to let σ_t^2 depend on its own lags as well as lags of the squared error. The GARCH(p, q) model is

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_p u_{t-p}^2 + \phi_1 \sigma_{t-1}^2 + \cdots + \phi_q \sigma_{t-q}^2, \tag{17.29}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_p, \phi_1, \ldots, \phi_q$ are unknown coefficients.

The ARCH model is analogous to a distributed lag model, and the GARCH model is analogous to an ADL model. As discussed in Chapter 16, the ADL model can provide a more parsimonious model of dynamic multipliers than can the distributed lag model. Similarly, by incorporating lags of σ_t^2, the GARCH model can capture slowly changing variances with fewer parameters than the ARCH model.

An important application of ARCH and GARCH models is to measuring and forecasting the time-varying volatility of returns on financial assets, particularly assets observed at high sampling frequencies such as the daily stock returns in Figure 17.2. In such applications, the return itself is often modeled as unpredictable, so the regression in Equation (17.27) includes only the intercept.

Estimation and inference. ARCH and GARCH models are estimated by the method of maximum likelihood (Appendix 11.2). The estimators of the ARCH and GARCH coefficients are normally distributed in large samples, so in large samples, t-statistics have standard normal distributions, and confidence intervals can be constructed as the maximum likelihood estimate ± 1.96 standard errors.

Application to Stock Price Volatility

A GARCH(1, 1) model of the Wilshire 5000 daily percentage stock price changes, R_t, estimated using data on all trading days from January 2, 1990, through December 29, 2017, is

$$\hat{R}_t = 0.063 \tag{17.30}$$
$$(0.010)$$

$$\hat{\sigma}_t^2 = 0.013 + 0.088\, u_{t-1}^2 + 0.908\sigma_{t-1}^2. \tag{17.31}$$
$$(0.002) \quad (0.008) \qquad (0.009)$$

No lagged predictors appear in Equation (17.30) because daily Wilshire 5000 percentage price changes are essentially unpredictable.

The two coefficients in the GARCH model (the coefficients on u_{t-1}^2 and σ_{t-1}^2) are both individually statistically significant at the 5% significance level. One measure of the persistence of movements in the variance is the sum of the coefficients on u_{t-1}^2 and σ_{t-1}^2 in the GARCH model (Exercise 17.9). This sum (0.99) is large, indicating that changes in the conditional variance are persistent. Said differently, the estimated GARCH model implies that periods of high volatility in stock prices will be long lasting. This implication is consistent with the long periods of volatility clustering seen in Figure 17.2.

The estimated conditional variance at date t, $\hat{\sigma}_t^2$, can be computed using the residuals from Equation (17.30) and the coefficients in Equation (17.31). For the Wilshire 5000 returns, the GARCH(1, 1) model and the 20-day realized volatility provide quantitatively similar estimates of the time-varying standard deviation of returns. This can be seen in Figure 17.3, which focuses on the 2015–2017 sample period. During the first half of 2015, the conditional standard deviation bands are relatively tight, indicating lower levels of risk for investors holding a portfolio of stocks making up the Wilshire 5000. But in the second half of 2015 these conditional standard deviations widened, indicating greater daily stook price volatility.

For these data, the realized volatility and GARCH bands are quantitatively similar to each other. An advantage of realized volatility is that it measures the changing variance without making any modeling assumptions. An advantage of the GARCH model is that it can be used to forecast volatility; another advantage is that it can be used in applications in which the data are observed infrequently—for example, monthly or quarterly. In general, realized volatility and GARCH models provide two complementary ways to quantify volatility clustering.

17.6   Forecasting with Many Predictors Using Dynamic Factor Models and Principal Components[2]

Statistical agencies in developed economies regularly report data on hundreds or thousands of time series describing the macroeconomy. These data include detailed information from the national income and product accounts (consumption, investment, imports, exports, and government spending), multiple series on price and wage

---

[2]This section draws on the material in Section 14.5, which should be read first.

inflation, output and production by industry or sector, data on specific markets such as housing, and data for asset markets including interest rates and asset prices. Each of these series could potentially contain information that could improve macroeconomic forecasts. But as explained in Chapter 14, with many predictors—potentially more than the number of available time series observations—regressions estimated by OLS will provide poor out-of-sample performance. To take advantage of this wealth of data, other methods must be used.

This section focuses on one such approach, which uses the principal components of the data set to reduce the number of coefficients to be estimated. The use of principal components for prediction was discussed in Section 14.5; that treatment is extended here to time series data. The framework for doing so is the dynamic factor model (DFM), which models the comovements of a large number of time series as arising from a small number of unobserved variables, the so-called dynamic factors. One of the steps in estimating a DFM is estimation of these unobserved factors using principal components. As discussed at the end of this section, the DFM can be used for purposes other than forecasting.

The DFM is a widely used approach for forecasting with many time series predictors, but it is not the only approach. Another method is to estimate a VAR with many predictors but to use shrinkage methods, including Bayesian methods, to estimate those coefficients. For a graduate textbook discussion of Bayesian estimation of VARs, see Kilian and Lütkepohl (2017).

The Dynamic Factor Model

A central empirical regularity of developed economies is that there are broad common movements among macroeconomic variables: When there is strength in one part of the economy, there often is strength in other parts as well. At a horizon of several years, the common swings in many economic variables give rise to what are referred to as business cycles. Macroeconomic variables also move together at shorter horizons (months or quarters) and at longer horizons (decadal movements in long-term growth rates). Theories of macroeconomic fluctuations build on this empirical regularity of broadly observed comovements and attribute these comovements to a relatively small number of driving forces, such as productivity improvements, monetary policy, fiscal policy, and changes in demand or consumer preferences.

The **dynamic factor model** captures this notion that there are a small number (r) of common factors, which drive the comovements among a large number (N) of time series variables. The DFM treats these driving factors as unobserved. Treating the factors as unobserved admits that macroeconomists do not know all the sources of macroeconomic fluctuations and that even if they did, those sources would be difficult to measure directly (for example, technological progress is very difficult to measure). In a DFM, observed macroeconomic variables, such as GDP growth and the unemployment rate, are modeled as depending on these common unobserved factors and on other omitted drivers or measurement error.

Stated mathematically, the DFM has two parts. The first relates each of the N observable variables, X_{it}, to the r factors $F_{1t}, \ldots, F_{rt}$ plus an error term u_{it}:

$$X_{it} = \Lambda_{i0} + \Lambda_{i1}F_{1t} + \cdots + \Lambda_{ir}F_{rt} + u_{it}, i = 1, \ldots, N, \tag{17.32}$$

where $\Lambda_{i1}, \ldots, \Lambda_{ir}$ are unknown coefficients relating the r factors to the ith observable variable and u_{it} is a mean 0 error term that represents omitted effects that are unique to X_{it} (that is, not common across variables) and measurement error.

The second part of the DFM specifies that the r factors follow a VAR. For notational convenience, we write the VAR here with a single lag [that is, as a VAR(1)]; however, more lags can be included:

$$\begin{aligned} F_{1t} &= A_{11}F_{1t-1} + A_{12}F_{2t-1} + \cdots + A_{1r}F_{rt-1} + \eta_{1t} \\ &\;\;\vdots \\ F_{rt} &= A_{r1}F_{1t-1} + A_{r2}F_{2t-1} + \cdots + A_{rr}F_{rt-1} + \eta_{rt}, \end{aligned} \tag{17.33}$$

where the A's are unknown VAR coefficients and the η's are mean 0 error terms. The factor VAR in Equation (17.33) is the extension to multiple variables (the r factors) of the two-variable VAR in Key Concept 17.1.

The error term u_{it} is assumed to be uncorrelated across series and to be uncorrelated with the factor VAR errors—that is, $E(u_{it}u_{jt+k}) = 0$, $i \neq j$, and $E(u_{it}\eta_{jt+k}) = 0$ for all k—so that all the common movements are associated with the common factors. Because there is no intercept in Equation (17.33), the factors have mean 0.

The **common component** of X_t is the part of X_{it} that is explained by the factors— that is, the predicted value of X_{it} given the factors, based on the population coefficients. In Equation (17.32), it is $\Lambda_{i1}F_{1t} + \cdots + \Lambda_{ir}F_{ir}$. The error term in Equation (17.32), u_{it}, is called the **idiosyncratic component** of X_{it} because it is the part of X_{it} not explained by the common factors. In general, the idiosyncratic component can be serially correlated, which affects how forecasts are made using the DFM.[3]

The DFM: Estimation and Forecasting

From the perspective of forecasting with many predictors, the DFM resolves the problem of having many predictors by replacing the many available time series with a small number of factors. If the factors were observed, the Λ coefficients in Equation (17.32) and the VAR coefficients in Equation (17.33) therefore could be estimated by OLS. The difficulty, however, is that the factors are not observed. The factors can, however, be estimated by the principal components of the N observed X's. These estimated factors can then be treated as data for the purpose of estimating the unknown DFM coefficients.

---

[3]Equations (17.32) and (17.33) are the so-called static form of the DFM, which is the version of the DFM most directly amenable to principal components estimation. Other forms of the DFM, and other ways to estimate the factors, are discussed in Stock and Watson (2016).

Estimation of the DFM and the factors using principal components. The method of principal components described in Section 14.5 extends directly to the time series setting. As discussed in Section 14.5, the X variables must first be standardized using their in-sample means and standard deviations; then the principal components are computed using the standardized X's. In Section 14.5, the first r principal components were denoted $PC_1, \ldots, PC_r$. In the context of the DFM, these principal components are the estimates of the common factors, and their value at date t is denoted $\hat{F}_{1t}, \ldots, \hat{F}_{rt}$, where the caret (^) indicates that the factor is estimated. If the factor model assumptions are, in fact, correct, then the principal components are consistent estimates of the factors in the sense that predictions made using the factors (were they observed) and using the principal components will be the same when both N and T are large.

Given the estimated factors $\hat{F}_{1t}, \ldots, \hat{F}_{rt}$, the Λ and A coefficients of the DFM in Equations (17.32) and (17.33) can be estimated by OLS, where the estimated factors are treated as data.

It is tempting to interpret the principal components themselves; for example, one might want to interpret the first principal component (the first estimated factor) as measuring overall economic activity, the second as measuring price inflation, and so forth. Unfortunately, such interpretations generally are not justified. The reason is that the factors are identified only up to linear combinations; without further assumptions, the factors themselves are not identified. Said differently, the common components of the series are identified in the dynamic factor model, but the factors themselves are not. For forecasting, this identification issue is irrelevant because the same forecasts will arise whether the factors or a linear combination of them is used (recall that, with OLS, the same prediction is made using, say, an intercept and the binary variable *male* as with an intercept and the binary variable *female*).

Determining the number of factors. In Chapter 14, the number of principal components was determined by leave-m-out cross validation. This method entails randomly assigning data to the m subsamples and then estimating the coefficients on the m subsamples that omit those observations. Unfortunately, leave-m-out cross validation has two problems in time series data. First, the time series observations are not independent, so the omitted data in the left-out subsample are not independent of the estimation sample. Second, if a subsample, even a contiguous subsample, is omitted, additional observations are lost because of the lag structure in the model.

For these reasons, determining the number of factors for DFMs tends to rely on scree plots and information criteria.

The scree plot with time series data is the same as that with cross-sectional data and is explained in Section 14.5.

Information criteria for determining the number of factors in a DFM have a similar structure to those used to determine the lag length for an autoregression [Equation (15.23)] or for a VAR [Equation (17.4)]. Specifically, the information criterion penalizes the sum of squared residuals for adding another factor. The information criterion approach to

determining r was introduced by Bai and Ng (2002). A specific criterion they propose, which has been found to work well in simulations, is

$$IC(r) = \ln\left\{ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[X_{it} - (\hat{\Lambda}_{i0} + \hat{\Lambda}_{i1}\hat{F}_{1t} + \cdots + \hat{\Lambda}_{ir}\hat{F}_{rt})]^2 \right\}$$
$$+ r\left(\frac{N+T}{NT}\right)\ln[\min(N, T)] \tag{17.34}$$

where the $\hat{\Lambda}$'s are the OLS estimates of the Λ's, estimated using the first r principal components as regressors, and the final term is the penalty for using r principal components.

The Bai–Ng penalty in Equation (17.34) increases proportionately to the number of factors r, with a constant of proportionality that depends on the number of variables as well as the number of time series observations. When $N = T$, this penalty simplifies to 2 times the BIC penalty, $[\ln(T)]/T$.

Estimation of the number of factors using the information criterion in Equation (17.34) proceeds as for autoregressions and VARs: Among a set of candidate values of r, the estimated lag length is the value of r that minimizes $IC(r)$.

Forecasting using the estimated factors. There are two approaches to forecasting using the estimated factors, which parallel the iterated and direct approaches to multi-period forecasting described in Section 17.2.

The starting point for both approaches is to extend Equation (17.32) to an autoregressive distributed lag model. Because u_{it} is, in general, serially correlated, past values of u_{it} are useful for forecasting u_{it} and thus X_{it}. Accordingly, the argument leading to Equation (16.21) applies here, so that the serial correlation in u_{it} implies that lagged values of X_{it} might be useful predictors as well. With these lagged terms added, Equation (17.32) becomes

$$X_{it} = \Lambda_{i0} + \Lambda_{i1}F_{1t} + \cdots + \Lambda_{ir}F_{rt} + \beta_1 X_{it-1} + \cdots + \beta_p X_{it-p} + u_{it}. \tag{17.35}$$

The right-hand side of Equation (17.35) depends on $F_{1t}, \ldots, F_{rt}$, which are unknown at date $t-1$; thus current values of the factors (or their principal components estimates) cannot be used as predictors. The iterated and direct forecasting approaches take two different tacks to address this problem.

In the iterated approach, the contemporaneous values of the factors in Equation (17.35) are replaced by their forecasts from the estimated factor VAR. Thus the one-step ahead forecast for period $T + 1$, using data through period T, is

$$\hat{X}_{iT+1|T} = \hat{\Lambda}_{i0} + \hat{\Lambda}_{i1}\hat{F}_{1T+1|T} + \cdots + \hat{\Lambda}_{ir}\hat{F}_{rT+1|T} + \hat{\beta}_1 X_{iT} + \cdots + \hat{\beta}_p X_{iT-p+1}, \tag{17.36}$$

where the $\hat{\Lambda}$'s and $\hat{\beta}$'s are the estimates of the Λ's and β's in Equation (17.32) using $\hat{F}_{1t}, \ldots, \hat{F}_{rt}$ and lagged X's as regressors and where $\hat{F}_{1T+1|T}, \ldots, \hat{F}_{rT+1|T}$ are the one-step ahead forecasts of the factors computed using the factor VAR. Forecasts for horizons $h > 1$ are computed using the iterated VAR forecasts of the factors and of X_i.

The direct approach builds on Key Concept 17.3. Specifically, the *h*-step ahead direct forecasting regression using the estimated factors is

$$X_{it} = \delta_0 + \delta_1 \hat{F}_{1t-h} + \cdots + \delta_r \hat{F}_{rt-h} + \delta_{r+1} X_{it-h} + \cdots + \delta_{r+p} X_{it-h-p} + u_{it},$$

$$\tag{17.37}$$

where there are different regressions, and thus different δ coefficients, at each forecasting horizon *h*. For a given horizon, the coefficients of Equation (17.37) can be estimated by OLS, and the direct forecasts are then made using those estimated coefficients.

Typically, the coefficients are estimated using data through a specific date, and then the coefficients are frozen and used for real-time forecasting. This introduces a subtlety for forecasting with DFMs: The final observations on the factors, which are used to make real-time forecasts, might not have appeared in the estimation data set. As discussed in Appendix 14.5, because the coefficients are estimated using the in-sample principal components, the same weights and standardizing means and variances must be used to construct the principal components in the out-of-sample period as were used in the estimation sample.

Other uses of DFMs. DFMs can be used for purposes other than forecasting.

One such use is to construct economic indexes. If one has a large number of similar series, it can be useful to have a single summary index that captures the common comovements. In this case, a model with a single factor can be appropriate. The estimate of the single factor (the first principal component) then summarizes the comovements of all the variables. This approach is commonly used to compute a coincident economic index from multiple measures of economic activity.

Another use of DFMs is to estimate the *current* value of a variable. This problem arises because economic data are typically released with a lag. For example, one might be interested in the change of employment in the current month, but those data will not be released until next month. The task of "forecasting" current values of economic data is called **nowcasting**. The main technical challenge of nowcasting is that data are released over the course of any month, so that the nowcasting model must be able to incorporate incoming data as they arrive. The DFM is well suited to doing so, but it must be adapted to handle missing observations, and those methods are beyond the scope of this book. The Federal Reserve Bank of New York uses a DFM to produce nowcasts of GDP, which it updates weekly based on that week's data.[4]

Application to U.S. Macroeconomic Data

We illustrate the estimation and use of the dynamic factor model using a data set comprised of 131 quarterly macroeconomic time series for the United States, spanning 1960:Q1–2017:Q4. The series are summarized in Table 17.2, with additional information provided in Appendix 17.1. The variables in the data set include standard

---

[4]The New York Fed GDP nowcasts are posted at https://www.newyorkfed.org/research/policy/nowcast.

TABLE 17.2 The Quarterly Macroeconomic Data Set	
Category	**Number of Series Used for Factor Estimation**
National Income and Product Accounts	13
Industrial Production	8
Employment and Unemployment	30
Orders, Inventories, and Sales	6
Housing Starts and Permits	6
Prices	22
Productivity and Labor Earnings	5
Interest Rates	10
Money and Credit	6
International	8
Asset Prices, Wealth, Household Balance Sheets	10
Other	2
Oil Market Variables	5
Total	**131**

measures of economic activity, wage and price inflation, interest rates, and data on large markets of macroeconomic importance including housing and oil markets. The data were transformed to eliminate stochastic trends, typically by transforming to growth rates (as for GDP) or first differences (interest rates). These transformed data were then standardized by subtracting their sample mean and dividing by their sample standard deviation prior to estimation.

In some categories, series are available at multiple levels of aggregation. For example, GDP is the sum of consumption, investment, government spending, and imports; thus GDP is perfectly collinear with its components. Similarly, total employment is the sum of employment across the sectors of the economy. For the purpose of estimating the factors, the aggregate series (GDP, total employment) provide no additional information beyond their components, so the aggregate series were excluded from the data set. The final column of Table 17.2 lists the number of series used to compute the principal component factor estimates.

Figure 17.4 presents the scree plot of the first 30 principal components of the 131 series in the data set, over the full 1960–2017 period. Evidently, a large amount of the variance of these series is captured by the first few principal components. The first principal component explains 20% of the total variance of the series, the second principal component explains 9%, and the first four collectively explain 39%.

The scree plot provides some guidance about the number of factors to include. Clearly, the first and second factors are important, and there are also substantial drops in the marginal R^2 after the third and fourth factors. The decline does not seem to stabilize, however, until the tenth factor, so this visual analysis is inconclusive. The Bai–Ng information criterion [Equation (17.34)] is minimized using $r = 4$ factors.

FIGURE 17.4 Scree Plot of First 30 Factors for the Macro Data Set, 1960–2017

The first factor explains 20% of the total variance of the series, and the first four factors collectively explain 39% of the total variance of the series.

Fraction of total variance of X explained

Factor number (principal component number)
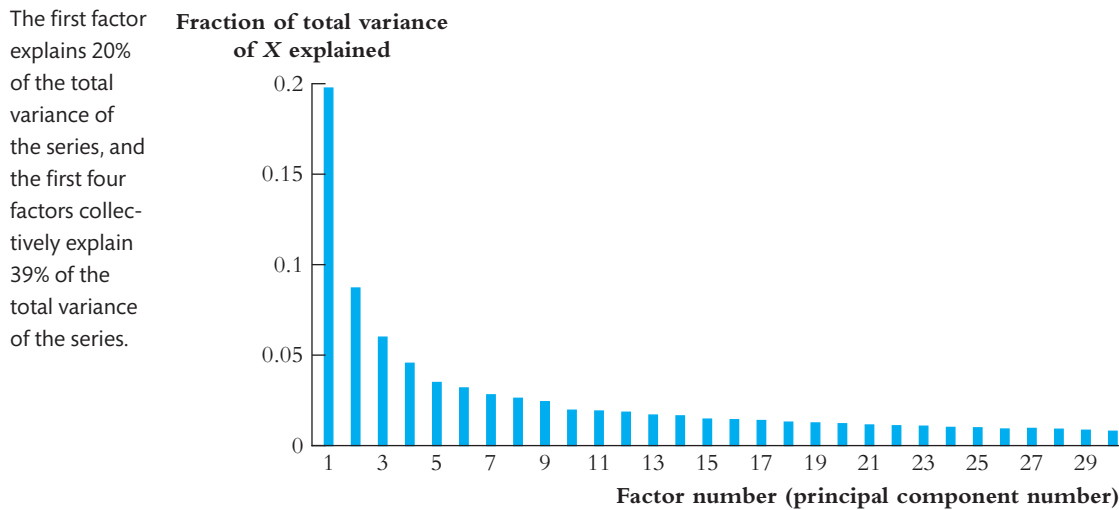
This estimate is within the plausible range from the inspection of the scree plot, so we adopt $r = 4$ for the rest of this example.

Figure 17.5 plots the four-quarter growth rate of GDP, employment, oil prices, and returns on the S&P 500 stock index (the four-quarter growth is the percentage growth of the series from quarter t to quarter $t + 4$, computed using the log approximation to percentage changes). The figure also plots the common component of each of the series, estimated using four factors. Of these series, GDP and employment are not in the data set used to estimate the factors because they are aggregates of other included series, while the oil prices and stock returns are among the 131 series used to estimate the factors.

The striking conclusion from Figure 17.5 is that the common component, computed using only the first 4 principal components of the 131 macro variables, captures a large amount of the variation in these series. Even a large fraction of the four-quarter returns on the S&P 500 are explained by these 4 factors. This does not imply that stock returns are predictable; rather, it implies that stock returns are heavily influenced by contemporaneous developments in aggregate economic activity.

We conclude by examining forecasts of GDP growth made using the four estimated factors and comparing those to the AR and ADL forecasts in Chapter 15. We consider direct forecasts of cumulative GDP growth at horizons $h = 1$, 4, and 8, where growth is measured at an annual rate. For example, at the four-quarter horizon, the dependent variable is $400\ln(GDP_t/GDP_{t-4})$, which equals the average of the quarterly growth in periods $t, t - 1, t - 2$, and $t - 3$ at an annual rate. The three forecasting models examined are direct forecasts of h-period growth corresponding to an AR(2), an ADL(2, 2) with the term spread, and a four-factor forecast that includes two lags of GDP growth.

FIGURE 17.5   Four-Quarter Growth Rates, Actual and Common Components, 1960–2017

(a) Real GDP

(b) Employment

(c) Oil prices

(d) Stock prices
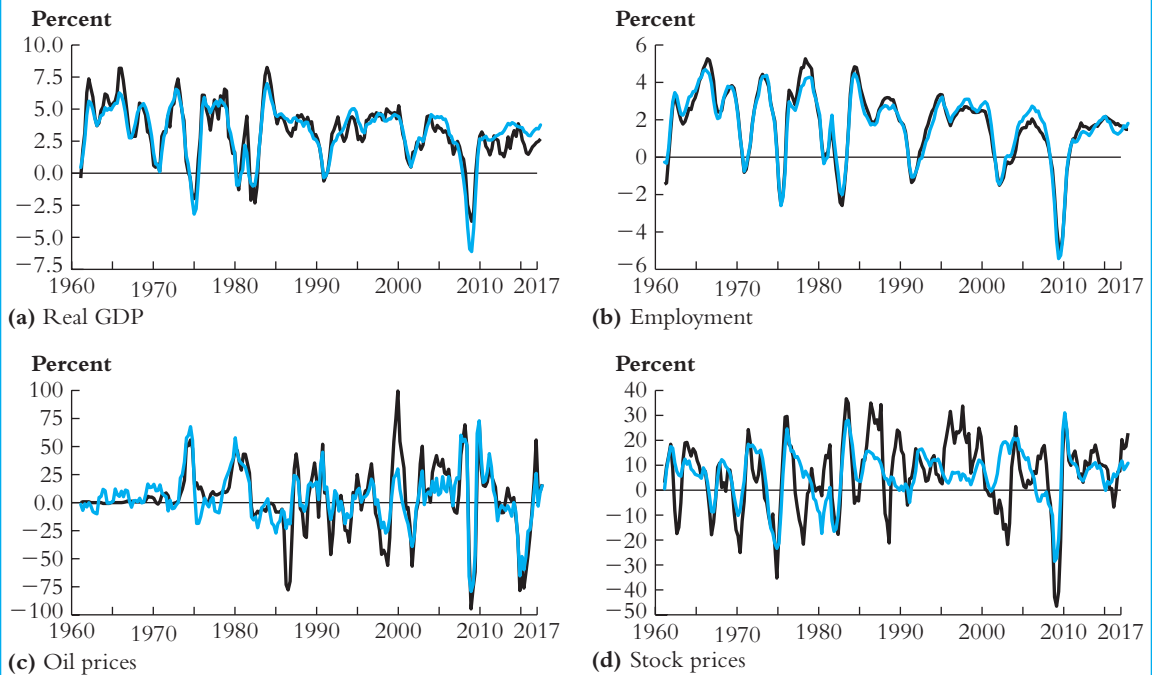
The series (black) and estimated common components (blue) of GDP, employment, oil prices, and returns on the S&P 500 based on a four-factor DFM, estimated using the 131-series macroeconomic data set, 1960–2017.

Table 17.3 reports the performance of the forecasts as measured by the pseudo out-of-sample root mean square forecast error, $\widehat{RMSFE}_{POOS}$ [Equation (15.22)]. The first column lists the regressors in the direct forecasting regressions. Following Section 15.8, the in-sample period starts in 1981:Q1 and ends h periods prior to 2002:Q4; the pseudo out-of-sample period is 2002:Q4–2017:Q4.

Three aspects of these results are noteworthy. First, the $\widehat{RMSFE}_{POOS}$ decreases as the horizon lengthens. One reason for this improvement at longer horizons is that quarterly GDP has a large amount of transitory measurement error, which is smoothed over (averaged out) by considering growth rates over one or two years. This quarterly "noise" is evident in the time series plot of quarterly GDP growth in Figure 15.1b.

Second, at all horizons the forecasts that use the term spread do worse in the out-of-sample period than the direct AR(2) forecasts. This would appear to contradict the improvement in in-sample fit provided by the term spread: The F-statistic testing whether the coefficients on $TSpread_{t-1}$ and $TSpread_{t-2}$ are 0 in the $h = 1$ estimation sample (1981:Q1–2002:Q3) is statistically significant at the 1% level. Evidently, the coefficients on the lagged term spread estimated in the in-sample period do not capture the relation between the term spread and GDP in the pseudo out-of-sample period, an indication that this relation is nonstationary. In real-world terms, one important difference between the in- and out-of-sample periods is that, starting

TABLE 17.3 Comparison of Direct Forecasts of Cumulative GDP Growth at an Annual Rate: Lagged GDP, Term Spread, and Principal Components, 2002:Q4–2017:Q4

	$\widehat{RMSFE}_{POOS}$		
Predictors	$h = 1$	$h = 4$	$h = 8$
$GDPGR_{t-h}, GDPGR_{t-h-1}$	2.25	1.91	1.74
$GDPGR_{t-h}, GDPGR_{t-h-1}, TSpread_{t-h}, TSpread_{t-h-1}$	2.29	1.94	1.77
$GDPGR_{t-h}, GDPGR_{t-h-1}, \hat{F}_{1t-h}, \hat{F}_{2t-h}, \hat{F}_{3t-h}, \hat{F}_{4t-h}$	2.14	1.40	1.48

Entries are root mean square forecast errors, estimated by pseudo out-of-sample forecasts for the forecast period 2002:Q4–2017:Q4 [Equation (15.22)]. The forecasting models were estimated using data from 1981:Q1 through h periods before 2002:Q4, where h is the forecast horizon. The dependent variable is the h-quarter cumulative growth in GDP at an annual rate, using log points—that is, $(400/h)\ln(GDP_t/GDP_{t-h})$. The regressors are given in the first column, where $\hat{F}_{1t}$ denotes the first factor estimated by the first principal component in the estimation sample and so on. All regressions include an intercept.

in 2008, the Federal Reserve Board introduced new monetary policy tools to manage long-term as well as short-term rates, thereby changing the relation between the term spread and economic activity.

Third, the factor forecasts improve upon the AR and ADL forecasts at all horizons. Closer inspection of the forecasts reveals that this improvement is due to much better performance of the factor forecasts during the recession and early recovery following the financial crisis in the fall of 2009. During this recession, the strong negative comovements across many macro variables pointed toward a deep recession, a feature missed by the AR forecast. In contrast, during the relatively quiescent periods of 2005 and after 2013, the AR(2) direct forecast actually performs slightly better than the factor forecast.

Nobel Laureates in Time Series Econometrics

In 2003, Robert Engle and Clive Granger won the Nobel Prize in Economics for fundamental theoretical research in time series econometrics. Engle's work was motivated by the volatility clustering evident in plots like Figure 17.2. Engle wondered whether series like these could be stationary and whether econometric models could be developed to explain and predict their time-varying volatility. Engle's answer was to develop the autoregressive conditional heteroskedasticity (ARCH) model, described in Section 17.5. The ARCH model and its extensions proved especially useful for modeling the volatility of asset returns, and the resulting volatility forecasts are used to price financial derivatives and to assess changes over time in the risk of holding financial assets. Today, measures and forecasts of volatility are a core component of financial econometrics, and the ARCH model and its descendants are the workhorse tools for modeling volatility.
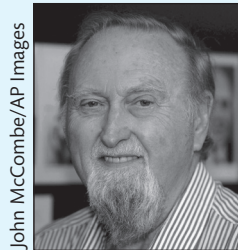
John McCombe/AP Images
Clive W. J. Granger

New York University/AFP/Newscom
Robert F. Engle

Granger's work focused on how to handle stochastic trends in economic time series data. From his earlier work, he knew that two unrelated series with stochastic trends could, by the usual statistical measures of *t*-statistics and regression R^2's, falsely appear to be meaningfully related; this is the "spurious regression" problem exemplified by the regressions in Equations (14.28) and (14.29). But are all regressions involving stochastic trending variables spurious? Granger discovered that when variables shared common trends—in his terminology, were "co-integrated"—meaningful relationships could be uncovered by regression analysis using a vector error correction model. The methods of cointegration analysis are now a staple in modern macroeconometrics.

In 2011, Thomas Sargent and Christopher Sims won the Nobel Prize for their empirical research on cause and effect in the macroeconomy. Sargent was recognized for developing models that featured the important role that expectations about the future play in disentangling cause and effect. Sims was recognized for developing structural VAR (SVAR) models. Sims's key insight concerned the forecast errors in a VAR model—the u_t errors in Equations (17.1) and (17.2). These errors, he realized, arose because of unforeseen "shocks" that buffeted the macroeconomy, and in many cases, these shocks had well-defined sources like the Organization of Petroleum Exporting Countries (oil price shocks), the Fed (interest rate shocks), or Congress (tax shocks). By disentangling the various sources of shocks that comprise the VAR errors, Sims was able to estimate the dynamic causal effect of these shocks on the variables appearing in the VAR. This disentangling of shocks is never without controversy, but SVARs are now a standard tool for estimating dynamic causal effects in macroeconomics.
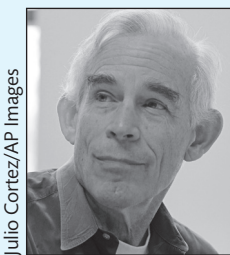
In 2013, Eugene Fama, Lars Peter Hansen, and Robert Shiller won the Nobel Prize for their empirical analysis of asset prices. The work in the box "Can You Beat the Market?" in Chapter 15 and the box "NEWS FLASH: Commodity Traders Send Shivers Through Disney World" in Chapter 16 was motivated in part by the "efficient markets" (unpredictability) work of Fama and the "irrational exuberance" (unexplained volatility) work of Shiller. Hansen was honored for developing generalized method of moments (GMM) methods to investigate whether asset returns are consistent with expected utility theory. Microeconomics says that investors should equate the marginal cost of an investment (today's foregone utility from investing rather than consuming) with its marginal benefit (tomorrow's boost in utility from consumption financed by the investment's return). But a simple test of this proposition is complicated because marginal utility is difficult to measure, asset returns are uncertain, and the argument should hold across all asset returns. Hansen developed GMM methods to test asset-pricing models. As it turned out, Hansen's GMM methods had applications well beyond finance and are now widely used in econometrics. Section 19.7 introduces GMM.

For more information on these and other Nobel laureates in economics, visit the Nobel Foundation website, **http://www.nobel.se/economics**.

Christopher A. Sims

Lars Peter Hansen

17.7 Conclusion

This part of the text has covered some of the most frequently used tools and concepts of time series regression. Many other tools for analyzing economic time series have been developed for specific applications. If you are interested in learning more about economic forecasting, see the introductory textbooks by Diebold (2017) and Enders (2009). For an advanced treatment of econometrics with time series data, see Hamilton (1994) and Hayashi (2000). For an advanced treatment of vector autoregressions, see Kilian and Lütkepohl (2017), and for more on dynamic factor models, see Stock and Watson (2016).

Summary

1. Vector autoregressions model k time series variables, with each depending on its own lags and the lags of the $k - 1$ other series. The forecasts of each of the time series produced by a VAR are mutually consistent in the sense that they are based on the same information.

2. Forecasts two or more periods ahead can be computed either by iterating forward a one-step ahead model (an AR or a VAR) or by estimating a multi-period ahead regression.

3. Two series that share a common stochastic trend are cointegrated; that is, Y_t and X_t are cointegrated if Y_t and X_t are $I(1)$ but $Y_t - \theta X_t$ is $I(0)$. If Y_t and X_t are cointegrated, the error correction term $Y_t - \theta X_t$ can help predict ΔY_t and/or ΔX_t. A vector error correction model is a VAR model of ΔY_t and ΔX_t, augmented to include the lagged error correction term.

4. Volatility clustering—in which the variance of a series is high in some periods and low in others—is common in economic time series, especially financial time series. Realized volatility is an estimate of time-varying volatility using a rolling root mean square estimator.

5. The ARCH model of volatility clustering expresses the conditional variance of the regression error as a function of recent squared regression errors. The GARCH model augments the ARCH model to include lagged conditional variances as well. Realized volatility and ARCH/GARCH models produce forecast intervals with widths that depend on the volatility of the most recent regression residuals.

6. The comovements of a large number of time series sometimes can be summarized by the first few principal components, which in turn can be used for forecasting. The framework for doing so is the dynamic factor model, which posits that a small number of unobserved factors drive the comovements of a large number of macroeconomic variables.

Key Terms

vector autoregression (VAR) (608)

iterated multi-period AR
forecast (614)

iterated multi-period VAR
forecast (614)

direct multi-period forecast (616)

second difference (617)

integrated of order zero [$I(0)$],
one [$I(1)$], or two [$I(2)$] (617)

order of integration (617)

integrated of order d [$I(d)$] (618)

common trend (621)

cointegration (621)

cointegrating coefficient (622)

error correction term (621)

vector error correction model
(VECM) (621)

EG-ADF test (623)

dynamic OLS (DOLS) estimator (623)

volatility clustering (625)

realized volatility (626)

autoregressive conditional
heteroskedasticity (ARCH) (627)

generalized ARCH (GARCH) (627)

dynamic factor model (630)

common component (631)

idiosyncratic component (631)

nowcasting (634)

MyLab Economics Can Help You Get a Better Grade

MyLab Economics If your exam were tomorrow, would you be ready? For each
chapter, **MyLab Economics** Practice Tests and Study Plan
help you prepare for your exams. You can also find the Exercises and all Review the Concepts Questions
available now in **MyLab Economics**. To see how it works, turn to the **MyLab Economics** spread on the
inside front cover of this text and then go to **www.pearson.com/mylab/economics**.

For additional Empirical Exercises and Data Sets, log on to the Companion Website at
www.pearsonhighered.com/stock_watson.

Review the Concepts

17.1 A macroeconomist wants to construct forecasts for the following macro-economic variables: GDP, consumption, investment, government purchases, exports, imports, short-term interest rates, long-term interest rates, and the rate of price inflation. He has quarterly time series for each of these variables from 1970 to 2017. Should he estimate a VAR for these variables and use this for forecasting? Why or why not? Can you suggest an alternative approach?

17.2 Suppose that Y_t follows a stationary AR(1) model with $\beta_0 = 0$ and $\beta_1 = 0.7$. If $Y_t = 5$, what is your forecast of Y_{t+2}? That is, what is $Y_{t+2|t}$? What is $Y_{t+h|t}$ for $h = 30$? Does this forecast for $h = 30$ seem reasonable to you?

17.3 A version of the permanent income theory of consumption implies that the logarithm of real GDP (Y) and the logarithm of real consumption (C) are cointegrated with a cointegrating coefficient equal to 1. Explain how you would investigate this implication by (a) plotting the data and (b) using a statistical test.

17.4 Consider the ARCH model, $\sigma_t^2 = 1.0 + 0.8\, u_{t-1}^2$. Explain why this will lead to volatility clustering. (*Hint:* What happens when u_{t-1}^2 is unusually large?)

17.5 Suppose a forecaster has 110 predictors (X's) she could use to predict Y and 150 monthly time series observations. Explain why an OLS regression of Y_t on the first lag of the X's is likely to produce poor one-step ahead forecasts. How does the dynamic factor model, estimated using principal components, address this problem?

Exercises

17.1 Suppose that Y_t follows a stationary AR(1) model, $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$.

    a. Show that the h-period ahead forecast of Y_t is given by
$Y_{t+h|t} = \mu_Y + \beta_1^h(Y_t - \mu_Y)$, where $\mu_Y = \beta_0/(1 - \beta_1)$.

    b. Suppose that X_t is related to Y_t by $X_t = \sum_{i=0}^{\infty} \delta^i Y_{t+i|t}$, where $|\delta| < 1$.
Show that $X_t = [\mu_Y/(1 - \delta)] + [(Y_t - \mu_Y)/(1 - \beta_1\delta)]$.

17.2 One version of the expectations theory of the term structure of interest rates holds that a long-term rate equals the average of the expected values of short-term interest rates into the future plus a term premium that is $I(0)$. Specifically, let Rk_t denote a k-period interest rate, let $R1_t$ denote a one-period interest rate, and let e_t denote an $I(0)$ term premium. Then $Rk_t = \frac{1}{k}\sum_{i=0}^{k-1} R1_{t+i|t} + e_t$, where $R1_{t+i|t}$ is the forecast made at date t of the value of $R1$ at date $t + i$. Suppose that $R1_t$ follows a random walk so that $R1_t = R1_{t-1} + u_t$.

    a. Show that $Rk_t = R1_t + e_t$.

    b. Show that Rk_t and $R1_t$ are cointegrated. What is the cointegrating coefficient?

    c. Now suppose that $\Delta R1_t = 0.5\Delta R1_{t-1} + u_t$. How does your answer to (b) change?

    d. Now suppose that $R1_t = 0.5R1_{t-1} + u_t$. How does your answer to (b) change?

17.3 Suppose that $E(u_t | u_{t-1}, u_{t-2}, \ldots) = 0$ and u_t follows the ARCH process, $\sigma_t^2 = 1.0 + 0.5\, u_{t-1}^2$.

    a. Let $E(u_t^2) = \text{var}(u_t)$ be the unconditional variance of u_t. Show that $\text{var}(u_t) = 2$. (*Hint:* Use the law of iterated expectations, $E(u_t^2) = E[Eu_t^2 | u_{t-1})]$.)

    b. Suppose that the distribution of u_t conditional on lagged values of u_t is $N(0, \sigma_t^2)$. If $u_{t-1} = 0.2$, what is $\Pr(-3 \le u_t \le 3)$? If $u_{t-1} = 2.0$, what is $\Pr(-3 \le u_t \le 3)$?

17.4 Suppose that Y_t follows the AR(p) model $Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + u_t$, where $E(u_t | Y_{t-1}, Y_{t-2}, \ldots) = 0$. Let $Y_{t+h|t} = E(Y_{t+h} | Y_t, Y_{t-1}, \ldots)$. Show that $Y_{t+h|t} = \beta_0 + \beta_1 Y_{t-1+h|t} + \cdots + \beta_p Y_{t-p+h|t}$ for $h > p$.

17.5 Verify Equation (17.20). [*Hint:* Use $\sum_{t=1}^{T} Y_t^2 = \sum_{t=1}^{T}(Y_{t-1} + \Delta Y_t)^2$ to show that $\sum_{t=1}^{T} Y_t^2 = \sum_{t=1}^{T} Y_{t-1}^2 + 2\sum_{t=1}^{T} Y_{t-1}\Delta Y_t + \sum_{t=1}^{T}\Delta Y_t^2$, and solve for $\sum_{t=1}^{T} Y_{t-1}\Delta Y_t$.]

17.6   A regression of Y_t onto current, past, and future values of X_t yields

$$Y_t = 3.0 + 1.7X_{t+1} + 0.8X_t - 0.2X_{t-1} + u_t.$$

   a. Rearrange the regression so that it has the form shown in Equation (17.25). What are the values of θ, δ_{-1}, δ_0, and δ_1?

   b.   i. Suppose that X_t is $I(1)$ and u_t is $I(1)$. Are Y and X cointegrated?

   ii. Suppose that X_t is $I(0)$ and u_t is $I(1)$. Are Y and X cointegrated?

   iii. Suppose that X_t is $I(1)$ and u_t is $I(0)$. Are Y and X cointegrated?

17.7   Suppose that $\Delta Y_t = u_t$, where u_t is i.i.d. $N(0, 1)$, and consider the regression $Y_t = \beta X_t + \ error$, where $X_t = \Delta Y_{t+1}$ and $error$ is the regression error. Show that $\hat{\beta} \xrightarrow{d} \frac{1}{2}(\chi_1^2 - 1)$. [*Hint:* Analyze the numerator of $\hat{\beta}$ using analysis like that in Equation (17.21). Analyze the denominator using the law of large numbers.]

17.8   Consider the following two-variable VAR model with one lag and no intercept:

$$Y_t = \beta_{11}Y_{t-1} + \gamma_{11}X_{t-1} + u_{1t}$$
$$X_t = \beta_{21}Y_{t-1} + \gamma_{21}X_{t-1} + u_{2t}.$$

   a. Show that the iterated two-period ahead forecast for Y can be written as $Y_{t|t-2} = \delta_1 Y_{t-2} + \delta_2 X_{t-2}$, and derive values for δ_1 and δ_2 in terms of the coefficients in the VAR.

   b. In light of your answer to (a), do iterated multi-period forecasts differ from direct multi-period forecasts? Explain.

17.9   **a.** Suppose that $E(u_t \mid u_{t-1}, u_{t-2}, \dots) = 0$, that $\text{var}(u_t \mid u_{t-1}, u_{t-2}, \dots)$ follows the ARCH(1) model $\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2$, and that the process for u_t is stationary. Show that $\text{var}(u_t) = \alpha_0/(1 - \alpha_1)$. (*Hint:* Use the law of iterated expectations, $E(u_t^2) = E[E(u_t^2 \mid u_{t-1})]$.)

   b. Extend the result in (a) to the ARCH(p) model.

   c. Show that $\sum_{i=1}^{p}\alpha_i < 1$ for a stationary ARCH(p) model.

   d. Extend the result in (a) to the GARCH(1, 1) model.

   e. Show that $\alpha_1 + \phi_1 < 1$ for a stationary GARCH(1, 1) model.

17.10 Consider the cointegrated model $Y_t = \theta X_t + v_{1t}$ and $X_t = X_{t-1} + v_{2t}$, where v_{1t} and v_{2t} are mean 0 serially uncorrelated random variables with $E(v_{1t}v_{2j}) = 0$ for all t and j. Derive the vector error correction model [Equations (17.22) and (17.23)] for X and Y.

Empirical Exercises

E17.1   This exercise is an extension of Empirical Exercise 14.1. On the text website, **http://www.pearsonhighered.com/stock_watson**, you will find the data file **USMacro_Quarterly**, which contains quarterly data on several macroeconomic series for the United States; the data are described in the file **USMacro_Description**. Compute

inflation, *Infl*, using the price index for personal consumption expenditures. For all regressions, use the sample period 1963:Q1–2017:Q4 (where data before 1963 may be used as initial values for lags in regressions).

a. Using the data on inflation through 2017:Q4 and an estimated AR(2) model:

    i. Forecast $\Delta Infl_{2018:Q1}$, the change in inflation from 2017:Q4 to 2018:Q1.

    ii. Forecast $\Delta Infl_{2018:Q2}$, the change in inflation from 2018:Q1 to 2018:Q2. (Use an iterated forecast.)

    iii. Forecast $Infl_{2018:Q2} - Infl_{2017:Q4}$, the change in inflation from 2017:Q4 to 2018:Q2.

    iv. Forecast $Infl_{2018:Q2}$, the rate of inflation in 2018:Q2.

b. Repeat (a) using the direct forecasting method.

E17.2 On the text website, **http://www.pearsonhighered.com/stock_watson,** you will find the data file **USMacro_Quarterly**, which contains quarterly data on real GDP, measured in 2009 dollars. Compute $GDPGR_t = 400 \times [\ln(GDP_t) - \ln(GDP_{t-1})]$, the GDP growth rate.

a. Using data on $GDPGR_t$ from 1960:Q1 to 2017:Q4, estimate an AR(2) model with GARCH(1, 1) errors.

b. Plot the residuals from the AR(2) model along with $\pm \hat{\sigma}_t$ bands as in Figure 17.3.

c. Some macroeconomists have claimed that there was a sharp drop in the variability of the growth rate of GDP around 1983, which they call the Great Moderation. Is this Great Moderation evident in your plot for (b)? Explain.

APPENDIX

17.1 The Quarterly U.S. Macro Data Set

The variables in the quarterly U.S. data set were obtained from the FRED online database of macroeconomic time series maintained by the Federal Reserve Bank of St. Louis. The categories of variables are listed in Table 17.2. The National Income and Product Account variables included in the data set for estimating the factors are three measures of personal consumption expenditures (durable goods, nondurable goods, and services); four measures of private investment (nonresidential structures, nonresidential intellectual property, nonresidential fixed equipment, and residential structures), federal government expenditures, federal government receipts, state and local government consumption, exports, and imports (all real). Stochastic trends were eliminated by (in most cases) computing quarterly growth rates or first differences. For details and for the full list of series, see the online documentation supporting this text.

The Theory of Linear Regression with One Regressor

Why should an applied econometrician bother learning any econometric theory? There are several reasons. Learning econometric theory turns your statistical software from a "black box" into a flexible tool kit from which you are able to select the right tool for the job at hand. Understanding econometric theory helps you appreciate why these tools work and what assumptions are required for each tool to work properly. Perhaps most importantly, knowing econometric theory helps you recognize when a tool will *not* work well in an application and when you should look for a different econometric approach.

This chapter provides an introduction to the econometric theory of linear regression with a single regressor. This introduction is intended to supplement—not replace—the material in Chapters 4 and 5, which should be read first.

This chapter extends Chapters 4 and 5 in two ways.

First, it provides a mathematical treatment of the sampling distribution of the ordinary least squares (OLS) estimator and *t*-statistic, both in large samples under the three least squares assumptions for causal inference of Key Concept 4.3 and in finite samples under the two additional assumptions of homoskedasticity and normal errors. These five extended least squares assumptions are laid out in Section 18.1. Sections 18.2 and 18.3, augmented by Appendix 18.2, mathematically develop the large-sample normal distributions of the OLS estimator and *t*-statistic under the first three assumptions (the least squares assumptions for causal inference of Key Concept 4.3). Section 18.4 derives the exact distributions of the OLS estimator and *t*-statistic under the two additional assumptions of homoskedasticity and normally distributed errors.

Second, this chapter extends Chapters 4 and 5 by providing an alternative method for handling heteroskedasticity. The approach of Chapters 4 and 5 is to use heteroskedasticity-robust standard errors to ensure that statistical inference is valid even if the errors are heteroskedastic. This method comes with a cost, however: If the errors are heteroskedastic, then in theory a more efficient estimator than OLS is available. This estimator, called weighted least squares, is presented in Section 18.5. Weighted least squares requires a great deal of prior knowledge about the precise nature of the heteroskedasticity—that is, about the conditional variance of *u* given *X*. When such knowledge is available, weighted least squares improves upon OLS. In most applications, however, such knowledge is unavailable; in those cases, using OLS with heteroskedasticity-robust standard errors is the preferred method.

18.1 The Extended Least Squares Assumptions and the OLS Estimator

This section introduces a set of assumptions that extend and strengthen the three least squares assumptions for causal inference of Chapter 4. These stronger assumptions are used in subsequent sections to derive stronger theoretical results about the OLS estimator than are possible under the weaker (but more realistic) assumptions of Chapter 4.

The Extended Least Squares Assumptions

Extended least squares Assumptions 1, 2, and 3. The first three extended least squares assumptions are the three assumptions given in Key Concept 4.3: The conditional mean of u_i given X_i is 0; (X_i, Y_i), $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) draws from their joint distribution; and X_i and u_i have nonzero finite fourth moments.

Under these three assumptions, the OLS estimator is unbiased, is consistent, and has a normal sampling distribution in large samples. If these three assumptions hold, then the methods for inference introduced in Chapter 4—hypothesis testing using the t-statistic and construction of 95% confidence intervals as ± 1.96 standard errors—are justified when the sample size is large. To develop a theory of efficient estimation using OLS or to characterize the exact sampling distribution of the OLS estimator, however, requires stronger assumptions.

Extended least squares assumption 4. The fourth extended least squares assumption is that u_i is homoskedastic; that is, $\text{var}(u_i \mid X_i) = \sigma_u^2$, where σ_u^2 is a constant. As seen in Section 5.5, if this additional assumption holds, then the OLS estimator is efficient among all linear estimators that are unbiased, conditional on $X_1, \ldots, X_n$.

Extended least squares assumption 5. The fifth extended least squares assumption is that the conditional distribution of u_i given X_i is normal.

Under extended least squares assumptions 1, 2, 4, and 5, u_i is i.i.d. $N(0, \sigma_u^2)$, and u_i and X_i are independently distributed. To see this, note that the fifth extended least squares assumption states that the conditional distribution of $u_i \mid X_i$ is $N(0, \text{var}(u_i \mid X_i))$, where the distribution has mean 0 by the first extended least squares assumption. By the fourth extended least squares assumption, however, $\text{var}(u_i \mid X_i) = \sigma_u^2$, so the conditional distribution of $u_i \mid X_i$ is $N(0, \sigma_u^2)$. Because this conditional distribution does not depend on X_i, u_i and X_i are independently distributed. By the second extended least squares assumption, u_i is distributed independently of u_j for all $j \neq i$. It follows that, under extended least squares assumptions 1, 2, 4, and 5, u_i and X_i are independently distributed and u_i is i.i.d. $N(0, \sigma_u^2)$.

The Extended Least Squares Assumptions for Regression with a Single Regressor

The linear regression model with a single regressor is

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n, \tag{18.1}$$

where β_1 is the causal effect on Y of X.

The extended least squares assumptions are

1. $E(u_i \mid X_i) = 0$ (conditional mean 0);
2. $(X_i, Y_i), i = 1, \ldots, n,$ are independent and identically distributed (i.i.d.) draws from their joint distribution;
3. X_i and u_i have nonzero finite fourth moments;
4. $\text{var}(u_i \mid X_i) = \sigma_u^2$ (homoskedasticity); and
5. The conditional distribution of u_i given X_i is normal (normal errors).

It is shown in Section 18.4 that, if all five extended least squares assumptions hold, the OLS estimator has an exact normal sampling distribution, and the homoskedasticity-only t-statistic has an exact Student t distribution.

The fourth and fifth extended least squares assumptions are much more restrictive than the first three. Although it might be reasonable to assume that the first three assumptions hold in an application, the final two assumptions are less realistic. Even though these final two assumptions might not hold in practice, they are of theoretical interest because if one or both of them hold, then the OLS estimator has additional properties beyond those discussed in Chapters 4 and 5. Thus we can enhance our understanding of the OLS estimator and the theory of estimation in the linear regression model by exploring estimation under these stronger assumptions.

The five extended least squares assumptions for the single-regressor model are summarized in Key Concept 18.1.

The OLS Estimator

For easy reference, we restate the OLS estimators of β_0 and β_1 here:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{18.2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{18.3}$$

Equations (18.2) and (18.3) are derived in Appendix 4.2.

18.2 Fundamentals of Asymptotic Distribution Theory

Asymptotic distribution theory is the theory of the distribution of statistics—estimators, test statistics, and confidence intervals—when the sample size is large. Formally, this theory involves characterizing the behavior of the sampling distribution of a statistic along a sequence of ever-larger samples. The theory is asymptotic in the sense that it characterizes the behavior of the statistic in the limit as $n \to \infty$.

Even though sample sizes are, of course, never infinite, asymptotic distribution theory plays a central role in econometrics and statistics for two reasons. First, if the number of observations used in an empirical application is large, then the asymptotic limit can provide a high-quality approximation to the finite sample distribution. Second, asymptotic sampling distributions typically are much simpler, and thus easier to use in practice, than exact finite-sample distributions. Taken together, these two reasons mean that reliable and straightforward methods for statistical inference—tests using t-statistics and 95% confidence intervals calculated as ± 1.96 standard errors—can be based on approximate sampling distributions derived from asymptotic theory.

The two cornerstones of asymptotic distribution theory are the law of large numbers and the central limit theorem, both introduced in Section 2.6. We begin this section by continuing the discussion of the law of large numbers and the central limit theorem, including a proof of the law of large numbers. We then introduce two more tools, Slutsky's theorem and the continuous mapping theorem, that extend the usefulness of the law of large numbers and the central limit theorem. As an illustration, these tools are then used to prove that the distribution of the t-statistic based on $\overline{Y}$ testing the hypothesis $E(Y) = \mu_0$ has a standard normal distribution under the null hypothesis.

Convergence in Probability and the Law of Large Numbers

The concepts of convergence in probability and the law of large numbers were introduced in Section 2.6. Here we provide a precise mathematical definition of convergence in probability, followed by a statement and proof of the law of large numbers.

Consistency and convergence in probability. Let $S_1, S_2, \ldots, S_n, \ldots$ be a sequence of random variables. For example, S_n could be the sample average $\overline{Y}$ of a sample of n observations of the random variable Y. The sequence of random variables $\{S_n\}$ is said to **converge in probability** to a limit, μ (that is, $S_n \xrightarrow{p} \mu$), if the probability that S_n is within $\pm \delta$ of μ tends to 1 as $n \to \infty$, as long as the constant δ is positive. That is,

$$S_n \xrightarrow{p} \mu \text{ if and only if } \Pr\left(\left|S_n - \mu\right| \geq \delta\right) \longrightarrow 0 \tag{18.4}$$

as $n \to \infty$ for every $\delta > 0$. If $S_n \xrightarrow{p} \mu$, then S_n is said to be a **consistent estimator** of μ.

The law of large numbers.   The law of large numbers says that, under certain conditions on $Y_1, \ldots, Y_n$, the sample average $\overline{Y}$ converges in probability to the population mean. Probability theorists have developed many versions of the law of large numbers, corresponding to various conditions on $Y_1, \ldots, Y_n$. The version of the law of large numbers used in this text is that $Y_1, \ldots, Y_n$ are i.i.d. draws from a distribution with finite variance. This law of large numbers (also stated in Key Concept 2.6) is

$$\text{if } Y_1, \ldots, Y_n \text{ are i.i.d., } E(Y_i) = \mu_Y, \text{ and var}(Y_i) < \infty, \text{ then } \overline{Y} \xrightarrow{p} \mu_Y \quad (18.5)$$

The idea of the law of large numbers can be seen in Figure 2.8: As the sample size increases, the sampling distribution of $\overline{Y}$ concentrates around the population mean, μ_Y. One feature of the sampling distribution is that the variance of $\overline{Y}$ decreases as the sample size increases; another feature is that the probability that $\overline{Y}$ falls outside $\pm \delta$ of μ_Y vanishes as n increases. These two features of the sampling distribution are, in fact, linked, and the proof of the law of large numbers exploits this link.

Proof of the law of large numbers.   The link between the variance of $\overline{Y}$ and the probability that $\overline{Y}$ is within $\pm \delta$ of μ_Y is provided by Chebychev's inequality, which is stated and proven in Appendix 18.2 [see Equation (18.42)]. Written in terms of $\overline{Y}$, Chebychev's inequality is

$$\Pr(|\overline{Y} - \mu_Y| \geq \delta) \leq \frac{\text{var}(\overline{Y})}{\delta^2} \quad (18.6)$$

for any positive constant δ. Because $Y_1, \ldots, Y_n$ are i.i.d. with variance σ_Y^2, $\text{var}(\overline{Y}) = \sigma_Y^2/n$; thus, for any $\delta > 0$, $\text{var}(\overline{Y})/\delta^2 = \sigma_Y^2/(\delta^2 n) \longrightarrow 0$. It follows from Equation (18.6) that $\Pr(|\overline{Y} - \mu_Y| \geq \delta) \longrightarrow 0$ for every $\delta > 0$, proving the law of large numbers.

Some examples.   Consistency is a fundamental concept in asymptotic distribution theory, so we present some examples of consistent and inconsistent estimators of the population mean, μ_Y. Suppose that $Y_i, i = 1, \ldots, n$, are i.i.d. $N(0, \sigma_Y^2)$, where $0 < \sigma_Y^2 < \infty$. Consider the following three estimators of μ_Y: (1) $m_a = Y_1$; (2) $m_b = (\frac{1 - a^n}{1 - a})^{-1} \sum_{i=1}^{n} a^{i-1} Y_i$, where $0 < a < 1$; and (3) $m_c = \overline{Y} + 1/n$. Are these estimators consistent?

The first estimator, m_a, is just the first observation, so $E(m_a) = E(Y_1) = \mu_Y$ and m_a is unbiased. However, m_a is not consistent: $\Pr(|m_a - \mu_Y| \geq \delta) = \Pr(|Y_1 - \mu_Y| \geq \delta)$, which must be positive for sufficiently small δ (because $\sigma_Y^2 > 0$), so $\Pr(|m_a - \mu_Y| \geq \delta)$ does not tend to 0 as $n \to \infty$ and m_a is not consistent. This inconsistency should not be surprising: Because m_a uses the information in only one observation, its distribution cannot concentrate around μ_Y as the sample size increases.

The second estimator, m_b, is unbiased but is not consistent. It is unbiased because

$$E(m_b) = E\left[\left(\frac{1 - a^n}{1 - a}\right)^{-1}\sum_{i=1}^{n}a^{i-1}Y_i\right] = \left(\frac{1 - a^n}{1 - a}\right)^{-1}\sum_{i=1}^{n}a^{i-1}\mu_Y = \mu_Y,$$

$$\text{since } \sum_{i=1}^{n}a^{i-1} = (1 - a^n)\sum_{i=0}^{\infty}a^i = \frac{1 - a^n}{1 - a}.$$

The variance of m_b is

$$\text{var}(m_b) = \left(\frac{1 - a^n}{1 - a}\right)^2\sum_{i=1}^{n}a^{2(i-1)}\sigma_Y^2 = \sigma_Y^2\frac{(1 - a^{2n})(1 - a)^2}{(1 - a^2)(1 - a^n)^2} = \sigma_Y^2\frac{(1 + a^n)(1 - a)}{(1 - a^n)(1 + a)},$$

which has the limit $\text{var}(m_b) \to \sigma_Y^2(1 - a)/(1 + a) > 0$ as $n \to \infty$. Because Y is normally distributed, m_b is normally distributed with mean μ_Y and the variance given above. Thus m_b has a positive probability of falling outside any interval around μ_Y, so $\text{Pr}(|m_b - \mu_Y| \geq \delta)$ does not tend to 0 and m_b is inconsistent. This is perhaps surprising because this estimator uses all the observations. Most of the observations, however, receive very small weight (the weight of the ith observation is proportional to a^{i-1}, a very small number when i is large), and for this reason, there is an insufficient amount of cancellation of sampling errors for the estimator to be consistent.

The third estimator, m_c, is biased but consistent. Its bias is $1/n$: $E(m_c) = E(\overline{Y} + 1/n) = \mu_Y + 1/n$, so the bias tends to 0 as the sample size increases. To see why m_c is consistent, $\text{Pr}(|m_c - \mu_Y| \geq \delta) = \text{Pr}(|\overline{Y} + 1/n - \mu_Y| \geq \delta)$. Now, from Equation (18.43) in Appendix 18.2, a generalization of Chebychev's inequality implies that for any random variable W, $\text{Pr}(|W| \geq \delta) \leq E(W^2)/\delta^2$ for any positive constant δ. Thus $\text{Pr}(|\overline{Y} + 1/n - \mu_Y| \geq \delta) \leq E[(\overline{Y} + 1/n - \mu_Y)^2]/\delta^2$. But $E[(\overline{Y} + 1/n - \mu_Y)^2] = \text{var}(\overline{Y}) + 1/n^2 = \sigma^2/n + 1/n^2 \longrightarrow 0$ as n grows large. It follows that $\text{Pr}(|\overline{Y} + 1/n - \mu_Y| \geq \delta) \longrightarrow 0$ and m_c is consistent. This example illustrates the general point that an estimator can be biased in finite samples but if that bias vanishes as the sample size gets large, the estimator can still be consistent (Exercise 18.10).

The Central Limit Theorem and Convergence in Distribution

If the distributions of a sequence of random variables converge to a limit as $n \to \infty$, then the sequence of random variables is said to converge in distribution. The central limit theorem says that, under general conditions, the standardized sample average converges in distribution to a normal random variable.

Convergence in distribution. Let $F_1, F_2, \ldots, F_n, \ldots$ be a sequence of cumulative distribution functions corresponding to a sequence of random variables, $S_1, S_2, \ldots, S_n, \ldots$. For example, S_n might be the standardized sample average, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$. Then the sequence of random variables S_n is said to **converge in distribution**

to S (denoted $S_n \xrightarrow{d} S$) if the distribution functions $\{F_n\}$ converge to F, the distribution of S. That is,

$$S_n \xrightarrow{d} S \text{ if and only if } \lim_{n \to \infty} F_n = F(t), \tag{18.7}$$

where the limit holds at all points t at which the limiting distribution F is continuous. The distribution F is called the **asymptotic distribution** of S_n.

It is useful to contrast the concepts of convergence in probability ($\xrightarrow{p}$) and convergence in distribution ($\xrightarrow{d}$). If $S_n \xrightarrow{p} \mu$, then S_n becomes close to μ with high probability as n increases. In contrast, if $S_n \xrightarrow{d} S$, then the *distribution* of S_n becomes close to the *distribution* of S as n increases.

The central limit theorem. We now restate the central limit theorem using the concept of convergence in distribution. The central limit theorem in Key Concept 2.7 states that if $Y_1, \ldots, Y_n$ are i.i.d. and $0 < \sigma_Y^2 < \infty$, then the asymptotic distribution of $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$ is $N(0, 1)$. Because $\sigma_{\overline{Y}} = \sigma_Y/\sqrt{n}$, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}} = \sqrt{n}(\overline{Y} - \mu_Y)/\sigma_Y$. Thus the central limit theorem can be restated as $\sqrt{n}(\overline{Y} - \mu_Y) \xrightarrow{d} \sigma_Y Z$, where Z is a standard normal random variable. This means that the distribution of $\sqrt{n}(\overline{Y} - \mu_Y)$ converges to $N(0, \sigma_Y^2)$ as $n \longrightarrow \infty$. Conventional shorthand for this limit is

$$\sqrt{n}(\overline{Y} - \mu_Y) \xrightarrow{d} N(0, \sigma_Y^2). \tag{18.8}$$

That is, if $Y_1, \ldots, Y_n$ are i.i.d. and $0 < \sigma_Y^2 < \infty$, then the distribution of $\sqrt{n}(\overline{Y} - \mu_Y)$ converges to a normal distribution with mean 0 and variance σ_Y^2.

Extensions to time series data. The law of large numbers and central limit theorem stated in Section 2.6 apply to i.i.d. observations. As discussed in Chapter 14, the i.i.d. assumption is inappropriate for time series data, and these theorems need to be extended before they can be applied to time series observations. Those extensions are technical in nature in the sense that the conclusion is the same—versions of the law of large numbers and the central limit theorem apply to time series data—but the conditions under which they apply are different. This is discussed briefly in Section 16.4, but a mathematical treatment of asymptotic distribution theory for time series variables is beyond the scope of this text, and interested readers are referred to Hayashi (2000, Chapter 2).

Slutsky's Theorem and the Continuous Mapping Theorem

Slutsky's theorem combines consistency and convergence in distribution. Suppose that $a_n \xrightarrow{p} a$, where a is a constant, and $S_n \xrightarrow{d} S$. Then

$$a_n + S_n \xrightarrow{d} a + S, \; a_n S_n \xrightarrow{d} aS, \text{ and, if } a \neq 0, S_n/a_n \xrightarrow{d} S/a. \tag{18.9}$$

These three results are together called Slutsky's theorem.

The **continuous mapping theorem** concerns the asymptotic properties of a continuous function, g, of a sequence of random variables, S_n. The theorem has two parts. The first is that if S_n converges in probability to the constant a, then $g(S_n)$ converges in probability to $g(a)$; the second is that if S_n converges in distribution to S, then $g(S_n)$ converges in distribution to $g(S)$. That is, if g is a continuous function, then

$$
\begin{aligned}
&\text{(i) if } S_n \xrightarrow{p} a, \text{ then } g(S_n) \xrightarrow{p} g(a), \text{ and} \\
&\text{(ii) if } S_n \xrightarrow{d} S, \text{ then } g(S_n) \xrightarrow{d} g(S).
\end{aligned}
\tag{18.10}
$$

As an example of (i), if $s_Y^2 \xrightarrow{p} \sigma_Y^2$, then $\sqrt{s_Y^2} = s_Y \xrightarrow{p} \sigma_Y$. As an example of (ii), suppose that $S_n \xrightarrow{d} Z$, where Z is a standard normal random variable, and let $g(S_n) = S_n^2$. Because g is continuous, the continuous mapping theorem applies and $g(S_n) \xrightarrow{d} g(Z)$; that is, $S_n^2 \xrightarrow{d} Z^2$. In other words, the distribution of S_n^2 converges to the distribution of a squared standard normal random variable, which in turn has a χ_1^2 distribution; that is, $S_n^2 \xrightarrow{d} \chi_1^2$.

Application to the t-Statistic Based on the Sample Mean

We now use the central limit theorem, the law of large numbers, and Slutsky's theorem to prove that, under the null hypothesis, the t-statistic based on $\overline{Y}$ has a standard normal distribution when $Y_1, \ldots, Y_n$ are i.i.d. and $0 < E(Y_i^4) < \infty$.

The t-statistic for testing the null hypothesis that $E(Y_i) = \mu_0$ based on the sample average $\overline{Y}$ is given in Equations (3.8) and (3.11), and can be written

$$
t = \frac{\overline{Y} - \mu_0}{s_Y/\sqrt{n}} = \frac{\sqrt{n}(\overline{Y} - \mu_0)}{\sigma_Y} \div \frac{s_Y}{\sigma_Y},
\tag{18.11}
$$

where the second equality uses the trick of dividing both the numerator and the denominator by σ_Y.

Because $Y_1, \ldots, Y_n$ have two moments (which is implied by their having four moments; see Exercise 18.5) and because $Y_1, \ldots, Y_n$ are i.i.d., the first term after the final equality in Equation (18.11) obeys the central limit theorem: Under the null hypothesis, $\sqrt{n}(\overline{Y} - \mu_0)/\sigma_Y \xrightarrow{d} N(0, 1)$. In addition, $s_Y^2 \xrightarrow{p} \sigma_Y^2$ (as proven in Appendix 3.3), so $s_Y^2/\sigma_Y^2 \xrightarrow{p} 1$ and the ratio in the second term in Equation (18.11) tends to 1 (Exercise 18.4). Thus the expression after the final equality in Equation (18.11) has the form of the final expression in Equation (18.9), where [in the notation of Equation (18.9)] $S_n = \sqrt{n}(\overline{Y} - \mu_0)/\sigma_Y \xrightarrow{d} N(0, 1)$ and $a_n = s_Y/\sigma_Y \xrightarrow{p} 1$. It follows by applying Slutsky's theorem that $t \xrightarrow{d} N(0, 1)$.

18.3 Asymptotic Distribution of the OLS Estimator and *t*-Statistic

Recall from Chapter 4 that, under the assumptions of Key Concept 4.3 (the first three assumptions of Key Concept 18.1), the OLS estimator $\hat{\beta}_1$ is consistent, and $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ has an asymptotic normal distribution. Moreover, the *t*-statistic testing the null hypothesis $\beta_1 = \beta_{1,0}$ has an asymptotic standard normal distribution under the null hypothesis. This section summarizes these results and provides additional details of their proofs.

Consistency and Asymptotic Normality of the OLS Estimators

The large-sample distribution of $\hat{\beta}_1$, originally stated in Key Concept 4.4, is

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N\left(0, \frac{\text{var}(v_i)}{[\text{var}(X_i)]^2}\right), \tag{18.12}$$

where $v_i = (X_i - \mu_X)u_i$. The proof of this result was sketched in Appendix 4.3, but that proof omitted some details and involved an approximation that was not formally shown. The missing steps in that proof are left as Exercise 18.3.

An implication of Equation (18.12) is that $\hat{\beta}_1$ is consistent (Exercise 18.4).

Consistency of Heteroskedasticity-Robust Standard Errors

Under the first three least squares assumptions, the heteroskedasticity-robust standard error for $\hat{\beta}_1$ forms the basis for valid statistical inferences. Specifically,

$$\frac{\hat{\sigma}^2_{\hat{\beta}_1}}{\sigma^2_{\hat{\beta}_1}} \xrightarrow{p} 1, \tag{18.13}$$

where $\sigma^2_{\hat{\beta}_1} = \text{var}(v_i)/\{n[\text{var}(X_i)]^2\}$ and $\hat{\sigma}^2_{\hat{\beta}_1}$ is the square of the heteroskedasticity-robust standard error defined in Equation (5.4); that is,

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \frac{\dfrac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \hat{u}_i^2}{\left[\dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}. \tag{18.14}$$

To show the result in Equation (18.13), first use the definitions of $\sigma^2_{\hat{\beta}_1}$ and $\hat{\sigma}^2_{\hat{\beta}_1}$ to rewrite the ratio in Equation (18.13) as

$$\frac{\hat{\sigma}^2_{\hat{\beta}_1}}{\sigma^2_{\hat{\beta}_1}} = \left[\frac{n}{n-2}\right]\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2}{\text{var}(v_i)}\right] \div \left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}{\text{var}(X_i)}\right]^2. \quad (18.15)$$

We need to show that each of the three terms in brackets on the right-hand side of Equation (18.15) converges in probability to 1. Clearly, the first term converges to 1, and by the consistency of the sample variance (Appendix 3.3), the final term converges in probability to 1. Thus all that remains is to show that the second term converges in probability to 1—that is, that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2 \xrightarrow{p} \text{var}(v_i)$.

The proof that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2 \xrightarrow{p} \text{var}(v_i)$ proceeds in two steps. The first shows that $\frac{1}{n}\sum_{i=1}^{n}v_i^2 \xrightarrow{p} \text{var}(v_i)$; the second shows that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2 - \frac{1}{n}\sum_{i=1}^{n}v_i^2 \xrightarrow{p} 0$.

For the moment, suppose that X_i and u_i have eight moments [that is, $E(X_i^8) < \infty$ and $E(u_i^8) < \infty$], which is a stronger assumption than the four moments required by the third least squares assumption. To show the first step, we must show that $\frac{1}{n}\sum_{i=1}^{n}v_i^2$ obeys the law of large numbers in Equation (18.5). To do so, v_i^2 must be i.i.d. (which it is by the second least squares assumption), and $\text{var}(v_i^2)$ must be finite. To show that $\text{var}(v_i^2) < \infty$, apply the Cauchy–Schwarz inequality (Appendix 18.2): $\text{var}(v_i^2) \le E(v_i^4) = E[(X_i - \mu_X)^4u_i^4] \le \{E[(X_i - \mu_X)^8]E(u_i^8)\}^{1/2}$ Thus, if X_i and u_i have eight moments, then v_i^2 has a finite variance and thus satisfies the law of large numbers in Equation (18.5).

The second step is to prove that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2 - \frac{1}{n}\sum_{i=1}^{n}v_i^2 \xrightarrow{p} 0$. Because $v_i = (X_i - \mu_X)u_i$, this second step is the same as showing that

$$\frac{1}{n}\sum_{i=1}^{n}[(X_i - \overline{X})^2\hat{u}_i^2 - (X_i - \mu_X)^2u_i^2] \xrightarrow{p} 0. \quad (18.16)$$

Showing this result entails setting $\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)X_i$, expanding the term in Equation (18.16) in brackets, repeatedly applying the Cauchy–Schwarz inequality, and using the consistency of $\hat{\beta}_0$ and $\hat{\beta}_1$. The details of the algebra are left as Exercise 18.9.

The preceding argument supposes that X_i and u_i have eight moments. This is not necessary, however, and the result $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2 \xrightarrow{p} \text{var}(v_i)$ can be proven under the weaker assumption that X_i and u_i have four moments, as stated in the third least squares assumption. That proof, however, is beyond the scope of this text; see Hayashi (2000, Section 2.5) for details.

Asymptotic Normality of the Heteroskedasticity-Robust *t*-Statistic

We now show that, under the null hypothesis, the heteroskedasticity-robust OLS *t*-statistic testing the hypothesis $\beta_1 = \beta_{1,0}$ has an asymptotic standard normal distribution if least squares assumptions 1, 2, and 3 hold.

The *t*-statistic constructed using the heteroskedasticity-robust standard error $SE(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}$ [defined in Equation (18.14)] is

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\sqrt{n}(\hat{\beta}_1 - \beta_{1,0})}{\sqrt{n\sigma_{\hat{\beta}_1}^2}} \div \sqrt{\frac{\hat{\sigma}_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2}}. \tag{18.17}$$

It follows from Equation (18.12) and the definition of $\sigma_{\hat{\beta}_1}^2$ that the first term after the second equality in Equation (18.17) converges in distribution to a standard normal random variable. In addition, because the heteroskedasticity-robust standard error is consistent in the sense of Equation (18.13), $\sqrt{\hat{\sigma}_{\hat{\beta}_1}^2/\sigma_{\hat{\beta}_1}^2} \xrightarrow{p} 1$ (Exercise 18.4). It follows from Slutsky's theorem that $t \xrightarrow{d} N(0, 1)$.

18.4   Exact Sampling Distributions When the Errors Are Normally Distributed

In small samples, the distribution of the OLS estimator and *t*-statistic depends on the distribution of the regressors and regression error and typically is complicated. As discussed in Section 5.6, however, if the regression errors are homoskedastic and normally distributed, then these distributions are simple. Specifically, if all five extended least squares assumptions in Key Concept 18.1 hold, then the OLS estimator has a normal sampling distribution, conditional on $X_1, \ldots, X_n$. Moreover, the *t*-statistic has a Student *t* distribution. We present these results here for $\hat{\beta}_1$.

Distribution of $\hat{\beta}_1$ with Normal Errors

If the errors are i.i.d. normally distributed and independent of the regressors, then the distribution of $\hat{\beta}_1$, conditional on $X_1, \ldots, X_n$, is $N(\beta_1, \sigma_{\hat{\beta}_{1|X}}^2)$, where

$$\sigma_{\hat{\beta}_{1|X}}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{18.18}$$

The derivation of the normal distribution $N(\beta_1, \sigma_{\hat{\beta}_{1|X}}^2)$, conditional on $X_1, \ldots, X_n$, entails (i) establishing that the distribution is normal; (ii) showing that $E(\hat{\beta}_1|X_1, \ldots, X_n) = \beta_1$; and (iii) verifying Equation (18.18).

To show (i), note that, conditional on $X_1, \ldots, X_n$, $\hat{\beta}_1 - \beta_1$ is a weighted average of $u_1, \ldots, u_n$:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{18.19}$$

This equation was derived in Appendix 4.3 [Equation (4.28)] and is restated here for convenience. By extended least squares assumptions 1, 2, 4, and 5, u_i is i.i.d. $N(0, \sigma_u^2)$, and u_i and X_i are independently distributed. Because weighted averages of normally distributed variables are themselves normally distributed, it follows that $\hat{\beta}_1$ is normally distributed, conditional on $X_1, \ldots, X_n$.

To show (ii), take conditional expectations of both sides of Equation (18.19):
$E[(\hat{\beta}_1 - \beta_1)|X_1, \ldots, X_n)] = E[\sum_{i=1}^{n}(X_i - \overline{X})u_i / \sum_{i=1}^{n}(X_i - \overline{X})^2 | X_1, \ldots, X_n] = [\sum_{i=1}^{n}(X_i - \overline{X}) E(u_i | X_1, \ldots, X_n)]/[\sum_{i=1}^{n}(X_i - \overline{X})^2] = 0$, where the final equality follows because $E(u_i | X_1, X_2, \ldots, X_n) = E(u_i | X_i) = 0$ and because $\sum_{i=1}^{n}(X_i - \overline{X})^2 \neq 0$ by assumption. Thus $\hat{\beta}_1$ is conditionally unbiased; that is,

$$E(\hat{\beta}_1 | X_1, \ldots, X_n) = \beta_1. \tag{18.20}$$

To show (iii), use the fact that the errors are independently distributed, conditional on $X_1, \ldots, X_n$, to calculate the conditional variance of $\hat{\beta}_1$ using Equation (18.19):

$$\mathrm{var}(\hat{\beta}_1 | X_1, \ldots, X_n) = \mathrm{var}\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \Bigg| X_1, \ldots, X_n \right]$$

$$= \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 \mathrm{var}(u_i | X_1, \ldots, X_n)}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2 \right]^2} \tag{18.21}$$

$$= \frac{\sigma_u^2 \sum_{i=1}^{n}(X_i - \overline{X})^2}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2 \right]^2}.$$

Canceling the term in the numerator in the final expression in Equation (18.21) yields the formula for the conditional variance in Equation (18.18).

Distribution of the Homoskedasticity-Only *t*-Statistic

The homoskedasticity-only *t*-statistic testing the null hypothesis $\beta_1 = \beta_{1,0}$ is

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}, \tag{18.22}$$

where $SE(\hat{\beta}_1)$ is computed using the homoskedasticity-only standard error of $\hat{\beta}_1$. Substituting the formula for $SE(\hat{\beta}_1)$ [Equation (5.29) of Appendix 5.1] into Equation (18.22) and rearranging yields

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{s_{\hat{u}}^2 / \sum_{i=1}^{n}(X_i - \overline{X})^2}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\sigma_u^2 / \sum_{i=1}^{n}(X_i - \overline{X})^2}} \div \sqrt{\frac{s_{\hat{u}}^2}{\sigma_u^2}}$$

$$= \frac{(\hat{\beta}_1 - \beta_{1,0}) / \sigma_{\hat{\beta}_{1|X}}}{\sqrt{W/(n-2)}}, \tag{18.23}$$

where $s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$ and $W = \sum_{i=1}^{n}\hat{u}_i^2/\sigma_u^2$. Under the null hypothesis, $\hat{\beta}_1$ has an $N(\beta_{1,0}, \sigma_{\hat{\beta}_{1|X}}^2)$ distribution, conditional on $X_1, \ldots, X_n$, so the distribution of the numerator in the final expression in Equation (18.23) is $N(0, 1)$. It is shown in Section 19.4 that W has a chi-squared distribution with $n-2$ degrees of freedom and moreover that W is distributed independently of the standardized OLS estimator in the numerator of Equation (18.23). It follows from the definition of the Student t distribution (Appendix 18.1) that, under the five extended least squares assumptions, the homoskedasticity-only t-statistic has a Student t distribution with $n-2$ degrees of freedom.

Where does the degrees of freedom adjustment fit in? The degrees of freedom adjustment in $s_{\hat{u}}^2$ ensures that $s_{\hat{u}}^2$ is an unbiased estimator of σ_u^2 and that the t-statistic has a Student t distribution when the errors are normally distributed.

Because $W = \sum_{i=1}^{n}\hat{u}_i^2/\sigma_u^2$ is a chi-squared random variable with $n-2$ degrees of freedom, its mean is $E(W) = n-2$. Thus $E[W/(n-2)] = (n-2)/(n-2) = 1$. Rearranging the definition of W, we have that $E(\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2) = \sigma_u^2$. Thus the degrees of freedom correction makes $s_{\hat{u}}^2$ an unbiased estimator of σ_u^2. Also, by dividing by $n-2$ rather than n, the term in the denominator of the final expression of Equation (18.23) matches the definition of a random variable with a Student t distribution given in Appendix 18.1. That is, by using the degrees of freedom adjustment to calculate the standard error, the t-statistic has the Student t distribution when the errors are normally distributed.

18.5   Weighted Least Squares

Under the first four extended least squares assumptions, the OLS estimator is efficient among the class of linear (in $Y_1, \ldots, Y_n$), conditionally (on $X_1, \ldots, X_n$) unbiased estimators; that is, the OLS estimator is the best linear unbiased estimator (BLUE). This result is the Gauss–Markov theorem, which was discussed in Section 5.5 and proven in Appendix 5.2. The Gauss–Markov theorem provides a theoretical justification for using the OLS estimator. A major limitation of the Gauss–Markov theorem is that it requires homoskedastic errors. If, as is often encountered in practice, the errors are heteroskedastic, the Gauss–Markov theorem does not hold, and the OLS estimator is not BLUE.

This section presents a modification of the OLS estimator, called **weighted least squares (WLS)**, which is more efficient than OLS when the errors are heteroskedastic.

WLS requires knowing quite a bit about the conditional variance function, $\text{var}(u_i \mid X_i)$. We consider two cases. In the first case, $\text{var}(u_i \mid X_i)$ is known up to a factor of proportionality, and WLS is BLUE. In the second case, the functional form of $\text{var}(u_i \mid X_i)$ is known, but this functional form has some unknown parameters that can be estimated. Under some additional conditions, the asymptotic distribution of WLS

in the second case is the same as if the parameters of the conditional variance function were, in fact, known, and in this sense, the WLS estimator is asymptotically BLUE. The section concludes with a discussion of the practical advantages and disadvantages of handling heteroskedasticity using WLS or, alternatively, heteroskedasticity-robust standard errors.

WLS with Known Heteroskedasticity

Suppose that the conditional variance $\text{var}(u_i \mid X_i)$ is known up to a factor of proportionality; that is,

$$\text{var}(u_i \mid X_i) = \lambda h(X_i), \tag{18.24}$$

where λ is a constant and h is a known function. In this case, the WLS estimator is the estimator obtained by first dividing the dependent variable and regressor by the square root of h and then regressing this modified dependent variable on the modified regressor using OLS. Specifically, divide both sides of the single-variable regressor model by $\sqrt{h(X_i)}$ to obtain

$$\widetilde{Y}_i = \beta_0 \widetilde{X}_{0i} + \beta_1 \widetilde{X}_{1i} + \widetilde{u}_i, \tag{18.25}$$

where $\widetilde{Y}_i = Y_i / \sqrt{h(X_i)}$, $\widetilde{X}_{0i} = 1 / \sqrt{h(X_i)}$, $\widetilde{X}_{1i} = X_i / \sqrt{h(X_i)}$, and $\widetilde{u}_i = u_i / \sqrt{h(X_i)}$.

The **WLS estimator** is the OLS estimator of β_1 in Equation (18.25); that is, it is the estimator obtained by the OLS regression of $\widetilde{Y}_i$ on $\widetilde{X}_{0i}$ and $\widetilde{X}_{1i}$, where the coefficient on $\widetilde{X}_{0i}$ takes the place of the intercept in the unweighted regression.

Under the first three least squares assumptions in Key Concept 18.1 plus the known heteroskedasticity assumption in Equation (18.24), WLS is BLUE. The reason that the WLS estimator is BLUE is that weighting the variables has made the error term $\widetilde{u}_i$ in the weighted regression homoskedastic. That is,

$$\text{var}(\widetilde{u}_i \mid X_i) = \text{var}\left[\frac{u_i}{\sqrt{h(X_i)}} \mid X_i\right] = \frac{\text{var}(u_i \mid X_i)}{h(X_i)} = \frac{\lambda h(X_i)}{h(X_i)} = \lambda, \tag{18.26}$$

so the conditional variance of $\widetilde{u}_i$, $\text{var}(\widetilde{u}_i \mid X_i)$, is constant. Thus the first four least squares assumptions apply to Equation (18.25). Strictly speaking, the Gauss–Markov theorem was proven in Appendix 5.2 for Equation (18.1), which includes the intercept β_0, so it does not apply to Equation (18.25), in which the intercept is replaced by $\beta_0 \widetilde{X}_{0i}$. However, the extension of the Gauss–Markov theorem for multiple regression (Section 19.5) does apply to estimation of β_1 in the weighted population regression, Equation (18.25). Accordingly, the OLS estimator of β_1 in Equation (18.25)—that is, the WLS estimator of β_1—is BLUE.

In practice, the function h typically is unknown, so neither the weighted variables in Equation (18.25) nor the WLS estimator can be computed. For this reason, the WLS estimator described here is sometimes called the **infeasible WLS** estimator.

To implement WLS in practice, the function h must be estimated, the topic to which we now turn.

WLS with Heteroskedasticity of Known Functional Form

If the heteroskedasticity has a known functional form, then the heteroskedasticity function h can be estimated, and the WLS estimator can be calculated using this estimated function.

Example 1: The variance of u is quadratic in X. Suppose that the conditional variance is known to be the quadratic function

$$\text{var}(u_i \mid X_i) = \theta_0 + \theta_1 X_i^2, \tag{18.27}$$

where θ_0 and θ_1 are unknown parameters, $\theta_0 > 0$, and $\theta_1 \geq 0$.

Because θ_0 and θ_1 are unknown, it is not possible to construct the weighted variables $\widetilde{Y}_i$, $\widetilde{X}_{0i}$, and $\widetilde{X}_{1i}$. It is, however, possible to estimate θ_0 and θ_1 and to use those estimates to compute estimates of $\text{var}(u_i \mid X_i)$. Let $\hat{\theta}_0$ and $\hat{\theta}_1$ be estimators of θ_0 and θ_1, and let $\widehat{\text{var}}(u_i \mid X_i) = \hat{\theta}_0 + \hat{\theta}_1 X_i^2$. Define the weighted regressors as $\hat{\widetilde{Y}}_i = Y_i / \sqrt{\widehat{\text{var}}(u_i \mid X_i)}$, $\hat{\widetilde{X}}_{0i} = 1 / \sqrt{\widehat{\text{var}}(u_i \mid X_i)}$, and $\hat{\widetilde{X}}_{0i} = X_{1i} / \sqrt{\widehat{\text{var}}(u_i \mid X_i)}$. The WLS estimator is the OLS estimator of the coefficients in the regression of $\hat{\widetilde{Y}}_i$ on $\hat{\widetilde{X}}_{0i}$ and $\hat{\widetilde{X}}_{1i}$ (where $\beta_0 \hat{\widetilde{X}}_{0i}$ takes the place of the intercept β_0).

Implementation of this estimator requires estimating the conditional variance function—that is, estimating θ_0 and θ_1 in Equation (18.27). One way to estimate θ_0 and θ_1 consistently is to regress $\hat{u}_i^2$ on X_i^2 using OLS, where $\hat{u}_i^2$ is the square of the ith OLS residual.

Suppose that the conditional variance has the form in Equation (18.27) and that $\hat{\theta}_0$ and $\hat{\theta}_1$ are consistent estimators of θ_0 and θ_1. Under assumptions 1 through 3 of Key Concept 18.1 plus additional moment conditions that arise because θ_0 and θ_1 are estimated, the asymptotic distribution of the WLS estimator is the same as if θ_0 and θ_1 were known. Thus the WLS estimator with θ_0 and θ_1 estimated has the same asymptotic distribution as the infeasible WLS estimator and is in this sense asymptotically BLUE.

Because this method of WLS can be implemented by estimating unknown parameters of the conditional variance function, this method is sometimes called **feasible WLS** or *estimated WLS*.

Example 2: The variance depends on a third variable. WLS also can be used when the conditional variance depends on a third variable, W_i, which does not appear in the regression function. Specifically, suppose that data are collected on three variables, Y_i, X_i, and W_i, $i = 1, \ldots, n$; the population regression function depends on X_i but not W_i; and the conditional variance depends on W_i but not X_i. That is, the population regression function is $E(Y_i \mid X_i, W_i) = \beta_0 + \beta_1 X_i$, and the conditional variance

is $\text{var}(u_i \mid X_i, W_i) = \lambda h(W_i)$, where λ is a constant and h is a function that must be estimated.

For example, suppose that a researcher is interested in modeling the relationship between the unemployment rate in a state and a state economic policy variable (X_i). The measured unemployment rate (Y_i), however, is a survey-based estimate of the true unemployment rate (Y_i^*). Thus Y_i measures Y_i^* with error, where the source of the error is random survey error, so $Y_i = Y_i^* + v_i$, where v_i is the measurement error arising from the survey. In this example, it is plausible that the survey sample size, W_i, is not itself a determinant of the true state unemployment rate. Thus the population regression function does not depend on W_i; that is, $E(Y_i^* \mid X_i, W_i) = \beta_0 + \beta_1 X_i$. We therefore have the two equations,

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i^* \text{ and} \tag{18.28}$$

$$Y_i = Y_i^* + v_i, \tag{18.29}$$

where Equation (18.28) models the relationship between the state economic policy variable and the true state unemployment rate and Equation (18.29) represents the relationship between the measured unemployment rate Y_i and the true unemployment rate Y_i^*.

The model in Equations (18.28) and (18.29) can lead to a population regression in which the conditional variance of the error depends on W_i but not on X_i. The error term u_i^* in Equation (18.28) represents other factors omitted from this regression, while the error term v_i in Equation (18.29) represents measurement error arising from the unemployment rate survey. If u_i^* is homoskedastic, then $\text{var}(u_i^* \mid X_i, W_i) = \sigma_{u^*}^2$ is constant. The survey error variance, however, depends inversely on the survey sample size W_i; that is, $\text{var}(v_i \mid X_i, W_i) = a/W_i$, where a is a constant. Because v_i is random survey error, it is safely assumed to be uncorrelated with u_i^*, so $\text{var}(u_i^* + v_i \mid X_i, W_i) = \sigma_{u^*}^2 + a/W_i$. Thus, substituting Equation (18.28) into Equation (18.29) leads to the regression model with heteroskedasticity:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{18.30}$$

$$\text{var}(u_i \mid X_i, W_i) = \theta_0 + \theta_1 \left(\frac{1}{W_i} \right), \tag{18.31}$$

where $u_i = u_i^* + v_i$, $\theta_0 = \sigma_{u^*}^2$, $\theta_1 = a$, and $E(u_i \mid X_i, W_i) = 0$.

If θ_0 and θ_1 were known, then the conditional variance function in Equation (18.31) could be used to estimate β_0 and β_1 by WLS. In this example, θ_0 and θ_1 are unknown, but they can be estimated by regressing the squared OLS residual [from OLS estimation of Equation (18.30)] on $1/W_i$. Then the estimated conditional variance function can be used to construct the weights in feasible WLS.

It should be stressed that it is critical that $E(u_i \mid X_i, W_i) = 0$; if not, the weighted error will have a nonzero conditional mean, and WLS will be inconsistent. Said differently, if W_i is, in fact, a determinant of Y_i, then Equation (18.30) should be a multiple regression equation that includes both X_i and W_i.

General method of feasible WLS. In general, feasible WLS proceeds in five steps:

1. Regress Y_i on X_i by OLS, and obtain the OLS residuals $\hat{u}_i$, $i = 1, \ldots, n$.

2. Estimate a model of the conditional variance function, $\text{var}(u_i \mid X_i)$. For example, if the conditional variance function has the form in Equation (18.27), this entails regressing $\hat{u}_i^2$ on X_i^2. In general, this step entails estimating a function for the conditional variance, $\text{var}(u_i \mid X_i)$.

3. Use the estimated function to compute predicted values of the conditional variance function, $\widehat{\text{var}}(u_i \mid X_i)$.

4. Weight the dependent variable and regressor (including the intercept) by the inverse of the square root of the estimated conditional variance function.

5. Estimate the coefficients of the weighted regression by OLS; the resulting estimators are the WLS estimators.

When the variance of u depends on variables other than X (such as W in example 2), steps 2 and 3 are modified accordingly.

Regression software packages typically include optional weighted least squares commands that automate the fourth and fifth of these steps.

Heteroskedasticity-Robust Standard Errors or WLS?

There are two ways to handle heteroskedasticity: estimating β_0 and β_1 by WLS or estimating β_0 and β_1 by OLS and using heteroskedasticity-robust standard errors. Deciding which approach to use in practice requires weighing the advantages and disadvantages of each.

The advantage of WLS is that it is more efficient than the OLS estimator of the coefficients in the original regressors, at least asymptotically. The disadvantage of WLS is that it requires knowing the conditional variance function and estimating its parameters. If the conditional variance function has the quadratic form in Equation (18.27), this is easily done. In practice, however, the functional form of the conditional variance function is rarely known. Moreover, if the functional form is incorrect, then the standard errors computed by WLS regression routines are invalid in the sense that they lead to incorrect statistical inferences (tests have the wrong size).

The advantage of using heteroskedasticity-robust standard errors is that they produce asymptotically valid inferences even if you do not know the form of the conditional variance function. An additional advantage is that heteroskedasticity-robust standard errors are readily computed as an option in modern regression packages, so no additional effort is needed to safeguard against this threat. The disadvantage of heteroskedasticity-robust standard errors is that the OLS estimator will have a larger variance than the WLS estimator (based on the true conditional variance function).

In practice, the functional form of $\text{var}(u_i \mid X_i)$ is rarely, if ever, known, which poses a problem for using WLS in real-world applications. This problem is difficult enough with a single regressor, but in applications with multiple regressors, it is

even more difficult to know the functional form of the conditional variance. For this reason, practical use of WLS confronts imposing challenges. In contrast, in modern statistical packages it is simple to use heteroskedasticity-robust standard errors, and the resulting inferences are reliable under very general conditions; in particular, heteroskedasticity-robust standard errors can be used without needing to specify a functional form for the conditional variance. For these reasons, it is our opinion that, despite the theoretical appeal of WLS, heteroskedasticity-robust standard errors provide a better way to handle potential heteroskedasticity in most applications.[1]

Summary

1. The asymptotic normality of the OLS estimator, combined with the consistency of heteroskedasticity-robust standard errors, implies that, if the first three least squares assumptions in Key Concept 18.1 hold, then the heteroskedasticity-robust *t*-statistic has an asymptotic standard normal distribution under the null hypothesis.
2. If the regression errors are i.i.d. and normally distributed, conditional on the regressors, then $\hat{\beta}_1$ has an exact normal sampling distribution, conditional on the regressors. In addition, the homoskedasticity-only *t*-statistic has an exact Student t_{n-2} sampling distribution under the null hypothesis.
3. The weighted least squares (WLS) estimator is OLS applied to a weighted regression, where all variables are weighted by the square root of the inverse of the conditional variance, $\text{var}(u_i \mid X_i)$, or its estimate. Although the WLS estimator is asymptotically more efficient than OLS, to implement WLS you must know the functional form of the conditional variance function, which usually is a tall order.

Key Terms

---

[1]This chapter has focused on the case of a single treatment effect, β_1. Heterogeneous treatment effects introduce additional complications for WLS. Suppose that the treatment X is randomly assigned and the observations (experimental units) are randomly drawn from the population (assumption 2 in Key Concept 18.1). Then OLS is a consistent estimator of the average causal effect, but WLS is not (Exercise 18.13).

Review the Concepts

18.1 Suppose that assumption 4 in Key Concept 18.1 is true but you construct a 95%
confidence interval for β_1 using the heteroskedastic-robust standard error in a
large sample. Would this confidence interval be valid asymptotically in the sense
that it contained the true value of β_1 in 95% of all repeated samples for large n?
Suppose instead that assumption 4 in Key Concept 18.1 is false but you construct
a 95% confidence interval for β_1 using the homoskedasticity-only standard error
formula in a large sample. Would this confidence interval be valid asymptotically?

18.2 Suppose that A_n is a sequence of random variables that converges in prob-
ability to 3. Suppose that B_n is a sequence of random variables that converges
in distribution to a standard normal. What is the asymptotic distribution of
$A_n B_n$? Use this asymptotic distribution to compute an approximate value of
$\Pr(A_n B_n < 2)$.

18.3 Suppose that Y and X are related by the regression $Y = 1.0 + 2.0X + u$.
A researcher has observations on Y and X, where $0 \le X \le 20$, where
the conditional variance is $\text{var}(u_i \mid X_i = x) = 1$ for $0 \le x \le 10$ and
$\text{var}(u_i \mid X_i = x) = 16$ for $10 < x \le 20$. Draw a hypothetical scatterplot of the
observations $(X_i, Y_i), i = 1, \ldots, n$. Does WLS put more weight on observa-
tions with $x \le 10$ or $x > 10$? Why?

18.4 Instead of using WLS, the researcher in the previous problem decides to com-
pute the OLS estimator using only the observations for which $x \le 10$, then
using only the observations for which $x > 10$, and then using the average the
two OLS of estimators. Is this estimator more efficient than WLS?

Exercises

18.1 Consider the regression model without an intercept term, $Y_i = \beta_1 X_i + u_i$ (so
the true value of the intercept, β_0, is 0).

    a. Derive the least squares estimator of β_1 for the restricted regression model
$Y_i = \beta_1 X_i + u_i$. This is called the restricted least squares estimator ($\hat{\beta}_1^{RLS}$) of
β_1 because it is estimated under a restriction, which in this case is $\beta_0 = 0$.

b. Derive the asymptotic distribution of $\hat{\beta}_1^{RLS}$ under assumptions 1 through 3 of Key Concept 18.1.

c. Show that $\hat{\beta}_1^{RLS}$ is linear [Equation (5.24)] and, under assumptions 1 and 2 of Key Concept 18.1, conditionally unbiased [Equation (5.25)].

d. Derive the conditional variance of $\hat{\beta}_1^{RLS}$ under the Gauss–Markov conditions (assumptions 1 through 4 of Key Concept 18.1).

e. Compare the conditional variance of $\hat{\beta}_1^{RLS}$ in (d) to the conditional variance of the OLS estimator $\hat{\beta}_1$ (from the regression including an intercept) under the Gauss–Markov conditions. Which estimator is more efficient? Use the formulas for the variances to explain why.

f. Derive the exact sampling distribution of $\hat{\beta}_1^{RLS}$ under assumptions 1 through 5 of Key Concept 18.1.

g. Now consider the estimator $\widetilde{\beta}_1 = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} X_i$. Derive an expression for $\text{var}(\widetilde{\beta}_1 \mid X_1, \ldots, X_n) - \text{var}(\hat{\beta}_1^{RLS} \mid X_1, \ldots, X_n)$ under the Gauss–Markov conditions, and use this expression to show that $\text{var}(\widetilde{\beta}_1 \mid X_1, \ldots, X_n) \geq \text{var}(\hat{\beta}_1^{RLS} \mid X_1, \ldots, X_n)$.

18.2 Suppose that (X_i, Y_i) are i.i.d. with finite fourth moments. Prove that the sample covariance is a consistent estimator of the population covariance—that is, that $s_{XY} \xrightarrow{p} \sigma_{XY}$, where s_{XY} is defined in Equation (3.24). (Hint: Use the strategy outlined in Appendix 3.3 and the Cauchy–Schwarz inequality.)

18.3. This exercise fills in the details of the derivation of the asymptotic distribution of $\hat{\beta}_1$ given in Appendix 4.3.

a. Use Equation (18.19) to derive the expression

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} v_i}}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} - \frac{(\overline{X} - \mu_X)\sqrt{\frac{1}{n}\sum_{i=1}^{n} u_i}}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

where $v_i = (X_i - \mu_X)u_i$.

b. Use the central limit theorem, the law of large numbers, and Slutsky's theorem to show that the final term in the equation converges in probability to 0.

c. Use the Cauchy–Schwarz inequality and the third least squares assumption in Key Concept 18.1 to prove that $\text{var}(v_i) < \infty$. Does the term $\sqrt{\frac{1}{n}}\sum_{i=1}^{n} v_i / \sigma_v$ satisfy the central limit theorem?

d. Apply the central limit theorem and Slutsky's theorem to obtain the result in Equation (18.12).

18.4 Show the following results:

 a. Show that $\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, a^2)$, where a^2 is a constant, implies that $\hat{\beta}_1$ is consistent. (*Hint:* Use Slutsky's theorem.)

 b. Show that $s_u^2/\sigma_u^2 \xrightarrow{p} 1$ implies that $s_u/\sigma_u \xrightarrow{p} 1$.

18.5 Suppose that W is a random variable with $E(W^4) < \infty$. Show that $E(W^2) < \infty$.

18.6 Show that if $\hat{\beta}_1$ is conditionally unbiased, then it is unbiased; that is, show that if $E(\hat{\beta}_1 \mid X_1, \ldots, X_n) = \beta_1$, then $E(\hat{\beta}_1) = \beta_1$.

18.7 Suppose that X and u are continuous random variables and (X_i, u_i), $i = 1, \ldots, n$, are i.i.d.

 a. Show that the joint probability density function (p.d.f.) of (u_i, u_j, X_i, X_j) can be written as $f(u_i, X_i) f(u_j, X_j)$ for $i \neq j$, where $f(u_i, X_i)$ is the joint p.d.f. of u_i and X_i.

 b. Show that $E(u_i u_j \mid X_i, X_j) = E(u_i \mid X_i) E(u_j \mid X_j)$ for $i \neq j$.

 c. Show that $E(u_i \mid X_1, \ldots, X_n) = E(u_i \mid X_i)$.

 d. Show that $E(u_i u_j \mid X_1, X_2, \ldots, X_n) = E(u_i \mid X_i) E(u_j \mid X_j)$ for $i \neq j$.

18.8 Consider the regression model in Key Concept 18.1, and suppose that assumptions 1, 2, 3, and 5 hold. Suppose that assumption 4 is replaced by the assumption that $\text{var}(u_i \mid X_i) = \theta_0 + \theta_1 |X_i|$, where $|X_i|$ is the absolute value of X_i, $\theta_0 > 0$, and $\theta_1 \geq 0$.

 a. Is the OLS estimator of β_1 BLUE?

 b. Suppose that θ_0 and θ_1 are known. What is the BLUE estimator of β_1?

 c. Derive the exact sampling distribution of the OLS estimator, $\hat{\beta}_1$, conditional on $X_1, \ldots, X_n$.

 d. Derive the exact sampling distribution of the WLS estimator (treating θ_0 and θ_1 as known) of β_1, conditional on $X_1, \ldots, X_n$.

18.9 Prove Equation (18.16) under assumptions 1 and 2 of Key Concept 18.1 plus the assumption that X_i and u_i have eight moments.

18.10 Let $\hat{\theta}$ be an estimator of the parameter θ, where $\hat{\theta}$ might be biased. Show that if $E[(\hat{\theta} - \theta)^2] \longrightarrow 0$ as $n \longrightarrow \infty$ (that is, if the mean squared error of $\hat{\theta}$ tends to 0), then $\hat{\theta} \xrightarrow{p} \theta$. [*Hint*: Use Equation (18.43) with $W = \hat{\theta} - \theta$.]

18.11 Suppose that X and Y are distributed bivariate normal with the density given in Equation (18.38).

 a. Show that the density of Y given $X = x$ can be written as

$$f_{Y|X=x}(y) = \frac{1}{\sigma_{Y|X}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y - \mu_{Y|X}}{\sigma_{Y|X}}\right)^2\right]$$

where $\sigma_{Y|X} = \sqrt{\sigma_Y^2(1 - \rho_{XY}^2)}$ and $\mu_{Y|X} = \mu_Y + (\sigma_{XY}/\sigma_X^2)(x - \mu_X)$. [*Hint:* Use the definition of the conditional probability density $f_{Y|X=x}(y) = g_{X,Y}(x, y)/f_X(x)$, where $g_{X,Y}$ is the joint density of X and Y and f_X is the marginal density of X.]

b. Use the result in (a) to show that $Y|X = x \sim N(\mu_{Y|X}, \sigma_{Y|X}^2)$.

c. Use the result in (b) to show that $E(Y|X = x) = a + bx$ for suitably chosen constants a and b.

18.12 a. Suppose that $u \sim N(0, \sigma_u^2)$. Show that $E(e^u) = e^{\frac{1}{2}\sigma_u^2}$.

b. Suppose that the conditional distribution of u given $X = x$ is $N(0, a + bx^2)$, where a and b are positive constants. Show that $E(e^u|X = x) = e^{\frac{1}{2}(a+bx^2)}$.

18.13 Consider the heterogeneous regression model $Y_i = \beta_{0i} + \beta_{1i}X_i + u_i$, where β_{0i} and β_{1i} are random variables that differ from one observation to the next. Suppose that $E(u_i|X_i) = 0$ and (β_{0i}, β_{1i}) are distributed independently of X_i and that the observational units are randomly drawn from the population.

a. Let $\hat{\beta}_1^{OLS}$ denote the OLS estimator of β_1 given in Equation (18.2). Show that $\hat{\beta}_1^{OLS} \xrightarrow{p} E(\beta_1)$, where $E(\beta_1)$ is the average value of β_{1i} in the population. [Hint: See Equation (13.10).]

b. Suppose that $\text{var}(u_i|X_i) = \theta_0 + \theta_1 X_i^2$, where θ_0 and θ_1 are known positive constants. Let $\hat{\beta}_1^{WLS}$ denote the weighted least squares estimator. Does $\hat{\beta}_1^{WLS} \xrightarrow{p} E(\beta_1)$? Explain.

18.14 Suppose that $Y_i, i = 1, 2, \ldots, n$, are i.i.d. with $E(Y_i) = \mu$, $\text{var}(Y_i) = \sigma^2$, and finite fourth moments. Show the following:

a. $E(Y_i^2) = \mu^2 + \sigma^2$.

b. $\overline{Y} \xrightarrow{p} \mu$.

c. $\dfrac{1}{n}\sum_{i=1}^{n} Y_i^2 \xrightarrow{p} \mu^2 + \sigma^2$.

d. $\dfrac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \dfrac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2$.

e. $\dfrac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \xrightarrow{p} \sigma^2$.

f. $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \xrightarrow{p} \sigma^2$.

18.15 Z is distributed $N(0, 1)$, W is distributed χ_n^2, and V is distributed χ_m^2. Show, as $n \to \infty$ and m is fixed, that

a. $W/n \xrightarrow{p} 1$.

b. $\dfrac{Z}{\sqrt{W/n}} \xrightarrow{\ d\ } N(0, 1)$. Use the result to explain why the t_∞ distribution is the same as the standard normal distribution.

c. $\dfrac{V/m}{W/n} \xrightarrow{\ d\ } \chi_m^2/m$. Use the result to explain why the $F_{m,\infty}$ distribution is the same as the χ_m^2/m distribution.

18.1 The Normal and Related Distributions and Moments of Continuous Random Variables

This appendix defines and discusses the normal and related distributions. The definitions of the chi-squared, F, and Student t distributions, given in Section 2.4, are restated here for convenient reference. We begin by presenting definitions of probabilities and moments involving continuous random variables.

Probabilities and Moments of Continuous Random Variables

As discussed in Section 2.1, if Y is a continuous random variable, then its probability is summarized by its probability density function (p.d.f.). The probability that Y falls between two values is the area under its p.d.f. between those two values. Because Y is continuous, however, the mathematical expressions for its probabilities involve integrals rather than the summations that are appropriate for discrete random variables.

Let f_Y denote the probability density function of Y. Because probabilities cannot be negative, $f_Y(y) \geq 0$ for all y. The probability that Y falls between a and b (where $a < b$) is

$$\Pr(a \leq Y \leq b) = \int_a^b f_Y(y)dy. \tag{18.32}$$

Because Y must take on some value on the real line, $\Pr(-\infty \leq Y \leq \infty) = 1$, which implies that $\int_{-\infty}^{\infty} f_Y(y)\, dy = 1$.

Expected values and moments of continuous random variables, like those of discrete random variables, are probability-weighted averages of their values except that summations [for example, the summation in Equation (2.3)] are replaced by integrals. Accordingly, the expected value of Y is

$$E(Y) = \mu_Y = \int y f_Y(y)dy, \tag{18.33}$$

where the range of integration is the set of values for which f_Y is nonzero. The variance is the expected value of $(Y - \mu_Y)^2$, the rth moment of a random variable is the expected value of Y^r, and the rth central moment is the expected value of $(Y - \mu_Y)^r$. Thus

$$\text{var}(Y) = E(Y - \mu_Y)^2 = \int (y - \mu_Y)^2 f_Y(y) dy, \tag{18.34}$$

$$E(Y^r) = \int y^r f_Y(y) dy, \tag{18.35}$$

and similarly for the rth central moment, $E(Y - \mu_Y)^r$.

The Normal Distribution

The normal distribution for a single variable. The probability density function of a normally distributed random variable (the **normal probability density function (p.d.f.)**) is

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y - \mu}{\sigma}\right)^2\right], \tag{18.36}$$

where $\exp(x)$ is the exponential function of x. The factor $1/(\sigma\sqrt{2\pi})$ in Equation (18.36) ensures that $\Pr(-\infty \le Y \le \infty) = \int_{-\infty}^{\infty} f_Y(y) dy = 1$.

The mean of the normal distribution is μ, and its variance is σ^2. The normal distribution is symmetric, so all odd central moments of order three and greater are 0. The fourth central moment is $3\sigma^4$. In general, if Y is distributed $N(\mu, \sigma^2)$, then its even central moments are given by

$$E(Y - \mu)^k = \frac{k!}{2^{k/2}(k/2)!} \sigma^k \ (k \text{ even}). \tag{18.37}$$

When $\mu = 0$ and $\sigma^2 = 1$, the normal distribution is called the standard normal distribution. The standard normal p.d.f. is denoted ϕ, and the standard normal cumulative distribution function (c.d.f.) is denoted Φ. Thus the standard normal density is $\phi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$ and $\Phi(y) = \int_{-\infty}^{y} \phi(s) ds$.

The bivariate normal distribution. The **bivariate normal p.d.f.** for the two random variables X and Y is

$$g_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - \rho_{XY}^2}} \times \exp\left\{\frac{1}{-2(1 - \rho_{XY}^2)}\left[\left(\frac{x - \mu_X}{\sigma_X}\right)^2 - \right.\right.$$
$$\left.\left. 2\rho_{XY}\left(\frac{x - \mu_X}{\sigma_X}\right)\left(\frac{y - \mu_Y}{\sigma_Y}\right) + \left(\frac{y - \mu_Y}{\sigma_Y}\right)^2\right]\right\}, \tag{18.38}$$

where ρ_{XY} is the correlation between X and Y.

When X and Y are uncorrelated ($\rho_{XY} = 0$), $g_{X,Y}(x, y) = f_X(x)f_Y(y)$, where f is the normal density given in Equation (18.36). This proves that if X and Y are jointly normally distributed and are uncorrelated, then they are independently distributed. This is a special feature of the normal distribution that is typically not true for other distributions.

The multivariate normal distribution extends the bivariate normal distribution to handle more than two random variables. This distribution is most conveniently stated using matrices and is presented in Appendix 19.1.

The conditional normal distribution. Suppose that X and Y are jointly normally distributed. Then the conditional distribution of Y given X is $N(\mu_{Y|X}, \sigma_{Y|X}^2)$, with mean

$\mu_{Y|X} = \mu_Y + (\sigma_{XY}/\sigma_X^2)(X - \mu_X)$ and variance $\sigma_{Y|X}^2 = (1 - \rho_{XY}^2)\sigma_Y^2$. The mean of this conditional distribution, conditional on $X = x$, is a linear function of x, and the variance does not depend on x (Exercise 18.11).

Related Distributions

The chi-squared distribution. Let $Z_1, Z_2, \ldots, Z_n$ be n i.i.d. standard normal random variables. The random variable

$$W = \sum_{i=1}^{n} Z_i^2 \tag{18.39}$$

has a chi-squared distribution with n degrees of freedom. This distribution is denoted χ_n^2. Because $E(Z_i^2) = 1$ and $E(Z_i^4) = 3$, $E(W) = n$ and $\text{var}(W) = 2n$.

The Student t distribution. Let Z have a standard normal distribution, let W have a χ_m^2 distribution, and let Z and W be independently distributed. Then the random variable

$$t = \frac{Z}{\sqrt{W/m}} \tag{18.40}$$

has a Student t distribution with m degrees of freedom, denoted t_m. The t_∞ distribution is the standard normal distribution. (See Exercise 18.15.)

The F distribution. Let W_1 and W_2 be independent random variables with chi-squared distributions with respective degrees of freedom n_1 and n_2. Then the random variable

$$F = \frac{W_1/n_1}{W_2/n_2} \tag{18.41}$$

has an F distribution with (n_1, n_2) degrees of freedom. This distribution is denoted F_{n_1, n_2}.

The F distribution depends on the numerator degrees of freedom n_1 and the denominator degrees of freedom n_2. As number of degrees of freedom in the denominator gets large, the F_{n_1, n_2} distribution is well approximated by a $\chi_{n_1}^2$ distribution, divided by n_1. In the limit, the $F_{n_1, \infty}$ distribution is the same as the $\chi_{n_1}^2$ distribution, divided by n_1; that is, it is the same as the $\chi_{n_1}^2/n_1$ distribution. (See Exercise 18.15.)

APPENDIX

18.2 Two Inequalities

This appendix states and proves Chebychev's inequality and the Cauchy–Schwarz inequality.

Chebychev's Inequality

Chebychev's inequality uses the variance of the random variable V to bound the probability that V is farther than $\pm \delta$ from its mean, where δ is a positive constant:

$$\Pr(|V - \mu_V| \geq \delta) \leq \frac{\text{var}(V)}{\delta^2} \text{ (Chebychev's inequality).} \tag{18.42}$$

To prove Equation (18.42), let $W = V - \mu_V$, let f be the p.d.f. of W, and let δ be any positive number. Now

$$
\begin{aligned}
E(W^2) &= \int_{-\infty}^{\infty} w^2 f(w) dw \\
&= \int_{-\infty}^{-\delta} w^2 f(w) dw + \int_{-\delta}^{\delta} w^2 f(w) dw + \int_{\delta}^{\infty} w^2 f(w) dw \\
&\geq \int_{-\infty}^{-\delta} w^2 f(w) dw + \int_{\delta}^{\infty} w^2 f(w) dw \\
&\geq \delta^2 \left[\int_{-\infty}^{-\delta} f(w) dw + \int_{\delta}^{\infty} f(w) dw \right] \\
&= \delta^2 \Pr(|W| \geq \delta),
\end{aligned}
\tag{18.43}
$$

where the first equality is the definition of $E(W^2)$, the second equality holds because the ranges of integration divide up the real line, the first inequality holds because the term that was dropped is nonnegative, the second inequality holds because $w^2 \geq \delta^2$ over the range of integration, and the final equality holds by the definition of $\Pr(|W| \geq \delta)$. Substituting $W = V - \mu_v$ into the final expression, noting that $E(W^2) = E[(V - \mu_V)^2] = \mathrm{var}(V)$, and rearranging yields the inequality given in Equation (18.42). If V is discrete, this proof applies with summations replacing integrals.

The Cauchy–Schwarz Inequality

The Cauchy–Schwarz inequality is an extension of the correlation inequality, $|\rho_{XY}| \leq 1$, to incorporate nonzero means. The Cauchy–Schwarz inequality is

$$
|E(XY)| \leq \sqrt{E(X^2)E(Y^2)} \quad \text{(Cauchy–Schwarz inequality)}. \tag{18.44}
$$

The proof of Equation (18.44) is similar to the proof of the correlation inequality in Appendix 2.1. Let $W = Y + bX$, where b is a constant. Then $E(W^2) = E(Y^2) + 2bE(XY) + b^2 E(X^2)$. Now let $b = -E(XY)/E(X^2)$, so that (after simplification) the expression becomes $E(W^2) = E(Y^2) - [E(XY)]^2 / E(X^2)$. Because $E(W^2) \geq 0$ (since $W^2 \geq 0$), it must be the case that $[E(XY)]^2 \leq E(X^2)E(Y^2)$, and the Cauchy–Schwarz inequality follows by taking the square root.

The Theory
of Multiple Regression

This chapter provides an introduction to the theory of multiple regression analysis. The chapter has four objectives. The first is to present the multiple regression model in matrix form, which leads to compact formulas for the ordinary least squares (OLS) estimator and test statistics. The second objective is to characterize the sampling distribution of the OLS estimator, both in large samples (using asymptotic theory) and in small samples (if the errors are homoskedastic and normally distributed). The third objective is to study the theory of efficient estimation of the coefficients of the multiple regression model and to describe generalized least squares (GLS), a method for estimating the regression coefficients efficiently when the errors are heteroskedastic and/or correlated across observations. The fourth objective is to provide a concise treatment of the asymptotic distribution theory of instrumental variables (IV) regression in the linear model, including an introduction to generalized method of moments (GMM) estimation in the linear IV regression model with heteroskedastic errors.

The chapter begins by laying out the multiple regression model and the OLS estimator in matrix form in Section 19.1. This section also presents the extended least squares assumptions for the multiple regression model. The first four of these assumptions are the same as the least squares assumptions of Key Concept 6.4 and underlie the asymptotic distributions used to justify the procedures described in Chapters 6 and 7. The remaining two extended least squares assumptions are stronger and permit us to explore in more detail the theoretical properties of the OLS estimator in the multiple regression model.

The next three sections examine the sampling distribution of the OLS estimator and test statistics. Section 19.2 presents the asymptotic distributions of the OLS estimator and t-statistic under the least squares assumptions of Key Concept 6.4. Section 19.3 unifies and generalizes the tests of hypotheses involving multiple coefficients presented in Sections 7.2 and 7.3 and provides the asymptotic distribution of the resulting F-statistic. In Section 19.4, we examine the exact sampling distributions of the OLS estimator and test statistics in the special case that the errors are homoskedastic and normally distributed. Although the assumption of homoskedastic normal errors is implausible in most econometric applications, the exact sampling distributions are of theoretical interest, and p-values computed using these distributions often appear in the output of regression software.

The next two sections turn to the theory of efficient estimation of the coefficients of the multiple regression model. Section 19.5 generalizes the Gauss–Markov theorem to multiple regression. Section 19.6 develops the method of generalized least squares (GLS).

The final section takes up IV estimation in the general IV regression model when the instruments are valid and strong. This section derives the asymptotic distribution of the two stage least squares (TSLS) estimator when the errors are heteroskedastic and provides expressions for the standard error of the TSLS estimator. The TSLS estimator is one of many possible GMM estimators, and this section provides an introduction to GMM estimation in the linear IV regression model. It is shown that the TSLS estimator is the efficient GMM estimator if the errors are homoskedastic.

Mathematical prerequisite. The treatment of the linear model in this chapter uses matrix notation and the basic tools of linear algebra and assumes that the reader has taken an introductory course in linear algebra. Appendix 19.1 reviews vectors, matrices, and the matrix operations used in this chapter. In addition, multivariate calculus is used in Section 19.1 to derive the OLS estimator.

19.1   The Linear Multiple Regression Model and OLS Estimator in Matrix Form

The linear multiple regression model and the OLS estimator can each be represented compactly using matrix notation.

The Multiple Regression Model in Matrix Notation

The population multiple regression model (Key Concept 6.2) is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, i = 1, \ldots, n. \qquad (19.1)$$

To write the multiple regression model in matrix form, define the following vectors and matrices:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{U} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix}, \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad (19.2)$$

so $\boldsymbol{Y}$ is $n \times 1$, $\boldsymbol{X}$ is $n \times (k + 1)$, $\boldsymbol{U}$ is $n \times 1$, and $\boldsymbol{\beta}$ is $(k + 1) \times 1$. Throughout we denote matrices and vectors by bold type. In this notation,

- $\boldsymbol{Y}$ is the $n \times 1$ dimensional vector of n observations on the dependent variable.

- $\boldsymbol{X}$ is the $n \times (k + 1)$ dimensional matrix of n observations on the $k + 1$ regressors (including the "constant" regressor for the intercept).

- The $(k + 1) \times 1$ dimensional column vector $\boldsymbol{X}_i$ is the ith observation on the $k + 1$ regressors; that is, $\boldsymbol{X}_i' = (1 \ X_{1i} \ldots X_{ki})$, where $\boldsymbol{X}_i'$ denotes the transpose of $\boldsymbol{X}_i$.

The Extended Least Squares Assumptions in the Multiple Regression Model

$$Y_i = X_i'\boldsymbol{\beta} + u_i, i = 1, \ldots, n., \tag{19.3}$$

where $\boldsymbol{\beta}$ is the vector of causal effects and

1. $E(u_i \mid X_i) = 0$ (u_i has conditional mean 0);
2. $(X_i, Y_i), i = 1, \ldots, n$, are independently and identically distributed (i.i.d.) draws from their joint distribution;
3. X_i and u_i have nonzero finite fourth moments;
4. X has full column rank (there is no perfect multicollinearity);
5. $\text{var}(u_i \mid X_i) = \sigma_u^2$ (homoskedasticity); and
6. The conditional distribution of u_i given X_i is normal (normal errors).

- U is the $n \times 1$ dimensional vector of the n error terms.
- $\boldsymbol{\beta}$ is the $(k + 1) \times 1$ dimensional vector of the $k + 1$ unknown regression coefficients.

The multiple regression model in Equation (19.1) for the ith observation, written using the vectors $\boldsymbol{\beta}$ and X_i, is

$$Y_i = X_i'\boldsymbol{\beta} + u_i, i = 1, \ldots, n. \tag{19.4}$$

In Equation (19.4), the first regressor is the "constant" regressor that always equals 1, and its coefficient is the intercept. Thus the intercept does not appear separately in Equation (19.4); rather, it is the first element of the coefficient vector $\boldsymbol{\beta}$.

Stacking all n observations in Equation (19.4) yields the multiple regression model in matrix form:

$$Y = X\boldsymbol{\beta} + U. \tag{19.5}$$

The Extended Least Squares Assumptions

The extended least squares assumptions for the multiple regression model are the four least squares assumptions for causal inference in the multiple regression model in Key Concept 6.4 plus the two additional assumptions of homoskedasticity and normally distributed errors. The assumption of homoskedasticity is used when we study the efficiency of the OLS estimator, and the assumption of normality is used when we study the exact sampling distribution of the OLS estimator and test statistics.

The extended least squares assumptions are summarized in Key Concept 19.1.

Except for notational differences, the first three assumptions in Key Concept 19.1 are identical to the first three assumptions in Key Concept 6.4.

The fourth assumptions in Key Concepts 6.4 and 19.1 might appear different, but, in fact, they are the same: They are simply different ways of saying that there cannot be perfect multicollinearity. Recall that perfect multicollinearity arises when one regressor can be written as a perfect linear combination of the others. In the matrix notation of Equation (19.2), perfect multicollinearity means that one column of X is a perfect linear combination of the other columns of X, but if this is true, then X does not have full column rank. Thus saying that X has rank $k + 1$—that is, rank equal to the number of columns of X—is just another way to say that the regressors are not perfectly multicollinear.

The fifth least squares assumption in Key Concept 19.1 is that the error term is conditionally homoskedastic, and the sixth assumption is that the conditional distribution of u_i given X_i is normal. These two assumptions are the same as the final two assumptions in Key Concept 18.1 except that they are now stated for multiple regressors.

Implications for the mean vector and covariance matrix of **U.** The least squares assumptions in Key Concept 19.1 imply simple expressions for the mean vector and covariance matrix of the conditional distribution of U given the matrix of regressors X. (The mean vector and covariance matrix of a vector of random variables are defined in Appendix 19.2.) Specifically, the first and second assumptions in Key Concept 19.1 imply that $E(u_i|X) = E(u_i|X_i) = 0$ and that $\text{cov}(u_i, u_j|X) = E(u_i u_j|X) = E(u_i u_j|X_i, X_j) = E(u_i|X_i)E(u_j|X_j) = 0$ for $i \neq j$ (Exercise 18.7). The first, second, and fifth assumptions imply that $E(u_i^2|X) = E(u_i^2|X_i) = \sigma_u^2$. Combining these results, we have that

$$\text{under assumptions 1 and 2, } E(U|X) = \mathbf{0}_n, \text{ and} \tag{19.6}$$

$$\text{under assumptions 1, 2, and 5, } E(UU'|X) = \sigma_u^2 I_n, \tag{19.7}$$

where $\mathbf{0}_n$ is the n-dimensional vector of zeros and I_n is the $n \times n$ identity matrix.

Similarly, the first, second, fifth, and sixth assumptions in Key Concept 19.1 imply that the conditional distribution of the n-dimensional random vector U, conditional on X, is the multivariate normal distribution (defined in Appendix 19.2). That is,

$$\text{under assumptions 1, 2, 5, and 6, the}$$

$$\text{conditional distribution of } U \text{ given } X \text{ is } N(\mathbf{0}_n, \sigma_u^2 I_n). \tag{19.8}$$

The OLS Estimator

The OLS estimator minimizes the sum of squared prediction mistakes, $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2$ [Equation (6.8)]. The formula for the OLS estimator is obtained by taking the derivative of the sum of squared prediction

mistakes with respect to each element of the coefficient vector, setting these deriva-tives to 0, and solving for the estimator $\hat{\boldsymbol{\beta}}$.

The derivative of the sum of squared prediction mistakes with respect to the jth regression coefficient, b_j, is

$$\frac{\partial}{\partial b_j} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2$$

$$= -2 \sum_{i=1}^{n} X_{ji}(Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki}) \tag{19.9}$$

for $j = 0, \ldots, k$, where, for $j = 0$, $X_{0i} = 1$ for all i. The derivative on the right-hand side of Equation (19.9) is the jth element of the $k + 1$ dimensional vector, $-2\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{Xb})$, where $\boldsymbol{b}$ is the $k + 1$ dimensional vector consisting of $b_0, \ldots, b_k$. There are $k + 1$ such derivatives, each corresponding to an element of $\boldsymbol{b}$. Combined, these yield the system of $k + 1$ equations that, when set to 0, constitute the first-order conditions for the OLS estimator $\hat{\boldsymbol{\beta}}$. That is, $\hat{\boldsymbol{\beta}}$ solves the system of $k + 1$ equations:

$$\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}_{k+1} \tag{19.10}$$

or, equivalently, $\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Solving the system of equations (19.10) yields the OLS estimator $\hat{\boldsymbol{\beta}}$ in matrix form:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{19.11}$$

where $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is the inverse of the matrix $\boldsymbol{X}'\boldsymbol{X}$.

The role of "no perfect multicollinearity." The fourth least squares assumption in Key Concept 19.1 states that $\boldsymbol{X}$ has full column rank. In turn, this implies that the matrix $\boldsymbol{X}'\boldsymbol{X}$ has full rank—that is, that $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular. Because $\boldsymbol{X}'\boldsymbol{X}$ is nonsingu-lar, it is invertible. Thus the assumption that there is no perfect multicollinearity ensures that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists, so Equation (19.10) has a unique solution and the for-mula in Equation (19.11) for the OLS estimator can actually be computed. Said dif-ferently, if $\boldsymbol{X}$ does *not* have full column rank, there is not a unique solution to Equation (19.10), and $\boldsymbol{X}'\boldsymbol{X}$ is singular. Therefore, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ cannot be computed, and thus $\hat{\boldsymbol{\beta}}$ cannot be computed from Equation (19.11).

19.2   Asymptotic Distribution of the OLS Estimator and *t*-Statistic

If the sample size is large and the first four assumptions of Key Concept 19.1 are satisfied, then the OLS estimator has an asymptotic joint normal distribution, the heteroskedasticity-robust estimator of the covariance matrix is consistent, and the

The Multivariate Central Limit Theorem

Suppose that $W_1, \ldots, W_n$ are i.i.d. m-dimensional random variables with mean vector $E(W_i) = \mu_W$ and covariance matrix $E[(W_i - \mu_W)(W_i - \mu_W)'] = \Sigma_W$, where Σ_W is positive definite and finite. Let $\overline{W} = \frac{1}{n}\sum_{i=1}^{n} W_i$. Then $\sqrt{n}(\overline{W} - \mu_W) \xrightarrow{d} N(0_m, \Sigma_W)$.

heteroskedasticity-robust OLS t-statistic has an asymptotic standard normal distribution. These results make use of the multivariate normal distribution (Appendix 19.2) and a multivariate extension of the central limit theorem.

The Multivariate Central Limit Theorem

The central limit theorem of Key Concept 2.7 applies to a one-dimensional random variable. To derive the *joint* asymptotic distribution of the elements of $\hat{\beta}$, we need a multivariate central limit theorem that applies to vector-valued random variables.

The multivariate central limit theorem extends the univariate central limit theorem to averages of observations on a vector-valued random variable, W, where W is m-dimensional. The difference between the central limit theorems for a scalar-valued random variable and that for a vector-valued random variable is the conditions on the variances. In the scalar case in Key Concept 2.7, the requirement is that the variance is both nonzero and finite. In the vector case, the requirement is that the covariance matrix is both positive definite and finite. If the vector-valued random variable W has a finite positive definite covariance matrix, then $0 < \text{var}(c'W) < \infty$ for all nonzero m-dimensional vectors c (Exercise 19.3).

The multivariate central limit theorem that we will use is stated in Key Concept 19.2.

Asymptotic Normality of $\hat{\beta}$

In large samples, the OLS estimator has the multivariate normal asymptotic distribution

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0_{k+1}, \Sigma_{\sqrt{n}(\hat{\beta}-\beta)}), \text{ where } \Sigma_{\sqrt{n}(\hat{\beta}-\beta)} = Q_X^{-1}\Sigma_V Q_X^{-1}, \quad (19.12)$$

where Q_X is the $(k + 1) \times (k + 1)$ dimensional matrix of second moments of the regressors—that is, $Q_X = E(X_i X_i')$—and Σ_V is the $(k + 1) \times (k + 1)$ dimensional covariance matrix of $V_i = X_i u_i$—that is, $\Sigma_V = E(V_i V_i')$. Note that the second least squares assumption in Key Concept 19.1 implies that V_i, $i = 1, \ldots, n$, are i.i.d.

Written in terms of $\hat{\beta}$ rather than $\sqrt{n}(\hat{\beta} - \beta)$, the normal approximation in Equation (19.12) is

$\hat{\beta}$, in large samples, is approximately distributed $N(\beta, \Sigma_{\hat{\beta}})$,

$$\text{where } \Sigma_{\hat{\beta}} = \Sigma_{\sqrt{n}(\hat{\beta}-\beta)}/n = Q_X^{-1}\Sigma_V Q_X^{-1}/n. \quad (19.13)$$

The covariance matrix $\Sigma_{\hat{\boldsymbol{\beta}}}$ in Equation (19.13) is the covariance matrix of the approximate normal distribution of $\hat{\boldsymbol{\beta}}$, whereas $\Sigma_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}$ in Equation (19.12) is the covariance matrix of the asymptotic normal distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. These two covariance matrices differ by a factor of n, depending on whether the OLS estimator is scaled by $\sqrt{n}$.

Derivation of Equation (19.12). To derive Equation (19.12), first use Equations (19.3) and (19.11) to write $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\boldsymbol{\beta} + U)$, so that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (X'X)^{-1}X'U. \tag{19.14}$$

Thus $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (X'X)^{-1}X'U$, so

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'U}{\sqrt{n}}\right). \tag{19.15}$$

The derivation of Equation (19.12) involves arguing first that the "denominator" matrix in Equation (19.15), $X'X/n$, is consistent for Q_X and second that the "numerator" matrix, $X'U/\sqrt{n}$, obeys the multivariate central limit theorem in Key Concept 19.2. The details are given in Appendix 19.3.

Heteroskedasticity-Robust Standard Errors

The heteroskedasticity-robust estimator of $\Sigma_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}$ is obtained by replacing the population moments in its definition [Equation (19.12)] by sample moments. Accordingly, the heteroskedasticity-robust estimator of the covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is

$$\hat{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})} = \left(\frac{X'X}{n}\right)^{-1}\hat{\Sigma}_{\hat{V}}\left(\frac{X'X}{n}\right)^{-1}, \text{ where } \hat{\Sigma}_{\hat{V}} = \frac{1}{n-k-1}\sum_{i=1}^{n}X_iX_i'\hat{u}_i^2, \tag{19.16}$$

The estimator $\hat{\Sigma}_{\hat{V}}$ incorporates the same degrees-of-freedom adjustment that is in the standard error of the regression (*SER*) for the multiple regression model (Section 6.4) to adjust for potential downward bias because of estimation of $k + 1$ regression coefficients.

The proof that $\hat{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})} \xrightarrow{p} \Sigma_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}$ is conceptually similar to the proof, presented in Section 18.3, of the consistency of heteroskedasticity-robust standard errors for the single-regressor model.

Heteroskedasticity-robust standard errors. The heteroskedasticity-robust estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}, \Sigma_{\hat{\boldsymbol{\beta}}}$, is

$$\hat{\Sigma}_{\hat{\boldsymbol{\beta}}} = n^{-1}\hat{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}. \tag{19.17}$$

The heteroskedasticity-robust standard error for the jth regression coefficient is the square root of the jth diagonal element of $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$. That is, the heteroskedasticity-robust standard error of the jth coefficient is

$$SE(\hat{\boldsymbol{\beta}}_j) = \sqrt{(\hat{\Sigma}_{\hat{\boldsymbol{\beta}}})_{jj}}, \tag{19.18}$$

where $(\hat{\Sigma}_{\hat{\boldsymbol{\beta}}})_{jj}$ is the (j, j) element of $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$.

Other heteroskedasticity-robust variance estimators. The variance estimator in Equation (19.16) is called the HC1 variance estimator. The HC1 estimator is the most commonly used in practice, but it is not the only heteroskedasticity-robust variance estimator. Simulation studies have found that, in small samples, the HC1 estimator can be biased down, yielding standard errors that are too small. Long and Ervin (2000) provide simulation evidence that in small samples HC1 can be improved upon by a variant that weights each squared residual by a function of the X's. Imbens and Kolesar (2016) point out that, in addition to this bias, in small samples the sampling variability of the variance estimator makes the normal approximation a poor one, and they suggest using instead a t approximation to the t-statistic, along with a different variance estimator than HC1 or that suggested by Long and Ervin (2000). Angrist and Pischke (2009) suggest, however, that when the sample size exceeds 50, the HC1 estimator leads to negligible size distortions. Consistent with modern econometric practice, this text focuses on large samples, for which the HC1 estimator works well.

Confidence Intervals for Predicted Effects

Section 8.1 describes two methods for computing the standard error of predicted effects that involve changes in two or more regressors. There are compact matrix expressions for these standard errors and thus for confidence intervals for predicted effects.

Consider a change in the value of the regressors for the ith observation from some initial value—say, $X_{i,0}$—to some new value—$X_{i,0} + d$—so that the change in X_i is $\Delta X_i = d$, where d is a $k + 1$ dimensional vector. This change in X can involve multiple regressors (that is, multiple elements of X_i). For example, if two of the regressors are the value of an independent variable and its square, then d is the difference between the subsequent and initial values of these two variables.

The expected effect of this change in X_i is $d'\beta$, and the estimator of this effect is $d'\hat{\beta}$. Because linear combinations of normally distributed random variables are themselves normally distributed, $\sqrt{n}(d'\hat{\beta} - d'\beta) = d'\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, d'\Sigma_{\sqrt{n}(\hat{\beta} - \beta)}d)$. Thus the standard error of this predicted effect is $(d'\hat{\Sigma}_{\hat{\beta}}d)^{1/2}$. A 95% confidence interval for this predicted effect is

$$d'\hat{\beta} \pm 1.96\sqrt{d'\hat{\Sigma}_{\hat{\beta}}d}. \tag{19.19}$$

Asymptotic Distribution of the t-Statistic

The t-statistic testing the null hypothesis that $\beta_j = \beta_{j,0}$, constructed using the heteroskedasticity-robust standard error in Equation (19.18), is given in Key Concept 7.1. The argument that this t-statistic has an asymptotic standard normal distribution parallels the argument given in Section 18.3 for the single-regressor model.

19.3 Tests of Joint Hypotheses

Section 7.2 considers tests of joint hypotheses that involve multiple restrictions, where each restriction involves a single coefficient, and Section 7.3 considers tests of a single restriction involving two or more coefficients. The matrix setup of Section 19.1 permits a unified representation of these two types of hypotheses as linear restrictions on the coefficient vector, where each restriction can involve multiple coefficients. Under the first four least squares assumptions in Key Concept 19.1, the heteroskedasticity-robust OLS F-statistic testing these hypotheses has an $F_{q,\infty}$ asymptotic distribution under the null hypothesis.

Joint Hypotheses in Matrix Notation

Consider a joint hypothesis that is linear in the coefficients and imposes q restrictions, where $q \leq k + 1$. Each of these q restrictions can involve one or more of the regression coefficients. This joint null hypothesis can be written in matrix notation as

$$\boldsymbol{R\beta} = \boldsymbol{r}, \tag{19.20}$$

where $\boldsymbol{R}$ is a $q \times (k + 1)$ nonrandom matrix with full row rank and $\boldsymbol{r}$ is a nonrandom $q \times 1$ vector. The number of rows of $\boldsymbol{R}$ is q, which is the number of restrictions being imposed under the null hypothesis.

The null hypothesis in Equation (19.20) subsumes all the null hypotheses considered in Sections 7.2 and 7.3. For example, a joint hypothesis of the type considered in Section 7.2 is that $\beta_0 = 0, \beta_1 = 0, \ldots, \beta_{q-1} = 0$. To write this joint hypothesis in the form of Equation (19.20), set $\boldsymbol{R} = \begin{bmatrix} \boldsymbol{I}_q & \boldsymbol{0}_{q \times (k+1-q)} \end{bmatrix}$ and $\boldsymbol{r} = \boldsymbol{0}_q$.

The formulation in Equation (19.20) also captures the restrictions of Section 7.3 involving multiple regression coefficients. For example, if $k = 2$, then the hypothesis that $\beta_1 + \beta_2 = 1$ can be written in the form of Equation (19.20) by setting $\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$, $r = 1$, and $q = 1$.

Asymptotic Distribution of the *F*-Statistic

The heteroskedasticity-robust F-statistic testing the joint hypothesis in Equation (19.20) is

$$F = (\boldsymbol{R\hat{\beta}} - \boldsymbol{r})'\big[\boldsymbol{R\hat{\Sigma}_{\hat{\beta}}R'}\big]^{-1}(\boldsymbol{R\hat{\beta}} - \boldsymbol{r})/q. \tag{19.21}$$

If the first four assumptions in Key Concept 19.1 hold, then under the null hypothesis

$$F \xrightarrow{d} F_{q,\infty}. \tag{19.22}$$

This result follows by combining the asymptotic normality of $\boldsymbol{\hat{\beta}}$ with the consistency of the heteroskedasticity-robust estimator $\boldsymbol{\hat{\Sigma}}_{\sqrt{n}(\hat{\beta}-\beta)}$ of the covariance matrix. Specifically, first note that Equation (19.12) and Equation (19.74) in

Appendix 19.2 imply that, under the null hypothesis, $\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) = \sqrt{n}\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{R}\boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}\boldsymbol{R}')$. It follows from Equation (19.77) that, under the null hypothesis, $(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) = [\sqrt{n}\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]'[\boldsymbol{R}\boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}\boldsymbol{R}']^{-1}[\sqrt{n}\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \xrightarrow{d} \chi_q^2$. However, because $\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})} \xrightarrow{p} \boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}$, it follows from Slutsky's theorem that $[\sqrt{n}\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]'[\boldsymbol{R}\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}\boldsymbol{R}']^{-1}[\sqrt{n}\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \xrightarrow{d} \chi_q^2$. or, equivalently (because $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}/n$), that $F \xrightarrow{d} \chi_q^2/q$, which is in turn distributed $F_{q, \infty}$.

Confidence Sets for Multiple Coefficients

As discussed in Section 7.4, an asymptotically valid confidence set for two or more elements of $\boldsymbol{\beta}$ can be constructed as the set of values that, when taken as the null hypothesis, are not rejected by the F-statistic. In principle, this set could be computed by repeatedly evaluating the F-statistic for many values of $\boldsymbol{\beta}$, but, as is the case with a confidence interval for a single coefficient, it is simpler to manipulate the formula for the test statistic to obtain an explicit formula for the confidence set.

Here is the procedure for constructing a confidence set for two or more of the elements of $\boldsymbol{\beta}$. Let $\boldsymbol{\delta}$ denote the q-dimensional vector consisting of the coefficients for which we wish to construct a confidence set. For example, if we are constructing a confidence set for the regression coefficients β_1 and β_2, then $q = 2$ and $\boldsymbol{\delta} = (\beta_1 \, \beta_2)'$. In general, we can write $\boldsymbol{\delta} = \boldsymbol{R}\boldsymbol{\beta}$, where the matrix $\boldsymbol{R}$ consists of 0's and 1's [as discussed following Equation (19.20)]. The F-statistic testing the hypothesis that $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ is $F = (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)'[\boldsymbol{R}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{R}']^{-1}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)/q$, where $\hat{\boldsymbol{\delta}} = \boldsymbol{R}\hat{\boldsymbol{\beta}}$. A 95% confidence set for $\boldsymbol{\delta}$ is the set of values $\boldsymbol{\delta}_0$ that are not rejected by the F-statistic. That is, when $\boldsymbol{\delta} = \boldsymbol{R}\boldsymbol{\beta}$, a 95% confidence set for $\boldsymbol{\delta}$ is

$$\{\boldsymbol{\delta} : (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})'[\boldsymbol{R}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{R}']^{-1}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})/q \le c\}, \tag{19.23}$$

where c is the 95th percentile (the 5% critical value) of the $F_{q, \infty}$ distribution.

The set in Equation (19.23) consists of all the points contained inside the ellipse determined when the inequality in Equation (19.23) is an equality (this is an ellipsoid when $q > 2$). Thus the confidence set for δ can be computed by solving Equation (19.23) for the boundary ellipse.

19.4 Distribution of Regression Statistics with Normal Errors

The distributions presented in Sections 19.2 and 19.3, which were justified by appealing to the law of large numbers and the central limit theorem, apply when the sample size is large. If, however, the errors are homoskedastic and normally distributed, conditional on $\boldsymbol{X}$, then the OLS estimator has a multivariate normal distribution in a finite sample, conditional on $\boldsymbol{X}$. In addition, the finite sample distribution of the

square of the standard error of the regression is proportional to the chi-squared distribution with $n - k - 1$ degrees of freedom, the homoskedasticity-only OLS t-statistic has a Student t distribution with $n - k - 1$ degrees of freedom, and the homoskedasticity-only F-statistic has an $F_{q,n-k-1}$ distribution. The arguments in this section employ some specialized matrix formulas for OLS regression statistics, which are presented first.

Matrix Representations of OLS Regression Statistics

The OLS predicted values, residuals, and sum of squared residuals have compact matrix representations. These representations make use of two matrices, $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$.

The matrices $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$. The algebra of OLS in the multivariate model relies on the two symmetric $n \times n$ matrices, $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$:

$$\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \text{ and} \tag{19.24}$$

$$\boldsymbol{M}_X = \boldsymbol{I}_n - \boldsymbol{P}_X. \tag{19.25}$$

A matrix $\boldsymbol{C}$ is idempotent if $\boldsymbol{C}$ is square and $\boldsymbol{C}\boldsymbol{C} = \boldsymbol{C}$ (see Appendix 19.1). Because $\boldsymbol{P}_X = \boldsymbol{P}_X\boldsymbol{P}_X$ and $\boldsymbol{M}_X = \boldsymbol{M}_X\boldsymbol{M}_X$ (Exercise 19.5) and because $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$ are symmetric, $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$ are symmetric idempotent matrices.

The matrices $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$ have some additional useful properties (Exercise 19.5), which follow directly from the definitions in Equations (19.24) and (19.25):

$$\boldsymbol{P}_X\boldsymbol{X} = \boldsymbol{X} \text{ and } \boldsymbol{M}_X\boldsymbol{X} = \boldsymbol{0}_{n \times (k+1)};$$

$$\text{rank}(\boldsymbol{P}_X) = k + 1 \text{ and rank}(\boldsymbol{M}_X) = n - k - 1, \tag{19.26}$$

where rank$(\boldsymbol{P}_X)$ is the rank of $\boldsymbol{P}_X$.

The matrices $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$ can be used to decompose an n-dimensional vector $\boldsymbol{Z}$ into two parts: a part that is spanned by the columns of $\boldsymbol{X}$ and a part that is orthogonal to the columns of $\boldsymbol{X}$. In other words, $\boldsymbol{P}_X\boldsymbol{Z}$ is the projection of $\boldsymbol{Z}$ onto the space spanned by the columns of $\boldsymbol{X}$, $\boldsymbol{M}_X\boldsymbol{Z}$ is the part of $\boldsymbol{Z}$ orthogonal to the columns of $\boldsymbol{X}$, and $\boldsymbol{Z} = \boldsymbol{P}_X\boldsymbol{Z} + \boldsymbol{M}_X\boldsymbol{Z}$.

OLS predicted values and residuals. The matrices $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$ provide some simple expressions for OLS predicted values and residuals. The OLS predicted values, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and the OLS residuals, $\hat{\boldsymbol{U}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, can be expressed as follows (Exercise 19.5):

$$\hat{\boldsymbol{Y}} = \boldsymbol{P}_X\boldsymbol{Y} \text{ and} \tag{19.27}$$

$$\hat{\boldsymbol{U}} = \boldsymbol{M}_X\boldsymbol{Y} = \boldsymbol{M}_X\boldsymbol{U}. \tag{19.28}$$

The expressions in Equations (19.27) and (19.28) provide a simple proof that the OLS residuals and predicted values are orthogonal—that is, that Equation (4.35) holds: $\hat{\boldsymbol{Y}}'\hat{\boldsymbol{U}} = \boldsymbol{Y}'\boldsymbol{P}_X'\boldsymbol{M}_X\boldsymbol{Y} = 0$, where the second equality follows from $\boldsymbol{P}_X'\boldsymbol{M}_X = \boldsymbol{0}_{n \times n}$, which in turn follows from $\boldsymbol{M}_X\boldsymbol{X} = \boldsymbol{0}_{n \times (k+1)}$ in Equation (19.26).

The standard error of the regression. The *SER*, defined in Section 4.3, is $s_{\hat{u}}$, where

$$s_{\hat{u}}^2 = \frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2 = \frac{1}{n-k-1} \hat{U}'\hat{U} = \frac{1}{n-k-1} U'M_X U, \quad (19.29)$$

where the final equality follows because $\hat{U}'\hat{U} = (M_X U)'(M_X U) = U'M_X M_X U = U'M_X U$ (because M_X is symmetric and idempotent).

Distribution of $\hat{\beta}$ with Independent Normal Errors

Because $\hat{\beta} = \beta + (X'X)^{-1}X'U$ [Equation (19.14)] and because the distribution of U, conditional on X, is, by assumption, $N(0_n, \sigma_u^2 I_n)$ [Equation (19.8)], the conditional distribution of $\hat{\beta}$ given X is multivariate normal with mean β. The covariance matrix of $\hat{\beta}$, conditional on X, is $\Sigma_{\hat{\beta}|X} = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X] = E[(X'X)^{-1} X'UU'X(X'X)^{-1}|X] = (X'X)^{-1}X'(\sigma_u^2 I_n)X(X'X)^{-1} = \sigma_u^2(X'X)^{-1}$. Accordingly, under all six assumptions in Key Concept 19.1, the finite-sample conditional distribution of $\hat{\beta}$ given X is

$$\hat{\beta} \sim N(\beta, \Sigma_{\hat{\beta}|X}), \text{ where } \Sigma_{\hat{\beta}|X} = \sigma_u^2(X'X)^{-1}. \quad (19.30)$$

Distribution of $s_{\hat{u}}^2$

If all six assumptions in Key Concept 19.1 hold, then $s_{\hat{u}}^2$ has an exact sampling distribution that is proportional to a chi-squared distribution with $n - k - 1$ degrees of freedom:

$$s_{\hat{u}}^2 \sim \frac{\sigma_u^2}{n-k-1} \times \chi_{n-k-1}^2 \quad (19.31)$$

The proof of Equation (19.31) starts with Equation (19.29). Because U is normally distributed, conditional on X, and because M_X is a symmetric idempotent matrix, the quadratic form $U'M_X U/\sigma_u^2$ has an exact chi-squared distribution with degrees of freedom equal to the rank of M_X [Equation (19.78) in Appendix 19.2]. From Equation (19.26), the rank of M_X is $n - k - 1$. Thus $U'M_X U/\sigma_u^2$ has an exact χ_{n-k-1}^2 distribution, from which Equation (19.31) follows.

The degrees-of-freedom adjustment ensures that $s_{\hat{u}}^2$ is unbiased. The expectation of a random variable with a χ_{n-k-1}^2 distribution is $n - k - 1$; thus $E(U'M_X U) = (n-k-1)\sigma_u^2$, so $E(s_{\hat{u}}^2) = \sigma_u^2$.

Homoskedasticity-Only Standard Errors

The homoskedasticity-only estimator $\widetilde{\Sigma}_{\hat{\beta}}$ of the covariance matrix of $\hat{\beta}$, conditional on X, is obtained by substituting the sample variance $s_{\hat{u}}^2$ for the population variance σ_u^2 in the expression for $\Sigma_{\hat{\beta}|X}$ in Equation (19.30). Accordingly,

$$\widetilde{\Sigma}_{\hat{\beta}} = s_{\hat{u}}^2(X'X)^{-1} \quad \text{(homoskedasticity-only)}. \quad (19.32)$$

The estimator of the variance of the normal conditional distribution of $\hat{\beta}_j$ given X is the (j, j) element of $\widetilde{\Sigma}_{\hat{\beta}}$. Thus the homoskedasticity-only standard error of $\hat{\beta}_j$ is the square root of the jth diagonal element of $\widetilde{\Sigma}_{\hat{\beta}}$. That is, the homoskedasticity-only standard error of $\hat{\beta}_j$ is

$$S\widetilde{E}(\hat{\beta}_j) = \sqrt{(\widetilde{\Sigma}_{\hat{\beta}})_{jj}} \quad \text{(homoskedasticity-only).} \tag{19.33}$$

Distribution of the *t*-Statistic

Let $\widetilde{t}$ be the *t*-statistic testing the hypothesis $\beta_j = \beta_{j,0}$, constructed using the homoskedasticity-only standard error; that is, let

$$\widetilde{t} = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{(\widetilde{\Sigma}_{\hat{\beta}})_{jj}}}. \tag{19.34}$$

Under all six of the extended least squares assumptions in Key Concept 19.1, the exact sampling distribution of $\widetilde{t}$ is the Student *t* distribution with $n - k - 1$ degrees of freedom; that is,

$$\widetilde{t} \sim t_{n-k-1}. \tag{19.35}$$

The proof of Equation (19.35) is given in Appendix 19.4.

Distribution of the *F*-Statistic

If all six least squares assumptions in Key Concept 19.1 hold, then the *F*-statistic testing the hypothesis in Equation (19.20), constructed using the homoskedasticity-only estimator of the covariance matrix, has an exact $F_{q, n-k-1}$ distribution under the null hypothesis.

The homoskedasticity-only F-statistic.   The homoskedasticity-only *F*-statistic is similar to the heteroskedasticity-robust *F*-statistic in Equation (19.21) except that the homoskedasticity-only estimator $\widetilde{\Sigma}_{\hat{\beta}}$ is used instead of the heteroskedasticity-robust estimator $\hat{\Sigma}_{\hat{\beta}}$. Substituting the expression $\widetilde{\Sigma}_{\hat{\beta}} = s_{\hat{u}}^2 (X'X)^{-1}$ into the expression for the *F*-statistic in Equation (19.21) yields the homoskedasticity-only *F*-statistic testing the null hypothesis in Equation (19.20):

$$\widetilde{F} = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{s_{\hat{u}}^2}. \tag{19.36}$$

If all six assumptions in Key Concept 19.1 hold, then under the null hypothesis

$$\widetilde{F} \sim F_{q,n-k-1}. \tag{19.37}$$

The proof of Equation (19.37) is given in Appendix 19.4.

The F-statistic in Equation (19.36) is called the Wald version of the F-statistic (named after the statistician Abraham Wald). Although the formula for the homoskedastic-only F-statistic given in Equation (7.13) appears quite different from the formula for the Wald statistic in Equation (19.36), the homoskedastic-only F-statistic and the Wald F-statistic are two versions of the same statistic. That is, the two expressions are equivalent, a result shown in Exercise 19.13.

19.5 Efficiency of the OLS Estimator with Homoskedastic Errors

Under the Gauss–Markov conditions for multiple regression, the OLS estimator of $\boldsymbol{\beta}$ is efficient among all linear conditionally unbiased estimators; that is, the OLS estimator is the best linear unbiased estimator (BLUE).

The Gauss–Markov Conditions for Multiple Regression

The **Gauss–Markov conditions for multiple regression** are

$$\text{(i) } E(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{0}_n,$$
$$\text{(ii) } E(\boldsymbol{U}\boldsymbol{U}'|\boldsymbol{X}) = \sigma_u^2 \boldsymbol{I}_n, \text{ and}$$
$$\text{(iii) } \boldsymbol{X} \text{ has full column rank.} \tag{19.38}$$

The Gauss–Markov conditions for multiple regression in turn are implied by the first five assumptions in Key Concept 19.1 [see Equations (19.6) and (19.7)]. The conditions in Equation (19.38) generalize the Gauss–Markov conditions for a single-regressor model to multiple regression. [By using matrix notation, the second and third Gauss–Markov conditions in Equation (5.31) are collected into the single condition (ii) in Equation (19.38).]

Linear Conditionally Unbiased Estimators

We start by describing the class of linear unbiased estimators and by showing that OLS is in that class.

The class of linear conditionally unbiased estimators. An estimator of $\boldsymbol{\beta}$ is said to be linear if it is a linear function of $Y_1, \ldots, Y_n$. Accordingly, the estimator $\widetilde{\boldsymbol{\beta}}$ is linear in $\boldsymbol{Y}$ if it can be written in the form

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{Y}, \tag{19.39}$$

where $\boldsymbol{A}$ is an $n \times (k + 1)$ dimensional matrix of weights that may depend on $\boldsymbol{X}$ and on nonrandom constants but not on $\boldsymbol{Y}$.

Gauss–Markov Theorem for Multiple Regression

Suppose that the Gauss–Markov conditions for multiple regression in Equation (19.38) hold. Then the OLS estimator $\hat{\boldsymbol{\beta}}$ is BLUE. That is, let $\tilde{\boldsymbol{\beta}}$ be a linear conditionally unbiased estimator of $\boldsymbol{\beta}$, and let $\boldsymbol{c}$ be a nonrandom $k + 1$ dimensional vector. Then $\text{var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}|\boldsymbol{X}) \leq \text{var}(\boldsymbol{c}'\tilde{\boldsymbol{\beta}}|\boldsymbol{X})$ for every nonzero vector $\boldsymbol{c}$, where the inequality holds with equality for all $\boldsymbol{c}$ only if $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.

An estimator is conditionally unbiased if the mean of its conditional sampling distribution given $\boldsymbol{X}$ is $\boldsymbol{\beta}$. That is, $\tilde{\boldsymbol{\beta}}$ is conditionally unbiased if $E(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta}$.

The OLS estimator is linear and conditionally unbiased. Comparison of Equations (19.11) and (19.39) shows that the OLS estimator is linear in $\boldsymbol{Y}$; specifically, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{A}}'\boldsymbol{Y}$, where $\hat{\boldsymbol{A}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. To show that $\hat{\boldsymbol{\beta}}$ is conditionally unbiased, recall from Equation (19.14) that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$. Taking the conditional expectation of both sides of this expression yields $E(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta} + E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}|\boldsymbol{X}] = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{\beta}$, where the final equality follows because $E(\boldsymbol{U}|\boldsymbol{X}) = 0$ by the first Gauss–Markov condition.

The Gauss–Markov Theorem for Multiple Regression

The **Gauss–Markov theorem for multiple regression** provides conditions under which the OLS estimator is efficient among the class of linear conditionally unbiased estimators. A subtle point arises, however, because $\hat{\boldsymbol{\beta}}$ is a vector and its "variance" is a covariance matrix. When the variance of an estimator is a matrix, just what does it mean to say that one estimator has a smaller variance than another?

The Gauss–Markov theorem handles this problem by comparing the variance of a candidate estimator of a *linear combination* of the elements of $\boldsymbol{\beta}$ to the variance of the corresponding linear combination of $\hat{\boldsymbol{\beta}}$. Specifically, let $\boldsymbol{c}$ be a $k + 1$ dimensional vector, and consider the problem of estimating the linear combination $\boldsymbol{c}'\boldsymbol{\beta}$ using the candidate estimator $\boldsymbol{c}'\tilde{\boldsymbol{\beta}}$ (where $\tilde{\boldsymbol{\beta}}$ is a linear conditionally unbiased estimator) on the one hand and $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ on the other hand. Because $\boldsymbol{c}'\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ are both scalars and are both linear conditionally unbiased estimators of $\boldsymbol{c}'\boldsymbol{\beta}$, it now makes sense to compare their variances.

The Gauss–Markov theorem for multiple regression says that the OLS estimator of $\boldsymbol{c}'\boldsymbol{\beta}$ is efficient; that is, the OLS estimator $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ has the smallest conditional variance of all linear conditionally unbiased estimators. Remarkably, this is true no matter what the linear combination is. It is in this sense that the OLS estimator is BLUE in multiple regression.

The Gauss–Markov theorem is stated in Key Concept 19.3 and proven in Appendix 19.5.

19.6   Generalized Least Squares[1]

The assumption of i.i.d. sampling fits many applications. For example, suppose that Y_i and X_i correspond to information about individuals, such as their earnings, education, and personal characteristics, where the individuals are selected from a population by simple random sampling. In this case, because of the simple random sampling scheme, (X_i, Y_i) are necessarily i.i.d. Because (X_i, Y_i) and (X_j, Y_j) are independently distributed for $i \neq j$, u_i and u_j are independently distributed for $i \neq j$. This in turn implies that u_i and u_j are uncorrelated for $i \neq j$. In the context of the Gauss–Markov assumptions, the assumption that $E(UU' | X)$ is diagonal therefore is appropriate if the data are collected in a way that makes the observations independently distributed.

Some sampling schemes encountered in econometrics do not, however, result in independent observations and instead can lead to error terms u_i that are correlated from one observation to the next. The leading example is when the data are sampled over time for the same entity—that is, when the data are time series data. As discussed in Section 16.3, in regressions involving time series data, many omitted factors are correlated from one period to the next, and this can result in regression error terms (which represent those omitted factors) that are correlated from one period of observation to the next. In other words, the error term in one period will not, in general, be distributed independently of the error term in the next period. Instead, the error term in one period could be correlated with the error term in the next period.

The presence of correlated error terms creates two problems for inference based on OLS. First, *neither* the heteroskedasticity-robust nor the homoskedasticity-only standard errors produced by OLS provide a valid basis for inference. The solution to this problem is to use standard errors that are robust to both heteroskedasticity and correlation of the error terms across observations. This topic—heteroskedasticity- and autocorrelation-consistent (HAC) covariance matrix estimation—is the subject of Section 16.4 and we do not pursue it further here.

Second, if the error term is correlated across observations, then $E(UU' | X)$ is not diagonal, the second Gauss–Markov condition in Equation (19.38) does not hold, and OLS is not BLUE. In this section, we study an estimator, **generalized least squares (GLS)**, that is BLUE (at least asymptotically) when the conditional covariance matrix of the errors is no longer proportional to the identity matrix. A special case of GLS is weighted least squares, discussed in Section 18.5, in which the conditional covariance matrix is diagonal and the ith diagonal element is a function of X_i. Like WLS, GLS transforms the regression model so that the errors of the transformed model satisfy the Gauss–Markov conditions. The GLS estimator is the OLS estimator of the coefficients in the transformed model.

---

[1]The GLS estimator was introduced in Section 16.5 in the context of distributed lag time series regression. The presentation here is a self-contained mathematical treatment of GLS that can be read independently of Section 16.5, but reading that section first will help to make these ideas more concrete.

The GLS Assumptions

In the linear regression model $Y = X\beta + U$, the GLS assumptions are

1. $E(U|X) = \mathbf{0}_n$;
2. $E(UU'|X) = \Omega(X)$, where $\Omega(X)$ is an $n \times n$ positive definite matrix that can depend on X;
3. X_i and u_i satisfy suitable moment conditions; and
4. X has full column rank (there is no perfect multicollinearity).

The GLS Assumptions

There are four assumptions under which GLS is valid. The first GLS assumption is that u_i has a mean of 0, conditional on $X_1, \ldots, X_n$; that is,

$$E(U|X) = \mathbf{0}_n. \tag{19.40}$$

This assumption is implied by the first two least squares assumptions in Key Concept 19.1; that is, if $E(u_i|X_i) = 0$ and $(X_i, Y_i), i = 1, \ldots, n,$ are i.i.d., then $E(U|X) = \mathbf{0}_n$. In GLS, however, we will not want to maintain the i.i.d. assumption; after all, one purpose of GLS is to handle errors that are correlated across observations. We discuss the significance of the assumption in Equation (19.40) after introducing the GLS estimator.

The second GLS assumption is that the conditional covariance matrix of U given X is some function of X:

$$E(UU'|X) = \Omega(X), \tag{19.41}$$

where $\Omega(X)$ is an $n \times n$ positive definite matrix-valued function of X.

There are two main applications of GLS that are covered by this assumption. The first is independent sampling with heteroskedastic errors, in which case $\Omega(X)$ is a diagonal matrix with diagonal element $\lambda h(X_i)$, where λ is a constant and h is a function. In this case, discussed in Section 18.5, GLS is WLS.

The second application is to homoskedastic errors that are serially correlated. In practice, in this case a model is developed for the serial correlation. For example, one model is that the error term is correlated with only its neighbor, so $\text{corr}(u_i, u_{i-1}) = \rho \neq 0$ but $\text{corr}(u_i, u_j) = 0$ if $|i - j| \geq 2$. In this case, $\Omega(X)$ has σ_u^2 as its diagonal element, $\rho\sigma_u^2$ in the first off-diagonal, and zeros elsewhere. Thus $\Omega(X)$ does not depend on X, $\Omega_{ii} = \sigma_u^2$, $\Omega_{ij} = \rho\sigma_u^2$ for $|i - j| = 1$, and $\Omega_{ij} = 0$ for $|i - j| > 1$. Other models for serial correlation, including the first-order autoregressive model, are discussed further in the context of GLS in Section 16.5 (also see Exercise 19.8).

One assumption that has appeared on all previous lists of least squares assumptions for cross-sectional data is that X_i and u_i have nonzero finite fourth moments. In the case of GLS, the specific moment assumptions needed to prove asymptotic results depend on the nature of the function $\Omega(X)$, whether $\Omega(X)$ is known or estimated, and the statistic under consideration (the GLS estimator, t-statistic, etc.). Because the assumptions are case- and model-specific, we do not present specific moment assumptions here, and the discussion of the large-sample properties of GLS assumes that such moment conditions apply for the relevant case at hand. For completeness, as the third GLS assumption, X_i and u_i are simply assumed to satisfy suitable moment conditions.

The fourth GLS assumption is that X has full column rank; that is, the regressors are not perfectly multicollinear.

The GLS assumptions are summarized in Key Concept 19.4.

We consider GLS estimation in two cases. In the first case, $\Omega(X)$ is known. In the second case, the functional form of $\Omega(X)$ is known up to some parameters that can be estimated. To simplify notation, we refer to the function $\Omega(X)$ as the matrix Ω, so the dependence of Ω on X is implicit.

GLS When Ω Is Known

When Ω is known, the GLS estimator uses Ω to transform the regression model to one with errors that satisfy the Gauss–Markov conditions. Specifically, let F be a matrix square root of Ω^{-1}; that is, let F be a matrix that satisfies $F'F = \Omega^{-1}$ (see Appendix 19.1). A property of F is that $F\Omega F' = I_n$. Now premultiply both sides of Equation (19.3) by F to obtain

$$\tilde{Y} = \tilde{X}\beta + \tilde{U}, \tag{19.42}$$

where $\tilde{Y} = FY$, $\tilde{X} = FX$, and $\tilde{U} = FU$.

The key insight of GLS is that, under the four GLS assumptions, the Gauss–Markov assumptions hold for the transformed regression in Equation (19.42). That is, by transforming all the variables by the matrix square root of the inverse of Ω, the regression errors in the transformed regression have a conditional mean of 0 and a covariance matrix that equals the identity matrix. To show this mathematically, first note that $E(\tilde{U}|\tilde{X}) = E(FU|FX) = FE(U|FX) = \mathbf{0}_n$ by the first GLS assumption [Equation (19.40)]. In addition, $E(\tilde{U}\tilde{U}'|\tilde{X}) = E[(FU)(FU)'|FX] = FE(UU'|FX)F' = F\Omega F' = I_n$, where the second equality follows because $(FU)' = U'F'$ and the final equality follows from the definition of F. It follows that the transformed regression model in Equation (19.42) satisfies the Gauss–Markov conditions in Key Concept 19.3.

The GLS estimator, $\tilde{\beta}^{GLS}$, is the OLS estimator of β in Equation (19.42); that is, $\tilde{\beta}^{GLS} = (\tilde{X}'\tilde{X})^{-1}(\tilde{X}'\tilde{Y})$. Because the transformed regression model satisfies the Gauss–Markov conditions, the GLS estimator is the best conditionally unbiased

estimator that is linear in $\widetilde{Y}$. But because $\widetilde{Y} = FY$ and F is (here) assumed to be known and because F is invertible (because Ω is positive definite), the class of estimators that are linear in $\widetilde{Y}$ is the same as the class of estimators that are linear in Y. Thus the OLS estimator of β in Equation (19.42) is also the best conditionally unbiased estimator among estimators that are linear in Y. In other words, under the GLS assumptions, the GLS estimator is BLUE.

The GLS estimator can be expressed directly in terms of Ω, so in principle there is no need to compute the square root matrix F. Because $\widetilde{X} = FX$ and $\widetilde{Y} = FY$, $\widetilde{\beta}^{GLS} = (X'F'FX)^{-1}(X'F'FY)$. But $F'F = \Omega^{-1}$, so

$$\widetilde{\beta}^{GLS} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y). \tag{19.43}$$

In practice, Ω is typically unknown, so the GLS estimator in Equation (19.43) typically cannot be computed and thus is sometimes called the **infeasible GLS** estimator. If, however, Ω has a known functional form but the parameters of that function are unknown, then Ω can be estimated, and a feasible version of the GLS estimator can be computed.

GLS When Ω Contains Unknown Parameters

If Ω is a known function of some parameters that in turn can be estimated, then these estimated parameters can be used to calculate an estimator of the covariance matrix Ω. For example, consider the time series application discussed following Equation (19.41), in which $\Omega(X)$ does not depend on X, $\Omega_{ii} = \sigma_u^2$, $\Omega_{ij} = \rho\sigma_u^2$ for $|i - j| = 1$, and $\Omega_{ij} = 0$ for $|i - j| > 1$. Then Ω has two unknown parameters, σ_u^2 and ρ. These parameters can be estimated using the residuals from a preliminary OLS regression; specifically, σ_u^2 can be estimated by $s_{\hat{u}}^2$, and ρ can be estimated by the sample correlation between all neighboring pairs of OLS residuals. These estimated parameters can in turn be used to compute an estimator of Ω, $\hat{\Omega}$.

In general, suppose that you have an estimator $\hat{\Omega}$ of Ω. Then the GLS estimator based on $\hat{\Omega}$ is

$$\hat{\beta}^{GLS} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}Y). \tag{19.44}$$

The GLS estimator in Equation (19.44) is sometimes called the **feasible GLS** estimator because it can be computed if the covariance matrix contains some unknown parameters that can be estimated.

The Conditional Mean Zero Assumption and GLS

For the OLS estimator to be consistent, the first least squares assumption must hold; that is, $E(u_i|X_i)$ must be 0. In contrast, the first GLS assumption is that $E(u_i|X_1, \ldots, X_n) = 0$. In other words, the first OLS assumption is that the error for the ith observation has a conditional mean of 0 given the values of the regressors for

that observation, whereas the first GLS assumption is that u_i has a conditional mean of 0 given the values of the regressors for *all* observations.

As discussed in Section 19.1, the assumptions that $E(u_i|X_i) = 0$ and that sampling is i.i.d. together imply that $E(u_i|X_1, \ldots, X_n) = 0$. Thus, when sampling is i.i.d., so that GLS is WLS, the first GLS assumption is implied by the first least squares assumption in Key Concept 19.1.

When sampling is not i.i.d., however, the first GLS assumption is not implied by the assumption that $E(u_i|X_i) = 0$; that is, the first GLS assumption is stronger. Although the distinction between these two conditions might seem slight, it can be very important in applications to time series data. This distinction is discussed in Section 16.5 in the context of whether the regressor is "past and present" exogenous or "strictly" exogenous; the assumption that $E(u_i|X_1, \ldots, X_n) = 0$ corresponds to strict exogeneity. Here, we discuss this distinction at a more general level using matrix notation. To do so, we focus on the case that U is homoskedastic, Ω is known, and Ω has nonzero off-diagonal elements.

The role of the first GLS assumption. To see the source of the difference between these assumptions, it is useful to contrast the consistency arguments for GLS and OLS.

We first sketch the argument for the consistency of the GLS estimator in Equation (19.43). Substituting Equation (19.3) into Equation (19.43), we have $\widetilde{\boldsymbol{\beta}}^{GLS} = \boldsymbol{\beta} + (X'\Omega^{-1}X/n)^{-1}(X'\Omega^{-1}U/n)$. Under the first GLS assumption, $E(X'\Omega^{-1}U) = E[X'\Omega^{-1}E(U|X)] = \mathbf{0}_n$. If in addition the variance of $X'\Omega^{-1}U/n$ tends to 0 and $X'\Omega^{-1}X/n \xrightarrow{p} \widetilde{Q}$, where $\widetilde{Q}$ is some invertible matrix, then $\widetilde{\boldsymbol{\beta}}^{GLS} \xrightarrow{p} \boldsymbol{\beta}$. Critically, when Ω has off-diagonal elements, the term $X'\Omega^{-1}U = \sum_{i=1}^{n} \sum_{j=1}^{n} X_i(\Omega^{-1})_{ij}u_j$ involves products of X_i and u_j for different i, j pairs, where $(\Omega^{-1})_{ij}$ denotes the (i, j) element of Ω^{-1}. Thus, for $X'\Omega^{-1}U$ to have a mean of 0, it is not enough that $E(u_i|X_i) = 0$; rather, $E(u_i|X_j)$ must equal 0 for all i, j pairs corresponding to nonzero values of $(\Omega^{-1})_{ij}$. Depending on the covariance structure of the errors, only some of or all the elements of $(\Omega^{-1})_{ij}$ might be nonzero. For example, if u_i follows a first-order autoregression (as discussed in Section 16.5), the only nonzero elements $(\Omega^{-1})_{ij}$ are those for which $|i - j| \leq 1$. In general, however, all the elements of Ω^{-1} can be nonzero, so, in general, for $X'\Omega^{-1}U/n \xrightarrow{p} \mathbf{0}_{(k+1)\times 1}$ (and thus for $\widetilde{\boldsymbol{\beta}}^{GLS}$ to be consistent), we need that $E(U|X) = \mathbf{0}_n$; that is, the first GLS assumption must hold.

In contrast, recall the argument that the OLS estimator is consistent. Rewrite Equation (19.14) as $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (X'X/n)^{-1}\frac{1}{n}\sum_{i=1}^{n}X_iu_i$. If $E(u_i|X_i) = 0$, then the term $\frac{1}{n}\sum_{i=1}^{n}X_iu_i$ has mean 0, and if this term has a variance that tends to 0, it converges in probability to 0. If in addition $X'X/n \xrightarrow{p} Q_X$, then $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

Is the first GLS assumption restrictive? The first GLS assumption requires that the errors for the ith observation be uncorrelated with the regressors for all other observations. This assumption is dubious in some time series applications. This issue is discussed in Section 16.6 in the context of an empirical example, the relationship

between the change in the price of a contract for future delivery of frozen orange concentrate and the weather in Florida. As explained there, the error term in the regression of price changes on the weather is plausibly uncorrelated with current and past values of the weather, so the first OLS assumption holds. However, this error term is plausibly correlated with future values of the weather, so the first GLS assumption does *not* hold.

This example illustrates a general phenomenon in economic time series data that arises when the value of a variable today is set in part based on expectations of the future: Those future expectations typically imply that the error term today depends on a forecast of the regressor tomorrow, which in turn is correlated with the actual value of the regressor tomorrow. For this reason, the first GLS assumption is, in fact, much stronger than the first OLS assumption. Accordingly, in some applications with economic time series data, the GLS estimator is not consistent even though the OLS estimator is.

19.7 Instrumental Variables and Generalized Method of Moments Estimation

This section provides an introduction to the theory of instrumental variables (IV) estimation and the asymptotic distribution of IV estimators. It is assumed throughout that the IV regression assumptions in Key Concepts 12.3 and 12.4 hold and, moreover, that the instruments are strong. These assumptions apply to cross-sectional data with i.i.d. observations. Under certain conditions, the results derived in this section are applicable to time series data as well, and the extension to time series data is briefly discussed at the end of this section. All asymptotic results in this section are developed under the assumption of strong instruments.

This section begins by presenting the IV regression model and the two stage least squares (TSLS) estimator and its asymptotic distribution in the general case of heteroskedasticity, all in matrix form. It is next shown that, in the special case of homoskedasticity, the TSLS estimator is asymptotically efficient among the class of IV estimators in which the instruments are linear combinations of the exogenous variables. Moreover, the *J*-statistic has an asymptotic chi-squared distribution in which the degrees of freedom equals the number of overidentifying restrictions. This section concludes with a discussion of efficient IV estimation and the test of overidentifying restrictions when the errors are heteroskedastic—a situation in which the efficient IV estimator is known as the efficient generalized method of moments (GMM) estimator [Hansen (1983)].

The IV Estimator in Matrix Form

In this section, we let X denote the $n \times (k + r + 1)$ matrix of the regressors in the equation of interest, so X contains the included endogenous regressors (the X's in Key Concept 12.1) *and* the included exogenous regressors (the W's in Key Concept 12.1).

That is, in the notation of Key Concept 12.1, the ith row of X is $X_i' = (1 \quad X_{1i} \quad X_{2i} \quad \ldots$ $X_{ki} \quad W_{1i} \quad W_{2i} \quad \ldots \quad W_{ri})$. Also, let Z denote the $n \times (m + r + 1)$ matrix of all the exogenous regressors, both those included in the equation of interest (the W's) *and* those excluded from the equation of interest (the instruments). That is, in the notation of Key Concept 12.1, the ith row of Z is $Z_i' = (1 \quad Z_{1i} \quad Z_{2i} \quad \ldots \quad Z_{mi}$ $W_{1i} \quad W_{2i} \quad \ldots \quad W_{ri})$.

With this notation, the IV regression model of Key Concept 12.1, written in matrix form, is

$$Y = X\beta + U, \tag{19.45}$$

where U is the $n \times 1$ vector of errors in the equation of interest, with i^{th} element u_i.

The matrix Z consists of all the exogenous regressors, so under the IV regression assumptions in Key Concept 12.4,

$$E(Z_i u_i) = 0 \quad \text{(instrument exogeneity).} \tag{19.46}$$

Because there are k included endogenous regressors, the first stage regression consists of k equations.

The TSLS estimator. The TSLS estimator is the instrumental variables estimator in which the instruments are the predicted values of X based on OLS estimation of the first-stage regression. Let $\hat{X}$ denote this matrix of predicted values, so that the ith row of $\hat{X}$ is $(\hat{X}_{1i} \quad \hat{X}_{2i} \quad \ldots \quad \hat{X}_{ki} \quad W_{1i} \quad W_{2i} \quad \ldots \quad W_{ri})$, where $\hat{X}_{1i}$ is the predicted value from the regression of X_{1i} on Z and so forth. Because the W's are contained in Z, the predicted value from a regression of W_{1i} on Z is just W_{1i} and so forth, so $\hat{X} = P_Z X$, where $P_Z = Z(Z'Z)^{-1}Z'$ [see Equation (19.27)]. Accordingly, the TSLS estimator is

$$\hat{\beta}^{TSLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y. \tag{19.47}$$

Because $\hat{X} = P_Z X$, $\hat{X}'\hat{X} = X'P_Z X$, and $\hat{X}'Y = X'P_Z Y$, the TSLS estimator can be rewritten as

$$\hat{\beta}^{TSLS} = (X'P_Z X)^{-1}X'P_Z Y. \tag{19.48}$$

Asymptotic Distribution of the TSLS Estimator

Substituting Equation (19.45) into Equation (19.48), rearranging, and multiplying by $\sqrt{n}$ yields the expression for the centered and scaled TSLS estimator:

$$\sqrt{n}(\hat{\beta}^{TSLS} - \beta) = \left(\frac{X'P_Z X}{n}\right)^{-1} \frac{X'P_Z U}{\sqrt{n}}$$

$$= \left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'X}{n}\right]^{-1}\left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'U}{\sqrt{n}}\right], \tag{19.49}$$

where the second equality uses the definition of P_Z. Under the IV regression assumptions, $X'Z/n \overset{p}{\longrightarrow} Q_{XZ}$ and $Z'Z/n \overset{p}{\longrightarrow} Q_{ZZ}$, where $Q_{XZ} = E(X_i Z_i')$ and $Q_{ZZ} = E(Z_i Z_i')$. In addition, under the IV regression assumptions, $Z_i u_i$ is i.i.d. with

mean 0 [Equation (19.46)] and a positive definite covariance matrix, so its sum, divided by $\sqrt{n}$, satisfies the conditions of the multivariate central limit theorem (Key Concept 19.2) and

$$\mathbf{Z}'\mathbf{U}/\sqrt{n} \xrightarrow{d} \mathbf{\Psi}_{ZU}, \text{ where } \mathbf{\Psi}_{ZU} \sim N(\mathbf{0}, \mathbf{H}), \mathbf{H} = E(\mathbf{Z}_i \mathbf{Z}_i' u_i^2) \quad (19.50)$$

and $\mathbf{\Psi}_{ZU}$ is $(m + r + 1) \times 1$.

Application of Equation (19.50) and of the limits $\mathbf{X}'\mathbf{Z}/n \xrightarrow{p} \mathbf{Q}_{XZ}$ and $\mathbf{Z}'\mathbf{Z}/n \xrightarrow{p} \mathbf{Q}_{ZZ}$ to Equation (19.49) yields the result that, under the IV regression assumptions, the TSLS estimator is asymptotically normally distributed:

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}^{TSLS} - \boldsymbol{\beta}) \xrightarrow{d} (\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{ZX})^{-1}\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{\Psi}_{ZU} \sim N(\mathbf{0}, \mathbf{\Sigma}^{TSLS}), \quad (19.51)$$

where

$$\mathbf{\Sigma}^{TSLS} = (\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{ZX})^{-1}\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{H}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{ZX}(\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{ZX})^{-1}, \quad (19.52)$$

where $\mathbf{H}$ is defined in Equation (19.50).

Standard errors for TSLS. The formula in Equation (19.52) is daunting. Nevertheless, it provides a way to estimate $\mathbf{\Sigma}^{TSLS}$ by substituting sample moments for the population moments. The resulting variance estimator is

$$\hat{\mathbf{\Sigma}}^{TSLS} = (\hat{\mathbf{Q}}_{XZ}\hat{\mathbf{Q}}_{ZZ}^{-1}\hat{\mathbf{Q}}_{ZX})^{-1}\hat{\mathbf{Q}}_{XZ}\hat{\mathbf{Q}}_{ZZ}^{-1}\hat{\mathbf{H}}\hat{\mathbf{Q}}_{ZZ}^{-1}\hat{\mathbf{Q}}_{ZX}(\hat{\mathbf{Q}}_{XZ}\hat{\mathbf{Q}}_{ZZ}^{-1}\hat{\mathbf{Q}}_{ZX})^{-1}, \quad (19.53)$$

where $\hat{\mathbf{Q}}_{XZ} = \mathbf{X}'\mathbf{Z}/n, \hat{\mathbf{Q}}_{ZZ} = \mathbf{Z}'\mathbf{Z}/n, \hat{\mathbf{Q}}_{ZX} = \mathbf{Z}'\mathbf{X}/n$, and

$$\hat{\mathbf{H}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i\hat{u}_i^2, \text{ where } \hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{TSLS}, \quad (19.54)$$

so that $\hat{\mathbf{U}}$ is the vector of TSLS residuals, and where $\hat{u}_i$ is the ith element of that vector (the TSLS residual for the ith observation).

The TSLS standard errors are the square roots of the diagonal elements of $\hat{\mathbf{\Sigma}}^{TSLS}/n$.

Properties of TSLS When the Errors Are Homoskedastic

If the errors are homoskedastic, then the TSLS estimator is asymptotically efficient among the class of IV estimators in which the instruments are linear combinations of the rows of $\mathbf{Z}$. This result is the IV counterpart to the Gauss–Markov theorem and constitutes an important justification for using TSLS.

The TSLS distribution under homoskedasticity. If the errors are homoskedastic—that is, if $E(u_i^2 | \mathbf{Z}_i) = \sigma_u^2$—then $\mathbf{H} = E(\mathbf{Z}_i \mathbf{Z}_i' u_i^2) = E[E(\mathbf{Z}_i \mathbf{Z}_i' u_i^2 | \mathbf{Z}_i)] = E[\mathbf{Z}_i \mathbf{Z}_i' E(u_i^2 | \mathbf{Z}_i)] = \mathbf{Q}_{ZZ}\sigma_u^2$. In this case, the variance of the asymptotic distribution of the TSLS estimator in Equation (19.52) simplifies to

$$\mathbf{\Sigma}^{TSLS} = (\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{ZX})^{-1}\sigma_u^2 \quad \text{(homoskedasticity only).} \quad (19.55)$$

The homoskedasticity-only estimator of the TSLS variance matrix is

$$\widetilde{\boldsymbol{\Sigma}}^{TSLS} = (\hat{\boldsymbol{Q}}_{XZ}\hat{\boldsymbol{Q}}_{ZZ}^{-1}\hat{\boldsymbol{Q}}_{ZX})^{-1}\hat{\sigma}_u^2, \text{ where } \hat{\sigma}_u^2 = \frac{\hat{\boldsymbol{U}}'\hat{\boldsymbol{U}}}{n - k - r - 1}$$

(homoskedasticity only),   (19.56)

and the homoskedasticity-only TSLS standard errors are the square roots of the diagonal elements of $\widetilde{\boldsymbol{\Sigma}}^{TSLS}/n$.

The class of IV estimators that use linear combinations of Z. The class of IV estimators that use linear combinations of $\boldsymbol{Z}$ as instruments can be generated in two equivalent ways. Both start with the same moment equation: Under the assumption of instrument exogeneity, the errors $\boldsymbol{U} = \boldsymbol{Y} - \boldsymbol{X\beta}$ are uncorrelated with the exogenous regressors; that is, at the true value of $\boldsymbol{\beta}$, Equation (19.46) implies that

$$E[(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{Z}] = 0. \tag{19.57}$$

Equation (19.57) constitutes a system of $m + r + 1$ equations involving the $k + r + 1$ unknown elements of $\boldsymbol{\beta}$. When $m > k$, these equations are redundant in the sense that all are satisfied at the true value of $\boldsymbol{\beta}$. When these population moments are replaced by their sample moments, the system of equations $(\boldsymbol{Y} - \boldsymbol{Xb})'\boldsymbol{Z} = 0$ can be solved for $\boldsymbol{b}$ when there is exact identification ($m = k$). This value of $\boldsymbol{b}$ is the IV estimator of $\boldsymbol{\beta}$. However, when there is overidentification ($m > k$), the equations in the system cannot be simultaneously satisfied by the same value of $\boldsymbol{b}$ because of sampling variation—there are more equations than unknowns—and, in general, this system does not have a solution.

The first approach to the problem of estimating $\boldsymbol{\beta}$ when there is overidentification is to trade off the desire to satisfy each equation by minimizing a quadratic form involving all the equations. Specifically, let $\boldsymbol{A}$ be an $(m + r + 1) \times (m + r + 1)$ symmetric positive semidefinite weight matrix, and let $\hat{\boldsymbol{\beta}}_A^{IV}$ denote the estimator that minimizes

$$\min_b (\boldsymbol{Y} - \boldsymbol{Xb})'\boldsymbol{ZAZ}'(\boldsymbol{Y} - \boldsymbol{Xb}). \tag{19.58}$$

The solution to this minimization problem is found by taking the derivative of the objective function with respect to $\boldsymbol{b}$, setting the result equal to 0, and rearranging. Doing so yields $\hat{\boldsymbol{\beta}}_A^{IV}$, the IV estimator based on the weight matrix $\boldsymbol{A}$:

$$\hat{\boldsymbol{\beta}}_A^{IV} = (\boldsymbol{X}'\boldsymbol{ZAZ}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{ZAZ}'\boldsymbol{Y}. \tag{19.59}$$

Comparison of Equations (19.59) and (19.48) shows that the TSLS estimator is the IV estimator with $\boldsymbol{A} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. That is, TSLS is the solution of the minimization problem in Equation (19.58) with $\boldsymbol{A} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

The calculations leading to Equations (19.51) and (19.52), applied to $\hat{\boldsymbol{\beta}}_A^{IV}$, show that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_A^{IV} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_A^{IV}), \text{ where}$$

$$\boldsymbol{\Sigma}_A^{IV} = (\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{H}\boldsymbol{A}\boldsymbol{Q}_{ZX}(\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}. \qquad (19.60)$$

The second way to generate the class of IV estimators that use linear combinations of $\boldsymbol{Z}$ is to consider IV estimators in which the instruments are $\boldsymbol{ZB}$, where $\boldsymbol{B}$ is an $(m + r + 1) \times (k + r + 1)$ matrix with full column rank. Then the system of $(k + r + 1)$ equations, $(\boldsymbol{Y} - \boldsymbol{Xb})'\boldsymbol{ZB} = 0$, can be solved uniquely for the $(k + r + 1)$ unknown elements of $\boldsymbol{b}$. Solving these equations for $\boldsymbol{b}$ yields $\hat{\boldsymbol{\beta}}^{IV} = (\boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{Y})$, and substitution of $\boldsymbol{B} = \boldsymbol{AZ}'\boldsymbol{X}$ into this expression yields Equation (19.59).

Thus the two approaches to defining IV estimators that are linear combinations of the instruments yield the same family of IV estimators. It is conventional to work with the first approach, in which the IV estimator solves the quadratic minimization problem in Equation (19.58), and that is the approach taken here.

Asymptotic efficiency of TSLS under homoskedasticity. If the errors are homoskedastic, then $\boldsymbol{H} = \boldsymbol{Q}_{ZZ}\sigma_u^2$, and the expression for $\boldsymbol{\Sigma}_A^{IV}$ in Equation (19.60) becomes

$$\boldsymbol{\Sigma}_A^{IV} = (\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZZ}\boldsymbol{A}\boldsymbol{Q}_{ZX}(\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\sigma_u^2. \qquad (19.61)$$

To show that TSLS is asymptotically efficient among the class of estimators that are linear combinations of $\boldsymbol{Z}$ when the errors are homoskedastic, we need to show that, under homoskedasticity,

$$\boldsymbol{c}'\boldsymbol{\Sigma}_A^{IV}\boldsymbol{c} \geq \boldsymbol{c}'\boldsymbol{\Sigma}^{TSLS}\boldsymbol{c} \qquad (19.62)$$

for all positive semidefinite matrices $\boldsymbol{A}$ and all $(k + r + 1) \times 1$ vectors $\boldsymbol{c}$, where $\boldsymbol{\Sigma}^{TSLS} = (\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX})^{-1}\sigma_u^2$ [Equation (19.55)]. The inequality (19.62), which is proven in Appendix 19.6, is the same efficiency criterion as is used in the multivariate Gauss–Markov theorem in Key Concept 19.3. Consequently, TSLS is the efficient IV estimator under homoskedasticity among the class of estimators in which the instruments are linear combinations of $\boldsymbol{Z}$.

The J-statistic under homoskedasticity. The *J*-statistic (Key Concept 12.6) tests the null hypothesis that all the overidentifying restrictions hold against the alternative that some or all of them do not hold.

The idea of the *J*-statistic is that, if the overidentifying restrictions hold, u_i will be uncorrelated with the instruments, and thus a regression of $\boldsymbol{U}$ on $\boldsymbol{Z}$ will have population regression coefficients that all equal 0. In practice, $\boldsymbol{U}$ is not observed, but it can be estimated by the TSLS residuals $\hat{\boldsymbol{U}}$, so a regression of $\hat{\boldsymbol{U}}$ on $\boldsymbol{Z}$ should yield statistically insignificant coefficients. Accordingly, the TSLS *J*-statistic is the homoskedasticity-only *F*-statistic testing the hypothesis that the coefficients on $\boldsymbol{Z}$ are all 0, in the regression of $\hat{\boldsymbol{U}}$ on $\boldsymbol{Z}$, multiplied by $(m + r + 1)$ so that the *F*-statistic is in its asymptotic chi-squared form.

An explicit formula for the *J*-statistic can be obtained using Equation (7.13) for the homoskedasticity-only *F*-statistic. The unrestricted regression is the regression of $\hat{U}$ on the $m + r + 1$ regressors Z, and the restricted regression has no regressors. Thus, in the notation of Equation (7.13), $SSR_{unrestricted} = \hat{U}'M_Z\hat{U}$, and $SSR_{restricted} = \hat{U}'\hat{U}$, so $SSR_{restricted} - SSR_{unrestricted} = \hat{U}'\hat{U} - \hat{U}'M_Z\hat{U} = \hat{U}'P_Z\hat{U}$ and the *J*-statistic is

$$J = \frac{\hat{U}'P_Z\hat{U}}{\hat{U}'M_Z\hat{U}/(n - m - r - 1)}. \tag{19.63}$$

The method for computing the *J*-statistic described in Key Concept 12.6 entails testing only the hypothesis that the coefficients on the excluded instruments are 0. Although these two methods have different computational steps, they produce identical *J*-statistics (Exercise 19.14).

It is shown in Appendix 19.6 that, under the null hypothesis that $E(u_i Z_i) = 0$,

$$J \xrightarrow{d} \chi^2_{m-k}. \tag{19.64}$$

Generalized Method of Moments Estimation in Linear Models

If the errors are heteroskedastic, then the TSLS estimator is no longer efficient among the class of IV estimators that use linear combinations of Z as instruments. The efficient estimator in this case is known as the efficient generalized method of moments (GMM) estimator. In addition, if the errors are heteroskedastic, then the *J*-statistic as defined in Equation (19.63) no longer has a chi-squared distribution. However, an alternative formulation of the *J*-statistic, constructed using the efficient GMM estimator, does have a chi-squared distribution with $m - k$ degrees of freedom.

These results parallel the results for the estimation of the usual regression model with exogenous regressors and heteroskedastic errors: If the errors are heteroskedastic, then the OLS estimator is not efficient among estimators that are linear in Y (the Gauss–Markov conditions are not satisfied), and the homoskedasticity-only *F*-statistic no longer has an *F* distribution, even in large samples. In the regression model with exogenous regressors and heteroskedasticity, the efficient estimator is weighted least squares; in the IV regression model with heteroskedasticity, the efficient estimator uses a different weighting matrix than TSLS, and the resulting estimator is the efficient GMM estimator.

GMM estimation. **Generalized method of moments (GMM)** estimation is a general method for the estimation of the parameters of linear or nonlinear models, in which the parameters are chosen to provide the best fit to multiple equations, each of which sets a sample moment to 0. These equations, which in the context of GMM are called moment conditions, typically cannot all be satisfied simultaneously. The GMM estimator trades off the desire to satisfy each of the equations by minimizing a quadratic objective function.

In the linear IV regression model with exogenous variables Z, the class of GMM estimators consists of all the estimators that are solutions to the quadratic minimization problem in Equation (19.58). Thus the class of GMM estimators based on the full set of instruments Z with different-weight matrices A is the same as the class of IV estimators in which the instruments are linear combinations of Z. In the linear IV regression model, GMM is just another name for the class of estimators we have been studying—that is, estimators that solve Equation (19.58).

The asymptotically efficient GMM estimator.   Among the class of GMM estimators, the **efficient GMM** estimator is the GMM estimator with the smallest asymptotic variance matrix [where the smallest variance matrix is defined as in Equation (19.62)]. Thus the result in Equation (19.62) can be restated as saying that TSLS is the efficient GMM estimator in the linear model when the errors are homoskedastic.

To motivate the expression for the efficient GMM estimator when the errors are heteroskedastic, recall that when the errors are homoskedastic, H [the variance matrix of $Z_i u_i$; see Equation (19.50)] equals $Q_{ZZ}\sigma_u^2$, and the asymptotically efficient weight matrix is obtained by setting $A = (Z'Z)^{-1}$, which yields the TSLS estimator. In large samples, using the weight matrix $A = (Z'Z)^{-1}$ is equivalent to using $A = (Q_{ZZ}\sigma_u^2)^{-1} = H^{-1}$. This interpretation of the TSLS estimator suggests that, by analogy, the efficient IV estimator under heteroskedasticity can be obtained by setting $A = H^{-1}$ and solving

$$\min_b (Y - Xb)'ZH^{-1}Z'(Y - Xb). \tag{19.65}$$

This analogy is correct: The solution to the minimization problem in Equation (19.65) is the efficient GMM estimator. Let $\widetilde{\beta}^{Eff.GMM}$ denote the solution to the minimization problem in Equation (19.65). By Equation (19.59), this estimator is

$$\widetilde{\beta}^{Eff.GMM} = (X'ZH^{-1}Z'X)^{-1}X'ZH^{-1}Z'Y. \tag{19.66}$$

The asymptotic distribution of $\widetilde{\beta}^{Eff.GMM}$ is obtained by substituting $A = H^{-1}$ into Equation (19.60) and simplifying; thus

$$\sqrt{n}(\widetilde{\beta}^{Eff.GMM} - \beta) \xrightarrow{d} N(0, \Sigma^{Eff.GMM}),$$

$$\text{where } \Sigma^{Eff.GMM} = (Q_{XZ}H^{-1}Q_{ZX})^{-1}. \tag{19.67}$$

The result that $\widetilde{\beta}^{Eff.GMM}$ is the efficient GMM estimator is proven by showing that $c'\Sigma_A^{IV}c \geq c'\Sigma^{Eff.GMM}c$ for all vectors c, where Σ_A^{IV} is given in Equation (19.60). The proof of this result is given in Appendix 19.6.

Feasible efficient GMM estimation.   The GMM estimator defined in Equation (19.66) is not a feasible estimator because it depends on the unknown variance matrix H. However, a feasible efficient GMM estimator can be computed by

substituting a consistent estimator of H into the minimization problem of Equation (19.65) or, equivalently, by substituting a consistent estimator of H into the formula for $\hat{\boldsymbol{\beta}}^{Eff.GMM}$ in Equation (19.66).

The efficient GMM estimator can be computed in two steps. In the first step, estimate $\boldsymbol{\beta}$ using any consistent estimator. Use this estimator of $\boldsymbol{\beta}$ to compute the residuals from the equation of interest, and then use these residuals to compute an estimator of H. In the second step, use this estimator of H to estimate the optimal weight matrix H^{-1} and to compute the efficient GMM estimator. To be concrete, in the linear IV regression model, it is natural to use the TSLS estimator in the first step and to use the TSLS residuals to estimate H. If TSLS is used in the first step, then the feasible efficient GMM estimator computed in the second step is

$$\hat{\boldsymbol{\beta}}^{Eff.GMM} = (\boldsymbol{X}'\boldsymbol{Z}\hat{\boldsymbol{H}}^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\hat{\boldsymbol{H}}^{-1}\boldsymbol{Z}'\boldsymbol{Y}, \tag{19.68}$$

where $\hat{\boldsymbol{H}}$ is given in Equation (19.54).

Because $\hat{\boldsymbol{H}} \xrightarrow{p} \boldsymbol{H}$, $\sqrt{n}(\hat{\boldsymbol{\beta}}^{Eff.GMM} - \tilde{\boldsymbol{\beta}}^{Eff.GMM}) \xrightarrow{p} 0$ (Exercise 19.12), and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{Eff.GMM} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}^{Eff.GMM}), \tag{19.69}$$

where $\boldsymbol{\Sigma}^{Eff.GMM} = (\boldsymbol{Q}_{XZ}\boldsymbol{H}^{-1}\boldsymbol{Q}_{ZX})^{-1}$ [Equation (19.67)]. That is, the feasible two-step estimator $\hat{\boldsymbol{\beta}}^{Eff.GMM}$ in Equation (19.68) is, asymptotically, the efficient GMM estimator.

The heteroskedasticity-robust J-statistic. The **heteroskedasticity-robust J-statistic**, also known as the **GMM J-statistic**, is the counterpart of the TSLS-based J-statistic, computed using the efficient GMM estimator and weight function. That is, the GMM J-statistic is given by

$$J^{GMM} = (\boldsymbol{Z}'\hat{\boldsymbol{U}}^{GMM})'\hat{\boldsymbol{H}}^{-1}(\boldsymbol{Z}'\hat{\boldsymbol{U}}^{GMM})/n, \tag{19.70}$$

where $\hat{\boldsymbol{U}}^{GMM} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{Eff.GMM}$ are the residuals from the equation of interest, estimated by (feasible) efficient GMM, and $\hat{\boldsymbol{H}}^{-1}$ is the weight matrix used to compute $\hat{\boldsymbol{\beta}}^{Eff.GMM}$.

Under the null hypothesis $E(\boldsymbol{Z}_i u_i) = \boldsymbol{0}$, $J^{GMM} \xrightarrow{d} \chi^2_{m-k}$ (see Appendix 19.6).

GMM with time series data. The results in this section were derived under the IV regression assumptions for cross–sectional data. In many applications, however, these results extend to time series applications of IV regression and GMM. Although a formal mathematical treatment of GMM with time series data is beyond the scope of this book (for such a treatment, see Hayashi, 2000, Chapter 6), we nevertheless will summarize the key ideas of GMM estimation with time series data. This summary assumes familiarity with the material in Chapters 14 and 16. For this discussion, it is assumed that the variables are stationary.

It is useful to distinguish between two types of applications: applications in which the error term u_t is serially correlated and applications in which u_t is serially uncorrelated. If the error term u_t is serially correlated, then the asymptotic distribution of the GMM estimator continues to be normally distributed, but the formula for $\boldsymbol{H}$ in Equation (19.50) is no longer correct. Instead, the correct expression for $\boldsymbol{H}$ depends on the autocovariances of $\boldsymbol{Z}_t u_t$ and is analogous to the formula given in Equation (16.14) for the variance of the OLS estimator when the error term is serially correlated. The efficient GMM estimator is still constructed using a consistent estimator of $\boldsymbol{H}$; however, that consistent estimator must be computed using the HAC methods discussed in Chapter 16.

If $\boldsymbol{Z}_t u_t$ is not serially correlated, then HAC estimation of $\boldsymbol{H}$ is unnecessary, and the formulas presented in this section all extend to time series GMM applications. In modern applications to finance and macroeconometrics, it is common to encounter models in which the error term represents an unexpected or unforecastable disturbance, in which case the model typically implies that $\boldsymbol{Z}_t u_t$ is serially uncorrelated. For example, consider a model with a single included endogenous variable and no included exogenous variables so that the equation of interest is $Y_t = \beta_0 + \beta_1 X_t + u_t$. Suppose that an economic theory implies that u_t is unpredictable given past information. Then the theory implies the moment condition

$$E(u_t | Y_{t-1}, X_{t-1}, Z_{t-1}, Y_{t-2}, X_{t-2}, Z_{t-2}, \dots) = 0, \qquad (19.71)$$

where Z_{t-1} is the lagged value of some other variable. The moment condition in Equation (19.71) implies that all the lagged variables $Y_{t-1}, X_{t-1}, Z_{t-1}, Y_{t-2}, X_{t-2}, Z_{t-2}, \dots$ are candidates for being valid instruments (they satisfy the exogeneity condition). Moreover, because $u_{t-1} = Y_{t-1} - \beta_0 - \beta_1 X_{t-1}$, the moment condition in Equation (19.71) is equivalent to $E(u_t | u_{t-1}, X_{t-1}, Z_{t-1}, u_{t-2}, X_{t-2}, Z_{t-2}, \dots) = 0$. Because u_t is serially uncorrelated, HAC estimation of $\boldsymbol{H}$ is unnecessary. The theory of GMM presented in this section, including efficient GMM estimation and the GMM J-statistic, therefore applies directly to time series applications with moment conditions of the form in Equation (19.71), under the hypothesis that the moment condition in Equation (19.71) is, in fact, correct.

Summary

1. The linear multiple regression model in matrix form is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}$, where $\boldsymbol{Y}$ is the $n \times 1$ vector of observations on the dependent variable, $\boldsymbol{X}$ is the $n \times (k + 1)$ matrix of n observations on the $k + 1$ regressors (including a constant), $\boldsymbol{\beta}$ is the $k + 1$ vector of unknown parameters, and $\boldsymbol{U}$ is the $n \times 1$ vector of error terms.

2. The OLS estimator is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$. Under the first four least squares assumptions in Key Concept 19.1, $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed. If in addition the errors are homoskedastic, then the conditional variance of $\hat{\boldsymbol{\beta}}$ is $\text{var}(\hat{\boldsymbol{\beta}} | \boldsymbol{X}) = \sigma_u^2 (\boldsymbol{X'X})^{-1}$.

3. General linear restrictions on $\boldsymbol{\beta}$ can be written as the q equations $\boldsymbol{R\beta} = \boldsymbol{r}$, and this formulation can be used to test joint hypotheses involving multiple coefficients or to construct confidence sets for elements of $\boldsymbol{\beta}$.

4. When the regression errors are i.i.d. and normally distributed, conditional on $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}}$ has an exact normal distribution, and the homoskedasticity-only t- and F-statistics have exact t_{n-k-1} and $F_{q, n-k-1}$ distributions, respectively.

5. The Gauss–Markov theorem says that, if the errors are homoskedastic and conditionally uncorrelated across observations and if $E(u_i|\boldsymbol{X}) = 0$, the OLS estimator is efficient among linear conditionally unbiased estimators (that is, OLS is BLUE).

6. If the error covariance matrix $\boldsymbol{\Omega}$ is not proportional to the identity matrix and if $\boldsymbol{\Omega}$ is known or can be estimated, then the GLS estimator is asymptotically more efficient than OLS. However, GLS requires that, in general, u_i be uncorrelated with *all* observations on the regressors, not just with $\boldsymbol{X}_i$, as is required by OLS, an assumption that must be evaluated carefully in applications.

7. The TSLS estimator is a member of the class of GMM estimators of the linear model. In GMM, the coefficients are estimated by making the sample covariance between the regression error and the exogenous variables as small as possible—specifically, by solving $\min_b [(\boldsymbol{Y} - \boldsymbol{Xb})'\boldsymbol{Z}]\boldsymbol{A}[\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{Xb})]$, where $\boldsymbol{A}$ is a non-random positive definite matrix. The asymptotically efficient GMM estimator sets $\boldsymbol{A} = [E(\boldsymbol{Z}_i\boldsymbol{Z}_i' u_i^2)]^{-1}$. When the errors are homoskedastic, the asymptotically efficient GMM estimator in the linear IV regression model is TSLS.

Key Terms

Gauss–Markov conditions for multiple regression (684)

Gauss–Markov theorem for multiple regression (685)

generalized least squares (GLS) (686)

infeasible GLS (689)

feasible GLS (689)

generalized method of moments (GMM) (696)

efficient GMM (697)

heteroskedasticity-robust J-statistic (698)

GMM J-statistic (698)

mean vector (710)

covariance matrix (710)

Review the Concepts

19.1 A researcher studying the relationship between earnings and workers' sex speci-
fies the regression model $Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + u_i$, where X_{1i} is a binary
variable that equals 1 if the ith person is a female and X_{2i} is a binary variable
that equals 1 if the ith person is a male. Write the model in the matrix form of
Equation (19.2) for a hypothetical set of $n = 5$ observations. Show that the col-
umns of X are linearly dependent, so that X does not have full rank. Explain how
you would respecify the model to eliminate the perfect multicollinearity.

19.2 You are analyzing a linear regression model with 500 observations and one
regressor. Explain how you would construct a confidence interval for β_1 if

    a. Assumptions 1 through 4 in Key Concept 19.1 are true but you think
assumption 5 or 6 might not be true.

    b. Assumptions 1 through 5 are true but you think assumption 6 might not
be true. (Give two ways to construct the confidence interval.)

    c. Assumptions 1 through 6 are true.

19.3 Suppose that assumptions 1 through 5 in Key Concept 19.1 are true but that
assumption 6 is not. Does the result in Equation (19.31) hold? Explain.

19.4 Can you compute the BLUE estimator of β if Equation (19.41) holds and you
do not know Ω? What if you know Ω?

19.5 Construct an example of a regression model that satisfies the assumption
$E(u_i \mid X_i) = 0$ but for which $E(U \mid X) \neq \mathbf{0}_n$.

Exercises

19.1 Consider the population regression of test scores against income and the
square of income in Equation (8.1).

    a. Write the regression in Equation (8.1) in the matrix form of Equation
(19.5). Define Y, X, U, and β.

    b. Explain how to test the null hypothesis that the relationship between test
scores and income is linear against the alternative that it is quadratic. Write
the null hypothesis in the form of Equation (19.20). What are R, r, and q?

19.2 Suppose that a sample of $n = 20$ households has the sample means and sam-
ple covariances below for a dependent variable and two regressors:

	Sample Means	Sample Covariances		
		Y	X_1	X_2
Y	6.39	0.26	0.22	0.32
X_1	7.24		0.80	0.28
X_2	4.00			2.40

   a. Calculate the OLS estimates of β_0, β_1, and β_2. Calculate $s_{\hat{u}}^2$. Calculate the R^2 of the regression.

   b. Suppose that all six assumptions in Key Concept 19.1 hold. Test the hypothesis that $\beta_1 = 0$ at the 5% significance level.

19.3 Let W be an $m \times 1$ vector with covariance matrix Σ_W, where Σ_W is finite and positive definite. Let c be a nonrandom $m \times 1$ vector, and let $Q = c'W$.

   a. Show that $\text{var}(Q) = c'\Sigma_W c$.

   b. Suppose that $c \neq 0_m$. Show that $0 < \text{var}(Q) < \infty$.

19.4 Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ from Chapter 4, and assume that the least squares assumptions in Key Concept 4.3 hold.

   a. Write the model in the matrix form given in Equations (19.2) and (19.3).

   b. Show that assumptions 1 through 4 in Key Concept 19.1 are satisfied.

   c. Use the general formula for $\hat{\beta}$ in Equation (19.11) to derive the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ given in Key Concept 4.2.

   d. Show that the $(1, 1)$ element of $\Sigma_{\hat{\beta}}$ in Equation (19.13) is equal to the expression for $\sigma_{\hat{\beta}_0}^2$ given in Key Concept 4.4.

19.5 Let P_X and M_X be as defined in Equations (19.24) and (19.25).

   a. Prove that $P_X M_X = 0_{n \times n}$ and that P_X and M_X are idempotent.

   b. Derive Equations (19.27) and (19.28).

   c. Show that $\text{rank}(P_X) = k + 1$ and $\text{rank}(M_X) = n - k - 1$. [*Hint:* First solve Exercise 19.10, and then use the fact that $\text{trace}(AB) = \text{trace}(BA)$ for conformable matrices A and B.]

19.6 Consider the regression model in matrix form, $Y = X\beta + W\gamma + U$, where X is an $n \times k_1$ matrix of regressors and W is an $n \times k_2$ matrix of regressors. Then, as shown in Exercise 19.17, the OLS estimator $\hat{\beta}$ can be expressed

$$\hat{\beta} = (X'M_W X)^{-1}(X'M_W Y).$$

Now let $\hat{\beta}_1^{BV}$ be the "binary variable" fixed effects estimator computed by estimating Equation (10.11) by OLS, and let $\hat{\beta}_1^{DM}$ be the "demeaning" fixed effects estimator computed by estimating Equation (10.14) by OLS, in which the entity-specific sample means have been subtracted from X and Y. Use the expression for $\hat{\beta}$ given above to prove that $\hat{\beta}_1^{BV} = \hat{\beta}_1^{DM}$. [*Hint:* Write Equation (10.11) using a full set of fixed effects, $D1_i, D2_i, \ldots, D_{ni}$ and no constant term. Include all of the fixed effects in W. Write out the matrix $M_W X$.]

19.7 Consider the regression model $Y_i = \beta_1 X_i + \beta_2 W_i + u_i$, where for simplicity the intercept is omitted and all variables are assumed to have a mean of 0. Suppose that X_i is distributed independently of (W_i, u_i) but W_i and u_i might be correlated, and let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the OLS estimators for this model.

a. Show that whether or not W_i and u_i are correlated, $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

b. Show that if W_i and u_i are correlated, then $\hat{\beta}_2$ is inconsistent.

c. Let $\hat{\beta}_1^r$ be the OLS estimator from the regression of Y on X (the restricted regression that excludes W). Will $\hat{\beta}_1$ have a smaller asymptotic variance than $\hat{\beta}_1^r$, allowing for the possibility that W_i and u_i are correlated? Explain.

19.8 Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $u_1 = \tilde{u}_1$ and $u_i = 0.5u_{i-1} + \tilde{u}_i$ for $i = 2, 3, \ldots, n$. Suppose that $\tilde{u}_i$ are i.i.d. with mean 0 and variance 1 and are distributed independently of X_j for all i and j.

a. Derive an expression for $E(UU') = \Omega$.

b. Explain how to estimate the model by GLS without explicitly inverting the matrix Ω. (*Hint:* Transform the model so that the regression errors are $\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_n$.)

19.9 This exercise shows that the OLS estimator of a subset of the regression coefficients is consistent under the conditional mean independence assumption stated in Key Concept 6.6. Consider the multiple regression model in matrix form $Y = X\beta + W\gamma + U$, where X and W are, respectively, $n \times k_1$ and $n \times k_2$ matrices of regressors. Let X_i' and W_i' denote the ith rows of X and W [as in Equation (19.4)]. Assume that (i) $E(u_i|X_i, W_i) = W_i'\delta$, where δ is a $k_2 \times 1$ vector of unknown parameters; (ii) (X_i, W_i, Y_i) are i.i.d.; (iii) (X_i, W_i, u_i) have four finite nonzero moments; and (iv) there is no perfect multicollinearity. These are assumptions 1 through 4 of Key Concept 19.1, with the conditional mean independence assumption (i) replacing the usual conditional mean 0 assumption.

a. Use the expression for $\hat{\beta}$ given in Exercise 19.6 to write $\hat{\beta} - \beta = (n^{-1}X'M_WX)^{-1}(n^{-1}X'M_WU)$.

b. Show that $n^{-1}X'M_WX \xrightarrow{p} \Sigma_{XX} - \Sigma_{XW}\Sigma_{WW}^{-1}\Sigma_{WX}$, where $\Sigma_{XX} = E(X_iX_i')$, $\Sigma_{XW} = E(X_iW_i')$, and so forth. [The matrix $A_n \xrightarrow{p} A$ if $A_{n,ij} \xrightarrow{p} A_{ij}$ for all i, j pairs, where $A_{n,ij}$ and A_{ij} are the (i, j) elements of A_n and A.]

c. Show that assumptions (i) and (ii) imply that $E(U|X, W) = W\delta$.

d. Use (c) and the law of iterated expectations to show that $n^{-1}X'M_WU \xrightarrow{p} 0_{k_1 \times 1}$.

e. Use (a) through (d) to conclude that, under assumptions (i) through (iv), $\hat{\beta} \xrightarrow{p} \beta$.

19.10 Let C be a symmetric idempotent matrix.

a. Show that the eigenvalues of C are either 0 or 1. (*Hint:* Note that $Cq = \gamma q$ implies $0 = Cq - \gamma q = CCq - \gamma q = \gamma Cq - \gamma q = \gamma^2 q - \gamma q$, and solve for γ.)

b. Show that $\text{trace}(C) = \text{rank}(C)$.

c. Let d be an $n \times 1$ vector. Show that $d'Cd \geq 0$.

19.11 Suppose that C is an $n \times n$ symmetric idempotent matrix with rank r, and let $V \sim N(\mathbf{0}_n, I_n)$.

    a. Show that $C = AA'$, where A is $n \times r$ with $A'A = I_r$. (*Hint: C* is positive semidefinite and can be written as $Q \Lambda Q'$, as explained in Appendix 19.1.)

    b. Show that $A'V \sim N(\mathbf{0}_r, I_r)$.

    c. Show that $V'CV \sim \chi_r^2$.

19.12 **a.** Show that $\widetilde{\boldsymbol{\beta}}^{Eff.GMM}$ is the efficient GMM estimator—that is, that $\widetilde{\boldsymbol{\beta}}^{Eff.GMM}$ in Equation (19.66) is the solution to Equation (19.65).

    b. Show that $\sqrt{n}(\hat{\boldsymbol{\beta}}^{Eff.GMM} - \widetilde{\boldsymbol{\beta}}^{Eff.GMM}) \xrightarrow{p} 0$.

    c. Show that $J^{GMM} \xrightarrow{d} \chi_{m-k}^2$.

19.13 Consider the problem of minimizing the sum of squared residuals, subject to the constraint that $Rb = r$, where R is $q \times (k + 1)$ with rank q. Let $\widetilde{\boldsymbol{\beta}}$ be the value of b that solves the constrained minimization problem.

    a. Show that the Lagrangian for the minimization problem is $L(b, \gamma) = (Y - Xb)'(Y - Xb) + \gamma'(Rb - r)$, where γ is a $q \times 1$ vector of Lagrange multipliers.

    b. Show that $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\boldsymbol{\beta}} - r)$.

    c. Show that $(Y - X\widetilde{\boldsymbol{\beta}})'(Y - X\widetilde{\boldsymbol{\beta}}) - (Y - X\hat{\boldsymbol{\beta}})(Y - X\hat{\boldsymbol{\beta}}) = (R\hat{\boldsymbol{\beta}} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\boldsymbol{\beta}} - r)$.

    d. Show that $\widetilde{F}$ in Equation (19.36) is equivalent to the homoskedasticity-only F-statistic in Equation (7.13).

19.14 Consider the regression model $Y = X\boldsymbol{\beta} + U$. Partition X as $[X_1 \ X_2]$ and $\boldsymbol{\beta}$ as $[\boldsymbol{\beta}_1' \ \boldsymbol{\beta}_2']'$, where X_1 has k_1 columns and X_2 has k_2 columns. Suppose that $X_2'Y = \mathbf{0}_{k_2 \times 1}$. Let $R = [I_{k_1} \ \mathbf{0}_{k_1 \times k_2}]$.

    a. Show that $\hat{\boldsymbol{\beta}}'(X'X)\hat{\boldsymbol{\beta}} = (R\hat{\boldsymbol{\beta}})'[R(X'X)^{-1}R]^{-1}(R\hat{\boldsymbol{\beta}})$.

    b. Consider the regression described in Equation (12.17). Let $W = [\mathbf{1} \ W_1 \ W_2 \ \ldots \ W_r]$, where $\mathbf{1}$ is an $n \times 1$ vector of 1's, W_1 is the $n \times 1$ vector with i^{th} element W_{1i}, and so forth. Let $\hat{U}^{TSLS}$ denote the vector of two stage least squares residuals.

       i. Show that $W'\hat{U}^{TSLS} = 0$.

       ii. Show that the method for computing the J-statistic described in Key Concept 12.6 (using a homoskedasticity-only F-statistic) and that using the formula in Equation (19.63) produce the same value for the J-statistic. [*Hint:* Use the results in (a), (b.i), and Exercise 19.13.]

19.15 (Consistency of clustered standard errors.) Consider the panel data model $Y_{it} = \beta X_{it} + \alpha_i + u_{it}$, where all variables are scalars. Assume that assumptions

1, 2, and 4 in Key Concept 10.3 hold and strengthen assumption 3, so that X_{it} and u_{it} have eight nonzero finite moments. Let $M = I_T - T^{-1} \boldsymbol{\iota} \boldsymbol{\iota}'$, where $\boldsymbol{\iota}$ is a $T \times 1$ vector of 1's. Also let $Y_i = (Y_{i1} \; Y_{i2} \; \cdots \; Y_{iT})'$, $X_i = (X_{i1} \; X_{i2} \; \cdots \; X_{iT})'$, $\boldsymbol{u}_i = (u_{i1} \; u_{i2} \; \cdots \; u_{iT})'$, $\tilde{Y}_i = M Y_i$, $\tilde{X}_i = M X_i$, and $\tilde{\boldsymbol{u}}_i = M \boldsymbol{u}_i$. For the asymptotic calculations in this problem, suppose that T is fixed and $n \longrightarrow \infty$.

a. Show that the fixed effects estimator of β from Section 10.3 can be written as $\hat{\beta} = (\sum_{i=1}^{n} \tilde{X}_i' \tilde{X}_i)^{-1} \sum_{i=1}^{n} \tilde{X}_i' \tilde{Y}_i$.

b. Show that $\hat{\beta} - \beta = (\sum_{i=1}^{n} \tilde{X}_i' \tilde{X}_i)^{-1} \sum_{i=1}^{n} \tilde{X}_i' \boldsymbol{u}_i$. (*Hint: M* is idempotent.)

c. Let $Q_{\tilde{X}} = T^{-1} E(\tilde{X}_i' \tilde{X}_i)$ and $\hat{Q}_{\tilde{X}} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{X}_{it}^2$. Show that $\hat{Q}_{\tilde{X}} \overset{p}{\longrightarrow} Q_{\tilde{X}}$.

d. Let $\eta_i = \tilde{X}_i' \boldsymbol{u}_i / \sqrt{T}$ and $\sigma_\eta^2 = \text{var}(\eta_i)$. Show that $\sqrt{\frac{1}{n}} \sum_{i=1}^{n} \eta_i \overset{d}{\longrightarrow} N(0, \sigma_\eta^2)$.

e. Use your answers to (b) through (d) to prove Equation (10.25); that is, show that $\sqrt{nT}(\hat{\beta} - \beta) \overset{d}{\longrightarrow} N(0, \sigma_\eta^2 / Q_{\tilde{X}}^2)$.

f. Let $\tilde{\sigma}_{\eta,clustered}^2$ be the infeasible clustered variance estimator, computed using the true errors instead of the residuals so that $\tilde{\sigma}_{\eta,clustered}^2 = \frac{1}{nT} \sum_{i=1}^{n} (\tilde{X}_i' \boldsymbol{u}_i)^2$. Show that $\tilde{\sigma}_{\eta,clustered}^2 \overset{p}{\longrightarrow} \sigma_\eta^2$.

g. Let $\hat{\tilde{\boldsymbol{u}}}_1 = \tilde{Y}_i - \hat{\beta} \tilde{X}_i$ and $\hat{\sigma}_{\eta, clustered}^2 = \frac{n}{n-1} \frac{1}{nT} \sum_{i=1}^{n} (\tilde{X}_i' \hat{\tilde{\boldsymbol{u}}}_i)^2$ [this is Equation (10.27) in matrix form]. Show that $\hat{\sigma}_{\eta, clustered}^2 \overset{p}{\longrightarrow} \sigma_\eta^2$. [*Hint:* Use an argument like that used in Equation (18.16) to show that $\hat{\sigma}_{\eta, clustered}^2 - \tilde{\sigma}_{\eta, clustered}^2 \overset{p}{\longrightarrow} 0$, and then use your answer to (f).]

19.16 This exercise takes up the problem of missing data discussed in Section 9.2. Consider the regression model $Y_i = X_i \beta + u_i$, $i = 1, \ldots, n$, where all variables are scalars and the constant term/intercept is omitted for convenience.

a. Suppose that the least squares assumptions in Key Concept 4.3 are satisfied. Show that the least squares estimator of β is unbiased and consistent.

b. Now suppose that some of the observations are missing. Let I_i denote a binary random variable that indicates the nonmissing observations; that is, $I_i = 1$ if observation i is not missing, and $I_i = 0$ if observation i is missing. Assume that $\{I_i, X_i, u_i\}$ are i.i.d.

   i. Show that the OLS estimator can be written as

   $$\hat{\beta} = \left(\sum_{i=1}^{n} I_i X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} I_i X_i Y_i \right) = \beta + \left(\sum_{i=1}^{n} I_i X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} I_i X_i u_i \right).$$

   ii. Suppose that data are missing "completely at random" in the sense that $\Pr(I_i = 1 | X_i, u_i) = p$, where p is a constant. Show that $\hat{\beta}$ is unbiased and consistent.

   iii. Suppose that the probability that the ith observation is missing depends of X_i but not on u_i; that is, $\Pr(I_i = 1 | X_i, u_i) = p(X_i)$. Show that $\hat{\beta}$ is unbiased and consistent.

iv. Suppose that the probability that the ith observation is missing depends on both X_i and u_i; that is, $\Pr(I_i = 1 \mid X_i, u_i) = p(X_i, u_i)$. Is $\hat{\beta}$ unbiased? Is $\hat{\beta}$ consistent? Explain.

c. Suppose that $\beta = 1$ and that X_i and u_i are mutually independent standard normal random variables [so that both X_i and u_i are distributed $N(0, 1)$]. Suppose that $I_i = 1$ when $Y_i \geq 0$ but that $I_i = 0$ when $Y_i < 0$. Is $\hat{\beta}$ unbiased? Is $\hat{\beta}$ consistent? Explain.

19.17 Consider the regression model in matrix form $Y = X\beta + W\gamma + U$, where X and W are matrices of regressors and β and γ are vectors of unknown regression coefficients. Let $\tilde{X} = M_W X$ and $\tilde{Y} = M_W Y$, where $M_W = I - W(W'W)^{-1}W$.

a. Show that the OLS estimators of β and γ can be written as

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'X & X'W \\ W'X & W'W \end{bmatrix}^{-1} \begin{bmatrix} X'Y \\ W'Y \end{bmatrix}$$

b. Show that

$$\begin{bmatrix} X'X & X'W \\ W'X & W'W \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} (X'M_WX)^{-1} & -(X'M_WX)^{-1}X'W(W'W)^{-1} \\ -(W'W)^{-1}W'X(X'M_WX)^{-1} & (W'W)^{-1} + (W'W)^{-1}W'X(X'M_WX)^{-1}X'W(W'W)^{-1} \end{bmatrix}.$$

(*Hint:* Show that the product of the two matrices is equal to the identity matrix.)

c. Show that $\hat{\beta} = (X'M_WX)^{-1}X'M_WY$.

d. The Frisch–Waugh theorem (Appendix 6.2) says that $\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}$. Use the result in (c) to prove the Frisch–Waugh theorem.

19.18 Consider the homoskedastic linear regression model with two regressors, and let $\rho_{X_1, X_2} = \text{corr}(X_1, X_2)$. Show that $\text{corr}(\hat{\beta}_1, \hat{\beta}_2) \to -\rho_{X_1,X_2}$ [Equation (6.21)] as n increases.

APPENDIX

19.1 Summary of Matrix Algebra

This appendix summarizes vectors, matrices, and the elements of matrix algebra used in Chapter 19. The purpose of this appendix is to review some concepts and definitions from a course in linear algebra, not to replace such a course.

Definitions of Vectors and Matrices

A **vector** is a collection of n numbers or elements, collected either in a column (a **column vector**) or in a row (a **row vector**). The n-dimensional column vector b and the n-dimensional row vector c are

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \text{ and } \boldsymbol{c} = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix},$$

where b_1 is the first element of $\boldsymbol{b}$ and, in general, b_i is the ith element of $\boldsymbol{b}$.

Throughout, a boldface denotes a vector or matrix.

A **matrix** is a collection, or an array, of numbers or elements, in which the elements are laid out in columns and rows. The dimension of a matrix is $n \times m$, where n is the number of rows and m is the number of columns. The $n \times m$ matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix},$$

where a_{ij} is the (i, j) element of $\boldsymbol{A}$; that is, a_{ij} is the element that appears in the ith row and j^{th} column. An $n \times m$ matrix consists of n row vectors or, alternatively, of m column vectors.

To distinguish one-dimensional numbers from vectors and matrices, a one-dimensional number is called a **scalar**.

Types of Matrices

Square, symmetric, and diagonal matrices. A matrix is said to be **square** if the number of rows equals the number of columns. A square matrix is said to be **symmetric** if its (i, j) element equals its (j, i) element. A **diagonal** matrix is a square matrix in which all the off-diagonal elements equal 0; that is, if the square matrix $\boldsymbol{A}$ is diagonal, then $a_{ij} = 0$ for $i \neq j$.

Special matrices. An important matrix is the **identity matrix, $\boldsymbol{I}_n$,** which is an $n \times n$ diagonal matrix with 1's on the diagonal. The **null matrix, $\boldsymbol{0}_{n \times m}$,** is the $n \times m$ matrix with all elements equal to 0.

The transpose. The **transpose** of a matrix switches the rows and the columns. That is, the transpose of a matrix turns the $n \times m$ matrix $\boldsymbol{A}$ into the $m \times n$ matrix, which is denoted by $\boldsymbol{A}'$, where the (i, j) element of $\boldsymbol{A}$ becomes the (j, i) element of $\boldsymbol{A}'$; said differently, the transpose of the matrix $\boldsymbol{A}$ turns the rows of $\boldsymbol{A}$ into the columns of $\boldsymbol{A}'$. If a_{ij} is the (i, j) element of $\boldsymbol{A}$, then $\boldsymbol{A}'$ (the transpose of $\boldsymbol{A}$) is

$$\boldsymbol{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}.$$

The transpose of a vector is a special case of the transpose of a matrix. Thus the transpose of a vector turns a column vector into a row vector; that is, if $\boldsymbol{b}$ is an $n \times 1$ column vector, then its transpose is the $1 \times n$ row vector:

$$\boldsymbol{b}' = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix}.$$

The transpose of a row vector is a column vector.

Elements of Matrix Algebra: Addition and Multiplication

Matrix addition. Two matrices A and B that have the same dimensions (for example, that are both $n \times m$) can be added together. The sum of two matrices is the sum of their elements; that is, if $C = A + B$, then $c_{ij} = a_{ij} + b_{ij}$. A special case of matrix addition is vector addition: If a and b are both $n \times 1$ column vectors, then their sum, $c = a + b$, is the element-wise sum; that is, $c_i = a_i + b_i$.

Vector and matrix multiplication. Let a and b be two $n \times 1$ column vectors. Then the product of the transpose of a (which is itself a row vector) and b is $a'b = \sum_{i=1}^{n} a_i b_i$. Applying this definition with $b = a$ yields $a'a = \sum_{i=1}^{n} a_i^2$.

Similarly, the matrices A and B can be multiplied together if they are conformable—that is, if the number of columns of A equals the number of rows of B. Specifically, suppose that A has dimension $n \times m$ and B has dimension $m \times r$. Then the product of A and B is an $n \times r$ matrix, C; that is, $C = AB$, where the (i, j) element of C is $c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$. Said differently, the (i, j) element of AB is the product of multiplying the row vector that is the ith row of A by the column vector that is the jth column of B.

The product of a scalar d with the matrix A has the (i, j) element da_{ij}; that is, each element of A is multiplied by the scalar d.

Some useful properties of matrix addition and multiplication. Let A and B be matrices. Then

    a. $A + B = B + A$;

    b. $(A + B) + C = A + (B + C)$;

    c. $(A + B)' = A' + B'$;

    d. If A is $n \times m$, then $AI_m = A$ and $I_n A = A$;

    e. $A(BC) = (AB)C$;

    f. $(A + B)C = AC + BC$; and

    g. $(AB)' = B'A'$.

In general, matrix multiplication does not commute; that is, in general $AB \neq BA$, although there are some special cases in which matrix multiplication commutes; for example, if A and B are both $n \times n$ diagonal matrices, then $AB = BA$.

Matrix Inverse, Matrix Square Roots, and Related Topics

The matrix inverse. Let A be a square matrix. Assuming that it exists, the **inverse** of the matrix A is defined as the matrix for which $A^{-1}A = I_n$. If, in, fact the inverse matrix A^{-1} exists, then A is said to be **invertible** or **nonsingular**. If both A and B are invertible, then $(AB)^{-1} = B^{-1}A^{-1}$.

Positive definite and positive semidefinite matrices. Let V be an $n \times n$ square matrix. Then V is **positive definite** if $c'Vc > 0$ for all nonzero $n \times 1$ vectors c. Similarly, V is **positive semidefinite** if $c'Vc \geq 0$ for all nonzero $n \times 1$ vectors c. If V is positive definite, then it is invertible.

Linear independence. The $n \times 1$ vectors a_1 and a_2 are **linearly independent** if there do not exist nonzero scalars c_1 and c_2 such that $c_1 a_1 + c_2 a_2 = \mathbf{0}_{n \times 1}$. More generally, the set of k vectors $a_1, a_2, \ldots, a_k$ is linearly independent if there do not exist nonzero scalars $c_1, c_2, \ldots, c_k$ such that $c_1 a_1 + c_2 a_2 + \cdots + c_k a_k = \mathbf{0}_{n \times 1}$.

The rank of a matrix. The **rank** of the $n \times m$ matrix A is the number of linearly independent columns of A. The rank of A is denoted rank(A). If the rank of A equals the number of columns of A, then A is said to have full column rank. If the $n \times m$ matrix A has full column rank, then there does not exist a nonzero $m \times 1$ vector c such that $Ac = \mathbf{0}_{n \times 1}$. If A is $n \times n$ with rank(A) $= n$, then A is nonsingular. If the $n \times m$ matrix A has full column rank, then $A'A$ is nonsingular.

The trace of a matrix. The **trace** of the $n \times n$ (square) matrix A is the sum of the diagonal elements; that is, trace(A) $= \sum_{i=1}^{n} a_{ii}$. For $n \times n$ matrices A and B and $n \times 1$ vector c, the trace satisfies these properties: trace(A) $=$ trace(A'), trace($A + B$) $=$ trace(A) $+$ trace(B), trace(AB) $=$ trace(BA), trace(BAB^{-1}) $=$ trace(A), and $c'Bc =$ trace(Bcc').

The matrix square root. Let V be an $n \times n$ square symmetric positive definite matrix. The matrix square root of V is defined to be an $n \times n$ matrix F such that $F'F = V$. The matrix square root of a positive definite matrix will always exist, but it is not unique. The matrix square root has the property that $FV^{-1}F' = I_n$. In addition, the matrix square root of a positive definite matrix is invertible, so $F'^{-1}VF^{-1} = I_n$.

Eigenvalues and eigenvectors. Let A be an $n \times n$ matrix. If the $n \times 1$ vector q and the scalar λ satisfy $Aq = \lambda q$, where $q'q = 1$, then λ is an **eigenvalue** of A, and q is the **eigenvector** of A associated with that eigenvalue. An $n \times n$ matrix has n eigenvalues, which need not take on distinct values, and n eigenvectors.

    If V is an $n \times n$ symmetric positive definite matrix, then the eigenvalues of V are positive real numbers, and the eigenvectors of V are real. Also, V can be written in terms of its eigenvalues and eigenvectors as $V = Q\Lambda Q'$, where Λ is a diagonal $n \times n$ matrix with diagonal elements that equal the eigenvalues of V and Q is an $n \times n$ matrix consisting of the eigenvectors of V, arranged so that the ith column of Q is the eigenvector corresponding to the eigenvalue λ_i, which is the ith diagonal element of Λ. The eigenvectors are orthonormal, so $Q'Q = I_n$. The trace of V equals the sum of its eigenvectors: trace(V) $=$ trace($Q\Lambda Q'$) $=$ trace($\Lambda Q'Q$) $=$ trace(Λ) $= \sum_{i=1}^{n} \lambda_i$.

Idempotent matrices. A matrix C is idempotent if C is square and $CC = C$. If C is an $n \times n$ idempotent matrix that is also symmetric, then C is positive semidefinite, and C has r eigenvalues that equal 1 and $n - r$ eigenvalues that equal 0, where $r =$ rank(C) (Exercise 19.10).

19.2 Multivariate Distributions

This appendix collects various definitions and facts about distributions of vectors of random variables. We start by defining the mean and covariance matrix of the n-dimensional random variable V. Next we present the multivariate normal distribution. Finally, we summarize some facts about the distributions of linear and quadratic functions of jointly normally distributed random variables.

The Mean Vector and Covariance Matrix

The first and second moments of an $m \times 1$ vector of random variables, $V = (V_1 \quad V_2 \quad \cdots \quad V_m)'$, are summarized by its mean vector and covariance matrix.

Because V is a vector, the vector of its means—that is, its **mean vector**—is $E(V) = \boldsymbol{\mu}_V$. The ith element of the mean vector is the mean of the ith element of V.

The **covariance matrix** of V is the matrix consisting of the variance $\text{var}(V_i), i = 1, \ldots, m$, along the diagonal and the (i, j) off-diagonal elements $\text{cov}(V_i, V_j)$. In matrix form, the covariance matrix V is

$$\boldsymbol{\Sigma}_V = E[(V - \boldsymbol{\mu}_V)(V - \boldsymbol{\mu}_V)'] = \begin{bmatrix} \text{var}(V_1) & \cdots & \text{cov}(V_1, V_m) \\ \vdots & \ddots & \vdots \\ \text{cov}(V_m, V_1) & \cdots & \text{var}(V_m) \end{bmatrix}. \quad (19.72)$$

The Multivariate Normal Distribution

The $m \times 1$ vector random variable V has a multivariate normal distribution with mean vector $\boldsymbol{\mu}_V$ and covariance matrix $\boldsymbol{\Sigma}_V$ if it has the joint probability density function

$$f(V) = \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma}_V)}} \exp\left[-\frac{1}{2}(V - \boldsymbol{\mu}_V)' \boldsymbol{\Sigma}_V^{-1}(V - \boldsymbol{\mu}_V)\right], \quad (19.73)$$

where $\det(\boldsymbol{\Sigma}_V)$ is the determinant of the matrix $\boldsymbol{\Sigma}_V$. The multivariate normal distribution is denoted $N(\boldsymbol{\mu}_V, \boldsymbol{\Sigma}_V)$.

An important fact about the multivariate normal distribution is that if two jointly normally distributed random variables are uncorrelated (or, equivalently, have a block-diagonal covariance matrix), then they are independently distributed. That is, let V_1 and V_2 be jointly normally distributed random variables with respective dimensions $m_1 \times 1$ and $m_2 \times 1$. Then if $\text{cov}(V_1, V_2) = E[(V_1 - \boldsymbol{\mu}_{V_1})(V_2 - \boldsymbol{\mu}_{V_2})'] = \mathbf{0}_{m_1 \times m_2}$, V_1 and V_2 are independent.

If $\{V_i\}$ are i.i.d. $N(0, \sigma_v^2)$, then $\boldsymbol{\Sigma}_V = \sigma_v^2 \boldsymbol{I}_m$, and the multivariate normal distribution simplifies to the product of m univariate normal densities.

Distributions of Linear Combinations and Quadratic Forms of Normal Random Variables

Linear combinations of multivariate normal random variables are themselves normally distributed, and certain quadratic forms of multivariate normal random variables have a chi-squared

distribution. Let V be an $m \times 1$ random variable distributed $N(\boldsymbol{\mu}_V, \boldsymbol{\Sigma}_V)$, let A and B be non-random $a \times m$ and $b \times m$ matrices, and let d be a nonrandom $a \times 1$ vector. Then

$$d + AV \text{ is distributed } N(d + A\boldsymbol{\mu}_V, A\boldsymbol{\Sigma}_V A'), \text{ and} \tag{19.74}$$

$$\text{cov}(AV, BV) = A\boldsymbol{\Sigma}_V B'; \tag{19.75}$$

$$\text{if } A\boldsymbol{\Sigma}_V B' = \mathbf{0}_{a \times b}, \text{ then } AV \text{ and } BV \text{ are independently distributed; and} \tag{19.76}$$

$$(V - \boldsymbol{\mu}_V)' \boldsymbol{\Sigma}_V^{-1} (V - \boldsymbol{\mu}_V) \text{ is distributed } \chi_m^2. \tag{19.77}$$

Let U be an m-dimensional multivariate standard normal random variable with distribution $N(\mathbf{0}, I_m)$. If C is symmetric and idempotent, then

$$U'CU \text{ has a } \chi_r^2 \text{ distribution, where } r = \text{rank}(C). \tag{19.78}$$

Equation (19.78) is proven as Exercise 19.11.

19.3 Derivation of the Asymptotic Distribution of $\hat{\beta}$

This appendix provides the derivation of the asymptotic normal distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ given in Equation (19.12). An implication of this result is that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

First consider the "denominator" matrix $X'X/n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$ in Equation (19.15). The (j, l) element of this matrix is $\frac{1}{n}\sum_{i=1}^{n} X_{ji} X_{li}$. By the second assumption in Key Concept 19.1, X_i is i.i.d., so $X_{ji} X_{li}$ is i.i.d. By the third assumption in Key Concept 19.1, each element of X_i has four moments, so, by the Cauchy–Schwarz inequality (Appendix 18.2), $X_{ji} X_{li}$ has two moments. Because $X_{ji} X_{li}$ is i.i.d. with two moments, $\frac{1}{n}\sum_{i=1}^{n} X_{ji} X_{li}$ obeys the law of large numbers, so $\frac{1}{n}\sum_{i=1}^{n} X_{ji} X_{li} \xrightarrow{p} E(X_{ji} X_{li})$. This is true for all the elements of $X'X/n$, so $X'X/n \xrightarrow{p} E(X_i X_i') = Q_X$.

Next consider the "numerator" matrix in Equation (19.15), $X'U/\sqrt{n} = \sqrt{\frac{1}{n}}\sum_{i=1}^{n} V_i$, where $V_i = X_i u_i$. By the first assumption in Key Concept 19.1 and the law of iterated expectations, $E(V_i) = E[X_i E(u_i | X_i)] = \mathbf{0}_{k+1}$. By the second least squares assumption, V_i is i.i.d. Let c be a finite $k + 1$ dimensional vector. By the Cauchy–Schwarz inequality, $E[(c'V_i)^2] = E[(c'X_i u_i)^2] = E[(c'X_i)^2(u_i)^2] \leq \sqrt{E[(c'X_i)^4] E(u_i^4)}$, which is finite by the third least squares assumption. This is true for every such vector c, so $E(V_i V_i') = \boldsymbol{\Sigma}_V$ is finite and, we assume, positive definite. Thus the multivariate central limit theorem of Key Concept 19.2 applies to $\sqrt{\frac{1}{n}}\sum_{i=1}^{n} V_i = \frac{1}{\sqrt{n}} X'U$; that is,

$$\frac{1}{\sqrt{n}} X'U \xrightarrow{d} N(\mathbf{0}_{k+1}, \boldsymbol{\Sigma}_V). \tag{19.79}$$

The result in Equation (19.12) follows from Equations (19.15) and (19.79), the consistency of $X'X/n$, the fourth least squares assumption (which ensures that $(X'X)^{-1}$ exists), and Slutsky's theorem.

19.4 Derivations of Exact Distributions of OLS Test Statistics with Normal Errors

This appendix presents the proofs of the distributions under the null hypothesis of the homoskedasticity-only t-statistic in Equation (19.35) and the homoskedasticity-only F-statistic in Equation (19.37), assuming that all six assumptions in Key Concept 19.1 hold.

Proof of Equation (19.35)

If (i) Z has a standard normal distribution, (ii) W has a χ_m^2 distribution, and (iii) Z and W are independently distributed, then the random variable $Z/\sqrt{W/m}$ has the t distribution with m degrees of freedom (Appendix 18.1). To put $\tilde{t}$ in this form, notice that $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = (s_{\hat{u}}^2/\sigma_u^2)\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}}$. Then rewrite Equation (19.34) as

$$\tilde{t} = \frac{(\hat{\beta}_j - \beta_{j,0})/\sqrt{(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}})_{jj}}}{\sqrt{W/(n-k-1)}}, \tag{19.80}$$

where $W = (n-k-1)(s_{\hat{u}}^2/\sigma_u^2)$, and let $Z = (\hat{\beta}_j - \beta_{j,0})/\sqrt{(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}})_{jj}}$ and $m = n-k-1$. With these definitions, $\tilde{t} = Z/\sqrt{W/m}$. Thus, to prove the result in Equation (19.35), we must show (i) through (iii) for these definitions of Z, W, and m.

 i. An implication of Equation (19.30) is that, under the null hypothesis, $Z = (\hat{\beta}_j - \beta_{j,0})/\sqrt{(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}})_{jj}}$ has an exact standard normal distribution, which shows (i).
 ii. From Equation (19.31), W is distributed as χ_{n-k-1}^2, which shows (ii).
 iii. To show (iii), it must be shown that $\hat{\beta}_j$ and $s_{\hat{u}}^2$ are independently distributed.

From Equations (19.14) and (19.29), $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$ and $s_{\hat{u}}^2 = (\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{U})'(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{U})/(n-k-1)$. Thus $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and $s_{\hat{u}}^2$ are independent if $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$ and $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{U}$ are independent. Both $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$ and $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{U}$ are linear combinations of $\boldsymbol{U}$, which has an $N(\boldsymbol{0}_{n\times 1}, \sigma_u^2\boldsymbol{I}_n)$ distribution, conditional on $\boldsymbol{X}$. But because $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{0}_{n\times(k+1)}$ [Equation (19.26)], it follows that $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$ and $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{U}$ are independently distributed [Equation (19.76)]. Consequently, under all six assumptions in Key Concept 19.1,

$$\hat{\boldsymbol{\beta}} \text{ and } s_{\hat{u}}^2 \text{ are independently distributed,} \tag{19.81}$$

which shows (iii) and thus proves Equation (19.35).

Proof of Equation (19.37)

The F_{n_1, n_2} distribution is the distribution of $(W_1/n_1)/(W_2/n_2)$, where (i) W_1 is distributed $\chi_{n_1}^2$; (ii) W_2 is distributed $\chi_{n_2}^2$; and (iii) W_1 and W_2 are independently distributed (Appendix 18.1). To express $\tilde{F}$ in this form, let $W_1 = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\sigma_u^2]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$ and $W_2 = (n-k-1)s_{\hat{u}}^2/\sigma_u^2$. Substitution of these definitions into Equation (19.36) shows that $\tilde{F} = (W_1/q)/[W_2/(n-k-1)]$. Thus, by the definition of the F distribution, $\tilde{F}$ has an $F_{q, n-k-1}$ distribution if (i) through (iii) hold with $n_1 = q$ and $n_2 = n-k-1$.

   i. Under the null hypothesis, $R\hat{\beta} - r = R(\hat{\beta} - \beta)$. Because $\hat{\beta}$ has the conditional normal distribution in Equation (19.30) and because R is a nonrandom matrix, $R(\hat{\beta} - \beta)$ is distributed $N(\mathbf{0}_{q\times1}, R(X'X)^{-1}R'\sigma_u^2)$, conditional on X. Thus, by Equation (19.77) in Appendix 19.2, $(R\hat{\beta} - r)'[R(X'X)R'\sigma_u^2]^{-1}(R\hat{\beta} - r)$ is distributed χ_q^2, proving (i).

   ii. Requirement (ii) is shown in Equation (19.31).

   iii. It has already been shown that $\hat{\beta} - \beta$ and $s_{\hat{u}}^2$ are independently distributed [Equation (19.81)]. It follows that $R\hat{\beta} - r$ and $s_{\hat{u}}^2$ are independently distributed, which in turn implies that W_1 and W_2 are independently distributed, proving (iii) and completing the proof.

APPENDIX

19.5 Proof of the Gauss–Markov Theorem for Multiple Regression

This appendix proves the Gauss–Markov theorem (Key Concept 19.3) for the multiple regression model. Let $\tilde{\beta}$ be a linear conditionally unbiased estimator of β so that $\tilde{\beta} = A'Y$ and $E(\tilde{\beta} \mid X) = \beta$, where A is an $n \times (k + 1)$ matrix that can depend on X and nonrandom constants. We show that $\text{var}(c'\hat{\beta}) \le \text{var}(c'\tilde{\beta})$ for all $k + 1$ dimensional vectors c, where the inequality holds with equality only if $\tilde{\beta} = \hat{\beta}$.

Because $\tilde{\beta}$ is linear, it can be written as $\tilde{\beta} = A'Y = A'(X\beta + U) = (A'X)\beta + A'U$. By the first Gauss–Markov condition, $E(U|X) = \mathbf{0}_{n\times1}$, so $E(\tilde{\beta} \mid X) = (A'X)\beta$, but because $\tilde{\beta}$ is conditionally unbiased, $E(\tilde{\beta} \mid X) = \beta = (A'X)\beta$, which implies that $A'X = I_{k+1}$. Thus $\tilde{\beta} = \beta + A'U$, so $\text{var}(\tilde{\beta} \mid X) = \text{var}(A'U|X) = E(A'UU'A|X) = A'E(UU'|X)A = \sigma_u^2 A'A$, where the third equality follows because A can depend on X but not U and the final equality follows from the second Gauss–Markov condition. That is, if $\tilde{\beta}$ is linear and unbiased, then under the Gauss–Markov conditions,

$$A'X = I_{k+1} \text{ and } \text{var}(\tilde{\beta}|X) = \sigma_u^2 A'A. \tag{19.82}$$

The results in Equation (19.82) also apply to $\hat{\beta}$ with $A = \hat{A} = X(X'X)^{-1}$, where $(X'X)^{-1}$ exists by the third Gauss–Markov condition.

Now let $A = \hat{A} + D$, so that D is the difference between the matrices A and $\hat{A}$. Note that $\hat{A}'A = (X'X)^{-1}X'A = (X'X)^{-1}$ [by Equation (19.82)] and $\hat{A}'\hat{A} = (X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}$, so $\hat{A}'D = \hat{A}'(A - \hat{A}) = \hat{A}'A - \hat{A}'\hat{A} = \mathbf{0}_{(k+1)\times(k+1)}$. Substituting $A = \hat{A} + D$ into the formula for the conditional variance in Equation (19.82) yields

$$\begin{aligned}
\text{var}(\tilde{\beta}|X) &= \sigma_u^2(\hat{A} + D)'(\hat{A} + D) \\
&= \sigma_u^2[\hat{A}'\hat{A} + \hat{A}'D + D'\hat{A} + D'D] \\
&= \sigma_u^2(X'X)^{-1} + \sigma_u^2 D'D, \tag{19.83}
\end{aligned}$$

where the final equality uses the facts $\hat{A}'\hat{A} = (X'X)^{-1}$ and $\hat{A}'D = \mathbf{0}_{(k+1)\times(k+1)}$.

Because $\text{var}(\hat{\boldsymbol{\beta}}|X) = \sigma_u^2(X'X)^{-1}$, Equations (19.82) and (19.83) imply that $\text{var}(\tilde{\boldsymbol{\beta}}|X) - \text{var}(\hat{\boldsymbol{\beta}}|X) = \sigma_u^2 D'D$. The difference between the variances of the two estimators of the linear combination $c'\boldsymbol{\beta}$ thus is

$$\text{var}(c'\tilde{\boldsymbol{\beta}}|X) - \text{var}(c'\hat{\boldsymbol{\beta}}|X) = \sigma_u^2 c'D'Dc \geq 0. \tag{19.84}$$

The inequality in Equation (19.84) holds for all linear combinations $c'\boldsymbol{\beta}$, and the inequality holds with equality for all nonzero c only if $D = \mathbf{0}_{n \times (k+1)}$—that is, if $A = \hat{A}$ or, equivalently, $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$. Thus $c'\hat{\boldsymbol{\beta}}$ has the smallest variance of all linear conditionally unbiased estimators of $c'\boldsymbol{\beta}$; that is, the OLS estimator is BLUE.

APPENDIX

19.6 Proof of Selected Results for IV and GMM Estimation

The Efficiency of TSLS Under Homoskedasticity [Proof of Equation (19.62)]

When the errors u_i are homoskedastic, the difference between $\boldsymbol{\Sigma}_A^{IV}$ [Equation (19.61)] and $\boldsymbol{\Sigma}^{TSLS}$ [Equation (19.55)] is given by

$$\boldsymbol{\Sigma}_A^{IV} - \boldsymbol{\Sigma}^{TSLS} = (Q_{XZ}AQ_{ZX})^{-1}Q_{XZ}AQ_{ZZ}AQ_{ZX}(Q_{XZ}AQ_{ZX})^{-1}\sigma_u^2 - (Q_{XZ}Q_{ZZ}^{-1}Q_{ZX})^{-1}\sigma_u^2$$
$$= (Q_{XZ}AQ_{ZX})^{-1}Q_{XZ}A[Q_{ZZ} - Q_{ZX}(Q_{XZ}Q_{ZZ}^{-1}Q_{ZX})^{-1}Q_{XZ}]AQ_{ZX}(Q_{XZ}AQ_{ZX})^{-1}\sigma_u^2, \tag{19.85}$$

where the second term within the brackets in the second equality follows from $(Q_{XZ}AQ_{ZX})^{-1}Q_{XZ}AQ_{ZX} = I_{(k+r+1)}$. Let F be the matrix square root of Q_{ZZ}, so $Q_{ZZ} = F'F$ and $Q_{ZZ}^{-1} = F^{-1}F'^{-1}$. [The latter equality follows from noting that $(F'F)^{-1} = F^{-1}F'^{-1}$ and $F'^{-1} = F^{-1'}$.] Then the final expression in Equation (19.85) can be rewritten to yield

$$\boldsymbol{\Sigma}_A^{IV} - \boldsymbol{\Sigma}^{TSLS} = (Q_{XZ}AQ_{ZX})^{-1}Q_{XZ}AF'[I - F^{-1'}Q_{ZX}(Q_{XZ}F^{-1}F^{-1'}Q_{ZX})^{-1}Q_{XZ}F^{-1}]$$
$$\times FAQ_{ZX}(Q_{XZ}AQ_{ZX})^{-1}\sigma_u^2, \tag{19.86}$$

where the second expression within the brackets uses $F'F^{-1'} = I$. Thus

$$c'(\boldsymbol{\Sigma}_A^{IV} - \boldsymbol{\Sigma}^{TSLS})c = d'[I - D(D'D)^{-1}D']d\sigma_u^2, \tag{19.87}$$

where $d = FAQ_{ZX}(Q_{XZ}AQ_{ZX})^{-1}c$ and $D = F^{-1'}Q_{ZX}$. Now $I - D(D'D)^{-1}D'$ is a symmetric idempotent matrix (Exercise 19.5). As a result, $I - D(D'D)^{-1}D'$ has eigenvalues that are either 0 or 1, and $d'[I - D(D'D)^{-1}D']d \geq 0$ (Exercise 19.10). Thus $c'(\boldsymbol{\Sigma}_A^{IV} - \boldsymbol{\Sigma}^{TSLS})c \geq 0$, proving that TSLS is efficient under homoskedasticity.

Asymptotic Distribution of the *J*-Statistic Under Homoskedasticity

The *J*-statistic is defined in Equation (19.63). First note that

$$
\begin{aligned}
\hat{U} &= Y - X\hat{\beta}^{TSLS} \\
&= Y - X(X'P_ZX)^{-1}X'P_ZY \\
&= (X\beta + U) - X(X'P_ZX)^{-1}X'P_Z(X\beta + U) \\
&= U - X(X'P_ZX)^{-1}X'P_ZU \\
&= [I - X(X'P_ZX)^{-1}X'P_Z]U.
\end{aligned} \tag{19.88}
$$

Thus

$$
\begin{aligned}
\hat{U}P_Z\hat{U} &= U'[I - P_ZX(X'P_ZX)^{-1}X']P_Z[I - X(X'P_ZX)^{-1}X'P_Z]U \\
&= U'[P_Z - P_ZX(X'P_ZX)^{-1}X'P_Z]U,
\end{aligned} \tag{19.89}
$$

where the second equality follows by simplifying the preceding expression. Because $Z'Z$ is symmetric and positive definite, it can be written in terms of its matrix square root, $Z'Z = (Z'Z)^{1/2\prime}(Z'Z)^{1/2}$, and this matrix square root is invertible, so $(Z'Z)^{-1} = (Z'Z)^{-1/2}(Z'Z)^{-1/2\prime}$, where $(Z'Z)^{-1/2} = [(Z'Z)^{1/2}]^{-1}$. Thus P_Z can be written as $P_Z = Z(Z'Z)^{-1}Z' = BB'$ where $B = Z(Z'Z)^{-1/2}$. Substituting this expression for P_Z into the final expression in Equation (19.89) yields

$$
\begin{aligned}
\hat{U}'P_Z\hat{U} &= U'[BB' - BB'X(X'BB'X)^{-1}X'BB']U \\
&= U'B[I - B'X(X'BB'X)^{-1}X'B]B'U \\
&= U'BM_{B'X}B'U,
\end{aligned} \tag{19.90}
$$

where $M_{B'X} = I - B'X(X'BB'X)^{-1}X'B$ is a symmetric idempotent matrix.

The asymptotic null distribution of $\hat{U}'P_Z\hat{U}$ is found by computing the limits in probability and in distribution of the various terms in the final expression in Equation (19.90) under the null hypothesis. Under the null hypothesis that $E(Z_iu_i) = 0$, $Z'U/\sqrt{n}$ has mean 0, and the central limit theorem applies, so $Z'U/\sqrt{n} \xrightarrow{d} N(0, Q_{ZZ}\sigma_u^2)$. In addition, $Z'Z/n \xrightarrow{p} Q_{ZZ}$ and $X'Z/n \xrightarrow{p} Q_{XZ}$. Thus $B'U = (Z'Z)^{-1/2}Z'U = (Z'Z/n)^{-1/2\prime}(Z'U/\sqrt{n}) \xrightarrow{d} \sigma_u z$, where z is distributed $N(\mathbf{0}_{m+r+1}, I_{m+r+1})$. In addition, $B'X/\sqrt{n} = (Z'Z/n)^{-1/2\prime}(Z'X/n) \xrightarrow{p} Q_{ZZ}^{-1/2\prime}Q_{ZX}$, so $M_{B'X} \xrightarrow{p} I - Q_{ZZ}^{-1/2\prime}Q_{ZX}(Q_{XZ}Q_{ZZ}^{-1/2\prime}Q_{ZZ}^{-1/2\prime}Q_{ZX})^{-1}Q_{XZ}Q_{ZZ}^{-1/2\prime} = M_{Q_{ZZ}^{-1/2}Q_{ZX}}$. Thus

$$
\hat{U}'P_Z\hat{U} \xrightarrow{d} (z'M_{Q_{XZ}Q_{ZZ}^{-1/2}}z)\sigma_u^2. \tag{19.91}
$$

Under the null hypothesis, the TSLS estimator is consistent, and the coefficients in the regression of $\hat{U}$ on Z converge in probability to 0 [an implication of Equation (19.91)], so the denominator in the definition of the *J*-statistic is a consistent estimator of σ_u^2:

$$
\hat{U}'M_Z\hat{U}/(n - m - r - 1) \xrightarrow{p} \sigma_u^2. \tag{19.92}
$$

From the definition of the *J*-statistic and Equations (19.91) and (19.92), it follows that

$$
J = \frac{\hat{U}'P_Z\hat{U}}{\hat{U}'M_Z\hat{U}/(n - m - r - 1)} \xrightarrow{d} z'M_{Q_{ZZ}^{-1/2}Q_{XZ}}z. \tag{19.93}
$$

Because z is a standard normal random vector and $M_{Q_{ZZ}^{-1/2}Q_{ZX}}$ is a symmetric idempotent matrix, J is distributed as a chi-squared random variable with degrees of freedom that equals the rank of $M_{Q_{ZZ}^{-1/2}Q_{ZX}}$ [Equation (19.78)]. Because $Q_{ZZ}^{-1/2}Q_{ZX}$ is $(m + r + 1) \times (k + r + 1)$ and $m > k$, the rank of $M_{Q_{ZZ}^{-1/2}Q_{ZX}}$ is $m - k$ [Exercise 19.5]. Thus $J \xrightarrow{d} \chi^2_{m-k}$, which is the result stated in Equation (19.64).

The Efficiency of the Efficient GMM Estimator

The infeasible efficient GMM estimator, $\widetilde{\boldsymbol{\beta}}^{Eff.GMM}$, is defined in Equation (19.66). The proof that $\widetilde{\boldsymbol{\beta}}^{Eff.GMM}$ is efficient entails showing that $\boldsymbol{c}'(\Sigma_A^{IV} - \Sigma^{Eff.GMM})\boldsymbol{c} \geq 0$ for all vectors $\boldsymbol{c}$. The proof closely parallels the proof of the efficiency of the TSLS estimator in the first section of this appendix, with the sole modification that $\boldsymbol{H}^{-1}$ replaces $\boldsymbol{Q}_{ZZ}\sigma_u^2$ in Equation (19.85) and subsequently.

Distribution of the GMM J-Statistic

The GMM J-statistic is given in Equation (19.70). The proof that, under the null hypothesis, $J^{GMM} \xrightarrow{d} \chi^2_{m-k}$ closely parallels the corresponding proof for the TSLS J-statistic under homoskedasticity.

19.7 Regression with Many Predictors: MSPE, Ridge Regression, and Principal Components Analysis

This appendix presents the derivations for various results used in Chapter 14 that rely on matrix calculations.

The MSPE for Linear Regression Estimated by OLS

We first derive Equation (14.4), the mean squared prediction error (MSPE) of the OLS estimator under homoskedasticity.

Let the $k \times 1$ vector $\boldsymbol{X}^{oos}$ denote the values of the X's for the out-of-sample observation ("oos") to be predicted. With this notation, the MSPE in Equation (14.3), written using matrix notation, is

$$\text{MSPE} = \sigma_u^2 + E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}^{oos}]^2, \tag{19.94}$$

where $\hat{\boldsymbol{\beta}}$ denotes any estimator of $\boldsymbol{\beta}$, not just the OLS estimator.

Under the least squares assumptions for prediction, the out-of-sample observation is assumed to be an i.i.d. draw from the same population as the estimation sample. Under this assumption, the MSPE in Equation (19.94) can be written

$$\text{MSPE} = \sigma_u^2 + \text{trace}\{E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})']\boldsymbol{Q}_X\}, \tag{19.95}$$

where $\boldsymbol{Q}_X = E(\boldsymbol{X}'\boldsymbol{X})$. Equation (19.95) follows from Equation (19.94) by writing, $E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}^{oos}]^2 = E[\boldsymbol{X}^{oos'}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}^{oos}] = \text{trace}E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \boldsymbol{X}^{oos}\boldsymbol{X}^{oos'}] = \text{trace}E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})']\boldsymbol{Q}_X$, where the second inequality uses the property of the trace that $\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a} = \text{trace}(\boldsymbol{B}\boldsymbol{a}\boldsymbol{a}')$ for $n \times n$ matrix $\boldsymbol{B}$ and $n \times 1$ vector $\boldsymbol{a}$ and where the final equality uses the assumptions that the out-of-sample observation is independent of the estimation observations and that it is drawn from the same distribution, so that $E(\boldsymbol{X}^{oos}\boldsymbol{X}^{oos'}) = \boldsymbol{Q}_X$.

The MSPE for OLS obtains by substituting the expression for OLS in Equation (19.14) into Equation (19.95) and simplifying. First note that, under the assumption of homoskedasticity, for the OLS estimator,

$$
\begin{aligned}
E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] \\
&= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] \\
&= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}]\sigma_u^2 = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}]\sigma_u^2,
\end{aligned}
$$

where the first equality uses Equation (19.14); the second equality uses the law of iterated expectations; the third equality uses the assumption of homoskedasticity, so $E(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}) = \sigma_u^2\boldsymbol{I}_n$; and the final equality simplifies. Substitution of $E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}]\sigma_u^2$ into Equation (19.95) and multiplying and dividing the second term by $1/n$ yields

$$
\text{MSPE}_{\text{OLS}} = \sigma_u^2 + \frac{1}{n}\text{trace}\left\{E\left[\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\right]\boldsymbol{Q}_X\right\}\sigma_u^2. \tag{19.96}
$$

Equation (19.96) is the MSPE for a prediction made using the OLS estimator under the least squares assumptions for prediction with homoskedastic errors.

Equation (14.4) is an approximation to Equation (19.96) when n is large relative to k. In that case, $\boldsymbol{X}'\boldsymbol{X}/n \cong \boldsymbol{Q}_X$ (specifically, for fixed k, $\boldsymbol{X}'\boldsymbol{X}/n \overset{p}{\longrightarrow} \boldsymbol{Q}_X$) so $\text{trace}\{E[(\boldsymbol{X}'\boldsymbol{X}/n)^{-1}]\boldsymbol{Q}_X\} \cong \text{trace}\{\boldsymbol{Q}_X^{-1}\boldsymbol{Q}_X\} = \text{trace}\{\boldsymbol{I}_k\} = k$. Substitution of this final expression into Equation (19.96) and collecting terms yields Equation (14.4):

$$
\text{MSPE}_{\text{OLS}} \cong \left(1 + \frac{k}{n}\right)\sigma_u^2. \tag{19.97}
$$

Connection to the final prediction error (FPE). Equation (19.97) is used in the derivation of the final prediction error (FPE) for time series forecasting given in Equation (15.21) (with a change in notation so that n is replaced by T and k is replaced by $p + 1$). The key difference between the cross-section and time-series cases is the relation of the out-of-sample observation to the in-sample observations. In the deriviation here, the in- and out-of-sample observations are independent. If the values of the predictors in the time series application are independent of the data used to estimate the coefficients, then the derivation here applies directly. Typically this will not be the case, however, because the final observations in the sample (the ones used to make the out-of-sample forecast) are correlated with the in-sample observations. If the sample size is large, however, then the dependence between the estimated regression coefficients and the out-of-sample predictors is small, so Equation (19.97) still holds as an approximation when the sample size is large relative to the number of regressors.

Ridge Regression

Equation (14.8) provides an expression for the ridge regression estimator with a single regressor. This appendix derives an expression for the case of multiple regressors.

The ridge regression estimator minimizes the penalized sum of squared residuals in Equation (14.7), written here using matrix notation:

$$S^{Ridge}(b; \lambda_{Ridge}) = (Y - Xb)'(Y - Xb) + \lambda_{Ridge}b'b. \tag{19.98}$$

Taking the derivative of the right-hand side of Equation (19.98) and setting it to 0 yields the system solved by the ridge regression estimator $\hat{\boldsymbol{\beta}}^{Ridge}$, $-2X'(Y - X\hat{\boldsymbol{\beta}}^{Ridge}) + 2\lambda_{Ridge}\hat{\boldsymbol{\beta}}^{Ridge} = \mathbf{0}$ [cf. Equations (19.9) and (19.10) for OLS)]. Solving this system yields the formula for the ridge regression estimator,

$$\hat{\boldsymbol{\beta}}^{Ridge} = (X'X + \lambda_{Ridge}I_k)^{-1}X'Y. \tag{19.99}$$

We note two implications of this formula that are discussed in Sections 14.3 and 14.4, respectively.

First, if the regressors are uncorrelated in the estimation sample, the ridge regression estimator can be written as the OLS estimator, shrunk toward 0 by a factor that depends on the data, that is, $\hat{\beta}_j^{Ridge} = (1 + \lambda^{Ridge}/\sum_{i=1}^n X_{ji}^2)^{-1}\hat{\beta}_j$, which is Equation (14.8). Moreover, if in addition the regressors are standardized using the sample standard deviation, as they are in the empirical work in Chapter 14, that shrinkage factor simplifies to $[1 + \lambda^{Ridge}/(n - 1)]^{-1}$. To show these results, note that if the regressors are uncorrelated, then $X'X$ is diagonal, so that $X'X + \lambda_{Ridge}I_k$ is diagonal with j^{th} diagonal element $\sum_{i=1}^n X_{ji}^2 + \lambda^{Ridge}$. Then Equation (19.99) simplifies, so that the ridge estimator of the jth coefficient β_j is $\hat{\beta}_j^{Ridge} = (\sum_{i=1}^n X_{ji}^2 + \lambda^{Ridge})^{-1}\sum_{i=1}^n X_{ji}Y_i = (1 + \lambda^{Ridge}/\sum_{i=1}^n X_{ji}^2)^{-1}(\sum_{i=1}^n X_{ji}^2)^{-1}\sum_{i=1}^n X_{ji}Y_i = (1 + \lambda^{Ridge}/\sum_{i=1}^n X_{ji}^2)^{-1}\hat{\beta}_j$, where $\hat{\beta}_j$ is the OLS estimator for these uncorrelated regressors. Thus, with uncorrelated regressors, the ridge regression estimator shrinks the OLS estimator toward 0 by the factor $(1 + \lambda^{Ridge}/\sum_{i=1}^n X_{ji}^2)^{-1}$. If in addition the regressors are standardized using the sample standard deviation, then $\sum_{i=1}^n X_{ji}^2 = n - 1$, in which case $\hat{\boldsymbol{\beta}}^{Ridge} = [1 + \lambda^{Ridge}/(n - 1)]^{-1}\hat{\boldsymbol{\beta}}$.

Second, as is discussed in Section 14.4, predictions made using the ridge regression estimator, in general, change if different linear combinations of the regressors are used as predictors. Specifically, if X denotes the matrix of predictors, then the ridge predictions made using X and using XA differ, where A is a nonsingular $k \times k$ matrix. This is an important difference between ridge and OLS because OLS yields the same predictions whether X or XA is used.

To show this result, consider the ridge regression estimator computed using XA, and denote that estimator by $\hat{\boldsymbol{\beta}}_A^{Ridge}$. In this notation, the ridge regression estimator computed using X without the linear transformation is $\hat{\boldsymbol{\beta}}_I^{Ridge}$. The same linear transformation must be applied to the out-of-sample and in-sample predictors, so the transformed out-of-sample observation is $A'X^{OOS}$. Thus the out-of-sample predicted value using $\hat{\boldsymbol{\beta}}_A^{Ridge}$ is $\hat{Y}_A^{OOS} = (A'X^{OOS})'\hat{\boldsymbol{\beta}}_A^{Ridge} = X^{OOS'}A\hat{\boldsymbol{\beta}}_A^{Ridge}$. In this notation, the out-of-sample predicted value using the original regressors X is $\hat{Y}_I^{OOS} = X^{OOS'}\hat{\boldsymbol{\beta}}_I^{Ridge}$. From Equation (19.99), the ridge estimator is $\hat{\boldsymbol{\beta}}_A^{Ridge} = [(XA)'(XA) + \lambda_{Ridge}I_k]^{-1}(XA)'Y = (A'X'XA + \lambda_{Ridge}I_k)^{-1}A'XY = [A'(X'X + $

$\lambda_{Ridge}\boldsymbol{A}'^{-1}\boldsymbol{A}^{-1})\boldsymbol{A}\,]^{-1}\boldsymbol{A}'\boldsymbol{XY} = \boldsymbol{A}^{-1}[\boldsymbol{X}'\boldsymbol{X} + \lambda_{Ridge}(\boldsymbol{AA}')^{-1}]^{-1}\boldsymbol{XY}$, where the equalities follow by collecting terms using the properties of matrix inverses. Thus the ridge prediction for the out-of-sample observation is $\hat{Y}_A^{OOS} = \boldsymbol{X}^{OOS'}\boldsymbol{A}\hat{\boldsymbol{\beta}}_A^{Ridge} = \boldsymbol{X}^{OOS'}[\boldsymbol{X}'\boldsymbol{X} + \lambda_{Ridge}(\boldsymbol{AA}')^{-1}]^{-1}\boldsymbol{XY}$, whereas using the X's without the linear rotation yields the prediction $\hat{Y}_I^{OOS} = \boldsymbol{X}^{OOS'}(\boldsymbol{X}'\boldsymbol{X} + \lambda_{Ridge}\boldsymbol{I}_k)^{-1}\boldsymbol{XY}$. The two predictions differ because the matrix $(\boldsymbol{AA}')^{-1}$ appears in the expression for $\hat{Y}_A^{OOS}$ but not in the expression for $\hat{Y}_I^{OOS}$. The only time that a linear transformation $\boldsymbol{A}$ does not change the ridge predicted value is when the linear transformation is orthonormal—that is, when $\boldsymbol{AA}' = \boldsymbol{I}_k$, so that $(\boldsymbol{AA}')^{-1} = \boldsymbol{I}_k$.

To see that OLS produces the same predicted value, regardless of the linear transformation $\boldsymbol{A}$ (as long as $\boldsymbol{A}$ is nonsingular), note that the OLS predicted value is the ridge predicted value when $\lambda_{Ridge} = 0$. The result follows from substituting $\lambda^{Ridge} = 0$ into the expressions for the ridge predictions $\hat{Y}_A^{OOS}$ and $\hat{Y}_I^{OOS}$ in the previous paragraph.

Principal Components Analysis

This section presents formulas for the principal components of X and shows that the sum of the variances of the principal components equals the sum of the variances of the X's [Equation (14.10)]. The section concludes with an expression for the out-of-sample prediction, computed using the first r principal components, as in Section 14.5, expressed in terms of the out-of-sample values of the predictors, $\boldsymbol{X}^{OOS}$.

In Key Concept 14.2, the jth principal component of X is defined to be the linear combination of X such that (a) the squared weights of the linear combinations sum to 1; (b) the jth principal component is uncorrelated with the previous $j - 1$ principal components; and (c) the jth principal component maximizes the variance of the linear combination, subject to (a) and (b). We now state these criteria mathematically and use them to derive explicit formulas for the principal components. In particular, we show that the linear combination weights used to form the first r principal components are the eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ corresponding to its r largest eigenvalues.

Let $\boldsymbol{PC}_j$ denote the jth principal component, and let $\boldsymbol{W}_j$ denote the $k \times 1$ vector of weights used to construct $\boldsymbol{PC}_j$, so that $\boldsymbol{PC}_j = \boldsymbol{XW}_j$. The sum of squares of $\boldsymbol{PC}_j$ is $\boldsymbol{PC}_j'\boldsymbol{PC}_j = \boldsymbol{W}_j'\boldsymbol{X}'\boldsymbol{XW}_j$, and the sum of squares weights is $\boldsymbol{W}_j'\boldsymbol{W}_j$. Because X has mean 0 (the X's are standardized), $\boldsymbol{PC}_j'\boldsymbol{PC}_j/(n-1)$ is the sample variance of the jth principal component. The weights $\boldsymbol{W}_j$ are chosen to solve

$$\max_{\boldsymbol{W}_j} \boldsymbol{PC}_j'\boldsymbol{PC}_j = \boldsymbol{W}_j'\boldsymbol{X}'\boldsymbol{XW}_j \text{ subject to } \boldsymbol{W}_j'\boldsymbol{W}_j = 1 \text{ and } \boldsymbol{PC}_j'\boldsymbol{PC}_i = 0 \text{ for } i < j. \quad (19.100)$$

For $j = 1$, the constrained maximization problem is to choose $\boldsymbol{W}_1$ to maximize $\boldsymbol{W}_1'\boldsymbol{X}'\boldsymbol{XW}_1$ subject to $\boldsymbol{W}_1'\boldsymbol{W}_1 = 1$. This constrained maximization is done by maximizing the Lagrangian, $\boldsymbol{W}_1'\boldsymbol{X}'\boldsymbol{XW}_1 - \lambda_1(\boldsymbol{W}_1'\boldsymbol{W}_1 - 1)$, where λ_1 is the Lagrange multiplier. Taking the derivative of the Lagrangian with respect to $\boldsymbol{W}_1$ and setting it to 0 yields

$$\boldsymbol{X}'\boldsymbol{XW}_1 = \lambda_1\boldsymbol{W}_1. \quad (19.101)$$

Equation (19.101) shows that $\boldsymbol{W}_1$ is an eigenvector of $\boldsymbol{X}'\boldsymbol{X}$ and λ_1 is its corresponding eigenvalue, where the eigenvector is normalized to have unit length. Moreover, multiplying

both sides of Equation (19.101) by W_1' shows that $W_1'X'XW_1 = PC_1'PC_1 = \lambda_1$, so that maximizing $PC_1'PC_1$ requires that λ_1 be the largest eigenvalue of $X'X$ and that W_1 be the eigenvector of $X'X$ corresponding to the largest eigenvalue.

Now consider W_2. There are two constraints, $W_2'W_2 = 1$ and $PC_2'PC_1 = W_2'X'XW_1 = 0$, so the Lagrangian is $W_2'X'XW_2 - \lambda_2(W_2'W_2 - 1) - \gamma_{21}W_2'X'XW_1$, where λ_2 and γ_{21} are Lagrange multipliers. Taking the derivative of the Lagrangian with respect to W_2 and setting it to 0 yields

$$X'XW_2 = \lambda_2 W_2 + \frac{1}{2}\gamma_{21}X'XW_1. \tag{19.102}$$

First note that multiplying both sides of Equation (19.101) by W_2' yields $W_2'X'XW_1 = \lambda_1 W_2'W_1$; because $W_2'X'XW_1 = 0$, it follows that $W_2'W_1 = 0$. Now multiplying both sides of Equation (19.102) by W_1' yields $W_1'X'XW_2 = \lambda_2 W_1'W_2 + \frac{1}{2}\gamma_{21}W_1'X'XW_1 = \frac{1}{2}\gamma_{21}W_1'X'XW_1$, but because $W_1'X'XW_2 = W_1'W_2 = 0$, it must be that $\gamma_{21} = 0$. Thus Equation (19.102) reduces to $X'XW_2 = \lambda_2 W_2$, so that W_2 is an eigenvector of $X'X$ and λ_2 is its corresponding eigenvalue. Multiplying both sides of $X'XW_2 = \lambda_2 W_2$ by W_2' and imposing the unit normalization yields $W_2'X'XW_2 = \lambda_2$. Thus, the Lagrangian is maximized by choosing W_2 to be the eigenvector corresponding to the largest of the remaining eigenvalues—that is, to the second-largest eigenvalue of $X'X$.

Continuing, these calculations shows that W_j is the unit-length eigenvector of $X'X$ associated with λ_j, the jth-largest eigenvalue of $X'X$; that $PC_j'PC_j = \lambda_j$; and that $PC_j'PC_i = 0$ for $i \neq j$. If $k < n$, only the first k eigenvalues of $X'X$ are nonzero, so the total number of principal components is $\min(n, k)$.

Because the trace of a matrix is equal to the sum of its eigenvalues,

$$\text{trace}(X'X) = \sum_{j=1}^{\min(n,k)} \lambda_j = \sum_{j=1}^{\min(n,k)} PC_j'PC_j. \tag{19.103}$$

Dividing the first and last expressions in Equation (19.103) by $n - 1$ yields Equation (14.10).

Finally, we provide an expression for the out-of-sample prediction in terms of the out-of-sample value of the predictors, X^{OOS}. The first r out-of-sample values of the principal components are $PC_{1:r}^{OOS} = [PC_1^{OOS} \quad PC_2^{OOS} \quad \cdots \quad PC_r^{OOS}] = W_{1:r}' X^{OOS}$, where $W_{1:r} = [W_1 \quad W_2 \quad \cdots \quad W_r]$ are the first r eigenvectors of $X'X$ in the estimation sample. Let $\hat{\gamma}$ denote the $r \times 1$ vector of OLS coefficients in the regression of Y on the first r principal components in the estimation sample. Then the principal components prediction of Y^{OOS} is $\hat{Y}^{OOS} = \hat{\gamma}'PC_{1:r}^{OOS}$. Written in terms of the original regressors, the principal components prediction is

$$\hat{Y}^{OOS} = \hat{\gamma}'W_{1:r}'X^{OOS}. \tag{19.104}$$

This expression was used to compute the entries in Table 14.4 for the principal components prediction.

Appendix

| TABLE 1 | The Cumulative Standard Normal Distribution Function, $\Phi(z) = \Pr(Z \leq z)$ |

Second Decimal Value of z
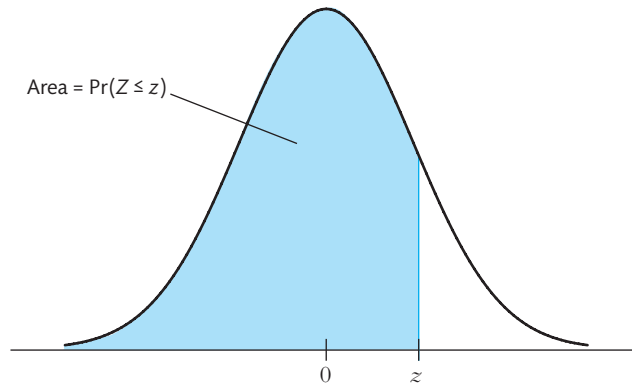
z	0	1	2	3	4	5	6	7	8	9
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611

(Table 1 continued)

(Table 1 continued)

z	0	1	2	3	4	5	6	7	8	9
					Second Decimal Value of z					
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

This table can be used to calculate $\Pr(Z \le z)$ where Z is a standard normal variable. For example, when $z = 1.17$, this probability is 0.8790, which is the table entry for the row labeled 1.1 and the column labeled 7.

| TABLE 2 | Critical Values for Two-Sided and One-Sided Tests Using the Student *t* Distribution |

Significance Level

Degrees of Freedom	20% (2-Sided) 10% (1-Sided)	10% (2-Sided) 5% (1-Sided)	5% (2-Sided) 2.5% (1-Sided)	2% (2-Sided) 1% (1-Sided)	1% (2-Sided) 0.5% (1-Sided)
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.32	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
30	1.31	1.70	2.04	2.46	2.75
60	1.30	1.67	2.00	2.39	2.66
90	1.29	1.66	1.99	2.37	2.63
120	1.29	1.66	1.98	2.36	2.62
∞	1.28	1.64	1.96	2.33	2.58

Values are shown for the critical values for two-sided ($\neq$) and one-sided ($>$) alternative hypotheses. The critical value for the one-sided ($<$) test is the negative of the one-sided ($>$) critical value shown in the table. For example, 2.13 is the critical value for a two-sided test with a significance level of 5% using the Student *t* distribution with 15 degrees of freedom.

| TABLE 3 | Critical Values for the χ^2 Distribution | | |

	Significance Level		
Degrees of Freedom	**10%**	**5%**	**1%**
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.41	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

This table contains the 90th, 95th, and 99th percentiles of the χ^2 distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.

TABLE 4	Critical Values for the $F_{m,\infty}$ Distribution

Area = Significance Level

0

Critical Value

Significance Level

Degrees of Freedom	10%	5%	1%
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78
4	1.94	2.37	3.32
5	1.85	2.21	3.02
6	1.77	2.10	2.80
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.18
13	1.52	1.72	2.13
14	1.50	1.69	2.08
15	1.49	1.67	2.04
16	1.47	1.64	2.00
17	1.46	1.62	1.97
18	1.44	1.60	1.93
19	1.43	1.59	1.90
20	1.42	1.57	1.88
21	1.41	1.56	1.85
22	1.40	1.54	1.83
23	1.39	1.53	1.81
24	1.38	1.52	1.79
25	1.38	1.51	1.77
26	1.37	1.50	1.76
27	1.36	1.49	1.74
28	1.35	1.48	1.72
29	1.35	1.47	1.71
30	1.34	1.46	1.70

This table contains the 90th, 95th, and 99th percentiles of the $F_{m,\infty}$ distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.

TABLE 5A   Critical Values for the F_{n_1,n_2} Distribution—10% Significance Level

Denominator Degrees of Freedom (n_2)	Numerator Degrees of Freedom (n_1)									
	1	2	3	4	5	6	7	8	9	10
1	39.86	49.50	53.59	55.83	57.24	58.20	58.90	59.44	59.86	60.20
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
∞	**2.71**	**2.30**	**2.08**	**1.94**	**1.85**	**1.77**	**1.72**	**1.67**	**1.63**	**1.60**

This table contains the 90th percentile of the F_{n_1,n_2} distribution, which serves as the critical values for a test with a 10% significance level.

| TABLE 5B | Critical Values for the F_{n_1, n_2} Distribution—5% Significance Level |

Denominator Degrees of Freedom (n_2)	Numerator Degrees of Freedom (n_1)									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	**3.84**	**3.00**	**2.60**	**2.37**	**2.21**	**2.10**	**2.01**	**1.94**	**1.88**	**1.83**

This table contains the 95th percentile of the distribution F_{n_1, n_2} which serves as the critical values for a test with a 5% significance level.

TABLE 5C   Critical Values for the F_{n_1, n_2} Distribution—1% Significance Level

Denominator Degrees of Freedom (n_2)	Numerator Degrees of Freedom (n_1)									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	4052.00	4999.00	5403.00	5624.00	5763.00	5859.00	5928.00	5981.00	6022.00	6055.00
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	**6.63**	**4.61**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**

This table contains the 99th percentile of the F_{n_1, n_2} distribution, which serves as the critical values for a test with a 1% significance level.

References

Acemoglu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared. 2008. "Income and Democracy." *American Economic Review* 98(3): 808–842.

Adda, Jérôme, and Francesca Cornaglia. 2006. "Taxes, Cigarette Consumption, and Smoking Intensity." *American Economic Review* 96(4): 1013–1028.

Aggarwal, Rajesh K., and Philippe Jorion. 2010. "The Performance of Emerging Hedge Funds and Managers." *Journal of Financial Economics* 96: 238–256.

Almond, Douglas, Kenneth Y. Chay, and David S. Lee. 2005. "The Costs of Low Birth Weight." *Quarterly Journal of Economics* 120(3): 1031–1083.

American Association for Public Opinion Research. 2017. "An Evaluation of 2016 Election Polls in the United States." http://www.aapor.org/getattachment/Education-Resources/Reports/AAPOR-2016-Election-Polling-Report.pdf.aspx.

Anderson, Theodore W., and Herman Rubin. 1949. "Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations." *Annals of Mathematical Statistics* 21: 570–582.

Andrews, Donald W. K. 1991. "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica* 59(3): 817–858.

Andrews, Donald W. K. 2003. "Tests for Parameter Instability and Structural Change with Unknown Change Point: A Corrigendum." *Econometrica* 71: 395–397.

Angrist, Joshua. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80(3): 313–336.

Angrist, Joshua, and William Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review* 88(3): 450–477.

Angrist, Joshua, Kathryn Graddy, and Guido Imbens. 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *Review of Economic Studies* 67(232): 499–527.

Angrist, Joshua and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Ayres, Ian, and John Donohue. 2003. "Shooting Down the 'More Guns Less Crime' Hypothesis." *Stanford Law Review* 55: 1193–1312.

Bai, Jushan, and Serena Ng. 2002. "Determining the Number of Factors in Approximate Factor Models." *Econometrica* 70: 191–221.

Barendregt, Jan J., L. Bonneux, and P. J. van der Maas. 1997. "The Health Care Costs of Smoking." *New England Journal of Medicine* 337(15): 1052–1057.

Beck, Thorsten, Ross Levine, and Norman Loayza. 2000. "Finance and the Sources of Growth." *Journal of Financial Economics* 58: 261–300.

Benartzi, Shlomo, and Richard H. Thaler. 2007. "Heuristics and Biases in Retirement Savings Behavior." *Journal of Economic Perspectives* 21(3): 81–104.

Benjamin, Daniel J., et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2: 6–10.

Bergstrom, Theodore A. 2001, Fall. "Free Labor for Costly Journals?" *Journal of Economic Perspectives* 15(4): 183–198.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119(1): 249–275.

Bertrand, Marianne, and Kevin Hallock. 2001. "The Gender Gap in Top Corporate Jobs." *Industrial and Labor Relations Review* 55(1): 3–21.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94(4): 991–1013.

Beshears, John, James J. Choi, David Laibson, and Brigitte C. Madrian. 2008. "The Importance of Default Options for Retirement Saving Outcomes: Evidence from the United States." In *Lessons from Pension Reform in the Americas*, edited by Stephen J. Kay and Tapen Sinha, 59–87. Oxford, UK: Oxford University Press.

Bettinger, Eric P., Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu. 2012. "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment." *Quarterly Journal of Economics* 127(3): 1205–1242.

Bollerslev, Tim. 1986. "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics* 31(3): 307–327.

Cameron, A. Colin and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50(2): 317–372.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review* 43(2): 245–257.

Card, David. 1999. "The Causal Effect of Education on Earnings." In *The Handbook of Labor Economics*, edited by Orley C. Ashenfelter and David Card, chap. 30. Amsterdam: Elsevier.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52(1): 57–82.

Carpenter, Christopher, and Philip J. Cook. 2008. "Cigarette Taxes and Youth Smoking: New Evidence from National, State, and Local Youth Risk Behavior Surveys," *Journal of Health* 27(2): 287–299.

Carpenter, Christopher, and Carlos Dobkin. 2011. "The Minimum Legal Drinking Age and Public Health." *Journal of Economic Perspectives* 25(2): 133–156.

Case, Anne, and Christina Paxson. 2008. "Stature and Status: Height, Ability, and Labor Market Outcomes." *Journal of Political Economy* 116(3): 499–532.

Chaloupka, Frank J., Michael Grossman, and Henry Saffer. 2002. "The Effect of Price on Alcohol Consumption and Alcohol-Related Problems." *Alcohol Research & Health* 26: 22–34.

Chaloupka, Frank J., and Kenneth E. Warner. 2000. "The Economics of Smoking." In *The Handbook of Health Economics*, edited by Joseph P. Newhouse and Anthony J. Cuyler, chap. 29. New York: North Holland.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *Quarterly Journal of Economics* 126(4): 1593–1660.

Chow, Gregory. 1960. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica* 28(3): 591–605.

Clay, Karen, Werner Troesken, and Michael Haines. 2014. "Lead and Mortality." *Review of Economics and Statistics* 96(3): 458–470.

Clements, Michael P. 2004. "Evaluating the Bank of England Density Forecasts of Inflation." *Economic Journal* 114: 844–866.

Cochrane, D., and Guy Orcutt. 1949. "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms." *Journal of the American Statistical Association* 44(245): 32–61.

Cook, Philip J., and Michael J. Moore. 2000. "Alcohol." In *The Handbook of Health Economics*, edited by Joseph P. Newhouse and Anthony J. Cuyler, chap. 30. New York: North Holland.

Cooper, Harris, and Larry V. Hedges. 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Dahl, Gordon, and Stefano DellaVigna. 2009. "Does Movie Violence Increase Violent Crime?" *Quarterly Journal of Economics* 124(2): 677–734.

Dang, Jennifer N. 2008. *Statistical Analysis of Alcohol-Related Driving Trends, 1982–2005*. Technical Report DOT HS 810 942. Washington, DC: U.S. National Highway Traffic Safety Administration.

Davis, Jonathan M. V., and Sara B. Heller. 2017. "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs." *American Economic Review* 107(5): 546–550.

Deaton, Angus. 2010, June. "Instruments, Randomization, and Learning About Development." *Journal of Economic Literature* 48: 424–455.

Dickey, David A., and Wayne A. Fuller. 1979. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." *Journal of the American Statistical Association* 74(366): 427–431.

Diebold, Francis X. 2017. *Forecasting*, Department of Economics, University of Pennsylvania, http://www.ssc.upenn.edu/~fdiebold/Textbooks.html

Ehrenberg, Ronald G., Dominic J. Brewer, Adam Gamoran, and J. Douglas Willms. 2001a. "Class Size and Student Achievement." *Psychological Science in the Public Interest* 2(1): 1–30.

Ehrenberg, Ronald G., Dominic J. Brewer, Adam Gamoran, and J. Douglas Willms. 2001b. "Does Class Size Matter?" *Scientific American* 285(5): 80–85.

Eicker, F. 1967. "Limit Theorems for Regressions with Unequal and Dependent Errors." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:59–82. Berkeley: University of California Press.

Enders, Walter. 2009. *Applied Econometric Time Series*. 3rd ed. New York: Wiley.

Engle, Robert F. 1982. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica* 50(4): 987–1007.

Engle, Robert F., and Clive W. J. Granger. 1987. "Cointegration and Error Correction: Representation, Estimation and Testing." *Econometrica* 55(2): 251–276.

Evans, William, Matthew Farrelly, and Edward Montgomery. 1999. "Do Workplace Smoking Bans Reduce Smoking?" *American Economic Review* 89(4): 728–747.

Fuller, Wayne A. 1976. *Introduction to Statistical Time Series.* New York: Wiley.

Gillespie, Richard. 1991. *Manufacturing Knowledge: A History of the Hawthorne Experiments.* New York: Cambridge University Press.

Goering, John, and Ron Wienk, eds. 1996. *Mortgage Lending, Racial Discrimination, and Federal Policy*. Washington, DC: Urban Institute Press.

Granger, Clive W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37(3): 424–438.

Granger, Clive W. J., and A. A. Weiss. 1983. "Time Series Analysis of Error-Correction Models." In *Studies in Econometrics: Time Series and Multivariate Statistics*, edited by S. Karlin, T. Amemiya, and L. A. Goodman, 255–278. New York: Academic Press.

Green, Richard K., and Susan M. Wachter. 2008. "The Housing Finance Revolution." In *Housing, Housing Finance, and Monetary Policy: Symposium Proceedings*, 21–67. Kansas City, MO: Federal Reserve Bank of Kansas City.

Greene, William H. 2018. *Econometric Analysis*. 8th ed. Upper Saddle River, NJ: Prentice Hall.

Gruber, Jonathan. 2001. "Tobacco at the Crossroads: The Past and Future of Smoking Regulation in the

United States." *Journal of Economic Perspectives* 15(2): 193–212.

Haldrup, Niels, and Michael Jansson. 2006. "Improving Size and Power in Unit Root Testing." In *Econometric Theory*. Vol. 1 of *Palgrave Handbook of Econometrics*, edited by Terrence Mills and Kerry Patterson, 255–277. Basingstoke, UK: Palgrave MacMillan.

Hamilton, James D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hansen, Bruce. 1992. "Efficient Estimation and Testing of Cointegrating Vectors in the Presence of Deterministic Trends." *Journal of Econometrics* 53(1–3): 86–121.

Hansen, Bruce. 2001, Fall. "The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity." *Journal of Economic Perspectives* 15(4): 117–128.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50(4): 1029–1054.

Hanushek, Eric. 1999a. "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects." *Educational Evaluation and Policy Analysis* 21: 143–164.

Hanushek, Eric. 1999b. "The Evidence on Class Size." In *Earning and Learning: How Schools Matter*, edited by S. Mayer and P. Peterson, chap. 7. Washington, DC: Brookings Institution Press.

Hayashi, Fumio. 2000. *Econometrics*. Princeton, NJ: Princeton University Press.

Heckman, James J. 1974. "Shadow Prices, Market Wages, and Labor Supply." *Econometrica* 42: 679–694.

Heckman, James J. 2001. "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture." *Journal of Political Economy* 109(4): 673–748.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics*, edited by Orley Ashenfelter and David Card, chap. 31. Amsterdam: Elsevier.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-analysis*. San Diego, CA: Academic Press.

Hetland, Lois. 2000. "Listening to Music Enhances Spatial-Temporal Reasoning: Evidence for the 'Mozart Effect.'" *Journal of Aesthetic Education* 34(3–4): 179–238.

Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics* 115(4): 1239–1285.

Huber, P. J. 1967. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–233. Berkeley: University of California Press.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62: 467–476.

Imbens, Guido. W. and Kolesár, Michal. 2016. "Robust standard errors in small samples: Some practical advice," *Review of Economics and Statistics*, 98(5): 701–712.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1): 5–86.

James, W., and Charles Stein. 1961. "Estimation with Quadratic Loss." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–379. Berkeley: University of California Press.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353(6301): 790–794.

Jones, Stephen R. G. 1992. "Was There a Hawthorne Effect?" *American Journal of Sociology* 98(3): 451–468.

Kilian, Lutz, and Helmut Lütkepohl. 2017. *Structural Vector Autoregressive Analysis*. Cambridge: Cambridge University Press.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133(1): 237–293.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to Learn." *Review of Economics and Statistics* 91: 437–456.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 14(2): 497–562.

Ladd, Helen. 1998, Spring. "Evidence on Discrimination in Mortgage Lending." *Journal of Economic Perspectives* 12(2): 41–62.

Levitt, Steven D. 1996. "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Litigation." *Quarterly Journal of Economics* 111(2): 319–351.

Levitt, Steven D., and Jack Porter. 2001. "How Dangerous Are Drinking Drivers?" *Journal of Political Economy* 109(6): 1198–1237.

List, John. 2003. "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics* 118(1): 41–71.

Long, J. Scott and Laurie H. Ervin. 2000. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *The American Statistician*, 543(3): 217–224.

Maddala, G. S., and In-Moo Kim. 1998. *Unit Roots, Cointegration, and Structural Change*. Cambridge: Cambridge University Press.

Madrian, Brigette C., and Dennis F. Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *Quarterly Journal of Economics* 116(4): 1149–1187.

Malkiel, Burton G. 2016. *A Random Walk Down Wall Street*. New York: Norton.

Manning, Willard G., Emmett B. Keeler, Joseph P. Newhouse, Elizabeth M. Sloss, and Jeffrey Wasserman. 1989. "The Taxes of Sin: Do Smokers and Drinkers Pay Their Way?" *Journal of the American Medical Association* 261(11): 1604–1609.

Matsudaira, Jordan D. 2008. "Mandatory Summer School and Student Achievement." *Journal of Econometrics* 142: 829–850.

McClellan, Mark, Barbara J. McNeil, and Joseph P. Newhouse. 1994. "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality?" *Journal of the American Medical Association* 272(11): 859–866.

Moreira, M. J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica* 71: 1027–1048.

Mosteller, Frederick. 1995, Summer/Fall. "The Tennessee Study of Class Size in the Early School Grades." *Future of Children: Critical Issues for Children and Youths* 5(2): 113–127.

Mosteller, Frederick, Richard Light, and Jason Sachs. 1996, Winter. "Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size." *Harvard Educational Review* 66(4): 631–676.

Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58: 275–309.

Munnell, Alicia H., Geoffrey M. B. Tootell, Lynne E. Browne, and James McEneaney. 1996. "Mortgage Lending in Boston: Interpreting HMDA Data." *American Economic Review* 86(1): 25–53.

Newey, Whitney, and Kenneth West. 1987. "A Simple Positive Semi-definite, Heteroskedastic and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55(3): 703–708.

Nobel Committee for the Economic Sciences Prize. 2017. "Easy Money or a Golden Pension? Integrating Economics and Psychology." https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2017/popular-economicsciences2017.pdf.

Phillips, Peter C. B., and Sam Ouliaris. 1990. "Asymptotic Properties of Residual Based Tests for Cointegration." *Econometrica* 58(1): 165–194.

Quandt, Richard. 1960. "Tests of the Hypothesis That a Linear Regression System Obeys Two Separate Regimes." *Journal of the American Statistical Association* 55(290): 324–330.

Rauscher, Frances, Gordon L. Shaw, and Katherine N. Ky. 1993. "Music and Spatial Task Performance." *Nature* 365(6447): 611.

Roll, Richard. 1984. "Orange Juice and Weather." *American Economic Review* 74(5): 861–880.

Ruhm, Christopher J. 1996. "Alcohol Policies and Highway Vehicle Fatalities." *Journal of Health Economics* 15(4): 435–454.

Ruud, Paul. 2000. *An Introduction to Classical Econometric Theory*. New York: Oxford University Press.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143(2): 534–547.

Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica* 48(1): 1–48.

Stock, James H. 1994. "Unit Roots, Structural Breaks, and Trends." In *Handbook of Econometrics*, vol. 4, edited by Robert Engle and Daniel McFadden, chap. 46. Amsterdam: Elsevier.

Stock, James H., and Francesco Trebbi. 2003. "Who Invented Instrumental Variable Regression?" *Journal of Economic Perspectives* 17: 177–194.

Stock, James H., and Mark W. Watson. 1988. "Variable Trends in Economic Time Series." *Journal of Economic Perspectives* 2(3): 147–174.

Stock, James H., and Mark W. Watson. 1993. "A Simple Estimator of Cointegrating Vectors in Higher-Order Integrated Systems." *Econometrica* 61(4): 783–820.

Stock, James H., and Mark W. Watson. 2001, Fall. "Vector Autoregressions." *Journal of Economic Perspectives* 15(4): 101–115.

Stock, James H., and Mark W. Watson. 2016. "Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics." In *Handbook of Macroeconomics*, vol. 2A, edited by John Taylor and Harald Uhlig, 415–525. Amsterdam: Elsevier.

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg*, edited by Donald W. K. Andrews and James H. Stock, chap. 5. Cambridge: Cambridge University Press.

Tobin, James. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrica* 26(1): 24–36.

Wagenaar, Alexander C., Matthew J. Salois, and Kelli A. Komro. 2009. "Effects of Beverage Alcohol Price and Tax Levels on Drinking: A Meta-Analysis of 1003 Estimates from 112 Studies." *Addiction* 104: 179–190.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48: 827–838.

Winner, Ellen, and Monica Cooper. 2000. "Mute Those Claims: No Evidence (Yet) for a Causal Link Between Arts Study and Academic Achievement." *Journal of Aesthetic Education* 34(3–4): 11–76.

Wooldridge, Jeffrey. 2010. *Economic Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

Wright, Philip G. 1915. "Moore's Economic Cycles." *Quarterly Journal of Economics* 29: 631–641.

Wright, Philip G. 1928. *The Tariff on Animal and Vegetable Oils*. New York: Macmillan.

Young, Douglas J., and Agnieszka Bielinska-Kwapisz. 2006. "Alcohol Prices, Consumption, and Traffic Fatalities." *Southern Economic Journal* 72: 690–703.

Glossary

Acceptance region: The set of values of a test statistic for which the null hypothesis is accepted (is not rejected).

ADF: See *augmented Dickey–Fuller (ADF) test*.

Adjusted R^2 ($\overline{R}^2$): A modified version of R^2 that does not necessarily increase when a new regressor is added to the regression.

ADL(p, q): See *autoregressive distributed lag (ADL) model*.

AIC: See *information criterion*.

Akaike information criterion (AIC): See *information criterion*.

Alternative hypothesis: The hypothesis that is assumed to be true if the null hypothesis is false. The alternative hypothesis is often denoted H_1.

ARCH: See *autoregressive conditional heteroskedasticity (ARCH)*.

AR(p): See *autoregression*.

Asymptotic distribution: The approximate sampling distribution of a random variable computed using a large sample. For example, the asymptotic distribution of the sample average is normal.

Asymptotic normal distribution: A normal distribution that approximates the sampling distribution of a statistic computed using a large sample.

Attrition: The loss of subjects from a study after assignment to the treatment or the control group.

Augmented Dickey–Fuller (ADF) statistic: A regression-based statistic used to test for a unit root in an AR(p) model.

Autocorrelation: The correlation between a time series variable and its lagged value. The jth autocorrelation of Y is the correlation between Y_t and Y_{t-j}.

Autocovariance: The covariance between a time series variable and its lagged value. The jth autocovariance of Y is the covariance between Y_t and Y_{t-j}.

Autoregression: A linear regression model that relates a time series variable to its past (that is, lagged) values. An autoregression with p lagged values as regressors is denoted AR(p).

Autoregressive conditional heteroskedasticity (ARCH): A time series model of conditional heteroskedasticity.

Autoregressive distributed lag (ADL) model: A linear regression model in which the time series variable Y_t is expressed as a function of lags of Y_t and of another variable, X_t. The model is denoted ADL(p, q), where p denotes the number of lags of Y_t and q denotes the number of lags of X_t.

Average causal effect: The population average of the individual causal effects in a heterogeneous population. Also called the average treatment effect.

Average treatment effect: See *average causal effect*.

Balanced panel: A panel data set with no missing observations; that is, the variables are observed for each entity and each time period.

Base specification: A baseline or benchmark regression specification that includes a set of regressors chosen using a combination of expert judgment, economic theory, and knowledge of how the data were collected.

Bayes information criterion (BIC): See *information criterion*.

Bayes rule: the conditional probability of Y given X is the conditional probability of X given Y times the relative marginal probabilities of Y and X:

Bernoulli distribution: The probability distribution of a Bernoulli random variable.

Bernoulli random variable: A random variable that takes on one of two values, 0 and 1. Also known as a binary random variable.

Best Linear Unbiased Estimator (BLUE): An estimator that has the smallest variance of any estimator that is a linear function of the sample values Y and is unbiased. Under the Gauss–Markov conditions, the ordinary least squares estimator is the Best Linear Unbiased Estimator of the regression coefficients conditional on the values of the regressors.

Bias: The expected value of the difference between an estimator and the parameter that it is estimating. If $\hat{\mu}_Y$ is an estimator of μ_Y, then the bias of $\hat{\mu}_Y$ is $E(\hat{\mu}_Y) - \mu_Y$.

BIC: See *information criterion*.

Binary variable: A variable that is either 0 or 1. A binary variable is used to indicate a binary outcome. For example, X is a binary (or indicator, or dummy) variable for a person's sex if $X = 1$ if the person is female and $X = 0$ if the person is male.

Bivariate normal distribution: A generalization of the normal distribution to describe the joint distribution of two random variables.

Bivariate normal p.d.f.: See *bivariate normal distribution*.

BLUE: See *Best Linear Unbiased Estimator (BLUE)*.

Bonferroni test: A way to test a joint hypothesis by testing the component individual hypotheses one at a time, using an adjusted critical value that accounts for the multiple hypotheses being tested.

Break date: The date of a discrete change in population time series regression coefficient(s).

Causal effect: The expected effect of a given intervention or treatment on an outcome as measured in an ideal randomized controlled experiment.

Causal inference: Tests, confidence intervals, and/or estimation of a causal effect.

c.d.f.: Cumulative distribution function. See *cumulative probability distribution*.

Central limit theorem: In mathematical statistics, under general conditions, the sampling distribution of the standardized sample average is well approximated by a standard normal distribution when the sample size is large.

Chi-squared distribution: The distribution of the sum of m squared independent standard normal random variables. The parameter m is called the degrees of the freedom of the chi-squared distribution.

Chow test: A test for a break in a time series regression at a known break date.

Classical measurement error model: The observed value of a random variable equals its true, unobserved value plus independent measurement error.

Clustered standard errors: A method of computing standard errors that is appropriate for panel data.

Coefficient of determination: See R^2.

Cointegration: When two or more time series variables share a common stochastic trend.

Common component: In a dynamic factor model, the part of a time series variable that is explained by the common unobserved factors.

Common trend: A trend shared by two or more time series.

Conditional distribution: The probability distribution of one random variable given that another random variable takes on a particular value.

Conditional expectation: The expected value of one random variable given that another random variable takes on a particular value.

Conditional heteroskedasticity: The variance, usually of an error term, depends on other variables.

Conditional mean: The mean of a conditional distribution. See *conditional expectation*.

Conditional mean independence: The conditional expectation of the regression error u_i given the regressors depends on some but not all of the regressors.

Conditional variance: The variance of a conditional distribution.

Confidence interval (confidence set): An interval (or set) constructed from sample data that contains the true value of a population parameter with a prespecified probability when computed over repeated samples.

Confidence level: The prespecified probability that a confidence interval (or set) contains the true value of the parameter.

Consistency: The property that an estimator is consistent. See *consistent estimator*.

Consistent estimator: An estimator that converges in probability to the parameter that it is estimating.

Constant regressor: The regressor associated with the regression intercept; this regressor is always equal to 1.

Constant term: The regression intercept.

Continuous mapping theorem: If a random variable S_n converges in distribution to S, then a continuous function of that random variable, $g(S_n)$, converges in distribution to $g(S)$.

Continuous random variable: A random variable that takes on a continuum of values.

Control group: The group that does not receive the treatment or intervention in an experiment.

Control variable: A regressor that controls for an omitted factor that determines the dependent variable.

Converge in probability: When a sequence of random variables converges to a specific value; for example, when the sample average becomes close to the population mean as the sample size increases; see Key Concept 2.6 and Section 18.2.

Convergence in distribution: When a sequence of distributions converges to a limit; a precise definition is given in Section 18.2.

Correlation: A unit-free measure of the extent to which two random variables move, or vary, together. The correlation (or correlation coefficient) between X and Y is $\sigma_{XY}/\sigma_X\sigma_Y$ and is denoted corr(X, Y).

Correlation coefficient: See *correlation*.

Covariance: A measure of the extent to which two random variables move together. The covariance between X and Y is the expected value $E\left[(X - \mu_X)(Y - \mu_Y)\right]$ and is denoted cov(X, Y) or σ_{XY}.

Covariance matrix: A matrix composed of the variances and covariances of a vector of random variables.

Coverage probability: The probability that a confidence interval contains the true value of the coefficient.

Critical value: The value of a test statistic for which the test just rejects the null hypothesis at the given significance level.

Cross-sectional data: Data collected for different entities in a single time period.

Cubic regression model: A nonlinear regression function that includes X, X^2, and X^3 as regressors.

Cumulative distribution function (c.d.f.): See *cumulative probability distribution*.

Cumulative dynamic multiplier: The cumulative effect of a unit change in the time series variable X on Y. The h-period cumulative dynamic multiplier is the effect of a unit change in X_t on $Y_t + Y_{t+1} + \cdots + Y_{t+h}$.

Cumulative probability distribution: A function showing the probability that a random variable is less than or equal to a given number.

Dependent variable: The variable to be explained in a regression or other statistical model; the variable appearing on the left-hand side in a regression.

Deterministic trend: A persistent long-term movement of a variable over time that can be represented as a nonrandom function of time.

DFM: See *dynamic factor model (DFM)*.

Dickey–Fuller statistic: A regression-based statistic used to test for a unit root in a first-order autoregression [AR(1)].

Differences estimator: An estimator of the causal effect constructed as the difference in the sample average outcomes between the treatment and control groups.

Differences-in-differences estimator: The average change in Y for those in the treatment group minus the average change in Y for those in the control group.

Discrete random variable: A random variable that takes on discrete values.

Distributed lag model: A regression model in which the regressors are current and lagged values of X.

Dummy variable: See *binary variable*.

Dummy variable trap: A problem caused by including a full set of binary variables in a regression together with a constant regressor (intercept), leading to perfect multicollinearity.

Dynamic causal effect: The causal effect of one variable on current and future values of another variable.

Dynamic factor model (DFM): A representation of N time series variables, where each variable is expressed as the sum of a reduced number r of common unobserved factors plus an idiosyncratic disturbance that is uncorrelated with the factors and the idiosyncratic disturbances of the other variables.

Dynamic multiplier: The h-period dynamic multiplier is the effect of a unit change in the time series variable X_t on Y_{t+h}.

Endogenous variable: A variable that is correlated with the error term.

Entity and time fixed effects regression model: A panel data regression that includes both entity fixed effects and time fixed effects.

Entity fixed effects: a set of variables that provide for each entity in a panel data regression to have its own intercept.

Errors-in-variables bias: The bias in an estimator of a regression coefficient that arises from measurement errors in the regressors.

Error term: The difference between Y and the population regression function, denoted u in this text.

ESS: See *explained sum of squares (ESS)*.

Estimate: The numerical value of an estimator computed using data from a specific sample.

Estimator: A function of a sample of data to be drawn randomly from a population. An estimator uses sample data to compute an educated guess of the value of a population parameter, such as the population mean.

Exact (finite-sample) distribution: The exact probability distribution of a random variable.

Exact identification: When the number of instrumental variables equals the number of endogenous regressors.

Exogenous variable: A variable that is uncorrelated with the regression error term.

Expectation: See *expected value*.

Expected value: The long-run average value of a random variable over many repeated trials or occurrences. It is the probability-weighted average of all possible values that the random variable can take on. The expected value of Y is denoted $E(Y)$ and is also called the expectation of Y.

Experimental data: Data obtained from an experiment designed to evaluate a treatment or policy or to investigate a causal effect.

Explained sum of squares (ESS): The sum of squared deviations of the predicted values of Y_i, $\hat{Y}_i$, from their average; see Equation (4.14).

Explanatory variable: See *regressor*.

External validity: Inferences and conclusions from a statistical study are externally valid if they can be generalized from the population and the setting studied to other populations and settings.

Fan chart: a time series plot that displays a forecast distribution (forecast uncertainty) as a function of the forecast horizon.

Feasible GLS estimator: A version of the generalized least squares (GLS) estimator that uses an estimator of the conditional variance of the regression errors and covariance between the regression errors at different observations.

Feasible WLS: A version of the weighted least squares (WLS) estimator that uses an estimator of the conditional variance of the regression errors.

Final prediction error (FPE): An estimator of the mean squared forecast error when the regression coefficients are estimated by ordinary least squares.

First difference: The first difference of a time series variable Y_t is $Y_t - Y_{t-1}$, denoted ΔY_t.

First-stage regression: The regression of an included endogenous variable on the included exogenous variables, if any, and the instrumental variable(s) in two stage least squares.

Fitted value: See *predicted value*.

Fixed effects: Binary variables indicating the entity or time period in a panel data regression.

Fixed effects regression model: A panel data regression that includes entity fixed effects.

***F_{m,n}* distribution:** The distribution of a ratio of independent random variables, where the numerator is a chi-squared random variable with m degrees of freedom, divided by m, and the denominator is an independently distributed chi-squared random variable with n degrees of freedom, divided by n.

***F_{m,\infty}* distribution:** The distribution of a random variable with a chi-squared distribution with m degrees of freedom, divided by m.

Forecast error: The difference between the value of the variable that actually occurs and its forecasted value.

Forecast interval: An interval that contains the future value of a time series variable with a prespecified probability.

FPE: See *final prediction error*.

***F*-statistic:** A statistic used to test a joint hypothesis concerning more than one of the regression coefficients.

Functional form misspecification: When the form of the estimated regression function does not match the form of the population regression function; for example, when a linear specification is used but the true population regression function is quadratic.

GARCH: See *generalized autoregressive conditional heteroskedasticity (GARCH)*.

Gauss–Markov theorem: Under certain conditions, the ordinary least squares estimator is the best linear unbiased estimator of the regression coefficients conditional on the values of the regressors.

Generalized autoregressive conditional heteroskedasticity (GARCH): A time series model for conditional heteroskedasticity.

Generalized least squares (GLS): A generalization of ordinary least squares that is appropriate when the regression errors have a known form of heteroskedasticity (in which case GLS is also referred to as weighted least squares, or WLS) or a known form of serial correlation.

Generalized method of moments (GMM): A method for estimating parameters by fitting sample moments to population moments that are functions of the unknown parameters. Instrumental variables estimators are an important special case.

GLS: See *generalized least squares (GLS)*.

GMM: See *generalized method of moments (GMM)*.

Granger causality test: A procedure for testing whether current and lagged values of one time series help predict future values of another time series.

HAC standard errors: See *heteroskedasticity- and autocorrelation-consistent (HAC) standard errors*.

Hawthorne effect: The phenomenon that experimental subjects change their behavior because they know they are subjects in an experiment.

Heteroskedasticity: The variance of the regression error term u_i, conditional on the regressors, is not constant.

Heteroskedasticity- and autocorrelation-consistent (HAC) standard errors: Standard errors for ordinary least squares estimators that are consistent whether or not the regression errors are heteroskedastic and/or autocorrelated.

Heteroskedasticity- and autocorrelation-robust (HAR) standard errors: Another term for HAC standard errors.

Heteroskedasticity-robust standard error: A standard error for the ordinary least squares estimator that is appropriate whether the error term is homoskedastic or heteroskedastic.

Heteroskedasticity-robust *t*-statistic: A *t*-statistic constructed using a heteroskedasticity-robust standard error.

Homoskedasticity: The variance of the regression error term u_i, conditional on the regressors, is constant.

Homoskedasticity-only *F*-statistic: A form of the *F*-statistic that is valid only when the regression errors are homoskedastic.

Homoskedasticity-only standard errors: Standard errors for the ordinary least squares estimator that are appropriate only when the error term is homoskedastic.

Hypothesis test: A procedure for using sample evidence to help determine if a specific hypothesis about a population is true or false.

***I*(0), *I*(1), and *I*(2):** See *order of integration*.

Identically distributed: When two or more random variables have the same distribution.

Idiosyncratic component: In a dynamic factor model, the part of a time series variable that is not explained by the common unobserved factors.

i.i.d. See *independently and identically distributed (i.i.d.)*.

Impact effect: The contemporaneous, or immediate, effect of a unit change in the time series variable X_t on Y_t.

Imperfect multicollinearity: The condition in which two or more regressors are highly correlated.

Included endogenous variables: Regressors that are correlated with the error term (usually in the context of instrumental variable regression).

Included exogenous variables: Regressors that are uncorrelated with the error term (usually in the context of instrumental variable regression).

Independence: When knowing the value of one random variable provides no information about the value of another random variable. Two random variables are independent if their joint distribution is the product of their marginal distributions.

Independently and identically distributed (i.i.d.): When two or more independent random variables have the same distribution.

Indicator variable: See *binary variable*.

Information criterion: A statistic used to estimate the number of lagged variables to include in an autoregression or a distributed lag model. Leading examples are the Akaike information criterion (AIC) and the Bayes information criterion (BIC).

In-sample prediction: The predicted value of the dependent variable for an observation in the sample used to estimate the prediction model.

Instrument: See *instrumental variable*.

Instrument exogeneity condition: The requirement that an instrumental variable is uncorrelated with the error term in the instrumental variables regression equation.

Instrument relevance condition: The requirement that an instrumental variable is correlated with the included endogenous regressor.

Instrumental variable: A variable that is correlated with an endogenous regressor (instrument relevance) and is uncorrelated with the regression error (instrument exogeneity).

Instrumental variables (IV) regression: A way to obtain a consistent estimator of the unknown coefficients of the function relating Y to X when the regressor, X, is correlated with the error term, u.

Interaction term: A regressor that is formed as the product of two other regressors, such as $X_{1i} \times X_{2i}$.

Intercept: The value of β_0 in the linear regression model.

Internal validity: When inferences about causal effects in a statistical study are valid for the population being studied.

IV: See *instrumental variables (IV) regression*.

Joint hypothesis: A hypothesis consisting of two or more individual hypotheses—that is, involving more than one restriction on the parameters of a model.

Joint probability distribution: The probability distribution determining the probabilities of outcomes involving two or more random variables.

***J*-statistic:** A statistic for testing overidentifying restrictions in instrumental variables regression.

Kurtosis: A measure of how much mass is contained in the tails of a probability distribution.

Lag: The value of a time series variable in a previous time period. The jth lag of Y_t is Y_{t-j}.

Lasso (least absolute shrinkage and selection operator): The regression estimator that minimizes a penalized sum of squared residuals, where the penalty term is proportional to the sum of the absolute values of the regression coefficients.

Law of iterated expectations: A result in probability theory that says that the expected value of Y is the expected value of its conditional expectation given X—that is, that $E(Y) = E[E(Y|X)]$.

Law of large numbers: According to this result from probability theory, under general conditions the sample average will be close to the population mean with very high probability when the sample size is large.

Least squares assumptions: The assumptions for the linear regression models listed in Key Concept 4.3 (single variable regression model) and Key Concept 6.4 (multiple regression model).

Least squares estimator: An estimator formed by minimizing the sum of squared residuals.

Leptokurtic: A distribution that has heavier tails than a normal, as measured by a kurtosis exceeding 3.

Likelihood function: The joint probability distribution of the data, treated as a function of the unknown coefficients.

Limited dependent variable: A dependent variable that can take on only a limited set of values. For example, the variable might be a 0–1 binary variable or arise from one of the models described in Appendix 11.3.

Linear-log model: A nonlinear regression function in which the dependent variable is Y and the independent variable is $\ln(X)$.

Linear probability model: A regression model in which Y is a binary variable.

Linear regression function: A regression function with a constant slope.

Local average treatment effect: A weighted average treatment effect estimated, for example, by two stage least squares.

Logarithm: See *natural logarithm*.

Logit regression: A nonlinear regression model for a binary dependent variable in which the population regression function is modeled using the cumulative logistic distribution function.

Log-linear model: A nonlinear regression function in which the dependent variable is $\ln(Y)$ and the independent variable is X.

Log-log model: A nonlinear regression function in which the dependent variable is $\ln(Y)$ and the independent variable is $\ln(X)$.

Longitudinal data: See *panel data*.

Long-run cumulative dynamic multiplier: The cumulative long-run effect on the time series variable Y of a change in X.

Marginal probability distribution: Another name for the probability distribution of a random variable Y, which distinguishes the distribution of Y alone (the marginal distribution) from the joint distribution of Y and another random variable.

Maximum likelihood estimator (MLE): An estimator of unknown parameters that is obtained by maximizing the likelihood function; see Appendix 11.2.

Mean: The expected value of a random variable. The mean of Y is denoted μ_Y.

Mean squared forecast error (MSFE): The expected value of the square of the time series forecast error for an observation not in the data set used for estimating the forecasting model.

Mean squared prediction error (MSPE): The expected value of the square of the prediction error for an observation not in the data set used for estimating the prediction model.

***m*-fold cross validation:** A method for estimating the mean squared prediction error by first dividing the in-sample data into m subsamples and then sequentially forming predictions for the observations in each subsample using the data not in that subsample.

MLE: See *maximum likelihood estimator (MLE)*.

Moments of a distribution: The expected value of a random variable raised to different powers. The rth moment of the random variable Y is $E(Y^r)$.

MSFE: See *mean squared forecast error (MSFE)*.

MSPE: See *mean squared prediction error (MSPE)*.

Multicollinearity: See *perfect multicollinearity* and *imperfect multicollinearity*.

Multiple regression model: An extension of the single variable regression model that allows Y to depend on k regressors.

Multi-step ahead forecast: A forecast made for more than one period beyond the final observation used to make the forecast.

Natural experiment: See *quasi-experiment*.

Natural logarithm: A mathematical function defined for a positive argument; its slope is always positive but tends to zero. The natural logarithm is the inverse of the exponential function; that is, $X = \ln(e^X)$.

95% confidence set: A confidence set with a 95% confidence level. See *confidence interval*.

Nonlinear least squares: The analog of ordinary least squares that applies when the regression function is a nonlinear function of the unknown parameters.

Nonlinear least squares estimator: The estimator obtained by minimizing the sum of squared residuals when the regression function is nonlinear in the parameters.

Nonlinear regression function: A regression function with a slope that is not constant.

Nonstationary: When the joint distribution of one or more time series variables and their lagged values changes over time.

Normal distribution: A commonly used bell-shaped distribution of a continuous random variable.

Nowcast: The forecast of the value of a time series variable for the current period—that is, the period in which the forecast is made.

Null hypothesis: The hypothesis being tested in a hypothesis test, often denoted H_0.

Observational data: Data based on observing, or measuring, actual behavior outside an experimental setting.

Observation number: The unique identifier assigned to each entity in a data set.

OLS estimator: See *ordinary least squares (OLS) estimator*.

OLS regression line: The regression line with population coefficients replaced by the ordinary least squares estimators.

OLS residual: The difference between Y_i and the ordinary least squares regression line, denoted $\hat{u}_i$ in this text.

Omitted variables bias: The bias in an estimator that arises because a variable that is a determinant of Y and is correlated with a regressor has been omitted from the regression.

One-sided alternative hypothesis: The parameter of interest is on one side of the value given by the null hypothesis.

One-step ahead forecast: A forecast made for the period immediately following the final observation used to make the forecast.

Oracle prediction: The infeasible best-possible prediction, which is made using the unknown conditional mean of the variable to be predicted given the predictors.

Order of integration: The number of times that a time series variable must be differenced to make it stationary. A time series variable that is integrated of order d must be differenced d times and is denoted $I(d)$.

Ordinary least squares (OLS) estimators: The estimators of the regression intercept and slope(s) that minimize the sum of squared residuals.

Out-of-sample prediction: The predicted value of the dependent variable for an observation not in the sample used to estimate the prediction model.

Outlier: An exceptionally large or small value of a random variable.

Overidentification: When the number of instrumental variables exceeds the number of included endogenous regressors.

Panel data: Data collected for multiple entities where each entity is observed in two or more time periods.

Parameters: Constants that determine a characteristic of a probability distribution or population regression function.

Partial compliance: The failure of some participants to follow the treatment protocol in a randomized experiment.

Partial effect: The effect on Y of changing one of the regressors while holding the other regressors constant.

p.d.f.: See *probability density function (p.d.f.)*.

Penalized sum of squared residuals: The sum of the sum of squared residuals and a penalty term that increases with the number and/or values of the regression coefficients.

Penalty term: A term that, when added to the sum of squared residuals, penalizes the estimator for choosing a large number of regressors and/or coefficients with large values.

Perfect multicollinearity: A situation in which one of the regressors is an exact linear function of the other regressors.

Polynomial regression model: A nonlinear regression function that includes $X, X^2, \ldots,$ and X^r as regressors, where r is an integer.

Population: The group of entities—such as people, companies, or school districts—being studied.

Population coefficients: See *population intercept and slope*.

Population intercept and slope: The true, or population, values of β_0 (the intercept) and β_1 (the slope) in a single-variable regression. In a multiple regression, there are multiple slope coefficients $(\beta_1, \beta_2, \ldots, \beta_k)$, one for each regressor.

Population multiple regression model: The multiple regression model in Key Concept 6.2.

Population regression line: In a single-variable regression, the population regression line is $\beta_0 + \beta_1 X_i$, and in a multiple regression, it is $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$.

Potential outcomes: The set of outcomes that might occur to an individual (treatment unit) after receiving, or not receiving, an experimental treatment.

Power of a test: The probability that a test correctly rejects the null hypothesis when the alternative is true.

Predicted value: The value of Y_i that is predicted by the ordinary least squares regression line, denoted $\hat{Y}_i$ in this text.

Price elasticity of demand: The percentage change in the quantity demanded resulting from a 1% increase in price.

Principal components: The linear combinations of a set of standardized variables for which the jth linear combination maximizes its variance, subject to being uncorrelated with the previous $j - 1$ linear combinations.

Probability: The proportion of time that an outcome (or event) from a random experiment occurs in the long run.

Probability density function (p.d.f.): For a continuous random variable, the area under the probability density function between any two points is the probability that the random variable falls between those two points.

Probability distribution: For a discrete random variable, a list of all values that a random variable can take on and the probability associated with each of these values.

Probit regression: A nonlinear regression model for a binary dependent variable in which the population regression function is modeled using the cumulative standard normal distribution function.

Program evaluation: The field of study concerned with estimating the effect of a program, policy, or some other intervention or "treatment."

Pseudo out-of-sample forecast: A forecast computed over part of the sample using a procedure that is *as if* these sample data have not yet been realized.

p-value (significance probability): The probability of drawing a statistic at least as adverse to the null hypothesis as the one actually computed, assuming the null hypothesis is correct. Also called the marginal significance probability, the p-value is the smallest significance level at which the null hypothesis can be rejected.

Quadratic regression model: A nonlinear regression function that includes X and X^2 as regressors.

Quandt likelihood ratio statistic: A statistic used with time series data to test for a break in the regression model at an unknown date.

Quasi-experiment: A circumstance in which randomness is introduced by variations in individual circumstances that make it appear *as if* the treatment is randomly assigned.

R^2: In a regression, the fraction of the sample variance of the dependent variable that is explained by the regressors.

$\bar{R}^2$: See *adjusted R^2*.

Randomized controlled experiment: An experiment in which participants are randomly assigned to a control group, which receives no treatment, or to a treatment group, which receives a treatment.

Random walk: A time series process in which the value of the variable equals its value in the previous period plus an unpredictable error term.

Random walk with drift: A generalization of the random walk in which the change in the variable has a nonzero mean but is otherwise unpredictable.

Realized volatility: The sample root mean square of a time series variable computed over consecutive time periods.

Regressand: See *dependent variable*.

Regression discontinuity: A regression involving a quasi-experiment in which treatment depends on whether an observable variable crosses a threshold.

Regression specification: A description of a regression that includes the set of regressors and any nonlinear transformation that has been applied.

Regressor: A variable appearing on the right-hand side of a regression; an independent variable in a regression.

Rejection region: The set of values of a test statistic for which the test rejects the null hypothesis.

Repeated cross-sectional data: A collection of cross-sectional data sets, where each cross-sectional data set corresponds to a different time period.

Residual: The difference between the observed value of the dependent variable and its value predicted by an estimated regression, for an observation in the sample used to estimate the regression coefficients, denoted $\hat{u}_i$ in the text.

Restricted regression: A regression in which the coefficients are restricted to satisfy some condition. For example, when computing the homoskedasticity-only F-statistic, it is the regression with coefficients restricted to satisfy the null hypothesis.

Ridge regression: The regression estimator that minimizes a penalized sum of squared residuals, where the penalty term is proportional to the sum of the squared regression coefficients.

RMSFE: See *root mean squared forecast error (RMSFE)*.

Root mean squared forecast error (RMSFE): The square root of the mean squared forecast error.

Sample correlation coefficient (sample correlation): An estimator of the correlation between two random variables.

Sample covariance: An estimator of the covariance between two random variables.

Sample selection bias: The bias in an estimator of a regression coefficient that arises when a selection process influences the availability of data and that process is related to the dependent variable. This bias induces correlation between one or more regressors and the regression error.

Sample standard deviation: An estimator of the population standard deviation of a random variable.

Sample variance: An estimator of the population variance of a random variable.

Sampling distribution: The distribution of a statistic over all possible samples; the distribution arising from repeatedly evaluating the statistic using a series of randomly drawn samples from the same population.

Scatterplot: A plot of n observations on X_i and Y_i, in which each observation is represented by the point (X_i, Y_i).

Scree plot: The normalized variance of the ordered principal components of a set of variables X, plotted against the principal component number, where the variance is normalized by the sum of the variances of the X's.

SER: See *standard error of the regression (SER)*.

Serial correlation: See *autocorrelation*.

Serially uncorrelated: A time series variable with all autocorrelations equal to 0.

Shrinkage estimator: An estimator that introduces bias by shrinking the OLS estimator toward a specific point (usually 0) and thereby reducing the variance of the estimator.

Significance level: The prespecified rejection probability of a statistical hypothesis test when the null hypothesis is true.

Significance probability: See *p-value (significance probability)*.

Simple random sampling: When entities are chosen independently from a population using a method that ensures that each entity is equally likely to be chosen.

Simultaneous causality: When, in addition to the causal link of interest from X to Y, there is a causal link from Y to X. Simultaneous causality makes X correlated with the error term in the function of interest that relates Y to X.

Simultaneous equations bias: See *simultaneous causality*.

Size of a test: The probability that a test incorrectly rejects the null hypothesis when the null hypothesis is true.

Skewness: A measure of the asymmetry of a probability distribution.

Sparse model: A regression model in which the coefficients are nonzero for only a small fraction of the predictors.

SSR: See *sum of squared residuals (SSR)*.

Standard deviation: The square root of the variance. The standard deviation of the random variable Y, denoted σ_Y, has the same units as Y and is a measure of the spread of the distribution of Y around its mean.

Standard error of an estimator: An estimator of the standard deviation of the estimator.

Standard error of the regression (*SER*): An estimator of the standard deviation of the regression error u.

Standardized predictive regression model: A special case of the linear multiple regression model in which the regressors are standardized and the dependent variable is demeaned so that it has mean 0.

Standardized random variable: Subtracting the mean and dividing by the standard deviation produces a standardized random variable with a mean of 0 and a standard deviation of 1. The standardized random variable computed from Y is $(Y - \mu_Y)/\sigma_Y$.

Standard normal distribution: The normal distribution with mean equal to 0 and variance equal to 1, denoted $N(0, 1)$.

Stationarity: When the joint distribution of a time series variable and its lagged values does not change over time.

Statistically insignificant: The null hypothesis (typically, that a regression coefficient is 0) cannot be rejected at a given significance level.

Statistically significant: The null hypothesis (typically, that a regression coefficient is 0) is rejected at a given significance level.

Stochastic trend: A persistent but random long-term movement of a variable over time.

Strict exogeneity: The requirement that the regression error have a mean of 0 conditional on current, future, and past values of the regressor in a distributed lag model.

Student t distribution: The Student t distribution with m degrees of freedom is the distribution of the ratio of a standard normal random variable, divided by the square root of an independently distributed chi-squared random variable with m degrees of freedom divided by m. As m gets large, the Student t distribution converges to the standard normal distribution.

Sum of squared residuals (*SSR*): The sum of the squared ordinary least squares residuals.

***t* distribution:** See *Student t distribution*.

Time effects: Binary variables indicating the time period in a panel data regression.

Time fixed effects: See *time effects*.

Time series data: Data collected for the same entity for multiple time periods.

Total sum of squares (*TSS*): The sum of squared deviations of Y_i from its average.

***t*-ratio:** See *t-statistic*.

Treatment effect: The causal effect in an experiment or a quasi-experiment. See *causal effect*.

Treatment group: The group that receives the treatment or intervention in an experiment.

TSLS: See *two stage least squares*.

***TSS*:** See *total sum of squares (TSS)*.

***t*-statistic:** A statistic used for hypothesis testing. See Key Concept 5.1.

Two-sided alternative hypothesis: When, under the alternative hypothesis, the parameter of interest is not equal to the value given by the null hypothesis.

Two stage least squares (TSLS): An instrumental variable estimator, described in Key Concept 12.2.

Type I error: In hypothesis testing, the error made when the null hypothesis is true but is rejected.

Type II error: In hypothesis testing, the error made when the null hypothesis is false but is not rejected.

Unbalanced panel: A panel data set in which data for some entities are missing for some time periods.

Unbiased estimator: An estimator with a bias that is equal to 0.

Uncorrelated: Two random variables are uncorrelated if their correlation is 0.

Underidentification: When the number of instrumental variables is less than the number of endogenous regressors.

Unit root: An autoregression with a largest root equal to 1.

Unrestricted regression: A regression in which the coefficients are not restricted to satisfy some condition. When computing the homoskedasticity-only F-statistic, it is the regression that applies under the alternative hypothesis, so that the coefficients are not restricted to satisfy the null hypothesis.

VAR: See *vector autoregression*.

Variance: The expected value of the squared difference between a random variable and its mean; the variance of Y is denoted σ_Y^2.

Vector autoregression (VAR): A model of k time series variables consisting of k equations, one for each variable, in which the regressors in all equations are lagged values of all the variables.

Volatility clustering: When a time series variable exhibits some clustered periods of high variance and other clustered periods of low variance.

Weak instruments: Instrumental variables that have a low correlation with the endogenous regressor(s).

Weighted least squares (WLS): An alternative to ordinary least squares that can be used when the regression error is heteroskedastic and the form of the heteroskedasticity is known or can be estimated.

WLS: See *weighted least squares (WLS)*.

This page intentionally left blank

Index

Page numbers followed by *f* indicate figures; *t* indicates tables.

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank

Large-Sample Critical Values for the t-statistic from the Standard Normal Distibution

	Significance Level		
	10%	**5%**	**1%**
2-Sided Test (≠)			
Reject if $|t|$ is greater than	1.64	1.96	2.58
1-Sided Test (>)			
Reject if t is greater than	1.28	1.64	2.33
1-Sided Test (<)			
Reject if t is less than	−1.28	−1.64	−2.33

Large-Sample Critical Values for the F-statistic from the $F_{m,\infty}$ Distribution

Reject if $F >$ Critical Value

Degrees of Freedom (m)	Significance Level		
	10%	5%	1%
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78
4	1.94	2.37	3.32
5	1.85	2.21	3.02
6	1.77	2.10	2.80
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.18
13	1.52	1.72	2.13
14	1.50	1.69	2.08
15	1.49	1.67	2.04
16	1.47	1.64	2.00
17	1.46	1.62	1.97
18	1.44	1.60	1.93
19	1.43	1.59	1.90
20	1.42	1.57	1.88
21	1.41	1.56	1.85
22	1.40	1.54	1.83
23	1.39	1.53	1.81
24	1.38	1.52	1.79
25	1.38	1.51	1.77
26	1.37	1.50	1.76
27	1.36	1.49	1.74
28	1.35	1.48	1.72
29	1.35	1.47	1.71
30	1.34	1.46	1.70