Chủ biên: PGS. TS. NGUYỄN TRỌNG HOÀI
Nhóm Tác giả: NGUYỄN TRỌNG HOÀI
PHÙNG THANH BÌNH - NGUYỄN KHÁNH DUY

DỰ BÁO VÀ PHÂN TÍCH DỮ LIỆU TRONG KINH TẾ VÀ TÀI CHÍNH

NHÀ XUẤT BẢN THỐNG KÊ
Năm 2009

LỜI NÓI ĐẦU

Sự phát triển của nền kinh tế Việt Nam đặt trong bối cảnh đầy biến động của nền kinh tế thế giới đã thúc đẩy các nhà kinh tế và quản trị phải quan tâm nhiều hơn đến việc dự báo ở nhiều lĩnh vực khác nhau. Đặc biệt, với sự ra đời của thị trường chứng khoán kèm theo sự biến động của nhiều chỉ số kinh tế trong và ngoài nước đã và đang thúc đẩy dự báo trở thành một hoạt động quan trọng trong các lĩnh vực kinh tế học, tài chính – ngân hàng, thương mại, tiếp thị, và nhiều lĩnh vực dịch vụ – sản xuất kinh doanh khác. Nhu cầu dự báo ở Việt Nam đang có xu hướng gia tăng bởi vì dự báo đã và đang ảnh hưởng đáng kể đến nhiều quyết định hàng ngày của các cơ quan quản lý Nhà nước và doanh nghiệp. Trước thực tế đó, ngày càng nhiều nhà nghiên cứu, nhà hoạch định chính sách, và sinh viên các chuyên ngành kinh tế – tài chính muốn trang bị một nền tảng cơ bản về các phương pháp dự báo định lượng. Tuy nhiên, nguồn tài liệu tham khảo bằng tiếng Việt còn hạn chế đã gây không ít khó khăn cho việc giảng dạy và học tập môn dự báo tại Việt Nam. Và điều này đã thôi thúc chúng tôi – nhóm giảng viên phụ trách môn *Dự báo và Phân tích dữ liệu trong kinh tế và tài chính*, Trường Đại học Kinh tế TP.HCM – cùng nhau nghiên cứu và viết cuốn giáo trình này.

Giáo trình này được biên soạn chủ yếu nhằm phục vụ cho việc giảng dạy và học tập môn *Dự báo và Phân tích dữ liệu trong kinh tế và tài chính* ở bậc Cử nhân của trường Đại học Kinh tế TP.HCM, và giáo trình này đã chính thức được Hội đồng khoa học Trường nghiệm thu đưa vào sử dụng kể từ năm học 2009. Ngoài ra, chúng tôi hy vọng đây cũng là một nguồn tài liệu tham khảo hữu ích cho tất cả các sinh viên kinh tế – tài chính ở nhiều cơ sở đào tạo khác trong cả nước và những ai hiện đang làm việc trong lĩnh vực kế hoạch, đầu tư, và phân tích chính sách hoặc thị trường của các cơ quan Nhà nước và doanh nghiệp. Phạm vi của cuốn giáo trình này chỉ giới hạn ở một số phương pháp dự báo định lượng phổ biến dựa trên các nguồn dữ liệu chuỗi thời gian.

Giáo trình này được biên soạn một cách hệ thống trên cơ sở tham khảo có chọn lọc nhiều nguồn tài liệu cập nhật của các học giả nổi tiếng trên thế giới. Mục tiêu chính của chúng tôi là muốn giới thiệu đến bạn đọc dưới dạng

"hướng dẫn thực hành" trên phần mềm Eviews và Crystal Ball một số phương pháp dự báo cơ bản có thể vận dụng tức thời cho công tác dự báo các biến quan trọng trong hoạt động kinh doanh của doanh nghiệp như doanh số và chi phí hoạt động; và các chỉ báo kinh tế chủ yếu của nền kinh tế như GDP, lãi suất, chỉ số giá chứng khoán, giá vàng, giá dầu, tỷ giá hối đoái và nhiều chỉ số quan trọng khác. Mặt khác, qua cuốn giáo trình này, chúng tôi cũng muốn chia sẻ những kinh nghiệm thực tế đúc kết từ quá trình mà chúng tôi đã từng nghiên cứu, giảng dạy, và tư vấn để có thể ứng dụng một cách hiệu quả dự báo vào quá trình ra quyết định.

Giáo trình này được chia thành 10 chương với các nội dung có liên quan chặt chẽ với nhau:

Chương 1 giới thiệu tổng quan về dự báo để giúp bạn đọc cảm nhận được vai trò của dự báo, nắm được sơ bộ cách phân loại dự báo và phương pháp luận của dự báo định lượng, hiểu được toàn bộ quy trình dự báo, và đặc biệt là biết cách đánh giá kết quả dự báo.

Chương 2 giới thiệu một số vấn đề cơ bản về xác suất thống kê để giúp bạn đọc ôn lại một số khái niệm cần thiết cho việc phân tích dữ liệu và dự báo như thống kê mô tả, phân phối xác suất, và kiểm định giả thiết. Chúng tôi không đề cập nhiều đến các công thức toán và thống kê phức tạp, mà chỉ quan tâm đến các ý tưởng ứng dụng thực tiễn trên phần mềm Eviews và Excel để bạn đọc với nền tảng khác nhau về xác suất thống kê đều có thể tiếp cận dự báo dễ dàng dưới góc độ ứng dụng thực sự. Trong chương này, chúng tôi cũng sẽ giới thiệu các loại dữ liệu kinh tế, cách nhập, quản lý, phân tích, và chuyển hóa dữ liệu trên phần mềm Eviews.

Chương 3 là một chương quan trọng nhất trong giáo trình này vì nó cung cấp một nền tảng cơ bản cho người phân tích biết cách lựa chọn những mô hình dự báo thích hợp với nguồn dữ liệu sẵn có. Trên cơ sở này, chúng ta sẽ lần lượt khảo sát từng phương pháp dự báo cụ thể ở các chương tiếp theo. Lưu ý rằng, không thể có một mô hình dự báo tốt nhất cho tất cả các trường hợp, mà tùy vào thực tế dữ liệu và nguồn lực sẵn có mà người làm dự báo chọn lựa ra các mô hình dự báo thích hợp cho mục đích sử dụng của mình.

Chương 4 sẽ hướng dẫn cách thức thực hiện các mô hình dự báo giản đơn như trung bình di động, san mũ giản đơn, san mũ Holt, và san mũ Winters. Mặc dù, đây là những phương pháp dự báo giản đơn, nhưng lại

được sử dụng rất rộng rãi trong phạm vi các doanh nghiệp sản xuất (dự báo doanh số, tồn kho, và chi phí hoạt động). Đặc biệt, trong chương này chúng tôi có trình bày một cách rất cụ thể cách thức thực hiện từng phương pháp trên phần mềm Crystal Ball. Có lẽ nhiều người trong chúng ta đã hiểu sức mạnh của Crystal Ball trong phân tích rủi ro, nhưng ít ai ngờ rằng Crystal Ball còn tuyệt vời như thế nào trong việc hỗ trợ thực hiện các mô hình dự báo giản đơn.

Chương 5 chuyên về các mô hình hàm xu thế, nghĩa là hồi quy một biến định lượng theo biến xu thế thời gian. Các mô hình hàm xu thế trở nên hữu ích trong việc dự báo xu hướng vận động của các chuỗi thời gian trong giai đoạn tăng trưởng của chu kỳ kinh doanh hoặc dự báo tốc độ tăng trưởng của một chỉ số kinh tế – xã hội.

Chương 6 sẽ giới thiệu hai mô hình chuyên cho các loại dữ liệu có xu hướng dao động theo mùa vụ. Thông qua các phương pháp này, chúng ta sẽ phân tích các thành phần cơ bản của một chuỗi thời gian để xác định các chỉ số mùa vụ (quý, tháng) một cách hợp lý. Mặc dù, đây không phải là các phương pháp duy nhất có thể mô hình hóa yếu tố mùa vụ của chuỗi thời gian, nhưng chúng được sử dụng phổ biến nhờ sự đơn giản ở khía cạnh toán học.

Chương 7 ứng dụng phương pháp phân tích hồi quy trong dự báo hệ số co giãn và xác định các nhân tố quan trọng ảnh hưởng đến đối tượng cần dự báo. Ở đây, chúng tôi có đề cập một số vấn đề cơ bản của phân tích hồi quy để giúp người phân tích nhận ra rằng để có thể dự báo một biến nhất định (biến phụ thuộc) từ các thông tin đã có (biến giải thích), chúng ta cần đảm bảo mô hình hồi quy cuối cùng phải là một mô hình tốt nhất. Chúng tôi thiết nghĩ rằng, chương này cũng là một tài liệu tham khảo rất hữu ích cho việc học môn Kinh tế lượng căn bản. Do đây chỉ là giáo trình cho bậc cử nhân, nên chúng tôi không trình bày các mô hình nâng cao như biến công cụ, VAR, nhân quả, hoặc các mô hình dữ liệu bảng.

Chương 8 sẽ trình bày các mô hình ARIMA vốn được sử dụng rất phổ biến trong rất nhiều lĩnh vực khác nhau khi dự báo các chỉ số có độ nhạy cao. Có thể trước khi đến với cuốn giáo trình này, nhiều bạn đọc có cảm giác hoài nghi về sự khó hiểu của các mô hình ARIMA. Tuy nhiên, chúng tôi tin rằng bạn đọc sẽ thật sự thấy thú vị vì có thể cảm nhận được sức mạnh của nhiều ứng dụng thực tiễn nhờ cách diễn giải hết sức gần gũi. Chúng tôi lấy làm vui mừng vì trong vài năm gần đây nhiều sinh viên năm

thứ ba của Trường Đại học Kinh tế TP.HCM đã thực hiện thành công nhiều nghiên cứu ứng dụng bằng các mô hình ARIMA.

Chương 9 sẽ đưa bạn đọc vào thế giới nghiên cứu ứng dụng hết sức hấp dẫn bằng các mô hình ARCH. Những mô hình này gần đây được ứng dụng phổ biến trong giới phân tích tài chính và chứng khoán vì chúng không chỉ giúp khắc phục nhược điểm của các mô hình ARIMA đối với các chuỗi dữ liệu có xu hướng biến động mạnh theo thời gian, mà còn giúp chúng ta có thể dự báo được yếu tố rủi ro trong một thế giới luôn tồn tại những điều không chắc chắn. Nhờ sự phát triển của phần mềm Eviews mà trước đây chúng tôi tưởng rằng các mô hình ARCH chỉ có thể thuộc về thế giới của các học giả đã trở thành một công cụ trong tầm tay của các sinh viên năm ba các chuyên ngành kinh tế – tài chính và nhiều nhà phân tích đầu tư ngay tại Việt Nam. Chúng tôi hy vọng rằng đây sẽ là một phương pháp hữu ích cho các học viên cao học và những người nghiên cứu trong lĩnh vực tài chính – chứng khoán ở Việt Nam trong một tương lai không xa.

Chương 10 sẽ tổng kết toàn bộ quy trình dự báo và đưa ra khung quản lý hiệu quả quy trình dự báo tại các tổ chức. Đặc biệt, chúng tôi đúc kết bảy yếu tố then chốt quyết định kết quả dự báo mà các tổ chức cần lưu ý để nâng cao hiệu quả quá trình ra quyết định. Mặc dù, cuốn giáo trình này chỉ tập trung vào các mô hình dự báo định lượng, nhưng chúng tôi cũng không quên vai trò và triển vọng của các phương pháp dự báo định tính. Cho nên, chúng tôi cho rằng sẽ là thiếu xót nếu bạn đọc bỏ qua những trang cuối cùng của cuốn giáo trình này.

Chúng tôi biết rằng còn rất nhiều mô hình dự báo khác rất mạnh và đang được nhiều học giả nghiên cứu phát triển, nhưng chúng tôi chỉ dừng lại ở các mô hình vừa nêu trên để bạn đọc có một nền tảng căn bản về dự báo. Ở cuối mỗi chương, chúng tôi có đưa ra nhận xét tóm tắt các nội dung chính, nhiều câu hỏi và bài tập dưới dạng tình huống kinh doanh. Qua các câu hỏi và bài tập này, bạn đọc không chỉ có cơ hội thực hành trên máy tính, mà còn tiếp cận được rất nhiều nguồn dữ liệu có thể rất cần thiết trong quá trình nghiên cứu và ứng dụng thực tế sau này. Đặc biệt, một số tình huống được thảo luận xuyên suốt qua hầu hết các chương để giúp bạn đọc dễ dàng hình dung một quy trình dự báo sẽ được thực hiện như thế nào tại doanh nghiệp. Chúng tôi hy vọng rằng sau khi học xong cuốn giáo trình này, bạn đọc sẽ có được niềm đam mê về phân tích định lượng và dự báo để sau này có thể tự nghiên cứu các mô hình nâng cao phù hợp với lĩnh vực chuyên môn của mình.

Giá trị của cuốn giáo trình này còn nằm ở những thông điệp được đúc kết từ kinh nghiệm thực tế trong quá trình nghiên cứu, giảng dạy, và tư vấn phân tích dữ liệu và dự báo của nhóm tác giả trong nhiều năm.

Trong quá trình viết cuốn giáo trình này, chúng tôi may mắn nhận được nhiều ý kiến đóng góp quý báu từ Hội đồng khoa học của Trường Đại học Kinh tế TP.HCM, bao gồm PGS. TS. Nguyễn Đình Thọ (Đại học Kinh tế TP.HCM), TS. Cao Hào Thi (Đại học Bách khoa TP.HCM), TS. Huỳnh Thị Thu Thủy (Đại học Kinh tế TP.HCM), TS. Trần Tiến Khai (Đại học Kinh tế TP.HCM), TS. Dư Quang Nam (Cục Thống kê TP.HCM), TS. Nguyễn Hữu Dũng (Đại học Kinh tế TP.HCM), ThS. Trương Thanh Vũ (Viện Chiến lược phát triển). Một số nội dung chính của giáo trình này đã được tham khảo và trích dẫn từ các tài liệu về kinh tế lượng và dự báo của GS. Dimitrios Asteriou, GS. Domodar Gujarati, GS. Francis Diebold, GS. Holton Wilson, GS. John Hanke, GS. Makridakis, GS. Mark Moon, GS. John Mentzer, GS. Michael Clements, GS. David Hendry, GS. Michael Evans, GS. Carter Hill, GS. Ramu Ramanathan, GS. Robert Pindyck, GS. Scott Amstrong, GS. Stephen Satchell, và nhiều học giả khác. Chúng tôi cũng xin cám ơn Phòng Nghiên cứu Khoa học Trường Đại học Kinh tế TP.HCM đã tạo điều kiện thuận lợi cho nhóm tác giả trong suốt quá trình nghiên cứu và viết cuốn giáo trình này. Ngoài ra, tinh thần say mê học tập và nghiên cứu của sinh viên Khoa Kinh tế Phát triển, Trung tâm Tư vấn doanh nghiệp và phát triển vùng thuộc Khoa cũng là một nguồn động lực mạnh mẽ cho sự hoàn thành cuốn giáo trình này.

Dữ liệu và các bài giảng bằng Powerpoint đã được đưa lên trang web của khoa Kinh tế Phát triển (www.fde.ueh.edu.vn). Đây là các dữ liệu thu thập từ các nhiều nguồn khác nhau, kể cả các giáo trình dự báo và kinh tế lượng như đã liệt kê ở Danh mục tài liệu tham khảo. Chúng tôi chỉ sử dụng các dữ liệu trên cho mục đích giáo dục, nên nhiều thông tin đã được điều chỉnh nhằm đảm bảo tính bảo mật của thông tin. Ngoài ra, nếu có nhu cầu về giảng dạy, tư vấn, cung cấp dữ liệu đã sử dụng trong cuốn sách này, xin quý đọc giả vui lòng liên hệ với nhóm tác giả hoặc Khoa Kinh tế Phát triển Đại học Kinh tế TP.HCM.

Mặc dù chúng tôi đã nỗ lực ở mức cao nhất, nhưng những sai sót vẫn có khả năng xảy ra, và đó là điều không tránh khỏi. Chính vì vậy, mọi sự đóng góp xây dựng của bạn đọc để hoàn thiện cuốn giáo trình là món quà vô cùng ý nghĩa đối với chúng tôi. Mọi ý kiến đóng góp, rất mong quý vị vui lòng gửi về các địa chỉ sau đây: Nguyễn Trọng Hoài (hoaianh@ueh.edu.vn),

Phùng Thanh Bình (ptbinh@ueh.edu.vn), và Nguyễn Khánh Duy (khanhduy@ueh.edu.vn).

Trân trọng!

TP.HCM, Tháng 8 Năm 2009
PGS. TS. NGUYỄN TRỌNG HOÀI
Trưởng Khoa Kinh tế Phát triển
Đại học Kinh tế TP.HCM

MỤC LỤC

CHƯƠNG 1 TỔNG QUAN VỀ DỰ BÁO

Để bắt đầu chương này, chúng tôi xin mượn lời của một vị giám đốc chiến lược chuỗi cung ứng của công ty Motts North America như sau: "Tôi tin rằng dự báo có lẽ có khả năng đóng góp vào giá trị của một doanh nghiệp nhiều hơn bất kỳ một hoạt động nào khác trong chuỗi cung ứng vì dự báo đúng sẽ làm cho mọi thứ khác trong chuỗi cung ứng được tiến hành một cách dễ dàng hơn"[1].

MỤC TIÊU HỌC TẬP

Sau khi học xong chương này, chúng ta kỳ vọng sẽ đạt được các nội dung sau đây:

- Hiểu được dự báo là gì.
- Biết được tại sao dự báo ngày càng đóng vai quan trọng trong quá trình ra quyết định của tổ chức.
- Hiểu được vì sao dự báo định lượng đang trở nên phổ biến.
- Biết được những lĩnh vực nào cần sự hỗ trợ của dự báo.
- Hiểu được các cách phân loại dự báo khác nhau.
- Hiểu được phương pháp luận của dự báo.
- Nắm vững quy trình thực hiện dự báo trên thực tế.
- Biết được cách thức đo lường độ chính xác của dự báo.

[1] Wilson, J. Holton & Barry Keating, 2007, Business Forecasting.

DỰ BÁO VÀ VAI TRÒ DỰ BÁO

"Dự báo" hàm ý dự đoán điều gì đó cho tương lai. Dự báo có thể là bất kỳ một phát biểu nào về tương lai và phát biểu đó có thể hoặc không dựa trên một hoặc một số căn cứ khoa học nào đó. Chính vì vậy, kết quả dự báo có thể chính xác hoặc không chính xác. Để tránh những dự báo thiếu căn cứ, người làm dự báo cần được trang bị các phương pháp dự báo khoa học và có hệ thống. Các phương pháp dự báo như thế được dùng trong kinh tế và kinh doanh là trọng tâm mà giáo trình này sẽ giới thiệu đến bạn đọc.

Dự báo được hiểu là việc ước lượng một sự kiện hoặc một điều kiện nào đó trong tương lai vốn nằm ngoài khả năng kiểm soát của tổ chức nhằm cung cấp cơ sở cho việc ra quyết định. Dự báo tốt có thể giúp tổ chức hình dung ra tương lai của mình sẽ như thế nào để hoạch định hướng đi phù hợp. Trong giáo trình này, cụm từ "tổ chức" được dùng để chỉ các doanh nghiệp, các cơ quan, cá nhân trong tổ chức, và nhiều tổ chức kinh tế xã hội khác có nhu cầu dự báo.

Cùng với sự phát triển của máy tính và nhiều phần mềm ứng dụng, dự báo ngày càng trở nên quan trọng và trở thành bộ phận không thể thiếu trong hầu hết các quyết định của mọi tổ chức. Dự báo có độ chính xác cao sẽ cung cấp cơ sở tin cậy cho hoạch định chính sách cũng như xây dựng các chiến lược kinh doanh. David (2000) cho rằng hiện nay nhiều tổ chức đang rất cần những chuyên viên biết kỹ thuật dự báo và nhu cầu tuyển dụng người làm dự báo đang có xu hướng gia tăng đáng kể, nhất là các đơn vị sản xuất kinh doanh do ba yếu tố sau. Thứ nhất, dự báo ngày càng được sử dụng phổ biến ở hầu hết các bộ phận của doanh nghiệp trong quá trình xây dựng kế hoạch chiến lược, phân tích tình huống kinh doanh, lập kế hoạch ngân sách vốn đầu tư, v.v... Vì thế, những người lập kế hoạch chiến lược, phân tích tài chính, kế toán, nghiên cứu thị trường, các nhà kinh tế, v.v..., đều cần biết các kỹ thuật dự báo. Thứ hai, để dự báo thực sự là cơ sở cho việc ra quyết định thì người làm dự báo và người sử dụng dự báo phải thường xuyên trao đổi qua lại. Cho nên, nếu những người sử dụng (thường là những nhà quản lý cấp cao của một tổ chức) có kiến thức về dự báo và tin cậy các kết quả dự báo sẽ có ý nghĩa rất lớn

trong quá trình ra quyết định. Thứ ba, dự báo có ý nghĩa sống còn đối với sự thành công của một tổ chức vì nhiều kết quả khảo sát ở Mỹ và các nước phát triển cho thấy khoảng 92% doanh nghiệp cho rằng dự báo rất quan trọng đối với sự thành công của doanh nghiệp.

Tóm lại, các tổ chức đang hoạt động trong một thế giới liên tục thay đổi nhưng các quyết định phải được thực hiện ngay hôm nay và ảnh hưởng sống còn đến tương lai của tổ chức, nên dự báo dĩ nhiên luôn luôn cần thiết nếu thực sự tổ chức muốn tồn tại và phát triển bền vững.

DỰ BÁO ĐANG ĐƯỢC SỬ DỤNG PHỔ BIẾN

Dự báo có thể được xem như một tập hợp các công cụ giúp người ra quyết định thực hiện các phán đoán tốt nhất có thể có về các sự kiện sẽ xảy ra trong tương lai. Và tương lai là bất định và nhiều rủi ro. Trong một thế giới kinh doanh liên tục biến đổi thì những phán đoán như thế có thể tạo ra sự khác biệt giữa thành công và thất bại. Thông điệp quan trọng nhất mà giáo sư Michael E. Porter, trường Kinh doanh, Đại học Harvard gửi đến các doanh nghiệp Việt Nam vào ngày 01 tháng 12 năm 2008 tại Việt Nam là "chúng ta chỉ có thể cạnh tranh và phát triển bằng cách tạo ra sự khác biệt". Cho nên chỉ dựa vào cảm tính để dự báo tình hình kinh tế của một quốc gia, doanh số tương lai, nhu cầu tồn kho, tuyển mới nhân sự, và nhiều biến kinh tế khác của một doanh nghiệp sẽ không còn hợp lý nữa. Wilson (2007) cho rằng các phương pháp định lượng rất hữu ích trong việc đưa ra các dự đoán tin cậy về tương lai, và nhiều phần mềm máy tính phức tạp đã và đang được phát triển nhằm đưa các phương pháp này ngày càng gần mọi người hơn. Tuy nhiên, hãy cẩn thận nếu không thì chúng ta sẽ trở thành những chiếc hộp đen. Chính vì thế, mục tiêu của giáo trình này là giúp bạn đọc hiểu rõ bản chất của từng phương pháp dự báo trước khi áp dụng trên các phần mềm đã được lập trình sẵn.

Giáo trình này được biên soạn mang tính ứng dụng nhằm giúp bạn đọc nắm vững cơ sở khái niệm và quy trình thực hiện từng phương

pháp dự báo định lượng đang được sử dụng phổ biến trên các phần mềm thông dụng, mà đặc biệt là phần mềm Eviews. Sau khi học xong giáo trình này, chúng tôi kỳ vọng bạn đọc có thể độc lập dự báo các chỉ báo kinh tế và kinh doanh với dữ liệu thực tế với độ chính xác cao. Và những dự báo chính xác như vậy chắc chắn sẽ làm gia tăng giá trị thị trường của doanh nghiệp. Mặc dù giáo trình chỉ thiên về các phương pháp định lượng, nhưng chúng tôi khuyên những người làm công tác dự báo hãy thận trọng. Đừng dựa quá nhiều vào các phương pháp định lượng và các kết quả từ máy tính mà không suy nghĩ một cách cẩn thận các dữ liệu mình đang dự báo. Các phán đoán cá nhân dựa trên kinh nghiệm thực tế và có nghiên cứu kỹ lưỡng ngữ cảnh của các biến số đang xem xét luôn có ý nghĩa quan trọng trong việc chuẩn bị dự báo, thực hiện dự báo, và ra quyết định.

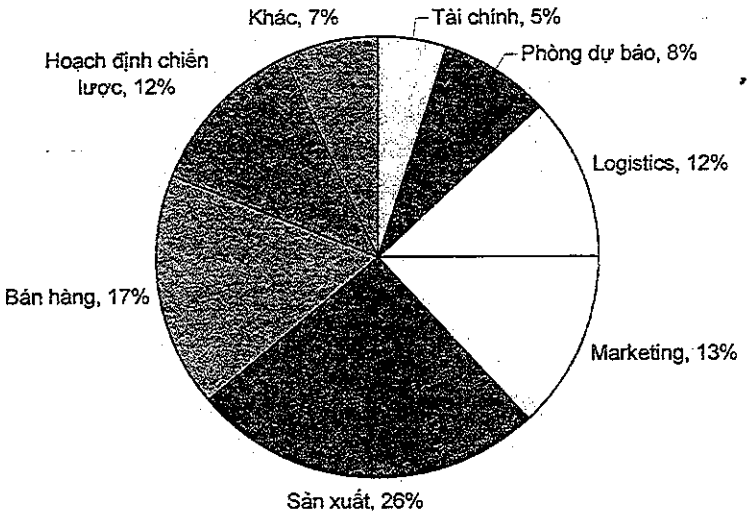
NHU CẦU DỰ BÁO

Nhắc lại rằng dự báo các chỉ báo kinh tế và kinh doanh đóng một vai trò rất quan trọng trong quá trình ra quyết định kinh doanh của doanh nghiệp, phân tích chính sách, và trong rất nhiều nghiên cứu kinh tế ứng dụng. Hầu như mỗi tổ chức, lớn hay nhỏ, công hay tư đều đang thực hiện dự báo theo một cách nào đó bởi vì hoạch định luôn luôn là một trong những chức năng cơ bản nhất của bất kỳ một tổ chức nào. Nhu cầu dự báo ngày càng gia tăng ở hầu hết các bộ phận chức năng của tổ chức để thực hiện các dự báo cho các quyết định về tài chính, tiếp thị, nhân sự, sản xuất, v.v... Trong kinh doanh, dự báo có ý nghĩa đặc biệt quan trọng trong mọi hoạt động của doanh nghiệp vì các doanh nghiệp cạnh tranh không chỉ thông qua nỗ lực đáp ứng tốt nhất nhu cầu khách hàng mà còn thông qua nỗ lực giảm các chi phí kinh doanh. Các quyết định kinh doanh của một doanh nghiệp thông thường dựa vào một dự báo nào đó. Ví dụ:

- Phòng kế toán dựa vào dự báo doanh thu hoặc chi phí để lên kế hoạch báo cáo thuế.
- Phòng nhân sự dựa vào dự báo thị trường lao động để có kế hoạch tuyển dụng nhân sự mới và các thay đổi khác trong lực lượng lao động của tổ chức.

- Phòng tài chính phải dự báo ngân lưu và chi phí sử dụng vốn để lập kế hoạch ngân sách vốn đầu tư, cơ cấu vốn tối ưu, tỷ lệ chia cổ tức, quản lý rủi ro và xác định giá trị doanh nghiệp.

■ HÌNH 1.1: **Chức năng thực hiện dự báo ở các công ty Mỹ năm 2005.**



Nguồn: Chaman (2006).

- Bộ phận sản xuất dựa vào dự báo để quyết định nhu cầu nguyên vật liệu và tồn kho.
- Phòng tiếp thị dựa vào dự báo doanh số để xây dựng ngân sách quảng cáo.
- Phòng nghiên cứu thị trường có thể dựa vào dự báo hệ số co giãn để đưa ra các quyết định về giá và quy mô sản lượng, hoặc dự báo nhân tố ảnh hưởng sự thỏa mãn nhu cầu khách hàng.

- Các phòng nghiên cứu và phân tích đầu tư của các công ty quỹ, công ty chứng khoán, và ngân hàng có thể dự báo giá chứng khoán, suất sinh lợi kỳ vọng, rủi ro thị trường, tốc độ tăng trưởng và triển vọng thị trường nhằm thực hiện các quyết định đầu tư và tư vấn khách hàng, hoặc dự báo các chỉ báo kinh tế vĩ mô khác như cung tiền, lãi suất, giá dầu, giá vàng, tỷ giá hối đoái, v.v..., để lập các báo cáo thị trường, báo cáo ngành, hoặc hỗ trợ cho nhiều quyết định đầu tư khác.

- Phòng kinh doanh, đặc biệt là của các công ty phụ thuộc nhiều vào tình hình thị trường thế giới như nhiên liệu, hóa chất, vật liệu xây dựng, v.v..., cần thực hiện nhiều dự báo về xu hướng biến động của các chỉ số giá thế giới. Ví dụ, thay vì dự báo giá dầu hoặc giá khí trong nước, các công ty dầu khí có xu hướng dự báo giá dầu, giá khí của thị trường Trung Đông để có thể đưa ra các kế hoạch kinh doanh hợp lý.

Tuy nhiên, dự báo doanh số thường là dự báo cơ bản nhất trong doanh nghiệp vì hầu hết các kế hoạch của doanh nghiệp đều dựa vào biến doanh số. Theo Wilson (2007), một nghiên cứu vào giữa thập niên 1980 của các tập đoàn lớn ở Mỹ cho biết có tới 94% các kế hoạch kinh doanh của họ dựa vào dự báo doanh số. Trong khi đó, con số này khoảng 98% ở các công ty Canada vào năm (Klassen, 2001). Dựa trên kết quả khảo sát năm 2005, Chaman (2006) đã cho biết hiện nay những bộ phận nào chuyên phụ trách công tác dự báo ở các công ty Mỹ (Hình 1.1).

Như vậy, các bộ phận sản xuất, bán hàng, marketing, logistics, và hoạch định chiến lược là các bộ phận thường phụ trách công việc dự báo cho các công ty ở Mỹ. Đặc biệt, ngày càng có nhiều công ty đã có phòng dự báo hoạt động độc lập và phụ trách toàn bộ các công việc dự báo của công ty.

Đối với các cơ quan Nhà nước, các tổ chức phi Chính phủ, những nhà nghiên cứu và phân tích chính sách thì dự báo các chỉ báo kinh tế vĩ mô như lạm phát, tốc độ tăng trưởng kinh tế, tỷ lệ thất nghiệp, lãi suất, tỷ lệ nghèo đói, chỉ số giá chứng khoán, v.v..., để xây dựng các

chiến lược kinh tế xã hội, các chương trình mục tiêu, các dự án phát triển, và để đưa ra được nhiều gợi ý chính sách quan trọng.

PHÂN LOẠI DỰ BÁO

Dự báo có thể được phân loại theo kết quả dự báo, phạm vi dự báo, và phương pháp dự báo.

DỰA TRÊN KẾT QUẢ DỰ BÁO

Dựa vào kết quả, người ta có thể chia dự báo thành dự báo điểm và dự báo khoảng. Dự báo điểm là kết quả dự báo được biểu hiện bằng một giá trị duy nhất, dự báo khoảng là kết quả dự báo được cho trong một khoảng giá trị với một xác suất tin cậy cho trước. Dự báo có thể là ngắn hạn, trung hạn hoặc dài hạn. Các dự báo dài hạn cần thiết khi doanh nghiệp xây dựng các kế hoạch chiến lược dài hạn, vì thế loại dự báo này thường được sử dụng bởi những người quản lý cấp cao. Các dự báo ngắn hạn chỉ được sử dụng để xây dựng các kế hoạch tức thời và thường được sử dụng bởi các quản trị viên cấp thấp. Nhìn chung, khoảng cách dự báo càng ngắn thì mức độ chính xác càng cao. Trong các mô hình dự báo định lượng, chúng ta sẽ đề cập chi tiết về dự báo điểm và dự báo khoảng.

DỰA TRÊN PHẠM VI DỰ BÁO

Nếu dựa vào quy mô có thể chia thành dự báo kinh tế vi mô và dự báo kinh tế vĩ mô. Khi nói về dự báo có lẽ chúng ta thường nghĩ ngay đến dự báo các biến quan trọng cho một doanh nghiệp hoặc một bộ phận nào đó của doanh nghiệp như doanh số, tồn kho, v.v... Tuy nhiên, việc dự báo các chỉ báo kinh tế quan trọng của nền kinh tế một quốc gia hay địa phương ngày càng được quan tâm, như dự báo tỷ lệ thất nghiệp, thu nhập quốc nội, lãi suất, chỉ số giá chứng khoán, v.v..., vì bất kỳ một hoạch định chính sách kinh tế nào cũng phải dựa vào giá trị ước đoán của các chỉ báo kinh tế quan trọng như thế. Chính vì thế, ngày càng có nhiều phương pháp dự báo được phát triển chuyên cho các lĩnh vực kinh tế vĩ mô. Một vấn đề tranh cãi trong dự báo các

chỉ báo kinh tế vĩ mô là làm sao đảm bảo mức độ chính xác nếu xảy ra một biến đổi đột ngột của một yếu tố kinh tế quan trọng như thay đổi giá dầu hay thay đổi chính sách. Các quản trị cấp cao của một tổ chức (nhất là phòng nghiên cứu của các công ty tài chính hoặc các công ty đa quốc gia) có thể quan tâm đến kết quả dự báo các chỉ báo kinh tế vĩ mô của nền kinh tế một địa phương, một quốc gia, và thậm chí toàn cầu trong việc cân nhắc đưa ra chiến lược kinh doanh của mình, nhất là trong một thế giới phẳng và không ngừng biến động như hiện nay. Ví dụ, một số công ty kinh doanh sản phẩm khí ở Việt Nam có thể nghĩ đến việc dự báo giá khí của các thị trường thế giới (đặc biệt là giá CP của công ty Saudi Aramco) dựa vào nhiều chỉ báo kinh tế toàn cầu như giá dầu, giá vàng, chỉ số giá tiêu dùng, tỷ giá USD/EUR, lãi suất SIBOR, các chỉ số chứng khoán như DJIA, FTSE, Nikkei, chỉ số Baltic Clean Tanker, và nhiều chỉ báo khác về lượng cung cầu về dầu hoặc khí ở Mỹ, Trung Đông, Nga, Trung Quốc, Châu Âu, và Nhật Bản.

DỰA TRÊN PHƯƠNG PHÁP DỰ BÁO

Dựa theo phương pháp có thể chia dự báo thành các nhóm phương pháp chính thức và các nhóm không chính thức. Các phương pháp không chính thức phần lớn dựa vào trực giác cảm tính, phụ thuộc vào kinh nghiệm và khả năng phán đoán của cá nhân. Các phương pháp này chỉ được sử dụng khi không có đủ thời gian, dữ liệu, và nhất là không được trang bị các phương pháp chính thức. Nói chung, các phán đoán cảm tính như thế thường không có độ tin cậy cao. Các phương pháp chính thức được sử dụng phổ biến vì có phương pháp luận rõ ràng. Các phương pháp chính thức được chia thành phương pháp dự báo định tính và dự báo định lượng.

Các phương pháp định tính

Các phương pháp định tính dựa vào kinh nghiệm và phán đoán của những chuyên viên, những người quản lý và những chuyên gia. Phương pháp định tính thường được sử dụng khi dữ liệu lịch sử không sẵn có hay có nhưng không đầy đủ, hoặc không đáng tin cậy, hoặc những đối tượng dự báo bị ảnh hưởng bởi những nhân tố không

thể lượng hóa được như sự thay đổi tiến bộ kỹ thuật. Để minh họa sự cần thiết của dự báo định tính, chúng ta hãy xem một số trường hợp điển hình sau đây. Trong quy trình thẩm định dự án, sau khi chuyên viên phân tích trình bày kết quả phân tích tài chính và rủi ro của một dự án cụ thể, thì để có thể ra quyết định cuối cùng, các nhà quản lý cấp cao cần dựa vào kinh nghiệm và những phán đoán định tính để cân nhắc các khía cạnh khác của dự án như vấn đề tác động môi trường, xã hội, pháp lý, đối tác chiến lược, sự khan hiếm của nguồn lực, v.v... Đối với một sản phẩm mới, có thể doanh nghiệp sẽ không có sẵn các dữ liệu về doanh số để có thể dự báo doanh số trong tương lai. Tương tự như vậy, các dữ liệu về doanh số quá khứ của một sản phẩm sẽ không còn phù hợp nếu một đối thủ cạnh tranh trực tiếp tung ra một sản phẩm mới với một đặc tính nào đó ưu việt hơn so với sản phẩm của công ty. Trong nhiều trường hợp khác, có thể doanh nghiệp không có đủ thời gian để thu thập dữ liệu hoặc sử dụng các kỹ thuật dự báo định lượng, hoặc tình hình ứng dụng công nghệ thông tin phát triển một cách quá nhanh chóng đến nỗi một kết quả dự báo trên cơ sở thống kê có thể không còn là một kết quả tin cậy. Các phương pháp định tính đôi khi cần thiết vì không đòi hỏi những người liên quan phải có kiến thức về các mô hình toán, mô hình thống kê hoặc kinh tế lượng. Ngoài ra, hiện nay các phương pháp định tính đang được chấp nhận rộng rãi nên ở nhiều nơi và nhiều lĩnh vực vẫn còn sử dụng khá phổ biến. Thậm chí khi có sẵn các kỹ thuật thống kê, thì phán đoán cá nhân vẫn là sự lựa chọn ưu tiên của nhiều nhà quản lý cấp cao. Tuy nhiên, kết quả dự báo định tính phụ thuộc vào ý kiến chủ quan nên có thể bị sai lệch, không chính xác một cách ổn định qua thời gian, không có phương pháp hệ thống để đánh giá và cải thiện mức độ chính xác, và đòi hỏi người tham gia phải mất nhiều thời gian để tích lũy kinh nghiệm về một lĩnh vực nhất định. Điều quan trọng cần lưu ý là, để có các quyết định sáng suốt, thì người sử dụng kết quả dự báo cần kết hợp giữa kết quả dự báo định lượng và định tính. Thông thường, các phán đoán cá nhân trong việc thực hiện nhiều dự báo có thể được biện hộ thông qua hai cách sau đây. Thứ nhất, so với các mô hình thống kê, con người có thể có khả năng phát hiện các xu hướng thay đổi trong chuỗi thời gian một cách tốt hơn vì các phán đoán đó đặt vấn đề dự báo trên một bình diện rộng hơn. Thứ

hai, con người có khả năng kết hợp các thông tin bên ngoài (ngoài bản thân chuỗi thời gian) vào quá trình dự báo.

Dayananda (2002) đã chia dự báo định tính thành hai nhóm. Thứ nhất, các phương pháp thu thập thông tin dự báo từ các cá nhân liên quan đến đối tượng dự báo. Các phương pháp này bao gồm khảo sát thị trường và tổng hợp lực lượng bán hàng. Thứ hai, các phương pháp dựa vào ý kiến của các nhóm chuyên gia am hiểu về lĩnh vực cần dự báo. Các phương pháp này bao gồm ý kiến ban quản lý, phương pháp Delphi, kỹ thuật nhóm định danh, và các kỹ thuật khác.

Thu thập thông tin cá nhân

Một khi đã quyết định dự báo định tính, người phân tích thường đối diện với nhiều sự lựa chọn khác nhau. Ví dụ, người phân tích phải quyết định nên ước lượng từ thông tin cá nhân hay thông tin nhóm, và sẽ thu thập thông tin như thế nào.

Thông tin hoặc ước lượng từ cá nhân có thể được thu thập theo các cách sau đây. Nếu các yêu cầu thông tin đơn giản, chẳng hạn ước lượng sự hiệu quả sử dụng nhiên liệu của một loại phương tiện, thì một cuộc gọi điện thoại hoặc gửi thư điện tử đến một chuyên gia thích hợp là được. Tuy nhiên, trong nhiều trường hợp các yêu cầu thông tin thường phức tạp và đòi hỏi có sự suy luận về phán đoán của chuyên gia. Trong các trường hợp như vậy, các kỹ thuật thu thập dữ liệu chính thức hơn cần phải được áp dụng. Và thông thường chúng ta phải tiến hành điều tra.

Phương pháp điều tra

Các dự báo dựa vào quá trình điều tra có thể bao gồm việc xây dựng và quản lý công cụ điều tra, thường là bảng câu hỏi phỏng vấn và phân tích dữ liệu thu thập được. Quy trình điều tra thông thường gồm một số bước khác nhau nhưng có liên hệ chặt chẽ với nhau, như được minh họa ở Hình 1.2. Các quyết định ở các giai đoạn đầu có thể ảnh hưởng đến sự lựa chọn ở các giai đoạn sau của quá trình. Ví dụ, các nhu cầu thông tin được xác định ở bước đầu tiên có thể ảnh hưởng

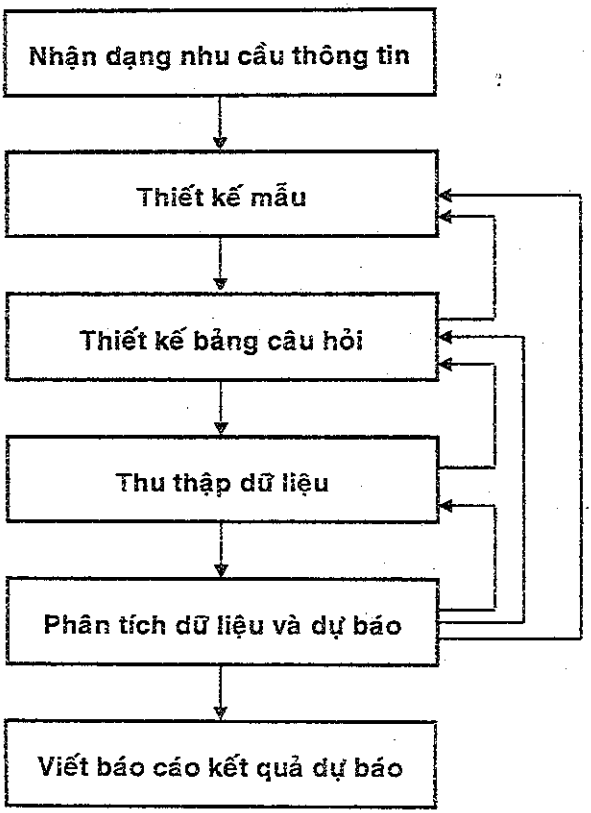
đến việc lựa chọn thiết kế mẫu điều tra, cách thiết kế bảng câu hỏi và lựa chọn các kỹ thuật phân tích dữ liệu.

Số lượng thông tin thu thập hầu như là vô hạn. Tuy nhiên, do hạn chế về mặt thời gian và nguồn lực, nên cần phải ưu tiên thu thập các thông tin cần thiết. Các nhu cầu thông tin có thể được phân loại thành ba mức độ quan trọng: (1) các thông tin chủ yếu vốn là lý do cần thực hiện điều tra, (2) các thông tin có giá trị cao cho các quyết định quan trọng, và (3) các thông tin bổ trợ. Một phần rất quan trọng của bất kỳ cuộc điều tra nào là phải xác định đúng nhóm đối tượng cần khảo sát. Tùy vào mục tiêu dự báo, chúng ta sẽ xác định nhóm cần khảo sát như nhân viên trong công ty, khách hàng, hay các chuyên gia. Một khi đã xác định nhóm đối tượng cần khảo sát, chúng ta cần xác định cỡ mẫu và phương pháp chọn mẫu thích hợp để đảm bảo tính đại diện. Để có bảng câu hỏi thích hợp, chúng ta cần thiết kế một số bảng câu hỏi nháp, đi phỏng vấn thử, rồi chỉnh sửa trước khi tiến hành khảo sát chính thức. Trước khi thực hiện khảo sát, chúng ta cần thống nhất để chọn phương pháp khảo sát hiệu quả nhất. Để biết thêm chi tiết về phương pháp khảo sát thị trường, chúng ta có thể tham khảo các tài liệu về nghiên cứu marketing hoặc phương pháp nghiên cứu.

Thông tin từ các cuộc điều tra nghiên cứu thị trường thứ cấp

Các cuộc điều tra khách hàng, thường nhằm nhận dạng khuynh hướng mua sắm của khách hàng trong tương lai, rất phổ biến và có thể cung cấp nguồn dữ liệu hữu ích cho công tác dự báo. Những người làm dự báo có thể dựa vào dữ liệu từ các cuộc điều tra do các tổ chức nghiên cứu thị trường thực hiện để phục vụ việc dự báo của mình. Từ những dữ liệu này, người làm dự báo có thể dự đoán được khả năng tăng hoặc giảm trong doanh số về một sản phẩm của công ty, và điều này rất hữu ích cho quá trình dự báo và ra quyết định.

HÌNH 1.2: Quy trình dự báo bằng khảo sát thị trường.



Nguồn: Dayananda, 2002, trang 57.

Tổng hợp lực lượng bán hàng

Phương pháp này thường được sử dụng trong các lĩnh vực sản xuất và bán lẻ để dự báo doanh số. Phương pháp này liên quan đến quan điểm của các nhân viên bán hàng riêng lẻ và quản lý bán hàng để dự đoán xu hướng doanh số của toàn công ty. Phương pháp này có thể được thực hiện theo ba cách tiếp cận khác nhau: (1) nhân viên bán hàng cơ sở, (2) quản lý bán hàng, và (3) nhà phân phối. Lực lượng bán hàng có thể là một nguồn thông tin phong phú về những xu hướng và thay

đổi trong tương lai về hành vi khách hàng. Dự báo có thể được thực hiện bởi những thành viên trong lực lượng bán hàng của một công ty. Những thành viên này sẽ sử dụng những đánh giá chủ quan về những điều mà họ mong đợi từ phía khách hàng sẽ mua trong tương lai. Dự báo tổng hợp một biến nào đó dựa trên dự báo những sản phẩm riêng lẻ trong những khu vực cụ thể. Nghĩa là, từ những dự báo về tình hình tiêu thụ những sản phẩm riêng lẻ sẽ được tổng hợp để dự báo tổng doanh số của một doanh nghiệp. Lợi thế của phương pháp này thể hiện ở chỗ những nhân viên bán hàng sẽ là người hiểu rõ nhất khách hàng của họ muốn gì trong tương lai. Tuy nhiên, hạn chế lớn nhất của phương pháp này là những đánh giá của các nhân viên bán hàng có thể sai lệch một cách chủ ý (thường bị ước lượng thấp) để có lợi cho họ nhờ chính sách thưởng dựa trên thành quả của công ty, và điều này có thể trái với kỳ vọng của ban giám đốc công ty.

Sử dụng các nhóm chuyên gia

Nhiều bằng chứng cho thấy các dự báo tạo ra từ các nhóm chuyên gia có độ chính xác cao hơn nhiều so với các dự báo từ một cá nhân. Các nhóm cũng cung cấp nhiều thông tin hơn, mặc dù lượng thông tin tăng thêm từ mỗi cá nhân sẽ giảm khi quy mô nhóm tăng lên. Sử dụng nhóm cũng cung cấp cơ hội để có nhiều thông tin hơn về nhiều kết quả dự báo có thể xảy ra hơn, và điều này giúp nhận diện được yếu tố rủi ro trong dự báo. Ngoài ra, kết quả dự báo nhóm làm gia tăng sự cam kết trong quá trình thực hiện và sử dụng kết quả dự báo. Hai phương pháp thường được sử dụng là đánh giá ý kiến ban quản trị và phương pháp Delphi.

Đánh giá ý kiến ban quản trị

Các phán đoán của chuyên gia ở bất kỳ một lĩnh vực nào đều là một nguồn lực quý giá. Dựa vào kinh nghiệm trong từng lĩnh vực mà các phán đoán có thể rất hữu ích cho quy trình dự báo. Dự báo bằng cách đánh giá ý kiến ban quản trị có thể được thực hiện bằng cách kết hợp ý kiến chủ quan của trưởng các bộ phận và những người quản lý vì họ là những người có am hiểu rõ nhất về tình hình kinh doanh và mục

tiêu của doanh nghiệp. Để tiến hành dự báo, người làm dự báo phải lựa chọn những chuyên gia ở nhiều lĩnh vực chuyên môn chức năng khác nhau như tiếp thị, nhân sự, sản xuất, tài chính và kinh tế để thu thập ý kiến thông qua các cuộc phỏng vấn riêng hoặc thông qua các cuộc họp và thảo luận.

Phương pháp Delphi

Từ những ứng dụng đầu tiên cho việc dự báo xu hướng thay đổi công nghệ vào những năm 1950, phương pháp Delphi được phát triển nhanh chóng trong nhiều lĩnh vực khác nhau như y tế, chăm sóc sức khỏe, giáo dục, sản phẩm mới, ảnh hưởng của toàn cầu hóa, v.v... Linstone và Turoff (1975) đã đưa ra một định nghĩa về phương pháp Delphi như sau:

> Delphi có thể được xem như một phương pháp giúp thiết lập một quá trình trao đổi thông tin nhóm một cách hiệu quả nhằm cho phép các thành viên trong nhóm giải quyết một vấn đề phức tạp.

Nhìn chung, phương pháp Delphi tương tự như phương pháp đánh giá ý kiến ban quản lý vì cũng dựa vào ý kiến của các chuyên gia có am hiểu lĩnh vực cần dự báo, nhưng lại khác ở cách thức tiến hành, và nhờ đó mà kết quả cuối cùng sẽ khách quan và tin cậy hơn. Quy trình thực hiện dự báo theo phương pháp Delphi bao gồm nhiều vòng, nhưng thường theo các bước sau đây:

(1) Xác định mục tiêu dự báo.

(2) Lựa chọn nhóm chuyên gia.

(3) Thiết lập bảng câu hỏi trưng cầu ý kiến về các biến dự báo và gửi đến từng thành viên trong nhóm chuyên gia (không yêu cầu khai báo tên).

(4) Các kết quả phản hồi từ mỗi chuyên gia được thu thập, lập bảng, và tổng hợp thành một báo cáo tóm tắt.

(5) Báo cáo tóm tắt kết quả sẽ được gởi trở lại các chuyên gia để lấy ý kiến nhận xét (lưu ý, tóm tắt này nên nhấn mạnh những ý kiến trái ngược, cực đoan, đặc biệt (khác với đa số)).

(6) Những chuyên gia có thể sẽ hiệu chỉnh lại các ước lượng lần trước của họ sau khi có xem xét thông tin nhận được từ những thành viên (không biết tên) khác.

(7) Lặp lại bước (3) đến bước (5) cho đến khi không còn sự thay đổi đáng kể nào (đi đến thống nhất). Lưu ý, cũng sẽ có trường hợp có vài chuyên gia không thay đổi ý kiến của họ trong quá trình thăm dò, điều này sẽ khó tìm một kết quả dự báo tập trung.

Mặc dù, Delphi là một phương pháp dự báo có lịch sử phát triển lâu đời, được áp dụng khá phổ biến và có độ tin cậy cao, nhưng nó vẫn tồn tại nhiều hạn chế đáng kể. Penelope (2003) cho rằng phương pháp Delphi thường bị chỉ trích ở những điểm liên quan đến nhóm chuyên gia, sự đồng thuận, xây dựng bảng câu hỏi, tình trạng nặc danh, và tương tác giữa các thành viên trong nhóm chuyên gia.

Chất lượng dự báo theo phương pháp Delphi phụ thuộc rất nhiều vào cách thức mà phương pháp đó được áp dụng như thế nào. Parente, Anderson, Myers và O'Brien (1984) đề xuất cần lưu ý các vấn đề sau đây:

- Tiêu chí lựa chọn nhóm chuyên gia (trình độ học thuật, kinh nghiệm) nên được xác định một cách cẩn thận và phải được trao đổi một cách rõ ràng.
- Tối thiểu nhóm có 10 chuyên gia, mặc dù đôi khi con số 5 chuyên gia hoặc ít hơn.
- Cam kết phục vụ trong nhóm chuyên gia phải được bảo đảm trước khi bắt đầu vòng dự báo thứ nhất.

- Một số các vấn đề dự báo có thể được trình bày, mặc dù số vấn đề này phải ít hơn 25. Nếu có thể, nội dung dự báo chính nên được chia thành nhiều vấn đề nhỏ.

- Các phát biểu vấn đề nên không được dài hơn 20 từ và nên sử dụng dữ liệu định lượng (ví dụ, tăng 50%) thay vì sử dụng một ngôn ngữ mơ hồ (ví dụ, tăng đáng kể).

- Các hướng dẫn trong việc thiết kế bảng câu hỏi phải được áp dụng khi trình bày các vấn đề. Lưu ý, không nên dùng các câu phức.

- Nếu mục đích của quá trình Delphi là phải tạo ra các vấn đề dự báo, thì đề nghị đưa ra các ví dụ về các kịch bản hấp dẫn và các kịch bản không như mong muốn.

- Cho dù sử dụng các phương tiện nào để quản lý quy trình Delphi – gửi thư, máy tính nối mạng, hoặc họp mặt trực tiếp – thì các bước tương tự như vậy phải được thực hiện trong suốt quá trình.

- Nguyên tắc ẩn danh nên được đảm bảo. Quan điểm của người tổ chức dự báo không bao giờ được trao đổi với các chuyên gia.

- Số lượng và mẫu phản hồi sẽ cần được quản lý một cách cẩn thận. Số lượng các vòng dự báo sẽ phụ thuộc vào các chuyên gia và cách thức mà tiến hành khảo sát Delphi. Lưu ý, nhiều vòng luôn luôn tốt hơn quá ít vòng, và phản hồi dưới dạng thống kê mô tả thường cung cấp thông tin tốt hơn.

- Các phản ứng cao quá hoặc thấp quá cần được xem xét lại để kiểm tra khả năng chuyên môn của chuyên gia. Nếu chuyên gia có chuyên môn tương đối thấp, thì ý kiến phản hồi nên có trọng số thấp hơn các chuyên gia khác.

- Nếu khảo sát Delphi nhằm mục đích ứng dụng nghiên cứu, thì một bản báo cáo chi tiết quá trình thực hiện (ngoài kết quả dự

báo) nên được công bố để những người nghiên cứu khác có thể học tập kinh nghiệm sau này.

Các phương pháp định lượng

Các phương pháp định lượng dựa vào các mô hình toán và giả định rằng dữ liệu quá khứ cũng như các yếu tố liên quan khác có thể được kết hợp để đưa ra các dự đoán tin cậy cho tương lai. Nói cách khác, dựa trên những dữ liệu quá khứ để phát hiện chiều hướng vận động của đối tượng phù hợp với một mô hình toán học nào đó và đồng thời sử dụng mô hình này là mô hình ước lượng. Tiếp cận định lượng dựa trên giả định rằng giá trị tương lai của biến dự báo sẽ phụ thuộc vào xu thế vận động của đối tượng đó trong quá khứ. Các phương pháp dự báo định lượng được chia thành hai nhóm: các mô hình chuỗi thời gian và các mô hình nhân quả (Hình 1.3).

Các mô hình dự báo chuỗi thời gian nghĩa là dự báo giá trị tương lai của một biến nào đó chỉ bằng cách phân tích số liệu quá khứ và hiện tại của chính biến số đó. Giả định chủ yếu là trong tương lai biến số dự báo sẽ giữ nguyên chiều hướng vận động đã xảy ra trong quá khứ và hiện tại. Chỉ có các chuỗi dữ liệu có tính ổn định thì mới có thể cho ra các dự báo tin cậy. Chính vì thế, như chúng ta sẽ biết, tính "dừng" là một điều kiện quan trọng nhất trong việc phân tích và dự báo chuỗi thời gian. Cho nên, trước khi xác định mô hình định lượng nào phù hợp với dữ liệu, người làm dự báo cần khảo sát dữ liệu một cách cẩn thận. Nội dung phân tích dữ liệu chuỗi thời gian sẽ được trình bày một cách chi tiết ở chương 3 và nhắc lại ở từng phương pháp cụ thể. Các công cụ phân tích dữ liệu thường sử dụng đối với chuỗi thời gian là vẽ đồ thị theo thời gian, giản đồ tự tương quan, và kiểm định nghiệm đơn vị. Giáo trình này sẽ trình bày các phương pháp chuỗi thời gian theo các nhóm sau đây: (1) Các mô hình dự báo giản đơn, (2) Các mô hình phân tích thành phần của chuỗi thời gian, (3) Các mô hình hàm xu thế tuyến tính, (4) Các mô hình ARIMA, và (5) Các mô hình ARCH/GARCH. Mặc dù, tùy thuộc vào bản chất dữ liệu sẵn có mà ta chọn mô hình dự báo thích hợp nhất thông qua các tiêu chính đánh giá dự báo, nhưng các mô hình ARIMA và

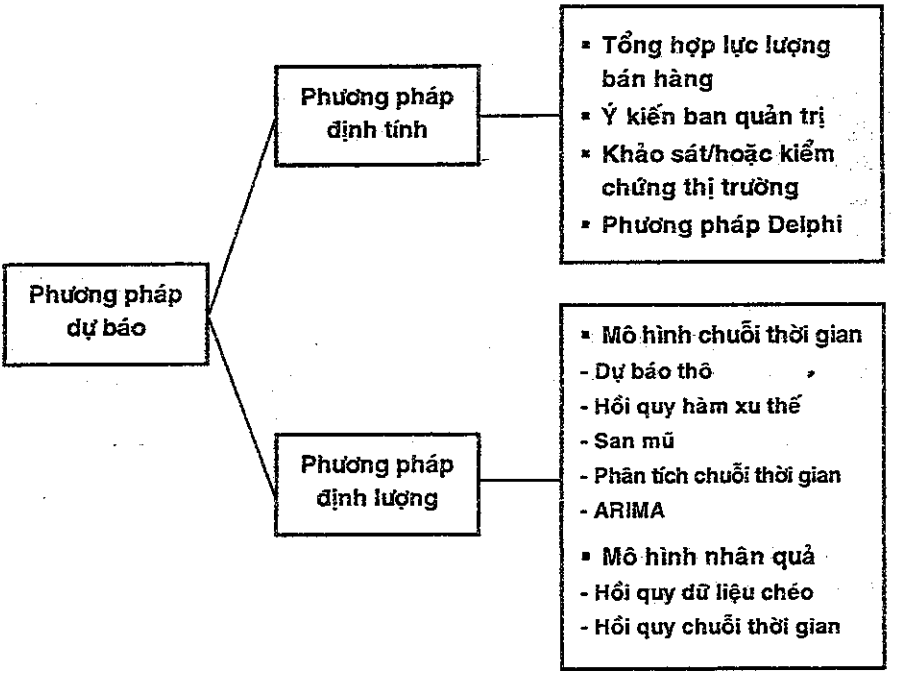
ARCH/GARCH đang được sử dụng khá phổ biến, nhất là trong việc dự báo tài chính và các chỉ báo kinh tế vĩ mô.

Ngược lại, các mô hình dự báo nhân quả dựa trên phân tích hồi quy. Chính vì vậy, chúng ta cần có kiến thức nền tảng nhất định về kinh tế lượng và thống kê để có thể dễ dàng tiếp cận và vận dụng các mô hình dự báo nhân quả. Trong giáo trình này, chúng tôi sẽ chia thành hai nhóm mô hình nhân quả sau đây: (1) Các mô hình kinh tế lượng về dữ liệu chéo, và (2) Các mô hình kinh tế lượng về chuỗi thời gian. Chúng tôi phân thành hai loại mô hình như vậy là vì mỗi loại dữ liệu có những tính chất và ứng dụng khác nhau, nên đòi hỏi chúng ta phải có cách tiếp cận khác nhau trong việc phân tích và kiểm định để đảm bảo có các mô hình dự báo thích hợp nhất cho từng trường hợp. Tuy nhiên, các mô hình nhân quả chuỗi thời gian sẽ được chúng tôi đề cập nhiều hơn.

Nhìn chung, phương pháp chuỗi thời gian và phương pháp nhân quả có những lợi thế sau đây:

- Một khi biến độc lập đã được chọn, dự báo sẽ hoàn toàn dựa trên những giá trị có trước của những biến số như vậy, do đó kết quả dự báo sẽ hoàn toàn khách quan.
- Có những phương pháp để đo lường độ chính xác dự báo nên có thể dễ dàng so sánh và lựa chọn mô hình dự báo tốt nhất.
- Khi mô hình dự báo đã được xây dựng, sẽ ít tốn thời gian để tìm ra kết quả dự báo.
- Những phương pháp dự báo này sẽ có thể dự báo điểm (một giá trị cụ thể) hay dự báo khoảng (một số giá trị trong khoảng tin cậy).

■ HÌNH 1.3: **Phân loại phương pháp dự báo.**



Nguồn: Wilson, 2007.

Tuy nhiên, các phương pháp định lượng cũng có một số hạn chế. Thứ nhất, cả mô hình chuỗi thời gian và mô hình nhân quả chỉ dự báo tốt trong ngắn hạn và trung hạn. Những nhà quản lý thông thường sẽ cân nhắc độ chính xác dự báo thông qua chiều dài chuỗi dữ liệu sẵn có và khoảng cách dự báo yêu cầu. Thứ hai, không có phương pháp nào có thể đưa đầy đủ những yếu tố bên ngoài có tác động đến kết quả dự báo vào mô hình.

Các chuyên gia dự báo cho rằng một phương pháp dự báo tốt thông thường sẽ phải kết hợp những phương pháp định lượng và định tính. Phương pháp định lượng sẽ dễ vướng phải sai lầm khi giả định rằng sự kiện tương lai sẽ phản ánh theo những hành vi quá khứ, do đó

kết quả dự báo của những phương pháp định lượng cần phải thông qua ý kiến đánh giá của chuyên gia thuộc những lĩnh vực liên quan.

Điều cần phải xem xét kỹ khi lựa chọn một phương pháp dự báo là các kết quả dự báo phải giúp quá trình ra quyết định của doanh nghiệp trở nên dễ dàng hơn. Rất hiếm khi áp dụng một phương pháp duy nhất cho tất cả các trường hợp. Người làm công tác dự báo phải xem xét các mục tiêu của doanh nghiệp cũng như những điều kiện ràng buộc nhất định khi quyết định lựa chọn phương pháp dự báo. Nhờ sự phát triển của máy tính và các phần mềm chuyên dụng nên các phương pháp dự báo định lượng ngày càng được sử dụng rộng rãi. Quyết định lựa chọn phương pháp dự báo thích hợp tùy thuộc vào ba yếu tố sau đây: mục tiêu dự báo dài hạn, trung hạn hay ngắn hạn; số lượng dữ liệu mà người dự báo có được; và bản chất dữ liệu là dừng hay không dừng.

PHƯƠNG PHÁP LUẬN CỦA DỰ BÁO ĐỊNH LƯỢNG

Một điểm khác biệt quan trọng nữa giữa các phương pháp định lượng với các phương pháp định tính là ở phương pháp luận của chúng. Hầu như không có một phương pháp luận cụ thể nào cho tất cả các phương pháp định tính vì mỗi phương pháp khác nhau có thể có cách thực hiện khác nhau. Ngược lại, nhờ có phương pháp luận rõ ràng nên các phương pháp định lượng trở nên đáng tin cậy hơn và ngày càng được các doanh nghiệp sử dụng phổ biến hơn.

Do phương pháp định lượng được chia thành hai nhóm khác nhau, nên phương pháp luận của chúng về cơ bản cũng khác nhau. Chính vì vậy, giáo trình này sẽ trình bày cả phương pháp luận của các mô hình dự báo chuỗi thời gian và phương pháp luận của các mô hình nhân quả. Nhìn chung, tất cả các phương pháp định lượng có điểm chung như sau: Xác định mục tiêu cần dự báo, thu thập và khảo sát dữ liệu để lựa chọn mô hình dự báo thích hợp, đánh giá mô hình, và dự báo dựa trên mô hình tốt nhất.

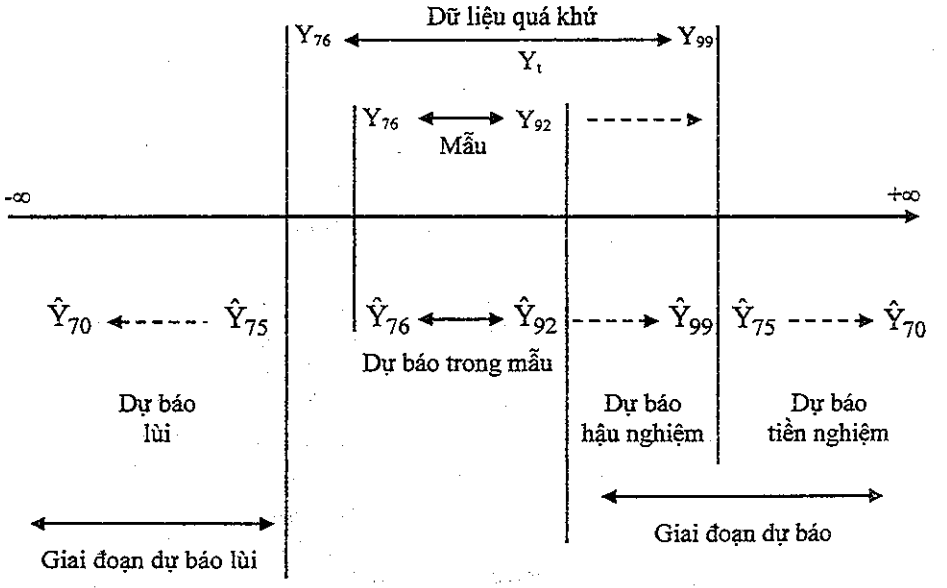
Phương pháp luận của dự báo chuỗi thời gian

Dự báo các giai đoạn quá khứ được gọi là dự báo hậu nghiệm và dự báo các giai đoạn tương lai được gọi là dự báo tiền nghiệm. Để hiểu rõ hơn về các loại dự báo này trong quy trình dự báo, chúng ta phân tích sơ đồ ở Hình 1.4.

- Dữ liệu lịch sử được cung cấp từ thời đoạn Y_{BEG} (Y_{76}) tới Y_{END} (Y_{99}). Chúng ta định nghĩa Y_{BEG} là thời đoạn bắt đầu của chuỗi thời gian. Y_{END} là thời đoạn mới nhất của chuỗi thời gian thu thập. Đối với chúng ta Y_{END} có thể là quan sát hiện tại.

- Dữ liệu mẫu phân tích, $Y_1, \ldots Y_n$, là những quan sát mà chúng ta sử dụng để xây dựng mô hình dự báo. Giữa Y_{BEG} và Y_1 không nhất thiết trùng khớp lẫn nhau.

■ HÌNH 1.4: **Các giai đoạn của chuỗi thời gian.**



Nguồn: Nguyễn Trọng Hoài, 2003.

- Tương ứng với giai đoạn ước lượng $Y_1, \ldots . Y_n$, những giá trị dự báo $\hat{Y}_1 \ldots\ldots\ldots\ldots \hat{Y}_n$.

  Những giá trị dự báo này được tìm trong mô hình hay trong mẫu dữ liệu khi tiến hành dự báo. Từ những giá trị thực tế và giá trị dự báo chúng ta có thể xác định sai số dự báo $e_1, \ldots . , e_n$ (với $e_n = Y_n - \hat{Y}_n$) cho mô hình (trong giai đoạn ước lượng), từ đó độ chính xác của mô hình có thể xác định. Tất cả những giá trị vượt ra ngoài Y_n phải là giá trị dự báo. Trong khuôn khổ của đường thời gian, những giá trị dự báo sẽ nằm trong giai đoạn ước lượng. Tất cả những giá trị dự báo hình thành trong giai đoạn dự báo được gọi là dự báo ngoài phạm vi mẫu bởi vì nó xuất hiện sau khi chấm dứt giai đoạn ước lượng.

- Toàn bộ giai đoạn dự báo sẽ được phân chia thành hai bộ phận phân biệt là dự báo hậu nghiệm (ex-post) và dự báo tiền nghiệm (ex-ante).

  - Giai đoạn dự báo hậu nghiệm là thời gian từ quan sát đầu tiên sau khi chấm dứt giai đoạn mẫu $\hat{Y}_{n+1}$ tới quan sát mới nhất $\hat{Y}_N$. Đặc trưng quan trọng trong giai đoạn này là nhà nghiên cứu đã có giá trị thực tế của đối tượng dự báo Y_t. Giai đoạn hậu nghiệm sẽ cung cấp cho nhà nghiên cứu cơ hội đánh giá mức độ chính xác của mô hình dự báo trong giai đoạn này bằng cách sử dụng chênh lệch giữa giá trị thực tế và giá trị dự báo hậu nghiệm. Nếu như độ chính xác của mô hình không thỏa mãn thì lúc đó nhà nghiên cứu sẽ lựa chọn hai giải pháp: tìm kiếm một mô hình thay thế với độ chính xác cao hơn hoặc mở rộng giai đoạn mẫu bao gồm cả những quan sát trong giai đoạn hậu nghiệm đang xét. Nếu nhà nghiên cứu mở rộng giai đoạn ước lượng tới hiện tại thì dự báo trong phạm vi mẫu sẽ hình thành từ $\hat{Y}_1 \ldots\ldots\ldots\ldots \hat{Y}_N$. Những quan sát giai đoạn mẫu và giá trị dự báo đã mở rộng trong giai đoạn hậu nghiệm được minh họa

bằng đường không liên tục trong hình vẽ. Chú ý rằng lúc này giai đoạn dự báo không bao gồm giá trị dự báo hậu nghiệm.

- Giai đoạn dự báo tiền nghiệm là giai đoạn không có giá trị thực tế về đối tượng dự báo (hay bất kỳ những biến số ảnh hưởng khác). Đây chính là dự báo cho tương lai. Chúng ta ký hiệu những dự báo tiền nghiệm là $\hat{Y}_{N+1}$.....$\hat{Y}_{N+K}$. Bởi vì trong giai đoạn này không có giá trị thực tế của đối tượng dự báo do đó sẽ không xác định được độ chính xác của những dự báo tiền nghiệm.

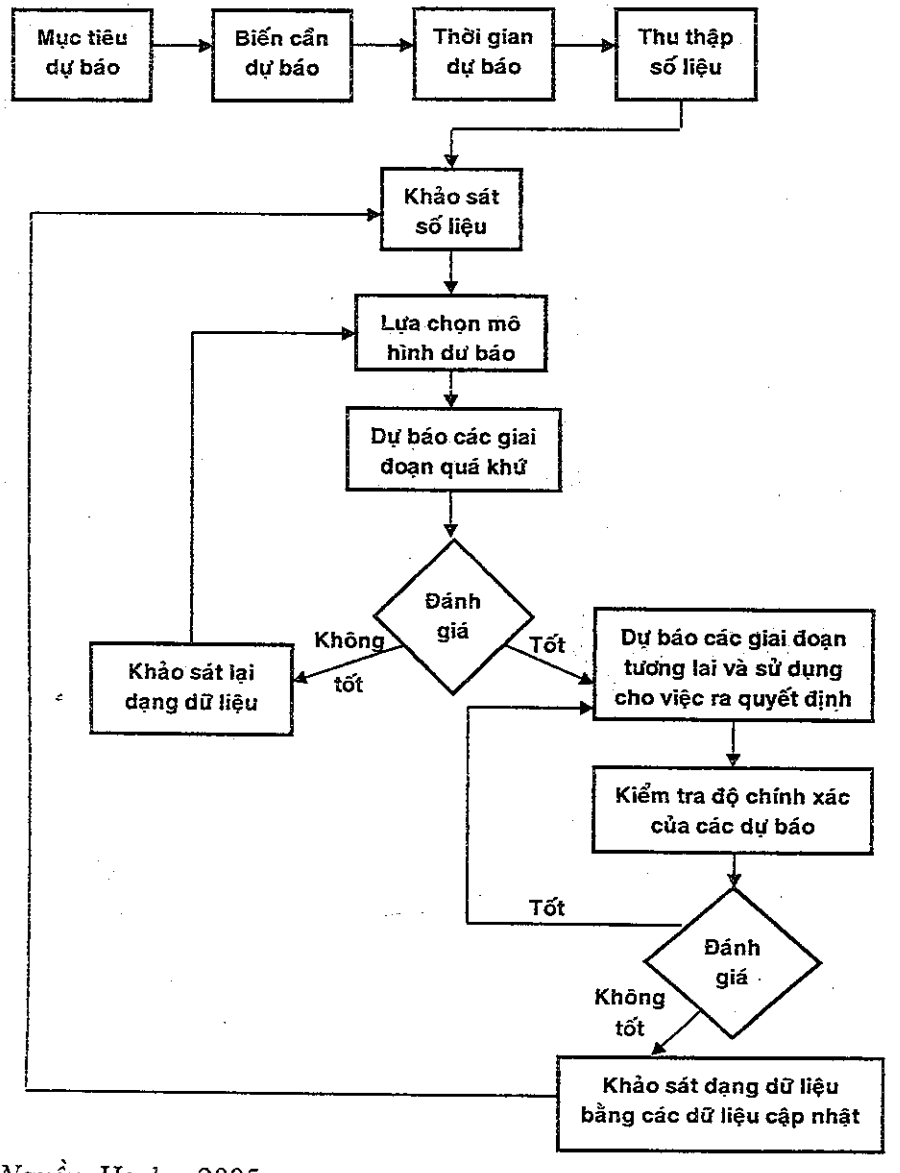
Phía dưới đường thời gian chúng ta mô tả một giai đoạn phía trước Y_{BEG} mà dữ liệu mô tả hầu như không có. Chúng ta có thể dự báo lùi (backcast) cho những thời đoạn trước Y_{BEG}. Chúng ta có thể sử dụng những dự báo lùi nhằm đạt những giá trị bổ sung cho thời đoạn lịch sử trong quá trình phân tích.

Phương pháp luận cho các mô hình dự báo chuỗi thời gian có thể được thể hiện như ở Hình 1.5.

Phương pháp luận của dự báo nhân quả

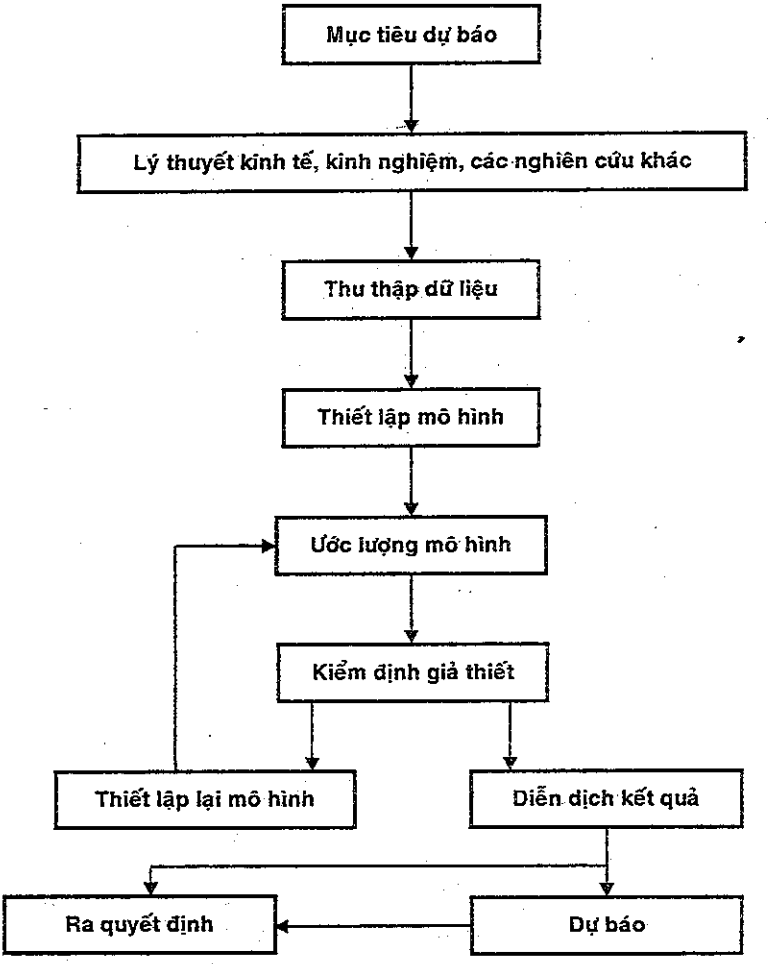
Khi dự báo bằng mô hình nhân quả, xuất phát từ mục tiêu dự báo; người làm dự báo cần dựa trên các lý thuyết kinh tế, các nghiên cứu thực nghiệm có liên quan, kinh nghiệm của các chuyên gia trong ngành, v.v..., để từ đó xác định các biến số (biến giải thích) có thể ảnh hưởng đến biến cần dự báo (biến phụ thuộc). Sau đó mới tiến hành thu thập dữ liệu; xây dựng, ước lượng mô hình; kiểm định giả giả thuyết, và thực hiện dự báo. Nói chung, phương pháp luận của dự báo bằng các mô hình nhân quả có thể được minh họa như ở Hình 1.6.

HÌNH 1.5: **Phương pháp luận của dự báo chuỗi thời gian.**



Nguồn: Hanke, 2005.

■ HÌNH 1.6: **Phương pháp luận của dự báo nhân quả.**



Nguồn: Ramanathan, 1998.

QUY TRÌNH THỰC HIỆN DỰ BÁO ĐỊNH LƯỢNG

Quy trình dự báo bắt đầu bằng bằng việc nhận ra nhu cầu ra các quyết định vốn phụ thuộc nhiều vào giá trị một số biến chưa biết trong tương lai. Điều quan trọng đối với người quản lý và nhà hoạch định chính sách - những người trực tiếp sử dụng kết quả dự báo trong việc ra các quyết định là họ cần phải hiểu các phương pháp đã được sử dụng trong quá trình dự báo. Ngoài ra, việc những người làm dự báo nhận thức được những người ra quyết định đang cần gì từ kết quả dự báo cũng không kém phần quan trọng. Như vậy, sự trao đổi giữa những người có liên quan trong suốt quy trình dự báo có ý nghĩa hết sức quan trọng trên cả phương diện mức độ chính xác và khả năng ứng dụng. Nói chung, một quy trình dự báo thường gồm các bước sau đây.

Xác định mục tiêu

Các mục tiêu liên quan đến các quyết định cần dựa vào kết quả dự báo nên được xác định rõ ràng. Những nhà quản trị và hoạch định chính sách cần khẳng định vai trò mà dự báo sẽ có trong quá trình ra quyết định. Nếu quyết định vẫn không thay đổi bất kể kết quả dự báo như thế nào thì bất kỳ nỗ lực nào cho việc chuẩn bị thực hiện dự báo đều trở nên lãng phí. Điều này có vẻ quá hiển nhiên, nhưng thực tế thường xảy ra do người quản lý hoặc hoạch định chính sách không hiểu và không tin vào dự báo. Chính vì thế, sự hiểu biết và tin tưởng nhất định phải là điều quan trọng nhất trong bất kỳ quy trình dự báo nào. Nếu nhà quản trị cần thông tin từ dự báo và người làm dự báo có cơ hội bàn bạc các mục tiêu sẽ sử dụng kết quả dự báo, thì nhà quản trị sẽ hiểu và tin tưởng sử dụng kết quả dự báo hơn.

Xác định biến số cần dự báo

Một khi các mục tiêu tổng quát đã được xác định rõ, người làm dự báo phải xác định chính xác sẽ dự báo cái gì. Ví dụ, nếu chúng ta chỉ nói dự báo doanh số thì chưa đủ vì cụ thể chúng ta muốn dự báo doanh thu hay sản lượng tiêu thụ, hoặc dự báo theo năm, quý hay

tháng. Thông thường ta nên dự báo doanh số theo sản lượng tiêu thụ hơn là đơn vị tiền tệ vì như thế sẽ tránh được những biến động do sự thay đổi giá cả. Cho nên, tốt nhất người làm dự báo phải bàn bạc thật kỹ với người sử dụng kết quả dự báo để quyết định chọn biến nào cần dự báo.

Nhận dạng các khía cạnh thời gian

Về mặt thời gian trong dự báo có hai vấn đề cần xem xét. Thứ nhất, người làm dự báo phải xác định độ dài dự báo. Đối với các dự báo theo năm, thì độ dài thời gian có thể từ 1 đến 5 năm hoặc hơn, mặc dù các dự báo có độ dài thời gian dài sẽ kém chính xác do những sự kiện không đoán trước không thể được đưa vào mô hình. Các dự báo theo quý có thể dự báo cho khoảng 1 đến 2 năm (4 đến 8 quý) và dự báo theo tháng có thể từ 12 đến 18 tháng. Thứ hai, người sử dụng và người làm dự báo phải thống nhất với nhau về tính cấp thiết của dự báo vì nó có ảnh hưởng đến việc chọn mô hình dự báo và kế hoạch tiến hành dự báo.

Thu thập và phân tích dữ liệu

Dữ liệu cần thiết cho dự báo có thể thu thập từ nội bộ doanh nghiệp hoặc từ các nguồn bên ngoài. Nhiều người vẫn nghĩ rằng dữ liệu nội bộ doanh nghiệp luôn có sẵn và dễ dàng đưa vào quy trình dự báo. Tuy nhiên, thực tế không phải như thế. Các dữ liệu có sẵn và dữ liệu cần cho việc dự báo có thể hoàn toàn khác nhau ở nhiều phương diện như đơn vị tính, thời gian, cách thức tổng hợp, v.v... Dữ liệu tốt nhất là được lưu giữ dưới dạng chưa tổng hợp (ví dụ dạng dữ liệu theo quý, tháng hơn là theo năm). Dữ liệu bên ngoài có thể được thu thập bằng nhiều nguồn khác nhau, nhưng cần phải có sự thống nhất giữa người làm dự báo và người sử dụng để thống nhất nên chọn từ những nguồn nào.

Có nhiều cách khảo sát dữ liệu của biến số cần dự báo như sẽ được trình bày ở các chương sau. Nhưng đối với các chuỗi thời gian thì việc xem xét một chuỗi dừng hay không dừng có ý nghĩa hết sức

quan trọng vì nó quyết định dạng mô hình thích hợp (đối với các mô hình chuỗi thời gian) và các kiểm định cần thiết trước khi có thể sử dụng kết quả dự báo (mô hình kinh tế lượng).

Lựa chọn mô hình

Đối với phương pháp định lượng thì việc lựa chọn mô hình thích hợp tùy thuộc vào các khía cạnh sau đây:

- Loại và số lượng dữ liệu sẵn có
- Dạng dữ liệu được thể hiện trong quá khứ (ví dụ chuỗi dừng hay không dừng, có yếu tố mùa vụ hay không)
- Tính cấp bách của dự báo
- Độ dài thời gian dự báo
- Năng lực và kiến thức về dự báo của cả người làm và người sử dụng dự báo

Đối với các phương pháp dự báo định tính thì việc lựa chọn mô hình tùy thuộc vào biến số sẽ dự báo là gì và năng lực của các chuyên gia ở từng lĩnh vực chuyên môn.

Đánh giá mô hình

Một khi đã xác định các phương pháp thích hợp cho việc dự báo biến mục tiêu, người làm công tác dự báo cần tiến hành một số đánh giá ban đầu để xem mức độ phù hợp của các mô hình đó như thế nào. Đối với các phương pháp chuỗi thời gian và phương pháp nhân quả với dữ liệu chuỗi thời gian, chúng ta nên thực hiện dự báo hậu nghiệm trước để đánh giá mức độ phù hợp của từng mô hình trong các giai đoạn quá khứ. Nếu mô hình nào không phù hợp đối với các dữ liệu thực trong quá khứ thì rất ít khả năng phù hợp cho tương lai. Đối với phương pháp nhân quả với dữ liệu chéo thì việc lựa chọn mô hình phù hợp cần phải nghiên cứu cơ sở lý thuyết và thực hiện nhiều kiểm định thống kê để đảm bảo đó là mô hình hồi quy tốt.

Chuẩn bị dự báo

Đến đây, chỉ một hoặc một vài phương pháp được chọn cho việc dự báo biến mục tiêu, và qua kiểm định người làm dự báo có những kỳ vọng hợp lý rằng các phương pháp đó sẽ cho kết quả dự báo tốt. Kinh nghiệm cho thấy nếu có thể, người làm dự báo nên sử dụng nhiều hơn một phương pháp, và tốt nhất các phương pháp đó có phân loại khác nhau. Hơn nữa, các phương pháp được chọn cũng nên sử dụng để đưa ra nhiều kết quả dự báo khác nhau từ trường hợp xấu nhất đến tốt nhất.

Trình bày kết quả dự báo

Nếu một kết quả dự định đưa vào sử dụng thì cần phải được trình bày một cách rõ ràng cho ban quản lý để cho người sử dụng hiểu toàn bộ quy trình đã được thực hiện và chứng minh sự tin cậy của kết quả dự báo. Lưu ý rằng, việc người làm dự báo đã bỏ ra bao nhiêu công sức để xây dựng dự báo, khả năng các kết quả có được sử dụng hay không, và mức độ phức tạp của phương pháp luận sử dụng cho dự báo như thế nào đều không thành vấn đề. Nhưng vấn đề ở chỗ người sử dụng có hiểu và có tin cậy vào dự báo hay không. Để làm được như thế, người làm dự báo cần phải trao đổi với ban quản lý bằng thứ ngôn ngữ không chỉ dễ hiểu nhất mà còn phù hợp với văn hóa doanh nghiệp.

Kết quả dự báo nên được trao đổi với ban quản trị hoặc các nhà hoạch định chính sách dưới cả hai hình thức văn bản và thuyết trình. Người sử dụng chỉ cần thông tin chứ ít quan tâm đến khía cạnh kỹ thuật, cho nên chỉ nên cung cấp một số ý tưởng chủ yếu đủ để hiểu phương pháp dự báo được sử dụng. Thông thường, chúng ta nên trình bày nhiều kịch bản khác nhau để thuận lợi cho người sử dụng cân nhắc ra quyết định. Ngoài ra, các bảng biểu trình bày dữ liệu phải gọn, chỉ đưa những thông tin cần thiết.

Theo dõi kết quả dự báo

Thông thường cả người dự báo và người sử dụng kết quả dự báo đều không theo dõi kết quả dự báo sau khi dự báo đã được trình bày và đã

được đưa vào các quyết định. Tuy nhiên, quy trình vẫn tiếp diễn. Những khác biệt giữa dự báo và thực tế nên được đưa ra thảo luận một cách mở, tích cực và có mục tiêu. Các mục tiêu của những thảo luận như thế nhằm tìm hiểu tại sao có sự sai số đó, xem độ lớn của những sai số đó có tạo ra sự khác biệt trong các quyết định dựa trên kết quả dự báo hay không, và xem xét lại toàn bộ quy trình dự báo để cải thiện kết quả dự báo trong tương lai.

Tóm lại, trong suốt quá trình thực hiện dự báo điều quan trọng nhất là phải liên tục có sự trao đổi, bàn bạc kỹ lưỡng giữa người làm dự báo và người sử dụng kết quả dự báo để tăng cường sự hiểu biết và tin cậy vào kết quả dự báo. Ngoài ra, tất cả các bộ phận liên quan cần tăng cường khả năng hợp tác và hỗ trợ lẫn nhau thì kết quả dự báo mới có thể có ích cho quá trình ra quyết định. Vấn đề này sẽ được chúng tôi trình bày một cách cụ thể ở chương 10 trong giáo trình này.

ĐO LƯỜNG MỨC ĐỘ CHÍNH XÁC CỦA DỰ BÁO

Sai số dự báo

Sai số dự báo sẽ là một thước đo tìm hiểu giá trị dự báo sẽ gần với giá trị thực tế bao nhiêu. Trong thực tế sai số dự báo là chênh lệch giữa những giá trị thực tế và giá trị dự báo tương ứng.

$$e_t = Y_t - \hat{Y}_t \qquad (1.1)$$

e_t là sai số dự báo trong giai đoạn t.

Y_t là giá trị thực tế trong giai đoạn t.

$\hat{Y}_t$ là giá trị dự báo.

Nếu một mô hình dự báo được đánh giá là tốt thì sai số dự báo phải tương đối nhỏ. Sự thực, nếu chúng ta đã xây dựng một mô hình một cách đúng đắn thì những dao động của sai số dự báo sẽ không theo một chiều hướng nào cả vì những dao động đó là do các hiện tượng

bên ngoài mà chúng ta không thể dự đoán được. Điều này có nghĩa rằng những dao động ngẫu nhiên của e_t trong mỗi thời đoạn chỉ thuần túy là dao động ngẫu nhiên quanh giá trị dự báo $\hat{Y}_t$, vì vậy tổng của sai số dự báo sẽ tiến về giá trị không. Chính vì thế, việc kiểm định sai số dự báo/phần dư có phải là một chuỗi ngẫu nhiên hay không sẽ là một tiêu chí quan trọng khi đánh giá mức độ chính xác của dự báo.

Đo lường độ chính xác dự báo bằng thống kê

(1) Sai số trung bình (Mean Error)

$$ME = \frac{\sum_{t=1}^{n} e_t}{n} \qquad (1.2)$$

Lưu ý: n là số quan sát của biến dự báo đã được ước lượng ($\hat{Y}_t$).

(2) Sai số phần trăm trung bình (Mean Percentage Error)

$$MPE = \frac{\sum_{t=1}^{n} e_t / Y_t}{n} \qquad (1.3)$$

ME và MPE ít được sử dụng để đo lường độ chính xác của dự báo vì các sai số lớn có giá trị dương có thể bị triệt tiêu bởi các sai số lớn có giá trị âm. Thực vậy, một mô hình xấu có thể có ME và MPE bằng không. Tuy nhiên, ME và MPE lại rất hữu ích trong việc đo lường sự sai lệch của dự báo. Một dự báo không sai lệch, ME và MPE có giá trị gần bằng không. Một dự báo có ME hay MPE âm có thể cho biết mô hình dự báo đang dự báo quá cao, ngược lại, ME hay MPE dương có thể cho biết mô hình dự báo đang dự báo quá thấp.

(3) Sai số tuyệt đối trung bình (Mean Absolute Error)

$$MAE = \frac{\sum_{t=1}^{n} |e_t|}{n} \quad (1.4)$$

MAE là một thước đo rất hữu ích khi người phân tích muốn đo lường sai số dự báo có cùng đơn vị tính với dữ liệu gốc.

(4) Sai số phần trăm tuyệt đối (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum \frac{|e_t|}{Y_t}}{n} \quad (1.5)$$

MAPE là thước đo hữu ích khi độ lớn của biến dự báo có ý nghĩa quan trọng trong việc đánh giá mức độ chính xác của dự báo. MAPE cho một chỉ số về độ lớn của sai số dự báo so với giá trị thực của biến số. Phương pháp này đặc biệt hữu ích khi Y_t có giá trị lớn. Ngoài ra, MAPE cũng có thể được dùng để so sánh các phương pháp giống hoặc khác nhau cho hai chuỗi dữ liệu hoàn toàn khác nhau.

(5) Sai số bình phương trung bình (Mean Squared Error)

$$MSE = \frac{\sum_{t=1}^{n} e_t^2}{n} \quad (1.6)$$

Do sai số được bình phương, nên thước đo MSE có vẻ như "trừng phạt" những sai số dự báo lớn. Và điều này rất quan trọng. Chẳng hạn một phương pháp dự báo có sai số vừa phải có vẻ tốt hơn một phương pháp có nhiều sai số nhỏ nhưng có một vài sai số lớn bất thường.

(6) Căn bậc hai của sai số bình phương trung bình (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} e_t^2}{n}} \tag{1.7}$$

(7) Hệ số không ngang bằng Theil's U

Hệ số không ngang bằng Theil's U là một thước đo khác về độ chính xác dự báo. Hệ số này chính là tỉ số giữa RMSE của mô hình dự báo và RMSE của mô hình dự báo thô giản đơn. Mô hình dự báo thô sử dụng giá trị thực tế Y_t là giá trị dự báo cho giai đoạn kế tiếp ($\hat{Y}_{t+1} = Y_t$). Mô hình dự báo thô sẽ được trình bày ở chương 4.

$$U = \frac{\sqrt{\sum (Y_t - \hat{Y}_t)^2}}{\sqrt{\sum (Y_t - Y_{t-1})^2}} \tag{1.8}$$

Nếu giá trị U càng tiến về không thì mô hình dự báo càng chính xác. Có thể có 3 trường hợp sau đây. Thứ nhất, nếu $U < 1$ thì mô hình dự báo tốt hơn mô hình dự báo thô giản đơn. Thứ hai, nếu $U = 1$ thì mô hình dự báo cũng như mô hình dự báo thô. Thứ ba, nếu $U > 1$ mô hình dự báo còn xấu hơn mô hình dự báo thô. Trong thực tế giá trị của $U < 0.55$ được đánh giá là rất tốt.

- Bảy thước đo độ chính xác dự báo này dùng để:
  - So sánh độ chính xác của hai hay nhiều phương pháp khác nhau.
  - Đo lường sự hữu ích hay độ tin cậy của một phương pháp cụ thể.
  - Giúp tìm ra một phương pháp tối ưu.

- Một số gợi ý quan trọng trong việc lựa chọn giữa các thước đo sai số dự báo như sau:
  - MAE, MAPE, MSE, RMSE và Theil's U có thể sử dụng để so sánh các mô hình dự báo khác nhau cho cùng một chuỗi dữ liệu.
  - Nếu các chuỗi khác nhau về đơn vị đo lường (triệu, %), đơn vị thời gian (năm, quý, tháng), dạng dữ liệu (dữ liệu gốc và dữ liệu chuyển hóa logarit) thì chỉ có MAPE và Theil's U có thể sử dụng được (ví doanh số có thể tính theo đơn vị triệu đồng, trong khi đó lãi suất được tính theo phần trăm).
  - Các phần mềm dự báo ứng dụng thường đưa sẵn giá trị thước đo RMSE về mức độ chính xác dự báo của một mô hình nhất định chỉ vì RMSE tương tự như khái niệm độ lệch chuẩn thông thường trong thống kê.

- Ngoài các thước đo thống kê nói trên, chúng ta có thể đánh giá độ chính xác của mô hình dự báo bằng đồ thị (vẽ các sai số dự báo theo thời gian):
  - Vẽ các sai số dự báo theo thời gian. Nếu những sai số này dao động ngẫu nhiên theo thời gian thì chúng ta có một mô hình dự báo tốt, ngược lại thì mô hình dự báo đã không mô tả đúng xu hướng của dữ liệu.
  - Vẽ giá trị thực tế và giá trị dự báo lên cùng một hệ trục, nếu hai giá trị này trên đồ thị càng gần nhau thì mô hình dự báo càng chính xác (đặc biệt là giai đoạn gần hiện tại hơn).
  - Quan sát những bước ngoặt: Một mô hình dự báo tốt sẽ dự đoán đúng những bước ngoặt theo mẫu dữ liệu thực tế. Một mô hình dự báo sẽ thất bại khi mà nó dự báo không đúng các bước ngoặt.

■ BẢNG 1.1: **Đánh giá độ chính xác của mô hình dự báo.**

t	Y_t	$\hat{Y}_t$	e_t	$\|e_t\|$	e_t^2	$\|e_t\|/Y_t$	e_t/Y_t
1	58	-	-	-	-	-	-
2	54	58	-4	4	16	.074	-.074
3	60	54	6	6	36	.100	.100
4	55	60	-5	5	25	.091	-.091
5	62	55	7	7	49	.113	.113
6	62	62	0	0	0	.000	.000
7	65	62	3	3	9	.046	.046
8	63	65	-2	2	4	.032	-.032
9	70	63	7	7	49	.100	.100
		Tổng	12	34	188	.556	.162

Trong đó:

Y_t = số khách đến cửa hàng vào thời điểm t.

$\hat{Y}_t$ = số khách dự báo đến cửa hàng vào thời điểm t (theo mô hình dự báo thô giản đơn).

$e_t = Y_t - \hat{Y}_t$.

Các thước đo độ chính xác của mô hình dự báo trên được tính toán như sau:

$$MPE = \frac{.162}{8} = .0203 \qquad MAE = \frac{34}{8} = 4.3 \qquad MAPE = \frac{.556}{8} = .0695$$

$$MSE = \frac{188}{8} = 23.5 \qquad RMSE = \sqrt{23.5} = 4.848 \qquad U = \frac{4.848}{4.848} = 1$$

Hầu hết các phần mềm kinh tế lượng và dự báo đều cung cấp các chỉ tiêu đánh giá mức độ chính xác của dự báo. Ví dụ, sử dụng tập tin

DATA1-1, trong đó CPI là chỉ số giá tiêu dùng theo tháng của Việt Nam giai đoạn từ tháng 11 năm 2004 đến tháng 12 năm 2008, và ước lượng hai mô hình dự báo sau đây trên phần mềm Eviews 6.0:

■ BẢNG 1.2: **Mô hình Holt-Winters.**

Sample: 2004M11 2008M12
Included observations: 50
Method: Holt-Winters No Seasonal
Original Series: CPI
Forecast Series: CPISM

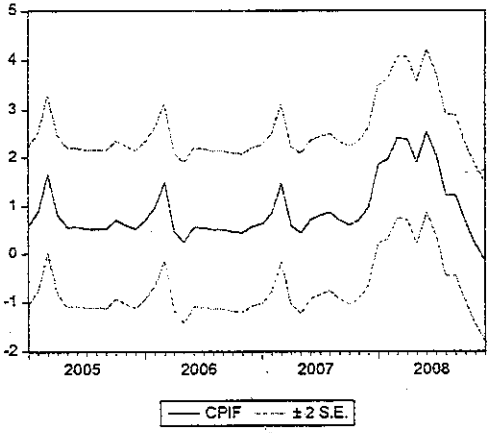
Parameters:	Alpha	0.6200
	Beta	0.0000
Sum of Squared Residuals		32.58313
Root Mean Squared Error		0.807256
End of Period Levels:	Mean	-0.547056
	Trend	0.013133

■ BẢNG 1.3: **Mô hình AR(2).**

Dependent Variable: CPI
Method: Least Squares
Sample (adjusted): 2005M01 2008M12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.826354	0.398618	2.073047	0.0439
AR(1)	0.488121	0.146510	3.331664	0.0017
AR(2)	0.215635	0.150305	1.434645	0.1583

R-squared	0.379261	Mean dependent var	0.903640
Adjusted R-squared	0.351672	S.D. dependent var	1.006776
S.E. of regression	0.810644	Akaike info criterion	2.478485
Sum squared resid	29.57144	Schwarz criterion	2.595435
Log likelihood	-56.48364	Hannan-Quinn criter.	2.522681
F-statistic	13.74710	Durbin-Watson stat	2.013320



Forecast: CPIF	
Actual: CPI	
Forecast sample: 2004M11 2008M12	
Adjusted sample: 2005M01 2008M12	
Included observations: 48	
Root Mean Squared Error	0.784902
Mean Absolute Error	0.561402
Mean Abs. Percent Error	121.8997
Theil Inequality Coefficient	0.322045
Bias Proportion	0.000000
Variance Proportion	0.237745
Covariance Proportion	0.762255

Mô hình Holt-Winters có giá trị RMSE là 0.8073 và mô hình AR(2) có giá trị RMSE là 0.785. Như vậy, trong hai mô hình này thì mô hình AR(2) có độ chính xác dự báo cao hơn so với mô hình Holt-Winters. Ở đây chúng ta chỉ đưa hai mô hình với mục đích minh họa việc đo lường độ chính xác của dự báo nhờ sự hỗ trợ của Eviews. Chúng ta sẽ khảo sát kỹ nội dung và các bước thực hiện các mô hình này trên Eviews lần lượt ở các chương 4 và chương 8. Ngoài ra, để hiểu rõ hơn quy trình thực hiện các mô hình dự báo này trên Eviews, bạn đọc có thể tham khảo thêm tài liệu "Sử dụng Eviews 6.0 trong kinh tế lượng và dự báo"[2].

[2] Xem "Phùng Thanh Bình và Nguyễn Trọng Hoài, 2010, Sử dụng Eviews 6.0 trong kinh tế lượng và dự báo kinh tế".

TÓM TẮT CHƯƠNG 1

Chương 1 đã trình bày tổng quát các chủ đề quan trọng liên quan đến dự báo như khái niệm dự báo, cách phân loại, hệ thống các phương pháp, phương pháp luận, trình tự dự báo, và các tiêu chí đo lường mức độ chính xác của dự báo. Qua chương này chúng ta nhận thức rằng dự báo cực kỳ quan trọng cho các nhà quản trị, các nhà hoạch định chính sách ở cả các lĩnh vực vi mô và vĩ mô. Điều quan trọng cần khẳng định là: cho dù chúng ta sử dụng phương pháp định tính hay định lượng hoặc là kết hợp cả hai thì kết quả dự báo sẽ vẫn có những sai số nhất định. Hơn nữa, kết quả dự báo cần được con người thảo luận và sử dụng có phê phán và cân nhắc theo một quy trình khoa học và thận trọng. Sau cùng là, mặc dù dự báo được rất nhiều phần mềm hỗ trợ với các kỹ thuật thống kê và mô hình kinh tế lượng phức tạp hay còn gọi là "hộp đen" vẫn không thể thay thế những đánh giá trực quan của con người. Ở chương 10, chúng tôi sẽ nhắc lại quy trình dự báo một cách hệ thống hơn sau khi bạn đọc đã hiểu được cách thức thực hiện từng mô hình dự báo cụ thể. Và hy vọng rằng qua cuốn giáo trình này, chúng ta có thể cảm nhận được tại sao dự báo định lượng thực sự quan trọng và cần thiết trong hầu hết các quyết định cho tương lai của tổ chức.

CÂU HỎI VÀ BÀI TẬP

1. Anh/Chị hãy trình bày ngắn gọn các cách thức phân loại dự báo?
2. Anh/Chị hãy trình bày ngắn gọn ưu nhược điểm của phương pháp dự báo định tính?
3. Anh/Chị cho biết phương pháp Delphi là gì? Phương pháp Delphi thường được sử dụng trong những trường hợp nào? Và quy trình dự báo theo phương pháp Delphi được thực hiện như thế nào?
4. Anh/Chị cho biết tại sao dự báo theo các phương pháp định lượng đang được sử dụng phổ biến?
5. Anh/Chị hãy trình bày ngắn gọn phương pháp luận của dự báo theo các mô hình chuỗi thời gian?
6. Anh/Chị hãy trình bày ngắn gọn phương pháp luận của dự báo theo các mô hình nhân quả?
7. Anh/Chị phân biệt dự báo theo các mô hình chuỗi thời gian và các mô hình nhân quả?
8. Anh/Chị hãy trình bày ngắn gọn quy trình thực hiện một mô hình dự báo chuỗi thời gian?
9. Anh/Chị hãy trình bày ngắn gọn quy trình thực hiện một mô hình dự báo kinh tế lượng?
10. Giả sử Anh/Chị được mời làm tư vấn – đào tạo và chuyển giao kỹ thuật dự báo giá khí đốt cho công ty kinh doanh sản phẩm khí, thì Anh/Chị thực hiện như thế nào?
11. Có kết quả dự báo doanh số như sau:

Năm	Doanh số thực tế (triệu đồng)	Doanh số dự báo (mô hình 1)	Doanh số dự báo (mô hình 2)
1	1.225	-	1.251
2	1.285	1.225	1.293
3	1.359	1.285	1.335

Năm	Doanh số thực tế (triệu đồng)	Doanh số dự báo (mô hình 1)	Doanh số dự báo (mô hình 2)
4	1.392	1.359	1.378
5	1.443	1.392	1.420
6	1.474	1.443	1.463
7	1.467	1.474	1.505
8		*1.467*	*1.548*

a. Anh/Chị hãy lập bảng tính các sai số dự báo (e_{1t} và e_{2t}) cho mô hình 1 và mô hình 2?

b. Anh/Chị hãy lập bảng và tính các tiêu chí MAE và RMSE cho mô hình 1 và mô hình 2?

c. Từ kết quả tính toán ở câu b, Anh/Chị đề xuất nên chọn kết quả dự báo nào? Tại sao?

12. Sử dụng tập tin **"EXIMVN.xls"**, trong đó Y là kim ngạch xuất khẩu theo quý của Việt Nam giai đoạn 1999Q1:2008Q4, $\hat{Y}_1$ và $\hat{Y}_2$ là giá trị dự báo tương ứng theo mô hình 1 và mô hình 2, Anh/Chị hãy trả lời các câu hỏi sau đây:

a. Vẽ trên cùng hệ trục các biến Y, $\hat{Y}_1$ và $\hat{Y}_2$, và nhận xét mô hình dự báo nào tốt hơn?

b. Tính các tiêu chí đánh giá độ chính xác của dự báo và nhận xét mô hình dự báo nào tốt hơn?

c. Anh/Chị cho biết có cần thêm thông tin nào khác để dự báo kim ngạch xuất khẩu hay không? Tại sao?

CHƯƠNG 2 VAI TRÒ CỦA THỐNG KÊ TRONG DỰ BÁO

Hầu hết các kỹ thuật dự báo mà chúng ta sẽ khảo sát trong cuốn giáo trình này đều dựa trên các khái niệm thống kê căn bản mà chúng ta đã được học ở giai đoạn đại cương. Thống kê căn bản xoay quanh ba nội dung: (1) phân tích dữ liệu, (2) phân phối xác suất, và (3) suy diễn thống kê. Để việc ôn tập các nội dung này thêm phần sinh động và dễ hiểu, chương này sẽ kết hợp hướng dẫn thao tác trên Eviews 6.0 và Excel.

Bất kể chúng ta sẽ sử dụng các mô hình chuỗi thời gian hay mô hình kinh tế lượng để dự báo một vấn đề nào đó, thì một sự phân tích sơ bộ có ý nghĩa vô cùng quan trọng để có cảm nhận sơ bộ về dữ liệu. Chương này sẽ bắt đầu bằng việc giới thiệu các cấu trúc dữ liệu kinh tế được sử dụng phổ biến trong các mô hình dự báo định lượng và cách tạo một tập tin Eviews. Trên cơ sở đó, chúng ta sẽ thảo luận một cách ngắn gọn các cách thức khác nhau để xem xét và phân tích dữ liệu thông qua việc khảo sát một số loại đồ thị quan trọng và thống kê mô tả trên Eviews. Đối với dữ liệu thời gian, chúng ta sẽ xem xét một số cách chuyển hóa dữ liệu khác nhau để có thể cô lập hoặc loại bỏ một hoặc một số thành phần nào đó trong chuỗi thời gian. Quy trình này sẽ cung cấp một nền tảng quan trọng cho việc lựa chọn các mô hình dự báo phù hợp được trình bày ở chương 3. Ngoài ra, để giúp bạn đọc cảm thấy dễ dàng hơn trong việc kiểm định các giả thiết thống kê và đánh giá kết quả dự báo, chương này sẽ trình bày một cách súc tích nhất các ý tưởng và ứng dụng quan trọng nhất của một số phân phối xác suất cơ bản và các phương pháp kiểm định giả thiết thường được sử dụng trong phân tích hồi quy và dự báo.

MỤC TIÊU HỌC TẬP

Sau khi học xong chương này, chúng ta kỳ vọng sẽ đạt được các nội dung sau đây:

- Biết được cấu trúc của các dữ liệu kinh tế và kinh doanh.
- Hiểu được các đặc điểm quan trọng của một phân phối xác suất.
- Nhận biết được một số phân phối xác suất quan trọng và các ứng dụng của các phân phối đó trong phân tích dữ liệu và dự báo.
- Hiểu được tầm quan trọng của suy diễn thống kê trong phân tích dữ liệu và dự báo.

CẤU TRÚC DỮ LIỆU KINH TẾ

Dữ liệu kinh tế có nhiều hình thức khác nhau. Trong khi một số phương pháp dự báo có thể được áp dụng trực tiếp cho các loại dữ liệu khác nhau, thì thông thường chúng ta phải phân tích các đặc điểm của dữ liệu để lựa chọn mô hình thích hợp. Đây là nội dung sẽ được phân tích chi tiết ở chương 3. Trong phần này chúng ta sẽ xem xét ba loại cấu trúc dữ liệu cơ bản thường được sử dụng trong dự báo và phân tích kinh tế lượng.

DỮ LIỆU CHÉO

Dữ liệu chéo là dữ liệu về một hay một số biến được thu thập tại cùng một thời điểm. Ví dụ, các cuộc điều tra thống kê của Việt Nam, điều tra về chỉ số cạnh tranh cấp tỉnh, điều tra nông hộ ở một địa phương, khảo sát mức độ thỏa mãn của khách hàng, v.v...

Dữ liệu chéo là loại dữ liệu được sử dụng phổ biến nhất trong kinh tế và nhiều lĩnh vực khoa học xã hội khác. Trong dự báo, dữ liệu chéo được sử dụng khá phổ biến để hỗ trợ cho các phương pháp dự báo định tính và một số mô hình dự báo nhân quả.

■ BẢNG 2.1: **Mẫu dữ liệu chéo.**

Quan sát	Y	X1	X2	X3	X4
1	1901	2250	600	1200	4
2	2126	11200	1400	3600	4
3	1780	0	4750	2160	4
4	1202	0	1200	1200	2
5	2223	2090	460	1200	3
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
9188	3687	0	0	0	5
9189	3094	0	1400	100	4

Nguồn: VHLSS, 2006.

Trong đó, Y, X1, X2, X3, X4 lần lượt là chi tiêu cho lúa gạo, chi tiêu cho giáo dục, chi tiêu cho y tế, chi tiêu cho tiền điện, và qui mô hộ gia đình (tập tin **DATA2-1**).

DỮ LIỆU CHUỖI THỜI GIAN

Dữ liệu chuỗi thời gian là các dữ liệu mà các biến quan sát được thu thập theo thời gian, chẳng hạn như GDP, CPI, việc làm, thất nghiệp, cung tiền, lãi suất, chỉ số giá chứng khoán, suất sinh lợi của một cổ phiếu, giá dầu, giá vàng, doanh số, v.v... Các dữ liệu thời gian có thể được thu thập theo một tần suất quan sát nhất định tùy đặc điểm của từng đối tượng nghiên cứu, ví dụ theo ngày (chứng khoán, lãi suất, tỷ giá hối đoái), theo tuần (lương tuần, cung tiền), theo tháng (tỷ lệ thất nghiệp, tỷ lệ lạm phát, sản lượng công nghiệp, doanh số), theo quý (GDP, doanh số), theo năm (ngân sách Chính phủ, tốc độ tăng trưởng kinh tế, tỷ lệ lạm phát, giá trị xuất khẩu).

■ BẢNG 2.2: **Mẫu dữ liệu chuỗi thời gian.**

Quan sát	Tháng	X1	X2	X3	X4
1	2005M1	80.38089	39.63368	54.20863	556.9369
2	2005M2	83.5175	44.61313	57.64914	549.3329
3	2005M3	95.45541	50.65714	60.31371	545.1831
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
49	2009M1	82.30484	85.77211	68.26367	464.5935
50	2009M2	78.27128	81.46765	67.22247	449.7209

Nguồn: IMF, 2009.

Trong đó, X1, X2, X3, và X4 lần lượt là chỉ số giá dầu, giá cà phê Robusta, giá cao su, và giá gỗ (tập tin **DATA2-2**).

Lưu ý, trong kinh tế lượng các biến chuỗi thời gian thường được ký hiệu bằng chữ *t* nhỏ dưới các biến, ví dụ Y_t, X_{2t}, X_{3t}, ... trong đó, *t* đại diện cho các quan sát từ 1 đến T. Chẳng hạn, 1 đại diện cho quan sát eủa tháng 1 năm 2005, 2 đại diện cho quan sát của tháng 2 năm 2005, ..., và 50 đại diện cho quan sát của tháng 2 năm 2009.

Do các sự kiện trong quá khứ có thể ảnh hưởng đến các sự kiện trong tương lai, nên trong các nghiên cứu dữ liệu thời gian thường có khái niệm biến có độ trễ, và các biến này thường được ký hiệu bằng Y_{t-1}, X_{2t-2}, Y_{t-3}, X_{2t-3}, v.v... Ngoài ra, dữ liệu thời gian còn phụ thuộc vào các yếu tố mùa vụ, chu kỳ, nên đôi khi trong nghiên cứu người ta có xem xét các biến giả mùa vụ hoặc các chỉ số mùa vụ.

DỮ LIỆU BẢNG

Dữ liệu bảng có các thành phần của cả dữ liệu thời gian và dữ liệu chéo. Ví dụ, nếu ta thu thập dữ liệu về tỷ lệ thất nghiệp của 10 quốc gia cho giai đoạn 20 năm, thì dữ liệu đó sẽ tạo thành một dữ liệu gộp - dữ liệu về tỷ lệ thất nghiệp của mỗi nước trong 20 năm là một dữ liệu

thời gian và tỷ lệ thất nghiệp của 10 nước tại một năm bất kỳ là dữ liệu chéo. Một ví dụ khác là hai cuộc điều tra mức sống gia định Việt Nam vào các năm 1993, 1998, 2002, 2004, và 2006 trong đó nội dung bảng câu hỏi là giống nhau, nhưng mẫu phỏng vấn khác nhau và lớn hơn.

Kết hợp các dữ liệu chéo ở các năm khác nhau thường là một cách rất hiệu quả trong việc phân tích các ảnh hưởng của một chính sách mới. Ý tưởng thu thập dữ liệu từ những năm trước và sau khi có một sự thay đổi chính sách quan trọng.

■ BẢNG 2.3: **Mẫu dữ liệu bảng.**

Quan sát	Năm	Y	X1	X2	X3
1	1993	85500	42	1600	3
2	1993	67300	36	1440	3
3	1993	134000	38	2000	4
.	.	.	.	.	.
.	.	.	.	.	.
250	1993	243600	41	2600	4
25i	1995	65000	16	1250	2
252	1995	182400	20	2200	4
.	.	.	.	.	.
.	.	.	.	.	.
520	1995	57200	16	1100	2

Nguồn: Wooldridge, 2003.

Trong đó, Y, X1, X2, X3 lần lượt là giá nhà, mức thuế tài sản, diện tích, và số phòng ngủ.

Quyết định chọn thu thập dữ liệu nào thường tùy thuộc vào bản chất của nghiên cứu. Để trả lời các câu hỏi ở cấp độ cá nhân hay hộ gia đình, ta thường chỉ sử dụng các dữ liệu chéo, cụ thể thu thập bằng

điều tra thực tế. Nếu mục đích nghiên cứu xem liệu có các mối quan hệ kinh tế có thay đổi theo thời gian hay không, ta có thể sử dụng loại dữ liệu bảng. Tuy nhiên, loại dữ liệu này rất khó thu thập. Dữ liệu chuỗi thời gian được sử dụng phổ biến trong dự báo và phân tích mối quan hệ dài hạn giữa các chỉ báo kinh tế. Dữ liệu thời gian thường được sử dụng trong các nghiên cứu kinh tế vĩ mô. Khi phân tích dữ liệu chuỗi thời gian người phân tích cần phải hết sức lưu ý xem các chuỗi thời gian đó dừng hay không dừng vì nó quyết định đến việc lựa chọn mô hình, ý nghĩa của mối quan hệ kinh tế, và các kiểm định chuyên biệt. Ngoài ra, một điều cũng đáng quan tâm là nếu chúng ta có chuyển hóa dữ liệu thời gian, ví dụ điều chỉnh yếu tố mùa, làm trơn dữ liệu, v.v..., thì chúng ta cũng cần phải tuyệt đối cẩn thận. Lưu ý, trong phạm vi cuốn giáo trình này, chúng tôi không trình bày các mô hình dự báo sử dụng dữ liệu bảng.

TẠO MỘT TẬP TIN EVIEWS

Có nhiều cách tạo một tập tin mới[1]. Việc đầu tiên khi tạo một tập tin Eviews là xác định cấu trúc của tập tin. Có hai cách tạo tập tin Eviews đang được sử dụng phổ biến. Thứ nhất là mô tả cấu trúc của tập tin Eviews. Theo cách này, Eviews sẽ tạo ra một tập tin mới để người sử dụng nhập dữ liệu một cách thủ công từ bàn phím hoặc copy và dán, ví dụ từ Excel. Thứ hai là mở và đọc dữ liệu từ một nguồn bên ngoài (không thuộc định dạng Eviews) như Text, Excel, Stata.

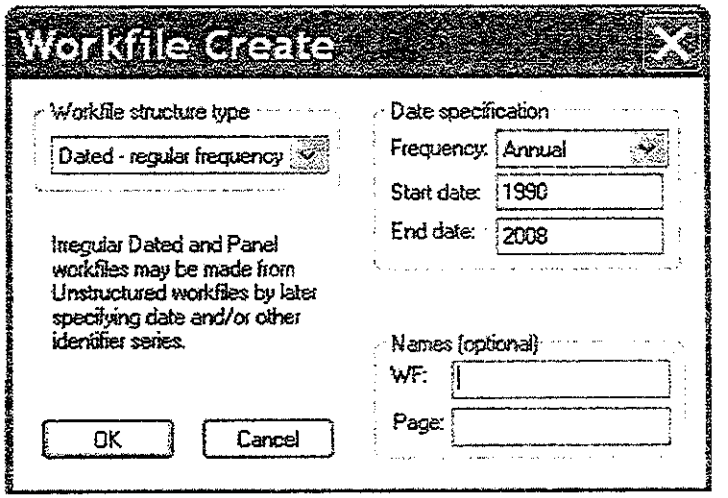
MÔ TẢ CẤU TRÚC DỮ LIỆU

Để mô tả cấu trúc của tập tin Eviews, ta phải cung cấp cho Eviews các thông tin về số quan sát và các nhận dạng liên quan. Để tạo một tập tin mới trên Eviews, ta chọn **File/New Workfile**, ... từ menu chính để mở hộp thoại **Workfile Create**. Ở góc trái của hộp thoại là một hộp nhỏ để mô tả cấu trúc cơ bản của bộ dữ liệu. Ta có thể chọn giữa **Dated-Regular Frequency**, **Unstructured**, và **Balanced Panel**. Nói

[1] Xem "Phùng Thanh Bình và Nguyễn Trọng Hoài, 2010, Sử dụng Eviews 6.0 trong kinh tế lượng và dự báo kinh tế".

chung, ta có thể sử dụng **Dated-regular frequency**[2] nếu ta có bộ dữ liệu thời gian, với bộ dữ liệu bảng đơn giản ta sử dụng **Balanced Panel**, và các trường hợp khác là dữ liệu chéo ta sử dụng **Unstructured**[3].

Sau khi đã xác định loại cấu trúc dữ liệu, chúng ta có thể nhập dữ liệu từ bàn phím hoặc copy dữ liệu từ Excel và dán vào, hoặc sử dụng import dữ liệu. Thông thường chúng ta chọn cách thứ hai hoặc thứ ba. Sau khi đã copy dữ liệu từ Excel, ta trở lại tập tin Eviews vừa được tạo ra, chọn **Quick/Empty Group (Edit Series)**, và Eviews sẽ dán các số liệu được chọn vào bảng tính theo một cấu trúc đã được xác định.



[2] Nếu là dữ liệu năm, thì ở ô Frequency ta chọn Annual; ở các ô Start date và End date ta nhập năm bắt đầu và năm kết thúc của các chuỗi dữ liệu. Nếu dữ liệu là quý, thì ở ô Frequency ta chọn Quarterly; ở các ô Start date và End date ta nhập quý bắt đầu và quý kết thúc của các chuỗi dữ liệu. Ở đây ta có thể chọn một trong hai cách sau (ví dụ quý 2 năm 2005): 2005:2 hoặc 2005Q2. Nếu là dữ liệu tháng, thì ở ô Frequency ta chọn Monthly; ở các ô Start date và End date ta nhập tháng bắt đầu và tháng kết thúc của các chuỗi dữ liệu. Tương tự, ta có thể chọn một trong hai cách sau (ví dụ tháng 8 năm 2008): 2008:8 hoặc 2008M8. Các ô đặt tên là tùy chọn (đặt tên tập tin và tên trang), nhưng thông thường không cần thiết.

[3] Sử dụng đối với loại dữ liệu chéo và ta chỉ cần nhập số quan sát của bộ dữ liệu vào ô Observations là xong.

MỞ VÀ ĐỌC TỪ MỘT NGUỒN BÊN NGOÀI

Thông thường, ta mở trực tiếp một nguồn dữ liệu bên ngoài như cách mở bất kỳ một tập tin nào. Để mở một file bên ngoài, trước hết ta chọn **File/Open/Foreign Data as Workfile**, ... để đến hộp thoại **Open**, chọn **Files of type,** mở file cần chuyển sang tập tin Eviews, và thực hiện một số điều chỉnh nếu cần thiết.

Đối với tập tin Stata

Khi chọn và mở tập tin Stata, ta thấy xuất hiện hộp thoại **Table Read Specification**. Trong đó, ta chọn **Select** hoặc **Unselect** để chọn các biến cần thiết chuyển sang dạng dữ liệu Eviews. Tuy nhiên, thông thường ta chọn tất cả các biến có sẵn theo mặc định của Eviews (rồi xóa sau, nếu cần). Ngoài ra, ta cũng có thể định nghĩa lại bộ dữ liệu của mình thông qua chọn các điều kiện cần cho phù hợp mục tiêu nghiên cứu (ví dụ chỉ chọn các quan sát ở vùng 8 trong tập tin **DATA2-1** bằng cách chọn reg8=8) bằng cách chọn **Filter Obs** và nhập điều kiện vào.

Đối với tập tin Text

Khi chọn và mở tập tin, ta thấy xuất hiện hộp thoại **ASCII Read**. Trong **Column specification** có ba lựa chọn: Delimiter ..., Fixed ..., và An explicit ... cho phép ta lựa chọn chiều rộng của các cột dữ liệu hiện trong tập tin. Tuy nhiên, thông thường Eviews sẽ mặc định ở dạng Delimiter ... Ở **Start date/header** ta thấy ô <Skip lines> cho phép ta lựa chọn bỏ các dòng đầu tiên (thường chỉ để lại dòng tên các biến), ví dụ ở đây ta chọn "2". Điều này chỉ có ý nghĩa giúp ta dễ dàng kiểm tra dữ liệu chứ không cần thiết lắm. Mục **Row specification** cho phép ta xác định số quan sát trong một dòng (thông thường là 1). Mục này nói chung cũng không cần thiết. Sau đó ta chọn Next qua bước 2, và lại chọn Next để qua bước 3. Ở bước 3 ta có thể đặt lại tên biến bằng cách chọn biến đó và thay bằng tên biến mong muốn. Ngoài ra, ta cũng có thể đặt tên nhãn và mô tả đặc điểm của biến đó. Cuối cùng ta chọn **Finish**.

Đối với tập tin Excel

Khi chọn và mở tập tin Eviews sẽ thực hiện thông qua hai bước. Bước một, ta thấy xuất hiện hộp thoại **Spreadsheet Read.** Bước hai, Eviews sẽ đưa ra các lựa chọn để đọc dữ liệu và những thay đổi theo ý người sử dụng như đặt lại tên và nhãn của các biến. Tuy nhiên, trong hầu hết các trường hợp người sử dụng chỉ cần chọn **Finish** để chấp nhận định dạng mặc định. Thực hành trên Eviews với tập tin **DATA2-3.**

PHÂN TÍCH DỮ LIỆU VỚI EVIEWS

CÁC ĐẶC ĐIỂM CỦA PHÂN PHỐI XÁC SUẤT

Mặc dù một hàm phân phối xác suất thể hiện các giá trị một biến ngẫu nhiên nhận được và các xác suất tương ứng của các giá trị đó, nhưng ta thường không quan trong đến toàn bộ hàm phân phối xác suất. Ta có thể chỉ quan tâm đến một số đặc điểm tóm tắt, hay nói theo ngôn ngữ thống kê là các giá trị kỳ vọng và phương sai. Ngoài ra, đôi khi ta vẫn quan tâm đến các đặc điểm khác của phân phối xác suất như độ nghiêng và độ nhọn của phân phối.

Giá trị kỳ vọng (trung bình tổng thể)

Giá trị kỳ vọng của một biến ngẫu nhiên rời rạc, ký hiệu là E(X), được định nghĩa như sau:

$$E(X) = \mu_X = \sum_x Xf(X) \qquad (2.1)$$

Trong đó, f(X) là hàm phân phối xác suất của X (hay còn được ký hiệu là xác suất P_i) và Σ_x là tổng tất cả các giá trị X. Đối với một biến ngẫu nhiên liên tục thì ta thay ký hiệu tổng bằng ký hiệu tích phân.

Ví dụ, kết quả thống kê cho biết trong điều kiện kinh tế khó khăn thì suất sinh lợi của một cổ phiếu X là -5%/năm, điều kiện bình thường thì suất sinh lợi là 15%/năm, và điều kiện thuận lợi thì suất sinh lợi là 25%/năm. Biết rằng xác suất ba điều kiện này xảy ra lần

lượt là 0.2, 0.5, và 0.3. Như vậy, suất sinh lợi kỳ vọng của cổ phiếu X sẽ được ước tính như sau:

BẢNG 2.4: Giá trị kỳ vọng suất sinh lợi của cổ phiếu X.

Điều kiện	Suất sinh lợi của X	Xác suất P_i	$X_i P_i$
Khó khăn	-5%	0.2	-1%
Bình thường	15%	0.5	7.5%
Thuận lợi	25%	0.3	7.5%
Suất sinh lợi kỳ vọng, E(X)			14%

Giá trị kỳ vọng của một biến ngẫu nhiên là trung bình có trọng số của các giá trị có thể có của biến đó, với xác suất của các giá trị này, f(X), đóng vai trò như các trọng số. Giá trị kỳ vọng của một biến ngẫu nhiên cũng được gọi là giá trị trung bình, mặc dù thuật ngữ chính xác hơn là giá trị trung bình tổng thể.

Tính chất của giá trị kỳ vọng

- $E(b) = b$ (2.2)
- $E(X+Y) = E(X) + E(Y)$ (2.3)
- $E(X/Y) \neq \dfrac{E(X)}{E(Y)}$ (2.4)
- $E(XY) \neq E(X)E(Y)$ (2.5)

  Nếu X và Y là hai biến ngẫu nhiên độc lập, thì

  $E(XY) = E(X)E(Y)$ (2.6)
- $E(X^2) \neq [E(X)]^2$ (2.7)
- $E(aX) = aE(X)$ (2.8)
- $E(aX+b) = aE(X) + b$ (2.9)

Phương sai tổng thể

Giá trị kỳ vọng của một biến ngẫu nhiên đơn giản chỉ cho biết trọng tâm của biến đó ở đâu chứ không cho biết các giá trị riêng lẻ của biến đó phân tán như thế nào xung quanh giá trị trung bình. Thước đo phổ biến nhất cho sự phân tán này là phương sai, và được định nghĩa như sau:

$$\text{var}(X) = \sigma_x^2 = E(X-\mu_x)^2 \qquad (2.10)$$

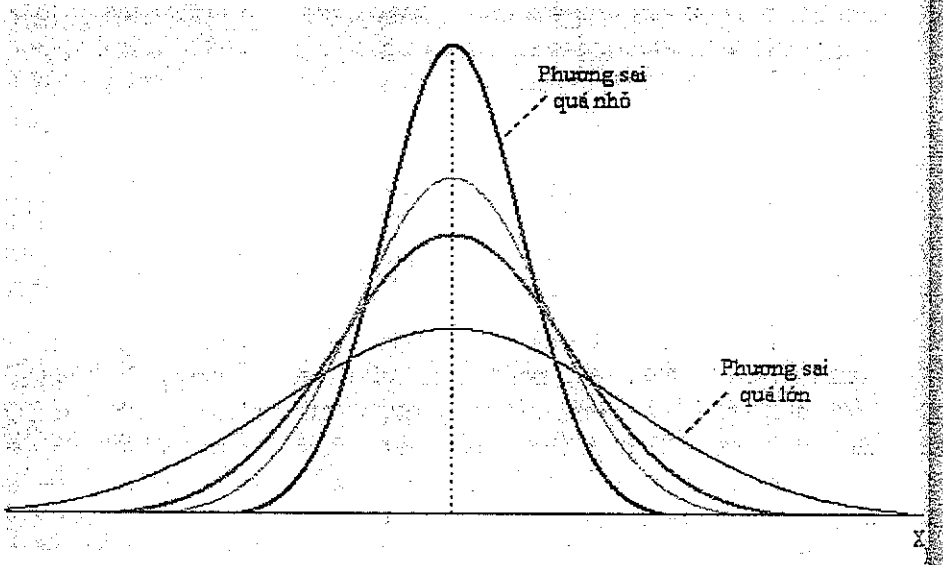
$$\text{var}(X) = \sum (X-\mu_x)^2 f(X) \qquad (2.11)$$

Nếu X là một biến ngẫu nhiên liên tục thì ta thay ký hiệu tổng bằng ký hiệu tích phân. Như vậy, phương sai cho biết các giá trị X riêng lẻ được phân phối hay phân tán xung quanh giá trị trung bình như thế nào. Nếu các giá trị X phân tán rộng quanh giá trị trung bình thì phương sai sẽ tương đối lớn (Hình 2.1). Căn bậc hai của phương sai là độ lệch chuẩn, ký hiệu là σ_x. Trở lại ví dụ ở Bảng 2.4, ta có thể tính phương sai của cổ phiếu X như sau:

■ BẢNG 2.5: **Giá trị kỳ vọng của cổ phiếu X.**

Điều kiện	X_i	P_i	$(X_i-\mu)^2$	$X_i P_i$
Khó khăn	-5%	0.2	$(-5\%-14\%)^2$	0.72%
Bình thường	15%	0.5	$(15\%-14\%)^2$	0.00%
Thuận lợi	25%	0.3	$(25\%-14\%)^2$	0.36%
Phương sai suất sinh lợi của cổ phiếu X, Var(X)				1.1%

■ HÌNH 2.1: **Hàm phân phối xác suất của các biến cùng giá trị kỳ vọng.**



Tính chất của phương sai

- Phương sai của một hằng số bằng không.
- Nếu X và Y là hai biến ngẫu nhiên độc lập, thì:

$$var(X+Y) = var(X) + var(Y) \quad (2.12)$$

$$var(X-Y) = var(X) - var(Y) \quad (2.13)$$

- Nếu b là hằng số, thì:

$$var(aX) = a^2 var(X) \quad (2.14)$$

- Nếu a và b là hằng số, thì:

$$var(aX+b) = a^2 var(X) \quad (2.15)$$

- Nếu X và Y là hai biến độc lập và a và b là hằng số, thì

$$var(aX+bY) = a^2 var(X) + b^2 var(Y) \quad (2.16)$$

Hệ số biến thiên tổng thể

Vì độ lệch chuẩn (hay phương sai) phụ thuộc vào các đơn vị đo lường khác nhau do các biến được đo lường khác nhau, cho nên sẽ khó cho việc so sánh giữa các độ lệch chuẩn nếu chúng có các thước đo khác nhau. Để giải quyết vấn đề này, ta có thể sử dụng hệ số biến thiên (V) như sau:

$$V = \frac{\sigma_x}{\mu_x}.100 \qquad (2.17)$$

Hiệp phương sai tổng thể

Giá trị kỳ vọng và phương sai là hai thước đo được sử dụng phổ biến nhất của hàm phân phối xác suất một biến. Nhưng khi muốn xem xét các hàm phân phối xác suất đa biến, ta thường quan tâm đến hai thước đo khác là hiệp phương sai và hệ số tương quan.

Giả sử X và Y là hai biến ngẫu nhiên với $E(X) = \mu_x$ và $E(Y) = \mu_y$, thì hiệp phương sai (cov) giữa hai biến sẽ như sau:

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] \qquad (2.18)$$

Hiệp phương sai giữa hai biến có thể dương, âm, hoặc bằng không. Nếu hai biến vận động theo cùng chiều, thì hiệp phương sai sẽ dương, nếu khác chiều, thì hiệp phương sai sẽ âm. Nếu hiệp phương sai giữa hai biến bằng không, thì điều này có nghĩa là không có mối quan hệ tuyến tính nào giữa hai biến đó. Nói cách khác, đó là hai biến độc lập hoàn toàn.

Tính chất của hiệp phương sai

- Nếu X và Y là hai biến ngẫu nhiên độc lập, hiệp phương sai của chúng bằng không.
- $cov(a+bX, c+dY) = bdcov(X,Y)$ (2.19)
- $cov(X,X) = var(X)$ (2.20)

- Nếu X và Y là hai biến ngẫu nhiên nhưng không nhất thiết phải độc lập, thì công thức tính phương sai được viết lại như sau:

$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y) \quad (2.21)$$

$$var(X-Y) = var(X) + var(Y) - 2cov(X,Y) \quad (2.22)$$

Hệ số tương quan tổng thể

Do hiệp phương sai phụ thuộc vào thước đo của hai biến, nên chúng ta khó có thể so sánh mối quan hệ giữa các 'cặp' biến có thước đo khác nhau. Để khắc phục vấn đề này, người ta thường sử dụng hệ số tương quan. Hệ số tương quan là thước đo mối quan hệ tuyến tính giữa hai biến ngẫu nhiên, nghĩa là nó cho biết hai đó có quan hệ với nhau như thế nào: mạnh hay yếu. Hệ số tương quan tổng thể (ρ) được xác định như sau:

$$\rho = \frac{cov(X,Y)}{\sigma_x \sigma_y} \quad (2.23)$$

Tính chất của hệ số tương quan

- Giống hiệp phương sai, hệ số tương quan có thể âm hoặc dương.
- Hệ số tương quan là một thước đo mối quan hệ tuyến tính giữa hai biến số.
- $-1 \le \rho \le 1$.
- Hệ số tương quan là một con số thuần túy không có đơn vị đo lường.
- Nếu hai biến độc lập, hệ số tương quan bằng không.
- Hệ số tương quan không hàm ý mối quan hệ nhân quả.

Độ nghiêng và độ nhọn tổng thể

Độ nghiêng và độ nhọn cho ta biết điều gì đó về hình dạng của phân phối xác suất. Độ nghiêng (S) là một thước đo sự mất cân xứng của đồ thị phân phối xác suất, và độ nhọn (K) là một thước đo độ cao hay thấp của đồ thị phân phối xác suất.

Để tính các thước đo độ nghiêng và độ nhọn (Hình 2.2 và Hình 2.3), ta cần biết mô men thứ ba và mô men thứ tư của một hàm phân phối xác suất. Ta đã biết mô men thứ nhất của hàm phân phối xác suất của một biến ngẫu nhiên được đo bằng $E(X) = \mu_x$, và mô men thứ hai xung quanh giá trị trung bình (phương sai) được đo bằng $E(X- \mu_x)^2$. Tương tự, mô men thứ ba và mô men thứ tư quanh giá trị trung bình được đo như sau:

Mô men thứ ba: $E(X-\mu_x)^3$ (2.24)

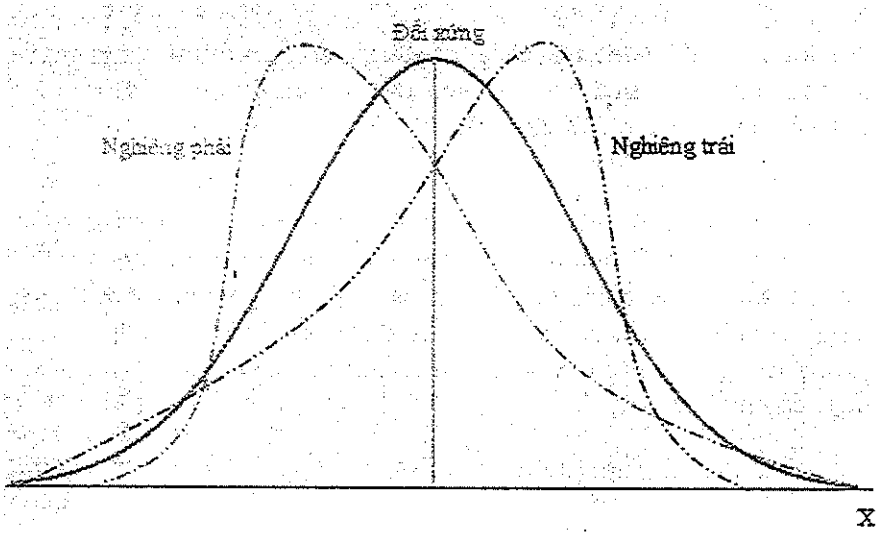
Mô men thứ tư: $E(X-\mu_x)^4$ (2.25)

$$S = \frac{E(X-\mu_x)^3}{\sigma_x^3} \quad (2.26)$$

Có ba khả năng xảy ra như sau:

- Nếu $S = 0$, hàm phân phối xác suất đối xứng quanh giá trị trung bình.
- Nếu $S > 0$, hàm phân phối xác suất bị nghiêng phải.
- Nếu $S < 0$, hàm phân phối xác suất bị nghiêng trái.

■ HÌNH 2.2: **Độ nghiêng của phân phối.**



$$K = \frac{E(X-\mu_X)^4}{[E(X-\mu_X)^2]^2} \qquad (2.27)$$

Có ba khả năng xảy ra như sau:

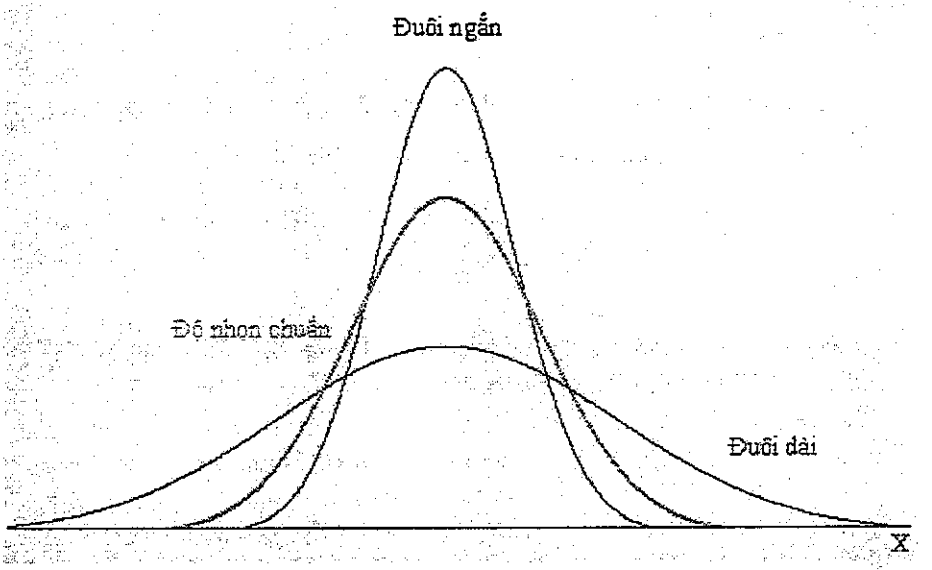
- Nếu K = 3, hàm phân phối xác suất có độ nhọn chuẩn và được gọi là mesokurtic.
- Nếu K < 3, hàm phân phối xác suất có đuôi ngắn và được gọi là platykurtic.
- Nếu K > 3, hàm phân phối xác suất có đuôi dài và được gọi là leptokurtic.

TỔNG THỂ VÀ MẪU DỮ LIỆU

Để tính các đặc điểm của các phân phối xác suất, như giá trị kỳ vọng, phương sai, hiệp phương sai, hệ số tương quan, và giá trị kỳ vọng có

điều kiện, hiển nhiên ta cần phải có hàm phân phối xác suất, nghĩa là, ta phải có sẵn dữ liệu mẫu hay tổng thể. Tổng thể bao hàm mẫu, và vì yếu tố tiết kiệm nguồn lực hoặc vì lý do bất khả kháng khi thu thập dữ liệu, do đó chúng ta thường nghiên cứu tổng thể từ dữ liệu mẫu. Ví dụ, để nghiên cứu đặc điểm của biến thu nhập trung bình của người dân thành phố Hồ Chí Minh tại một thời điểm, ta cần phải có thông tin về toàn bộ dân số ở thành phố Hồ Chí Minh. Mặc dù trên lý thuyết dân số ở thành phố Hồ Chí Minh là xác định, nhưng ta không thể nào thu thập thông tin của từng người dân vì điều đó quá tốn kém và không khả thi. Điều ta có thể làm trên thực tế là rút một mẫu ngẫu nhiên hay đại diện từ dân số thành phố Hồ Chí Minh và tính thu nhập trung bình của số người trong mẫu khảo sát. Trên cơ sở đó ta suy ra các đặc điểm của phân phối tổng thể từ một mẫu khảo sát. Tất cả các mô hình kinh tế lượng và dự báo điều được thực hiện trên cơ sở này.

■ HÌNH 2.3: **Độ nhọn của phân phối.**



Trung bình mẫu

Trung bình mẫu của một biến ngẫu nhiên X có n quan sát được ký hiệu là $\overline{X}$ và được định nghĩa như sau:

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \tag{2.28}$$

Trung bình mẫu được xem là một ước lượng của E(X), tức trung bình tổng thể. Một ước lượng đơn giản là một qui tắc, một công thức, hay một thống kê cho ta biết làm sao để ước lượng một đại lượng của tổng thể. Giả sử X có 7 quan sát với các giá trị như sau: 8, 9, 10, 11, 12, 13, 14. Vậy $\overline{X}$ = 11, và con số 11 này được gọi là một giá trị ước lượng của trung bình tổng thể. Nói cách khác, giá trị ước lượng đơn giản là giá trị bằng số nào đó của một ước lượng. Trong Excel, chúng ta sử dụng hàm =Average(X_1:X_7).

Phương sai mẫu

Phương sai mẫu được ký hiệu bằng S_x^2, là ước lượng của phương sai tổng thể σ_x^2. Phương sai mẫu được định nghĩa như sau:

$$S_x^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1} \tag{2.29}$$

Trong đó, n-1 được gọi là số bậc tự do (d.f.). Bậc tự do là số nguồn thông tin (piece of information) về một biến ngẫu nhiên. Để hiểu khái niệm này, chúng ta xét ví dụ sau đây.

■ BẢNG 2.6: **Định nghĩa khái niệm bậc tự do.**

Quan sát	X	$(X-\bar{X})$	$(X-\bar{X})^2$
1	8	-3	9
2	9	-2	4
3	10	-1	1
4	11	0	0
5	12	1	1
6	13	2	4
7	14	3	9
	Tổng	0	28

Ta biết rằng tổng độ lệch luôn luôn bằng không[4], nên để xem độ lệch của các giá trị X so với giá trị trung bình ta phải lấy độ lệch bình phương. Tổng của 7 độ lệch bình phương là 28, nhưng thực sự con số 28 này chỉ do 6 "nguồn" đóng góp, vì quan sát thứ tư trùng với giá trị trung bình. Như vậy, để xem độ lệch trung bình ta chỉ lấy 28 chia cho số nguồn thực sự tạo ra nó, tức 7-1 = 6. Vậy phương sai là 4,67 (là một giá trị ước lượng của phương sai tổng thể) và căn bậc hai của phương sai mẫu được gọi là độ lệch chuẩn mẫu. Độ lệch chuẩn (bằng 2.16) được xem như một thước đo xấp xỉ cho trung bình của 6 độ lệch tuyệt đối ở trên. Mở rộng cho trường hợp một biến ngẫu nhiên liên tục. Ta biết rằng trong tất cả các giá trị của một biến ngẫu nhiên liên tục chắc chắn sẽ có một giá trị trùng với giá trị trung bình. Cho nên, khi tính thước đo độ lệch trung bình, thì số bậc tự do luôn là (n-1). Hơn nữa, nếu ta không hiệu chỉnh mẫu của công thức phương sai mẫu thành (n-1) thì trung bình của tất cả các phương sai mẫu của tất cả các mẫu với n quan sát rút ra từ một tổng thể nhất định sẽ không trùng với phương sai tổng thể. Điều đó có nghĩa là phương sai mẫu đã ước lượng chệch phương sai tổng thể. Khi cỡ mẫu tăng lên, ví dụ n = 100, thì n-1 = 99 sẽ không có sự khác biệt nhiều so với n = 100. Tuy nhiên, khi cỡ mẫu nhỏ, thì đó có thể là một sự khác biệt đáng kể. Trong Excel, ta sử dụng hàm =Var(X_1:X_7) để tính phương sai mẫu.

[4] Chứng minh: $\sum(X-\bar{X}) = \sum X - \sum \bar{X} = \sum X - n\bar{X} = \sum X - \sum X = 0$.

Hiệp phương sai mẫu

Hiệp phương sai mẫu giữa hai biến ngẫu nhiên X và Y là ước lượng của hiệp phương sai tổng thể, và được định nghĩa như sau:

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \qquad (2.30)$$

Trong Excel, ta sử dụng hàm = Covar(X_1:X_n;Y_1:Y_n) để tính giá trị của hiệp phương sai mẫu.

Hệ số biến thiên mẫu

Hệ số biến thiên mẫu của X được xác định bằng công thức sau đây:

$$V = \frac{S_X}{\overline{X}}.100 \qquad (2.31)$$

Hệ số tương quan mẫu

Hệ số tương quan mẫu giữa hai biến ngẫu nhiên X và Y là ước lượng của hệ số tương quan tổng thể, và được định nghĩa như sau:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})/(n-1)}{S_X S_Y} \qquad (2.32)$$

Trong Excel, ta sử dụng hàm =Correl(X_1:X_n;Y_1:Y_n) để tính giá trị của hệ số tương quan mẫu.

Độ nghiêng và độ nhọn mẫu

Để tính độ nghiêng và độ nhọn mẫu, ta sử dụng các mô men mẫu thứ ba và thứ tư như sau:

$$\text{Mô men thứ ba: } \frac{\sum(X - \overline{X})^3}{(n-1)} \qquad (2.33)$$

$$\text{Mô men thứ tư: } \frac{\sum(X-\overline{X})^4}{(n-1)} \qquad (2.34)$$

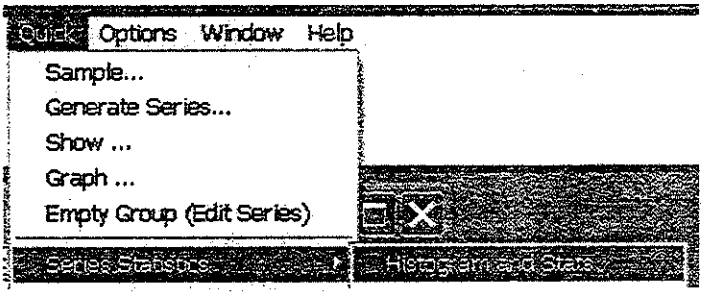
THAO TÁC TRÊN EVIEWS

Sử dụng tập tin **DATA2-1** vừa được tạo ra ở trên, và thực hiện theo các hướng dẫn sau đây:

Vẽ đồ thị tần suất và thống kê mô tả biến foodreal

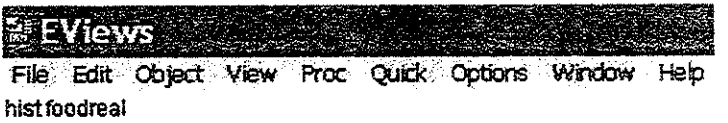
Chúng ta có thể thực hiện theo ba cách sau đây (tùy vào mức độ thông thạo trong việc sử dụng các phím tắt trên Eviews):

Cách 1: Quick/Series Statistics/Histogram and Stats

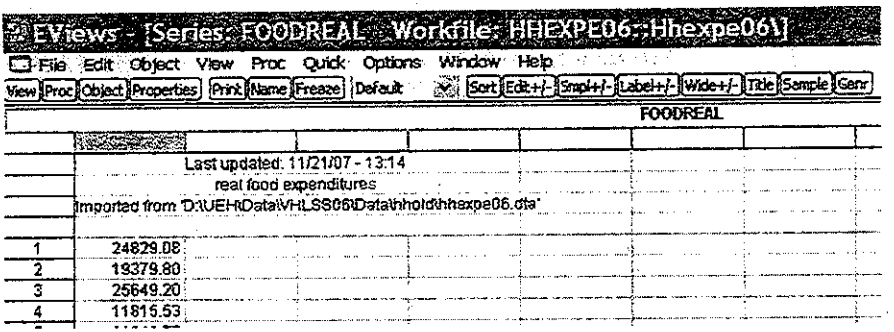


Sau đó nhập tên biến foodreal vào ô "Series Name", rồi chọn "OK".

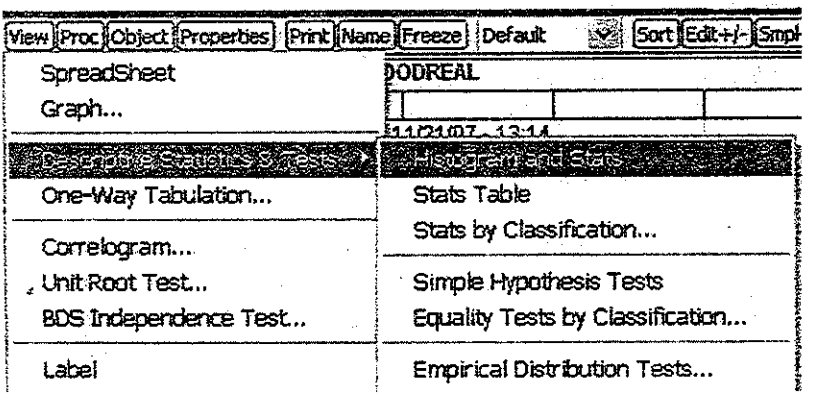
Cách 2: Từ cửa sổ lệnh ta nhập HIST FOODREAL, và nhấn Enter.



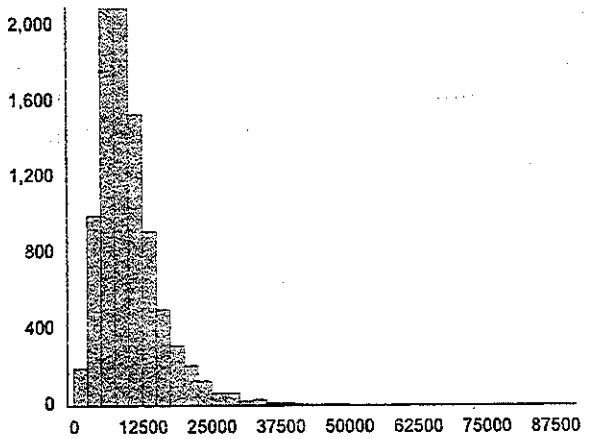
Cách 3: Chọn và mở biến FOODREAl (thường bằng cách chọn và nhấp đúp vào biến mà chúng ta quan tâm), ta sẽ có được một bảng tính (như dạng bảng tính trong Excel) như sau:



Từ 'View' trên góc trái của bảng tính, chọn **Descriptive Statistics & Tests/Histogram and Stats**, và ta sẽ có được kết quả như ở Hình 2.4.



■ HÌNH 2.4: **Kết quả thống kê mô tả trên Eviews.**



Series: FOODREAL	
Sample 1 9189	
Observations 9189	
Mean	10179.27
Median	8934.865
Maximum	88267.90
Minimum	632.0662
Std. Dev.	5799.511
Skewness	2.172539
Kurtosis	14.04996
Jarque-Bera	53978.19
Probability	0.000000

Trong bảng tóm tắt thống kê ở Hình 2.4, Jarque-Bera (hay thường gọi thống kê JB) là một thống kê rất quan trọng cho chúng ta biết chuỗi FOODREAL có phân phối chuẩn hay không. Thống kê JB được sử dụng rất phổ biến trong việc phân tích dữ liệu trước khi lựa chọn mô hình và kiểm định phần dư trong phân tích hồi quy. Thống kê JB có phân phối χ^2 với d.f. = 2. Một cách ngắn gọn, thống kê JB được tính theo công thức sau đây (đối với một biến ngẫu nhiên bất kỳ):

$$JB = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \qquad (2.35)$$

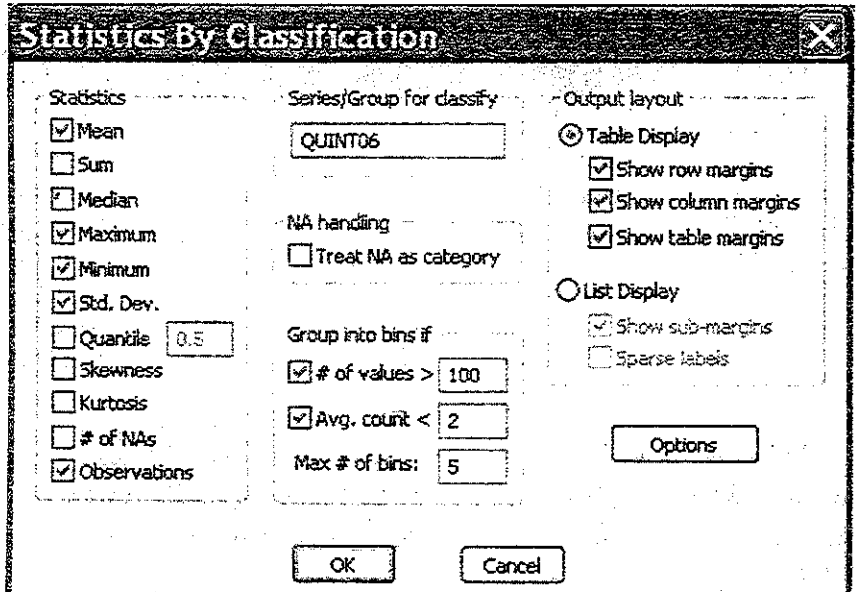
Và 'Probability' là xác suất để χ^2 lớn hơn 53978,19. Như chúng ta sẽ được giới thiệu ở phần sau, đây chính là giá trị xác suất *p*. Giá trị xác suất *p* là một thông tin rất quan trọng thường được sử dụng để kiểm định giả thiết.

Thống kê mô tả theo nhóm

Một câu hỏi quan trọng mà chúng ta rất hay đặt ra khi phân tích dữ liệu (ví dụ biến FOODREAL) là liệu có sự khác biệt nào trong giá trị trung bình và phương sai giữa các nhóm khác nhau trong dữ liệu (ví dụ giữa thành thị và nông thôn, giữa năm nhóm thu nhập (QUINT06), hoặc giữa tám vùng trong cả nước). Bước đầu tiên để kiểm định loại giả thiết như vậy thường đòi hỏi chúng ta phải lập bảng so sánh để có cái nhìn sơ bộ. Và thao tác trên Eviews như sau:

Bước 1: Mở biến FOODREAL dưới dạng bảng tính (như bước 3 ở trên).

Bước 2: View/Statistics and Tests/Stats by Classification, và chúng ta thấy xuất hiện một hộp thoại như sau:



Trong hộp thoại này, chúng ta nhập tên biến cần phân loại vào ô "Series/Group for classify", chọn OK và ta có kết quả như sau:

■ BẢNG 2.7: **Thống kê mô tả FOODREAL theo QUINT06.**

QUINT06	Mean	Median	Max	Min.	Std. Dev.	Obs.
1	6534.200	6296.216	21794.90	632.0662	2750.651	1799
2	8024.485	7742.791	24340.25	741.2529	3218.301	1820
3	9495.203	9124.661	46164.63	1232.651	3918.774	1856
4	11464.19	10842.28	39560.23	1018.557	4948.864	1871
5	15249.67	13772.25	88267.90	1482.371	8024.478	1843
All	10179.27	8934.865	88267.90	632.0662	5799.511	9189

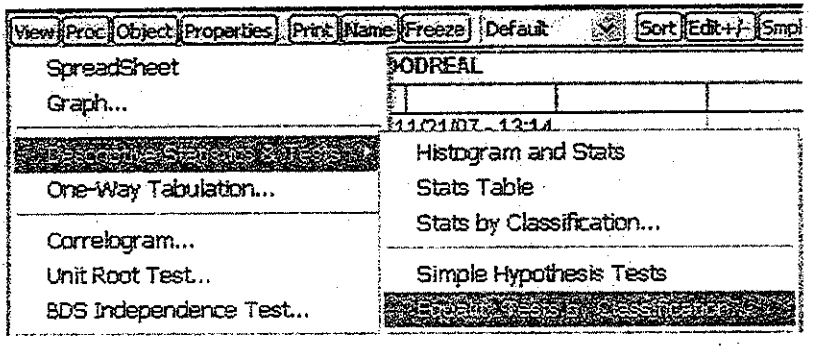
Trong Bảng 2.7, ta có thể quan sát thấy có sự khác biệt trong FOODREAL giữa các nhóm thu nhập khác nhau (từ nhóm 20% nghèo nhất đến nhóm 20% giàu nhất).

Kiểm định trung bình bằng nhau

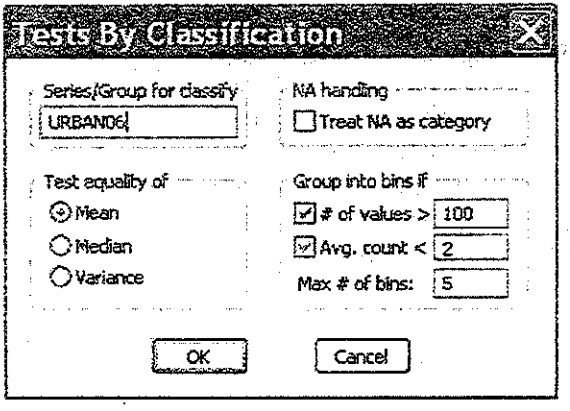
Từ giá trị thống kê mô tả như ở Bảng 2.7, chúng ta có thể đặt thêm câu hỏi rằng liệu sự khác biệt trong giá trị trung bình giữa năm nhóm thu nhập có ý nghĩa thống kê hay không. Lưu ý rằng, nếu chỉ so sánh trung bình giữa hai nhóm (ví dụ giữa thành thị và nông thôn), thì chúng ta chỉ cần sử dụng thống kê t (với H_0: $\mu_1 = \mu_2$ và H_1: $\mu_1 \neq \mu_2$). Nhưng, nếu chúng ta muốn kiểm định có sự khác biệt trong giá trị trung bình giữa nhiều nhóm khác nhau (ví dụ giữa năm nhóm thu nhập), thì chúng ta phải sử dụng giả thiết đồng thời và thống kê *F* ((với H_0: $\mu_1 = \mu_2 = \ldots = \mu_5$ và H_1: Ít nhất có một nhóm nào đó khác nhau). Thao tác trên Eviews cho hai loại kiểm định này là như nhau, nhưng kết quả kiểm định có khác nhau một chút.

Kiểm định xem có sự khác nhau trong FOODREAL giữa thành thị và nông thôn (URBAN06)

Bước 1: View/Descriptive Statistics and Tests/Equality Tests by Classification



Bước 2: Nhập tên biến cần phân loại vào ô "Series/Group for classify"



Sau khi chọn OK, ta sẽ có kết quả kiểm định như ở Bảng 2.8 như sau:

■ BẢNG 2.8: **Kiểm định trung bình bằng nhau giữa hai nhóm.**

Test for Equality of Means of FOODREAL
Categorized by values of URBAN06
Sample: 1 9189

Method	df	Value	Probability
t-test	9187	27.02506	0.0000
Satterthwaite-Welch t-test*	2995.189	22.08290	0.0000
Anova F-test	(1, 9187)	730.3538	0.0000
Welch F-test*	(1, 2995.19)	487.6547	0.0000

Kiểm định xem có sự khác nhau trong FOODREAL giữa tám vùng kinh tế (REG8)

■ BẢNG 2.9: **Kiểm định trung bình bằng nhau giữa nhiều nhóm.**

Test for Equality of Means of FOODREAL
Categorized by values of REG8
Sample: 1 9189

Method	df	Value	Probability
Anova F-test	(7, 9181)	53.45420	0.0000
Welch F-test*	(7, 2953.51)	51.91643	0.0000

Hệ số tương quan và hiệp phương sai

Để xác định hệ số tương quan và hiệp phương sai giữa các biến trên Eviews, ta vào Quick/Group Statistics/Correlations hoặc Quick/Statistics/Covariances.

PHÂN TÍCH ĐỒ THỊ VÀ CHUYỂN HÓA DỮ LIỆU

Một công cụ khá phổ biến khác trong phân tích kinh tế lượng và dự báo là xem xét các dạng đồ thị và chuyển hóa dữ liệu. Đối với dữ liệu chéo, việc xem xét đồ thị giúp chúng ta nhận diện các quan sát có giá trị bất thường (outliers), vốn có thể dẫn đến sự sai lệch trong phân tích hoặc phương sai thay đổi trong phân tích hồi quy. Nhờ việc xem xét đồ thị giữa các biến, chúng ta có thể lựa chọn được các dạng hàm thích hợp cho dữ liệu. Đối với dữ liệu chuỗi thời gian (sẽ được phân tích một cách chi tiết ở chương 3), thì xem xét đồ thị là một bước không thể thiếu trong việc lựa chọn các mô hình dự báo phù hợp. Ngoài ra, từ thông tin dưới dạng đồ thị có thể cho phép chúng ta biết nên chuyển hóa dữ liệu như thế nào cho thích hợp với các mục tiêu dự báo.

PHÂN TÍCH ĐỒ THỊ

Trong Eviews, chúng ta có thể vẽ đồ thị theo nhiều cách khác nhau. Tương tự như thống kê mô tả, chúng ta có thể có ba cách sau đây (sử dụng tập tin **DATA2-3**):

Cách 1: Quick/Graph, rồi nhập tên biến (ví dụ GDP) vào ô "Series Name", rồi chọn "OK".

Cách 2: Từ cửa sổ lệnh ta nhập PLOT GDP (hoặc HIST GDP, SCAT GDP, v.v...), và nhấn Enter.

Cách 3: Chọn và mở biến GDP (thường bằng cách chọn và nhấp đúp vào biến mà ta quan tâm), ta sẽ có được một bảng tính (như ở phần trên), rồi từ 'View' trên gốc trái của bảng tính, chúng ta chọn **GRAPH**.

Tuy nhiên, thông thường chúng ta nên chọn theo cách 1, vì cách này sẽ cung cấp cho chúng ta nhiều sự lựa chọn khác nhau. Theo cách này, chúng ta có thể liệt kê cùng một lúc nhiều biến khác nhau hơn. Dưới đây là các dạng đồ thị quan trọng trong phân tích dự báo.

Đồ thị hệ trục kép

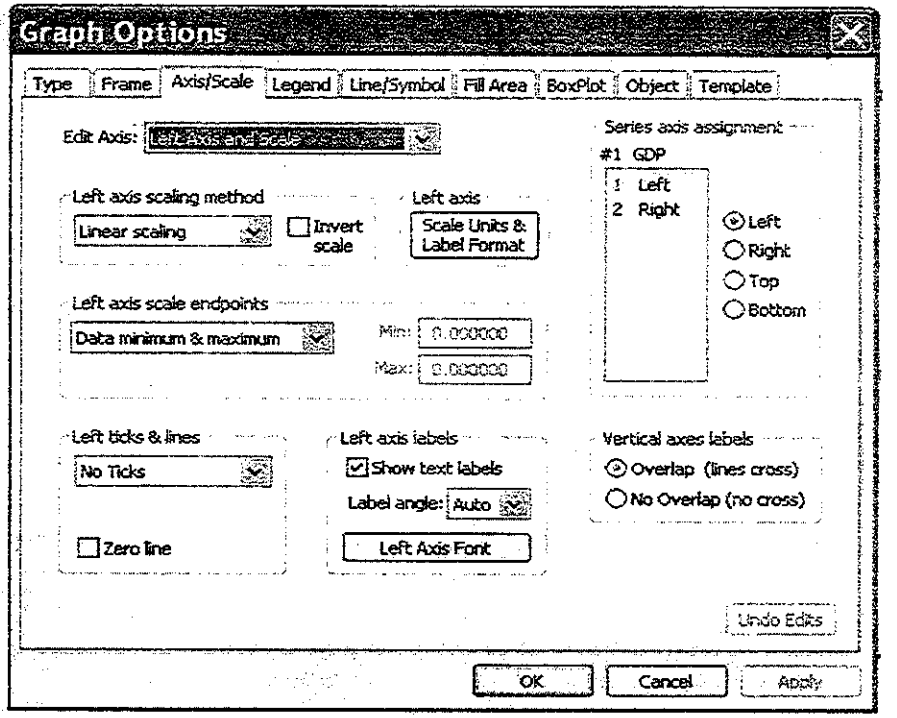
Khi muốn xem xét và đánh giá mối quan hệ giữa hai hay nhiều chuỗi thời gian với các thước đo khác nhau (ví dụ GDP đo bằng triệu đôla và lãi suất (RS) đo bằng phần trăm), thì ta nên sử dụng đồ thị hệ trục kép.

Bước 1: Quick/Graph, nhập tên hai biến GDP và RS, và thấy xuất hiện hộp thoại Chart Options.

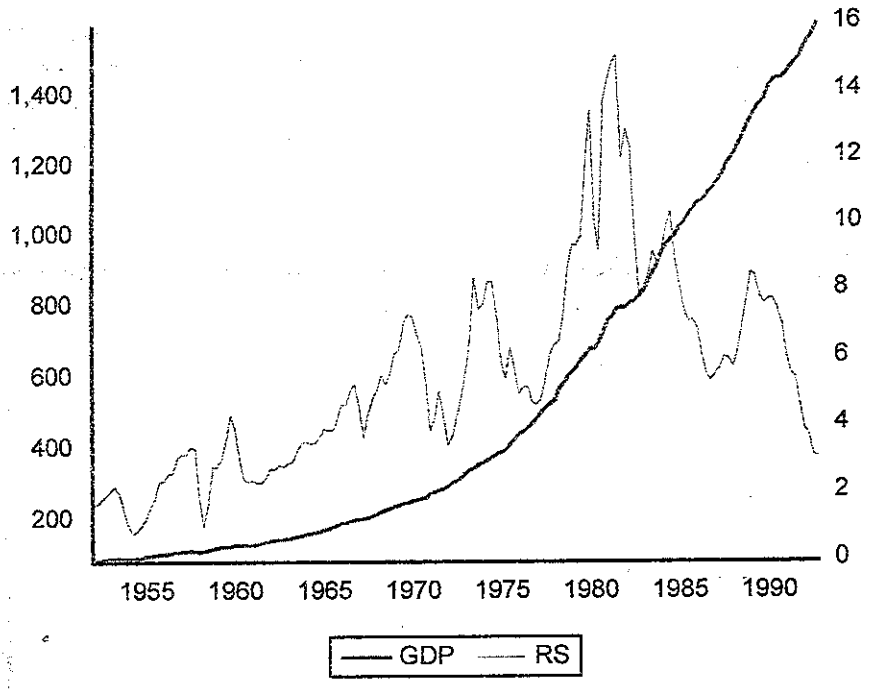
Bước 2: Khai báo một số lựa chọn quan trọng như sau:

- Type: Loại đồ thị. Ở đây ta chọn dạng Line & Symbol.
- Frame: Khung ngoài đồ thị. Ở đây chúng ta có thể chọn màu và khung nền cho đồ thị (chủ yếu là color và frame border).

- Axis/Scale: Trục và thước đo. Ở đây chúng ta tập trung vào các vấn đề sau đây: (1) Edit Axis: Chọn trục trái, phải, trên, hay dưới để định dạng trục theo tùy chọn; (2) Ticks and Lines: Tùy vào việc chúng ta muốn có các lằn gạch phân cách hay không (nếu không, ta chọn No ticks); (3) Top Axis Labels: Cho biết có biểu hiện tên nhãn hay không và đặc biệt là chọn font; (4) Axis Scale Endpoints: Thông thường ta nên chọn "Data Minimum & Maximum"; (5) Series Axis Assignment:



■ HÌNH 2.5: **Đồ thị hệ trục kép.**



Thông thường thì Eviews sẽ biểu hiện tất cả các chuỗi trên cùng trục trái (left), nên nếu ta chọn biến RS và nhấp vào "Right" thì ta sẽ có đồ thị hệ trục kép (bên trái là GDP và bên phải là RS); (6) Vertical Axes Labels: Thông thường ta nên chọn "Overlap". Sau khi chọn "OK", ta sẽ có đồ thị như ở Hình 2.5.

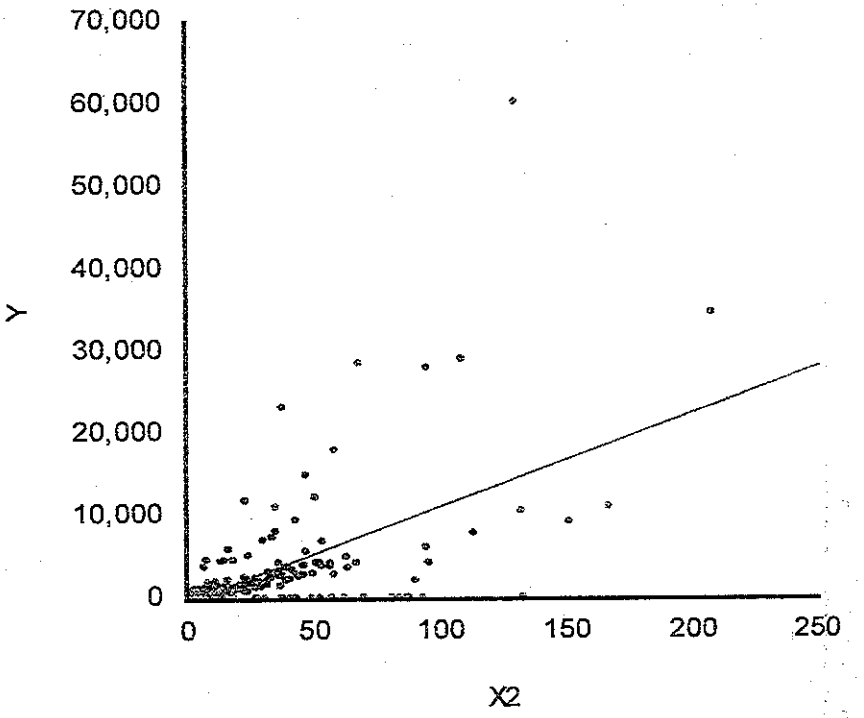
- Line/Symbol: Chọn dạng đường nét đồ thị theo tùy chọn.

Đồ thị phân tán có đường hồi quy

Đây là dạng đồ thị sử dụng phổ biến nhằm xem hai biến bất kỳ, thông thường giữa biến phụ thuộc và biến giải thích, có mối quan hệ với

nhau hay không. Ngoài ra, thông qua đồ thị này chúng ta cũng có thể phát hiện những quan sát có giá trị bất thường có thể ảnh hưởng đến kết quả phân tích. Ví dụ, sử dụng tập tin **DATA2-4** (gồm 266 công ty bất động sản), với các biến được định nghĩa như sau: Y = Doanh số năm 2008 (triệu đôla); X_2 = Tổng số lao động (ngàn người); X_3 = Chi tiêu vốn hữu hình (triệu đôla); X_4 = Chi tiêu vốn vô hình (triệu đôla); X_5 = Giá vốn hàng bán (triệu đôla); X_6 = Chi phí quản lý (triệu đôla); X_7 = Chi phí quảng cáo và bán hàng (triệu đôla); và X_8 = Chi phí nghiên cứu & phát triển (triệu đôla).

■ HÌNH 2.6: **Đồ thị phân tán có đường hồi quy.**

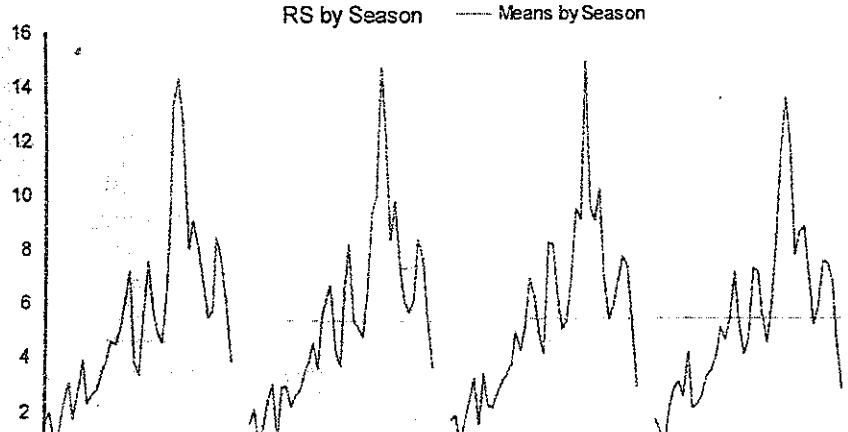


Tương tự, nếu ta lần lượt vẽ các đồ thị phân tán giữa Y và các biến giải thích khác thì nhận thấy rằng trong bộ số liệu gồm 266 quan sát này có một quan sát bất thường. Đúng như thế, nếu không loại bỏ quan sát này thì mô hình hồi quy sẽ bị hiện tượng phương sai thay đổi. Khi vẽ đồ thị này thì chúng ta cần lưu ý hai điểm sau: (1) Ở phần loại đồ thị ta chọn "Scatter"; và (2) Ở "Details" ta chọn "Regression Line" ở ô "Fit Lines".

Đồ thị một chuỗi theo mùa

Đây là loại đồ thị rất hữu ích đối với các chuỗi thời gian có yếu tố mùa vụ, đặc biệt là theo quý. Quan sát đồ thị này có thể cung cấp chúng ta thông tin có nên quan tâm đến các mô hình có yếu tố mùa vụ hay không. Để vẽ đồ thị này, ta chọn Quick/Graph, nhập tên biến RS (tập tin **DATA2-3**), rồi chọn "Seasonal Graph".

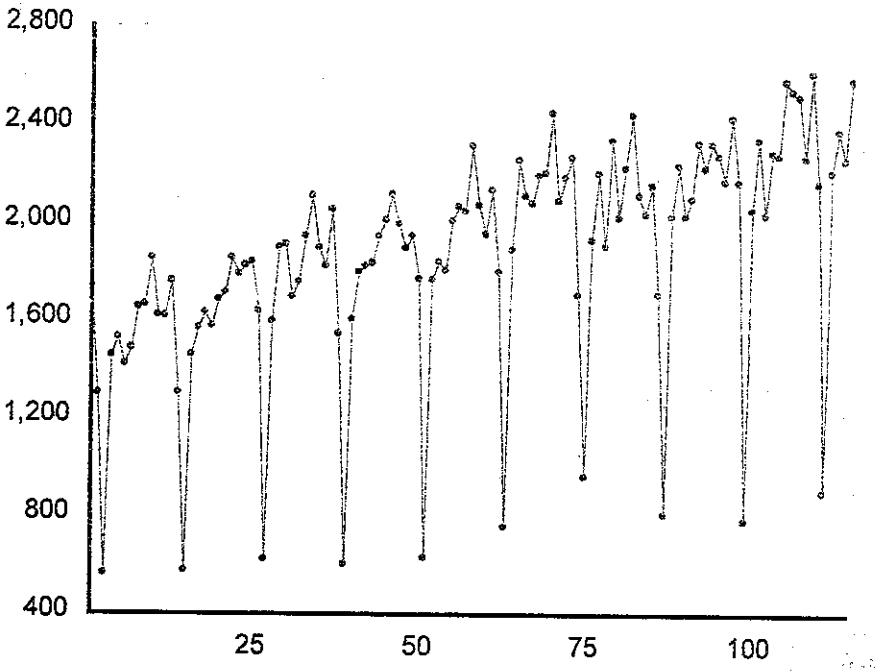
■ HÌNH 2.7: **Đồ thị theo mùa vụ.**



Đồ thị một chuỗi theo thời gian

Trong phân tích dự báo chuỗi thời gian, việc vẽ và quan sát đồ thị của một chuỗi dữ liệu cần dự báo là một công việc gần như là bắt buộc. Như chúng ta sẽ thảo luận ở chương 3, một chuỗi thời gian có thể bao gồm nhiều thành phần khác nhau. Cho nên, việc quan sát đồ thị sẽ có ý nghĩa hết sức quan trọng trong việc giới hạn số các mô hình có thể phù hợp với dữ liệu quá khứ (chương 3). Ví dụ, sử dụng tập tin **DATA2-5** và chọn Quick/Graph/Sales, ta có đồ thị sau đây:

■ HÌNH 2.8: **Đồ thị theo thời gian.**



Như vậy, quan sát đồ thị ta thấy doanh số là một chuỗi vừa số yếu tố xu thế vừa có yếu tố mùa vụ.

CHUYỂN HÓA DỮ LIỆU

Chuyển hóa dữ liệu là một công việc phổ biến đối với cả các mô hình dữ liệu chéo và dữ liệu chuỗi thời gian nhằm giúp người phân tích lựa chọn mô hình phù hợp nhất với dữ liệu sẵn có hoặc mục đích nghiên cứu. Đối với dữ liệu chéo, chúng ta thường chuyển hóa dữ liệu gốc thành dữ liệu dạng logarith để chuyển các mô hình phi tuyến thành các mô hình tuyến tính hoặc để khắc phục một số hiện tượng hay gặp trong phân tích hồi quy như đa cộng tuyến, phương sai thay đổi, và tự tương quan. Đối với dữ liệu chuỗi thời gian, việc chuyển hóa dữ liệu có tính thường xuyên hơn không chỉ nhằm tạo ra các chuỗi dữ liệu mới từ các dữ liệu có sẵn như tạo ra các chỉ số kinh tế, tốc độ tăng trưởng, suất sinh lợi, hoặc các chuỗi dữ liệu thực, mà còn giúp người phân tích xác định các mô hình dự báo phù hợp với dữ liệu như lấy sai phân, tạo biến trễ, biến mùa vụ, hay biến xu thế. Trong nhiều trường hợp chúng ta có thể cần làm trơn dữ liệu bằng các kỹ thuật lấy giá trị trung bình, hoặc khử yếu tố mùa, và tạo các biến tương tác. Ngoài ra, chúng ta có thể tạo các biến tổng hợp từ tháng, quý thành dữ liệu năm hoặc ngược lại[5].

Chuyển sang dạng logarith

Hai trường hợp cần được chuyển hóa các biến dữ liệu gốc sang dữ liệu dạng logarith sẽ được đề cập nhiều trong cuốn sách này là ước lượng các hệ số co giãn trong mô hình hồi quy bội và nhận dạng độ trễ thích hợp trong các mô hình ARIMA(p,d,q). Bây giờ, giả sử chúng ta có hàm sản xuất cà phê, trong đó Y_t là sản lượng cà phê (tấn/năm), K_t là vốn (triệu đồng/năm), và L_t là lao động (1.000 giờ/năm) như sau:

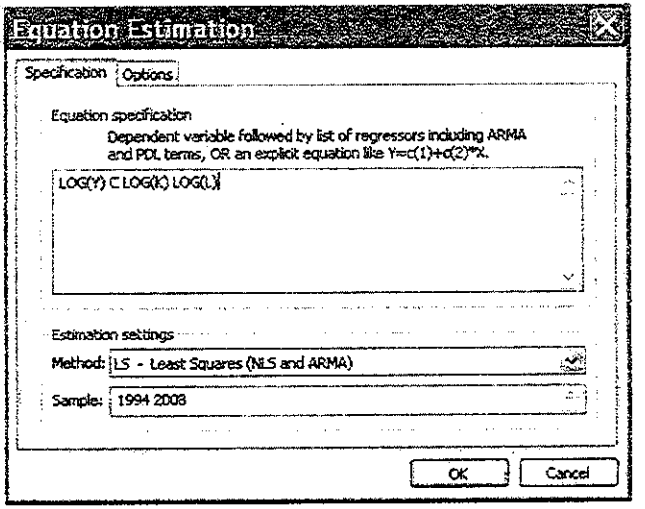
$$Y_t = AK_t^{\beta_2} L_t^{\beta_3} e^{u_t} \qquad (2.36)$$

Nếu lấy logarith tự nhiên hai vế của phương trình (2.36), ta sẽ có được một hàm hồi quy tuyến tính như sau:

$$\ln Y_t = \beta_1 + \beta_2 K_t + \beta_3 L_t + u_i \qquad (2.37)$$

[5] Xem "Nguyễn Trọng Hoài, 2003, Mô hình hóa chuỗi thời gian và dự báo kinh tế, NXB Đại học Quốc Gia TP.HCM".

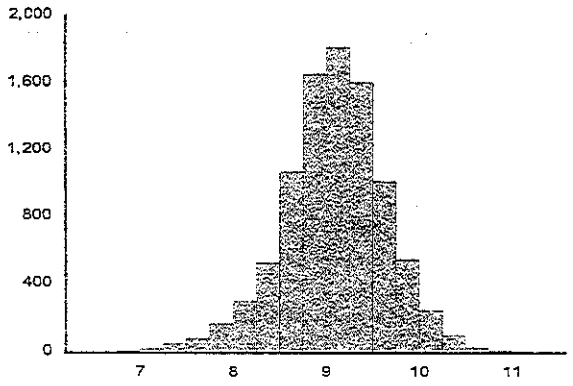
Trong Eviews, logarith tự nhiên được qui ước là hàm "log". Ví dụ, muốn tạo biến lnY, ta nhấp vào "GENR" rồi nhập lnY=log(Y), OK. Thông thường, ta sử dụng dạng hàm này một cách trực tiếp mà không cần tạo thêm một biến dạng logarith như trong Excel hay Stata. Để minh họa, chúng ta sử dụng tập tin **DATA2-6** và chọn Quick/Estimate Equation, rồi nhập vào hộp thoại ước lượng phương trình hồi quy như sau:



Như vậy, ta có thể ước lượng mô hình hồi quy (vẽ đồ thị, hay bất cứ thao tác nào) dạng logarith một cách trực tiếp mà không cần tao ra các biến mới (để việc quản lý dữ liệu được thuận lợi hơn).

Khi dữ liệu gốc có mức độ phân tán cao hoặc có một số quan sát có giá trị bất thường (nếu loại bỏ có thể dẫn đến hậu quả thiết quan sát), chúng ta có thể chuyển hóa sang dữ liệu dạng logarith để thuận lợi trong việc nhận dạng và phân tích dữ liệu hơn. Ví dụ, sử dụng tập tin **DATA2-1**, và chọn Quick/Series Statistics/Histogram and Stats, rồi nhập biến log(FOODREAL) vào ô "Series Name", ta có kết quả như ở Hình 2.9.

■ HÌNH 2.9: **Thống kê mô tả của biến log(FOODREAL).**



Series: LOG(FOODREAL)	
Sample 1 9189	
Observations 9189	
Mean	9.084815
Median	9.097716
Maximum	11.38813
Minimum	6.448994
Std. Dev.	0.546392
Skewness	-0.324339
Kurtosis	3.853278
Jarque-Bera	439.8723
Probability	0.000000

Nếu so với Hình 2.4, thì đồ thị trong Hình 2.9 có vẻ rõ ràng hơn vì việc lấy logarith đã làm giảm giá trị của các quan sát quá cao.

Biến trễ, biến tới, sai phân, mùa vụ, và biến xu thế

Khi làm việc với dữ liệu chuỗi thời gian, ta thường xử lý dữ liệu bằng cách chuyển hóa sang dạng trễ, tới, sai phân, hoặc tạo thêm các biến giả mùa vụ.

Biến trễ, tới và sai phân

- Biến trễ một giai đoạn (X_{t-1}): x(-1)
- Biến trễ k giai đoạn (X_{t-k}): x(-k)
- Biến tới một giai đoạn (X_{t+1}): x(1)
- Biến tới k giai đoạn (X_{t+k}): x(k)
- Sai phân bậc một ($\Delta X = X_t - X_{t-1}$): d(x)
- Sai phân bậc k ($\Delta^k X = X_t - X_{t-k}$): d(x,k)
- Sai phân bậc một của biến trễ dạng log tự nhiên: dlog(x)
- Trung bình trượt k giai đoạn: @movav(x,k)

Để tạo một biến mới (ví dụ sai phân của X) ta có thể chọn một trong hai cách sau đây. Thứ nhất, trên cửa sổ lệnh ta nhập **genr dx=d(x)**. Thứ hai, ta có thể nhấp vào **genr** trên thanh công cụ của cửa sổ tập tin Eviews và nhập **dx=d(x)**.

Biến giả mùa vụ

- Tạo ra một biến giả theo quý có giá trị là 1 đối với quý 2 và giá trị là 0 đối với các quý khác: @seas(2) hoặc @quarter=2.
- Tạo ra một biến giả theo tháng có giá trị là 1 đối với tháng 2 và giá trị 0 đối với các tháng khác: @month(2) hoặc @month=2. Lưu ý, đối với dữ liệu theo tháng, ta cũng tạo được các biến giả theo quý theo hàm @seas hoặc @quarter.

Biến xu thế

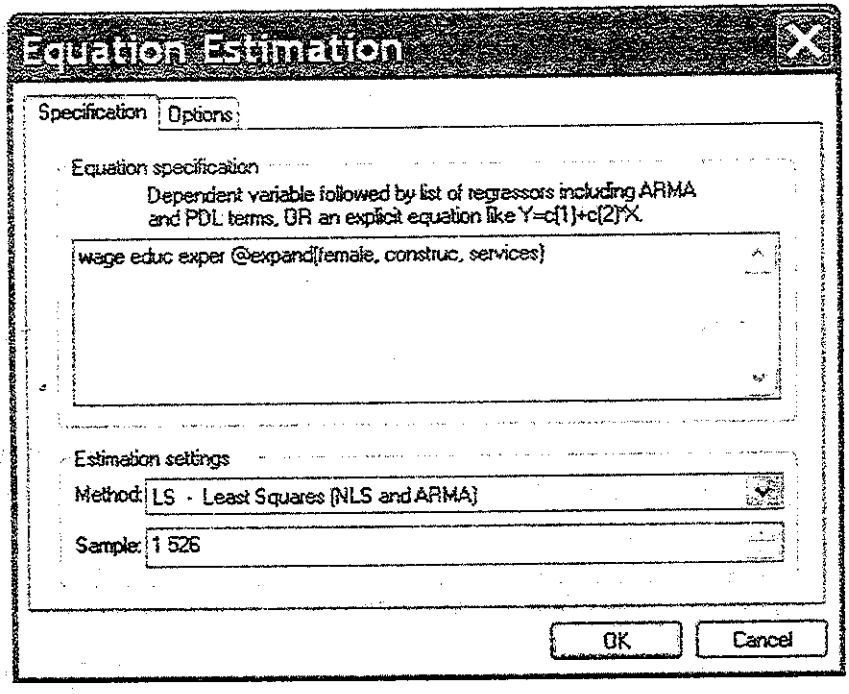
Biến xu thế là một biến có giá trị từ 1 đến n, trong đó 1 đại diện cho quan sát đầu tiên trong dữ liệu và n đại diện cho quan sát cuối cùng trong chuỗi dữ liệu.

- Tạo biến xu thế đối với dữ liệu theo năm, ví dụ bắt đầu từ năm 1990 đến 2008, ta làm như sau: Trên cửa sổ lệnh ta nhập **genr t=@trend(1989)**.
- Tạo biến xu thế đối với dữ liệu theo quý, ví dụ bắt đầu từ 2000Q2 đến 2008Q3, ta làm như sau: Trên cửa sổ lệnh ta nhập **genr t=@trend(2000Q1)**.
- Tạo biến xu thế đối với dữ liệu theo tháng, ví dụ bắt đầu từ 2002M3 đến 2008M2, ta làm như sau: Trên cửa sổ lệnh ta nhập **genr t=@trend(2002M2)**.

Nhắc lại rằng, khi phân tích dữ liệu, dự báo và hồi quy, ta không cần phải tạo thêm các biến mới như vậy mà thường sử dụng trực tiếp các hàm từ các dữ liệu gốc. Ví dụ, ta có thể hồi quy trực tiếp như sau: **y c x @trend(1989)** hoặc **log(y) c log(x)**. Lý do ta không cần tạo thêm biến mới là để cho tập tin Eviews được đơn giản và dễ quản lý hơn.

Biến giả trong Eviews

Để đưa biến giả vào mô hình hồi quy, thay vì phải tạo ra các biến này trên Excel, rồi chuyển qua Eviews, chúng ta có thể sử dụng hàm sau đây: @EXPAND(D1, D2, v.v). Ví dụ, sử dụng tập tin **DATA2-7** để xây dựng hàm hồi quy biến wage theo các biến giáo dục, năm kinh nghiệm, giới tính, ngành xây dựng, và ngành dịch vụ, thực hiện như sau: (1) Quick/Estimate Equation, (2) Khai báo biến phụ thuộc và các biến giải thích vào hộp thoại ước lượng phương trình. Sau khi chọn "OK", ta có kết quả như trong Bảng 2.10 (tập tin **DATA2-7**).



■ BẢNG 2.10: **Minh họa hàm @Expand trong Eviews.**

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
EDUC	0.641829	0.053761	11.93861	0.0000
EXPER	0.072268	0.010938	6.607226	0.0000
CONSTRUC=0,SERVICES=0	-3.298405	0.775144	-4.255218	0.0000
CONSTRUC=0,SERVICES=1	-4.693916	0.855767	-5.485042	0.0000
CONSTRUC=1,SERVICES=0	-2.373707	0.944839	-2.512288	0.0123

R-squared	0.241855	Mean dependent var	5.896103
Adjusted R-squared	0.235833	S.D. dependent var	3.693086
S.E. of regression	3.228372	Akaike info criterion	5.191293
Sum squared resid	5430.062	Schwarz criterion	5.231838
Log likelihood	-1360.310	Durbin-Watson stat	1.819868

Các hàm phổ biến trong Eviews

Hầu hết các hàm trong Eviews đều bắt đầu bằng ký hiệu @, ví dụ @mean(y) nghĩa là lấy giá trị trung bình của chuỗi y cho toàn bộ mẫu hiện hành. Có ba nhóm hàm chuỗi hay sử dụng trong Eviews: hàm toán (mathematical functions), hàm tập tin Eviews (workfile functions), và hàm dãy số (string functions). Để tìm hiểu thêm về các hàm này, ta có thể tham khảo ở **Help/Command & Programming Reference**, hoặc **Help/Quick Help Reference**, ở đây chỉ trình bày một số hàm hay sử dụng trong giáo trình này.

■ BẢNG 2.11: **Các hàm phổ biến trong Eviews.**

Tên hàm	Thao tác trên Eviews
Hệ số tương quan giữa X và Y	@cor(x,y)
Hiệp phương sai giữa X và Y	@cov(x,y)
Số quan sát của biến X	@obs(x)
Số quan sát của hồi quy	@regobs
Giá trị trung bình của X	@mean(x)

Tên hàm	Thao tác trên Eviews
Giá trị trung vị của X	@median(x)
Giá trị nhỏ nhất của X	@min(x)
Giá trị lớn nhất của X	@max(x)
Độ lệch chuẩn của X	@stdev(x)
Phương sai của X	@var(x)
Độ nghiêng của X	@skew(x)
Độ nhọn của X	@kurt(x)
Giá trị tổng của X	@sum(x)
Giá trị tổng bình phương của X	@sumsq(x)
Tổng bình phương phần dư	@sse
Giá trị tuyệt đối của X	@abs(x)
Antilog của X (e^X)	@exp(x)
Hàm nghịch đảo của X (1/X)	@inv(x)
Hàm logarith tự nhiên của X (ln(X))	@log(x) hoặc log(x)
Căn bậc hai của X	@sqrt(x) hoặc sqrt(x)
Tạo biến xu thế	@trend(base date)
Làm tròn số	@round(x)
Lũy thừa của X	x^2, x^3, ...

Nguồn: Eviews 6 User Guide.

Tốc độ tăng trưởng và suất sinh lợi

Trong các mô hình hồi quy dữ liệu chuỗi thời gian như CAPM, hoặc các mô hình ARIMA và ARCH, chúng ta thường xuyên sử dụng các biến dưới dạng tốc độ tăng trưởng hoặc suất sinh lợi, thì sự kết hợp giữa các hàm logarith và biến trễ có ý nghĩa hết sức quan trọng. Giả sử ta muốn sử dụng mô hình ARCH để dự báo suất sinh lợi trung bình và rủi ro của thị trường chứng khoán Việt Nam từ biến chỉ số giá chứng khoán VN-Index (ký hiệu là VNI_t). Nếu gọi R_{mt} là suất sinh lợi thị trường (theo ngày, tuần, hoặc tháng), thì theo lý thuyết tài chính ta có thể tính R_{mt} theo công thức truyền thống sau đây:

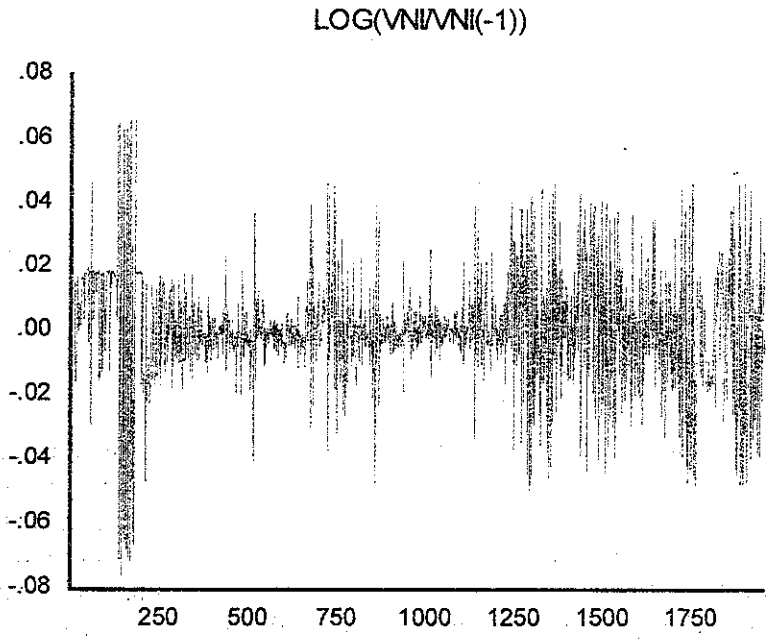
$$R_{mt} = \frac{VNI_t - VNI_{t-1}}{VNI_{t-1}} \tag{2.38}$$

Tuy nhiên, trong các nghiên cứu về tài chính, người ta thường sử dụng công thức mang tính xấp xỉ dưới dạng log sau đây hơn:

$$R_{mt} = \log\left(\frac{VNI_t}{VNI_{t-1}}\right) \tag{2.39}$$

Trong Eviews, ta có thể tạo biến R_{mt} theo cách đơn giản là $\log(VNI_t/VNI_t(-1))$. Sử dụng tập tin **DATA2-8** để vẽ đồ thị biến R_{mt} theo thời gian như Hình 2.10.

■ HÌNH 2.10: **Suất sinh lợi của thị trường chứng khoán Việt Nam.**



Nguồn: Reuters, 2009.

MỘT SỐ PHÂN PHỐI XÁC SUẤT CƠ BẢN

Cho đến đây, chúng ta đã đề cập đến một số thống kê quan trọng như thống kê *t*, thống kê *F*, hay thống kê χ^2. Nếu chúng ta là một nhà nghiên cứu có tư duy hợp lý, thì chúng ta sẽ có thể thắc mắc rằng tại sao lại sử dụng thống kê *t* cho các kiểm định các hệ số hồi quy? Tại sao sử dụng phân phối *t* cho kiểm định so sánh giá trị trung của hai nhóm? Tại sao sử dụng thống kê χ^2 để kiểm định xem một biến có phân phối chuẩn hay không? Tại sao sử dụng phân phối F để kiểm định các giả thiết đồng thời? Và nhiều câu hỏi tương tự như vậy.

Trong phần này ta sẽ xem xét một cách ngắn gọn bốn phân phối xác suất quan trọng thường gặp trong kinh tế lượng và dự báo; đó là, phân phối chuẩn, phân phối *t*, phân phối χ^2, và phân phối *F*. Các phân phối này rất quan trọng vì chúng giúp người phân tích tìm ra các phân phối xác suất của các ước lượng đang xem xét để biết cách suy diễn thống kê hợp lý. Nghĩa là, nếu chúng ta muốn kiểm định một giả thiết nào đó, thì điều đầu tiên chúng ta cần phải biết là ước lượng đó theo loại phân phối gì. Ngoài ra, đối với các phân phối *t*, *F*, và chi bình phương, chúng ta cũng cần lưu ý cách xác định số bậc tự do trong từng trường hợp vì nó rất quan trọng khi ta xác định các giá trị tra bảng tương ứng. Một khi chúng ta đã hiểu được luật phân phối, thì việc kiểm định giả thiết ở nội dung phân tích hồi quy sẽ trở nên dễ dàng hơn. Đừng bỏ qua nội dung này, nếu chúng ta muốn đi nghiên cứu sâu về phân tích dữ liệu và dự báo.

PHÂN PHỐI CHUẨN

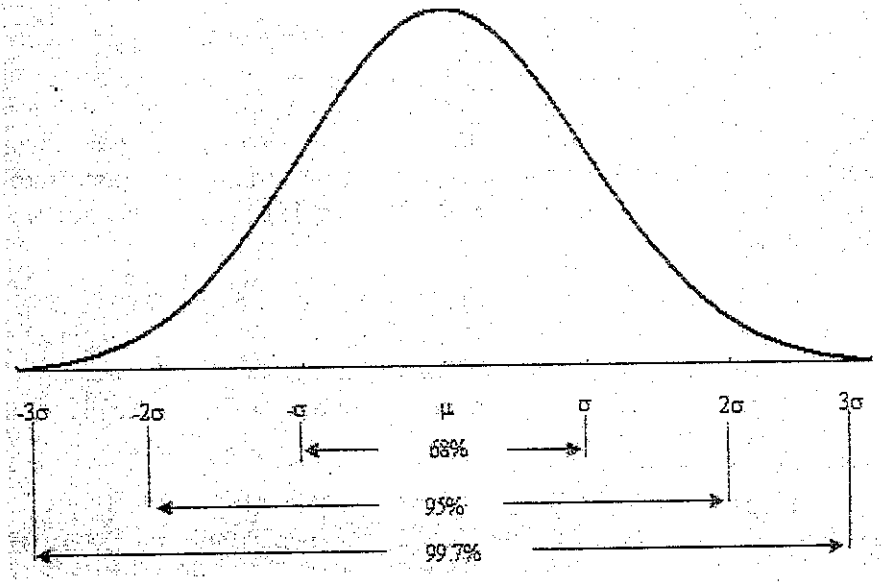
Có lẽ phân phối xác suất quan trọng nhất của một biến ngẫu nhiên liên tục là phân phối chuẩn. Kinh nghiệm cho thấy rằng phân phối chuẩn là một mô hình hợp lý cho một biến ngẫu nhiên liên tục với giá trị của nó phụ thuộc vào nhiều yếu tố, nhưng mỗi yếu tố chỉ có ảnh hưởng tương đối nhỏ lên giá trị của biến số đó (Gujarati, 2006). Phân phối chuẩn của một biến ngẫu nhiên X được thể hiện thông qua hai tham số cơ bản là giá trị trung bình và phương sai. Cụ thể như sau:

$$X \sim N(\mu_x, \sigma_x^2) \qquad (2.40)$$

Tính chất của phân phối chuẩn

- Đường phân phối chuẩn đối xứng quanh giá trị trung bình μ_x.
- Hàm phân phối xác suất của một biến ngẫu nhiên theo phân phối chuẩn cao nhất tại giá trị trung bình nhưng nhỏ dần về các cực trị của nó. Nghĩa là, xác suất để có một giá trị của một biến ngẫu nhiên theo phân phối chuẩn càng xa giá trị trung bình càng nhỏ.
- Theo kinh nghiệm, khoảng 68% diện tích dưới đường phân phối chuẩn nằm giữa giá trị $\mu_x \pm \sigma_x$, khoảng 95% diện tích nằm giữa $\mu_x \pm 2\sigma_x$, và khoảng 99.7% diện tích nằm giữa $\mu_x \pm 3\sigma_x$.

■ HÌNH 2.11: **Đồ thị phân phối chuẩn.**



- Một phân phối chuẩn được định nghĩa hoàn toàn bởi hai tham số μ_x và σ_x^2. Trong Excel, ta có thể dễ dàng tạo ra một hoặc nhiều biến có phân phối chuẩn bằng cách vào Tools/Data Analysis/Random Number Generation. Chúng ta hãy dành ít thời gian thực hành tạo các biến có phân phối chuẩn như hướng dẫn trên!

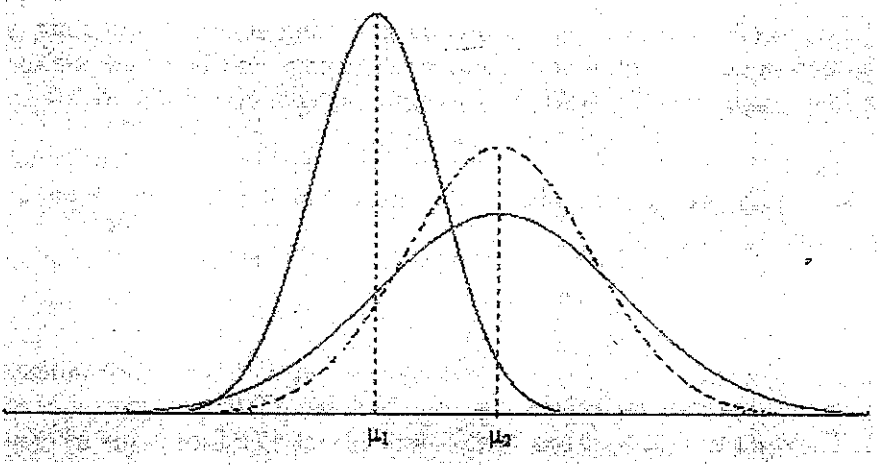
- Một kết hợp (hay một hàm) tuyến tính của một hay nhiều biến ngẫu nhiên theo phân phối chuẩn sẽ theo phân phối chuẩn – đây là một tính chất đặc biệt quan trọng của phân phối chuẩn trong kinh tế lượng và dự báo. Có thể nói, kết hợp với các giả định trong mô hình hồi quy tuyến tính cổ điển, thì đặc điểm này được xem như chìa khóa để hiểu tại sao các ước lượng OLS có phân phối chuẩn, và tại sao khi kiểm định lại chọn phân phối *t*.

- Đối với phân phối chuẩn, thì độ nghiêng S là 0 và độ nhọn K là 3.

PHÂN PHỐI CHUẨN HÓA

Mặc dù một phân phối chuẩn hoàn toàn được xác định bằng hai tham số, giá trị trung bình và phương sai tổng thể, nhưng các phân phối chuẩn có thể khác nhau hoặc ở giá trị trung bình, hoặc phương sai, hoặc cả hai.

■ HÌNH 2.12: **Phân phối chuẩn có trung bình và phương sai khác nhau.**



Ta không thể so sánh các phân phối chuẩn có các tính chất khác nhau. Cho nên, người ta quy về cùng một biến chuẩn hóa Z như sau:

$$Z = \frac{X - \mu_X}{\sigma_X} \tag{2.41}$$

Theo tính chất của phân phối chuẩn, nếu X là một biến ngẫu nhiên có trung bình là μ_x và phương sai là σ_x, $X \sim N(\mu_X, \sigma^2_X)$, thì Z là một kết hợp tuyến tính của X sẽ là một biến ngẫu nhiên có phân phối chuẩn với trung bình là không và phương sai là một, $Z \sim N(0, 1)$[6].

Như vậy, bất kỳ một biến ngẫu nhiên theo phân phối chuẩn với một giá trị trung bình và phương sai nhất định đều có thể được chuyển

---

[6] Chứng minh: $E(Z) = E\left(\frac{X-\mu_X}{\sigma_X}\right) = \frac{1}{\sigma_X}E(X-\mu_X) = 0$ do $E(X-\mu_x) = E(X) - E(\mu_x) = \mu_x - \mu_x = 0$. Và $Var(Z) = E[Z-E(Z)]^2 = E(Z^2)$, do $E(Z) = 0$, vậy $E(Z^2) = E\left(\frac{X-\mu_X}{\sigma_X}\right)^2 = \frac{1}{\sigma_X^2}E(X-\mu_X)^2 = \frac{1}{\sigma_X^2}\sigma_X^2 = 1$.

đổi thành một biến chuẩn hóa, điều này giúp đơn giản hóa rất nhiều việc tính xác suất. Để hiểu vai trò của phân phối chuẩn hóa, ta xem xét ví dụ sau đây.

Giả sử X, số lượt khách du lịch quốc tế hàng ngày của một công ty du lịch, theo phân phối chuẩn với giá trị trung bình là 70 và phương sai là 9; nghĩa là, $X \sim N(70,9)$. Hãy tính xác suất cho một ngày bất kỳ công ty có số khách du lịch quốc tế nhiều hơn 75 khách?

Ta thấy, do X theo phân phối chuẩn với giá trị trung bình và phương sai đã biết, nên ta có:

$$Z = \frac{75-70}{3} = \approx 1.67$$

sẽ theo phân phối chuẩn hóa với trung bình bằng 0 và phương sai bằng 1. Thay vì tìm $P(X > 75)$, ta có thể tìm $P(Z > 1.67)$. Lưu ý, giá trị hàm phân phối xác suất tích lũy (CDF) hay giá trị xác suất tích lũy của phân phối chuẩn hóa giữa các giá trị $Z = -3$ và $Z = 3$ (tại sao?). Theo kết quả tính toán trên Excel, thì xác suất Z nằm từ -3 đến 1.67 là 0.9525[7]. Cho nên,

$$P(Z > 1.67) = 1 - P(Z < 1.67) = 1 - 0.9525 = 0.0475$$

Vậy xác suất để một ngày bất kỳ công ty có số lượt khách du lịch nhiều hơn 75 người là 4.75%. Nếu đóng vai trò là một người ra quyết định, thì chúng ta sẽ suy nghĩ gì với con số này?

[7] Trong Excel, ta sử dụng hàm = NORMDIST(X, Mean, Standard_dev, Cumulative). Trong đó, "X" là giá trị cần tính xác suất tích lũy (1.67), "Mean" và "Standard_dev" ở đây lần lượt là trung bình (0) và độ lệch chuẩn (1) của biến X, và "Cumulative" có hai lựa chọn là "True" (đồng ý tính xác suất tích lũy) và "False" (không tính xác suất tích lũy). Ở trường hợp đang xét, ta chọn "True". Ngược lại, nếu ta đã biết xác suất tích lũy, giá trị trung bình và phương sai thì ta dễ dàng tính giá trị của biến đó như sau: =NORMINV(0.9525,0,1) = 1.67.

PHÂN PHỐI XÁC SUẤT CỦA TRUNG BÌNH MẪU

Đây là cầu nối giữa phân phối chuẩn và phân phối *t*. Cho nên, chúng ta nên dành ít thời gian xem qua phần giới thiệu dưới đây.

Giả sử ta chọn ngẫu nhiên một mẫu với *n* quan sát gồm các giá trị $X_1, X_2, \ldots, X_n$ từ một tổng thể có cùng hàm phân phối xác suất. Nếu ta thực hiện m mẫu như thế thì giá trị trung bình mẫu $\overline{X}$ sẽ là một biến ngẫu nhiên. Như vậy, vấn đề đặt ra là $\overline{X}$ sẽ có phân phối như thế nào?

Lý thuyết thống kê cho rằng, nếu $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên từ một tổng thể có phân phối chuẩn với trung bình μ_x và phương sai σ_x^2, thì trung bình mẫu, $\overline{X}$, cũng theo phân phối chuẩn với trung bình μ_x nhưng phương sai[8] $\frac{\sigma_x^2}{n}$. Nghĩa là,

$$\overline{X} \sim N(\mu_x, \frac{\sigma_x^2}{n}) \qquad (2.42)$$

Căn bậc hai của phương sai trung bình mẫu, $\frac{\sigma_x}{\sqrt{n}}$, được gọi là sai số chuẩn (se) của $\overline{X}$, tương tự như khái niệm độ lệch chuẩn. Lưu ý, căn bậc hai của phương sai của một biến ngẫu nhiên được gọi là độ lệch chuẩn (s.d.), và căn bậc hai của một ước lượng được gọi là sai số chuẩn (se).

---

[8] Chứng minh: Do $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ nên ta có:

$$E(\overline{X}) = \frac{1}{n}[E(X_1) + E(X_2) + \ldots + E(X_n)] = \frac{1}{n}[\mu_X + \mu_X + \ldots + \mu_X] = \frac{1}{n}(n\mu_X) = \mu_X$$

$$var(\overline{X}) = var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n^2}[var(X_1) + var(X_2) + \ldots + var(X_n)]$$

$$= \frac{1}{n^2}(n\sigma_X^2) = \frac{\sigma_X^2}{n}$$

PHÂN PHỐI t

Phân phối xác suất được sử dụng rất nhiều trong phần kinh tế lượng căn bản là phân phối *t*, cũng được gọi là phân phối *t* Student. Student là biệt danh của W.S.Gosset, người phát hiện ra phân phối xác suất này năm 1908. Phân phối *t* có mối quan hệ rất gần với phân phối chuẩn như sau. Nếu $\overline{X} \sim N(\mu_x, \frac{\sigma_x^2}{n})$, thì biến chuẩn hóa Z được định nghĩa như sau: $Z = \frac{(\overline{X} - \mu_X)}{\sigma_x / \sqrt{n}} \sim N(0,1)$ nếu cả hai tham số μ_x và $\frac{\sigma_x^2}{n}$ đều được biết. Nhưng giả sử ta chỉ biết μ_x và giá trị ước lượng của σ_x^2 bởi ước lượng mẫu $S_x^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$. Như vậy, nếu thay σ_x bằng S_x ta sẽ có một biến mới như sau:

$$t = \frac{(\overline{X} - \mu_X)}{S_x / \sqrt{n}} \qquad (2.43)$$

Lý thuyết thống kê cho rằng biến *t* sẽ theo phân phối *t* với số bậc tự do là (n-1), đây là tham số duy nhất của phân phối *t*.

Tính chất của phân phối t

- Giống như phân phối chuẩn, phân phối *t* đối xứng quanh giá trị trung bình.

- Trung bình của phân phối *t*, giống như phân phối chuẩn hóa, là không, nhưng phương sai là k/(k-2), với *k* là số bậc tự do. Vì vậy, phương sai của phân phối *t* chỉ được xác định khi số bậc tự do d.f. > 2. Và khi *k* đủ lớn, thì phương sai của phân phối *t* sẽ gần bằng 1.

Để minh họa ứng dụng của phân phối *t* trên thực tế ta xét tiếp ví dụ về số lượt khách du lịch quốc tế tại một công ty du lịch như đã đề cập. Biết rằng, trong giai đoạn 15 ngày qua, số lượt khách quốc tế trung bình một ngày là 72 và phương sai mẫu là 4. Hãy tính xác suất để có được số lượt khách trung bình đó, biết rằng giá trị trung bình thực là 70 khách một ngày?

Nếu biết độ lệch thực của tổng thể (σ) thì ta có thể dễ dàng sử dụng phân phối chuẩn hóa để tính xác suất trên. Nhưng ở đây ta có S, là một ước lượng của σ, nên ta có thể sử dụng phân phối *t* như sau:

$$t = \frac{72-70}{2/\sqrt{15}} = 3.87 \qquad (2.44)$$

sẽ theo phân phối t với trung bình bằng 0, phương sai bằng 1.17, và d.f. = 14. Thay vì tìm $P(\overline{X} > 72)$, ta có thể tìm P(t > 3.87). Áp dụng hàm phân phối t[9] cho trường hợp một đuôi ta có:

$$P(t > 3.87) = 1 - P(t < 3.87) = 0.0085$$

Vậy xác suất để số lượt khách trung bình một ngày của công ty du lịch này là 0.085%. Nếu là một người ra quyết định của công ty, thì bạn sẽ có nghĩ gì với con số này?

Tóm lại, trong phân tích kinh tế lượng và dự báo, ta sẽ biết rằng các hệ số hồi quy từ mô hình hồi quy mẫu là các ước lượng của các hệ số hồi quy tổng thể. Do ta không biết giá trị sai số chuẩn của từng hệ số hồi quy tổng thể nhưng có thể ước lượng từ mẫu, nên các hệ số hồi quy ước lượng sẽ có phân phối *t*. Đây là một cơ sở rất quan trọng cho

[9] Hàm phân phối xác suất t trên Excel là: =TDIST(X, Deg_freedom, Tails). "X" nghĩa là giá trị t cần tính xác suất (3.87), nghĩa là diện tích dưới đường phân phối *t* từ t đến +∞ (ta sẽ biết đây chính là vùng bác bỏ giả thiết H_0). "Deg_freedom" là số bậc tự do (14). "Tails" có hai lựa chọn: "1" (một đuôi), và "2" (hai đuôi). Giá trị xác suất ta tính được từ công thức này chính là P-Value. Nếu ta đã biết mức ý nghĩa và số bậc tự do, ta sẽ tìm được giá trị t theo công thức sau: =TINV(Probability, Deg_freedom). Ví dụ, =TINV(0.085%,14) =3.87.

việc kiểm định giả thiết về các hệ số hồi quy sẽ được trình bày ở chương 7.

PHÂN PHỐI CHI BÌNH PHƯƠNG (χ^2)

Ta đã xác định phân phối xác suất của trung bình mẫu, $\overline{X}$, vậy còn phương sai mẫu, $S^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$ sẽ có phân phối như thế nào? Phân phối xác suất cần cho mục đích này chính là phân phối xác suất χ^2, cũng là một phân phối có mối quan hệ rất gần với phân phối chuẩn. Lưu ý, giống như trung bình mẫu, phương sai mẫu cũng thay đổi từ mẫu này qua mẫu khác. Cho nên, giống như trung bình mẫu, phương sai mẫu cũng là một biến ngẫu nhiên.

Ta biết rằng nếu một biến ngẫu nhiên X theo phân phối chuẩn với trung bình là μ_x và phương sai là σ_x^2, $X \sim N(\mu_x, \sigma_x^2)$, thì biến chuẩn hóa $Z \sim N(0,1)$. Lý thuyết thống kê cho rằng bình phương của một biến chuẩn hóa có phân phối χ^2 với một bậc tự do. Ký hiệu như sau:

$$Z^2 \sim \chi^2_{(1)} \tag{2.45}$$

Giống như phân phối *t*, bậc tự do là tham số của phân phối χ^2. Ở phương trình (2.45) chỉ có một bậc tự do vì ta đang xét bình phương của một biến chuẩn hóa.

Giả sử $Z_1, Z_2, \ldots, Z_k$ là các biến chuẩn hóa độc lập (mỗi biến Z là một biến ngẫu nhiên có phân phối chuẩn với trung bình bằng 0 và phương sai bằng 1). Nếu ta lấy bình phương từng biến này, thì tổng của các biến Z bình phương này cũng theo phân phối Chi bình phương với k bậc tự do.

$$\sum Z_i^2 = Z_1^2 + Z_1^2 + \ldots + Z_1^2 \sim \chi^2_{(k)} \tag{2.46}$$

Tính chất của phân phối χ^2

- Khác phân phối chuẩn, phân phối[10] χ^2 chỉ có giá trị dương từ 0 đến vô cùng.
- Khác phân phối chuẩn, phân phối χ^2 là một phân phối nghiêng, độ nghiêng của phân phối phụ thuộc vào số bậc tự do. Khi bậc tự do thấp, phân phối χ^2 bị nghiêng phải, nhưng khi bậc tự do tăng lên, phân phối sẽ đối xứng và dần về phân phối chuẩn.
- Giá trị trung bình của một biến ngẫu nhiên theo phân χ^2 là k và phương sai là $2k$ (k là số bậc tự do). Đây là một tính chất đáng chú ý của phân phối χ^2 vì phương sai gấp đôi giá trị trung bình.
- Nếu Z_1 và Z_2 là hai biến có phân phối Chi bình phương độc lập với k_1 và k_2 bậc tự do, thì (Z_1+Z_2) cũng là một biến có phân phối chi bình phương với bậc tự do là (k_1+k_2).

Tóm lại, đây là một phân phối rất hay sử dụng trong phân tích kinh tế lượng và dự báo cho các biến ngẫu nhiên dạng bình phương như kiểm định JB, kiểm định phương sai của hạn nhiễu, các kiểm định phần dư có sử dụng phương trình hồi quy phụ (nR^2) như kiểm định về phương sai thay đổi.

PHÂN PHỐI F

Một phân phối xác suất khác cũng rất quan trọng trong kinh tế lượng là phân phối F với ý tưởng như sau. Giả sử $X_1, X_2, \ldots, X_m$ là một mẫu ngẫu nhiên với cỡ mẫu m từ một tổng thể có phân phối chuẩn với trung bình μ_X và σ^2_X; và $Y_1, Y_2, \ldots, Y_n$ là một mẫu ngẫu nhiên với cỡ

[10] Hàm phân phối xác suất Chi bình phương trên Excel là: =CHIDIST(X, Deg_freedom). "X" nghĩa là giá trị χ^2 cần tính xác suất (ví dụ 6), nghĩa là diện tích dưới đường phân phối Chi bình phương từ χ^2 đến $+\infty$ (ta sẽ biết đây chính là vùng bác bỏ giả thiết H_0). "Deg_freedom" là số bậc tự do (ví dụ 2). Giá trị xác suất ta tính được từ công thức này (4.98%) chính là P-Value. Nếu ta đã biết mức ý nghĩa và số bậc tự do, ta sẽ tìm được giá trị χ^2 theo công thức sau: =CHIINV(Probability, Deg_freedom). Ví dụ, =CHIINV(4.98%,2) =6.

mẫu n từ một tổng thể phân phối chuẩn với trung bình μ_Y và phương sai σ^2_Y. Giả sử hai mẫu độc này lập nhau và được lấy từ hai tổng thể có phân phối chuẩn. Giả sử ta muốn xem phương sai của hai tổng thể trên có giống nhau hay không ($\sigma^2_X = \sigma^2_Y$). Do ta không thể quan sát trực tiếp phương sai tổng thể nên ta suy ra từ các ước lượng phương sai như sau:

$$S^2_X = \frac{\sum(X_i - \overline{X})^2}{m-1} \tag{2.47}$$

$$S^2_Y = \frac{\sum(Y_i - \overline{Y})^2}{n-1} \tag{2.48}$$

Bây giờ ta xét tỷ số sau đây:

$$F = \frac{S^2_X}{S^2_Y} = \frac{\sum(X_i - \overline{X})^2/(m-1)}{\sum(Y_i - \overline{Y})^2/(n-1)} \tag{2.49}$$

Lưu ý, khi tính tỷ số F[11], ta quy ước giá trị phương sai lớn hơn sẽ nằm trên tử số của phân số F. Nếu hai phương sai bằng nhau, thì tỷ số F bằng 1, nhưng nếu khác nhau, tỷ số F sẽ khác 1 (lớn hơn 1). Các phương sai càng khác nhau thì tỷ số F càng lớn. Phân phối F phụ thuộc vào hai tham số là bậc tự do của tử (m-1) và bậc tự do của mẫu (n-1). Trong kinh tế lượng và dự báo, tỷ số F thông thường số bậc tự do của tử số nhỏ hơn số bậc tự do của mẫu số. Bậc tự do của tử số có thể là số biến giải thích trong mô hình (phân tích ANOVA) hoặc số điều kiện ràng buộc trong kiểm định Wald, và số bậc tự do của mẫu có thể là số bậc tự do của tổng bình phương phần dư (phân tích ANOVA) hoặc bậc tự do của tổng bình phương phần dư trong mô hình không bị ràng buộc (kiểm định Wald). Ngoài ra, hãy lưu ý rằng, bản thân tỷ số F chính là tỷ số của hai biến có phân phối χ^2, vì chúng là các biến ở dạng bình phương. Cho nên, phân phối F không chỉ có mối quan hệ gần gũi với phân phối chuẩn mà còn với phân phối χ^2.

[11] F chính là tỷ số của hai biến có phân phối χ^2.

Đặc điểm của phân phối F

- Giống phân phối chi bình phương, phân phối F[12] cũng bị nghiêng phải và nằm trong khoảng từ 0 đến vô cùng.
- Giống phân phối *t* và phân phối chi bình phương, phân phối *F* sẽ dần về phân phối chuẩn khi k_1 và k_2 tăng lên vô cùng.
- Bình phương của một biến ngẫu nhiên có phân phối *t* với *k* bậc tự do sẽ có phân phối *F* với 1 và k bậc tự do.

$$t_k^2 \sim F_{1,k} \qquad (2.50)$$

- Có một mối quan hệ rất chặt chẽ giữa phân phối *F* và phân phối chi bình phương.

Tóm lại, phân phối *F* rất quan trọng trong kinh tế lượng khi chúng ta thực hiện phân tích phương sai (ANOVA) và kiểm định các biến dưới dạng tỷ số giữa các phương sai ví dụ kiểm định giả thiết đồng thời, kiểm định Wald, kiểm định Chow, kiểm định nhân quả Granger, v.v...

Một lần nữa chúng tôi xin khuyên bạn đọc rằng nếu chúng ta muốn học tốt kinh tế lượng và dự báo thì hãy dành ít thời gian xem lại các nội dung cơ bản của thống kê, đặc biệt các các phân phối vừa được đề cập ở trên để có thể hiểu cách thức suy luận thống kê trong phân tích kinh tế lượng và dự báo. Chúng ta có thể không cần quan tâm đến các công thức phức tạp của thống kê toán, nhưng vấn đề quan trọng là chúng ta phải hiểu và phân biệt được khi nào thì nên sử dụng loại

[12] Hàm phân phối xác suất F trên Excel là: =FDIST(X, Deg_freedom1, Deg_freedom2). "X" nghĩa là giá trị F cần tính xác suất (ví dụ 4), nghĩa là diện tích dưới đường phân phối F từ F đến +∞ (ta sẽ biết đây chính là vùng bác bỏ giả thiết H_0). "Deg_freedom1" là số bậc tự do của tử số (ví dụ 2). "Deg_freedom2" là số bậc tự do của mẫu số (ví dụ 14). Giá trị xác suất ta tính được từ công thức này (4.23%) chính là P-Value. Nếu ta đã biết mức ý nghĩa (sẽ được trình bày ở chương 4) và số bậc tự do của tử và mẫu số, ta sẽ tìm được giá trị F theo công thức sau: =FINV(Probability, Deg_freedom1, Deg_freedom2). Ví dụ, =FINV(4.23%,2,14) = 4.

phân phối nào và biết cách tính giá trị xác suất của các loại phân phối nêu trên.

SUY DIỄN THỐNG KÊ

Suy diễn thống kê là việc rút ra các kết luận về bản chất của một tổng thể nào đó trên cơ sở phân tích một dữ liệu mẫu ngẫu nhiên được rút ra chính từ tổng thể đó. Giả sử từ một mẫu nhất định được thu thập từ một tổng thể có phân phối chuẩn ta tính được giá trị trung bình và phương sai mẫu, thì điều mà ta quan tâm là muốn biết giá trị trung bình và phương sai thực của tổng thể đó sẽ như thế nào. Thực tế cho thấy mục tiêu của bất kỳ nghiên cứu nào đều muốn tìm hiểu bản chất của toàn bộ tổng thể đang xem xét nhằm có những quyết định thích hợp. Tuy nhiên, việc thu thập toàn bộ dữ liệu của tổng thể rất tốn kém thời gian và chi phí. Cho nên ta thường chỉ nghiên cứu dựa trên một hoặc hai dữ liệu mẫu ngẫu nhiên được lấy từ tổng thể để suy diễn cho tổng thể.

SUY DIỄN THỐNG KÊ LÀ GÌ?

Các khái niệm tổng thể và mẫu rất quan trọng trong thống kê. Tổng thể là toàn bộ tất cả các kết quả có thể có của một hiện tượng được quan tâm. Một mẫu là một tập hợp con của tổng thể. Suy diễn thống kê là việc nghiên cứu mối quan hệ giữa một tổng thể và một mẫu được lấy ra từ tổng thể đó. Để hiểu rõ khái niệm này, ta xem xét ví dụ cụ thể sau đây về ROE (suất sinh lợi trên vốn chủ sở hữu, %) trên Sở giao dịch chứng khoán TP.HCM (HOSE, tập tin **DATA2-9**).

■ BẢNG 2.12: **ROE của 30 công ty trên HOSE.**

Công ty	ROE	Công ty	ROE
ABT	13.7	VFC	13.2
SC5	29.2	VGP	10.5
SFI	41.9	VHC	32.5

Công ty	ROE	Công ty	ROE
BHS	14.1	VHG	16.8
BMP	22.8	VIC	14.6
DHA	17.4	VIS	12.6
DHG	19.7	TSC	44.3
DQC	29.3	VNE	16.6
ITA	11.1	VNM	22.3
PAC	22.9	VPL	8.1
REE	13.0	VSC	26.1
SGH	17.6	VSH	12.7
STB	19.0	BMC	56.2
TAC	34.8	VTB	11.2
UNI	32.5	VTO	30.2
Trung bình = 22.2	Phương sai = 131.9		Độ lệch = 11.5

Nguồn: Reuters, 2009.

Giả sử quan tâm chính của ta không phải là ROE của một công ty nhất định nào đó, mà là ROE trung bình của toàn bộ chứng khoán niêm yết trên thị trường TP.HCM. Trên nguyên tắc, việc thu thập dữ liệu tỷ số ROE của tất cả các công ty niêm yết để tính ROE trung bình là hoàn toàn có thể thực hiện được, nhưng thực tế, việc làm này rất tốn thời gian và chi phí (giả sử trong tương lai có thêm rất nhiều công ty niêm yết thì sao?). Cho nên liệu ta có thể sử dụng ROE trung bình của 30 công ty như *một giá trị ước lượng* của ROE trung bình tổng thể hay không. Cụ thể, nếu ta đặt X = ROE của một chứng khoán và $\overline{X}$ là ROE trung bình của 30 chứng khoán, thì liệu ta có thể nói gì về giá trị kỳ vọng của ROE, E(X) của toàn bộ thị trường chứng khoán TP.HCM hay không. Quá trình khái quát hóa từ giá trị mẫu, ví dụ $\overline{X}$, cho giá trị tổng thể, E(X), là nội dung chủ yếu của suy diễn thống kê. Ta sẽ thảo luận chi tiết chủ đề này.

Việc thu thập thông tin về tất cả các ROE của cả thị trường chứng khoán để tính ROE trung bình là rất tốn kém, nên ta có thể thu thập một mẫu ngẫu nhiên một số chứng khoán để tính ROE trung bình mẫu, $\overline{X}$. $\overline{X}$ là ước lượng của ROE trung bình tổng thể, E(X), và cũng được gọi là tham số tổng thể. Một giá trị bằng số của ước lượng được gọi là một giá trị ước lượng. Vì thế, ước lượng là bước đầu tiên của suy diễn thống kê. Sau khi đã có giá trị ước lượng của tham số, ta cần phải tìm hiểu xem giá trị đó có phải là một giá trị ước lượng tốt cho tham số tổng thể hay không, vì một giá trị ước lượng có thể không bằng giá trị tham số thực.

Bước thứ hai của suy diễn thống kê là kiểm định giả thiết. Trong kiểm định giả thiết ta có thể đã có phán đoán hay kỳ vọng trước về giá trị của một tham số nhất định. Ví dụ, dựa vào kiến thức đã có hoặc ý kiến chuyên gia chúng ta có thể biết được ROE trung bình thực của thị trường là 17.4. Như vậy, giá trị 22.2 từ mẫu 30 công ty có khác về mặt thống so với giá trị 17.4 hay không.

Ước lượng tham số

Thông thường, một biến ngẫu nhiên X được cho rằng sẽ theo một phân phối nhất định, nhưng ta không biết giá trị các tham số của phân phối đó. Chẳng hạn, nếu X theo phân phối chuẩn, ta muốn biết giá trị của hai tham số trung bình $E(X) = \mu_x$ và phương sai σ_x^2. Để ước lượng các tham số này, quy trình thông thường là lấy một mẫu ngẫu nhiên với n quan sát từ một phân phối xác suất đã biết và sử dụng mẫu để ước lượng các tham số chưa biết. Vì thế, ta có thể sử dụng trung bình mẫu như một giá trị ước lượng của trung bình tổng thể và phương sai mẫu như một giá trị ước lượng của phương sai tổng thể. Quy trình này được gọi là vấn đề ước lượng. Vấn đề ước lượng được chia thành hai loại: ước lượng điểm và ước lượng khoảng.

Để cụ thể hóa ý tưởng này, giả sử biến ngẫu nhiên X (ROE) là một biến có phân phối chuẩn với một giá trị trung bình và một giá trị phương sai nhất định, nhưng ta không biết giá trị của các tham số này. Giả sử ta có một mẫu ngẫu nhiên gồm 30 ROE từ một tổng thể có phân phối chuẩn như Bảng 2.12.

Ta có thể sử dụng mẫu này để tính giá trị trung bình tổng thể μ_x = E(X) và phương sai tổng thể σ_x^2 như thế nào. Cụ thể hơn, giả sử quan tâm của ta bây giờ là tìm giá trị trung bình μ_x. Ta thấy rằng giá trị trung bình mẫu, $\overline{X}$, là 22.2. Lưu ý rằng ước lượng điểm là một biến ngẫu nhiên vì giá trị của nó sẽ khác nhau ở các mẫu khác nhau. Vì thế, làm sao có thể tin một giá trị cụ thể như 22.2 là giá trị ước lượng của μ_x. Nói cách khác, làm thế nào ta có thể chỉ dựa vào một giá trị ước lượng của trung bình tổng thể. Tuy nhiên, có vẻ tốt hơn nếu cho rằng một khoảng giá trị nào đó có chứa trung bình tổng thể. Đó là ý tưởng của khái niệm ước lượng khoảng. Ước lượng khoảng là một khoảng các giá trị sẽ chứa giá trị thực của tổng thể với một mức tin cậy nhất định.

Ý tưởng nền tảng của ước lượng khoảng là khái niệm phân phối mẫu của một ước lượng. Giả sử, một biến ngẫu nhiên X có phân phối chuẩn, $X \sim N(\mu_X, \sigma^2)$, thì:

$$\overline{X} \sim N(\mu_X, \frac{\sigma_X^2}{n}) \qquad (2.42)$$

$$Z = \frac{\left(\overline{X} - \mu_X\right)}{\sigma_X / \sqrt{n}} \sim N(0, 1) \qquad (2.51)$$

Tuy nhiên, dù không biết giá trị phương sai σ_x^2, nhưng ta có thể sử dụng ước lượng của nó là $S_x^2 = \frac{\sum(X_x - \overline{X})^2}{n-1}$ thì:

$$t = \frac{(\overline{X} - \mu_X)}{S_x / \sqrt{n}} \sim t_{d.f.=(n-1)} \qquad (2.43)$$

theo phân phối *t* với (n-1) bậc tự do, (d.f.).

Để biết phương trình (2.43) được sử dụng như thế nào cho ước lượng khoảng của giá trị trung bình tổng thể μ_x của ROE. Với bậc tự do là 29, ta có khoảng tin cậy 95% như sau:

$$P(-2.045 \leq t \leq 2.045) = 0.95 \qquad (2.52)$$

Với d.f. = 29, xác suất là 0.95 (hay 95%), thì khoảng (-2.045, 2.045) sẽ chứa giá trị *t* tính từ công thức (2.43). Các giá trị *t* trên được gọi là các giá trị *t* phê phán[13] (critical t values) sẽ cho biết phần trăm diện tích dưới đường phân phối *t* giữa hai giá trị phê phán này. Trong đó, t = -2.045 được gọi là giá trị *t* phê phán chặn dưới (lower critical t value) và t = 2.045 được gọi là giá trị *t* phê phán chặn trên (upper critical t value).

Thế giá trị *t* từ (2.43) vào (2.52) ta có:

$$P(-2.045 \leq \frac{(\overline{X} - \mu_X)}{S_x / \sqrt{n}} \leq 2.045) = 0.95 \qquad (2.53)$$

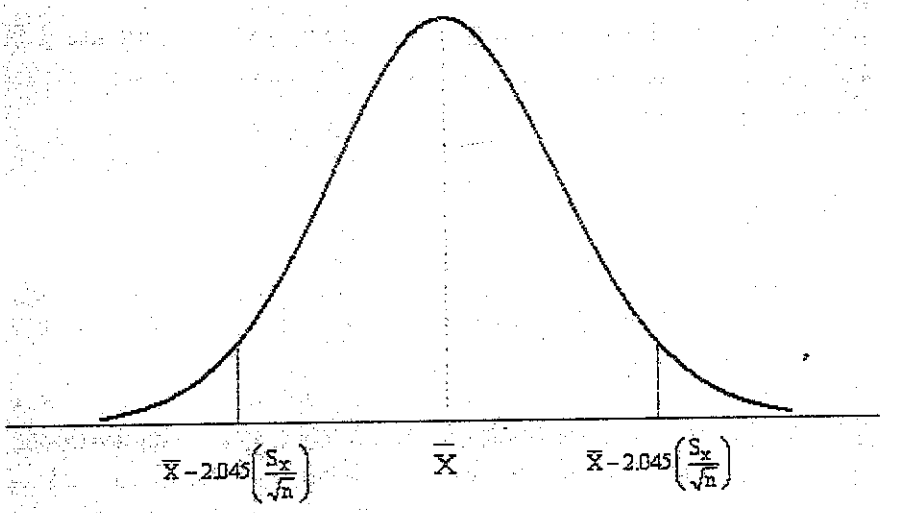
$$P(\overline{X} - 2.045\frac{S_x}{\sqrt{n}} \leq \mu_X \leq \overline{X} + 2.045\frac{S_x}{\sqrt{n}}) = 0.95 \qquad (2.54)$$

Quay lại ví dụ ở trên với n = 30, $\overline{X}$ = 22.2, và S_x = 11.5 thì ta xác định khoảng ngẫu nhiên như sau:

$$17.91 \leq \mu \leq 26.49 \qquad (2.55)$$

[13] Để đơn giản, chúng ta thường sử dụng tên gọi khác là giá trị tra bảng.

■ HÌNH 2.12: **Khoảng tin cậy 95% của μ_X với d.f.=29.**



Như vậy, khác với ước lượng điểm, ước lượng khoảng đưa ra một khoảng các giá trị sẽ chứa giá trị thực với một độ tin cậy hay xác suất nhất định. Ước lượng khoảng được xem như một cầu nối quan trọng giữa ước lượng và kiểm định trong suy luận thống kê.

Kiểm định giả thiết

Kiểm định giá trị trung bình

Sau khi đã xem xét nhánh ước lượng của suy diễn thống kê, bây giờ chúng ta sẽ xem xét chi tiết hơn nhánh thứ hai của nó là kiểm định giả thiết. Trở lại ví dụ về ROE ở trên, thay vì tìm khoảng tin cậy cho μ_x, giả sử ta giả thiết rằng giá trị thực μ_X bằng một giá trị bằng số cụ thể, ví dụ $\mu_X = 17,4$. Công việc của ta bây giờ là kiểm định giả thiết này. Ta sẽ kiểm định giả thiết này như thế nào – đó là ủng hộ hay bác bỏ nó?

Theo ngôn ngữ kiểm định giả thiết thì một giả thiết như $\mu_x = 17,4$ được gọi là giả thiết không và được ký hiệu là H_0. Như vậy, H_0: $\mu_x = 17,4$. Giả thiết không luôn được kiểm định ngược lại với giả thiết khác, ký hiệu là H_1 hoặc H_a. Giả thiết khác có thể dưới một số hình thức khác nhau như sau:

- H_1: $\mu_x > 17,4$, đây là giả thiết khác một phía hay một đuôi.
- H_1: $\mu_x < 17,4$, cũng là giả thiết khác một phía hay một đuôi.
- H_1: $\mu_x \neq 17,4$, đây được gọi là giả thiết khác kết hợp, hai phía, hay hai đuôi.

Để kiểm định H_0, ta sử dụng dữ liệu mẫu và lý thuyết thống kê để xây dựng các quy tắc quyết định nhằm xem chứng cứ của mẫu có ủng hộ giả thiết không hay không. Nếu chứng cứ mẫu ủng hộ giả thiết không, ta không bác bỏ giả thiết H_0, ngược lại, ta bác bỏ giả thiết H_0, điều này cũng có nghĩa ta chấp nhận giả thiết H_1.

Vấn đề đặt ra là chúng ta xây dựng các quy tắc quyết định này như thế nào? Có hai cách tiếp cận kiểm định có tính bổ sung cho nhau: (i) Khoảng tin cậy, và (ii) Kiểm định ý nghĩa. Chúng tả sử dụng ví dụ về ROE để minh họa cho các phương pháp kiểm định này. Giả sử ta có các giả thiết kiểm định như sau:

$$H_0: \mu_x = 17.4$$

$$H_1: \mu x \neq 17.4$$

Kiểm định dựa vào khoảng tin cậy

Kiểm định dựa vào khoảng tin cậy có nghĩa là ta cố gắng xây dựng khoảng tin cậy cho một ước lượng, rồi kiểm tra xem giá trị thực được giả định theo giả thiết H_0 sẽ nằm trong hay ngoài khoảng tin cậy đó.

Và trên cơ sở đó ta sẽ có quyết định bác bỏ hay không bác bỏ giả thiết H_0.

Để kiểm định giả thiết H_0, giả sử ta có dữ liệu mẫu như trong Bảng 2.12. Từ dữ liệu này, ta tính được giá trị trung bình mẫu là 22.2. Ta biết rằng trung bình mẫu có phân phối chuẩn với trung bình là μ_x và phương sai là σ_x^2/n. Nhưng do ta không biết giá trị phương sai thực của tổng thể, nên ta thay thế phương sai này bằng phương sai mẫu, và như thế thì trung bình mẫu sẽ theo phân phối *t*. Dựa vào phân phối *t* ta xây dựng được khoảng tin cậy 95% cho μ_X như sau:

$$17.91 \le \mu_X \le 26.49 \qquad (2.56)$$

Như vậy, khoảng tin cậy này không giá trị giả thiết không, $\mu_x = 17.4$, ta có thể bác bỏ giả thiết H_0 này không?

Theo ngôn ngữ kiểm định giả thiết, khoảng tin cậy 95% được gọi là vùng chấp nhận và vùng ngoài vùng chấp nhận là vùng phê phán hay vùng bác bỏ giả thiết H_0. Giá trị chặn dưới và giá trị chặn trên của vùng chấp nhận được gọi là các giá trị phê phán[14]. Theo ngôn ngữ thống kê, nếu vùng chấp nhận có chứa giá trị tham số ở giả thiết H_0, ta không bác bỏ H_0 (nghĩa là chấp nhận H_0 là đúng). Nhưng nếu rơi ngoài vùng chấp nhận (tức nằm trong vùng bác bỏ), ta bác bỏ H_0. Ở ví dụ đang xét, ta bác bỏ giả thiết H_0: $\mu_x = 17.4$ vì vùng chấp nhận, như ở phương trình (2.56), không chứa giá trị giả thiết không này.

[14] Critical value. Lưu ý, các sách thống kê hoặc kinh tế lượng ở Việt Nam dịch thuật ngữ "critical" là "phê phán" hoặc "tới hạn". Trong thống kê, các ranh giới của vùng chấp nhận được gọi là các giá trị phê phán, vì các ranh giới này là đường phân chia giữa việc chấp nhận và bác bỏ giả thiết H_0. Chúng ta thường tra nhanh các giá trị này bằng các công thức như TINV, FINV, CHIINV, ... trên Excel.

Kiểm định dựa vào ý nghĩa

Kiểm định ý nghĩa là một cách kiểm định khác có tính bổ sung và ngắn gọn hơn so với kiểm định dựa vào khoảng tin cậy[15]. Xin nhắc lại rằng $t = \frac{(\overline{X} - \mu_X)}{S_x / \sqrt{n}}$ có phân phối *t* với (n-1) bậc tự do. Các thông tin $\overline{X}$, n, và S đã biết từ mẫu, nên nếu có giá trị μ_x dưới giả thiết không, ta sẽ tính được giá trị thống kê *t* (còn được gọi là giá trị *t* tính toán). Hơn nữa, với (n-1) bậc tự do ta có thể xác định được các giá trị *t* phê phán[16].

Nếu khác biệt giữa $\overline{X}$ và μ_x là nhỏ (tử số sẽ nhỏ và mẫu số không đổi), thì giá giá trị tuyệt đối của t tính toán sẽ nhỏ. Nếu $\overline{X} = \mu_x$, giá trị *t* tính toán sẽ bằng không, và chắc chắn ta sẽ không bác bỏ giả thiết H_0. *Cho nên, khi giá trị tuyệt đối của t tính toán càng khác không, ta sẽ có xu hướng bác giả thiết H_0.* Nói cách khác, khi giá trị tuyệt đối của *t* tính toán càng lớn, ta càng có xu hướng bác bỏ giả thiết H_0. Tuy nhiên, quyết định bác bỏ hay chấp nhận H_0 tùy thuộc vào mức ý nghĩa được chọn. Nếu giá trị *t* tính toán nằm giữa các giá trị *t* phê phán chặn dưới và chặn trên thì ta chấp nhận H_0 ($|t| \leq t_{\alpha/2}$). Ngược lại, nếu giá trị thống kê *t* nằm ngoài các giá trị *t* phê phán chặn dưới và chặn trên thì ta bác bỏ H_0 ($|t| > t_{\alpha/2}$).

---

[15] Vì các phần mềm kinh tế lượng đều cung cấp các thông tin giá trị thống kê *t* và giá trị xác suất tương ứng, nên phương pháp kiểm định này được sử dụng phổ biến.

[16] Biết rằng $t = \frac{(\overline{X} - \mu_X)}{S_x / \sqrt{n}}$ theo phân phối *t* với bậc tự do là (n-1), nên ta có thể dễ dàng tính được giá trị xác suất tương ứng của nó bằng công thức =TDIST(x,Deg_Freedom,Tails). Ví dụ, =TDIST(2.045,29,2) = 5%. Ngoài ra, để tính giá trị *t* phê phán ta dùng công thức =TINV(Probability,Deg_Freedom). Ví dụ, =TINV(5%,29) = 2.045.

Ví dụ, $\overline{X} = 22.2$, $S_x = 11.5$, và $n = 30$. Giả sử H_0: $\mu_x = 17.4$ và H_1: $\mu x \neq 17.4$. Ta có:

$$t = \frac{(22.2 - 18.5)}{11.5/\sqrt{30}} = 2.286 \qquad (2.57)$$

Với số bậc tự do là 29, các giá trị *t* phê phán với mức ý nghĩa 5% là -2.045 và 2.045. Như vậy, giá trị *t* tính toán bằng 2.286 nằm ở phía đuôi phải vùng bác bỏ của phân phối *t*, nên ta bác bỏ giả thiết H_0. Ngoài ra, ta dễ dàng tính xác suất để $|t| > 2.286$ chỉ là 2,974%.

Các bước thực hiện:

- Xác định giả thiết H_0 và H_1.
- Tính giá trị thống kê *t*.
- Chọn mức ý nghĩa α với phân phối *t* hai đuôi và xác định các giá trị *t* phê phán chặn dưới và chặn trên.
- So sánh giá trị thống kê *t* và giá trị *t* phê phán.
- Đưa ra quyết định.

Trong phương pháp kiểm định ý nghĩa, thay vì xác định một dãy các giá trị hợp lý cho tham số tổng thể chưa biết, ta chọn một giá trị cụ thể của tham số được đưa ra ở giả thiết H_0, tính thống kê kiểm định, ví dụ thống kê *t*; và tìm phân phối mẫu của nó và xác suất để có giá trị cụ thể của thống kê kiểm định (được gọi là p-value). Nếu xác suất này rất thấp, ví dụ, nhỏ thua $\alpha = 5\%$ hay 1%, ta bác bỏ giả thiết H_0. Ngược lại, nếu xác suất này lớn hơn giá trị α, ta không bác bỏ giả thiết H_0.

Mức ý nghĩa α và giá trị xác suất p

Giá trị xác suất *p* (p-value) cũng được gọi là mức ý nghĩa chính xác của thống kê kiểm định (ví dụ thống kê *t*). Giá trị xác suất *p* có thể

được định nghĩa là mức ý nghĩa thấp nhất tại đó giả thiết H_0 có thể bị bác bỏ. Quy tắc quyết định với giá trị xác suất *p* như sau: *Giá trị xác suất p càng nhỏ, thì bằng chứng để bác bỏ giả thiết H_0 càng mạnh.* Các phần mềm kinh tế lượng đều có báo cáo giá trị xác suất *p*.

Kiểm định phương sai

Như đã biết, nếu S^2 là phương sai mẫu từ một mẫu được lấy ngẫu nhiên với *n* quan sát từ một tổng thể có phân phối chuẩn với phương sai là σ^2, thì

$$(n-1)\left(\frac{S^2}{\sigma^2}\right) \sim \chi^2_{(n-1)} \qquad (2.58)$$

Nghĩa là, tỷ số của phương sai mẫu/phương sai tổng thể được nhân với số bậc tự do (n-1) có phân phối χ^2 với số bậc tự do là (n-1). Lưu ý rằng, (n-1) và σ^2 chỉ là các con số nhất định, riêng bản thân S^2 là một biến ngẫu nhiên vì giá trị của S^2 sẽ thay đổi từ mẫu này qua mẫu khác (tương tự như $\overline{X}$). Do $\overline{X}$ là một biến ngẫu nhiên có phân phối chuẩn, thì S^2 được xem gần như một $\overline{X}^2$, nên S^2 sẽ có phân phối χ^2. Hơn nữa, vế trái của phương trình (2.58) là một biến tổng của (n-1) biến S^2/σ^2, nên số bậc tự do sẽ là (n-1).

Ví dụ, từ một mẫu ngẫu nhiên như Bảng 2.12 ta có phương sai mẫu là $S^2 = 131.9$, ta kiểm định xem giá trị này có khác gì ở mức ý nghĩa $\alpha = 5\%$ so với giá trị phương sai thực của tổng thể là 99.1 hay không.

$$H_0: \sigma^2 = 99.1 \text{ và } H_1: \sigma^2 \neq 99.1$$

Ta có n = 30 nên thay vào phương trình (2.58) ta sẽ có giá trị χ^2 tính toán là $(30-1)\frac{131.9}{99.1}$ = 38.598 với số bậc tự do là 29. Từ công thức =CHIINV(2.5%,29) = 45.72 > 38.598 > =CHIINV(97.5%,29) = 16.05, ta không bác bỏ giả thiết H_0 ở mức ý nghĩa 5%. Ngoài ra, ta có thể tính xác suất để có được giá trị χ^2 bằng hoặc lớn hơn 38.598 (với 29 bậc tự do) theo công thức =CHIDIST(38.598,29) = 0.1096 hay 10.96%. Vì xác suất này lớn hơn mức ý nghĩa được chọn là 5%, nên ta không bác bỏ giả thiết H_0 cho rằng phương sai thực là 99.1.

TÓM TẮT CHƯƠNG 2

Nền tảng cho một nhà nghiên cứu tiến hành các dự báo chính là thống kê và kinh tế lượng. Hơn nữa, để tiến hành dự báo dưới các nghiên cứu định lượng thì việc thu thập và phân tích dữ liệu là các nội dung hết sức cần thiết. Việc thu thập và phân tích dữ liệu, sau đó trước khi đưa ra các mô hình dự báo thích hợp thì yêu cầu nhà nghiên cứu phải hiểu cấu trúc cụ thể của từng loại dữ liệu mẫu và hơn nữa là phải có những suy luận ra các thông tin của tổng thể từ các dữ liệu mẫu. Để đạt được các yêu cầu như vậy thì các nhà nghiên cứu trước khi tiến hành dự báo phải am hiểu các nội dung quan trọng là: hiểu được đặc điểm biến động của dữ liệu thông qua các đại lượng đo lường từ tổng thể như phương sai và độ lệch chuẩn; cách thể hiện dữ liệu thông qua các loại đồ thị khác nhau nhằm tìm ra các hiện tượng cá biệt, tính xu hướng, tính dao động ngẫu nhiên, các yếu tố mùa vụ tiềm ẩn; các kiểm định và các phân phối xác xuất thống kê cần thiết để kiểm định các giả thiết mang tính suy luận cho các thông tin của tổng thể từ dữ liệu mẫu; và sau cùng là các công việc phục vụ cho nền tảng dự báo nói ở đây cần được hỗ trợ bằng một phần mềm thống kê nào đó mà chúng tôi giới thiệu ở đây là Eviews nhằm giảm thiểu thời gian tác nghiệp cho các nhà nghiên cứu dự báo không chuyên nhưng vẫn đảm bảo tính hiệu quả cuối cùng của nó là nhanh gọn và đơn giản.

CÂU HỎI VÀ BÀI TẬP

1. Cho X và Y là hai biến ngẫu nhiên, a và b là hai hằng số. Anh/Chị hãy chứng minh các tính chất sau đây:

   a. $E(a) = a$ và $Var(a) = 0$?

   b. $E(a+bX) = a + bE(X)$ và $Var(a+bX) = b^2Var(X)$

   c. $Var(X) = E(X^2) - [E(X)]^2$

   d. Hãy tìm $E(aX+bY)$ và $Var(aX+bY)$

   e. $Cov(X,Y) = E(XY) - E(X)E(Y)$

   f. Nếu X và Y độc lập về mặt thống kê, hãy chứng tỏ rằng

   $E(XY) = E(X)E(Y)$

   $Cov(X,Y) = 0$

2. Sử dụng tập tin "**VLSS04.wf1**" và thực hiện các thống kê mô tả trên Eviews, ta có các kết quả sau đây:

■ BẢNG 2.13: **Tổng thu nhập của hộ theo 8 vùng kinh tế.**

REGION	Mean	Median	Max	Min.	Std. Dev.	Obs.
1	24207.70	18632.00	412790.0	250.0000	22485.97	1943
2	21981.03	16780.00	469400.0	1840.000	21255.87	1317
3	17221.88	13973.00	120416.0	3359.000	12676.27	429
4	18060.34	14934.00	243390.0	550.0000	15933.52	1013
5	24582.21	18070.50	1380200.	20.00000	50644.44	852
6	24560.55	20072.00	174480.0	1375.000	18565.19	581
7	39117.46	30080.00	298940.0	250.0000	33673.56	1187
8	24640.09	18496.00	471274.0	10.00000	26211.76	1859
All	24955.88	18584.00	1380200.	10.00000	28066.18	9181

■ BẢNG 2.14: **Tổng thu nhập của hộ theo 5 nhóm chi tiêu.**

QUANTILE	Mean	Median	Max	Min.	Std. Dev.	Obs.
1	12079.60	10707.50	96848.00	10.00000	7140.992	1738
2	16337.79	14585.00	90858.00	200.0000	9474.912	1816
3	20454.96	18370.50	98940.00	20.00000	12034.53	1876
4	27379.28	24292.00	263750.0	20.00000	17874.60	1964
5	48298.70	38720.00	1380200.	250.0000	50940.98	1787
All	24955.88	18584.00	1380200.	10.00000	28066.18	9181

■ BẢNG 2.15: **Tổng thu nhập của hộ theo thành thị và nông thôn.**

URBAN	Mean	Median	Max	Min.	Std. Dev.	Obs.
0	20525.09	16340.50	412790.0	10.00000	18312.26	6934
1	38628.84	29720.00	1380200.	250.0000	44009.19	2247
All	24955.88	18584.00	1380200.	10.00000	28066.18	9181

■ BẢNG 2.16: **Tổng thu nhập của hộ theo giới tính chủ hộ.**

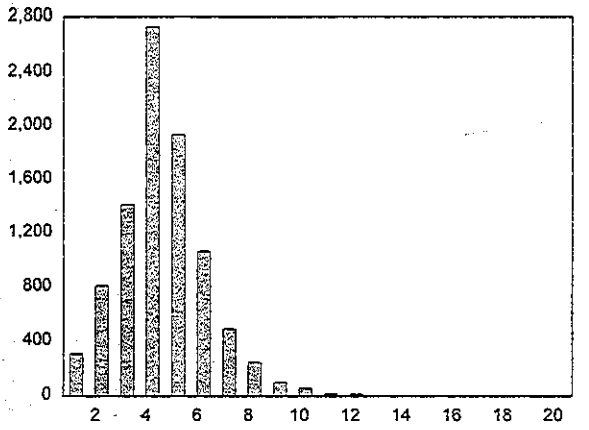
GENDER	Mean	Median	Max	Min.	Std. Dev.	Obs.
0	25199.08	18210.00	469400.0	300.0000	27204.48	2229
1	24877.91	18660.00	1380200.	10.00000	28338.39	6952
All	24955.88	18584.00	1380200.	10.00000	28066.18	9181

a. Trình bày cách lập bảng thống kê theo nhóm trên Eviews?

b. Anh/Chị hãy giải thích các kết quả trên như thế nào?

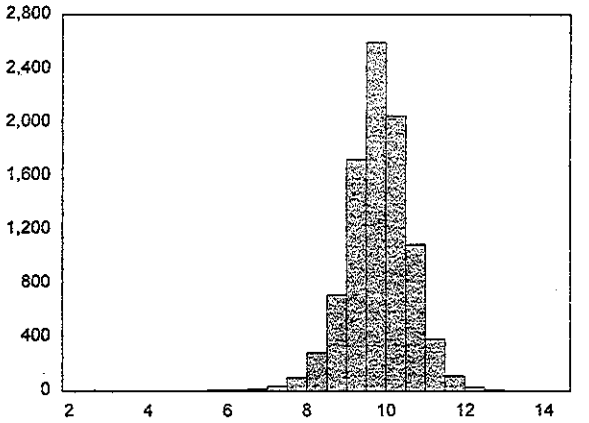
3. Sử dụng tập tin "**VLSS04.wf1**" và thực hiện các thống kê mô tả trên Eviews, ta có các kết quả sau đây:

■ HÌNH 2.17: **Quy mô hộ gia đình Việt Nam năm 2004.**



Series: FAMILYSIZE	
Sample 1 9189	
Observations 9189	
Mean	4.400914
Median	4.000000
Maximum	20.00000
Minimum	1.000000
Std. Dev.	1.731455
Skewness	0.797609
Kurtosis	5.449923
Jarque-Bera	3272.373
Probability	0.000000

■ HÌNH 2.18: **Logarith của thu nhập năm 2004.**



Series: LOG(INCOME)	
Sample 1 9189	
Observations 9181	
Mean	9.811172
Median	9.830056
Maximum	14.13774
Minimum	2.302585
Std. Dev.	0.814295
Skewness	-0.706313
Kurtosis	7.465823
Jarque-Bera	8392.614
Probability	0.000000

a. Trình bày cách vẽ đồ thị tần suất trên Eviews?

b. Anh/Chị cho biết hai đồ thị trên khác nhau ở điểm gì?

c. Anh/Chị cho biết tại sao lấy log(INCOME) như ở Hình 2.18?

d. Anh/Chị hãy giải thích các giá trị trong bảng thống kê tóm tắt?

4. Sử dụng tập tin **"VLSS04.wf1"** và thực hiện thống kê mô tả trên Eviews, ta có kết quả sau đây:

■ BẢNG 2.19: **Phân bố nhóm chi tiêu theo thành thị và nông thôn.**

Variable	Categories
URBAN	2
QUANTILE	5
Product of Categories	10

Count % Total % Row % Col		QUANTILE					
		1	2	3	4	5	Total
URBAN	0	1660	1677	1595	1430	577	6939
		18.07	18.25	17.36	15.56	6.28	75.51
		23.92	24.17	22.99	20.61	8.32	100.00
		95.51	92.29	84.98	72.66	32.25	75.51
	1	78	140	282	538	1212	2250
		0.85	1.52	3.07	5.85	13.19	24.49
		3.47	6.22	12.53	23.91	53.87	100.00
		4.49	7.71	15.02	27.34	67.75	24.49

a. Trình bày cách lập bảng 2.19 trên Eviews?

b. Anh/Chị giải thích các con số trong bảng 2.19 như thế nào?

c. Anh/chị cho biết có sự khác biệt gì giữa các nhóm chi tiêu ở thành thị và nông thôn?

d. Anh/Chị cho biết có sự khác biệt gì giữa từng nhóm chi tiêu ở thành thị và nông thôn?

5. Doanh số trung bình của một mẫu gồm 100 đại diện bán hàng là \$25.350/tháng và độ lệch chuẩn của mẫu là \$7.490. Giám đốc kinh doanh muốn biết liệu kết quả này có khác \$24.000 với khoảng tin cậy 95% hay không. Anh/Chị hãy thiết lập giả thiết H_0 và thực hiện kiểm định thống kê thích hợp để trả lời câu hỏi của vị giám đốc kinh doanh?

6. Giả sử Anh/Chị được yêu cầu đánh giá lượng hàng tồn kho ở các đại lý bán lẻ của một công ty sản xuất vỏ xe hơi. Từ một mẫu gồm 120 cửa hàng, Anh/Chị biết được lượng hàng tồn kho trung bình là 310 vỏ xe/tháng. Biết rằng, lượng tồn kho trung bình của ngành là

325 vỏ xe/tháng. Nếu độ lệch chuẩn của mẫu là 72, thì Anh/Chị có cho rằng lượng tồn kho của công ty này có khác lượng tồn kho của ngành ở mức ý nghĩa 5% hay không? Tại sao?

7. Giám đốc kinh doanh của một công ty bất động sản gần đây đã đọc một báo cáo phân tích ngành và nhận thấy rằng doanh số/một địa ốc viên của các công ty bất động sản khác có phân phối chuẩn với trung bình là \$255.000. Giám đốc kinh doanh muốn biết liệu rằng các địa ốc viên của công ty mình có đạt được mức doanh số tương tự như vậy không. Vị giám đốc kinh doanh thu thập một mẫu ngẫu nhiên gồm 16 địa ốc viên. Dữ liệu được lưu trong tập tin **"RE_SALES.xls"**. Anh/Chị cho biết, với mức ý nghĩa 5% thì Anh/Chị có bác bỏ giả thiết H_0 cho rằng doanh số trung bình của các địa ốc viên của công ty cũng là \$255.000. Anh/Chị hãy minh họa câu trả lời trên đồ thị?

8. Ngân hàng VCB thực hiện một cuộc nghiên cứu thị trường nhằm tìm hiểu mức độ đánh giá của khách hàng về hình ảnh của VCB. Câu hỏi dưới dạng thang đo từ 1 đến 10, trong đó 10 là mức đánh giá cao nhất. Kết quả khảo sát từ một mẫu 400 người cho biết mức đánh giá trung bình là 7.25 và độ lệch chuẩn là 2.51. Biết rằng mức đánh giá trung bình của tất cả các ngân hàng thương mại là 7.01.

   a. Giám đốc kinh doanh của VCB muốn kiểm định có phải mức đánh giá của khách hàng về VCB cao hơn mức trung bình của các ngân hàng thương mại hay không. Anh/Chị hãy thực hiện kiểm định giả thiết phù hợp với khoảng tin cậy 95%?

   b. Anh/Chị hãy minh họa kết quả kiểm định trên đồ thị?

   c. Giả sử mẫu khảo sát là 100 người (thay vì 400 người), Anh/Chị cho biết kết quả kiểm định có thay đổi gì không? Tại sao?

9. Dựa trên kinh nghiệm quá khứ, công ty điện lực TP.HCM dự báo rằng lượng điện tiêu thụ trung bình/hộ gia đình sẽ là 700 kWh trong tháng Giêng tới. Trong tháng Giêng, công ty thu thập một mẫu ngẫu nhiên gồm 50 hộ gia đình và tính được giá trị trung bình và độ lệch

chuẩn lần lượt là 715 và 50. Anh/Chị hãy kiểm định dự báo của công ty điện lực có hợp lý hay không nếu mức ý nghĩa được chọn là 5%? Anh/Chị hãy tính toán và giải thích ý nghĩa giá trị xác suất *p* của thống kê kiểm định bằng Excel?

10. Giả sử sau khi thảo luận với ban giám đốc và các chuyên viên phòng kinh doanh của Công ty kinh doanh sản phẩm khí, Anh/Chị xác định được một cơ sở dữ liệu như trong tập tin **"GAS.xls"**. Anh/Chị hãy xây dựng một bảng tính các giá trị thống kê tổng hợp cho các biến chứa trong tập tin này? Anh/Chị có nhận xét gì về kết quả của các giá trị thống kê trên?

11. GAP[17] là một thương hiệu đứng thứ hai trong lĩnh vực kinh doanh quần áo ở Mỹ. Năm 1994, GAP đứng thứ 25 trong nhóm 50 nhãn hiệu nổi tiếng nhất nước Mỹ. Cuối năm 1995, GAP có hơn 1.500 cửa hàng trong nước và 164 cửa hàng ở nước ngoài như Canada, Anh, Pháp, Đức, và Nhật Bản. Với sự mở rộng thị trường và gia tăng thị phần nhanh chóng, nên công việc lập kế hoạch kinh doanh có ý nghĩa đặc biệt quan trọng đối với GAP. Chính vì vậy, dự báo doanh số trở thành một hoạt động thường xuyên và rất cần thiết đối với GAP. Từ chương này, chúng ta sẽ sử dụng tình huống này như một phần trong toàn bộ cuốn giáo trình để thực hành phân tích dữ liệu và dự báo doanh số với đầy đủ các tính chất đặc thù cho lĩnh vực kinh doanh khác. Dữ liệu về doanh số của GAP được cho trong tập tin **"GAP.xls"**.

    a. Anh/Chị hãy lập bảng thống kê mô tả doanh số theo quý của GAP?

    b. Anh/Chị hãy vẽ doanh số của GAP theo quý trên cùng một đồ thị? Anh/Chị có nhận xét gì về các đồ thị này? Tại sao?

    c. Sử dụng doanh số theo quý trong năm 2004, Anh/Chị hãy xây dựng khoảng tin cậy 95% cho doanh số theo quý của GAP?

[17] Wilson, 2007, Business Forecasting with Accompanying Excel-Based ForecastX™ Software, pp.39-42.

12. Công ty tư vấn tín dụng tiêu dùng CCC[18], một tổ chức tư vấn phi lợi nhuận tư nhân được thành lập năm 1982, chuyên cung cấp dịch vụ tư vấn miễn phí trong việc lập và theo dõi các kế hoạch ngân sách nhằm hỗ trợ khách hàng dễ dàng thương lượng với tổ chức tín dụng về các khoản trả nợ vay có vấn đề. Ngoài ra, CCC còn cung cấp miễn phí dịch vụ giáo dục về kỹ năng quản lý tiền cho các cá nhân và hộ gia đình khó khăn về mặt tài chính. Thông qua chương trình này, CCC, nhân danh khách hàng, còn thương lượng trực tiếp với tổ chức tín dụng về các hợp đồng thanh toán đặc biệt. Nhân viên của CCC phần lớn là các tình nguyện viên và cộng tác viên. Cho nên, để thuận lợi cho việc hoạch định kế hoạch nhân sự và sắp xếp chương trình làm việc với các tổ chức tín dụng, CCC muốn dự báo trước số lượng khách hàng mới trong thời gian một hoặc hai tháng kế tiếp. Từ chương này, chúng ta sẽ sử dụng tình huống trình bày ở câu 12 như một phần trong toàn bộ cuốn giáo trình để thực hành phân tích dữ liệu và dự báo số lượng khách hàng mới với đầy đủ các tính chất đặc thù cho lĩnh vực kinh doanh khác. Sử dụng tập tin "**CCC.xls**", Anh/Chị hãy phân tích thống kê mô tả số lượng khách hàng mới của CCC.

[18] Hanke, 2005, Business Forecasting, 8th Edition, Person, pp.10-11.

CHƯƠNG 3 PHÂN TÍCH DỮ LIỆU VÀ LỰA CHỌN MÔ HÌNH

Vấn đề quan trọng nhất trong dự báo không nằm ở việc sử dụng loại kỹ thuật dự báo nào. Nhiều người vẫn nhầm lẫn rằng các mô hình dự báo càng phức tạp càng cho các kết quả có độ chính xác cao. Thực tế, không có một mô hình dự báo nào là tốt nhất cho mọi trường hợp. Nhiều sinh viên và nhà phân tích thường hay nghĩ rằng để có thể dự báo tốt cần phải giỏi kinh tế lượng và biết sử dụng các mô hình phức tạp như ARIMA, SARIMA, ARCH/GARCH, Holt-Winters, hay VAR. Hiểu như vậy là một sai lầm nghiêm trọng. Thật vậy, công việc quan trọng nhất, khó khăn nhất và cũng tốn nhiều thời gian nhất trong dự báo là thu thập được nguồn dữ liệu tin cậy và thích hợp cho mục đích dự báo. Hanke (2005) cho rằng, dù có sử dụng mô hình dự báo phức tạp đến mức nào đi nữa, thì kết quả dự báo cũng sẽ không có giá trị nếu dựa trên nguồn dữ liệu không tin cậy và áp dụng phương pháp dự báo phức tạp nhưng độ chính xác có thể lại kém.

Ngày nay, nhờ sự phát triển của máy tính và công nghệ thông tin, nên các tổ chức có thể tạo ra và lưu trữ dữ liệu một cách đầy đủ ở hầu hết các lĩnh vực. Đứng trước một "rừng" dữ liệu như vậy, vấn đề khó khăn đối với những người làm công tác dự báo là làm sao có thể chọn lọc những dữ liệu phù hợp để hỗ trợ các vấn đề ra quyết định của doanh nghiệp và những nhà hoạch định chính sách. Có phải tất cả dữ liệu quá khứ đều được sử dụng cho dự báo tương lai? Có nên áp dụng thử tất cả các mô hình dự báo cho dữ liệu sẵn có rồi lựa chọn mô hình có sai số dự báo bé nhất? Có nên bắt chước một cách máy móc các mô hình dự báo trước đây? Và nhiều câu hỏi tương tự như vậy có thể đang được đặt ra. Nên nhớ rằng, chúng ta luôn đối diện với các nguồn lực

khan hiếm. Vì vậy, các kỹ thuật phân tích dữ liệu và lựa chọn mô hình dự báo sẽ được giới thiệu trong chương này sẽ là chìa khóa của các vấn đề nói trên.

MỤC TIÊU HỌC TẬP

Sau khi học xong chương này, chúng ta kỳ vọng sẽ đạt được các nội dung sau đây:

- Biết được các tiêu chí để xác định nguồn dữ liệu đáng tin cậy.
- Hiểu được các thành phần cơ bản trong chuỗi thời gian.
- Khảo sát dữ liệu bằng phân tích tự tương quan.
- Hiểu rõ bản chất một chuỗi dữ liệu có tính dừng.
- Biết được các tiêu chí lựa chọn mô hình dự báo thích hợp.

CHẤT LƯỢNG DỮ LIỆU

Như đã giới thiệu ở chương 1, thì kết quả dự báo thành công hay không tùy thuộc rất nhiều vào quá trình trao đổi qua lại giữa người làm công tác dự báo và người sử dụng kết quả dự báo. Tuy nhiên, điều này đã giả định rằng nguồn dữ liệu phục vụ việc dự báo được thu thập một cách chính xác và đáng tin cậy. Cho nên, nếu dữ liệu không đáng tin cậy và thiếu chính xác đã cho biết trước một kết quả dự báo tồi. Hanke (2005) cho rằng bốn tiêu chí sau đây có thể được sử dụng để đánh giá dữ liệu có hữu ích cho việc dự báo hay không.

- **Dữ liệu phải tin cậy và chính xác.** Trước khi thực hiện bất kỳ một dự báo nào, người làm dự báo cần phải quan tâm xem liệu dữ liệu có được thu thập từ một nguồn đáng tin cậy hay không. Tùy vào mục đích dự báo và loại dữ liệu mà chúng ta có thể đánh giá nguồn dữ liệu như thế nào. Đối với dữ liệu chéo, chúng ta cần kiểm tra thật kỹ cách thức thiết kế bảng câu hỏi có dựa trên cơ sở lý thuyết và khung phân tích hợp lý chưa? Phương pháp chọn mẫu có tính đại diện cho tổng thể không? Cách thức huấn luyện người phỏng vấn có bài bản và khoa học

hay không? Phương pháp thu thập dữ liệu có thích hợp và đáng tin cậy hay không? Quy trình nhập liệu và quản lý dữ liệu có chính xác hay không? Đối với dữ liệu chuỗi thời gian, thì nguồn cung cấp dữ liệu là yếu tố quan trọng đầu tiên cần phải xem xét. Khi dự báo sử dụng các nguồn bên ngoài, như các chỉ báo kinh tế vĩ mô, các chỉ số tài chính, v.v..., thì tốt nhất người làm dự báo và người sử dụng kết quả dự báo cần nhất trí nên chọn nguồn dữ liệu nào. Vấn đề có thể đơn giản hơn nếu dự báo sử dụng các nguồn bên trong doanh nghiệp. Tuy nhiên, doanh nghiệp cần tránh hiện tượng "ốc đảo", vì điều này có thể gây trở ngại trong quá trình tổng hợp dữ liệu từ các bộ phận khác nhau. Vấn đề này sẽ được đề cập chi tiết hơn ở chương 10. Nếu xác định đúng vai trò của dự báo trong quá trình ra quyết định, thì mỗi doanh nghiệp nên xây dựng một cơ sở dữ liệu thống nhất thông qua một quy trình cụ thể và chặt chẽ. Lưu ý, chưa chắc các dữ liệu được cung cấp bởi các hãng tin nổi tiếng như Reuters, Bloomberg, hay các công ty nghiên cứu thị trường được xem là chính xác hoàn toàn. McCommick (2009) cho rằng 25% dữ liệu của Forturn 1000 là không chính xác, không đầy đủ hoặc bị điều chỉnh bằng các hệ số gán ghép và nhiều khi là bóp méo dữ liệu.

- **Dữ liệu phải phù hợp.** Dữ liệu phải có tính đại diện cho vấn đề dự báo đang được xem xét. Ví dụ, ở chương 1, khi đã xác định dự báo doanh số, thì cần xác định rõ loại doanh số nào: tổng giá trị doanh số, tổng sản lượng tiêu thụ, doanh số xuất khẩu, doanh số của mặt hàng quan trọng nhất, doanh số theo quý hay theo năm. Đối với các dự báo liên quan đến chỉ số giá, thì cần cân nhắc loại chỉ số giá tiêu dùng tổng hợp, chỉ số giá tiêu dùng của một nhóm mặt hàng (lương thực, phi lương thực, dầu, vàng, rượu bia, v.v...). Đối với các biến kinh tế vĩ mô khác, ví dụ: cung tiền thì cần xác định giá trị các đại lượng M1, M2 theo đơn vị gốc hay phần trăm thay đổi, v.v..., giá trị GDP hay tăng trưởng GDP, v.v... Đối với các mô hình kinh tế lượng với dữ liệu chéo, cần xác định các biến đại diện phù hợp với mục tiêu nghiên cứu trên cơ sở lý thuyết và khung phân tích chặt chẽ.

- **Dữ liệu phải nhất quán.** Khi các định nghĩa thay đổi liên quan đến cách thức các dữ liệu được thu thập, thì các điều chỉnh cần thiết phải được thực hiện để duy trì sự nhất quán trong các mô hình dữ liệu quá khứ. Ví dụ, điều này có thể là vấn đề khi tổng cục thống kê thay đổi rổ hàng hóa trong việc tính chỉ số giá tiêu dùng. Ví dụ, ở Việt Nam, hiện tại có khoảng 660 mặt hàng được sử dụng để tính chỉ số giá tiêu dùng, và trọng số của các mặt hàng sẽ được điều chỉnh hai năm một lần theo kết quả điều tra mức sống dân cư Việt Nam. Đối với các chỉ số tài chính, như chỉ số giá chứng khoán theo tháng, thì khi phân tích cần nhất quán cách lấy dữ liệu: chỉ số giá đóng cửa cuối tháng hay chỉ số giá trung bình trong tháng. Ngoài ra, khi nghiên cứu tài chính cần lưu ý trường hợp sáp nhập hay phân tách cổ phiếu.

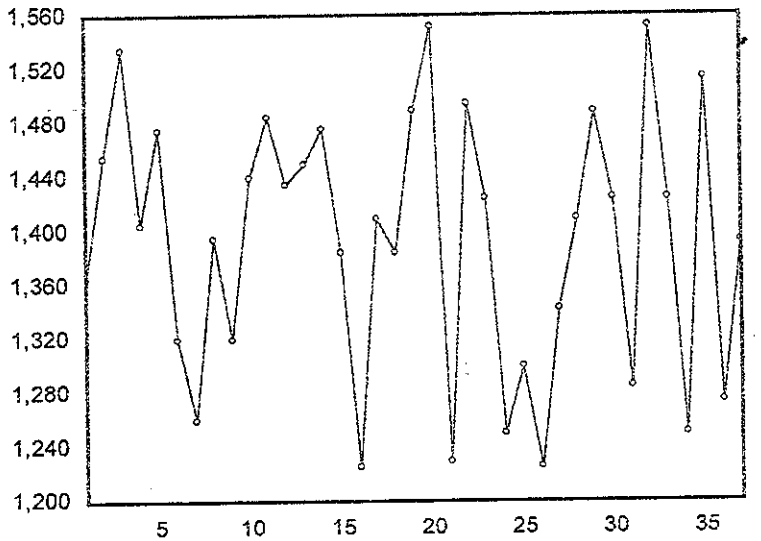
- **Dữ liệu phải kịp thời.** Dữ liệu được thu thập, tóm tắt và công bố kịp thời sẽ có giá trị rất lớn đối với kết quả dự báo. Có nhiều trường hợp, người làm dự báo đối diện với việc có quá ít dữ liệu hoặc có quá nhiều dữ liệu quá khứ. Ví dụ, khi sử dụng các mô hình ARIMA và ARCH, thông thường người phân tích cần xác định một số quan sát theo thời gian thích hợp đủ lớn để xác định sự biến thiên trong dữ liệu. Nếu ít quan sát quá sẽ khó nhận biết chuỗi dữ liệu có phải là chuỗi dừng hay không. Ngoài ra, đối với các dữ liệu có tính mùa vụ, chúng ta cần có số quan sát đủ lớn để có thể phát hiện và đánh giá thực sự có yếu tố mùa vụ hay không.

CÁC THÀNH PHẦN CỦA MỘT CHUỖI THỜI GIAN

Một trong những khía cạnh quan trọng nhất trong việc lựa chọn một phương pháp dự báo thích hợp cho chuỗi thời gian là phải xem xét chuỗi thời gian đó thuộc dạng dữ liệu nào. Thông thường, một chuỗi thời gian có thể có một hoặc một số trong bốn dạng dữ liệu sau đây: dữ liệu dừng, dữ liệu có tính xu thế, dữ liệu có yếu tố mùa vụ, và dữ liệu có tính chu kỳ.

Khi các quan sát của dữ liệu dao động xung quanh một giá trị cố định hay giá trị trung bình, thì dữ liệu có thể được xem có dạng dữ liệu dừng. Ví dụ, doanh số của một sản phẩm không tăng hay không giảm đáng kể qua thời gian có thể được xem là một chuỗi có dạng dữ liệu phẳng. Hình 3.1 minh họa doanh số của một cây xăng trong 36 tuần qua, trong đó, doanh số hình như dao động quanh mức 1.400 (**DATA3-1**).

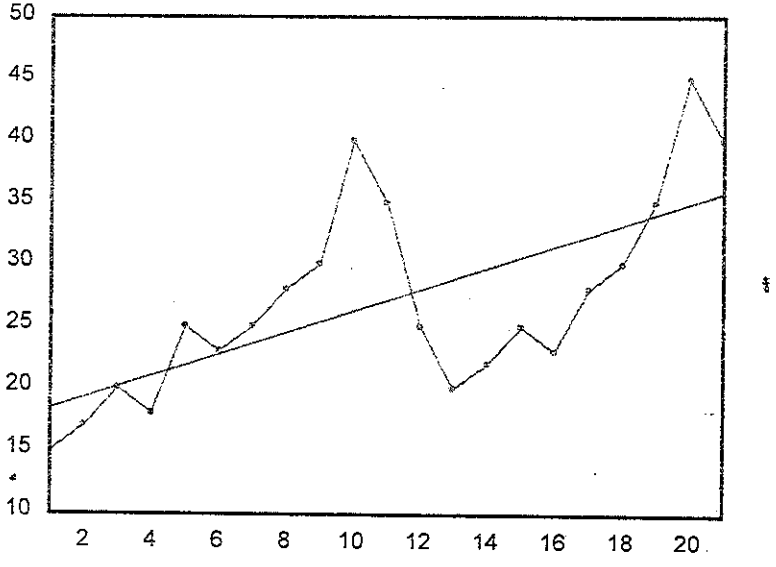
HÌNH 3.1: **Doanh số/tuần của một cây xăng.**



Khi các quan sát của dữ liệu tăng hoặc giảm qua thời gian, thì dữ liệu đó có thể được xem như có yếu tố xu thế. Hình 3.2 cho thấy xu hướng tăng trưởng trong dài hạn của chi phí thuê lao động phổ thông trung bình/ngày ở nông thôn từ năm 1990 đến nay (**DATA3-2**). Đường xu thế được vẽ cùng với dữ liệu gốc để giúp nhận dạng xu thế một cách rõ ràng hơn. Mặc dù chi phí thuê lao động phổ thông không gia tăng mỗi năm, nhưng xu hướng của biến số này nói chung là tăng lên trong giai đoạn 1990-2009. Rất nhiều các ví dụ khác như doanh số của một công ty, số lao động của một công ty, chi phí xây dựng/m^2, chi phí vận

chuyến/km, v.v..., có thể có yếu tố xu thế. Các yếu tố ảnh hưởng và giải thích yếu tố xu thế của một chuỗi thời gian bất kỳ có thể là do tăng dân số, lạm phát, thay đổi công nghệ, sở thích người tiêu dùng, và tăng năng suất.

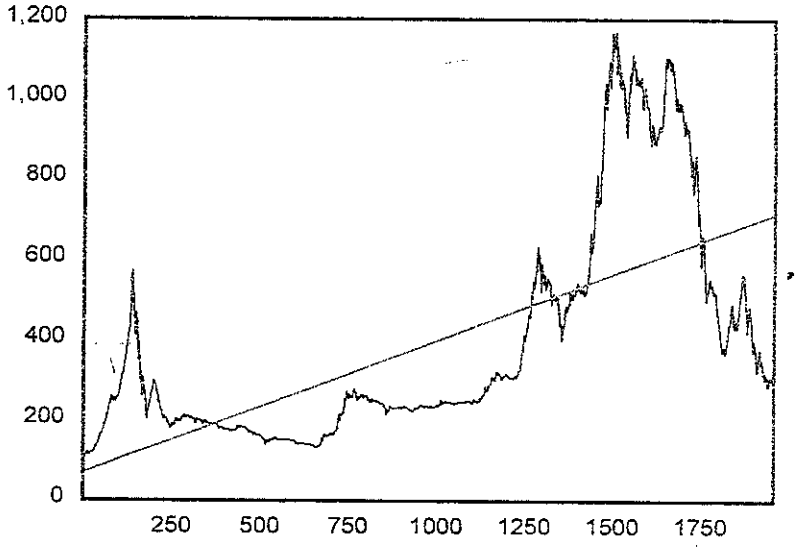
HÌNH 3.2: **Chi phí thuê lao động ở nông thôn.**



Nhiều chỉ báo kinh tế vĩ mô như GDP, GNP, việc làm, sản lượng công nghiệp, cung tiền, vốn đầu tư trực tiếp nước ngoài, chỉ số giá chứng khoán có khả năng thể hiện yếu tố xu thế rõ nét trong một giai đoạn nhất định. Hình 3.3 thể hiện xu hướng vận động của chỉ số giá chứng khoán VN-Index của thị trường chứng khoán Việt Nam giai đoạn 2000-2008 (**DATA3-3**).

Để vẽ đường xu thế cùng với dữ liệu gốc trong Eviews, ta thực hiện như sau (như Hình 3.3).

■ HÌNH 3.3: **Chỉ số giá chứng khoán VN-Index giai đoạn 2000-2008.**



- Bước 1: Tạo biến xu thế (với tên TREND) trên Eviews bằng cách chọn "GENR", nhập vào hợp thoại như sau: TREND=C(1)+C(2)*@TREND(0). Lưu ý, do ta đang sử dụng tập tin theo thứ tự quan sát, nên ta sử dụng hàm @TREND(0). Nếu ta sử dụng dữ liệu theo thứ tự thời gian (vị dụ tháng, quý), ta phải sử dụng hàm @TREND(tháng hoặc quý trước tháng hoặc quý đứng trước quan sát đầu tiên trong chuỗi dữ liệu).

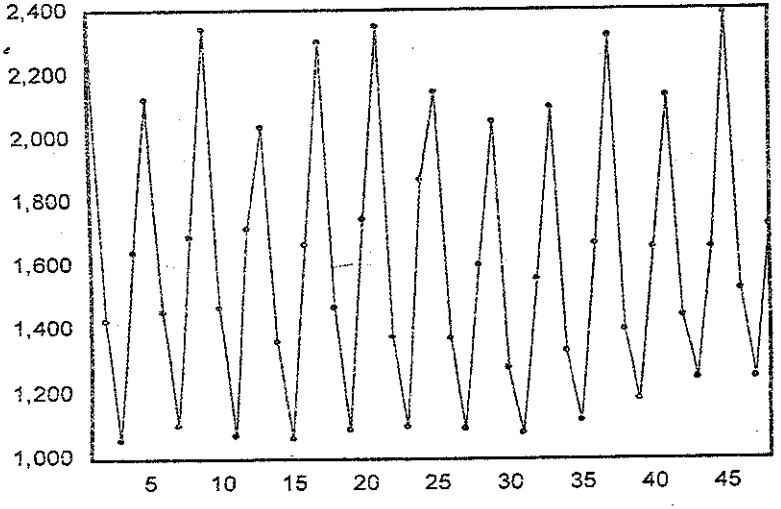
- Bước 2: Quick/Graph, nhập tên hai biến, ví dụ VNI và TREND vào hợp thoại (như đã hướng dẫn ở chương 2).

Khi các quan sát thể hiện xu hướng vận động của một giai đoạn như thể lặp lại xu hướng vận động của giai đoạn trước, thì ta có thể xem dữ liệu có tính chu kỳ. Thành phần chu kỳ của dữ liệu là một sự dao động theo dạng bước sóng xung quanh yếu tố xu thế thường chịu ảnh hưởng của các điều kiện kinh tế. Các dao động chu kỳ thường chịu

ảnh hưởng bởi những thay đổi trong những sự mở rộng hoặc thu hẹp hoạt động kinh tế, và phổ biến nhất là do chu kỳ kinh doanh. Hình 3.2 thể hiện tính chu kỳ trong chi phí lao động phổ thông ở nông thôn. Ở dữ liệu chu kỳ, thông thường xuất hiện các điểm 'đỉnh' và 'đáy'.

Khi các quan sát bị ảnh hưởng bởi các yếu tố mùa vụ, thì chúng ta có thể kỳ vọng sự tồn tại dạng dữ liệu mùa vụ. Thành phần mùa vụ cho biết một dạng thay đổi có tính lặp đi lặp lại năm này qua năm khác. Đối với dữ liệu tháng, thành phần mùa vụ đo lường sự biến thiên của chuỗi dữ liệu mỗi tháng Một, mỗi tháng Hai, v.v... Đối với dữ liệu quý, thường có bốn yếu tố mùa, mỗi yếu tố đại diện cho một quý trong năm. Dữ liệu có yếu tố mùa thường tồn tại đối với các dữ liệu về doanh số, du lịch, hoặc những hoạt động sản xuất kinh doanh phụ thuộc vào thời tiết, văn hóa, lễ hội. Hình 3.4 cho thấy sản lượng điện dao động lặp đi lặp lại theo quý, trong đó cao nhất là quý I hàng năm (**DATA3-4**).

■ HÌNH 3.4: **Sản lượng điện của công ty điện lực XYZ (kWh/năm).**



TỰ TƯƠNG QUAN VÀ GIẢN ĐỒ TỰ TƯƠNG QUAN

HỆ SỐ TỰ TƯƠNG QUAN

Khi một biến được đo lường theo thời gian, thì các quan sát ở các giai đoạn thời gian khác nhau thường tương quan với nhau. Sự tương quan này thường được đo bằng hệ số tự tương quan. Tự tương quan là sự tương quan giữa một biến trễ một hoặc *k* giai đoạn với chính bản thân biến đó. Các dạng dữ liệu, thường bao gồm các thành phần xu thế và mùa vụ, có thể được nhận dạng dựa trên các hệ số tự tương quan. Trong phần này chúng ta sẽ thảo luận các hệ số tự tương quan cho các độ trễ khác nhau của một chuỗi thời gian nhất định được sử dụng như thế nào để nhận biết các dạng dữ liệu khác nhau.

Hệ số tự tương quan tổng thể có độ trễ bậc k (ký hiệu là ρ_k) được xác định theo công thức sau đây:

$$\rho_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2} \qquad (3.1)$$

Nếu ta chia cả tử và mẫu của phương trình (3.1) cho *n*, thì hệ số tự tương quan trên có thể được viết lại như sau:

$$\rho_k = \frac{\text{Cov}(Y_t, Y_{t-k})}{\text{Var}(Y_t)} \qquad (3.2)$$

Các phương trình (3.1) và (3.2) được gọi là hàm tự tương quan, ký hiệu là ACF.

Do thực tế ta chỉ làm việc với dữ liệu mẫu, nên ta chỉ có thể ước lượng được hệ số tự tương quan mẫu (ký hiệu r_k) theo công thức sau đây:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2} \qquad (3.3)$$

Trong đó, $\overline{Y}$ là giá trị trung bình mẫu của chuỗi Y_t, k là độ trễ, n là số quan sát của mẫu.

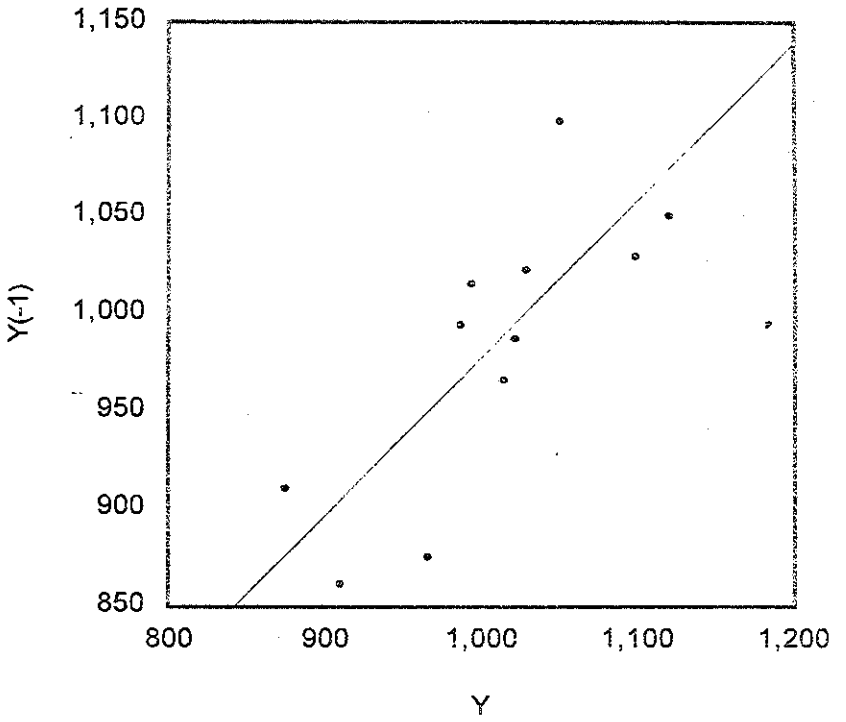
Khái niệm tự tương quan và hệ số tự tương quan sẽ được làm rõ hơn thông qua ví dụ sau đây. Bảng 3.1 thể hiện dữ liệu Y_t và hai biến trễ Y_{t-1} và Y_{t-2} của Y_t (**DATA3-5**).

■ BẢNG 3.1: **Khái niệm biến trễ.**

Tháng	Y_t	Y_{t-1}	Y_{t-2}	Y_{t-3}
1	861	NA	NA	NA
2	910	861	NA	NA
3	875	910	861	NA
4	966	875	910	861
5	1015	966	875	910
6	994	1015	966	875
7	987	994	1015	966
8	1022	987	994	1015
9	1029	1022	987	994
10	1099	1029	1022	987
11	1050	1099	1029	1022
12	1120	1050	1099	1029

Để xem biến Y_t và biến Y_{t-1} có tương quan với nhau hay không, trước hết ta vẽ đồ thị phân tán (scatter) giữa Y_t và Y_{t-1} như sau:

■ HÌNH 3.5: **Mối quan hệ giữa Y_t và Y_{t-1}.**



Hình 3.5 cho thấy giữa Y_t và Y_{t-1} có mối quan hệ đồng biến. Nghĩa là, khi Y_t tăng thì Y_{t-1} cũng tăng. Như vậy, hệ số tự tương quan giữa Y_t và Y_{t-1} sẽ dương. Từ Bảng 3.1, ta lập bảng tính các hệ số tự tương quan như sau.

■ BẢNG 3.2: **Tính hệ số tự tương quan bậc 1.**

t	Y_t	Y_{t-1}	$(Y_t-\overline{Y})$	$(Y_{t-1}-\overline{Y})$	$(Y_t-\overline{Y})^2$	$(Y_t-\overline{Y})(Y_{t-1}-\overline{Y})$
1	861	-	-133	-	17689	-
2	910	861	-84	-133	7056	11172

t	Y_t	Y_{t-1}	$(Y_t-\overline{Y})$	$(Y_{t-1}-\overline{Y})$	$(Y_t-\overline{Y})^2$	$(Y_t-\overline{Y})(Y_{t-1}-\overline{Y})$
3	875	910	-119	-84	14161	9996
4	966	875	-28	-119	784	3332
5	1015	966	21	-28	441	-588
6	994	1015	0	21	0	0
7	987	994	-7	0	49	0
8	1022	987	28	-7	784	-196
9	1029	1022	35	28	1225	980
10	1099	1029	105	35	11025	3675
11	1050	1099	56	105	3136	5880
12	1120	1050	126	56	15876	7056
$\overline{Y}=994$			Tổng		72226	41307

Áp dụng công thức (3.3) ta có hệ số tự tương quan bậc một sẽ được tính như sau:

$$r_1=\frac{\sum_{t=1+1}^{n}(Y_t-\overline{Y})(Y_{t-1}-\overline{Y})}{\sum_{t=1}^{n}(Y_t-\overline{Y})^2}=\frac{41.307}{72.226}=0.572 \qquad (3.4)$$

Thực hiện tương tự, ta có hệ số tự tương quan bậc hai sẽ được tính như sau:

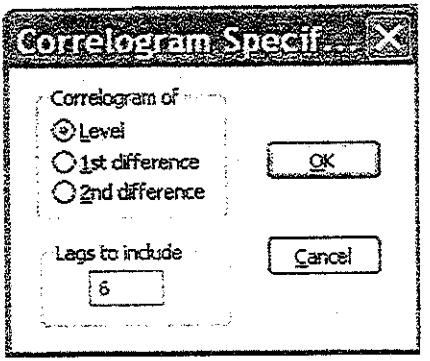
$$r_1=\frac{\sum_{t=2+1}^{n}(Y_t-\overline{Y})(Y_{t-2}-\overline{Y})}{\sum_{t=1}^{n}(Y_t-\overline{Y})^2}=\frac{33.418}{72.226}=0.463 \qquad (3.5)$$

GIẢN ĐỒ TỰ TƯƠNG QUAN

Nếu thực hiện như vậy cho đến độ trễ *k*, ta sẽ có *k* hệ số tự tương quan. Nếu vẽ đồ thị giữa các hệ số tự tương quan theo *k*, chúng ta sẽ

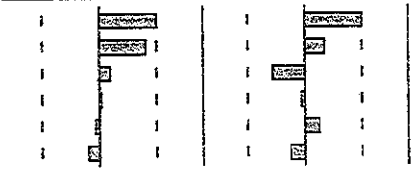
có giản đồ tự tương quan. Giản đồ tự tương quan là một công cụ quan trọng hữu ích giúp chúng ta xác định hệ số tự tương quan một cách nhanh chóng. Các bước vẽ giản đồ tự tương quan trên Eviews như sau:

- Bước 1: Quick/Series Statistics/Correlogram, nhập tên biến vào ô "Series Name".
- Bước 2: Chọn dạng dữ liệu cần vẽ giản đồ tự tương quan (dữ liệu gốc, sai phân bậc một, sai phân bậc 2) và số độ trễ như sau:



Sau khi chọn "OK" chúng ta sẽ có đồ thị như sau:

HÌNH 3.6: Giản đồ tự tương quan trên Eviews.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.572	0.572	4.9955	0.025
		2	0.463	0.202	8.5920	0.014
		3	0.111	-0.335	8.8203	0.032
		4	0.016	-0.026	8.8254	0.066
		5	-0.033	0.143	8.8519	0.115
		6	-0.102	-0.152	9.1419	0.166

Đồ thị đầu tiên chính là đồ thị giữa hệ số tự tương quan (AC) và độ trễ k (ở đây k = 6). Đồ thị thứ hai là đồ thị giữa hệ số tự tương quan riêng (PAC) và độ trễ. Chúng ta sẽ đề cập một cách chi tiết về hệ số tự

tương quan riêng ở chương 8 (các mô hình ARIMA). Hai cột cuối cho biết giá trị thống kê Q và xác suất của nó. Chúng ta sẽ trình bày cách phân tích các thống kê này ở phần sau.

KIỂM ĐỊNH HỆ SỐ TỰ TƯƠNG QUAN

Sau khi đã biết cách tính các hệ số tự tương quan, vấn đề đặt ra cho người phân tích dữ liệu và dự báo là làm sao có thể xác định một hoặc một nhóm hệ số tự tương quan khác không một cách có ý nghĩa thống kê? Tại sao điều này lại quan trọng? Bởi vì, đây là cơ sở quan trọng nhất để giúp người phân tích dữ liệu và dự báo biết một chuỗi thời gian đang xem xét thuộc dạng dữ liệu nào: ngẫu nhiên, dừng, có yếu tố xu thế (không dừng), hoặc có yếu tố mùa vụ. Đây là chủ đề sẽ được thảo luận kỹ hơn ở phần sau. Dựa vào giản đồ tự tương quan, có hai phương pháp kiểm định thường được sử dụng để xem hệ số tự tương quan có ý nghĩa thống kê hay không: Thống kê *t*, và Thống kê *Q*.

Thống kê t

Quenouille (1949) cho rằng các hệ số tự tương quan của một dữ liệu ngẫu nhiên có phân phối mẫu (xem chương 2) với trung bình bằng không và phương sai mẫu bằng 1/N. Trong đó, N là số quan sát trong mẫu. Lưu ý, khi ta biết giá trị của trung bình tổng thể (ở đây là 0), nhưng không biết giá trị phương sai tổng thể mà chỉ có ước lượng của nó là 1/N, thì việc kiểm định sẽ theo phân phối *t*. Biết được điều này, người phân tích có thể so sánh các hệ số tự tương quan mẫu với phân phối mẫu lý thuyết (phân phối *t*) để quyết định xem có nên chấp nhận hay bác bỏ giả thiết H_0 hay không.

Gọi ρ_k là hệ số tự tương quan tổng thể (r_k là ước lượng không chệch của ρ_k), ta có các giả thiết sau đây:

$$H_0: \rho_k = 0$$

$$H_1: \rho_k \neq 0$$

Như vậy, với sai số chuẩn của hệ số tự tương quan $se(r_k)$[1] là $\sqrt{1/N}$, ta có thể xây dựng khoảng tin cậy cho ρ_k (kiểm định theo phương pháp xây dựng khoảng tin cậy) hoặc tìm được giá trị thống kê *t* tính toán (kiểm định mức ý nghĩa) ở một mức ý nghĩa xác định. Nếu ρ_k nằm ngoài khoảng tin cậy đó hoặc giá trị *t* tính toán lớn hơn giá trị *t* quan sát ta bác bỏ giả thiết H_0.

Kiểm định dựa vào khoảng tin cậy

Tiếp tục ví dụ ở phần trên, chúng ta tiến hành kiểm định theo phương pháp khoảng tin cậy theo các bước sau đây:

- Bước 1: Xác định giả thiết H_0 và H_1

$$H_0: \rho_1 = 0$$

$$H_1: \rho_1 \neq 0$$

- Bước 2: Tính sai số chuẩn $se(r_1) = 1/sqrt(12) = 0.289$.

- Bước 3: Xây dựng khoảng tin cậy 95% (tức ta chọn mức ý nghĩa $\alpha = 5\%$).

$$Prob[r_1 - se(r_1)*t_{\alpha/2} < \rho_1 < r_1 - se(r_1)*t_{\alpha/2}] = 1 - 0.05$$

$$Prob[0.572 - 0.289*2.2 < \rho_1 < 0.572 - 0.289*2.2] = 0.95$$

---

[1] Các phần mềm kinh tế lượng sử dụng một công thức hơi khác một chút để tính sai số chuẩn của các hệ số tự tương quan. Theo Hanke (2005), công thức này giả định rằng bất kỳ hệ số tự tương quan nào trước độ trễ *k* được cho là khác 0 một cách có ý nghĩa thống kê, và bất kỳ hệ số tự tương quan nào ở độ trễ *k* hoặc lớn hơn k đều bằng 0. Công thức được xác định như sau:

$$se(r_k) = \sqrt{\frac{1 + 2\sum_{i=1}^{k-1} r_i^2}{n}} \qquad (3.6)$$

Trong đó, $se(r_k)$ = sai số chuẩn của hệ số tự tương quan ở độ trễ *k*; r_i = hệ số tự tương quan ở độ trễ i; k = độ trễ; và n = số quan sát của chuỗi dữ liệu đang xét. Đối với hệ số tự tương quan ở độ trễ k = 1, luôn luôn có sai số chuẩn, $se(r_1)$, bằng $1/\sqrt{n}$.

Lưu ý, giá trị *t* phê phán 2 đuôi tại mức ý nghĩa $\alpha = 5\%$, với bậc tự do d.f. = 12 – 1 được tính trên Excel bằng công thức =TINV(5%,11) = 2.2. Như vậy, ta có:

$$\text{Prob}[0.572 - 0.289*2.2 < \rho_1 < 0.572 + 0.289*2.2] = 0.95$$

$$\text{Prob}[-0.06 < \rho_1 < 1.21] = 0.95 \qquad (3.7)$$

- Bước 4: So sánh giá trị của giả thiết H_0 với khoảng tin cậy (3.7), ta dễ dàng thấy rằng khoảng tin cậy này có chứa giá trị 0, nên ta kết luận chấp nhận giả thiết H_0. Điều này có nghĩa rằng, hệ số tự tương quan $r_1 = 0.572$ bằng không một cách có ý nghĩa thống kê ở mức ý nghĩa 5%.

Nếu ta chọn mức ý nghĩa $\alpha = 10\%$, thì khoảng tin cậy (3.7) sẽ được tính lại như sau:

$$\text{Prob}[0.05 < \rho_1 < 1.09] = 0.90 \qquad (3.8)$$

Như vậy, khoảng tin cậy (3.8) không chứa giá trị 0, nên ta có thể kết luận bác bỏ giả thiết H_0. Điều này có nghĩa rằng, hệ số tự tương quan $r_1 = 0.572$ khác không một cách có ý nghĩa thống kê ở mức ý nghĩa 10%.

Tương tự, giả sử bây giờ ta có n = 20, thì $se(r_1) = 1/sqrt(20) = 0.224$ và TINV(5%,19) = 2.09, nên khoảng tin cậy (3.7) sẽ được tính lại như sau:

$$\text{Prob}[0.104 < \rho_1 < 1.04] = 0.95 \qquad (3.9)$$

Như vậy, nếu n = 20, thì khoảng tin cậy (3.9) không chứa giá trị 0, nên ta có thể kết luận bác bỏ giả thiết H_0. Điều này có nghĩa rằng, hệ số tự tương quan $r_1 = 0.572$ khác không một cách có ý nghĩa thống kê ở mức ý nghĩa 5%.

Từ các ví dụ trên, ta có thể thấy rằng việc bác bỏ hay không bác bỏ một giả thiết H_0 phụ thuộc vào cỡ mẫu và mức ý nghĩa được chọn.

Kiểm định mức ý nghĩa

Kiểm định ý nghĩa cũng sẽ cho kết quả bác bỏ hay không bác bỏ giả thiết H_0 giống như kiểm định dựa vào khoảng tin cậy. Tuy nhiên,

phương pháp kiểm định này đơn giản hơn và cho kết quả nhanh hơn. Các bước thực hiện kiểm định ý nghĩa sẽ như sau:

- Bước 1: Xác định giả thiết H_0 và H_1

$$H_0: \rho_1 = 0$$

$$H_1: \rho_1 \neq 0$$

- Bước 2: Tính sai số chuẩn $se(r_1) = 1/sqrt(12) = 0.289$.

- Bước 3: Tính giá trị *t* tính toán theo công thức sau đây:

$$t = \frac{r_1 - \rho_1}{se(r_1)} = \frac{r_1 - 0}{se(r_1)} = \frac{r_1}{se(r_1)} \qquad (3.10)$$

$$t = \frac{0.572}{0.289} = 1.98 \qquad (3.11)$$

- Bước 4: So sánh giá trị *t* tính toán với giá trị *t* phê phán ở mức ý nghĩa được chọn, ví dụ 5%, ta thấy giá trị tuyệt đối của *t* tính toán bằng 1.98 < giá trị *t* phê phán 2.2 (Giá trị t phê phán 2 đuôi tại mức ý nghĩa $\alpha = 5\%$, với bậc tự do d.f. = 12 – 1 được tính trên Excel bằng công thức =TINV(5%,11) = 2.2). Như vậy, ta chấp nhận giả thiết H_0 ở mức ý nghĩa 5%. Ngược lại, nếu mức ý nghĩa $\alpha = 10\%$, với bậc tự do d.f. = 12 – 1 được tính trên Excel bằng công thức =TINV(10%,11) = 1.796, thì giá trị tuyệt đối của *t* tính toán bằng 1.98 > giá trị t tra bảng từ bảng phân phối *t*. Như vậy, ta bác bỏ giả thiết H_0 ở mức ý nghĩa 10%.

Trên cơ sở này, ta tiếp tục kiểm định xem hệ số tự tương quan ở độ trễ k = 2 có ý nghĩa thống kê hay không.

- Bước 1: Xác định giả thiết H_0 và H_1

$$H_0: \rho_2 = 0$$

$$H_1: \rho_2 \neq 0$$

- Bước 2: Tính sai số chuẩn $se(r_2)$:

$$se(r_2) = \sqrt{\frac{1+2\sum_{i=1}^{2-1} r_i^2}{n}} = \sqrt{\frac{1+2\sum_{i=1}^{2-1} 0.572^2}{12}} = \sqrt{\frac{1.6544}{12}} = 0.371$$

- Bước 3: Tính giá trị t tính toán theo công thức sau đây:

$$t = \frac{r_2 - \rho_2}{se(r_2)} = \frac{0.463}{0.371} = 1.25 \qquad (3.12)$$

- Bước 4: So sánh giá trị *t* tính toán với giá trị *t* phê phán ở mức ý nghĩa được chọn, ví dụ 5%, ta thấy giá trị tuyệt đối của *t* tính toán bằng 1.25 < giá trị *t* phê phán 2.2. Như vậy, ta chấp nhận giả thiết H_0 ở mức ý nghĩa 5%. Tương tự, nếu mức ý nghĩa α = 10%, với bậc tự do d.f. = 12 – 1 được tính trên Excel bằng công thức =TINV(10%,11) = 1.796, thì giá trị tuyệt đối của *t* tính toán bằng 1.25 < giá trị *t* phê phán. Như vậy, ta cũng chấp nhận giả thiết H_0 ở mức ý nghĩa 10%.

Thống kê Q

Hai cột cuối trong biểu đồ tự tương quan là thống kê *Q* của Ljung-Box và giá trị xác suất tương ứng. Thống kê *Q* kiểm định giả thiết đồng thời là tất cả các hệ số ρ_k cho tới một độ trễ *k* đồng thời bằng không. Giá trị thống kê *Q* tính toán theo công thức sau đây[2]:

$$Q = n\sum_{k=1}^{m} \rho_k^2 \qquad (3.13)$$

Với cỡ mẫu lớn, *Q* có phân phối theo χ^2 với bậc tự do bằng số độ trễ. Nếu giá trị thống kê *Q* tính toán lớn hơn giá trị thống kê *Q* phê phán ở một mức ý nghĩa xác định, ta bác bỏ giả thiết H_0.

---

[2] Các phần mềm kinh tế lượng thường sử dụng một biến dạng của công thức (3.13) như sau:

$$Q = n(n+2)\sum_{k=1}^{m} \frac{r_k^2}{n-k} \qquad (3.14)$$

Trong đó, n = số quan sát trong chuỗi thời gian; k = số độ trễ; m = số số độ trễ sẽ được kiểm định; rk = hàm hệ số tự tương quan mẫu ở độ trễ k.

Các bước thực hiện kiểm định Q như sau:

- Bước 1: Xác định giả thiết H_0 và H_1

$$H_0: \rho_1 = \rho_2 = 0$$

H_1: Ít nhất một hệ số khác không

- Bước 2: Tính giá trị thống kê Q (tức χ^2 tính toán). Từ kết quả ở Hình 3.6 ta thấy giá trị thống kê $Q = 8.59$.
- Bước 3: Tính giá trị χ^2 phê phán ở mức ý nghĩa $\alpha = 5\%$ và số bậc tự do d.f. = 2. Áp dụng công thức CHIINV(5%,2) trên Excel ta có giá trị χ^2 phê phán này bằng 5.99.
- Bước 4: So sánh giá trị χ^2 tính toán (8.59) và χ^2 phê phán (5.99), ta kết luận bác bỏ giả thiết H_0.

HỆ SỐ TỰ TƯƠNG QUAN VÀ NHẬN DẠNG DỮ LIỆU

Sử dụng các hệ số tự tương quan ở các độ trễ khác nhau của một biến số nhất định (Y_t) nhằm trả lời các câu hỏi sau đây:

1. Có phải Y_t là một chuỗi ngẫu nhiên?
2. Có phải Y_t là một chuỗi dừng?
3. Có phải Y_t là một chuỗi có xu thế (không dừng)?
4. Có phải Y_t là một chuỗi có yếu tố mùa vụ?

Hanke (2005) đưa ra các kết luận như sau:

- Nếu một chuỗi được cho là "ngẫu nhiên", thì các hệ số tự tương quan giữa Y_t và Y_{t-k} cho bất kỳ độ trễ k nào đều gần bằng không. Điều này có nghĩa là các giá trị kế nhau trong một chuỗi thời

gian không có liên quan gì với nhau. Kết luận này rất có ý nghĩa trong việc kiểm định phần dư của một mô hình hồi quy có thể theo phân phối chuẩn.

- Nếu một chuỗi được cho là "dừng", thì hệ số tự tương quan bậc một khác không một cách có ý nghĩa thống kê, nhưng các hệ số tự tương quan bậc hai hoặc bậc ba bằng không. Như vậy, khi quan sát giản đồ tự tương quan, ta nhận thấy các hệ số tự tương quan giảm xuống bằng không một cách nhanh chóng sau hai hoặc ba độ trễ. Như vậy, một chuỗi ngẫu nhiên hiển nhiên là một chuỗi dừng. Lưu ý, chúng ta sẽ thảo luận các đặc điểm của một chuỗi dừng ở phần sau.

- Nếu một chuỗi được cho là "có xu thế", nghĩa là các giá trị kế nhau trong chuỗi thời gian có mức độ tương quan cao với nhau, thì các hệ số tự tương quan khác không một cách có ý nghĩa thống kê cho một số các độ trễ đầu tiên và sẽ dần dần giảm về không khi số độ trễ tăng lên. Trong trường hợp này, hệ số tự tương quan bậc một thường rất cao (gần bằng 1). Hệ số tự tương quan bậc hai cũng rất cao, nhưng thấp hơn hệ số tự tương quan bậc một.

- Nếu một chuỗi thời gian được cho là "có yếu tố mùa vụ", thì hệ số tự tương quan tại một độ trễ mùa (hoặc một số độ trễ mùa) khác không một cách có ý nghĩa thống kê. Độ trễ mùa cho dữ liệu theo quý bằng 4 và cho dữ liệu theo tháng bằng 12.

CHUỖI DỮ LIỆU NGẪU NHIÊN

Một chuỗi ngẫu nhiên Y_t thường chứa đựng hai thành phần: c, giá trị trung bình không đổi; và ε_t, sai số ngẫu nhiên (thường là nhiễu trắng). Thành phần ε_t được giả định là không có sự tương quan giữa các độ trễ khác nhau. Như vậy, nếu Y_t là ngẫu nhiên, thì Y_t có thể được viết như sau:

$$Y_t = c + \varepsilon_t \qquad (3.15)$$

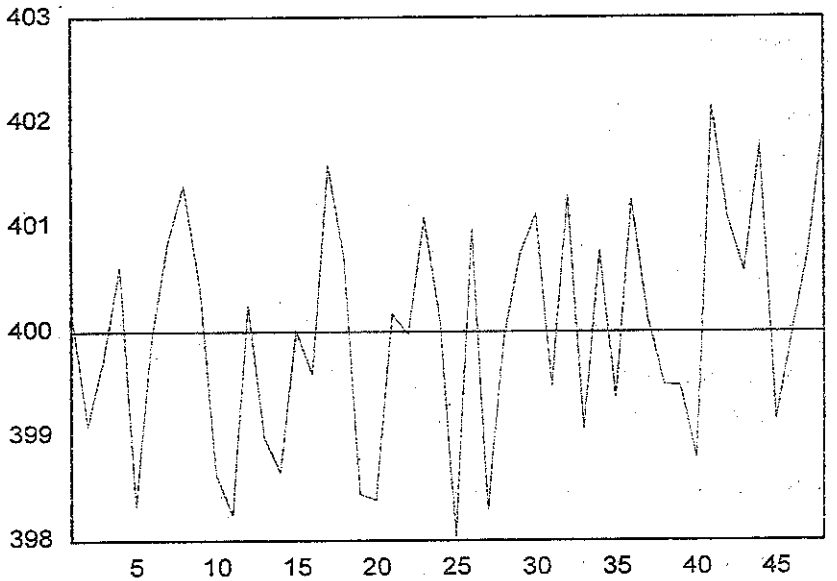
Để minh họa chuỗi ngẫu nhiên, chúng ta có thể tạo một chuỗi Y_t có giá trị trung bình bằng 400 trên Eviews theo hướng dẫn sau đây. Mở một tập tin Eviews mới, theo dữ liệu chéo (Undated), với n = 48. Sau đó thực hiện các lệnh sau đây trên cửa sổ lệnh:

Genr X_t=nrnd

Genr Y_t=X_t+400

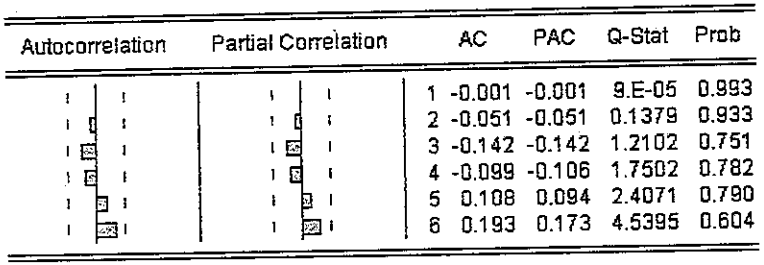
Plot Y_t

■ HÌNH 3.7: **Biến ngẫu nhiên** Y_t.



Để vẽ đường giá trị trung bình bằng 400 trên cùng đồ thị (Hình 3.7), chúng ta tạo thêm một biến (ví dụ Z=400), rồi vẽ Y_t và Z trên cùng đồ thị. Để kiểm định xem Y_t có phải là một chuỗi ngẫu nhiên hay không, ta có thể vẽ và quan sát giản đồ tự tương quan. Vào Quyck/Series Statistics/Correlogram, nhập tên biến "Y_t" vào, ta sẽ có đồ thị sau đây:

■ HÌNH 3.8: **Giản đồ tự tương quan biến ngẫu nhiên Y_t.**



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.001	-0.001	9.E-05	0.993
		2	-0.051	-0.051	0.1379	0.933
		3	-0.142	-0.142	1.2102	0.751
		4	-0.099	-0.106	1.7502	0.782
		5	0.108	0.094	2.4071	0.790
		6	0.193	0.173	4.5395	0.604

Như vậy, các hệ số tự tương quan r_1 và r_2 bằng không một cách có ý nghĩa thống kê (do giá trị xác suất p khá cao, 0.933). Theo kết luận của Hanke (2005), ta có thể nói rằng Y_t là một chuỗi ngẫu nhiên.

CHUỖI DỮ LIỆU DỪNG

Một khái niệm quan trọng trong các quy trình phân tích chuỗi thời gian là tính dừng. Một chuỗi dừng có các đặc điểm sau đây:

- Thể hiện xu hướng trở lại trạng thái trung bình theo một cách trong đó dữ liệu dao động xung quanh một giá trị trung bình cố định trong dài hạn.
- Có một giá trị phương sai xác định không thay đổi theo thời gian.
- Có một giản đồ tự tương quan với các hệ số tự tương quan sẽ giảm dần khi độ trễ tăng lên.

Theo ngôn ngữ thống kê, các đặc điểm trên của một chuỗi thời gian Y_t được thể hiện như sau:

- $E(Y_t)$ là một hằng số cho tất cả các thời điểm *t*

$$E(Y_t) = \mu \tag{3.16}$$

- $Var(Y_t)$ là một hằng số cho tất cả các thời điểm *t*

$$Var(Y_t) = E(Y_t-\mu)^2 = \sigma^2 \tag{3.17}$$

- $Cov(Y_t, Y_{t+k})$ là một hằng số cho tất cả các thời điểm t và k khác không. Lưu ý, giá trị của hiệp phương sai giữa hai giai đoạn chỉ phụ thuộc vào khoảng cách giữa hai giai đoạn.

$$Cov(Y_t, Y_{t+k}) = \gamma_k = E[(Y_t-\mu)(Y_{t+k}-\mu)] \quad (3.18)$$

Trong đó, γ_k là hiệp phương sai ở độ trễ k, là hiệp phương sai giữa các giá trị Y_t và Y_{t+k} (hoặc Y_{t-k}); nghĩa là, giữa hai giá trị Y cách nhau k thời đoạn. Nếu $k = 0$, ta có γ_0, đó cũng chính là phương sai của Y (σ^2); nếu $k = 1$, γ_1 là hiệp phương sai giữa hai giá trị Y liền nhau (xem lại chương 2).

Giả sử khi ta di chuyển giá trị gốc của Y từ Y_t sang Y_{t+k} (ví dụ, từ quý I năm 1975 sang quý I năm 1985). Nếu Y_t là một chuỗi dừng, thì giá trị trung bình, phương sai, và hiệp phương sai của Y_{t+m} phải bằng giá đại lượng này của Y_t. Tóm lại, nếu một chuỗi dừng, thì giá trị trung bình, phương sai, và hiệp phương sai (ở các độ trễ khác nhau) sẽ giống nhau không cần biết ta đang đo lường chúng tại thời điểm nào; điều này có nghĩa là, các đại lượng này không thay đổi theo thời gian. Một chuỗi dữ liệu như vậy sẽ có xu hướng trở về giá trị trung bình và những dao động xung quanh giá trị trung bình (đo bằng phương sai) sẽ là như nhau. Trong khi đó, nếu một chuỗi thời gian không dừng theo cách ta vừa định nghĩa ở trên, thì ta gọi đó là chuỗi không dừng[3]. Nói cách khác, một chuỗi thời gian không dừng sẽ có giá trị trung bình thay đổi theo thời gian, hoặc giá trị phương sai thay đổi theo thời gian, hoặc cả hai.

Tại sao chuỗi thời gian dừng lại quan trọng? Có hai lý do quan trọng khi biết một chuỗi thời gian là dừng hay không. Thứ nhất, Gujarati (2003) cho rằng nếu một chuỗi thời gian không dừng, chúng ta chỉ có thể nghiên cứu hành vi của nó chỉ trong khoảng thời gian đang được xem xét. Vì thế, mỗi một mẫu dữ liệu thời gian sẽ mang một tình tiết nhất định. Kết quả là, chúng ta không thể khái quát hóa cho các giai đoạn thời gian khác. Đối với mục đích dự báo, các chuỗi thời gian không dừng như vậy có thể sẽ không có giá trị thực tiễn. Vì như chúng ta đã biết, trong dự báo chuỗi thời gian, chúng ta luôn giả

[3] Sẽ được trình bày một cách chi tiết hơn ở chương 8.

định rằng xu hướng vận động của dữ liệu trong quá khứ và hiện tại được duy trì cho các giai đoạn tương lai. Và như vậy, chúng ta không thể dự báo được điều gì cho tương lai nếu như bản thân dữ liệu luôn thay đổi. Hơn nữa, đối với phân tích hồi quy, nếu chuỗi thời gian không dừng thì tất cả các kết quả điển hình của một phân tích hồi quy tuyến tính cổ điển sẽ không có giá trị, không có ý nghĩa và thường được gọi là hiện tượng "hồi quy giả mạo". Thứ hai, khi biết dữ liệu dừng hay không, chúng ta sẽ giới hạn được số mô hình dự báo phù hợp nhất cho dữ liệu.

Để tạo một chuỗi dừng Y_t có giá trị trung bình bằng 100 trên Eviews, trước hết ta tạo một tập tin mới với 48 quan sát, rồi thực hiện theo hướng dẫn sau đây:

Smpl 1 1

Genr X_t=0

Smpl 2 48

Genr X_t=0.5*X_t(-1)+nrnd

Smpl 1 48

Genr Y_t=X_t+100

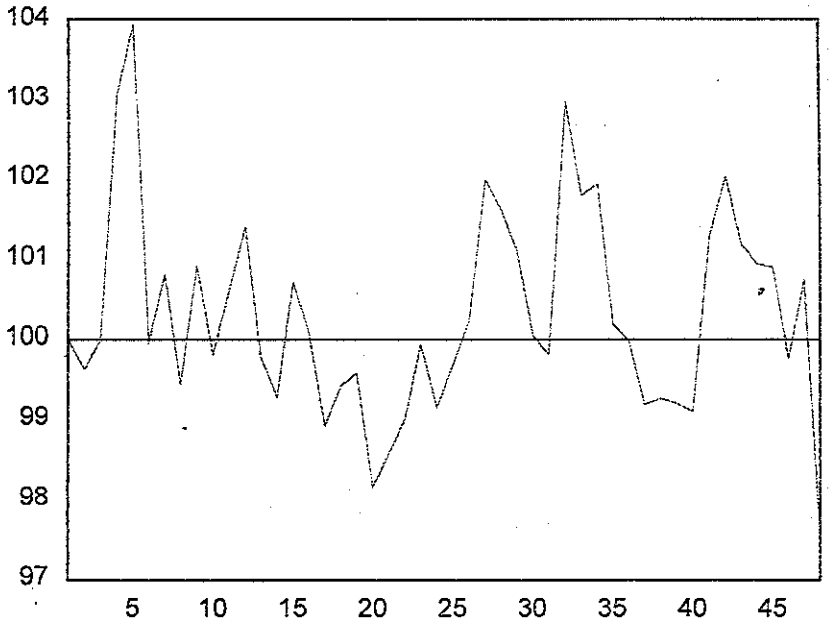
Quyck/Series Statistics/Correlogram

HÌNH 3.9: Giản đồ tự tương quan của của dừng Y_t.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.408	0.408	8.5211	0.004
		2	0.155	-0.014	9.7759	0.008
		3	-0.027	-0.102	9.8141	0.020
		4	-0.021	0.034	9.8378	0.043
		5	-0.062	-0.058	10.050	0.074
		6	-0.113	-0.092	10.786	0.095
		7	-0.054	0.042	10.954	0.141
		8	-0.053	-0.043	11.123	0.195
		9	-0.076	-0.072	11.480	0.244
		10	-0.073	-0.013	11.818	0.297

Chuỗi Y_t này có đồ thị như ở Hình 3.10:

■ HÌNH 3.10: **Đồ thị chuỗi dừng Y_t.**



CHUỖI DỮ LIỆU CÓ XU THẾ

Nếu một chuỗi thời gian có yếu tố xu thế, thì các giá trị liên tiếp của nó có mối quan hệ với nhau khá có ý nghĩa. Các hệ số tự tương quan của các độ trễ đầu tiên rất lớn và sẽ giảm dần về không khi số độ trễ tăng lên. Một chuỗi có yếu tố xu thế được gọi là chuỗi không dừng. Thông thường, khi phân tích các chuỗi không dừng, chúng ta cần loại bỏ yếu tố xu thế trước khi xác định mô hình dự báo. Có nhiều cách loại bỏ yếu tố xu thế trong chuỗi thời gian, nhưng thông thường nhất là lấy sai phân. Sai phân nghĩa là gì? Giả sử ta có chuỗi dữ liệu Y_t, thì sai phân của Y_t sẽ được định nghĩa như sau:

- Sai phân bậc 1: $\Delta Y_t = Y_t - Y_{t-1}$

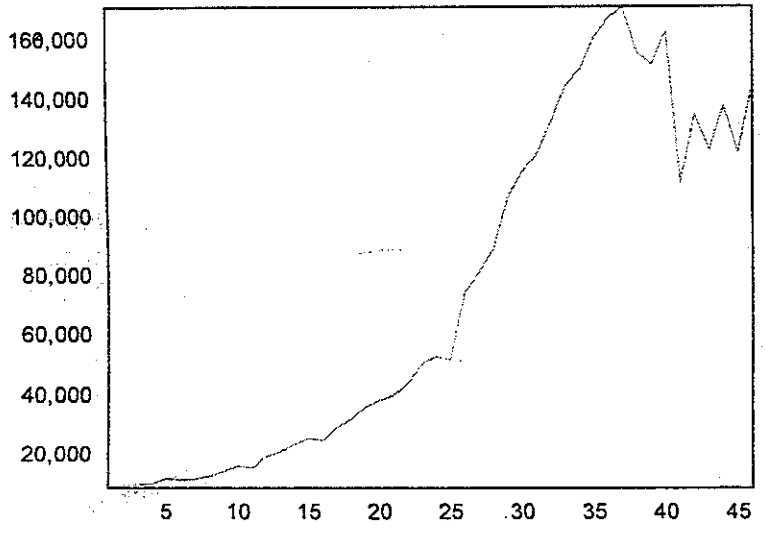
- Sai phân bậc 2: $\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1}$
- Sai phân bậc 1 với độ trễ theo quý: $\Delta Y_t = Y_t - Y_{t-4}$
- Sai phân bậc 1 với độ trễ theo tháng: $\Delta Y_t = Y_t - Y_{t-12}$

Trên Eviews, sai phân được tính như sau:

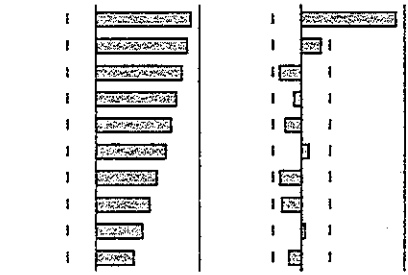
- Sai phân bậc 1: $dY_t=d(Y_t)$
- Sai phân bậc 2: $d2Y_t=d(Y_t,2)$
- Sai phân bậc 1 với độ trễ theo quý: $dqY_t=Y_t-Y_t(-4)$
- Sai phân bậc 1 với độ trễ theo tháng: $dmY_t=Y_t-Y_t(-12)$

Sử dụng tập tin **DATA3-6**, ta vẽ được đồ thị như sau:

HÌNH 3.11: **Đồ thị chuỗi có yếu tố xu thế Y_t.**



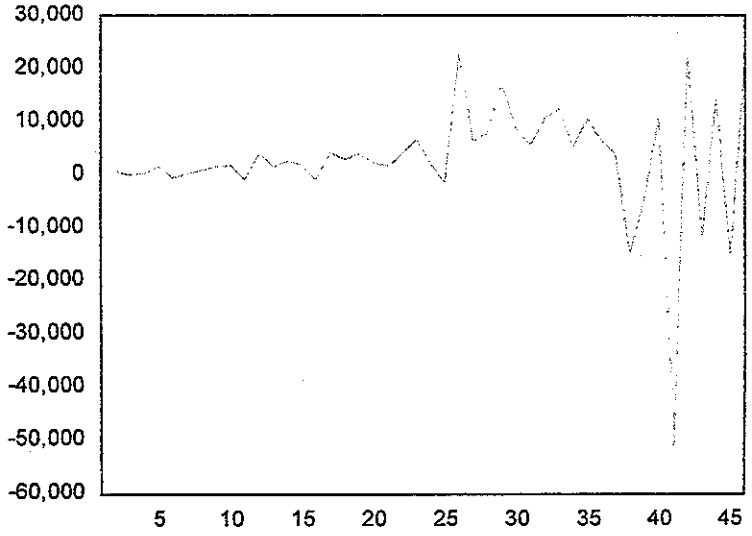
■ HÌNH 3.12: **Giản đồ tự tương quan của chuỗi xu thế Y_t.**



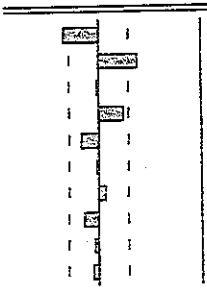
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.948	0.948	44.115	0.000
		2	0.919	0.201	86.528	0.000
		3	0.865	-0.226	124.96	0.000
		4	0.816	-0.068	160.00	0.000
		5	0.750	-0.164	190.28	0.000
		6	0.699	0.071	217.25	0.000
		7	0.621	-0.215	239.10	0.000
		8	0.543	-0.199	256.23	0.000
		9	0.466	0.033	269.18	0.000
		10	0.381	-0.125	278.06	0.000

Như vậy, nhìn vào đồ thị Hình 3.11 và giản đồ tự tương quan Hình 3.12, ta nhận thấy rằng Y_t là một chuỗi có xu thế tăng và đó là một chuỗi không dừng. Nếu lấy sai phân bậc một của Y_t, ta có chuỗi dữ liệu chuyển hóa là một chuỗi dừng như ở Hình 3.13 và Hình 3.14.

■ HÌNH 3.13: **Đồ thị sai phân bậc 1 của chuỗi xu thế Y_t.**



■ HÌNH 3.14: **Giản đồ tự tương quan của chuỗi $d(Y_t)$.**



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.364	-0.364	6.3675	0.012
		2	0.390	0.297	13.855	0.001
		3	-0.016	0.243	13.868	0.003
		4	0.257	0.269	17.287	0.002
		5	-0.185	-0.161	19.103	0.002
		6	-0.013	-0.435	19.112	0.004
		7	0.076	-0.041	19.438	0.007
		8	-0.151	0.095	20.746	0.008
		9	-0.046	0.074	20.872	0.013
		10	-0.049	-0.003	21.019	0.021

Như vậy, sai phân bậc một chuỗi xu thế có thể là một chuỗi dừng. Lưu ý, trong kinh tế lượng chuỗi thời gian, người ta còn phân biệt giữa một chuỗi dừng sai phân và một chuỗi dừng xu thế[4]. Như vậy, khi gặp một chuỗi không dừng, chúng ta có thể chuyển sai dạng sai phân, rồi xác định mô hình dự báo phù hợp cho chuỗi sai phân. Khi đó, chúng ta ước tính giá trị dự báo cho chuỗi sai phân ($\Delta\hat{Y}_{t+1}$) trước, rồi từ đó dự báo cho chuỗi dữ liệu gốc ($\hat{Y}_{t+1}$).

CHUỖI DỮ LIỆU CÓ YẾU TỐ MÙA

Nếu một chuỗi có yếu tố mùa, thì dạng dữ liệu của nó sẽ được lặp đi lặp lại qua một khoảng thời gian nhất định (thường là một năm). Các quan sát trong các "mùa" giống nhau (ví dụ quý I năm 2006 và quý I năm 2007) có xu hướng tương quan với nhau. Nếu là dữ liệu theo quý, thì các quý I trông có vẻ giống nhau, các quý II trông có vẻ giống nhau, v.v... Khi đó, các hệ số tự tương quan với độ trễ k = 4 có thể có ý nghĩa thống kê. Tương tự, nếu là dữ liệu theo tháng, thì các tháng Một trông có vẽ giống nhau, các tháng Hai trông có vẽ giống nhau, v.v... Khi đó, các hệ số tự tương quan với độ trễ k = 12 có thể có ý nghĩa thống kê. Trong các mô hình dự báo doanh số, lượng khách du lịch, v.v..., thông thường chúng ta nên lưu ý đến yếu tố mùa vụ. Ví dụ,

[4] Xem "Gujarati, 2009, Basic Econometrics, 5th Edition, McGraw-Hill".

xem xét dữ liệu doanh số của công ty XYZ (**DATA3-7**, Bảng 3.3) có thể phù hợp với dạng dữ liệu theo quý.

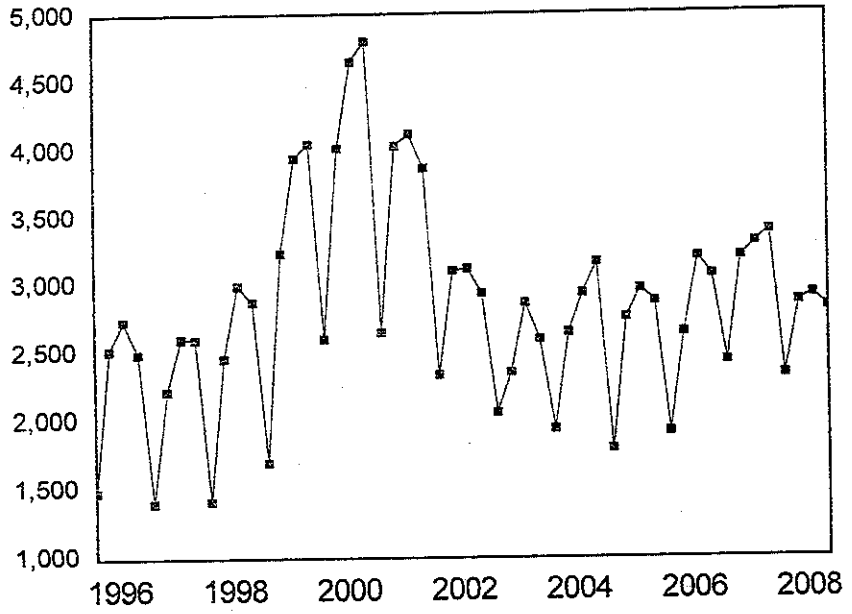
■ BẢNG 3.3: **Doanh số của công ty XYZ (triệu đồng).**

Năm	Quý I	Quý II	Quý III	Quý IV
1996	1483	2525	2738	2498
1997	1400	2219	2609	2602
1998	1412	2462	2995	2877
1999	1695	3233	3942	4050
2000	2604	4018	4653	4804
2001	2651	4033	4120	3866
2002	2334	3099	3114	2937
2003	2058	2351	2861	2594
2004	1939	2644	2932	3159
2005	1790	2752	2961	2871
2006	1915	2642	3195	3062
2007	2433	3195	3303	3389
2008	2328	2863	2917	2821

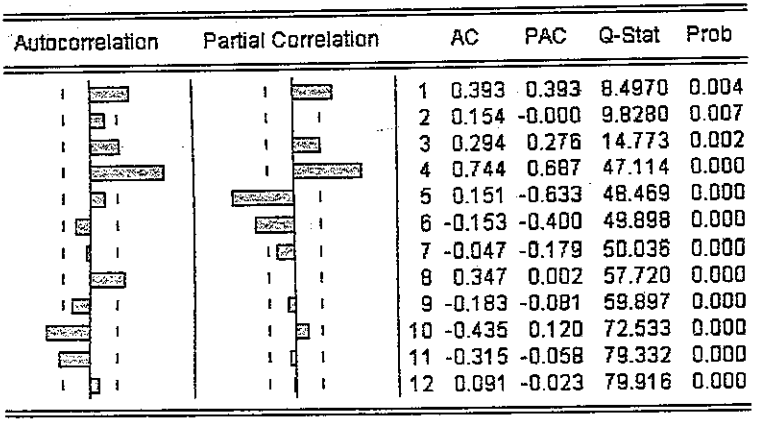
Quan sát dữ liệu theo quý trong từng năm, chúng ta nhận thấy dữ liệu có dạng đặc trưng như sau: Doanh số quý I thấp nhất, quý II và quý III tăng lên, sau đó có xu hướng giảm xuống trong quý IV.

Hình 3.15 và Hình 3.16 cho thấy doanh số của công ty XYZ có yếu tố mùa vụ. Trong giản đồ tự tương quan ta thấy rằng các hệ số tự tương quan ở các độ trễ k = 4 (giữa Y_t và Y_{t-4})và k = 8 (giữa Y_t và Y_{t-8}) khác không một cách có ý nghĩa thống kê. Nếu không kể yếu tố mùa vụ, thì dữ liệu này là một chuỗi dừng vì hệ số tự tương quan đầu tiên khác không, nhưng các hệ số tự tương quan bậc hai và bậc ba bằng không một cách có ý nghĩa thống kê.

■ HÌNH 3.15: **Giản đồ tự tương quan của chuỗi $d(Y_t)$.**



■ HÌNH 3.16: **Giản đồ tự tương quan của chuỗi $d(Y_t)$.**



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.393	0.393	8.4970	0.004
		2	0.154	-0.000	9.8280	0.007
		3	0.294	0.276	14.773	0.002
		4	0.744	0.687	47.114	0.000
		5	0.151	-0.633	48.469	0.000
		6	-0.153	-0.400	49.898	0.000
		7	-0.047	-0.179	50.036	0.000
		8	0.347	0.002	57.720	0.000
		9	-0.183	-0.081	59.897	0.000
		10	-0.435	0.120	72.533	0.000
		11	-0.315	-0.058	79.332	0.000
		12	0.091	-0.023	79.916	0.000

LỰA CHỌN MÔ HÌNH DỰ BÁO

Trên cơ sở tìm hiểu hệ số tự tương quan, giản đồ tự tương quan và cách nhận biết dạng dữ liệu của các chuỗi thời gian khác nhau, phần này sẽ trình bày các kỹ thuật dự báo khác nhau tương ứng với từng dạng dữ liệu cũng như các tính chất đặc trưng khác của từng mô hình dự báo.

Trước khi quyết định lựa chọn mô hình dự báo nào, theo Hanke (2005), người làm công tác dự báo cần trả lời được các câu hỏi sau đây:

1. Tại sao cần thực hiện việc dự báo này?
2. Ai sẽ là người sử dụng kết quả dự báo?
3. Các đặc điểm cơ bản của dữ liệu sẵn có là gì?
4. Thời đoạn dự báo là gì?
5. Dữ liệu tối thiểu đòi hỏi phải có là bao nhiêu?
6. Mức độ chính xác mong muốn của dự báo là bao nhiêu?
7. Chi phí để thực hiện việc dự báo là bao nhiêu?

Để lựa chọn kỹ thuật dự báo phù hợp, Hanke (2005) cũng đề xuất người làm dự báo cũng cần phải đáp ứng các yêu cầu sau đây:

1. Xác định bản chất của vấn đề dự báo.
2. Giải thích bản chất của dữ liệu đang được xem xét.
3. Mô tả được các khả năng và các hạn chế của các kỹ thuật dự báo có thể áp dụng.
4. Xây dựng các tiêu chí đánh giá độ chính xác để lựa chọn giữa các mô hình dự báo.

Một nhân tố chủ yếu ảnh hưởng việc lựa chọn một kỹ thuật dự báo thích hợp là việc nhận dạng và hiểu được các dạng dữ liệu quá khứ của dữ liệu. Tùy vào dữ liệu có dạng gì: dừng, xu thế, mùa vụ, hay chu kỳ, hay kết hợp các dạng trên, mà chúng ta có thể xác định được giới hạn các mô hình dự báo thích hợp cho từng trường hợp cụ thể. Dưới đây là bốn yếu tố quan trọng nhất quyết định việc lựa chọn kỹ thuật dự báo thích hợp cho một dữ liệu sẵn có:

1. Dạng dữ liệu quá khứ: Bốn dạng dữ liệu cơ bản là dữ liệu dừng, dữ liệu xu thế, dữ liệu mùa, và dữ liệu chu kỳ.

2. Thời đoạn dự báo: Dự báo có thể là dự báo ngắn hạn (dưới 3 tháng), dự báo trung hạn (từ 3 đến 18 tháng), và dự báo dài hạn.

3. Loại mô hình dự báo: Đối với dự báo định lượng, hai dạng mô hình hay được sử dụng là các mô hình chuỗi thời gian và các mô hình nhân quả.

4. Dữ liệu tối thiểu cần thiết: Tùy thuộc vào dạng mô hình, dạng dữ liệu, và mục đích dự báo mà số lượng quan sát đòi hỏi phải có là khác nhau.

Dưới đây là bảng tóm tắt các yếu tố ảnh hưởng đến việc chọn lựa kỹ thuật dự báo cho từng trường hợp nhất định.

■ BẢNG 3.4: **Lựa chọn mô hình dự báo.**

Phương pháp	Dạng dữ liệu	Thời đoạn dự báo	Loại mô hình	Dữ liệu tối thiểu	
				Không có mùa vụ	Có mùa vụ
Dự báo thô	ST, T, S	S	TS	1	
Trung bình giản đơn	ST	S	TS	30	
Trung bình di động	ST	S	TS	4-20	
San mũ giản đơn	ST	S	TS	2	

San mũ Holt	T	S	TS	3	
San mũ Winters	S	S	TS		2 x s
Hồi quy đơn	T	I	C	10	
Hồi quy bội	C, S	I	C	10xV	
Phân tích	S	S	TS		5 x s
Xu thế	T	I, L	TS	10	
ARIMA	ST, T, C, S	S	TS	24	
ARCH	ST, T, C, S	S	TS	24	
Hồi quy chuỗi thời gian	T, S	I, L	C		6 x s

Dạng dữ liệu: ST = Chuỗi dừng, T = Xu thế, S = Mùa vụ, C = Chu kỳ
Thời đoạn dự báo: S = Ngắn hạn, I = Trung hạn, L = Dài hạn
Loại mô hình: TS = Chuỗi thời gian, C = Nhân quả
Yêu cầu dữ liệu: V = Số biến giải thích, s = Số mùa vụ (4 hoặc 12)

Nguồn: Hanke, 2005, và Wilson, 2007.

Lưu ý, trên đây chỉ xét các mô hình với dữ liệu chuỗi thời gian. Đối với các mô hình dữ liệu chéo, chúng ta cần dựa vào cơ sở lý thuyết và khung phân tích để xác định các biến cũng như cỡ mẫu. Ngoài ra, còn một số các mô hình dự báo khác, nhưng phạm vi cuốn sách này chỉ tập trung vào các kỹ thuật được đề cập trong Bảng 3.4.

CÁC KỸ THUẬT DỰ BÁO VỚI DỮ LIỆU DỪNG

Như định nghĩa ở trên, một chuỗi dừng là một chuỗi có giá trị trung bình không đổi theo thời gian. Các trường hợp này xảy ra khi môi trường ảnh hưởng đến chuỗi dữ liệu tương đối ổn định. Ở dạng đơn giản nhất, dự báo một chuỗi dữ liệu dừng liên quan đến việc sử dụng thông tin quá khứ của dữ liệu để ước tính giá trị trung bình, và giá trị trung bình này sẽ được sử dụng làm giá trị dự báo cho các giai đoạn tương lai. Các kỹ thuật phức tạp hơn có xu hướng cập nhật giá trị ước lượng khi có thêm các thông tin mới cập nhật. Các kỹ thuật này sẽ trở nên hữu ích khi các ước lượng ban đầu tỏ ra không còn đáng tin cậy hoặc có sự hoài nghi về mức độ ổn định của giá trị trung bình. Ngoài ra, các kỹ thuật mới hơn trong chừng mực nào đó có đề cập đến phản

ứng của những thay đổi trong cấu trúc dữ liệu. Các kỹ thuật dự báo với dữ liệu dừng được sử dụng khi:

- Các nhân tố tạo nên chuỗi dữ liệu có tính ổn định và môi trường trong đó chuỗi dữ liệu tồn tại tương đối không đổi. Ví dụ, số ca làm việc/tuần của một chuyền sản xuất tương đối ổn định, doanh số đơn vị của một sản phẩm hay dịch vụ đang trong giai đoạn bảo hòa của chu kỳ kinh doanh, và số cửa hàng của một công ty bán lẻ tương đối ổn định qua thời gian.
- Một mô hình dự báo đơn giản nhất được sử dụng bởi vì thiếu dữ liệu hoặc nhằm dễ giải thích cho người sử dụng kết quả dự báo, hoặc dễ thực hiện việc dự báo. Ví dụ, khi một doanh nghiệp mới hoạt động nên có rất ít dữ liệu quá khứ.
- Có thể đạt được sự ổn định bằng cách thực hiện các điều chỉnh giản đơn các yếu tố như tốc độ tăng dân số hay lạm phát. Ví dụ, thay đổi thu nhập bình quân đầu người hoặc thay đổi giá trị doanh số sau khi đã khử lạm phát.
- Chuỗi dữ liệu có thể được chuyển hóa sang một chuỗi có tính ổn định hơn. Ví dụ, dữ liệu được chuyển sang dạng sai phân, lấy logarith, hay lấy căn bậc hai.
- Dữ liệu là một tập hợp của các sai số dự báo từ một kỹ thuật dự báo nào đó cũng được xem như có tính ổn định.

Các kỹ thuật dự báo có thể phù hợp với dạng dữ liệu này bao gồm các mô hình dự báo thô, các phương pháp trung bình giản đơn, các mô hình trung bình di động, các mô hình ARIMA.

CÁC KỸ THUẬT DỰ BÁO VỚI DỮ LIỆU XU THẾ

Như đã định nghĩa ở trên, một chuỗi xu thế là một chuỗi có một thành phần dài hạn có xu hướng tăng hoặc giảm theo thời gian. Nói cách khác, một chuỗi thời gian được cho là có yếu tố xu thế nếu giá trị trung bình của nó thay đổi qua thời gian (có thể tăng hoặc giảm). Các kỹ thuật dự báo với chuỗi xu thế phù hợp trong các trường hợp sau đây:

- Tăng năng suất hay thay đổi công nghệ dẫn đến thay đổi trong lối sống. Ví dụ, nhu cầu mua sắm các thiết bị điện tử gai tăng khi công nghệ máy tính phát triển nhanh chóng, nhu cầu sử dụng phương tiện đi lại bằng đường sắt giảm khi ngành hàng không phát triển.

- Gia tăng dân số làm tăng nhu cầu hàng hóa và dịch vụ. Ví dụ, doanh số các hàng hóa tiêu dùng, nhu cầu năng lượng, và các nguyên vật liệu có xu hướng gia tăng.

- Sức mua của một đồng tiền ảnh hưởng đến nhiều chỉ báo kinh tế do yếu tố lạm phát. Ví dụ, tiền lương, chi phí sản xuất, và giá hàng hóa có xu hướng tăng do lạm phát.

- Sự chấp nhận của thị trường gia tăng. Ví dụ, trong giai đoạn tăng trưởng của một sản phẩm trong chu kỳ kinh doanh của một sản phẩm mới.

Các kỹ thuật dự báo phù hợp với dạng dữ liệu này bao gồm các mô hình trung bình di động, san mũ Holt, hồi quy đơn, mô hình hàm xu thế, mô hình ARIMA.

CÁC KỸ THUẬT DỰ BÁO VỚI DỮ LIỆU MÙA

Như đã định nghĩa ở trên, một chuỗi có yếu tố mùa là một chuỗi có dạng dữ liệu thay đổi có tính lặp đi lặp lại từ năm này sang năm khác. Các kỹ thuật này được sử dụng khi:

- Thời tiết, văn hóa, và lễ hội ảnh hưởng đến biến số cần dự báo. Ví dụ, lượng tiêu dùng điện, các hoạt động theo mùa đông hoặc mùa hè (thể thao, du lịch), thời trang, sản xuất nông nghiệp.

- Niên lịch ảnh hưởng đến biến số cần dự báo. Ví dụ, doanh số bán lẻ chịu ảnh hưởng bởi các kỳ nghĩ, ngày nghĩ cuối tuần, hoặc niên học.

Các kỹ thuật dự báo phù hợp với dạng dữ liệu này bao gồm các mô hình phân tích, san mũ Winters, hồi quy bội, các mô hình ARIMA.

CÁC KỸ THUẬT DỰ BÁO VỚI DỮ LIỆU CHU KỲ

Như được định nghĩa, một chuỗi thời gian có yếu tố chu kỳ thường có xu hướng dao động dạng bước sóng xung quanh một xu thế. Các dạng dữ liệu có tính chu kỳ thường rất khó mô hình hóa bởi vì các dạng dữ liệu không có tính ổn định. Các kỹ thuật dự báo này thường được sử dụng trong các trường hợp sau đây:

- Chu kỳ kinh doanh ảnh hưởng đến biến cần dự báo. Ví dụ, các yếu tố kinh tế, thị trường và cạnh tranh có thể ảnh hưởng đến doanh số.
- Xảy ra các xu hướng dịch chuyển trong sở thích của người tiêu dùng. Ví dụ như thời trang, âm nhạc và thức ăn.
- Xảy ra các dịch chuyển trong dân số. Ví dụ như chiến tranh, nghèo đói, bệnh dịch và thiên tai.
- Xảy ra các dịch chuyển trong vòng đời sản phẩm.

Các kỹ thuật phù hợp với dạng dữ liệu này bao gồm các mô hình phân tích, các mô hình kinh tế lượng, hồi quy bội, và các mô hình ARIMA.

CÁC YẾU TỐ KHÁC ẢNH HƯỞNG VIỆC CHỌN KỸ THUẬT DỰ BÁO

Độ dài dự báo có ảnh hưởng trực tiếp đến việc lựa chọn một kỹ thuật dự báo. Đối với các dự báo ngắn hạn và trung hạn, có rất nhiều mô hình dự báo định lượng có thể được áp dụng. Tuy nhiên, khi độ dài dự báo tăng lên, chỉ có một số ít trong các kỹ thuật này có thể áp dụng được. Ví dụ, các mô hình bình quân di động, san mũ, và ARIMA trở nên kém chính xác khi độ dài dự báo tăng lên. Ngược lại, các mô hình kinh tế lượng có thể hữu ích hơn khi độ dài dự báo tăng. Các mô hình hồi quy có thể phù hợp cho cả các dự báo ngắn hạn, trung hạn, và dài hạn. Các giá trị trung bình, bình quân di động, phân tích, và hồi quy hàm xu thế chỉ phù hợp đối với các dự báo ngắn và trung hạn. Các mô hình phức tạp hơn như ARIMA và các kỹ thuật kinh tế lượng cũng chỉ phù hợp đối với các dự báo ngắn và trung hạn. Lưu ý, các phương pháp định tính lại rất hữu ích cho các dự báo dài hạn.

Khả năng áp dụng các kỹ thuật dự báo nói chung phụ thuộc nhiều vào mức độ kinh nghiệm và khả năng của người làm dự báo. Nhiều giám đốc ở các doanh nghiệp thường cần các dự báo tương đối ngắn hạn. Các mô hình san mũ, hàm xu thế, các mô hình hồi quy, và phân tích thường rất hữu ích trong những trường hợp này.

Chi phí máy tính và phần mềm hiện nay không còn là vấn đề quan trọng trong việc lựa chọn các kỹ thuật dự báo. Ngày nay, rất nhiều phần mềm chuyên dụng đang được phát triển và áp dụng rộng rãi ở nhiều doanh nghiệp và tổ chức. Cho nên, đây không còn là một tiêu chí quan trọng khi lựa chọn kỹ thuật dự báo.

Kết quả dự báo sẽ được trình lên ban giám đốc để được xét duyệt và chấp nhận. Vì thế, việc dễ hiểu và dễ giải thích kết quả dự báo cũng là một vấn đề quan trọng cần được xem xét khi lựa chọn kỹ thuật dự báo. Các mô hình hồi quy, hàm xu thế, phân tích, và san mũ là các kỹ thuật dự báo chiếm ưu thế trong trường hợp này.

Ngoài ra, các tiêu chí đo lường độ chính xác của dự báo như đã được trình bày trong chương 1 cũng là nội dung cần xem xét khi lựa chọn các kỹ thuật dự báo. Nói chung, cùng một dữ liệu, mô hình nào cho giá trị dự báo bé hơn được cho là mô hình tốt hơn. Tuy nhiên, người làm công tác dự báo cần so sánh các giá trị dự báo với dữ liệu thực tế để có cái nhìn toàn diện hơn trong việc lựa chọn các kỹ thuật dự báo.

XÁC ĐỊNH ĐỘ CHÍNH XÁC CỦA KỸ THUẬT DỰ BÁO

Trước khi bắt đầu dự báo với một kỹ thuật được chọn, mức độ chính xác của sự lựa chọn cần được đánh giá. Hanke (2005) đề nghị người làm dự báo nên trả lời các câu hỏi sau đây:

1. Các hệ số tự tương quan của phần dư trong mô hình dự báo có ngẫu nhiên hay chưa? Câu hỏi này có thể được trả lời bằng cách xem xét giản đồ tự tương quan cho phần dư của mô hình dự báo.

2. Các phần dư của dự báo đã có phân phối chuẩn hay chưa? Câu hỏi này có thể được trả lời bằng cách phân tích đồ thị tần xuất và thống kê của phần dư. Cụ thể, chúng ta sử dụng kiểm định Jarque-Bera để kiểm định phần dư (như giới thiệu ở chương 2).
3. Các hệ số ước lượng (trong các mô hình hồi quy, ARIMA, ARCH, v.v..., có ý nghĩa thống kê hay không? Câu hỏi này có thể được trả lời bằng việc sử dụng thống kê t để kiểm định các hệ số hồi quy.
4. Các mô hình (chủ yếu các mô hình hồi quy) có bị các hiện tượng đa cộng tuyến, phương sai thay đổi, tự tương quan, hay hồi quy giả mạo hay không? Các vấn đề này sẽ được trình bày ở các chương 7.
5. Các mô hình dự báo (các mô hình ARIMA, ARCH, v.v...) có hiệu quả nhất chưa? Các tiêu chí thường được sử dụng là AIC, SBC, và HQ. Ngoài ra, chúng ta cũng nên quan sát đồ thị giá trị dự báo và giá trị thực để hỗ trợ cho các tiêu chí thống kê.
6. Kỹ thuật được chọn có đơn giản và dễ hiểu đối với những người ra quyết định hay không?

TÓM TẮT CHƯƠNG 3

Chương này đã cung cấp cho chúng ta một phương pháp luận tổng quát khá hoàn chỉnh để tiến hành hiệu quả một nghiên cứu dự báo cụ thể. Bất kỳ một nghiên cứu dự báo cụ thể nào đều đòi hỏi chuyên gia dự báo phải xác định được chất lượng của dữ liệu mà mình thu thập được, sau đó là phải phân tích tiền dự báo để hiểu được bản chất của dữ liệu đang chứa đựng những yếu tố cụ thể nào bằng các công cụ phân tích tương quan và tự tương quan với sự hỗ trợ của phần mềm. Đa số những dạng dữ liệu thời gian thường thể hiện ở dưới các dạng khác nhau hay là kết hợp như xu thế, mùa, chu kỳ, dừng. Khi hiểu được bản chất dữ liệu qua định dạng các yếu tố chứa đựng trong nó thì chuyên gia dự báo có khả năng áp dụng những phương pháp dự báo thích hợp tùy theo mục tiêu nghiên cứu và thời đoạn dự báo. Việc lựa chọn kết quả dự báo lại có những tiêu chí xem xét các sai số vốn có của những mô hình dự báo khác nhau. Sai số dự báo là điều không thể tránh khỏi vì ngoài các yếu tố mà chuỗi dữ liệu chứa đựng như đã thảo luận trong chương này thì các dao động ngẫu nhiên là một yếu tố mà chuyên gia dự báo không thể kiểm soát hiệu quả.

CÂU HỎI VÀ BÀI TẬP

1. Anh/Chị hãy trình bày các tiêu chí quan trọng để đánh giá chất lượng của dữ liệu dùng cho dự báo?
2. Anh/Chị hãy trình bày các thành phần cơ bản của một chuỗi thời gian?
3. Anh/Chị cho biết thế nào là một chuỗi dừng? Làm sao biết một chuỗi thời gian có tính dừng hay không? Và tại sao tính dừng có ý nghĩa quan trọng trong dự báo?
4. Anh/Chị cho biết việc lựa chọn mô hình dự báo thích hợp phụ thuộc vào các yếu tố nào?
5. Anh/Chị cho biết làm thế nào chúng ta có thể biết được độ chính xác của một mô hình dự báo?
6. Anh/Chị hãy liệt kê các phương pháp dự báo phù hợp với các loại dữ liệu sau đây:
   a. Dữ liệu dừng? Cho ví dụ?
   b. Dữ liệu có xu thế? Cho ví dụ?
   c. Dữ liệu có yếu tố mùa? Cho ví dụ?
7. Anh/Chị cho biết làm thế nào để đánh giá độ chính xác của dự báo trong các trường hợp sau đây:
   a. Người phân tích dự báo muốn xác định xem liệu một phương pháp dự báo có bị chệch hay không?
   b. Người phân tích dự báo cảm nhận rằng độ lớn của biến cần dự báo là yếu tố quan trọng trong việc đánh giá mức độ chính xác của dự báo?
   c. Người phân tích dự báo quan tâm nhiều đến các sai số dự báo lớn?
8. Tập tin **"REVENUE.xls"** chứa dữ liệu về doanh thu theo quý của 20 công ty niêm yết trên thị trường chứng khoán Việt Nam. Sử dụng tập tin này, Anh/Chị hãy trả lời các câu hỏi sau đây:
   a. Vẽ đồ thị doanh số của các công ty theo thời gian và nhận xét đặc điểm của các chuỗi dữ liệu này?

b. Vẽ giản đồ tự tương quan của từng chuỗi dữ liệu và cho biết đây có phải là các chuỗi dừng hay không? Tại sao?

c. Anh/Chị cho biết có thể sử dụng các mô hình dự báo nào để dự báo doanh số của các công ty này?

9. Tập tin **"PRICE.xls"** chứa dữ liệu về nhiều loại giá và chỉ số giá khác nhau (tổng hợp từ nguồn IMF và Reuters). Sử dụng tập tin này, Anh/Chị hãy trả lời các câu hỏi sau đây:

a. Lựa chọn và vẽ đồ thị một số giá/chỉ số giá theo thời gian và nhận xét đặc điểm của các chuỗi dữ liệu này?

b. Vẽ giản đồ tự tương quan của từng chuỗi dữ liệu và cho biết đây có phải là các chuỗi dừng hay không? Tại sao?

c. Anh/Chị cho biết trong các biến giá/chỉ số giá này, thì những biến nào có yếu tố mùa vụ?

10. Tập tin **"IMF.xls"** chứa các dữ liệu về kinh tế Việt Nam và một số dữ liệu của các nước trên thế giới (tổng hợp từ CD-ROM năm 2008 của IMF). Sử dụng tập tin này, Anh/Chị hãy khảo sát dạng dữ liệu của các biến trong tập tin này và đề xuất các mô hình dự báo thích hợp cho từng biến?

11. Sử dụng dữ liệu **"GAS.xls"**, Anh/Chị hãy trả lời các câu hỏi sau đây:

a. Vẽ đồ thị giá CP theo thời gian và giải thích tại sao giá CP có yếu tố mùa?

b. Vẽ đồ thị phân tán giữa biến CP và biến OIL, và giải thích mối quan hệ giữa hai biến này?

c. Lập bảng tính các hệ số tự tương quan của CP (với độ trễ là 12), và cho biết chuỗi dữ liệu CP có đặc điểm gì?

d. Từ kết quả ở câu c), Anh/Chị cho biết có thể sử dụng các mô hình nào để dự báo giá CP?

12. Sử dụng dữ liệu **"GAP.xls"**, Anh/Chị hãy trả lời những câu hỏi sau đây:

a. Vẽ đồ thị doanh số theo thời gian và giải thích tại sao doanh số của GAP có yếu tố mùa vụ?

b. Lập bảng tính các hệ số tự tương quan của doanh số (với độ trễ là 12), và cho biết chuỗi dữ liệu doanh số của GAP có đặc điểm gì?

c. Từ kết quả ở câu b), Anh/Chị cho biết có thể sử dụng các mô hình nào để dự báo doanh số của GAP?

d. Tính sai phân bậc 1 của doanh số và cho biết sai phân bậc 1 của doanh số là một chuỗi như thế nào?

13. Giám đốc điều hành của công ty CCC quan tâm đến số lượng và lập kế hoạch nhân sự cho các tháng còn lại trong năm. Cho nên Ông yêu cầu giám đốc nhân sự chuẩn bị việc dự báo. Cô trưởng phòng nhân sự đề xuất nên dựa vào dữ liệu khách hàng trong quá khứ để dự báo số khách hàng mới cho các tháng còn lại, rồi từ đó dự đoán được nhu cầu nhân sự. Cô trưởng phòng nhân sự đã thu thập dữ liệu về số khách hàng mới theo tháng, rồi sử dụng phương pháp quan sát đồ thị và phân tích hệ số tự tương quan. Sử dụng tập tin **"CCC.xls"**, Anh/Chị hãy trả lời các câu hỏi sau đây:

a. Cách lập luận của cô trưởng phòng nhân sự có hợp lý hay không? Tại sao?

b. Anh/Chị cho biết cô trưởng phòng nhân sự sử dụng phân tích hệ số tự tương quan như thế nào để khảo sát dạng dữ liệu về số khách hàng mới?

c. Anh/Chị cho biết cô trưởng phòng sẽ kết luận như thế nào sau khi phân tích hệ số tự tương quan?

d. Anh/Chị cho biết cô trưởng phòng nhân sự sẽ lựa chọn những mô hình dự báo nào?

14. Công ty Murphy Brothers[5], thành lập năm 1958, là một công ty nổi tiếng ở lĩnh vực kinh doanh hàng trang trí nội thất ở Mỹ. Đến năm 1996, Murphy Brothers đã thiết lập một hệ thống cửa hàng với mạng lưới dầy đặc ở 36 bang của Mỹ. Julie Murphy, con gái của một sáng lập viên công ty Murphy Brothers, vừa tốt nghiệp MBA và quyết định làm việc cho Murphy Brothers. Cha của cô (Glen Murphy) và những đồng sự khác vốn rất giỏi trong việc kinh doanh, nhưng không am hiểu nhiều về các phương pháp định lượng. Cụ thể, họ không thể dự báo doanh số tương lai của công ty bằng các kỹ thuật máy tính

[5] Hanke, 2007, Business Forecasting, 8th Edition, Pearson, pp.90-91.

hiện đại. Chính vì thế, họ quyết định Julie chuyên trách công việc dự báo cho công ty. Julie quyết định chọn doanh số theo tháng của công ty làm biến dự báo chính. Tuy nhiên, do không am hiểu nhiều về tầm quan trọng của dự báo chuỗi thời gian, các thành viên của công ty trước đây chỉ dựa vào kinh nghiệm để dự đoán và ra quyết định, nên công công ty thiếu quá nhiều dữ liệu doanh số quá khứ. Julie đề nghị phòng kinh doanh thu thập và xây dựng cơ sở dữ liệu về doanh số hàng tháng để phục vụ cho những năm sau này. Trước mắt, Julie quyết định chọn dữ liệu doanh số toàn quốc làm biến dự báo vì cô cho rằng doanh số của công ty Murphy Brothers có quan hệ chặt chẽ với doanh số toàn quốc. Sau khi thu thập dữ liệu doanh số toàn quốc (tập tin **"MURPHY.xls"**), Julie khảo sát đồ thị và hệ số tự tương quan. Julie nhận thấy dữ liệu có yếu tố xu thế, và quyết định lấy sai phân bậc 1 để lựa chọn mô hình dự báo tốt.

a. Anh/Chị cho biết việc Julie chọn biết doanh số toàn quốc làm biến dự báo có phù hợp không? Tại sao?

b. Anh/Chị dự đoán xem Julie sẽ kết luận như thế nào về dữ liệu doanh số bán lẻ toàn quốc?

c. Anh/Chị cho biết quy trình của Julie có phải là một quy trình đúng cho việc xác định một mô hình dự báo thích hợp hay không?

d. Anh/Chị dự đoán xem Julie sẽ chọn các phương pháp dự báo gì?

15. Ông Glen, cha của Julie, có vẻ không hài lòng với cách dự báo của Julie, nên yêu cầu phòng kinh doanh lục lại các chứng từ trước đây của công ty từ năm 1992 đến 1995 để cung cấp dữ liệu doanh số thực tế của Murphy Brothers cho Julie (cùng tập tin **"MURPHY.xls"**). Julie biết rằng thu thập doanh số thực tế chỉ trong vòng bốn năm qua thường dẫn đến xu hướng thay đổi cùng chiều. Cho nên, cô không chắc chắn lắm với bộ dữ liệu doanh số của công ty.

a. Anh/Chị cho biết Julie sẽ kết luận như thế nào về dữ liệu của công ty Murphy Brothers?

b. Anh/Chị cho biết dạng dữ liệu doanh số của công ty Murphy Brothers có khác gì so với dạng dữ liệu doanh số toàn quốc?

c. Anh/Chị cho biết Julie sẽ sử dụng bộ dữ liệu nào để xây dựng mô hình dự báo cho công ty?

CHƯƠNG 4 CÁC MÔ HÌNH DỰ BÁO GIẢN ĐƠN

Trên cơ sở khảo sát dữ liệu và lựa chọn mô hình dự báo đã được đề cập ở chương 3, bây giờ chúng ta sẽ bắt đầu tìm hiểu từng mô hình dự báo cụ thể. Và chương này sẽ tập trung phân tích nhóm các mô hình dự báo giản đơn nhất vốn đã trở nên phổ biến trong tất cả các phần mềm phân tích dữ liệu như Excel, Eviews, Crystal Ball, v.v... Mặc dù, với sự phát triển vượt bậc của máy tính và kinh tế lượng đã giới thiệu nhiều mô hình dự báo phức tạp và có độ chính xác cao, nhưng các mô hình dự báo giản đơn, trong một chừng mực nào đó, vẫn luôn là một sự lựa chọn hữu ích đối với nhiều tổ chức và cá nhân vì chúng có các đặc điểm như sau. Thứ nhất, các doanh nghiệp mới được thành lập nên chưa sẵn có nhiều dữ liệu quá khứ. Thứ hai, các doanh nghiệp phải cùng lúc đương đầu với việc dự báo rất nhiều vấn đề khác nhau như doanh số, tồn kho, mua sắm, v.v..., của rất nhiều sản phẩm và dịch vụ khác nhau. Thứ ba, không phải bất kỳ tổ chức nào cũng có sẵn các chuyên gia phân tích dữ liệu chuyên nghiệp có đủ khả năng thực hiện dự báo bằng các mô hình phức tạp.

Chương này sẽ trình bày ba nhóm phương pháp dự báo chuỗi thời gian giản đơn: các phương pháp dự báo thô, các phương pháp trung bình, và các phương pháp san mũ. Các phương pháp dự báo thô được sử dụng để phát triển các mô hình giản đơn trong đó giả định rằng các dữ liệu gần nhất là các dự đoán tốt nhất cho tương lai. Các phương pháp trung bình đưa ra các dự báo dựa trên giá trị trung bình của các quan sát quá khứ với trọng số như nhau. Các phương pháp san mũ đưa ra các dự báo dựa trên giá trị trung bình có trọng số của các quan sát quá khứ với điều kiện là các trọng số có xu hướng giảm dần.

Phương pháp luận chung nhất cho tất cả các phương pháp dự báo sẽ được trình bày trong chương này đã được giới thiệu ở chương 1 như sau: (1) Chia bộ dữ liệu quá khứ làm hai giai đoạn: giai đoạn dữ liệu mẫu và giai đoạn dự báo hậu nghiệm (ước lượng trong giai đoạn quá khứ và hiện tại) để kiểm chứng kết quả dự báo của từng mô hình, (2) Thực hiện các mô hình dự báo cho giai đoạn dự báo mẫu, (3) Đánh giá kết quả dự báo hậu nghiệm bằng việc phân tích đồ thị và các tiêu chí thống kê, (4) Dự báo tiền nghiệm (dự báo cho các giai đoạn tương lai) đối với mô hình tốt nhất trong giai đoạn dự báo hậu nghiệm.

MỤC TIÊU HỌC TẬP

Sau khi học xong chương này, chúng ta kỳ vọng sẽ hiểu và thực hiện được các kỹ thuật dự báo sau đây:

- Mô hình dự báo thô giản đơn.
- Mô hình dự báo thô điều chỉnh.
- Mô hình dự báo trung bình giản đơn.
- Mô hình dự báo trung bình di động.
- Mô hình dự báo trung bình di động kép.
- Mô hình dự báo san mũ giản đơn.
- Mô hình dự báo Holt.
- Mô hình dự báo Winters.
- Quy trình chuẩn thực hiện dự báo bằng các mô hình dự báo giản đơn trên Crystal Ball.

CÁC MÔ HÌNH DỰ BÁO THÔ

MÔ HÌNH DỰ BÁO THÔ GIẢN ĐƠN

Nhiều doanh nghiệp mới thành lập thường đối diện với một vấn đề hết sức khó khăn trong việc dự báo và lập kế hoạch kinh doanh do có quá

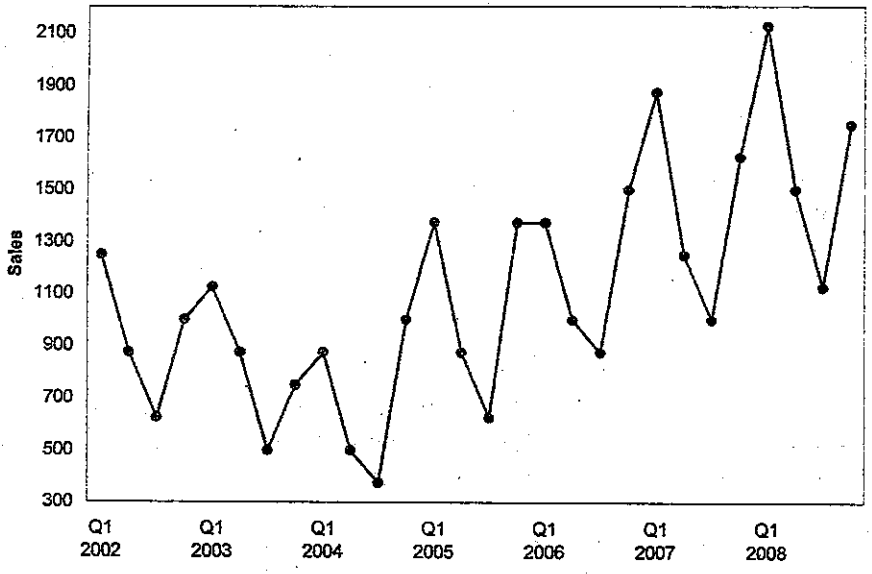
ít dữ liệu quá khứ. Điều này trở thành một vấn đề khó khăn thực sự bởi vì nhiều kỹ thuật dự báo đòi hỏi một lượng dữ liệu quá khứ đủ lớn để đảm bảo tính chính xác trong việc nhận dạng và lựa chọn mô hình dự báo thích hợp. Tuy nhiên, đối với các doanh nghiệp mới thành lập thì các mô hình dự báo thô có thể là một giải pháp khả dĩ nhất vì chúng chỉ dựa trên các thông tin sẵn có gần nhất. Các mô hình dự báo thô giả định rằng các giai đoạn gần nhất là các ước lượng tốt nhất cho tương lai. Mô hình dự báo thô giản đơn nhất có thể được biểu diễn như sau:

$$\hat{Y}_{t+1} = Y_t \qquad (4.1)$$

Trong đó, $\hat{Y}_{t+1}$ là giá trị dự báo ở giai đoạn $t+1$ trên cơ sở giá trị thực tế ở giai đoạn t. Giá trị dự báo thô giản đơn của mỗi giai đoạn đơn giản chỉ là giá trị của quan sát của giai đoạn ngay trước đó. Như vậy, 100% trọng số được gán cho giá trị hiện tại của dữ liệu (Y_t) khi dự báo cho giai đoạn $t+1$. Thực vậy, rất nhiều người trong chúng ta hiện đang áp dụng phương pháp dự báo thô giản đơn trong công việc kinh doanh hàng này của mình nhưng lại không nghĩ đó là một phương pháp dự báo được đề cập trong các sách dự báo. Chẳng hạn, một quầy báo dự kiến sẽ lấy bao nhiêu tờ Tuổi Trẻ vào ngày mai có thể đã cân nhắc trong ngày hôm nay đã tiêu thụ hết bao nhiêu.

Ví dụ, Bảng 4.1 (**DATA4-1**) trình bày dữ liệu về doanh số theo quý của công ty ABC giai đoạn 2002-2008. Dữ liệu này được thể hiện trong Hình 4.1. Nếu sử dụng phương pháp dự báo thô giản đơn, thì giá trị dự báo của quý I năm 2009 sẽ là 1.750 triệu đồng. Kết quả dự báo được thể hiện ở Hình 4.2 và Bảng 4.2.

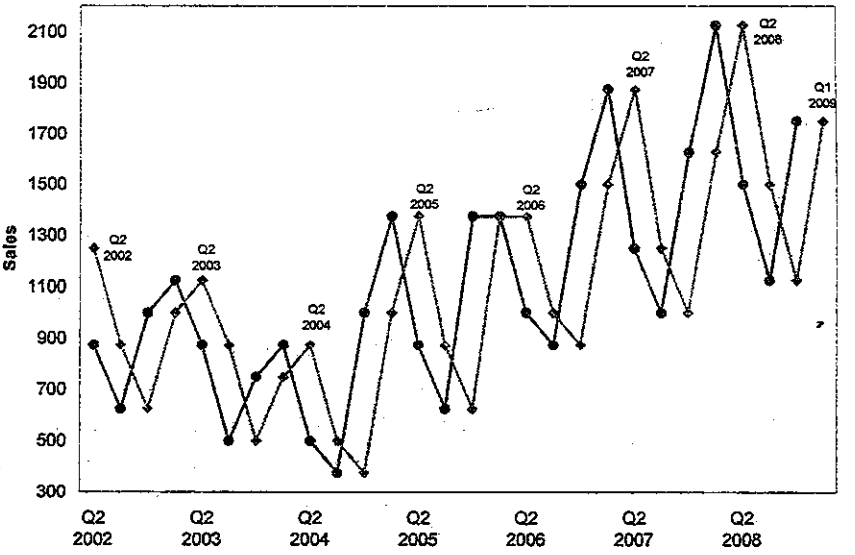
■ HÌNH 4.1: **Doanh số theo quý của công ty ABC (triệu đồng).**



■ BẢNG 4.1: **Doanh số theo quý của công ty ABC (triệu đồng).**

Năm	Quý	t	Y_t	Năm	Quý	t	Y_t
2002	1	1	1250		3	15	625
	2	2	875		4	16	1375
	3	3	625	2006	1	17	1375
	4	4	1000		2	18	1000
2003	1	5	1125		3	19	875
	2	6	875		4	20	1500
	3	7	500	2007	1	21	1875
	4	8	750		2	22	1250
2004	1	9	875		3	23	1000
	2	10	500		4	24	1625
	3	11	375	2008	1	25	2125
	4	12	1000		2	26	1500
2005	1	13	1375		3	27	1125
	2	14	875		4	28	1750

■ HÌNH 4.2: **Doanh số dự báo của công ty ABC theo công thức (4.1).**



■ BẢNG 4.2: **Đánh giá độ chính xác của mô hình dự báo thô.**

T	Y_t	$\hat{Y}_t$	e_t	$\|e_t\|$	e_t^2	$\|e_t\|/Y_t$	e_t/Y_t
1	1.250	-	-	-	-	-	-
2	875	1.250	375	375	140.625	0.429	0.429
3	625	875	250	250	62.500	0.400	0.400
4	1.000	625	-375	375	140.625	0.375	-0.375
5	1.125	1.000	-125	125	15.625	0.111	-0.111
.	.	.	.	.	.	.	.
27	1.125	1.500	375	375	140.625	0.333	0.333
28	1.570	1.125	-625	625	390.625	0.357	-0.357
	Tổng		-500	10.250	4.906.250	9.94	1.94

Trong đó:

Y_t = doanh số của quý t

$\hat{Y}_t$ = doanh số dự báo của quý t

$e_t = Y_t - \hat{Y}_t$

Các thước đo độ chính xác của mô hình dự báo trên được tính toán như sau:

MAE	MAPE	MPE	MSE	RMSE	U
379.6	0.368	0.072	181.713	426	1

MÔ HÌNH DỰ BÁO THÔ ĐIỀU CHỈNH

Hình 4.1 cho thấy chuỗi dữ liệu Y_t vừa có yếu tố xu thế và vừa có yếu tố mùa vì các quý có những cách biệt khá lớn và lặp lại một mức độ giống nhau cho cùng một quý giữa các năm. Trong trường hợp như vậy thì mô hình dự báo thô giản đơn có thể điều chỉnh yếu tố mùa và yếu tố xu thế theo các cách sau đây:

Điều chỉnh xu thế

Cổng thức (4.1) có thể được điều chỉnh yếu tố xu thế như sau:

$$\hat{Y}_{t+1} = Y_t + (Y_t - Y_{t-1}) \qquad (4.2)$$

Hoặc

$$\hat{Y}_{t+1} = Y_t \frac{Y_t}{Y_{t-1}} \qquad (4.3)$$

Ví dụ, doanh số của quý I năm 2009 sẽ được dự báo như sau:

$$\hat{Y}_{28+1} = Y_{28} + (Y_{28} - Y_{27})$$

$$\hat{Y}_{29} = Y_{28} + (Y_{28} - Y_{27})$$

$$\hat{Y}_{29} = 1.570 + (1.570 - 1.125)$$

$$\hat{Y}_{29} = 1.570 + 625 = 2.375$$

Để xem mô hình dự báo thô điều chỉnh xu thế nào có tốt hơn mô hình dự báo thô giản đơn hay không, chúng ta lập bảng đánh giá như Hình 4.3.

■ BẢNG 4.3: **Đánh giá độ chính xác của mô hình dự báo thô.**

T	Y_t	$\hat{Y}_t$	e_t	$\|e_t\|$	e_t^2	$\|e_t\|/Y_t$	e_t/Y_t
1	1.250	-	-	-	-	-	-
2	875	-	-	-	-	-	-
3	625	500	-125	125	15.625	0.200	-0.200
4	1.000	375	-625	625	390.625	0.625	-0.625
5	1.125	1.375	250	250	62.500	0.222	0.222
.	.	.	.	.	.	.	.
27	1.125	875	-250	250	62.599	0.222	-0.222
28	1.570	750	-1.000	1.000	1.000.000	0.571	-0.571
Tổng			-1.000	13.250	9.437.500	12.6	-0.81

Các thước đo độ chính xác của mô hình dự báo trên được tính toán như sau:

MAE	MAPE	MPE	MSE	RMSE	U
509.6	0.484	-0.031	362.981	602	1.41

Như vậy, mô hình dự báo thô điều chỉnh xu thế như trên không tốt bằng mô hình dự báo thô giản đơn. Điều này hoàn toàn hợp lý vì chúng ta đã bỏ qua yếu tố mùa vụ trong dữ liệu, mà đây lại là yếu tố khá nổi trội trong bộ dữ liệu ví dụ.

Điều chỉnh mùa vụ

Đối với dữ liệu theo quý, thì mô hình dự báo thô có thể được điều chỉnh như sau:

$$\hat{Y}_{t+1} = Y_{t-3} \qquad (4.4)$$

Điều này có nghĩa là: khi dự báo có yếu tố mùa thì giá trị dự báo cho một quý nào đó (mùa nào đó) của năm sau chính là giá trị thực tế của chúng của chính quý đó vào năm trước.

Ví dụ, doanh số của quý I năm 2009 sẽ được dự báo như sau:

$$\hat{Y}_{28+1} = Y_{28-3}$$
$$\hat{Y}_{29} = Y_{25}$$
$$\hat{Y}_{29} = 2.125$$

MAE	MAPE	MPE	MSE	RMSE	U
218.8	0.225	-0.053	65.104	255.2	0.599

Điều chỉnh yếu tố mùa đã cải thiện rất nhiều sự chính xác của mô hình dự báo thô này. Tuy nhiên, phương pháp này có hạn chế là đã bỏ qua yếu tố xu thế giữa các các quý trong năm qua. Chính vì thế, đối với dữ liệu vừa có yếu tố xu thế, vừa có yếu tố quý, chúng ta có thể điều chỉnh như sau:

$$\hat{Y}_{t+1} = Y_{t-3} + \frac{(Y_t - Y_{t-1}) + ... + (Y_{t-3} - Y_{t-4})}{4} = Y_{t-3} + \frac{(Y_t - Y_{t-4})}{4} \quad (4.5)$$

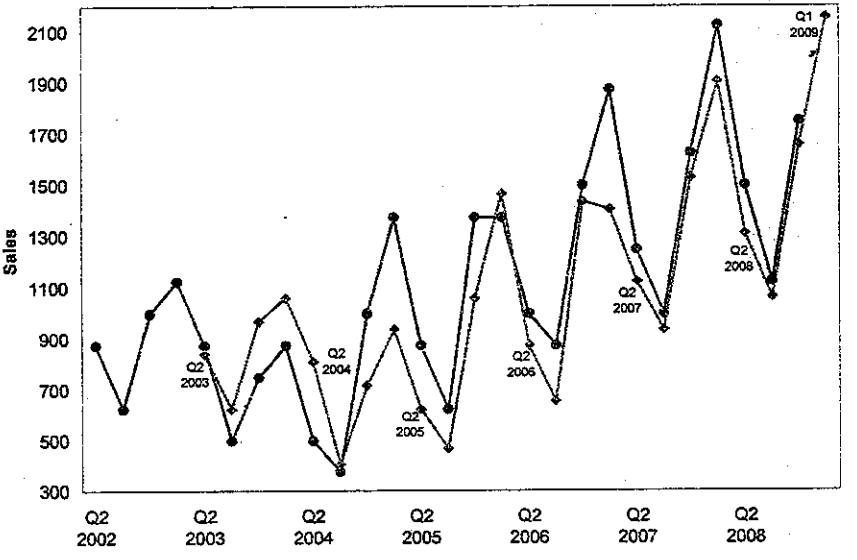
Ví dụ, doanh số của quý I năm 2009 sẽ được dự báo như sau:

$$\hat{Y}_{28+1} = Y_{28-3} + \frac{(Y_{28} - Y_{28-1}) + ... + (Y_{28-3} - Y_{28-4})}{4}$$
$$\hat{Y}_{29} = Y_{28-3} + \frac{(Y_{28} - Y_{28-4})}{4}$$
$$\hat{Y}_{29} = Y_{25} + \frac{(Y_{28} - Y_{24})}{4} = 2.125 + \frac{(1.750 - 1.625)}{4}$$
$$\hat{Y}_{29} = Y_{25} + \frac{(Y_{28} - Y_{24})}{4} = 2.125 + 31 = 2.156$$

MAE	MAPE	MPE	MSE	RMSE	U
180	0.181	-0.047	46.663	216	0.507

Như vậy, phương pháp dự báo thô có điều chỉnh xu thế và mùa vụ là mô hình dự báo tốt nhất trong tất cả các mô hình dự báo thô.

■ HÌNH 4.2: **Doanh số dự báo của công ty ABC theo công thức (4.5).**



CÁC PHƯƠNG PHÁP DỰ BÁO TRUNG BÌNH

Hanke (2005) cho rằng thông thường người làm dự báo gặp phải trường hợp trong đó các dự báo cần phải được cập nhật hàng ngày, hàng tuần, hay hàng tháng đối với hàng trăm hoặc hàng ngàn hạng mục như hàng tồn kho, doanh số. Thường thì rất khó để có thể phát triển các kỹ thuật dự báo phức tạp cho từng hạng mục. Chính vì vậy, các phương pháp dự báo giản đơn, ít tốn kém, và nhanh có thể sẽ hữu

hiệu cho mục đích dự báo ngắn hạn này. Tại sao các phương pháp dự báo giản đơn như trung bình di động và san mũ lại có ích trong các trường hợp như vậy? Vì đây là những phương pháp dự báo không đòi hỏi nhiều về các kiến thức thống kê, kinh tế lượng, và đã được lập trình hóa trên tất cả các phần mềm phân tích dữ liệu như Excel, Crystal Ball, hay Eviews. Các phương pháp dự báo này đều sử dụng hình thức bình quân hoặc bình quân gia quyền của các quan sát quá khứ để "nhẵn" hoặc "san" (smooth) các dao động ngắn hạn của dữ "iệu. Cũng theo Hanke (2005), thì giả định cơ bản của các phương pháp này cho rằng các dao động trong các dữ liệu quá khứ chỉ thể hiện tính ngẫu nhiên xoay quanh một cấu trúc ổn định. Một khi cấu trúc dữ liệu quá khứ được nhận diện, thì việc dự báo tương lai trở nên dễ dàng.

TRUNG BÌNH GIẢN ĐƠN

Mô hình dự báo trung bình giản đơn có thể được biểu hiện qua công thức đơn giản sau đây:

$$\hat{Y}_{t+1} = \frac{1}{t}\sum_{t=1}^{t} Y_t \qquad (4.6)$$

Trong đó, t có thể là quan sát cuối cùng trong mẫu hoặc toàn bộ mẫu dữ liệu quá khứ sẵn có. Khi một quan sát mới được đưa thêm vào, thì giá trị dự báo cho giai đoạn tiếp theo, $\hat{Y}_{t+2}$, chỉ đơn giản là trung bình của $\hat{Y}_{t+1}$ và quan sát mới thêm vào. Cho nên, khi cập nhật thông tin, thì công thức (4.6) sẽ được điều chỉnh như sau:

$$\hat{Y}_{t+1} = \frac{t\hat{Y}_{t+1} + Y_{t+1}}{t+1} \qquad (4.7)$$

Phương pháp dự báo trung bình giản đơn chỉ phù hợp đối với các chuỗi dữ liệu không có biến động lớn, và thuật ngữ chuỗi thời gian gọi là có tính 'dừng'. Điều này có nghĩa rằng, các yếu tố và môi trường kinh doanh ảnh hưởng lên đối tượng dự báo có tính ổn định. Ví dụ, khối lượng doanh số của một nhãn hiệu thời trang được kỳ vọng sẽ ổn định nếu các yếu tố như nỗ lực của lực lượng bán hàng, số cửa hàng,

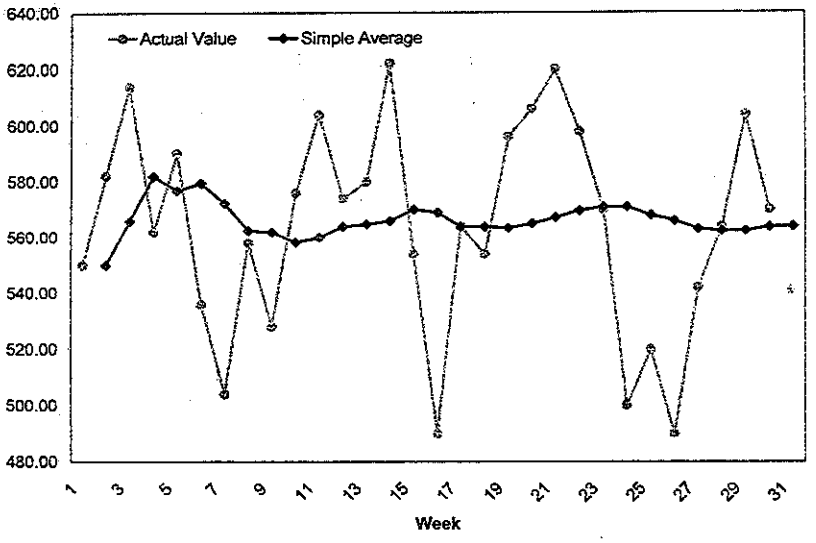
áp lực của các đối thủ cạnh tranh, v.v..., không có nhiều thay đổi hoặc đang trong trạng thái rất ổn định. Ngoài ra, nhãn hiệu thời trang này phải đang ở giai đoạn ổn định và đi tới bão hòa trong vòng đời sản phẩm.

Ví dụ, Bảng 4.4 (**DATA4-2**) là dữ liệu số lượng quần Jean tiêu thụ/tuần của một của hàng của nhãn hiệu thời trang Nino-Max trong 30 tuần qua.

■ BẢNG 4.4: **Số quần Jean bán/tuần tại một cửa hàng của Nino-Max.**

Tuần	Y_t	Tuần	Y_t	Tuần	Y_t
1	550	11	604	21	620
2	582	12	574	22	598
3	614	13	580	23	570
4	562	14	622	24	500
5	590	15	554	25	520
6	536	16	490	26	490
7	504	17	564	27	542
8	558	18	554	28	564
9	528	19	596	29	604
10	576	20	606	30	570

■ HÌNH 4.3: **Dự báo số quần Jean bán/tuần theo trung bình giản đơn.**



Để ước tính giá trị dự báo theo phương pháp trung bình giản đơn trên Excel (để vẽ đồ thị như Hình 4.3), ta thực hiện như sau:

Bước 1: Tạo giá trị dự báo đầu tiên ($\hat{Y}_2$) bằng ô B2 (550).

Bước 2: Tạo giá trị dự báo thứ ba ($\hat{Y}_3$) theo công thức (4.7) như sau:

	A	B	C
1	**Week**	**Actual**	Simple Average
2	1	550.00	
3	2	582.00	550.00
4	3		=(A2*C3+B3)/A3
5	4	562.00	582.00

Bước 3: Copy công thức của ô C4 cho các ô tiếp theo của cột C.

Như vậy, giá trị dự báo cho giai đoạn tiếp theo ($\hat{Y}_{31}$) sẽ được tính như sau:

$$\hat{Y}_{31} = \frac{1}{30}\sum_{t=1}^{30} Y_t = \frac{29 * Y_{30} + Y_{30}}{30} = 564.07$$

Để làm cơ sở so sánh với các mô hình khác, ta ước tính các sai số dự báo của mô hình trung bình giản đơn như sau:

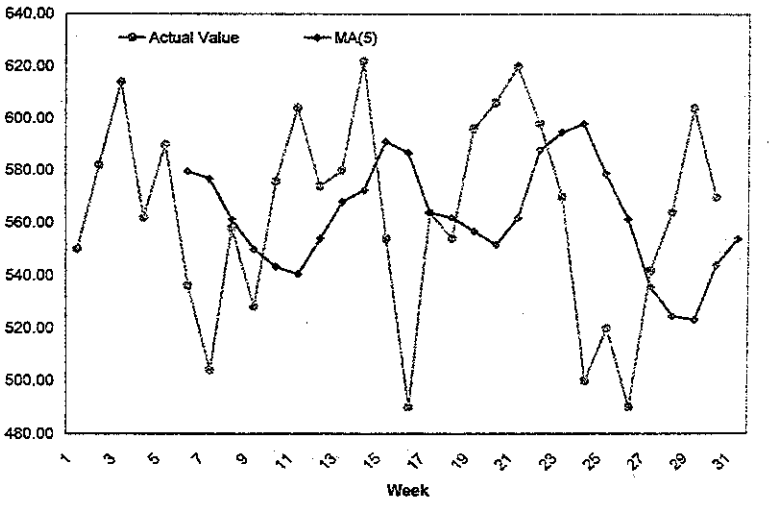
MAE	MAPE	MPE	MSE	RMSE	U
32.15	0.058	-0.008	1572.7	39.66	0.97

TRUNG BÌNH DI ĐỘNG

Phương pháp trung bình giản đơn sử dụng giá trị trung bình của toàn bộ dữ liệu quá khứ làm giá trị dự báo. Ngược lại, phương pháp trung bình di động chỉ sử dụng một số quan sát gần nhất làm giá trị dự báo. Phương pháp trung bình di động cũng thích hợp đối với các chuỗi dừng. Với hệ số số trượt k, trung bình di động bậc k, ký hiệu là MA(k) được thể hiện theo công thức sau đây:

$$\hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + ... + Y_{t-k+1}}{k} \qquad (4.8)$$

■ HÌNH 4.4: **Dự báo số quần Jean bán/tuần theo MA(5).**



Như vậy, trung bình di động cho giai đoạn *t* là giá trị trung bình số học của *k* quan sát gần nhất. Trong một giá trị trung bình di động, thì trọng số của mỗi quan sát đều bằng nhau và bằng *1/k*.

Sử dụng dữ liệu trong Bảng 4.4 và giả sử k = 5, thì giá trị dự báo cho giai đoạn tiếp theo ($\hat{Y}_{31}$) sẽ được thực hiện như sau (sử dụng hàm AVERAGE):

$$\hat{Y}_{30+1} = \frac{Y_{30} + Y_{30-1} + ... + Y_{30-5+1}}{5}$$

$$\hat{Y}_{31} = \frac{Y_{30} + Y_{29} + Y_{28} + Y_{27} + Y_{26}}{5}$$

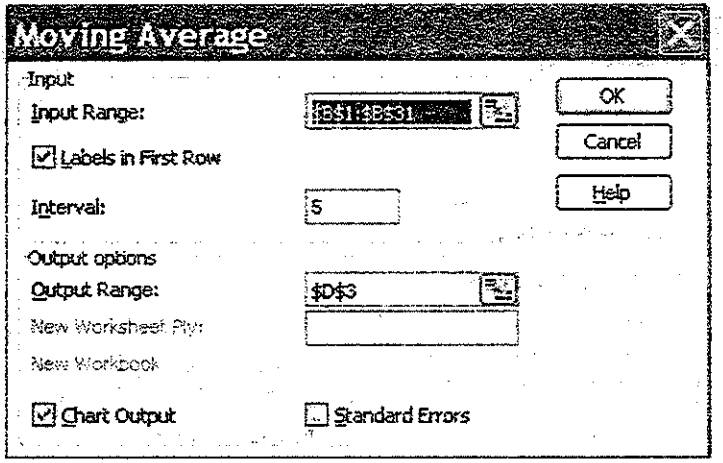
$$\hat{Y}_{31} = \frac{570 + 604 + 564 + 542 + 490}{5} = 554$$

MAE	MAPE	MPE	MSE	RMSE	U
45.7	0.083	-0.004	2644.5	51.4	1.25

Như vậy, quan sát đồ thị Hình 4.4 có vẻ phương pháp trung bình di động cho kết quả dự báo tốt hơn so với phương pháp trung bình giản đơn. Tuy nhiên, các tiêu chí thống kê lại cho thấy phương pháp trung bình giản đơn cho kết quả dự báo tốt hơn phương pháp trung bình di động.

Quy trình thực hiện trên Excel:

Bước 1: Tools/Data Analysis/Moving Average



Bước 2: Nhập dữ liệu vào ô 'Input Range' (nếu bao gồm cả ô tiêu đề thì chọn 'Labels in First Row').

Bước 3: Nhập hệ số trượt k vào ô 'Interval'.

Bước 4: Xác định ô đặt giá trị dự báo vào ô 'Output Range', rồi chọn OK.

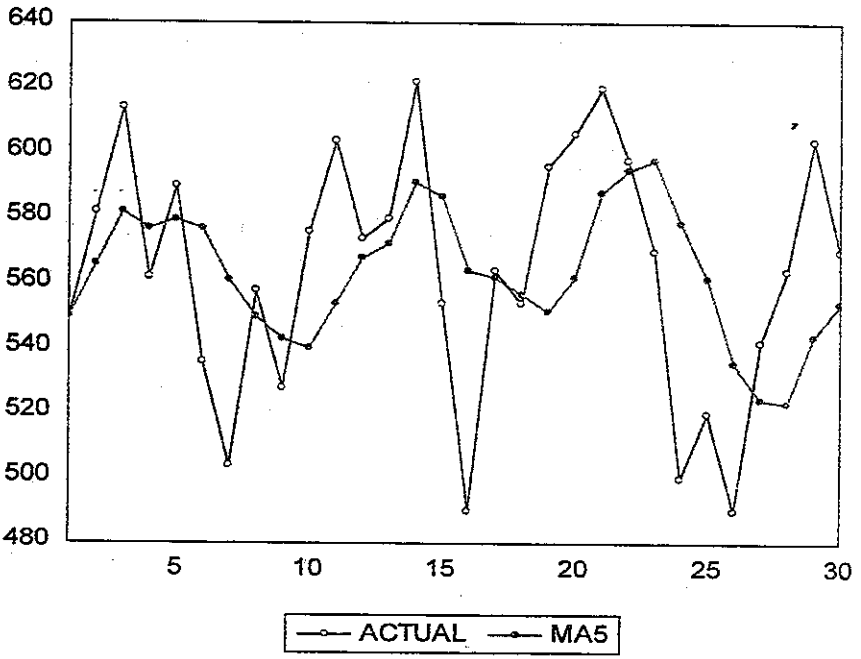
Lưu ý: Excel chỉ cung cấp giá trị dự báo và đồ thị. Cho nên, chúng ta phải tự lập bảng tính các tiêu chí đánh giá độ chính xác của dự báo. Tuy nhiên, đối với phương pháp trung bình di động, ta nên sử dụng hàm AVERAGE.

Quy trình thực hiện trên Eviews:

Bước 1: Mở và chuyển tập tin "**DATA4-2.xls**" qua Eviews

Lưu ý, đối với phương pháp trung bình di động, chúng ta chưa nhất thiết phải thực hiện trên Eviews, vì các thao tác tính toán sai số dự báo trên Eviews vốn rất tốn thời gian.

■ HÌNH 4.5: **Dự báo số quần Jean bán/tuần theo MA(5) trên Eviews.**

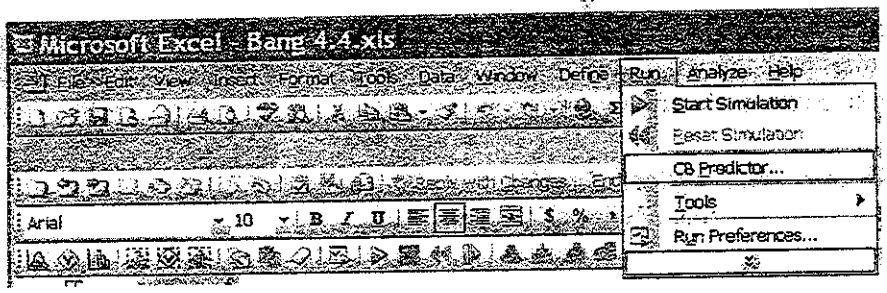


Quy trình thực hiện trên Crystal Ball:

Bước 1: Khởi động Crystal Ball (7.2 hoặc 7.3)

Bước 2: Mở tập tin "**DATA4-2.xls**"

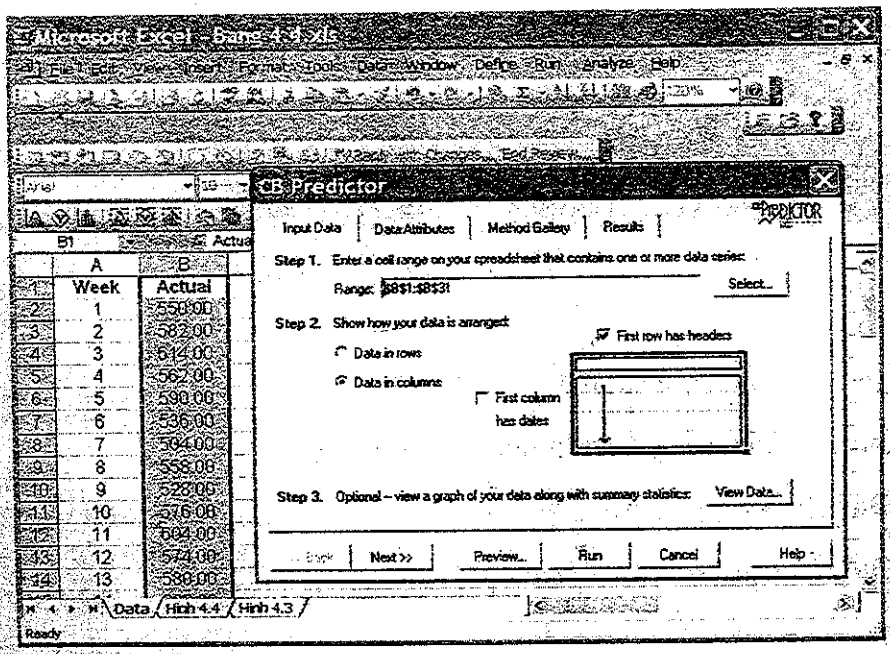
Bước 3: Vào "Run/CB Predictor..."



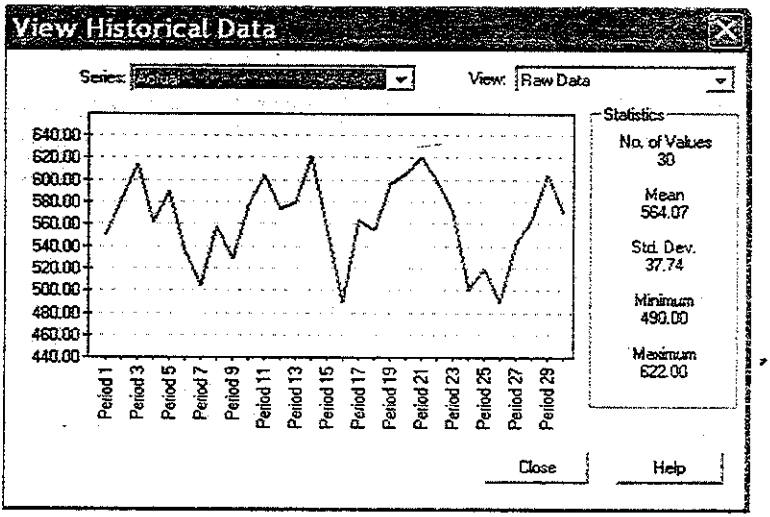
Bước 4: Nhập dữ liệu vào "Input Data" như ở Hình 4.6

- Trong **Step 1**, nhập B1:B31 vào ô "Range"
- Trong **Step 2**, chọn "First row has headers" và "Data in columns"
- Trong **Step 3** (tùy chọn), nhấp vào "View Data" để xem đồ thị của dữ liệu quá khứ (xem Hình 4.7), và chọn "Next"

■ HÌNH 4.6: **Thực hiện phương pháp MA trên Crystal Ball.**



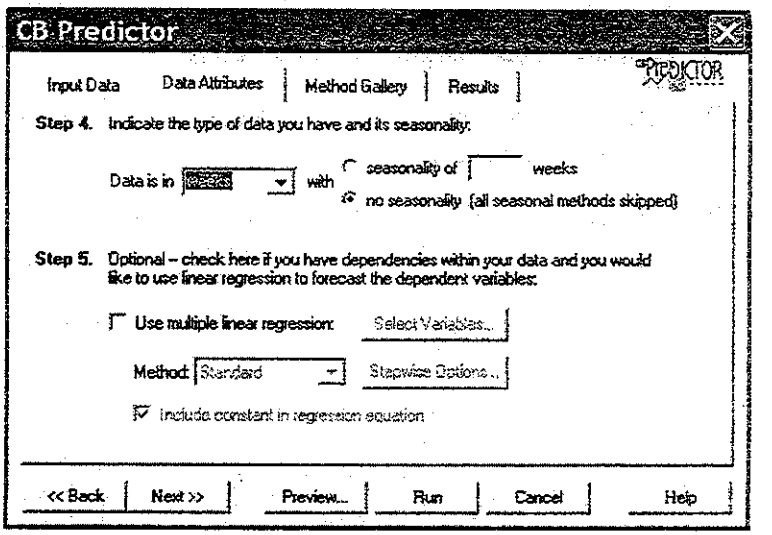
■ HÌNH 4.7: **Xem dữ liệu quá khứ trên Crystal Ball.**



Bước 5: Khai báo đặc điểm dữ liệu vào "Data Attributes"

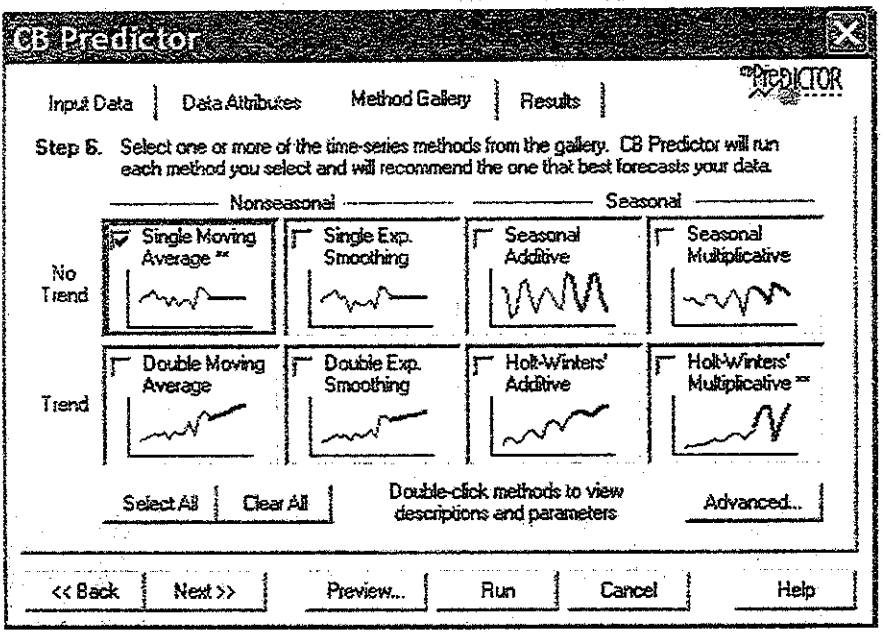
- Trong **Step 4**, ta chọn "weeks" và "no seasonality"
- Bỏ qua **Step 5** và chọn "Next"

■ HÌNH 4.8: **Khai báo đặc điểm dữ liệu trên Crystal Ball.**



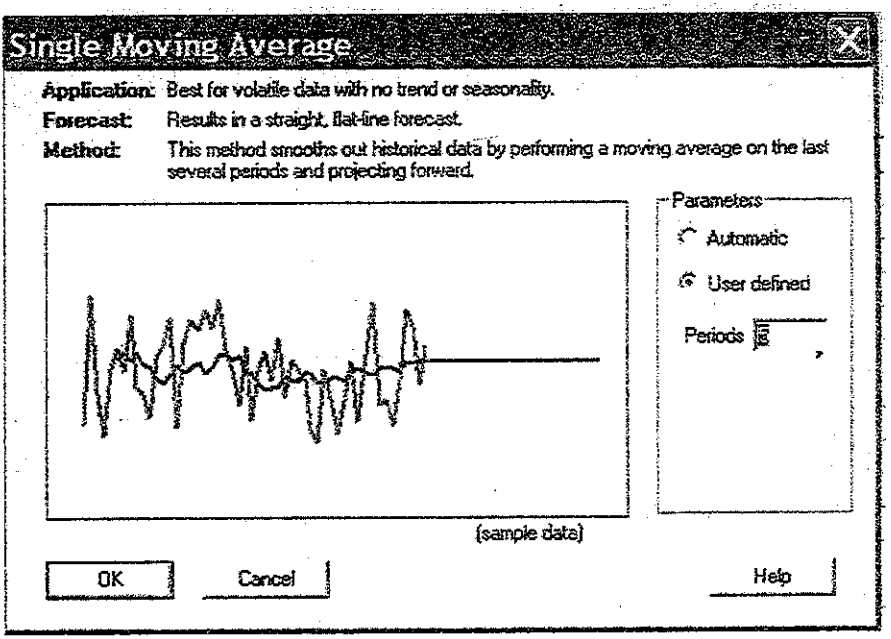
Bước 6: Chọn phương pháp dự báo trong "Method Gallery"

■ HÌNH 4.9: **Lựa chọn phương pháp dự báo trên Crystal Ball.**



- Trong **Step 6**, chọn "Single Moving Average"
- Nếu nhấp đúp vào "Single Moving Average" ta sẽ thấy xuất hiện Hình 4.10 và nhập 5 vào ô "Periods", OK.
- Chọn "Next".

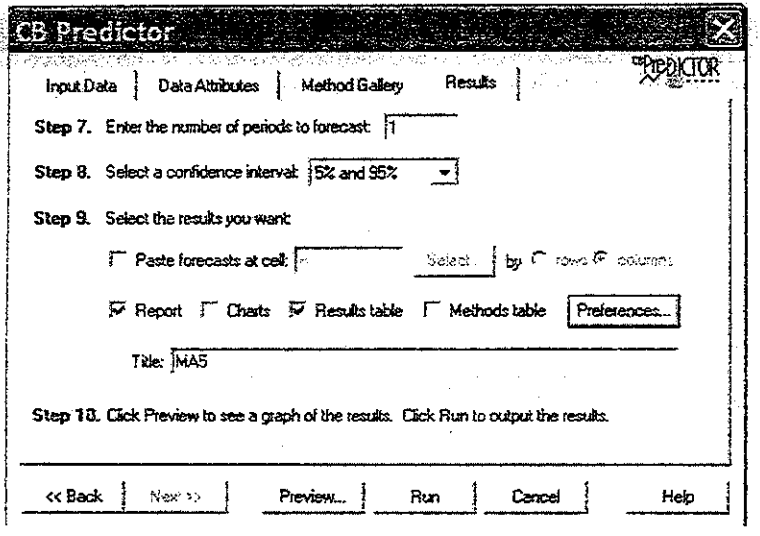
■ HÌNH 4.10: **Xác định hệ số trượt k trên Crystal Ball.**



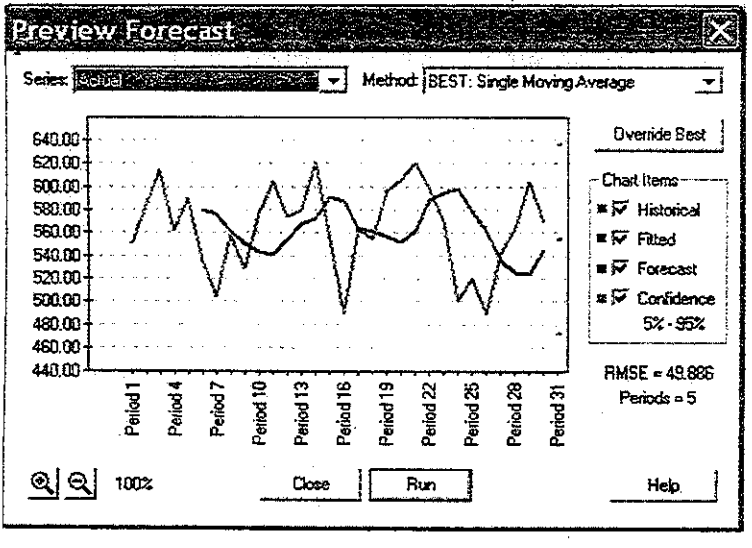
Bước 7: Khai báo kết quả dự báo vào "Results"

- Trong **Step 7**, ta nhập số giai đoạn cần dự báo (Hình 4.11).
- Trong **Step 8**, ta xác định khoảng tin cậy cho dự báo (mặc định là 5% và 95%).
- Trong **Step 9**, ta xác định vị trí đặt kết quả dự báo và các loại kết quả ta cần có báo cáo.
- Trong **Step 10**, ta có thể chọn "Preview" để xem đồ thị dự báo (Hình 4.12) hoặc chọn "Run" để có các kết quả dự báo.

■ HÌNH 4.11: **Xác định kết quả dự báo trên Crystal Ball.**



■ HÌNH 4.12: **Xem đồ thị dự báo trên Crystal Ball.**



■ BẢNG 4.5: **Kết quả dự báo MA(5) trên Crystal Ball.**

	A	B	C	D	E	F
1	**Results Table for MA5**					
2	Created: 5/2/2009 at 3:13:52 PM					
3						
36	Series	Actual				
37						
38		Data				
39	Date	Historical Data	Lower: 5%	Fit & Forecast	Upper: 95%	Residuals
40	Period 1	550.00				550.00
41	Period 2	582.00				582.00
42	Period 3	614.00				614.00
43	Period 4	562.00				562.00
44	Period 5	590.00				590.00
45	Period 6	536.00		579.60		-43.60
46	Period 7	504.00		576.80		-72.80
47	Period 8	558.00		561.20		-3.20
48	Period 9	528.00		550.00		-22.00
66	Period 27	542.00		535.60		6.40
67	Period 28	564.00		524.40		39.60
68	Period 29	604.00		523.20		80.80
69	Period 30	570.00		544.00		26.00
70	Period 31		471.94	554.00	636.06	

Tóm lại, phương pháp bình quân di động thường được sử dụng với các dữ liệu theo quý hoặc tháng nhằm mục đích làm trơn các thành phần trong một chuỗi thời gian. Và điều này rất cần thiết cho việc thực hiện dự báo theo phương pháp phân tích mà chúng ta sẽ phải tiến hành ở chương sau nữa của cuốn sách. Đối với dữ liệu theo quý, chúng ta thường sử dụng hệ số trượt bằng 4, MA(4), để tạo ra giá trị trung bình của bốn quý. Tương tự, đối với dữ liệu theo tháng, chúng ta thường sử dụng hệ số trượt bằng 12, MA(12), để tạo ra giá trị trung bình của mười hai tháng. Mục đích của việc làm trơn dữ liệu này là nhằm loại bỏ các ảnh hưởng mùa vụ hoặc ảnh hưởng ngẫu nhiên của dữ liệu.

TRUNG BÌNH DI ĐỘNG KÉP

Phương pháp bình quân di động kép nhằm sử dụng dự báo dữ liệu chuỗi thời gian có yếu tố xu thế. Bản thân 'tên' của phương pháp này đã có hàm ý như sau:

Bước 1: Tính giá trị bình quân di động cho chuỗi dữ liệu gốc (MA).

Bước 2: Tính giá trị bình quân di động cho chuỗi bình quân di động thứ nhất (MA').

Bảng 4.6 (**DATA4-3**) là dữ liệu về số lượng người thuê đĩa DVD tại một quầy dịch vụ cho thuê băng đĩa. Hình 4.13 cho thấy lượng người thuê đĩa DVD có xu hướng tăng lên qua thời gian (tuần). Chính vì vậy, có thể phương pháp bình quân di động, MA(3), sẽ không phải là mô hình dự báo phù hợp. Trong ví dụ này, chúng ta sẽ so sánh độ chính xác của dự báo giữa hai mô hình MA(3) và mô hình bình quân di động kép, DMA(3).

■ BẢNG 4.6: **So sánh MA(3) và DMA(3).**

t	Y_t	MA(3)		DMA(3)					
		$\hat{Y}_t$	e_t	MA(3)	MA(3)'	a	b	$\hat{Y}_t$	e_t
1	654	-	-	-	-	-	-	-	-
2	658	-	-	-	-	-	-	-	-
3	665	-	-	659	-	-	-	-	-
4	672	659	13	665	-	-	-	-	-
5	673	665	8	670	665	675	5	-	-
6	671	670	1	672	669	675	3	681	-10
7	693	672	21	679	674	684	5	678	15
8	694	679	15	686	679	693	7	690	4
9	701	686	15	696	687	705	9	700	1
10	703	696	7	699	694	705	6	714	-11
11	702	699	3	702	699	705	3	710	-8
12	710	702	8	705	702	708	3	708	2
13	712	705	7	708	705	711	3	711	1
14	711	708	3	711	708	714	3	714	-3
15	728	711	17	717	712	722	5	717	11
16	-	**717**	-	-	-	-	-	**727**	-
MSE		**133**				**63.7**			

Bước 1: Tính giá trị bình quân di động cho chuỗi dữ liệu gốc (MA).

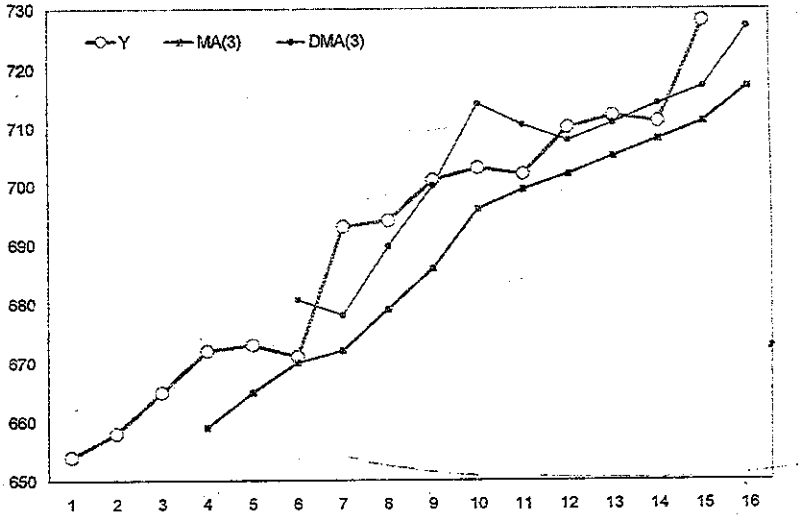
Bước 2: Tính giá trị bình quân di động cho chuỗi bình quân di động thứ nhất (MA').

Bảng 4.6 (**DATA4-3**) là dữ liệu về số lượng người thuê đĩa DVD tại một quầy dịch vụ cho thuê băng đĩa. Hình 4.13 cho thấy lượng người thuê đĩa DVD có xu hướng tăng lên qua thời gian (tuần). Chính vì vậy, có thể phương pháp bình quân di động, MA(3), sẽ không phải là mô hình dự báo phù hợp. Trong ví dụ này, chúng ta sẽ so sánh độ chính xác của dự báo giữa hai mô hình MA(3) và mô hình bình quân di động kép, DMA(3).

■ BẢNG 4.6: **So sánh MA(3) và DMA(3).**

t	Y_t	MA(3)		DMA(3)					
		$\hat{Y}_t$	e_t	MA(3)	MA(3)'	a	b	$\hat{Y}_t$	e_t
1	654	-	-	-	-	-	-	-	-
2	658	-	-	-	-	-	-	-	-
3	665	-	-	659	-	-	-	-	-
4	672	659	13	665	-	-	-	-	-
5	673	665	8	670	665	675	5	-	-
6	671	670	1	672	669	675	3	681	-10
7	693	672	21	679	674	684	5	678	15
8	694	679	15	686	679	693	7	690	4
9	701	686	15	696	687	705	9	700	1
10	703	696	7	699	694	705	6	714	-11
11	702	699	3	702	699	705	3	710	-8
12	710	702	8	705	702	708	3	708	2
13	712	705	7	708	705	711	3	711	1
14	711	708	3	711	708	714	3	714	-3
15	728	711	17	717	712	722	5	717	11
16	-	**717**	-	-	-	-	-	**727**	-
MSE		**133**				**63.7**			

HÌNH 4.13: So sánh MA(3) với DMA(3).



Các bước thực hiện dự báo theo DMA trên Excel:

Bước 1: Tính chuỗi trung bình di động bậc 1, MA(k), theo công thức (4.8):

$$MA(k)_t = \hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + ... + Y_{t-k+1}}{k} \qquad (4.8)$$

Bước 2: Tính chuỗi trung bình di động bậc 2, MA(k)', theo công thức (4.9) như sau:

$$MA(k)'_t = \frac{MA(k)_t + MA(k)_{t-2} + ... + MA(k)_{t-k+1}}{k} \qquad (4.9)$$

Bước 3: Tính chênh lệch giữa $MA(k)_t$ và $MA(k)'_t$ để xác định 'vị trí' trung bình của chuỗi dữ liệu khi có yếu tố xu thế (a_t):

$$a_t = MA(k)_t - [MA(k)_t - MA(k)'_t] = 2MA(k)_t - MA(k)'_t \qquad (4.10)$$

Bước 4: Ước tính hệ số điều chỉnh yếu tố xu thế của dữ liệu, hệ số này được xem như độ dốc (b_t). Lưu ý rằng, hệ số độ dốc này có thể thay đổi theo thời gian.

$$b_t = \frac{2}{k-1}\left[MA(k)_t - MA(k)'_t\right] \quad (4.10)$$

Bước 5: Dự báo cho giai đoạn tiếp theo:

$$\hat{Y}_{t+p} = a_t + b_t P \quad (4.11)$$

Trong đó, *k* là hệ số trượt và *P* là số giai đoạn dự báo. Ở ví dụ trên, ta chọn P = 1, nghĩa là, chỉ dự báo cho một giai đoạn tiếp theo.

Kết quả dự báo cho thấy DMA(3) có sai số dự báo nhỏ hơn so với MA(3) (xem Bảng 4.6), và đồ thị DMA(3) thể hiện rất đúng xu hướng vận động của dữ liệu gốc (xem Hình 4.13). Như vậy, khi dữ liệu có yếu tố xu thế thì DMA là mô hình thích hợp hơn.

Quy trình thực hiện dự báo bằng phương pháp bình quân di động kép trên Crystal Ball giống như quy trình thực hiện dự báo theo phương pháp bình quân di động, nhưng thay vì chọn "Single Moving Average" ở Bước 6, thì ta chọn "Double Moving Average".

CÁC PHƯƠNG PHÁP SAN MŨ

SAN MŨ GIẢN ĐƠN

Trong khi các phương pháp trung bình di động chỉ quan tâm đến các quan sát gần nhất, thì phương pháp san mũ giản đơn lại đưa ra một giá trị trung bình di động với trọng số giảm dần cho tất cả các quan sát trong quá khứ. Mô hình san mũ giản đơn thường phù hợp với loại dữ liệu không thể dự đoán được có xu hướng tăng hay giảm. Mục tiêu của phương pháp này là ước lượng giá trị trung bình hiện tại và sử dụng giá trị này làm giá trị dự báo cho tương lai.

Phương pháp san mũ vẫn dựa trên cơ sở lấy trung bình tất cả các giá trị quá khứ của chuỗi dữ liệu dưới dạng trọng số giảm dần theo hàm mũ. Quan sát gần nhất (với giá trị dự báo) nhận trọng số α (với $0 < \alpha < 1$) lớn nhất, quan sát tiếp theo nhận trọng số nhỏ hơn một chút, $\alpha(1-\alpha)$, quan sát tiếp theo nữa nhận trọng số nhỏ hơn nữa, $\alpha(1-\alpha)^2$, và cứ tiếp diễn như thế cho đến quan sát cuối cùng trong dữ liệu quá khứ.

Cách thể hiện đơn giản nhất của phương pháp này được biểu hiện theo công thức sau đây:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t \qquad (4.12)$$

Trong đó,

$\hat{Y}_{t+1}$ = Giá trị dự báo (mới) ở giai đoạn $t+1$

α = Hệ số san mũ

Y_t = Giá trị quan sát hoặc giá trị thực ở giai đoạn t

$\hat{Y}_t$ = Giá trị dự báo (cũ) ở giai đoạn t

Như vậy, ý tưởng của phương pháp san mũ giản đơn cho rằng giá trị dự báo mới là một giá trị trung bình có trọng số giữa giá trị thực tế và giá trị dự báo ở giai đoạn t. Một khi đã có hệ số san mũ và giá trị dự báo trước đó thì việc ước lượng giá trị dự báo mới trở nên hết sức dễ dàng.

Công thức (4.12) có thể được triển khai theo cách sau đây:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t = \alpha Y_t + \hat{Y}_t - \alpha\hat{Y}_t$$

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t)$$

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha e_t \qquad (4.13)$$

Theo công thức (4.13), thì giá trị dự báo mới bằng giá trị dự báo cũ được điều chỉnh theo sai số dự báo cũ (αe_t).

Giá trị hệ số san mũ α đóng vai trò như một yếu tố xác định mức độ ảnh hưởng của quan sát hiện tại lên giá trị dự báo của quan sát tiếp theo. Khi α gần bằng 1, thì giá trị dự báo sẽ hầu như chính là giá trị

của quan sát hiện tại (hoặc giá trị dự báo mới sẽ bằng giá trị dự báo cũ cộng với một giá trị điều chỉnh rất đáng kể của sai số dự báo trước đó). Ngược lại, nếu α gần bằng 0, thì giá trị dự báo mới sẽ rất giống giá trị dự báo cũ và quan sát hiện tại sẽ có ảnh hưởng rất ít lên giá trị dự báo mới.

Nếu công thức (4.12) đúng với giai đoạn $t+1$, thì cũng đúng với giai đoạn t, và nếu đúng với giai đoạn t, thì cũng sẽ đúng với giai đoạn $t-1$, v.v... Nói cách khác, công thức (4.12) có thể được viết lại như sau:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t \tag{4.12}$$

$$\hat{Y}_t = \alpha Y_{t-1} + (1-\alpha)\hat{Y}_{t-1} \tag{4.14}$$

$$\hat{Y}_{t-1} = \alpha Y_{t-2} + (1-\alpha)\hat{Y}_{t-2} \tag{4.15}$$

Thế công thức (4.14) vào (4.12) ta sẽ có:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)[\alpha Y_{t-1} + (1-\alpha)\hat{Y}_{t-1}]$$

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + (1-\alpha)^2 \hat{Y}_{t-1} \tag{4.16}$$

Thế công thức (4.15) vào (4.16) ta sẽ có:

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + (1-\alpha)^2[\alpha Y_{t-2} + (1-\alpha)\hat{Y}_{t-2}]$$

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + (1-\alpha)^3 \hat{Y}_{t-2} \tag{4.17}$$

Nếu cứ tiếp tục thay thế vào như vậy ta sẽ có công thức tổng quát như sau:

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \alpha(1-\alpha)^3 Y_{t-3} +$$
$$+ \alpha(1-\alpha)^4 Y_{t-4} + \alpha(1-\alpha)^5 Y_{t-5} + \ldots + \alpha(1-\alpha)^n Y_{t-n} \tag{4.18}$$

Trong đó, n là số quan sát có sẵn trong mẫu dữ liệu quá khứ. Như vậy, với bất kỳ giá trị α bằng bao nhiêu, thì quan sát càng lùi sâu về quá khứ thì trọng số của nó trong giá trị dự báo càng nhỏ. Rõ ràng, theo công thức (4.18) thì giá trị dự báo $\hat{Y}_{t+1}$ là một giá trị san mũ, nghĩa là,

$\hat{Y}_{t+1}$ là một giá trị bình quân gia quyền của tất cả các quan sát quá khứ, trong đó trọng số của từng quan sát sẽ giảm theo hàm mũ khi quan sát đó dần xa về quá khứ. Tuy nhiên, tổng trọng số của tất cả các quan sát quá khứ phải bằng 1. Hơn nữa, nếu giá trị α lớn thì giá trị dự báo $\hat{Y}_{t+1}$ thực sự chỉ phụ thuộc vào một số quan sát gần nhất; ngược lại, nếu giá trị α nhỏ thì giá trị dự báo $\hat{Y}_{t+1}$ sẽ phụ thuộc vào tất cả các quan sát quá khứ. Để minh họa cho trọng số trong giá trị dự báo $\hat{Y}_{t+1}$, ta xem ví dụ trong Bảng 4.7.

■ BẢNG 4.7: **So sánh các hằng số san mũ.**

Giai đoạn	$\alpha = 0.1$		$\alpha = 0.6$	
		Trọng số		Trọng số
T		0.100		0.600
t-1	0.9*0.1	0.090	0.4*0.6	0.240
t-2	0.9*0.9*0.1	0.081	0.4*0.4*0.6	0.096
t-3	0.9*0.9*0.9*0.1	0.073	0.4*0.4*0.4*0.6	0.038
t-4	0.9*0.9*0.9**0.9*0.1	0.066	0.4*0.4*0.4**0.4*0.6	0.015
còn lại		0.590		0.011
Tổng		1.00	Tổng	1.00

■ BẢNG 4.8: **So sánh hai mô hình san mũ giản đơn (triệu đồng).**

Năm	Quý	Y_t	$\alpha = 0.1$		$\alpha = 0.6$	
			$\hat{Y}_t$	e_t	$\hat{Y}_t$	e_t
2002	1	500	500.0[a]	0.0	500.0	0.0
	2	350	500.0[b]	-150.0[c]	500.0	-150.0
	3	250	485.0[d]	-235.0	410.0	-160.0
	4	400	461.5[e]	-61.5	314.0	86.0

Năm	Quý	Y_t	$\alpha = 0.1$		$\alpha = 0.6$	
			$\hat{Y}_t$	e_t	$\hat{Y}_t$	e_t
2003	5	450	455.4	-5.4	365.6	84.4
	6	350	454.8	-104.8	416.2	-66.2
	7	200	444.3	-244.3	376.5	-176.5
	8	300	419.9	-119.9	270.6	29.4
2004	9	350	407.9	-57.9	288.2	61.8
	10	200	402.1	-202.1	325.3	-125.3
	11	150	381.9	-231.9	250.1	-100.1
	12	400	358.7	41.3	190.0	210.0
2005	13	550	362.8	187.2	316.0	234.0
	14	350	381.6	-31.6	456.4	-106.4
	15	250	378.4	-128.4	392.6	-142.6
	16	550	365.6	184.4	307.0	243.0
2006	17	550	384.0	166.0	452.8	97.2
	18	400	400.6	-0.6	511.1	-111.1
	19	350	400.5	-50.5	444.4	-94.4
	20	600	395.5	204.5	387.8	212.2
2007	21	750	415.9	334.1	515.1	234.9
	22	500	449.3	50.7	656.0	-156.0
	23	400	454.4	-54.4	562.4	-162.4
	24	650	449.0	201.0	465.0	185.0
2008	25	850	469.1	380.9	576.0	274.0
RMSE			163.48		156.67	

Yếu tố quan trọng nhất trong phương pháp san mũ giản đơn là việc xác định giá trị của hệ số san mũ α. Kinh nghiệm cho thấy rằng, nếu dữ liệu tương đối ổn định với mức biến thiên thấp, thì chúng ta có thể chọn giá trị α nhỏ. Ngược lại, nếu dữ liệu có mức biến thiên cao, thì

chúng ta có thể chọn giá trị α lớn. Để làm rõ điều này, chúng ta có thể xem xét ví dụ sau đây (**DATA4-4**).

Bảng 4.8 cung cấp dữ liệu về doanh thu (Y_t) theo quý của công ty bất động sản Hoàng Gia, và giá trị dự báo $\hat{Y}_t$ theo hai mô hình san mũ giản đơn với hệ số san mũ α lần lượt là 0.1 và 0.6. Trong Bảng 4.8, các giá trị dự báo và sai số dự báo được tính theo trình tự sau đây:

(a) Giá trị dự báo đầu tiên được chọn đúng bằng giá trị thực tế đầu tiên (500 ở quý I năm 2002).

(b) Sử dụng công thức (4.12) để tính giá trị dự báo thứ 2. Cụ thể như sau:

$$\hat{Y}_{1+1} = \alpha Y_1 + (1-\alpha)\hat{Y}_1$$

$$\hat{Y}_2 = 0.1*500 + 0.9*500 = 500$$

(c) Sai số dự báo được tính như sau:

$$e_2 = Y_2 - \hat{Y}_2 = 350 - 500 = 150$$

(d) Sử dụng công thức (4.12) để tính giá trị dự báo thứ 3. Cụ thể như sau:

$$\hat{Y}_{2+1} = \alpha Y_2 + (1-\alpha)\hat{Y}_2$$

$$\hat{Y}_3 = 0.1*350 + 0.9*500 = 485$$

(e) Sử dụng công thức (4.12) để tính giá trị dự báo thứ 4. Cụ thể như sau:

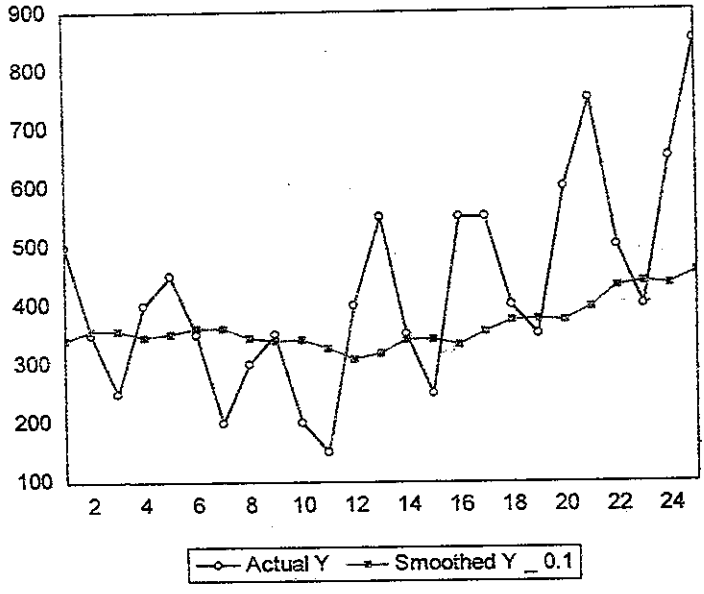
$$\hat{Y}_{3+1} = \alpha Y_3 + (1-\alpha)\hat{Y}_3$$

$$\hat{Y}_4 = 0.1*250 + 0.9*485 = 461.5$$

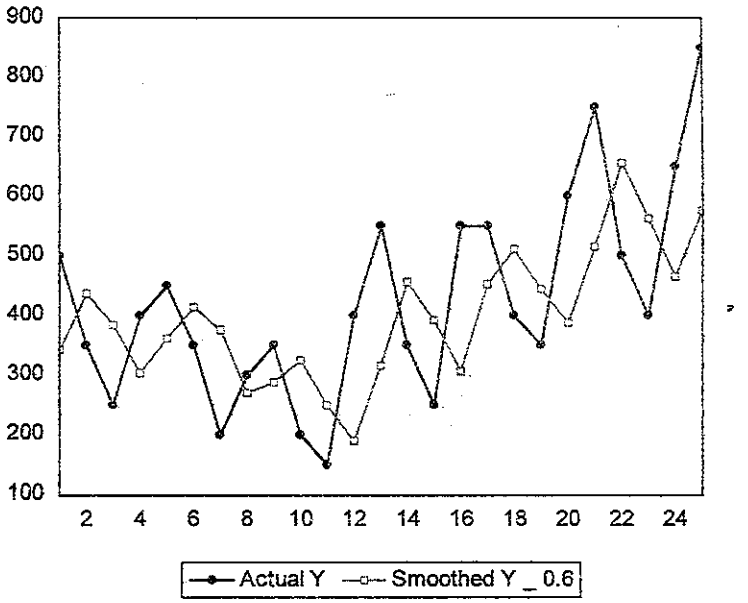
Quy trình này được thực hiện tiếp tục cho các giá trị dự báo còn lại của cả hai mô hình dự báo. Lưu ý, trong Excel, chúng ta nên cố định ô chứa giá trị α và copy công thức cho các giá trị dự báo và sai số dự báo còn lại.

Kết quả hai mô hình dự báo này được thể hiện trên các đồ thị ở Hình 4.14 và 4.15.

HÌNH 4.14: **Mô hình san mũ với $\alpha = 0.1$.**



■ HÌNH 4.15: **Mô hình san mũ với $\alpha = 0.6$.**



Như vậy, mô hình san mũ với hệ số $\alpha = 0.6$ tốt hơn so với mô hình san mũ với $\alpha = 0.1$. Điều này có vẻ phù hợp bởi vì dữ liệu thực Y_t có xu hướng biến thiên cao theo quý. Đến đây, vấn đề đặt ra là làm sao để xác định được hệ số α tối ưu? Cách duy nhất để xác định α tối ưu là phương pháp lặp đi lặp lại sao cho sai số dự báo tính toán được lỗ (có thể là RMSE) bé nhất. Trong Excel, chúng ta có thể sử dụng kỹ thuật phân tích độ nhạy một chiều.

Quy trình phân tích độ nhạy một chiều trên Excel:

Bước 1: Chọn một giá trị α bất kỳ (ví dụ $\alpha = 0.5$) đặt tại ô G5.

Bước 2: Tính giá trị dự báo như hướng dẫn ở Bảng 4.8.

Bước 3: Tính sai số dự báo như hướng dẫn ở Bảng 4.8.

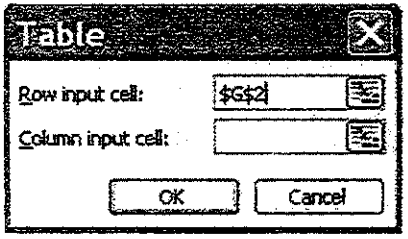
Bước 4: Tính RMSE (như hướng dẫn ở chương 1) tại ô K29.

Bước 5: Lập bảng theo dòng (2 dòng), với dòng trên là các giá trị α tùy chọn và dòng dưới để trống như sau:

	0.2	0.3	0.35	0.38	0.4	0.5	0.6
154.038							

Trong bảng này, giá trị 154.038 là RMSE ở Bước 4, được tính theo công thức =K29.

Bước 6: Chọn khối toàn bộ bảng trên (kể cả ô chứa giá trị 154.038), vào Data/Table, rồi nhập địa chỉ ô G5 vào ô "Row input cell" như sau:



Sau khi chọn "OK", ta có kết quả sau đây:

α	0.2	0.3	0.35	**0.38**	0.4	0.5	0.6
RMSE	158.87	153.56	152.76	**152.67**	152.73	154.04	156.08

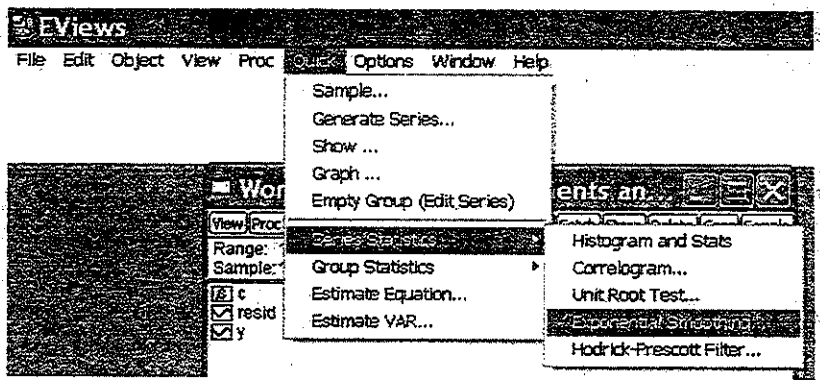
Như vậy, kết quả cho thấy có thể α = 0.38 là mức tối ưu vì có RMSE bé nhất (152.67).

May mắn thay, chúng ta có thể dễ dàng tìm ra mức α tối ưu bằng cách thực hiện mô hình san mũ trên Eviews.

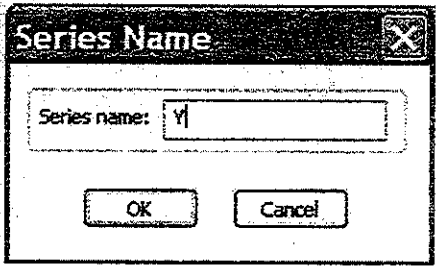
Quy trình thực hiện mô hình san mũ giản đơn trên Eviews:

Bước 1: Mở tập tin "**DATA4-4**" trên Eviews.

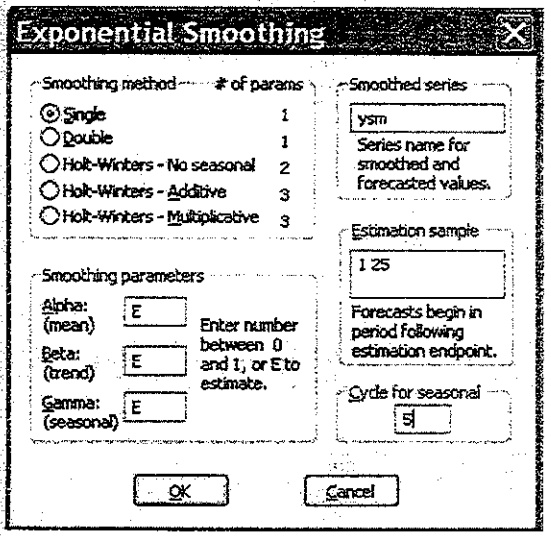
Bước 2: Vào Quick/Series Statistics/Exponential Smoothing.



Nhập tên biến cần dự báo vào ô "Series name" như sau:



Bước 3: Chọn "Single" như trong hộp thoại sau đây:



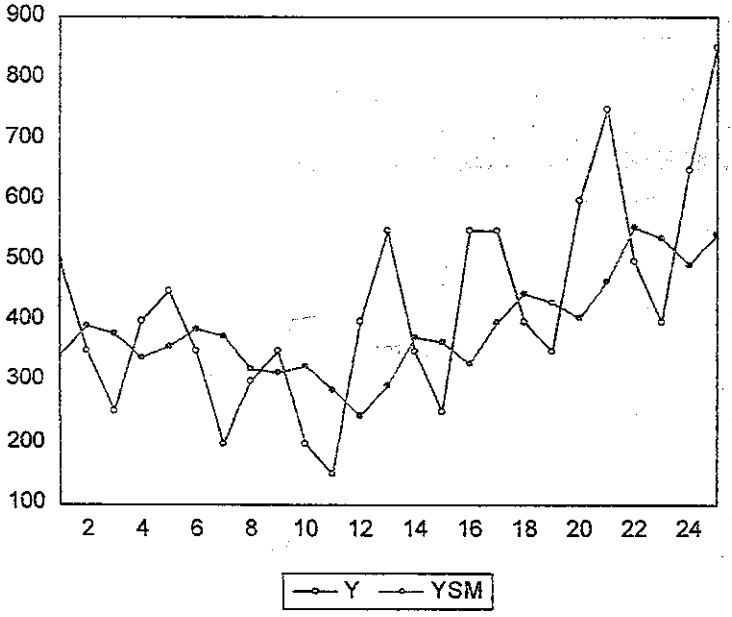
Sau khi chọn "OK", Eviews sẽ tự tạo một biến dự báo có tên YSM và bảng kết quả dự báo như sau:

Sample: 1 25		
Method: Single Exponential		
Original Series: Y		
Forecast Series: YSM		
Parameters:	Alpha	0.3140
Sum of Squared Residuals		566487.9
Root Mean Squared Error		150.5308
End of Period Levels:	Mean	639.8012

Như vậy, hệ số α **tối ưu** theo ước tính của Eviews sẽ bằng **0.314**.

Bước 4: Vẽ đồ thị với Y và YSM bằng cách chọn Quick/Graph và nhập tên biến Y YSM vào. Ta có đồ thị như sau:

■ HÌNH 4.16: **Mô hình san mũ với $\alpha = 0.314$.**



Quy trình thực hiện san mũ giản đơn trên Crystal Ball:

Bước 1: Khởi động Crystal Ball (7.2 hoặc 7.3).

Bước 2: Mở tập tin "**DATA4-4.xls**".

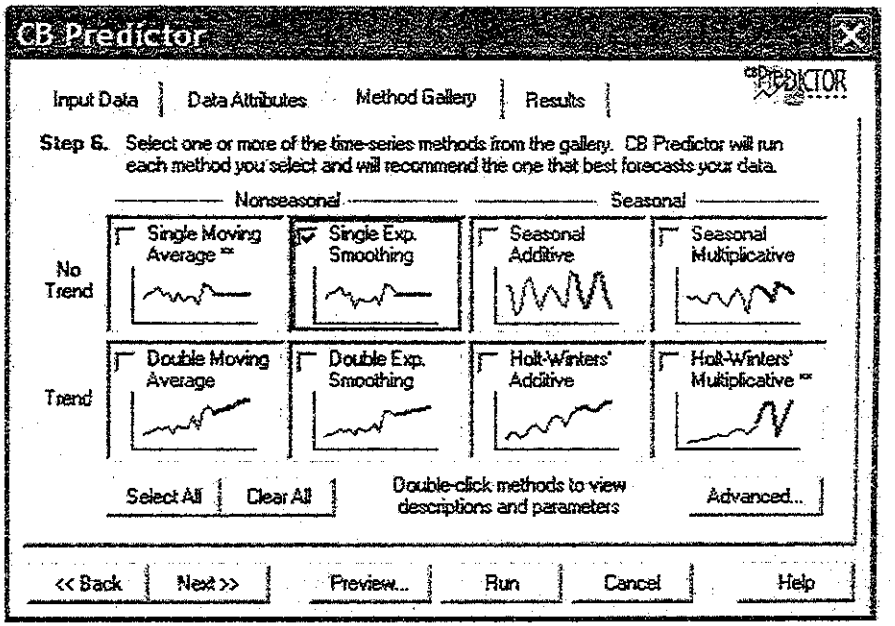
Bước 3: Vào "Run/CB Predictor..." (giống như ở mô hình MA).

Bước 4: Nhập dữ liệu vào "Input Data" (giống như ở mô hình MA).

Bước 5: Khai báo đặc điểm dữ liệu vào "Data Attributes" (giống như ở mô hình MA).

Bước 6: Chọn phương pháp dự báo trong "Method Gallery".

■ HÌNH 4.17: **Lựa chọn phương pháp dự báo trên Crystal Ball.**

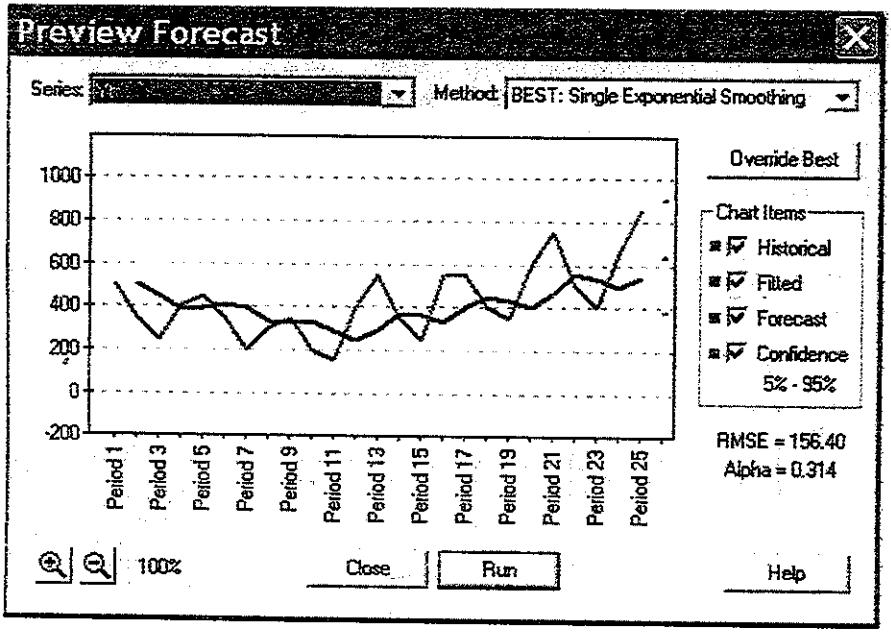


- Trong **Step 6**, chọn "Single Exp.Smoothing".

- Nếu nhấp đúp vào "Single Exp.Smoothing" ta sẽ thấy xuất hiện một hộp thoại (tương tự như Hình 4.10) và nhập 0.314 vào ô "Alpha", OK
- Chọn "Next"

Bước 7: Khai báo kết quả dự báo vào "Results" (giống như ở mô hình trung bình di động - MA)

HÌNH 4.18: **Xem đồ thị dự báo trên Crystal Ball.**



■ BẢNG 4.9: **Kết quả dự báo san mũ giản đơn trên Crystal Ball.**

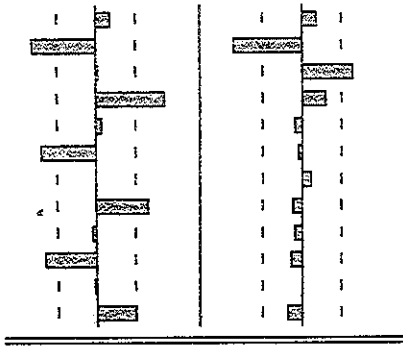
	A	B	C	D	E	F
1	**Results Table**					
2	Created: 5/3/2009 at 9:17:02 PM					
3						
31	Series	Yt				
32						
33		Data				
34	Date	Historical Data	Lower: 5%	Fit & Forecast	Upper: 95%	Residuals
35	Period 1	500				500
36	Period 2	350		500		-150
37	Period 3	250		452.9		-202.9
38	Period 4	400		389.1894		10.8106
51	Period 17	550		399.1658073		150.8341927
52	Period 18	400		446.5277438		-46.5277438
53	Period 19	350		431.9180323		-81.9180323
54	Period 20	600		406.1957701		193.8042299
55	Period 21	750		467.0502983		282.9497017
56	Period 22	500		555.8965046		-55.8965046
57	Period 23	400		538.3450022		-138.345002
58	Period 24	650		494.9046715		155.0953285
59	Period 25	850		543.6046046		306.3953954
60	Period 26		382.092001	639.8127588	897.533517	

Một yếu tố khác, ngoài hệ số α, có ảnh hưởng đáng kể đến kết quả dự báo theo mô hình san mũ giản đơn là việc chọn giá trị dự báo đầu tiên, $\hat{Y}_1$. Trong Bảng 4.8 ta đã chọn $\hat{Y}_1 = Y_1$, và sự lựa chọn này có xu hướng gán cho Y_1 một trọng số tương đối cao trong các giá trị dự báo tiếp sau đó. Tuy nhiên, nếu dữ liệu đủ lớn thì đây không phải là vấn đề quan trọng vì ảnh hưởng của giá trị dự báo đầu tiên sẽ giảm đáng kể khi t tăng lên. Một cách khác có thể được sử dụng trong mô hình san mũ là lấy giá trị trung bình của k quan sát đầu tiên. Ở đây, k có thể được chọn tùy ý. Tuy nhiên, nếu dữ liệu theo quý, thì ta nên chọn k = 4, nếu dữ liệu theo tháng thì ta nên chọn k = 12. Nếu ta giữ nguyên hệ số α, thì giá trị dự báo tối ưu đầu tiên vẫn có thể được thực hiện bằng kỹ thuật phân tích độ nhạy một chiều. Lưu ý, Eviews cũng sẽ giúp chúng ta chọn giá trị dự báo đầu tiên sao cho RMSE bé nhất.

Một vấn đề hết sức quan trọng nữa, ngoài xem xét đồ thị và các tiêu chí thống kê về sai số dự báo, là làm sao nhận biết một mô hình dự báo nào đó (kể cả MA, DMA) có phải là một mô hình phù hợp đối

với dữ liệu sẵn có hay không? Hanke (2005) đề xuất nên sử dụng giản đồ tự tương quan để xem xét phần dư (sai số dự báo) có thực sự là một ngẫu nhiên chưa. Nếu phần dư là ngẫu nhiên, thì mô hình dự báo thực sự là mô hình phù hợp với dữ liệu. Ngược lại, nếu có một hoặc một số hệ số tự tương quan nằm ngoài 'biên độ' của giản đồ tự tương quan thì điều này cảnh báo rằng đó chưa phải là một mô hình dự báo tốt. Chẳng hạn, mô hình san mũ giản đơn với hệ số $\alpha = 0.314$ chưa phải là một mô hình dự báo tốt nhất đối với dữ liệu doanh số của công ty bất động sản Hoàng Gia vì có một số hệ số tự tương quan nằm ngoài 'biên độ' (hai đường biên không liên tục ở cột 'Autocorrelation') theo như kết quả của giản đồ tự tương quan trình bày ở Hình 4.19.

■ HÌNH 4.19: **Giản đồ tự tương quan của e_t với $\alpha = 0.314$.**



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.137	0.137	0.5246	0.469
		2	-0.649	-0.681	12.904	0.002
		3	0.024	0.505	12.921	0.005
		4	0.687	0.222	28.099	0.000
		5	0.061	-0.071	28.224	0.000
		6	-0.561	-0.042	39.422	0.000
		7	0.004	0.091	39.422	0.000
		8	0.519	-0.089	50.064	0.000
		9	-0.028	-0.071	50.096	0.000
		10	-0.520	-0.110	62.253	0.000
		11	-0.011	0.002	62.259	0.000
		12	0.393	-0.140	70.274	0.000

SAN MŨ HOLT

Khác với san mũ giản đơn, mô hình san mũ Holt được sử dụng đối với dữ liệu có yếu tố xu thế. Hanke (2005) cho rằng bởi vì hầu hết các chuỗi dữ liệu kinh tế và kinh doanh hiếm khi theo một xu thế cố định, nên chúng ta cần xem xét khả năng mô hình hóa các xu thế mang tính cục bộ và thay đổi theo thời gian. Holt (1957) đã phát triển một phương pháp san mũ, được gọi là phương pháp san mũ tuyến tính Holt, cho phép suy diễn các xu thế cục bộ và có thể được sử dụng cho dự báo tương lai.

Khi chuỗi thời gian có yếu tố xu thế (cục bộ), thì chúng ta cần phải dự báo cả giá trị trung bình (giá trị san mũ) và độ dốc (xu thế) hiện tại để làm cơ sở cho dự báo tương lai. Ý tưởng cơ bản của phương pháp Holt là sử dụng các hệ số san mũ khác nhau để ước lượng giá trị trung bình và độ dốc của chuỗi thời gian (theo mô hình san mũ giản đơn). Trên cơ sở san mũ giản đơn, các hệ số san mũ này sẽ đưa ra các giá trị ước lượng về mức trung bình và độ dốc ngay khi có sẵn một quan sát mới. Nói cách khác, giá trị trung bình hiện tại vẫn là trung bình với trọng số giảm dần của tất cả các giá trị trung bình quá khứ; và độ dốc hiện tại sẽ là trung bình với trọng số giảm dần của tất cả các độ dốc quá khứ. Mô hình san mũ Holt được thể hiện qua ba phương trình sau đây:

1. Ước lượng giá trị trung bình hiện tại[1]:

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{4.19}$$

2. Ước lượng xu thế (độ dốc):

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \tag{4.20}$$

3. Dự báo p giai đoạn trong tương lai:

$$\hat{Y}_{t+p} = L_t + pT_t \tag{4.21}$$

Trong đó:

L_t = Giá trị san mũ mới (hoặc giá trị ước lượng trung bình hiện tại).

α = Hệ số san mũ của giá trị trung bình ($0 < \alpha < 1$).

Y_t = Giá trị quan sát hoặc giá trị thực tế vào thời điểm t.

β = Hệ số san mũ của giá trị xu thế ($0 < \beta < 1$).

[1] Lưu ý, ta sử dụng chữ L vì đó là chữ viết tắt của từ "Level", nghĩa là giá trị trung bình.

T_t = Giá trị ước lượng của xu thế.

p = Thời đoạn dự báo trong tương lai.

$\hat{Y}_{t+p}$ = Giá trị dự báo cho p giai đoạn trong tương lai.

Phương trình (4.19) rất giống với phương trình (4.12) của san mũ giản đơn. Theo 'ngôn ngữ' của san mũ giản đơn, thì $(L_{t-1} + T_{t-1})$ chính là $\hat{Y}_{t-1}$ điều chỉnh. Như vậy, giá trị san mũ (trung bình) dự báo tại thời điểm t được tính bằng bình quân gia quyền giữa hai giá trị ước lượng của trung bình: một ước lượng chính là giá trị quan sát Y_t và một ước lượng khác được tính bằng cách cộng thêm yếu tố xu thế T_{t-1} vào giá trị san mũ trước đó (L_{t-1}); nghĩa là dự báo cho giá trị ước lượng của trung bình tại thời điểm *t*. So với phương trình (4.12), thì giá trị dự báo ở giai đoạn *t* đã cập nhật yếu tố xu thế (T_{t-1}) vào giá trị trung bình. Nếu không có yếu tố xu thế trong dữ liệu, thì không cần thiết đưa T_{t-1} vào phương trình (4.19), và khi đó phương trình (4.19) và (4.13) sẽ như nhau. Nếu triển khai phương trình (4.19) như cách đã thực hiện ở phương trình (4.18), thì giá trị san mũ mới của giá trị trung bình chính là bình quân gia quyền của tất cả các giá trị quan sát trong quá khứ với trọng số giảm theo hàm mũ.

$$L_t = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \alpha(1-\alpha)^3 Y_{t-3} + \alpha(1-\alpha)^4 Y_{t-4} + \\ + \alpha(1-\alpha)^5 Y_{t-5} + ... + \alpha(1-\alpha)^n Y_{t-n} \qquad (4.22)$$

Hệ số san mũ thứ hai, β, được sử dụng để tạo ra các ước lượng xu thế. Phương trình (4.20) cho thấy giá trị ước lượng của xu thế mới là bình quân gia quyền của hai giá trị ước lượng của xu thế: một giá trị ước lượng được tính bằng thay đổi trong giá trị của trung bình từ thời điểm t-1 sang t $(L_t - L_{t-1})$, và một giá trị ước lượng là xu thế san mũ trước đó (T_{t-1}), nghĩa là giá trị dự báo cho $(L_t - L_{t-1})$. Về mặt ý tưởng, thì phương trình (4.20) cũng giống như phương trình (4.19), nhưng khác biệt ở đây chỉ là san mũ đối với 'xu thế' chứ không phải san mũ đối với dữ liệu quan sát thực. Nếu triển khai phương trình (4.20) như cách đã thực hiện ở phương trình (4.18), thì giá trị san mũ mới của xu thế chính là giá trị trung bình gia quyền của tất cả các giá trị san mũ của xu thế trong quá khứ với trọng số giảm theo hàm mũ.

$$T_t = \beta[L_t - L_{t-1}] + \beta(1-\beta)[L_{t-1} - L_{t-2}] + \beta(1-\beta)^2[L_{t-2} - L_{t-3}] + ...$$
$$...+ \beta(1-\beta)^n[L_{t-n} - L_{t-n-1}] \quad (4.23)$$

$$T_t = \beta\Delta L_t + \beta(1-\beta)\Delta L_{t-1} + \beta(1-\beta)^2\Delta L_{t-2} + \beta(1-\beta)^3\Delta L_{t-3} + ...$$
$$...+ \beta(1-\beta)^n\Delta L_{t-n} \quad (4.24)$$

Phương trình (4.21) hàm ý rằng, nếu ta dự báo cho một giai đoạn tiếp theo ($\hat{Y}_{t+1}$), thì p sẽ bằng 1, hai giai đoạn tiếp theo ($\hat{Y}_{t+2}$), thì p sẽ bằng 2, ... Như vậy, kết quả dự báo chính là giá trị san mũ mới (trung bình) cò điều chỉnh yếu tố xu thế cục bộ.

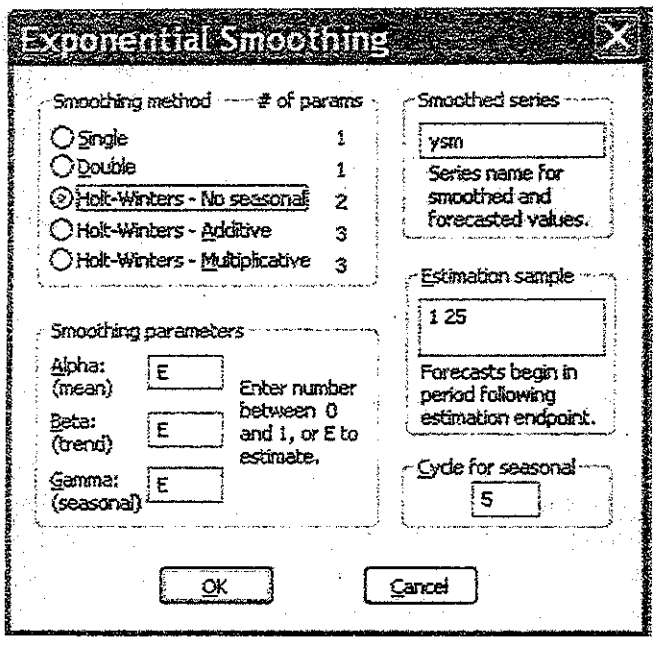
Tương tự như san mũ giản đơn, các hệ số san mũ α và β trong san mũ Holt có thể được chọn một cách chủ quan hoặc bằng cách tối thiểu hóa sai số dự báo (RMSE). Ở đây, chúng ta có thể sử dụng kỹ thuật phân tích độ nhạy hai chiều để xác định α và β tối ưu. Như đã đề cập ở san mũ giản đơn, nhờ sự phát triển của các phần mềm kinh tế lượng, nên chúng ta không nhất thiết phải mất nhiều thời gian cho vấn đề này. Tuy nhiên, kinh nghiệm cho thấy rằng, nếu dữ liệu có mức độ biến thiên cao, thì chúng ta nên chọn các hệ số san mũ lớn, và ngược lại.

Quy trình thực hiện mô hình san mũ Holt trên Eviews:

Bước 1: Mở tập tin "**DATA4-4**" trên Eviews (giống san mũ giản đơn).

Bước 2: Vào Quick/Series Statistics/Exponential Smoothing (giống san mũ giản đơn).

Bước 3: Chọn "Holt-Winters – No seasonal" như trong hộp thoại sau đây:



Sau khi chọn "OK", Eviews sẽ tự tạo một biến dự báo có tên YSM và bảng kết quả dự báo như sau:

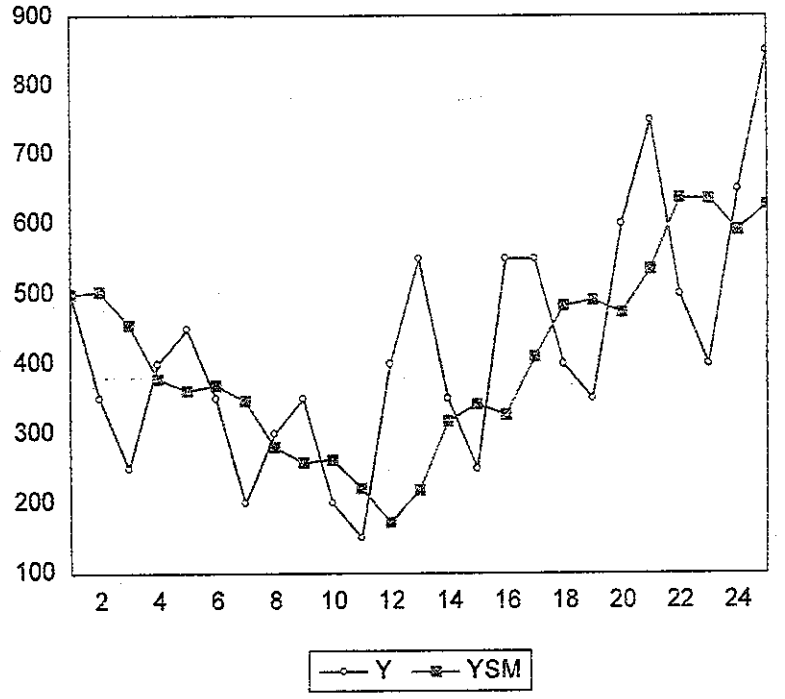
Sample: 1 25
Method: Holt-Winters No Seasonal
Original Series: Y
Forecast Series: YSM

Parameters:	Alpha	0.2600
	Beta	0.3000
Sum of Squared Residuals		570515.6
Root Mean Squared Error		151.0650
End of Period Levels:	Mean	685.5251
	Trend	38.38341

Như vậy, hệ số α và β tối ưu theo ước tính của Eviews sẽ lần lượt bằng 0.26 và 0.30.

Bước 4: Vẽ đồ thị với Y và YSM bằng cách chọn Quick/Graph và nhập tên biến Y YSM vào. Ta có đồ thị như sau:

■ HÌNH 4.20: **Mô hình san mũ Holt với $\alpha = 0.26$ và $\beta = 0.30$.**



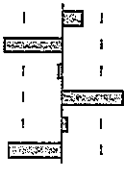
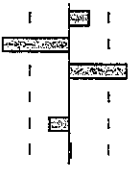
Quy trình thực hiện trên Crystal Ball:

Bước 1 - Bước 5 và Bước 7: Giống như ở san mũ giản đơn

Bước 6: Chọn phương pháp dự báo trong "Method Gallery"

- Trong **Step 6**, chọn "Double Exp.Smoothing"
- Nếu nhấp đúp vào "Double Exp.Smoothing" ta sẽ thấy xuất hiện một hộp thoại (tương tự như Hình 4.10), nhập 0.26 vào ô "Alpha", và 0.3 vào ô "Beta", OK.

■ HÌNH 4.21: **Giản đồ tự tương quan của e_t với $\alpha = 0.314$.**

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.200	0.200	1.1209	0.290
		2	-0.600	-0.666	11.679	0.003
		3	-0.030	0.567	11.706	0.008
		4	0.607	0.005	23.538	0.000
		5	0.047	-0.195	23.613	0.000
		6	-0.544	0.024	34.133	0.000

Như vậy, sai số dự báo không phải là một chuỗi ngẫu nhiên, nên san mũ Holt với $\alpha = 0.26$ và $\beta = 0.3$ chưa phải là mô hình dự báo tốt. Bởi vì, mô hình san mũ Holt nói chung chỉ có thể phù hợp với các dữ liệu có xu hướng dao động ngẫu nhiên quanh yếu tố xu thế. Cho nên, với dữ liệu có dạng mùa vụ như Hình 4.20 thì chúng ta nên sử dụng loại mô hình khác.

SAN MŨ WINTERS

San mũ Winters là một phương pháp mở rộng của san mũ Holt đối với các dữ liệu có chứa yếu tố mùa. Yếu tố mùa trong chuỗi thời gian có thể thuộc dạng phép cộng hoặc phép nhân. Dạng phép cộng có nghĩa là yếu tố mùa ở các năm khác nhau được lặp đi lặp lại một cách đều đặn. Ngược lại, dạng phép nhân có nghĩa là yếu tố mùa ở năm sau được lặp đi lặp lại nhưng với một cường độ cao hơn hoặc thấp hơn so với từng mùa trong năm trước. Mô hình san mũ Winters tổng quát nhất là mô hình dạng nhân tính. Mô hình này được ước lượng thông qua hệ bốn phương trình sau đây:

1. Ước lượng giá trị trung bình hiện tại:

$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + T_{t-1}) \qquad (4.25)$$

2. Ước lượng giá trị xu thế (độ dốc):

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \qquad (4.26)$$

3. Ước lượng giá trị chỉ số mùa

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma) S_{t-s} \qquad (4.27)$$

4. Dự báo p giai đoạn trong tương lai

$$\hat{Y}_{t+p} = (L_t + pT_t) S_{t-s+p} \qquad (4.28)$$

Trong đó,

L_t = Giá trị san mũ mới (hoặc giá trị ước lượng trung bình hiện tại).

α = Hệ số san mũ của giá trị trung bình ($0 < \alpha < 1$).

Y_t = Giá trị quan sát hoặc giá trị thực tại thời điểm t.

β = Hệ số san mũ của giá trị xu thế ($0 < \beta < 1$).

T_t = Giá trị ước lượng của xu thế.

γ = Hệ số san mũ của chỉ số mùa.

S_t = Giá trị ước lượng của chỉ số mùa.

p = Thời đoạn dự báo trong tương lai.

s = Độ dài của yếu tố mùa.

$\hat{Y}_{t+p}$ = Giá trị dự báo cho p giai đoạn trong tương lai.

Phương trình (4.27) cho rằng giá trị dự báo chỉ số mùa vụ hiện tại, S_t, được tính bằng bình quân gia quyền giữa giá trị ước lượng của chỉ số mùa vụ hiện tại, Y_t/L_t, và giá trị dự báo chỉ số mùa vụ trước đó, S_{t-s}. Nói cách khác, phương trình này chính là san mũ giản đơn giữa Y_t/L_t và Y_{t-1}/L_{t-1}. Khác với san mũ Holt, phương trình (4.25) có điều chỉnh loại ảnh hưởng mùa vụ để ước lượng giá trị trung bình. Do đây là mô hình nhân tính, nên để dự báo cho các giai đoạn tương lai, chúng ta nhân giá trị ước lượng của trung bình có điều chỉnh xu thế với chỉ số mùa vụ của từng mùa vụ riêng biệt.

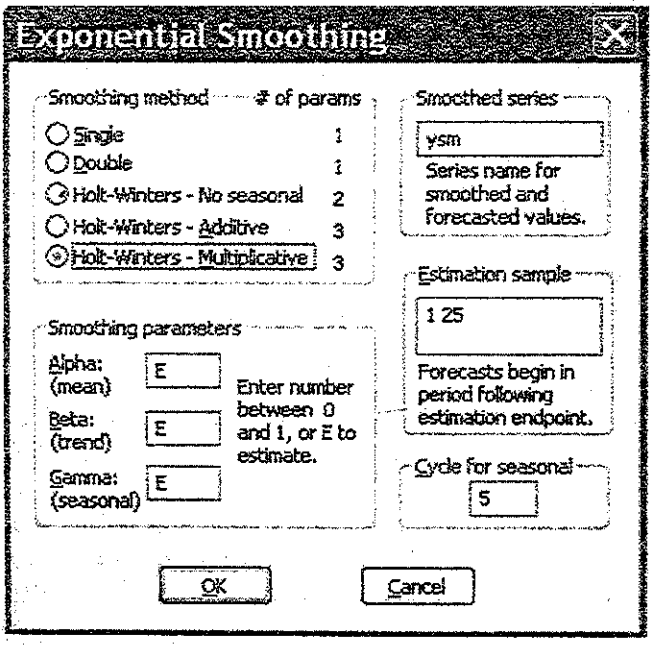
Đối với mô hình cộng tính, thì có ba điểm khác biệt so với mô hình nhân tính là: (i) Giá trị san mũ tại thời điểm t là $(Y_t - S_{t-s})$, chứ không phải (Y_t/S_{t-s}), (ii) chỉ số mùa vụ tại thời điểm t là $(Y_t - L_t)$, chứ không phải (Y_t/L_t), và (iii) Giá trị dự báo mới sẽ là tổng của ba thành phần $(L_t + pT_t + S_{t-s+p})$.

Quy trình thực hiện mô hình san mũ Winters trên Eviews:

Bước 1: Mở tập tin "**DATA4-4**" trên Eviews (giống san mũ giản đơn và san mũ Holt).

Bước 2: Vào Quick/Series Statistics/Exponential Smoothing (giống san mũ giản đơn và san mũ Holt).

Bước 3: Chọn "Holt-Winters – Multiplicative" (mô hình nhân tính) như trong hộp thoại sau đây:



Sau khi chọn "OK", Eviews sẽ tự tạo một biến dự báo có tên YSM và bảng kết quả dự báo như sau:

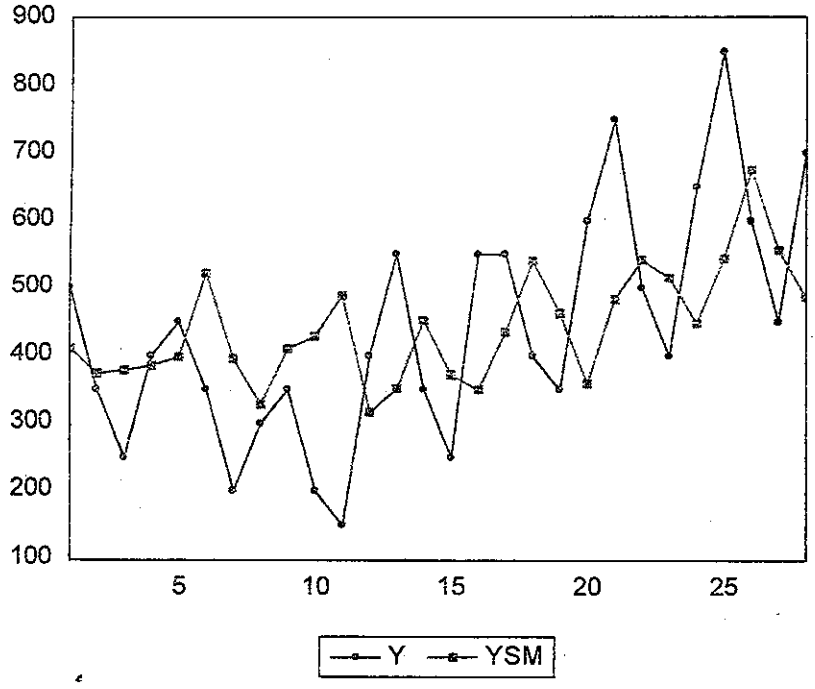
Sample: 1 25
Method: Holt-Winters Multiplicative Seasonal
Original Series: Y
Forecast Series: YSM

Parameters:	Alpha		0.2700
	Beta		0.0000
	Gamma		0.0000
Sum of Squared Residuals			481368.1
Root Mean Squared Error			138.7814
End of Period Levels:	Mean		660.2188
	Trend		12.00000
	Seasonals:	21	1.122707
		22	0.968152
		23	0.954173
		24	0.977381
		25	0.977587

Như vậy, hệ số α tối ưu là 0.27, nhưng các hệ số β và γ có giá trị rất nhỏ. Điều này chúng tỏ sự biến thiên trong xu thế và chỉ số mùa vụ trong dữ liệu là rất thấp.

Bước 4: Vẽ đồ thị với Y và YSM bằng cách chọn Quick/Graph và nhập tên biến Y YSM vào. Ta có đồ thị như sau:

■ HÌNH 4.22: **Mô hình san mũ Winters.**



Quy trình thực hiện trên Crystal Ball:

Bước 1 - Bước 5 và Bước 7: Giống như ở san mũ giản đơn và san mũ Holt.

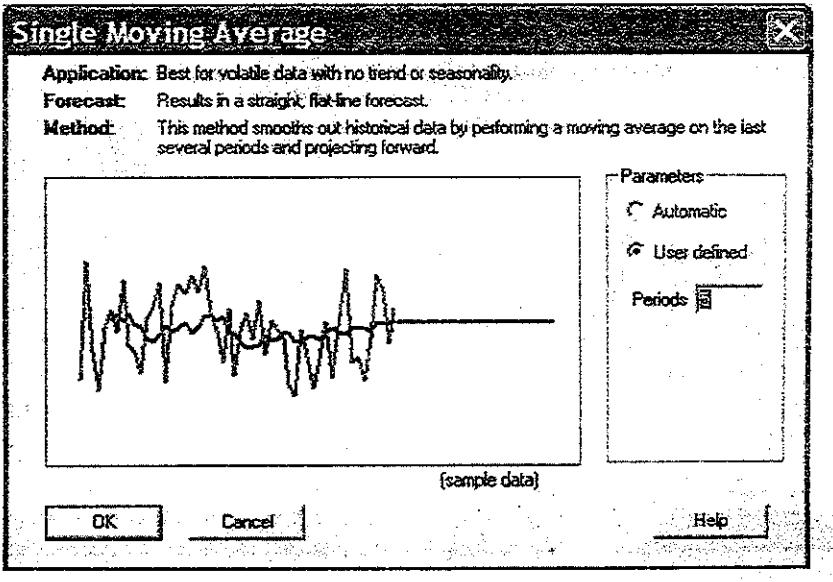
Bước 6: Chọn phương pháp dự báo trong "Method Gallery".

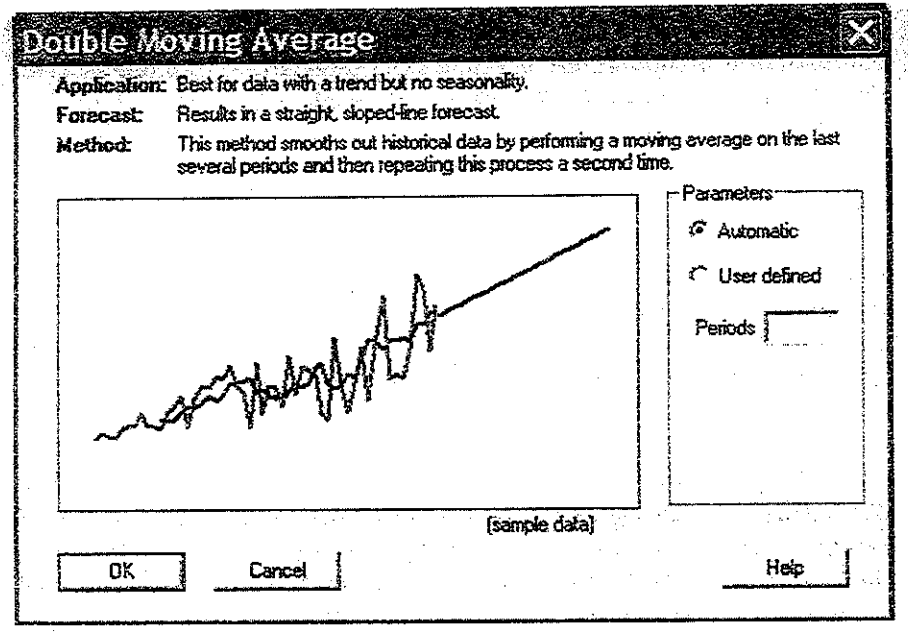
- Trong **Step 6**, chọn "Holt-Winters' Multiplicative".
- Nếu nhấp đúp vào "Holt-Winters' Multiplicative" ta sẽ thấy xuất hiện một hộp thoại (tương tự như Hình 4.10), nhập 0.4 vào ô "Alpha", 0.1 vào ô "Beta", và 0.3 vào ô "Gamma", OK.

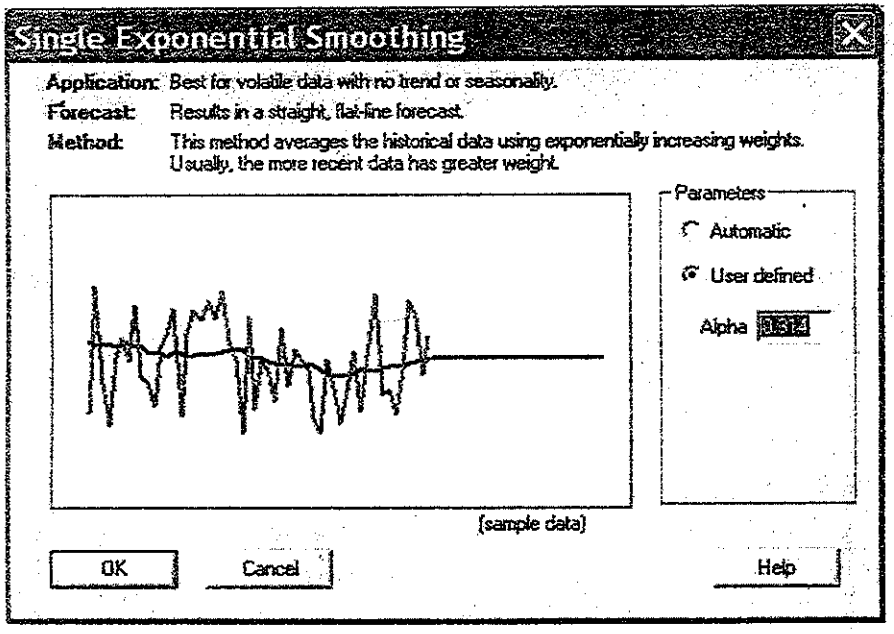
Mặc dù, Eviews là một phần mềm rất mạnh về phân tích chuỗi thời gian, đặc biệt là các mô hình kinh tế lượng và dự báo phức tạp, nhưng

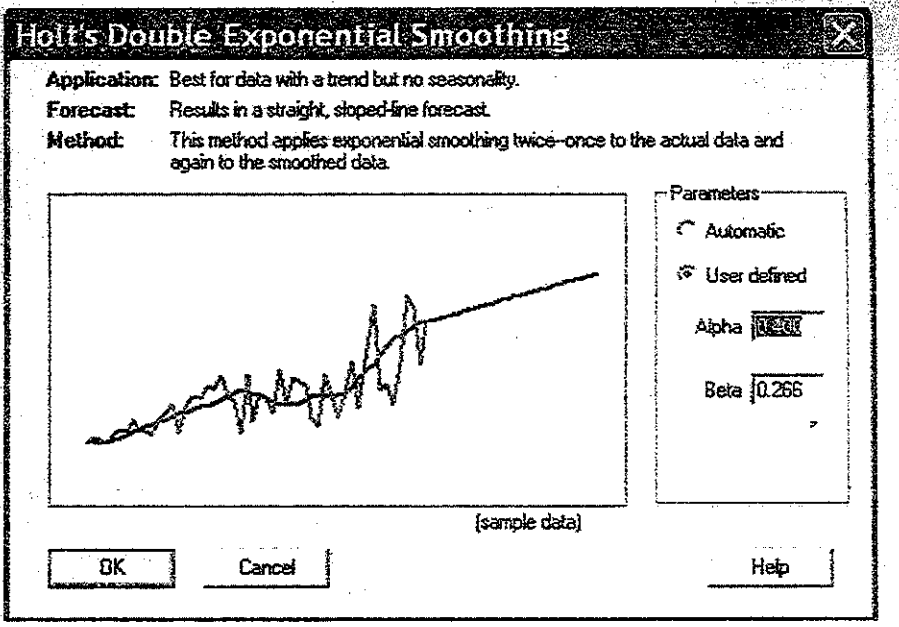
Eviews không phải là công cụ hỗ trợ tốt nhất đối với các mô hình dự báo giản đơn. Kinh nghiệm cho thấy, Crystal Ball rất hữu ích cho việc dự báo các mô hình giản đơn vừa được trình bày trong chương này. Tại sao chúng ta nên sử dụng Crystal Ball? Có ba lý do sau đây. Thứ nhất, Crystal Ball là một phần mềm chuyên dụng chạy trên nền Excel sau khi cài đặt bổ xung chức năng này. Thứ hai, hầu hết các sinh viên ngành kinh tế - quản trị đều ngày càng có xu hướng sử dụng Crystal Ball trong việc hỗ trợ nhiều môn học chuyên ngành cần thực hiên phân tích mô phỏng như tài chính, thẩm định, phân tích rủi ro, và dự báo. Thứ ba, Crystal Ball sẽ tự động phân tích dữ liệu quá khứ và đề xuất thứ tự ưu tiên mô hình nào phù hợp nhất đối với dữ liệu sẵn có.

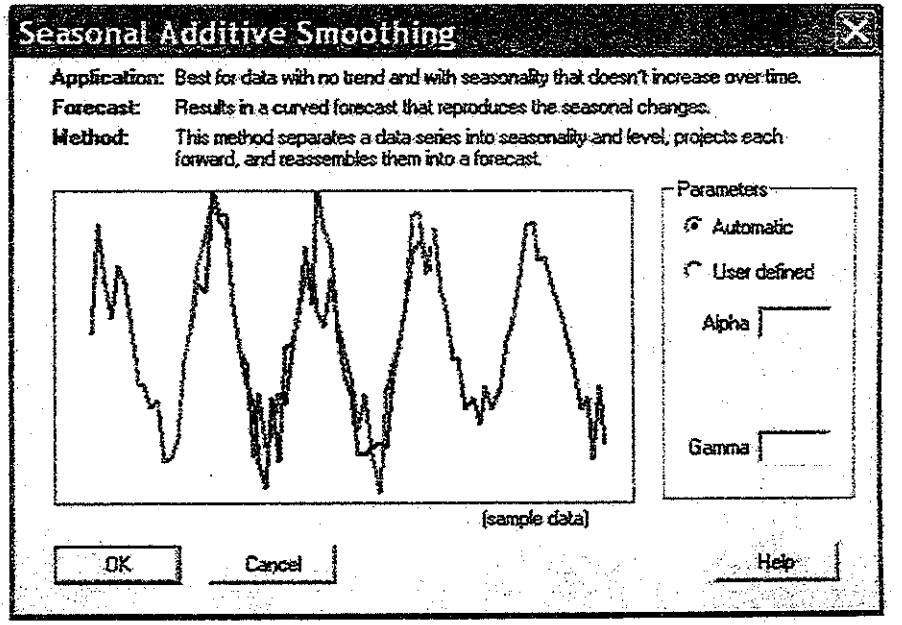
Crystal Ball chia các mô hình dự báo giản đơn thành tám phương pháp riêng biệt, tùy vào dạng dữ liệu quá khứ. Từng phương pháp cụ thể được mô tả thông qua 8 hộp thoại trình bày dưới đây bao gồm: Single Moving Averrage, Double Moving Averrage, Single Exponetial Smoothing, Double Exponetial Smoothing, Seasonal Additive Smoothing, Holt Winters Seasonal Additive Smoothing, Seasonal Multiplicative Smoothing, Holt Winters Seasonal Multiplicative Smoothing, v.v.

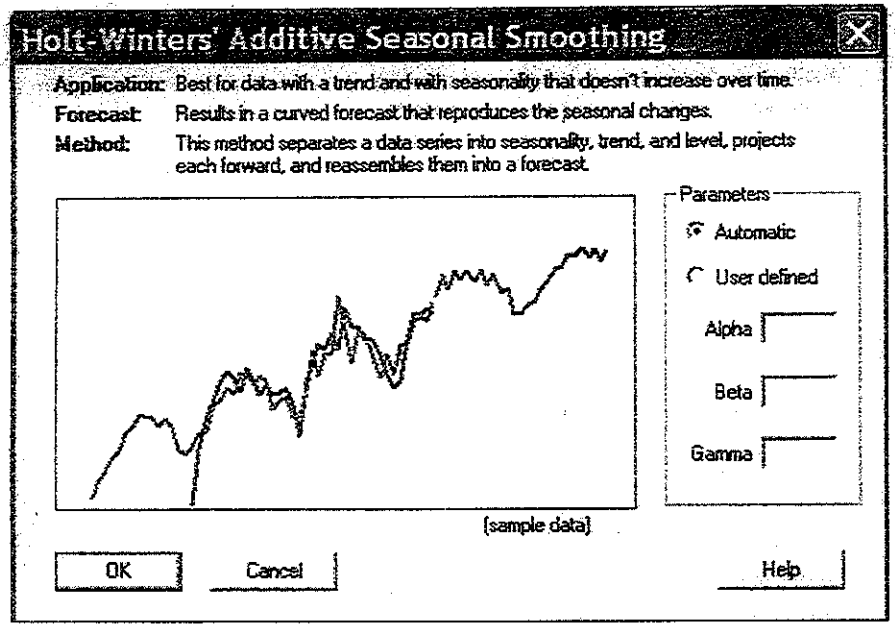


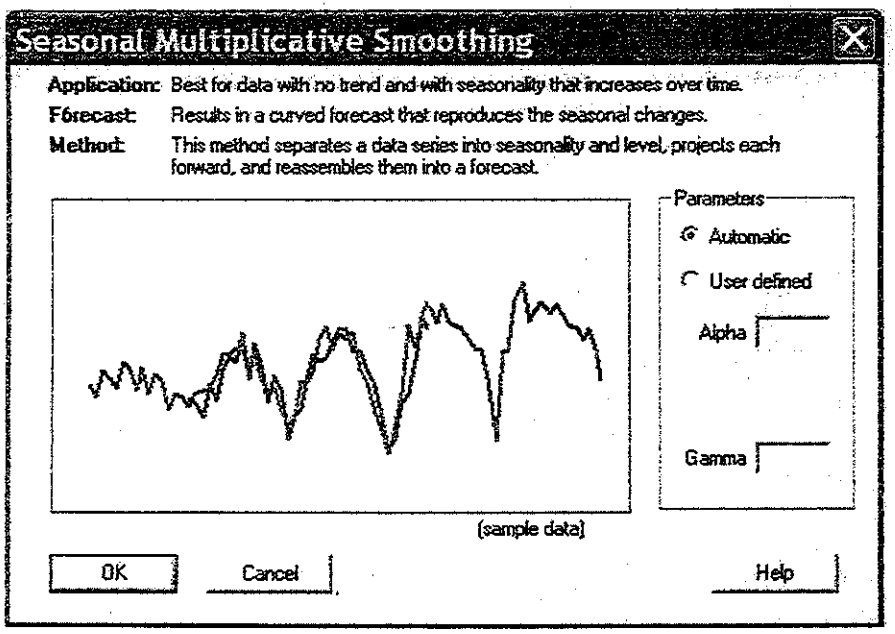

Double Moving Average
Application: Best for data with a trend but no seasonality.
Forecast: Results in a straight, sloped-line forecast.
Method: This method smooths out historical data by performing a moving average on the last several periods and then repeating this process a second time.
Parameters
Automatic
User defined
Periods
(sample data)
OK
Cancel
Help

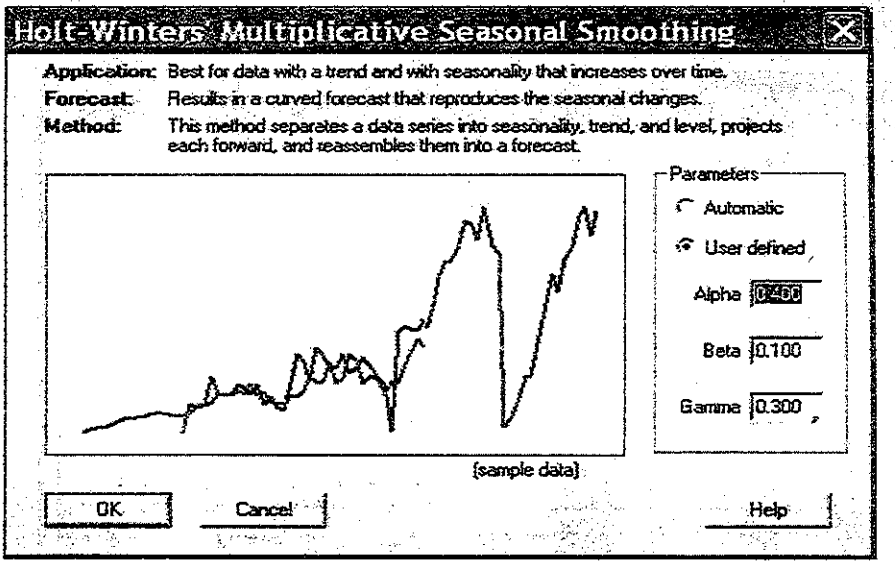

Single Exponential Smoothing
Application: Best for volatile data with no trend or seasonality.
Forecast: Results in a straight, flat-line forecast.
Method: This method averages the historical data using exponentially increasing weights. Usually, the more recent data has greater weight.
Parameters
Automatic
User defined
Alpha
(sample data)
OK
Cancel
Help


Holt's Double Exponential Smoothing
Application: Best for data with a trend but no seasonality.
Forecast: Results in a straight, sloped-line forecast.
Method: This method applies exponential smoothing twice--once to the actual data and again to the smoothed data.
Parameters
Automatic
User defined
Alpha
Beta
0.266
(sample data)
OK
Cancel
Help


Seasonal Additive Smoothing
Application: Best for data with no trend and with seasonality that doesn't increase over time.
Forecast: Results in a curved forecast that reproduces the seasonal changes.
Method: This method separates a data series into seasonality and level, projects each forward, and reassembles them into a forecast.
Parameters
Automatic
User defined
Alpha
Gamma
(sample data)
OK
Cancel
Help


Holt-Winters' Additive Seasonal Smoothing
Application: Best for data with a trend and with seasonality that doesn't increase over time.
Forecast: Results in a curved forecast that reproduces the seasonal changes.
Method: This method separates a data series into seasonality, trend, and level, projects each forward, and reassembles them into a forecast.
Parameters
Automatic
User defined
Alpha
Beta
Gamma
(sample data)
OK
Cancel
Help


Seasonal Multiplicative Smoothing
Application: Best for data with no trend and with seasonality that increases over time.
Forecast: Results in a curved forecast that reproduces the seasonal changes.
Method: This method separates a data series into seasonality and level, projects each forward, and reassembles them into a forecast.
Parameters
Automatic
User defined
Alpha
Gamma
(sample data)
OK
Cancel
Help



Quy trình chuẩn thực hiện dự báo giản đơn trên Crystal Ball:

Bước 1: Khởi động Crystal Ball (7.2 hoặc 7.3).

Bước 2: Mở tập tin "**DATA4-4.xls**".

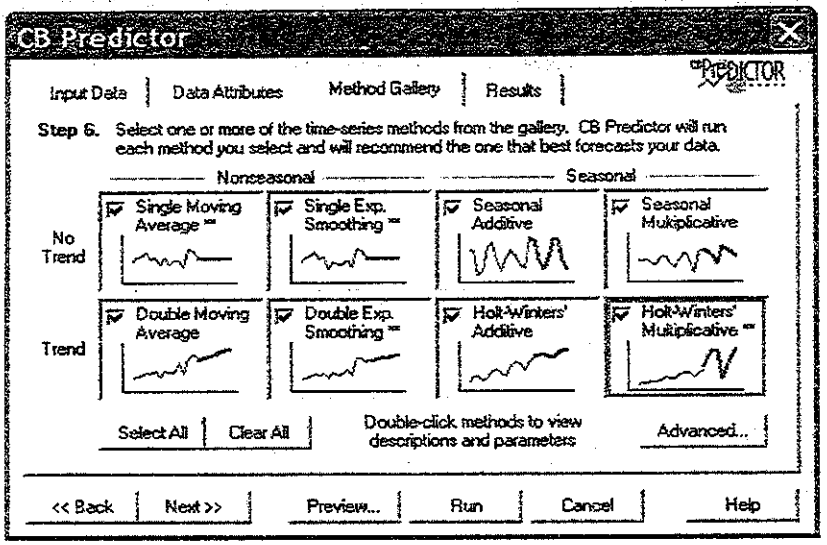
Bước 3: Vào "Run/CB Predictor…".

Bước 4: Nhập dữ liệu vào "Input Data".

Bước 5: Khai báo đặc điểm dữ liệu vào "Data Attributes".

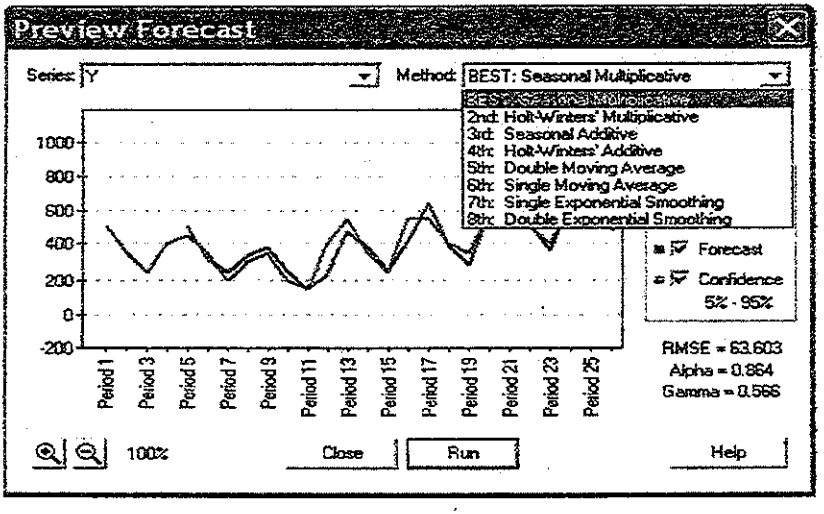
Bước 6: Chọn phương pháp dự báo trong "Method Gallery".

■ HÌNH 4.23: **8 lựa chọn phương pháp dự báo trên Crystal Ball.**



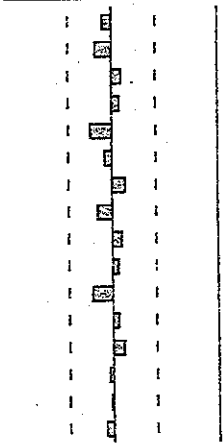
- Trong **Step 6**, chọn tất cả tám phương pháp dự báo trong Hình 4.23.
- Chọn "Next".

■ HÌNH 4.24: **Xem đồ thị dự báo trên Crystal Ball.**



Bước 7: Khai báo kết quả dự báo vào "Results" (như hướng dẫn ở mô hình trung bình giản đơn), sau đó chọn "preview", Crystal Ball sẽ đề xuất mô hình dự báo tốt nhất đối với dữ liệu sẵn có. Nếu chọn "Method" như trong Hình 4.24, chúng ta sẽ thấy thứ tự ưu tiên từ 1 đến 8 theo đề xuất của Crystal Ball. Và hiển nhiên, chúng ta nên chọn "BEST". Kết quả dự báo được trình bày ở Bảng 4.10.

■ HÌNH 4.25: **Giản đồ tự tương quan của mô hình tốt nhất.**



	AC	PAC	Q-Stat	Prob
1	-0.084	-0.084	0.1685	0.681
2	-0.160	-0.168	0.8192	0.664
3	0.090	0.063	1.0383	0.792
4	0.067	0.056	1.1641	0.884
5	-0.219	-0.191	2.6072	0.760
6	-0.080	-0.110	2.8140	0.832
7	0.121	0.040	3.3159	0.854
8	-0.140	-0.140	4.0477	0.853
9	0.084	0.124	4.3298	0.888
10	0.044	-0.021	4.4130	0.927
11	-0.189	-0.217	6.1389	0.864
12	0.062	0.074	6.3419	0.898
13	0.112	0.025	7.0967	0.897
14	-0.040	0.603	7.2061	0.926
15	-0.024	0.071	7.2529	0.950
16	-0.068	-0.239	7.6956	0.957

Như vậy, mô hình Winters không có yếu tố xu thế, thì sai số dự báo là một chuỗi ngẫu nhiên.

■ BẢNG 4.10: **Kết quả dự báo giản đơn trên Crystal Ball.**

	A	B	C	D	E	F
33		Data				
34	Date	Historical Data	Lower: 5%	Fit & Forecast	Upper: 95%	Residuals
35	Period 1	500				500
36	Period 2	350				350
37	Period 3	250				250
38	Period 4	400				400
39	Period 5	450		500		-50
40	Period 6	350		319.772994		30.22700605
41	Period 7	200		247.0556466		-47.05564659
42	Period 8	300		330.2672602		-30.26726019
43	Period 9	350		376.9483652		-26.94836523
44	Period 10	200		251.3654106		-51.36541061
45	Period 11	150		144.286188		5.713812026
46	Period 12	400		241.1622419		158.8377581
47	Period 13	550		469.7798743		80.22012569
48	Period 14	350		378.0094562		-28.00945621
49	Period 15	250		252.1767147		-2.17671469
50	Period 16	550		416.3912392		133.6087608
51	Period 17	550		646.928593		-96.92859303
52	Period 18	400		388.0802308		11.91976922
53	Period 19	350		285.4856492		64.51435077
54	Period 20	600		579.0147649		20.98523506
55	Period 21	750		703.1577256		46.84227441
56	Period 22	500		520.4745446		-20.47454457
57	Period 23	400		364.7295237		35.27047631
58	Period 24	650		662.7792943		-12.77929427
59	Period 25	850		769.1020465		80.89795353
60	Period 26		477.9128756	582.5393753	687.165875	

Ngoài ra, Crystal Ball cũng trình bày kết quả báo cáo gồm đồ thị dự báo của mô hình tốt nhất, các tiêu chí đánh giá mức độ chính xác của dự báo của cả tám mô hình dự báo giản đơn (Bảng 4.11, 4.12), và tham số dự báo của tám mô hình dự báo giản đơn (Bảng 4.13).

■ BẢNG 4.11: **Sai số dự báo của tám mô hình giản đơn.**

Method Errors:

	Method	RMSE	MAD	MAPE
Best:	Seasonal Multiplicative	63.603	49.288	11.96%
2nd:	Holt-Winters' Multiplicative	71.012	53.967	13.51%
3rd:	Seasonal Additive	71.553	59.545	13.97%
4th:	Holt-Winters' Additive	72.92	60.238	15.50%
5th:	Double Moving Average	145.94	119.73	28.04%
6th:	Single Moving Average	148.46	124	31.48%
7th:	Single Exponential Smoothing	156.4	132.76	35.27%
8th:	Double Exponential Smoothing	158.67	141.79	36.11%

■ BẢNG 4.12: **Thống kê dự báo của tám mô hình giản đơn.**

Method Statistics:

	Method	Durbin-Watson	Theil's U
Best:	Seasonal Multiplicative	1.956	0.495
2nd:	Holt-Winters' Multiplicative	1.194	0.529
3rd:	Seasonal Additive	1.743	0.437
4th:	Holt-Winters' Additive	1.474	0.505
5th:	Double Moving Average	1.616	0.756
6th:	Single Moving Average	1.333	0.781
7th:	Single Exponential Smoothing	1.351	0.825
8th:	Double Exponential Smoothing	1.675	0.91

■ BẢNG 4.13: **Tham số dự báo của tám mô hình giản đơn.**

Method Parameters:

	Method	Parameter	Value
Best:	Seasonal Multiplicative	Alpha	0.864
		Gamma	0.566
2nd:	Holt-Winters' Multiplicative	Alpha*	0.4
		Beta*	0.1
		Gamma*	0.3
3rd:	Seasonal Additive	Alpha	0.999
		Gamma	0.001
4th:	Holt-Winters' Additive	Alpha	0.529
		Beta	0.068
		Gamma	0.999
5th:	Double Moving Average	Periods	4
6th:	Single Moving Average	Periods*	5
7th:	Single Exponential Smoothing	Alpha*	0.314
8th:	Double Exponential Smoothing	Alpha*	0.4
		Beta*	0.266

*=user defined

TÓM TẮT CHƯƠNG 4

Các mô hình dự báo giản đơn đã trình bày ở trên dưới hai khía cạnh là thuật toán và ứng dụng vào các phần mềm Excel, Eviews, Crystal Ball thông qua các ví dụ cụ thể. Tính hiệu quả của các phương pháp dự báo giản đơn này là những cá nhân làm nhiệm vụ phân tích dữ liệu và dự báo ở các lĩnh vực như các doanh nghiệp thông thường, cũng như các nhà phân tích kinh tế và tài chính trong các thị trường có khả năng vừa tiếp thu và tác nghiệp chúng một cách dễ dàng, và đồng thời có khả năng đưa ra kết quả dự báo nhanh chóng trên cơ sở một bộ dữ liệu vừa phải. Mặc dù đơn giản theo tính học thuật nhưng các phương pháp dự báo giản đơn này vẫn có khả năng điều chỉnh một cách hiệu quả các yếu tố căn bản của việc phân tích dữ liệu là tính xu thế, tính mùa, và các dao động ngẫu nhiên. Theo đánh giá kinh nghiệm của các nhà phân tích dữ liệu thì các phương pháp dự báo thô này đã và đang được áp dụng phổ biến hơn là các phương pháp dự báo hàn lâm phức tạp khác ở bản chất tính thích nghi và đơn giản của chúng.

CÂU HỎI VÀ BÀI TẬP

1. Anh/Chị hãy trả lời các câu hỏi sau đây:
   a. Loại phương pháp dự báo nào sử dụng giai đoạn hiện tại làm giá trị dự báo cho giai đoạn tiếp theo?
   b. Loại phương pháp dự báo nào chủ yếu dựa vào các dữ liệu gần nhất?
   c. Loại phương pháp dự báo nào gán giá trị trọng số bằng nhau cho mỗi quan sát?
   d. Loại phương pháp dự báo nào xem nặng vai trò của quan sát hiện tại và gần hiện tại hơn so với các quan sát xa dần trọng quá khứ?
   e. Loại phương pháp dự báo nào nên được sử dụng nếu dữ liệu có yếu tố xu thế?

2. Suất sinh lời của một loại cổ phiếu ABC (ký hiệu là Y) được cho trong file **"ABC.xls"**. Anh/Chị hãy cho biết:
   a. Suất sinh lời của cổ phiếu ABC có phải là một chuỗi dừng hay không? Tại sao?
   b. Tìm giá trị dự báo của suất sinh lời hàng tháng của ABC theo phương pháp bình quân di động với hệ số trượt k = 3, từ tháng 4/2007 đến tháng 12/2007?
   c. Tìm giá trị dự báo của suất sinh lời hàng tháng của ABC theo phương pháp bình quân di động với hệ số trượt k = 5, từ tháng 6/2007 đến tháng 12/2007?
   d. Đánh giá hai mô hình dự báo trên bằng tiêu chí MAE, MSE, MAPE, và MPE?
   e. Anh chị cho biết nên chọn mô hình nào để dự báo suất sinh lời của ABC cho tháng 1/2008?
   f. Với hằng số san mũ $\alpha = 0.2$ và giá trị dự báo đầu tiên $\hat{Y}_1 = 9.29$, hãy dự báo suất sinh lời của ABC bằng phương pháp san mũ giản đơn? Anh/Chị cho biết phương pháp san mũ giản đơn này có cho kết quả dự báo tốt hơn phương pháp bình quân di động hay không?

g. Giả sử giá trị dự báo ban đầu $\hat{Y}_1 = 9.29$, theo Anh/Chị hằng số san mũ α là bao nhiêu thì kết quả dự báo cho tháng 1/2008 là tốt nhất?

3. Nhu cầu hàng tồn kho theo tháng của một công ty được ghi nhận trong năm 2007 được cho trong file **"INVENTORY.xls"**.

   a. Anh/Chị cho biết dữ liệu nhu cầu tồn kho của công ty có phải là một chuỗi dừng hay không? Tại sao?

   b. Áp dụng phương pháp san mũ giản đơn với hằng số san mũ $\alpha = 0.5$ và giá trị dự báo đầu tiên $\hat{Y}_1 = 205$ để dự báo nhu cầu tồn kho cho tháng 1/2008?

   c. Giả sử $\hat{Y}_1 = 205$, Anh/Chị cho biết hằng số san mũ α là bao nhiêu thì kết quả dự báo cho tháng 1/2008 là tốt nhất?

   d. Giả sử $\alpha = \alpha_{(c)}$, thì giá trị dự báo đầu tiên ($\hat{Y}_1$) là bao nhiêu thì kết quả dự báo tháng 1/2008 là tốt nhất?

4. Các chuyên viên phân tích của công ty quản lý quỹ đầu tư XYZ đang nghiên cứu giá trị tài sản/cổ phiếu (NAVPS, ký hiệu là Y) của các cổ phiếu Blue-Chip nhằm dự báo NAVPS cho các quý của năm 2008. Dữ liệu được tổng hợp theo quý từ năm 1997 đến 2007 được cho trong file **"XYZ.xls"**. Ông trưởng phòng phân tích yêu cầu 2 chuyên viên tiến hành dự báo và nhận được hai kết quả khác nhau như sau:

   Chuyên viên A cho rằng chỉ cần áp dụng 1 trong 3 mô hình sau đây là có thể dự báo tốt cho các quý năm 2008: Dự báo thô giản đơn, bình quân di động giản đơn với $k = 4$, san mũ giản đơn với $\alpha = 0.5$ và giá trị dự báo đầu tiên $\hat{Y}_1 = Y_1$.

   Chuyên gia B cho rằng do đây là dữ liệu theo quý và có xu hướng tăng nên phải chọn 1 trong 2 phương pháp sau: Dự báo thô có điều chỉnh theo quý, san mũ Winter với $\alpha = 0.5$, $\gamma = 0.1$, $\beta = 0.2$, $\hat{Y}_1 = Y_1$, $T_1 = 0$, và $S_1 = 1$.

   a. Anh/Chị hãy khảo sát dữ liệu và cho biết trong 5 mô hình trên thì mô hình nào có thể thích hợp nhất?

b. Theo Anh/Chị có mô hình nào tốt hơn 5 mô hình trên hay không? Trình bày kết quả dự báo của Anh/Chị?

5. HC là một công ty sản xuất xi măng có quy mô rất lớn nhưng qua khảo sát cho thấy công ty hiện đang sử dụng nhiên liệu rất lãng phí. Trước tình hình giá nhiên liệu ngày càng tăng, Ban giám đốc quyết định có sự điều chỉnh thích hợp và đề nghị phòng kinh doanh dự báo doanh thu của quý I năm 2008. Kết quả tổng hợp dữ liệu doanh thu theo tháng được cho trong file **"HC.xls"**.

a. Anh/Chị sử dụng mô hình san mũ với hằng số san mũ $\alpha = 0.4$ và giá trị dự báo đầu tiên = 77.4 (Y_1) để dự báo doanh thu theo quý cho quý I năm 2008?

b. Anh/Chị sử dụng mô hình san mũ với hằng số san mũ $\alpha = 0.6$ và giá trị dự báo đầu tiên = 77.4 (Y_1) để dự báo doanh thu theo quý cho quý I năm 2008?

c. Anh/Chị sử dụng mô hình san mũ với hằng số san mũ $\alpha = 0.4$ và giá trị dự báo đầu tiên bằng giá trị trung bình của năm đầu tiên để dự báo doanh thu theo quý cho quý I năm 2008?

d. Anh/Chị sử dụng mô hình san mũ với hằng số san mũ $\alpha = 0.6$ và giá trị dự báo đầu tiên bằng giá trị trung bình của năm đầu tiên để dự báo doanh thu theo quý cho quý I năm 2008?

e. Theo Anh/Chị có mô hình san mũ giản đơn nào khác cho kết quả dự báo tốt hơn hay không? Trình bày kết quả so sánh các mô hình trên với mô hình của Anh/Chị?

f. Với mô hình tốt nhất, Anh/Chị hãy vẽ giản đồ tự tương quan và phân tích xem sai số dự báo có phải là một chuỗi ngẫu nhiên chưa? Nếu chưa thì Anh/Chị có nhận xét gì không?

6. File **"RETAIL.xls"** là dữ liệu về doanh số theo tháng của một cửa hàng bán lẻ được tổng hợp trong hai năm 2006 và 2007.

a. Vẽ doanh số theo thời gian. Anh/Chị cho biết dữ liệu có yếu tố mùa hay không? (Lưu ý: "mùa" ở đây có nghĩa là có dao động lặp đi lặp lại theo tháng hay không)

b. Sử dụng mô hình dự báo thô giản đơn để ước lượng giá trị doanh số dự báo. Tính RMSE.

c. Sử dụng mô hình san mũ giản đơn vời hằng số mũ $\alpha = 0.5$ và giá trị dự báo đầu tiên là 430 để ước lượng giá trị doanh số dự báo? Tính RMSE.

d. Anh/Chị cho biết mô hình dự báo ở câu (b) và (c) có thể cho kết quả dự báo chính xác không? Tại sao?

e. Sử dụng Eviews và mô hình san mũ Winter nhân tính với $\alpha = \gamma = \beta = 0.5$ để ước lượng giá trị doanh số dự báo. Tính RMSE.

f. So sánh với các mô hình trên?

g. Từ sai số dự báo ở mô hình san mũ Winter nhân tính, hãy vẽ giản đồ tự tương quan với $k = 6$ và cho biết mô hình san mũ Winter nhân tính có cho kết quả dự báo tốt hơn không?

h. Anh/Chị có mô hình dự báo nào tốt hơn không? Trình bày kết quả?

7. Công ty CEC chuyên bán các sản phẩm điện (82% doanh thu), gas (13% doanh thu), và các sản phẩm khác (5% doanh thu) ở TP.HCM. Chuyên viên phân tích phòng kinh doanh yêu cầu bạn (sinh viên đang thực tập tại CEC) dự báo doanh thu theo quý cho các quý còn lại năm 2002 và bốn quý năm 2003. Chuyên viên này cung cấp dữ liệu cho bạn như trong tập tin **"CEC.xls"**. Anh/Chị hãy lựa chọn mô hình dự báo tốt nhất và dự báo doanh thu theo yêu cầu của chuyên viên phòng kinh doanh?

8. Sử dụng tập tin **"REVENUE.xls"**, Anh/Chị hãy trả lời các câu hỏi sau đây:

a. Sử dụng Crystal Ball để lựa chọn mô hình dự báo tốt nhất (trong các mô hình dự báo giản đơn) cho doanh thu quý 2/2009 của các công ty trên?

b. Sử dụng ForecastX để thực hiện lại mô hình dự báo tốt nhất cho doanh thu quý 2/2009 (từ câu a) của các công ty trên?

c. Sử dụng Eviews để thực hiện lại mô hình dự báo tốt nhất cho doanh thu quý 2/2009 (từ phần a) của các công ty trên?

d. Anh/Chị nhận xét như thế nào về kết quả dự báo của 3 phần mềm này?

9. Sử dụng tập tin **"PRICE.xls"**, Anh/Chị hãy trả lời các câu hỏi sau đây:

   a. Sử dụng Crystal Ball để lựa chọn mô hình dự báo tốt nhất (trong các mô hình dự báo giản đơn) cho giá vàng và giá dầu tháng 12/2008?

   b. Sử dụng ForecastX để thực hiện lại mô hình dự báo tốt nhất cho giá vàng và giá dầu tháng 12/2008? (từ phần a)?

   c. Sử dụng Eviews để thực hiện lại mô hình dự báo tốt nhất cho giá vàng và giá dầu tháng 12/2008? (từ phần a)?

   d. Anh/Chị nhận xét như thế nào về kết quả dự báo của 3 phần mềm này?

10. Sử dụng dữ liệu **"IMF.xls"**, Anh/Chị hãy lựa chọn mô hình giản đơn tốt nhất để dự báo kim ngạch xuất khẩu, nhập khẩu, CPI, và lãi suất tháng 12/2008?

11. Sử dụng dữ liệu **"GAS.xls"**, Anh/Chị hãy trả lời các câu hỏi sau đây:

   a. Sử dụng Crystal Ball để lựa chọn mô hình dự báo tốt nhất (trong các mô hình dự báo giản đơn) cho giá CP tháng 12/2008?

   b. Sử dụng ForecastX để thực hiện lại mô hình dự báo tốt nhất cho giá CP tháng 12/2008? (từ câu a)?

   c. Sử dụng Eviews để thực hiện lại mô hình dự báo tốt nhất cho giá CP tháng 12/2008? (từ câu a)?

   d. Anh/Chị nhận xét như thế nào về kết quả dự báo của 3 phần mềm này?

12. Sử dụng dữ liệu **"GAP.xls"**, Anh/Chị hãy trả lời những câu hỏi sau đây:

   a. Anh/Chị hãy thực hiện mô hình dự báo giản đơn theo quý cho bốn quý năm 2004? Anh/Chị nhận xét như thế nào về độ chính xác của mô hình này?

b. Sử dụng Crystal Ball để xác định mô hình dự báo giản đơn thích hợp nhất cho doanh số các quý năm 2004?

c. Sử dụng ForecastX để thực hiện lại mô hình dự báo tốt nhất cho doanh số bốn quý năm 2004 của GAP (từ phần a)?

d. Sử dụng Eviews để thực hiện lại mô hình dự báo tốt nhất cho doanh số bốn quý năm 2004 của GAP (từ phần b)?

e. Anh/Chị nhận xét như thế nào về kết quả dự báo của 3 phần mềm này?

13. Tiếp tục sử dụng tập tin **"CCC.xls"**, Anh/Chị hãy trả lời các câu hỏi sau đây:

a. Thực hiện các mô hình dự báo giản đơn trên Crystal Ball?

b. Phân tích các tiêu chí đánh giá mức độ chính xác của các mô hình này và cho biết CCC nên sử dụng mô hình nào để dự báo lượng khách hàng mới cho tháng sau?

14. Cô Julie của công ty Murphy Brothers biết rằng các quyết định kinh doanh quan trọng nhất của công ty phụ thuộc rất nhiều vào kết quả dự báo doanh số. Đối với công ty kinh doanh hàng trang trí nội thất, các dự báo doanh số ảnh hưởng đến các quyết định phát triển các dòng sản phẩm mới hoặc cắt giảm một số dòng sản phẩm lỗi thời, lập kế hoạch mua sắm, xác định hạn mức doanh số, chuẩn bị nhân sự, quảng cáo, và các quyết định tài chính. Julie biết rằng bộ phận sản xuất cần kết quả dự báo để lên kế hoạch nhân sự và mua vật tư cho một hoặc hai tháng tới. Julie cũng biết rằng Hội đồng quản trị, trong đó cha cô với vai trò Chủ tịch kiêm Giám đốc điều hành, cần kết quả dự báo để ra các quyết định đầu tư quan trọng và định hướng chiến lược cho công ty. Cô nghiên cứu thật kỹ hai bộ dữ liệu và cuối cùng thực hiện các mô hình dự báo khác nhau cho cả hai chuỗi dữ liệu. Thực hành trên Crystal Ball, Anh/Chị hãy trả lời các câu hỏi sau đây:

a. Anh/Chị cho biết trong tất cả các mô hình dự báo giản đơn, thì Julie sẽ chọn mô hình nào để dự báo doanh số toàn quốc? Tại sao?

b. Anh/Chị cho biết trong tất cả các mô hình dự báo giản đơn, thì Julie sẽ chọn mô hình nào để dự báo doanh số của công ty? Tại sao?

c. Anh/Chị dự đoán xem Julie sẽ quyết định chọn mô hình nào để dự báo doanh số của công ty trong năm 1996? Tại sao?

CHƯƠNG 5 DỰ BÁO BẰNG CÁC MÔ HÌNH XU THẾ

Chương này trình bày một phương pháp dự báo đơn giản nhưng không kém phần hiệu quả, mà người ta thường gọi nó là dự báo bằng các mô hình xu thế, hay còn gọi là phương pháp dự báo xu thế.

Mô hình nhân quả đòi hỏi chúng ta phải có các lý thuyết, các nghiên cứu trước, hay kinh nghiệm của các chuyên gia chỉ ra những yếu tố nào có thể ảnh hưởng đến biến mà chúng ta cần dự báo, rồi phải thu thập dữ liệu cho biến đó, và số quan sát cần có trong quá khứ cũng ít nhất bằng 10 lần số biến độc lập. Mô hình ARIMA cũng đòi hỏi dữ liệu trong quá khứ cần ít nhất là 50 quan sát. Trong thực tế kinh doanh ở công ty hay thực tế quản lý ở tổ chức của mình, cần phải dự báo một chỉ tiêu nào đó nhưng dữ liệu trong quá khứ có không nhiều, và cũng khó có thể thu thập được số liệu của nhiều biến khác có khả năng ảnh hưởng đến biến số cần dự báo trong giới hạn về điều kiện thời gian, kinh phí, v.v... Lúc đó, chúng ta hãy nghĩ đến phương pháp dự báo bằng các mô hình xu thế.

MỤC TIÊU HỌC TẬP

Sau khi học xong chương này, chúng ta kỳ vọng sẽ đạt được các nội dung sau đây:

- Hiểu được tổng quan về các mô hình dự báo bằng phương pháp hồi quy hàm xu thế.
- Lý giải được những trường hợp nào có thể áp dụng mô hình xu thế trong dự báo.
- Phân biệt được những dạng hàm xu thế thường sử dụng.

- Sử dụng được EViews để thực hiện dự báo bằng các mô hình xu thế.

TỔNG QUAN VỀ HÀM XU THẾ

Trước tiên, chúng ta cũng nên làm quen với một số khái niệm, một số công thức toán học có vẻ hơi khô khan một chút, nhưng đó là kiến thức nền giúp chúng ta hiểu vấn đề một cách rõ ràng hơn và có thể áp dụng vào nhiều tình huống đa dạng trong công việc.

Xu thế là sự vận động tăng hay giảm của dữ liệu trong một thời gian dài. Sự vận động này có thể được mô tả bằng một đường thẳng (*xu thế tuyến tính*) hoặc bởi một vài dạng đường cong toán học (*xu thế phi tuyến*). Có thể mô hình hoá xu thế bằng cách thực hiện một hàm hồi quy thích hợp giữa biến cần dự báo (biến Y) và thời gian (biến *t*). Sau đó, hàm hồi quy này được sử dụng để tạo ra các giá trị dự báo trong tương lai.

Khi thực hiện dự báo bằng mô hình nhân quả, người làm dự báo cần dựa trên một lý thuyết nào đó về sự ảnh hưởng của các biến độc lập (biến giải thích) lên biến phụ thuộc (biến được giải thích)[1]. Thế nhưng, phương pháp dự báo bằng mô hình hàm xu thế không cần phải dựa trên điều đó mà dựa trên một giả định rằng dạng thức vận động của dữ liệu trong quá khứ sẽ còn tiếp tục trong tương lai[2]. Nó sử dụng thời gian (biến Time) là biến giải thích, với Time bằng 1 tương ứng với quan sát đầu tiên, tăng dần theo chuỗi thời gian, và bằng *n* tương ứng với quan sát cuối cùng (Ramanathan, 2002).

NHẬN DẠNG

Giả sử chúng ta có sẵn dữ liệu của biến Y_t theo thời gian, thì làm sao chúng ta biết được xu thế trong dữ liệu sẽ tuân theo dạng hàm nào? Cách đơn giản nhất, người làm dự báo thường vẽ đồ thị của biến phụ thuộc (Y_t) theo thời gian (Time), sau đó nhận dạng xem đồ thị đó biến động gần với dạng đồ thị của hàm số tương ứng với dạng hàm toán

[1] Sẽ được trình bày ở chương 7.

[2] Giả định này có thể được hiểu là môi trường dự báo ở tương lai ít có sự thay đổi.

học nào. Dưới đây là một số dạng hàm xu hướng được sử dụng phổ biến (Bảng 5.1) và đồ thị tương ứng (Hình 5.1).

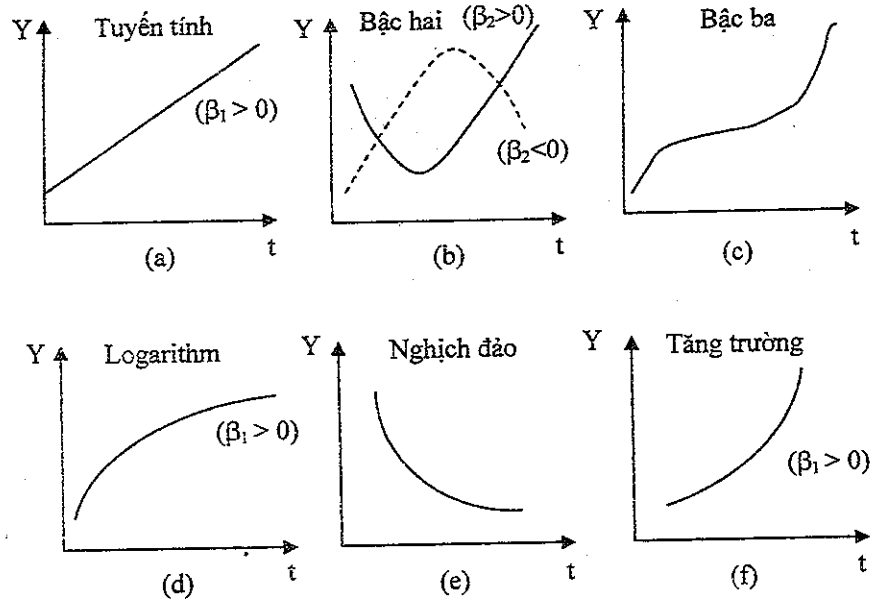
■ BẢNG 5.1: **Một số dạng hàm xu thế điển hình.**

	Dạng hàm xu thế	Phương trình hồi quy tổng thể	
A	Tuyến tính	$Y_t=\beta_0+\beta_1 Time+u_t$	(5.1)
B	Bậc hai	$Y_t=\beta_0+\beta_1 Time+\beta_2 Time^2+u_t$	(5.2)
C	Bậc ba	$Y_t=\beta_0+\beta_1 Time+\beta_2 Time^2+\beta_3 Time^3+u_t$	(5.3)
D	Tuyến tính-log	$Y_t=\beta_0+\beta_1 \ln(Time)+u_t$	(5.4)
E	Nghịch đảo	$Y_t=\beta_0+\beta_1(1/Time)+u_t$	(5.5)
F	Tăng trưởng mũ	$Y_t=e^{\beta_0+\beta_1 Time+u_t}$	(5.6)
G	Log-tuyến tính	$\ln(Y_t)=\beta_0+\beta_1 Time+u_t$	(5.7)

Ba dạng hàm đầu tiên, được gọi là các hàm đa thức. Ngoại trừ mô hình F là mô hình hồi quy phi tuyến theo các tham số, các mô hình còn lại đều là các mô hình hồi quy tuyến tính theo tham số. Người ta không ước lượng mô hình F một cách trực tiếp bằng phương pháp OLS được, mà ước lượng nó gián tiếp thông qua mô hình G. Dễ dàng nhận thấy rằng, nếu lấy ln hai vế của phương trình hồi quy ở mô hình F, sẽ có được kết quả như mô hình G.

Trong các phương trình ở trên, chúng ta gặp một số hạng được ký hiệu là u_t - sai số của mô hình. Trong các phương trình dự báo luôn có nó vì dữ liệu trong thực tế không phải lúc nào cũng hoàn toàn nằm trên đường xu thế của bạn, nói cách khác thường tồn tại một sai số. Sai số này càng nhỏ càng tốt.

HÌNH 5.1: Đồ thị một số dạng hàm xu thế điển hình.



Cũng có khi, bằng đồ thị, chúng ta chưa phân biệt được dữ liệu có xu thế tương ứng với dạng hàm toán học nào. Lúc đó, chúng ta có thể ước lượng một số mô hình mà mình cho rằng có khả năng phù hợp, sau đó kiểm định, tính toán các chỉ tiêu đo lường độ chính xác, v.v..., và chọn ra mô hình phù hợp nhất. Chúng ta cũng có thể kết hợp nhiều cách nhận diện khác nhau như quan sát đồ thị, hệ số tương quan, sai phân, hoặc tốc độ phát triển.

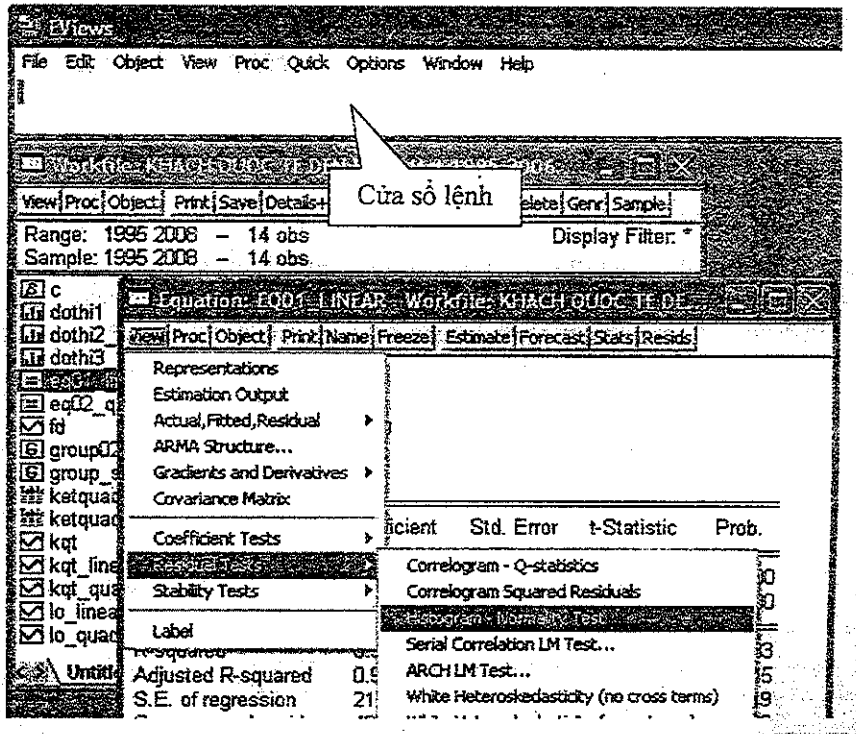
ƯỚC LƯỢNG VÀ KIỂM ĐỊNH

Các mô hình xu thế có thể là mô hình hồi quy bội, cũng có thể là mô hình hồi quy đơn. Với các mô hình xu thế tuyến tính theo tham số, chúng ta có thể dùng phương pháp OLS để ước lượng (xem Bảng 5.2). Sau đó, cần kiểm định ý nghĩa thống kê của các hệ số độ dốc, đánh giá mức độ phù hợp chung, đánh giá độ chính xác của mô hình, và dò tìm xem mô hình có bị vi phạm các giả định của phương pháp OLS không (Gaynor & Kirkpatrick, 1994).

Ba trong số các giả định của mô hình hồi quy tuyến tính cổ điển[1] là: (1) sai số dự báo tuân theo quy luật phân phối chuẩn, (2) phương sai của sai số không đổi, (3) mô hình không bị hiện tượng tự tương quan (Gaynor & Kirkpatrick, 1994). Nếu một trong số giả định này bị vi phạm, kết quả kiểm định hệ số độ dốc sẽ không còn hiệu lực nữa vì các hệ số độ dốc ước lượng sẽ bị chệch.

Với **Eviews**, chúng ta rất dễ dàng ước lượng các mô hình bằng cách gõ lệnh tương ứng vào cửa sổ lệnh; đọc bảng kết quả hồi quy (viết phương trình, kiểm định hệ số độ dốc, kiểm định mức độ phù hợp chung, v.v) từ cửa sổ **Equation**; và kiểm định các giả định của mô hình hồi quy tuyến tính theo tham số (liên quan đến phần dư), bằng cách chọn các mục phù hợp trong Menu **View\ Residual test** từ cửa sổ **Equation** (Hình 5.2).

■ HÌNH 5.2: **Cửa sổ lệnh của Eviews.**



[1] Sẽ được đề cập ở chương 7.

Trong cửa sổ lệnh này chúng ta dễ dàng thực hiện cách lệnh thường hay sử dụng trong suốt chuốn giáo trình này như hồi quy OLS (lệnh ls), vẽ đồ thị theo thời gian (lệnh plot), vẽ đồ thị phân tán (lệnh scat), vẽ đồ thị tần suất (lệnh hist), tạo một biến mới (lệnh genr hoặc series). Ngoài ra, chúng ta có thể dễ dàng thực hiện các lệnh trên bằng cách vào Quick/Estimate Equation, Quick/Graph, hoặc Quick/Series Statistics, v.v.

■ BẢNG 5.2: **Ước lượng các hàm xu thế trên Eviews.**

Phương trình hồi quy tổng thể	Các lệnh trên Eviews
$Y_t=\beta_0+\beta_1 Time + u_t$	LS Y C Time
$Y_t=\beta_0+\beta_1 Time+\beta_2 Time^2 + u_t$	LS Y C Time Time^2
$Y_t=\beta_0+\beta_1 Time+\beta_2 Time^2 + \beta_3 Time^3 + u_t$	LS Y C Time Time^2 Time^3
$Y_t=\beta_0+\beta_1 \ln(Time) + u_t$	LS Y C LOG(Time)
$Y_t=\beta_0+\beta_1 (1/Time) + u_t$	LS Y C 1/Time
$Y_t = e^{\beta_0+\beta_1 Time+u_t}$	LS LOG(Y) C Time
$\ln(Y_t) = \beta_0+\beta_1 Time + u_t$	LS LOG(Y) C Time

Ghi chú: Khi sử dụng các lệnh của Eviews, biến Y là biến phụ thuộc, Time là thứ tự thời gian hay còn gọi là biến xu thế (trend). Để tạo biến xu thế trên Eviews, chúng ta sử dụng hàng @Trend(*), trong đó, * là mốc thời gian liền trước thời điểm bắt đầu của chuỗi dữ liệu đang xét. Ngoài ra, các kiểm định cơ bản về mô hình và các hệ số hồi quy sẽ được chúng tôi trình bày một cách cụ thể ở chương 7.

THỰC HIỆN DỰ BÁO

Dự báo điểm

Dưới đây là các công thức để tính dự báo điểm đối với từng dạng mô hình[1].

[1] Ramanathan, 2002, PP.499.

■ BẢNG 5.3: **Dự báo điểm với hàm xu thế.**

Dạng hàm	Hàm hồi quy mẫu	
Tuyến tính	$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 \text{Time}$	(5.8)
Bậc hai	$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 \text{Time} + \hat{\beta}_2 \text{Time}^2$	(5.9)
Bậc ba	$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 \text{Time} + \beta_2 \text{Time}^2 + \hat{\beta}_3 \text{Time}^3$	(5.10)
Tuyến tính-log	$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 \ln(\text{Time})$	(5.11)
Nghịch đảo	$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 (1/\text{Time})$	(5.12)
Log-tuyến tính	$\ln(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 \text{Time}$	(5.13)
Tăng trưởng mũ[1]	$\hat{Y}_t = e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Time} + (\hat{\sigma}^2/2)} = e^{\ln(\hat{Y}) + (\hat{\sigma}^2/2)}$	(5.14)
	Trong đó $\hat{\sigma}^2 = \frac{\sum(\ln(Y_t) - \ln(\hat{Y}_t))^2}{n-2}$	(5.15)

Dự báo khoảng

Khoảng dự báo của năm mô hình đầu tiên trong Bảng 5.1 (từ 5.8 đến 5.12) được tính theo công thức sau:

$$[\hat{Y}_t - t_{\alpha/2,n-k}\, se(\hat{u}_t), \hat{Y}_t + t_{\alpha/2,n-k}\, se(\hat{u}_t)] \qquad (5.17)$$

Trong đó,

- $\hat{Y}_t$ là giá trị dự báo điểm tại thời điểm dự báo.
- $se(\hat{u}_t)$ là sai số chuẩn của hàm dự báo cho các giá trị cá biệt tại thời điểm dự báo t. Công thức tính toán giá trị này khá phức tạp, tuy vậy các phần mềm Eviews, hoặc SPSS đều tự động tính toán giúp chúng ta giá trị này[2].

[1] Để đơn giản, Theo Gaynor & Kirkpatrick (1994) ta có thể tính ta có thể tính giá trị dự báo điểm của hàm tăng trưởng mũ theo công thức:

$$\hat{Y}_t = e^{\hat{\beta}_0 + \hat{\beta}_1 t} \qquad (5.16)$$

[2] Chương 7 cũng sẽ đề cập chi tiết vấn đề này.

Khoảng dự báo cho mô hình tăng trưởng mũ được tính theo công thức sau[1]:

$$Exp[\ln(\hat{Y}_t) \pm t_{\alpha/2,n-2}\, s_t + \hat{\sigma}^2/2] \qquad (5.18)$$

Trong đó,

- $\hat{\sigma}^2 = \dfrac{\sum\left(\ln(Y_t)-\ln(\hat{Y}_t)\right)}{n-2}$ (5.19)
- S_t là sai số chuẩn của hàm dự báo cho các giá trị cá biệt khi dự báo $\ln(Y_t)$. S_t và $\hat{\sigma}^2$ được phần mềm máy tính tự động tính toán.
- Exp(X) là e^X

Để đơn giản hơn, khi chúng ta áp dụng dự báo điểm cho mô hình tăng trưởng mũ với công thức (5.16), thì với Eviews chúng ta sẽ dễ dàng có được giá trị này, cũng như giá trị $se(\hat{u}_t)$. Lúc đó, chúng ta có thể áp dụng công thức (5.17) để đưa ra khoảng dự báo cho mô hình tăng trưởng mũ.

Nếu mô hình thỏa mãn các giả định hồi quy tuyến tính theo tham số, chúng ta có thể trình bày cả kết quả dự báo điểm, và kết quả dự báo khoảng ở một độ tin cậy nào đó (90%, 95%, 99%) tùy theo mức ý nghĩa được lựa chọn. Ngược lại, chúng ta chỉ nên trình bày kết quả dự báo điểm (Gaynor & Kirkpatrick, 1994). Và dĩ nhiên, với kết quả dự báo điểm, chúng ta không xác định được độ tin cậy là bao nhiêu. Trong trường hợp này, có thể dạng hàm chúng ta chọn là chưa phù hợp, cũng có thể số quan sát chưa đủ lớn, hoặc phương pháp dự báo bằng hàm xu thế không thích hợp.

Đọc các phần trên, có thể chúng ta chưa hiểu lắm. Đừng lo lắng! Vì những ví dụ tiếp theo sau sẽ giúp chúng ta dễ hiểu hơn, và nó cũng giúp chúng ta biết cách thực hiện các thao tác trên Eviews trong dự báo cho các mô hình xu thế.

[1] Ramanathan, 2002, pp.252.

VÍ DỤ DẠNG HÀM XU THẾ BẬC NHẤT, BẬC HAI

Cuối mùa mưa năm 2006, một công ty du lịch ở Việt Nam muốn dự báo khách quốc tế đến Việt Nam trong tương lai để từ đó đưa ra kế hoạch kinh doanh. Chị Hương là sinh viên vừa mới ra trường và mới vào làm việc tại công ty chưa đầy một tuần lễ. Giám đốc muốn thử thách và giao cho chị công việc thật khó khăn này. Đi tìm khắp nơi, chị đã thu thập được số liệu từ Niên Giám Thống Kê, cũng như từ một số nguồn khác về khách quốc tế đến Việt Nam (biến có ký hiệu là KQT, có đơn vị là 1000 lượt người); sau đó nhập vào Eviews như Bảng 5.4 (**DATA5-1**). Và bây giờ, chị đang phân vân không biết làm gì tiếp theo. Nếu chúng ta được giao nhiệm vụ như thế, chúng ta sẽ làm gì tiếp tục?

Bên cạnh cột KQT có biến T, với T=1 ở năm đầu tiên (1995), T tăng dần ở các năm các năm tiếp theo, và bằng 12 ở năm 2006. Sau khi nhập dữ liệu cho biến KQT, biến T được tạo ra rất dễ dàng bằng cách gõ lệnh GENR T=@TREND(1994) vào cửa sổ lệnh của Eviews.

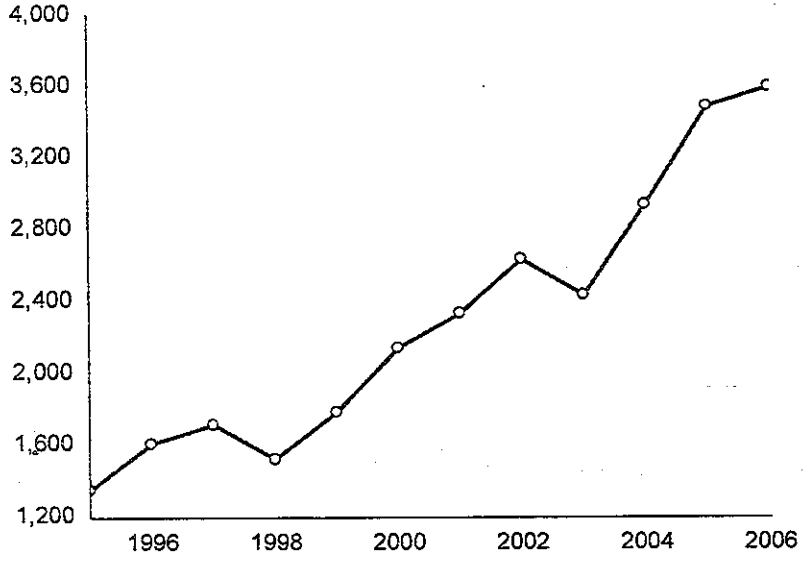
■ BẢNG 5.4: **Khách quốc tế đến Việt Nam (1000 người).**

Năm	Lượng khách quốc tế, (KQT)	T
1995	1351.3	1
1996	1607.2	2
1997	1715.6	3
1998	1520.1	4
1999	1781.8	5
2000	2140.1	6
2001	2330.8	7
2002	2628.2	8
2003	2429.6	9
2004	2927.9	10
2005	3477.5	11
2006	3583.7	12

Nguồn: Tổng Cục du lịch Việt Nam.

Để xem xét số liệu về KQT biến động theo thời gian như thế nào, cách đơn giản và thường được áp dụng nhất là đồ thị như Hình 5.3. Đồ thị này cho thấy có thể dữ liệu biến động theo thời gian dưới dạng hàm tuyến tính bậc nhất, cũng có thể là một nhánh của đường xu thế bậc hai. Chúng ta sẽ thực hiện ước lượng và kiểm định cả hai dạng hàm, sau đó sẽ so sánh xem dạng hàm nào phù hợp hơn.

■ HÌNH 5.3: **Đồ thị KQT đến Việt Nam 1995-2006.**



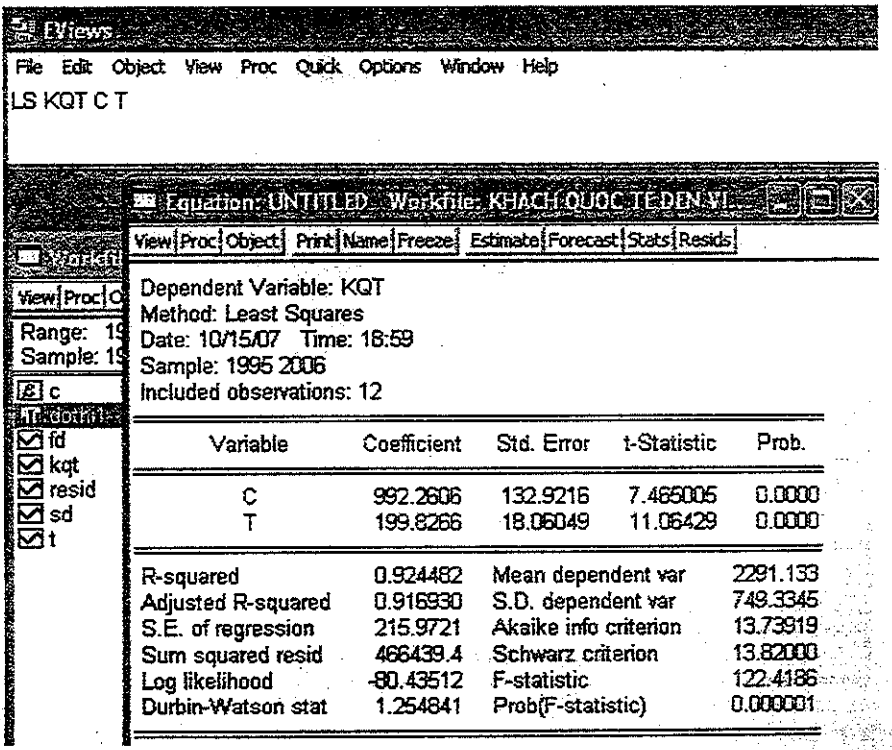
Có nhiều cách khác nhau có thể sử dụng để nhận diện dạng hàm xu thế tuyến tính. Ngoài việc sử dụng đồ thị, chúng ta có thể tính sai phân bậc 1 của biến Y (ở đây là biến KQT), sau đó xem sai phân bậc 1 này có ổn định không (có dao động nhỏ xung quanh một hằng số nào đó không). Hoặc cũng có thể sử dụng hệ số tương quan giữa Y và Time, nếu trị tuyệt đối của hệ số tương quan này lớn (thường > 0.9) và có ý nghĩa thống kê thì dạng hàm xu thế tuyến tính cũng có thể phù hợp. Phương pháp sai phân cũng thường sử dụng để nhận diện các hàm đa thức, nếu sai phân bậc 2 ổn định thì dạng hàm có thể là hàm xu thế bậc 2; nếu sai phân bậc 3 ổn định thì dạng hàm phù hợp có khả năng là dạng hàm bậc 3.

DẠNG HÀM BẬC NHẤT

Trong cửa sổ lệnh của Eviews, chúng ta gõ lệnh **LS KQT C T** (hoặc LS KQT C @Trend(1995)), rồi nhấn phím Enter để ước lượng hàm hồi quy tổng thể $KQT_t = \beta_0 + \beta_1 T + u_t$. Để đơn giản, chúng ta thay tên biến "Time" thành "T" cho các ước lượng đưới đây. Hình 5.4 thể hiện kết quả ước lượng. Từ đó, kết quả hồi quy có thể viết dạng đơn giản hơn như sau:

$$\widehat{KQT_t} = 992.26 + 199.82T$$
(t-Stat) (7.46) (11.06)
R^2=0.926, Adj R^2=0.917, ESS = 446439
DW = 1.25, n = 12

■ HÌNH 5.4: **Kết quả ước lượng hàm xu thế bậc một.**

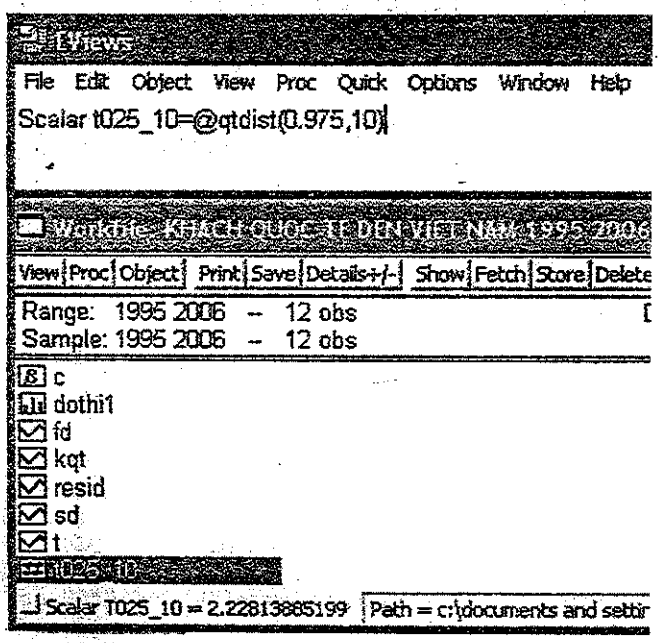


Kiểm định hệ số hồi quy

Hệ số hồi quy $\hat{\beta}_1$ có ý nghĩa thống kê ở độ tin cậy 95%, do Prob ($\hat{\beta}_1$) bằng 0.000 (nhỏ hơn 0.05).

Chúng ta cũng có thể so sánh trị tuyệt đối của thống kê t-stat($\hat{\beta}_1$) với giá trị t tra bảng ở mức ý nghĩa 5% (dùng lệnh Scalar t025_10=@qtdist(0.975,10)) để tính toán giá trị tra bảng trên Eviews và lưu giá trị tra bảng này với tên t025_10. Sau đó, chúng ta hãy nhấp vào biểu tượng T025_10 ở cửa sổ Workfile, kết quả hiển thị ở thanh trạng thái cho biết giá trị t tra bảng này là 2.228. Để đơn giản, chúng ta có thể sử dụng hàm =TINV(%5,10) trên Excel. |T-stat($\hat{\beta}_1$)| bằng 11.06 > 2.228, nên ta có thể nói rằng $\hat{\beta}_1$ có ý nghĩa thống kê ở độ tin cậy 95%.

■ HÌNH 5.5: **Tìm giá trị t tra bảng bằng Eviews.**



Đánh giá mức độ phù hợp chung của mô hình hồi quy

R^2=0.926 cho thấy 92.6% biến thiên của biến KQT được giải thích bởi mô hình. Prob (F-statistic) = 0.000 nên mô hình phù hợp với dữ liệu.

Mở rộng Workfile

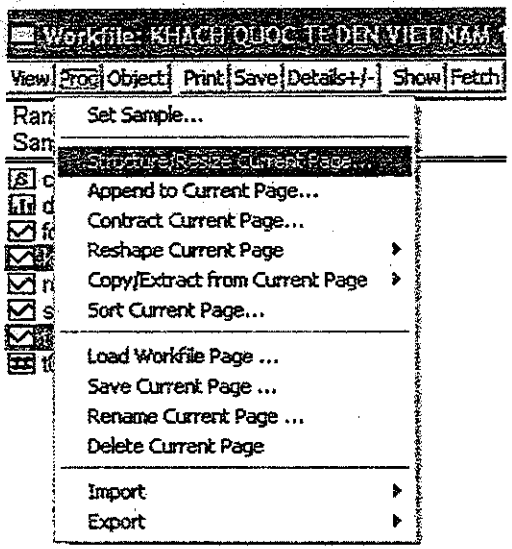
Giả sử, chúng ta muốn dự báo cho hai năm tiếp theo (2007, 2008)[1]. Kích cỡ của Workfile bây giờ chỉ là từ 1995 đến 2006 (12 quan sát), vì vậy chúng ta cần mở rộng kích cỡ của Workfile.

Từ thanh công cụ của cửa sổ Workfile, chọn **Proc/ Structure/Resize Current Page.** Hộp thoại Workfile structure, xuất hiện; trong hộp thoại này, Nhập lại năm 2008 vào ô **End date;** nhấp nút **OK.** Khi đó, máy tính sẽ hiện ra hộp thoại và thông báo "Resize involves inserting 2 observations", chúng ta hãy chọn nút **Yes.**

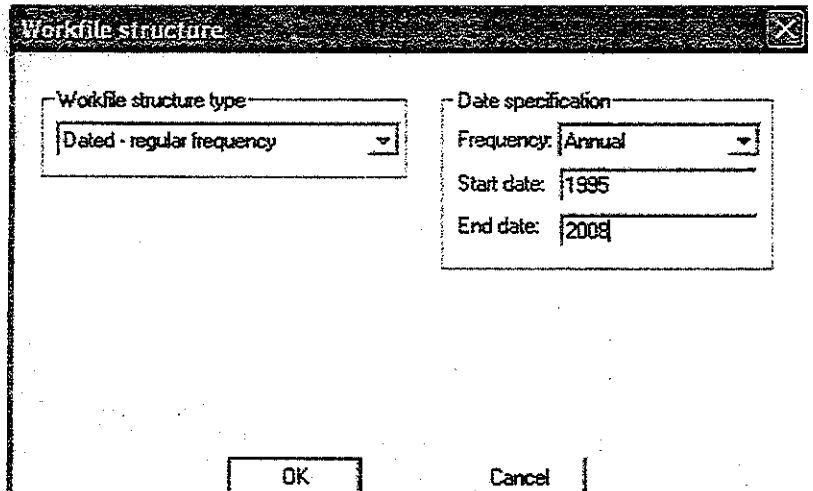
Sau khi đã mở rộng kích cỡ cho Workfile, chúng ta hãy tạo một Group cho biến KQT và biến T (Hình 5.8), sau đó nhập số 13, 14 tương ứng với giá trị của biến T ở các năm 2007, 2008 vì các giá trị này sẽ được sử dụng để thực hiện dự báo KQT cho năm 2007, 2008. (Để bật/ tắt chế độ nhập liệu trong cửa sổ Group, chúng ta phải bật/tắt nút Edit+/-).

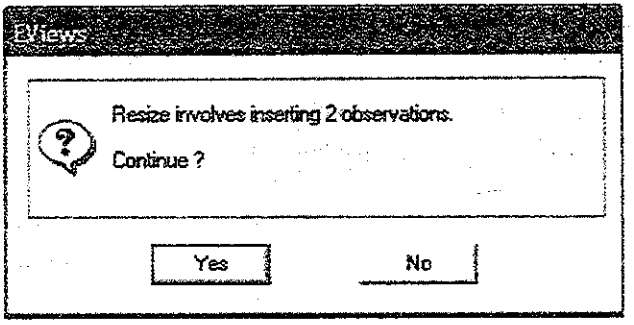
[1] Chúng ta không nên dự báo quá xa. Tầm xa của dự báo tối đa khoảng bằng 1/3 dữ liệu có được trong quá khứ. Nếu chuỗi dữ liệu trong quá khứ của chúng ta có 12 quan sát, bạn không nên dự báo quá 4 năm tiếp theo trong tương lai.

■ HÌNH 5.6: **Mở rộng tập tin sẵn có.**



■ Hình 5.7: **Cửa sổ Workfile structure.**





■ HÌNH 5.8: **Mở rộng dữ liệu thêm hai năm.**

Group: UNTITLED Workfile: KHACH QUOC TE DEN VIE

View | Proc | Object | Print | Name | Freeze | Default | Sort | Transpos

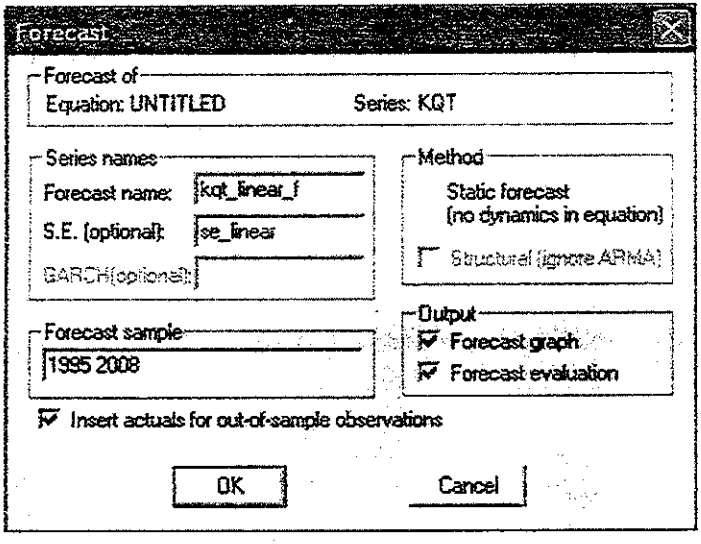
obs	obs	KQT	T
1995	1995	1351.300	1.000000
1996	1996	1607.200	2.000000
1997	1997	1715.600	3.000000
1998	1998	1520.100	4.000000
1999	1999	1781.800	5.000000
2000	2000	2140.100	6.000000
2001	2001	2330.600	7.000000
2002	2002	2628.200	8.000000
2003	2003	2429.600	9.000000
2004	2004	2927.900	10.00000
2005	2005	3477.500	11.00000
2006	2006	3583.500	12.00000
2007	2007	NA	13.00000
2008	2008	NA	14.00000

Thực hiện dự báo

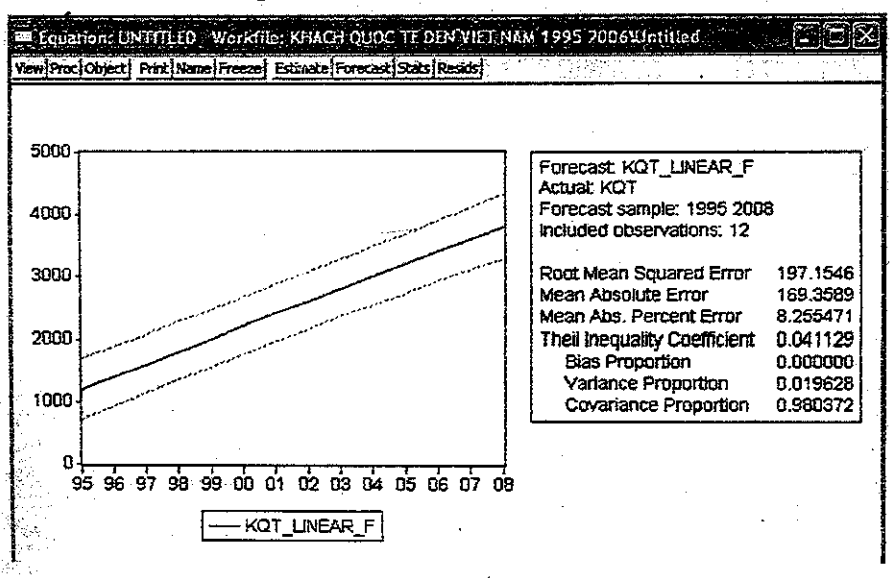
Tại cửa sổ Equation của hàm hồi quy tuyến tính (Hình 5.4), bấm nút Forecast, hộp thoại Forecast xuất hiện (Hình 5.9). Trong hộp thoại này, nhập tên biến mà chúng ta sử dụng để lưu giá trị dự báo vào khung **Forecast name,** ví dụ là biến **kqt_linear_f**; nhập tên biến mà chúng ta sử dụng để lưu giá trị sai số chuẩn của giá trị dự báo vào

khung **S.E. (Optional)**, ví dụ là biến **se_linear**, sau đó bấm **OK.** Kết quả xuất hiện như Hình 5.10.

■ HÌNH 5.9: **Thực hiện dự báo trên Eviews.**



■ HÌNH 5.10: **Kết quả dự báo trên Eviews.**



Cửa sổ Equation bây giờ sẽ hiện ra như Hình 5.10. Để lưu lại kết quả này, chúng ta hãy bấm vào nút **Freeze;** một cửa sổ mới tên là **Graph** có nội dung hoàn toàn giống như cửa sổ Equation xuất hiện; trong cửa sổ Graph, bấm Name, đặt tên cho Graph để lưu đồ thị này vào Workfile. Kết quả trên cũng cho thấy các chỉ tiêu đo lường độ chính xác của hàm xu thế tuyến tính, RMSE=197.15, MAE=169.36, MAPE=8.26, Theil's U=0.041. Có thể sử dụng các chỉ tiêu này để so sánh với các mô hình khác (các mô hình có cùng biến phụ thuộc), mô hình nào có các chỉ số này nhỏ hơn thì độ chính xác của mô hình đó tốt hơn. Ngoài ra, ở đây ta thấy Theil's U=0.041 (nhỏ hơn rất nhiều so với 0.55) nên độ chính xác của mô hình xu thế tuyến tính là rất tốt.

Nếu chúng ta quan sát cửa sổ Workfile, chúng ta sẽ thấy có hai biến mới xuất hiện **kqt_linear_f, se_linear.** Biến **kqt_linear_f** lưu giá trị dự báo điểm của mô hình. Bây giờ, nếu chúng ta muốn tính giá trị dự báo khoảng (ở độ tin cậy 95%) thì thực hiện các bước dưới đây.

- Để tính giá trị cận dưới của khoảng dự báo, rồi lưu giá trị này vào một biến nào đó, biến LO_LINEAR chẳng hạn; hãy gõ lệnh:

GENR LO_LINEAR=kqt_linear_f-@qtdist(0.975,10)*se_linear

vào cửa sổ lệnh, sau đó nhấn Enter.

Cách khác, chúng ta cũng sẽ tính tính được cận dưới này khi chọn nút Genr trong cửa sổ Workfile, sau đó gõ dòng:

LO_LINEAR=kqt_linear_f-@qtdist(0.975,10)*se_linear

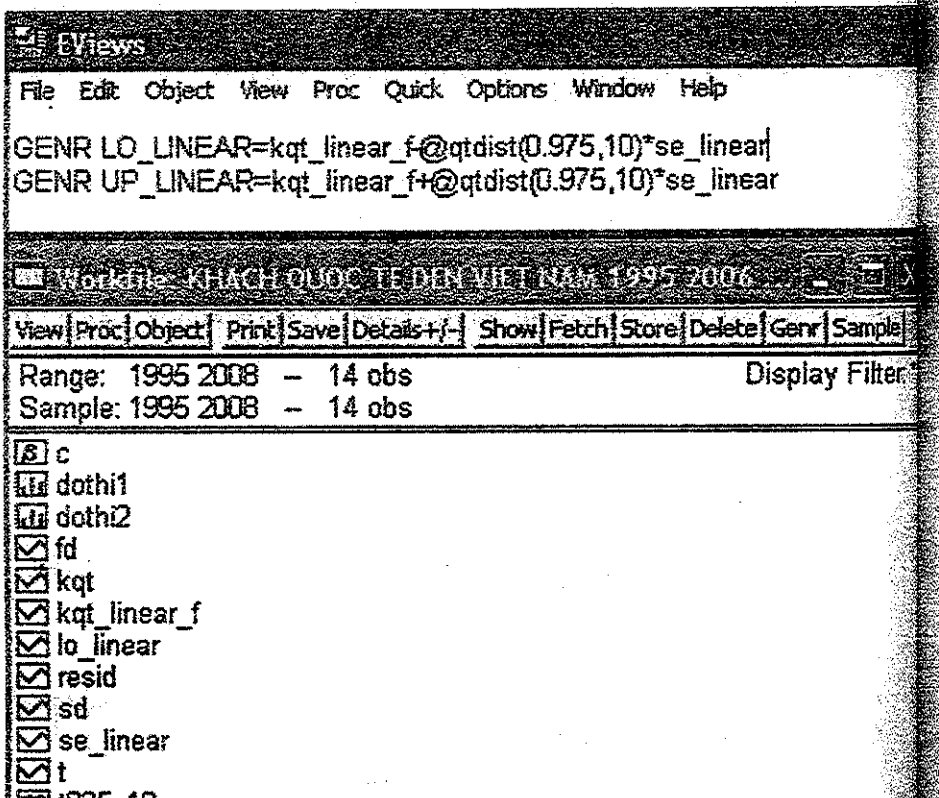
- Tương tự, tính giá trị cận trên của khoảng dự báo, gõ lệnh sau vào cửa sổ lệnh:

GENR UP_LINEAR=kqt_linear_f+@qtdist(0.975,10)*se_linear

Bây giờ trong cửa sổ Workfile, đối với giá trị dự báo điểm, chúng ta đã lưu nó trong biến **kqt_linear_f,** cận dưới và cận trên của khoảng

dự báo đã lưu trong 2 biến tương ứng là **lo_linear** và **up_linear.** Hãy mở Group cho các biến này, chúng ta sẽ thấy kết quả dự báo được tính toán trong cột dữ liệu.

■ HÌNH 5.11: **Kết quả dự báo trên Eviews.**



■ HÌNH 5.12: **Kết quả dự báo trên Eviews.**

Group: UNTITLED Workfile: KHACH QUOC TE DEN VIET NAM 1995 2006\Unti

View | Proc | Object | Print | Name | Freeze | Default | Sort | Transpose | Edit+/- | Smpl+/- | InsDel | Titl

obs	obs	KQT	T	KQT_LINEAR_F	LO_LINEAR	UP_LINEAR
1995	1995	1351.3	1	1192.1	644.5	1739.7
1996	1996	1607.2	2	1391.9	859.3	1924.5
1997	1997	1715.6	3	1591.7	1071.4	2112.0
1998	1998	1520.1	4	1791.6	1280.7	2302.4
1999	1999	1781.8	5	1991.4	1486.9	2495.9
2000	2000	2140.1	6	2191.2	1690.0	2692.5
2001	2001	2330.8	7	2391.0	1889.8	2892.3
2002	2002	2628.2	8	2590.9	2086.4	3095.4
2003	2003	2429.6	9	2790.7	2279.8	3301.6
2004	2004	2927.9	10	2990.5	2470.2	3510.8
2005	2005	3477.5	11	3190.4	2657.8	3722.9
2006	2006	3583.5	12	3390.2	2842.6	3937.8
2007	2007	NA	13	3590.0	3025.0	4155.1
2008	2008	NA	14	3789.8	3205.1	4374.6

Nếu sử dụng mô hình xu thế tuyến tính để dự báo Khách quốc tế đến Việt Nam vào năm 2008; dự báo điểm là 3789.8 nghìn lượt người; ở độ tin cậy 95%, lượt khách quốc tế đến Việt Nam vào năm 2008 có khả năng nằm trong khoảng từ 3205.1 đến 4374.6 nghìn lượt người.

Kiểm định chẩn đoán

Về mặt logic, chúng ta cần thực hiện các kiểm định này trước khi thực hiện kiểm định hệ số hồi quy, và dự báo. Nếu mô hình vi phạm các giả định của mô hình hồi quy tuyến tính (theo tham số) thì chúng ta không nên thực hiện dự báo khoảng, mà chỉ sử dụng dự báo điểm. Một cách thận trọng hơn khi thực hiện dự báo, nếu các giả định này bị vi phạm, kiểm định hệ số hồi quy mà chúng ta thực hiện phía trên cũng không đáng tin cậy, và dạng hàm của chúng ta sẽ có thể là dạng hàm khác.

Có nhiều cách khác nhau để kiểm định hiện tượng tự tương quan (tương quan chuỗi). Kiểm định BG (Hình 5.13) về tương quan chuỗi bậc 1, hoặc bậc 2 cho thấy Prob của F-statistic bằng 0.224 (lớn hơn so với 0.05), nên ở độ tin cậy 95%, mô hình không bị tương quan chuỗi (bậc 1, hoặc bậc 2). Đồ thị hệ số tự tương quan (HÌNH 5.14), tự tương

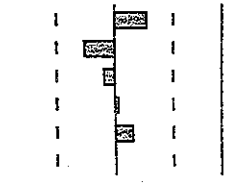
quan riêng phần cho thấy các cột autocorrelation đều nằm trong giới hạn của các đường nét đứt, và Prob của thống kê Q-Stat đều lớn 0.05 nên không có dấu hiệu về hiện tượng tự tương quan. Để kiểm định tương quan chuỗi bậc nhất, đơn giản hơn, có thể sử dụng thống kê Durbin-Watson. DW=1.25; mô hình có k'=1 (k' là số biến độc lập trong mô hình), n=12 nên tra bảng Durbin-Watson ở mức ý nghĩa 5% ta được d_L=0.971, d_U=1.331. DW nằm trong khoảng từ d_L đến d_U nên bằng cách kiểm định này chưa thể kết luận được gì về hiện tượng tương quan chuỗi, điều này có thể là do số quan sát của ví dụ này chưa đạt được bậc tự do tối thiểu là 30. Lưu ý, các kiểm định này sẽ được trình bày cụ thể hơn ở chương 7.

■ HÌNH 5.13: **Kiểm định LM của Breusch – Godfrey.**

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	1.812713	Probability	0.224246
Obs*R-squared	3.742238	Probability	0.153951

■ HÌNH 5.14: **Giản đồ tự tương quan của phần dư.**



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.305	0.305	1.4240	0.233
		2	-0.293	-0.425	2.8620	0.239
		3	-0.114	0.193	3.1047	0.376
		4	0.036	-0.158	3.1324	0.536
		5	0.175	0.310	3.8658	0.569
		6	-0.003	-0.351	3.8661	0.695

Kiểm định White cho thấy Prob của F-statistic bằng 0.824 (lớn hơn 0.05), nên mô hình có phương sai đồng nhất.

■ HÌNH 5.15: **Kiểm định phương sai thay đổi.**

White Heteroskedasticity Test:

F-statistic	0.197583	Probability	0.824179
Obs*R-squared	0.504726	Probability	0.776963

Ngoài ra, kiểm định Jarque-Bera cho thấy Prob của thống kê này bằng 0.68 (lớn hơn 0.05), nên ở độ tin cậy 95%, sai số dự báo của mô hình xấp xỉ phân phối chuẩn.

DẠNG HÀM BẬC HAI

Để ước lượng hàm xu thế bậc 2, gõ lệnh LS KQT C T T^2 vào cửa sổ lệnh (hoặc gõ LS KQT C T T*T cũng được). Kết quả ước lượng mô hình xu thế bậc hai như Hình 5.16.

$$\widehat{KQT_t} = 1383.61 + 32.10T + 12.9T^2$$

(t-Stat) (8.12) (0.53) (2.86)

R^2=0.960, Adj R^2 = 0.951, ESS = 244277

DW = 2.25, n = 12

■ HÌNH 5.16: **Ước lượng mô hình xu thế bậc hai trên Eviews.**

EViews

File Edit Object View Proc Quick Options Window Help

LS KQT C T T^2

Equation: EQ02_QUADRATIC Workfile: KHACH QUOC TE DEN VIET NAM

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: KQT
Method: Least Squares
Date: 10/16/07 Time: 01:52
Sample (adjusted): 1995 2006
Included observations: 12 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1383.614	170.2719	8.125909	0.0000
T	32.10385	60.22130	0.533098	0.6069
T^2	12.90175	4.509557	2.860979	0.0188

R-squared	0.960451	Mean dependent var	2291.133
Adjusted R-squared	0.951662	S.D. dependent var	749.3345
S.E. of regression	164.7481	Akaike info criterion	13.25903
Sum squared resid	244277.3	Schwarz criterion	13.38026
Log likelihood	-76.55418	F-statistic	109.2820
Durbin-Watson stat	2.249035	Prob(F-statistic)	0.000000

Kiểm định ý nghĩa thống kê của các hệ số hồi quy

Hệ số hồi quy $\hat{\beta}_1$ không có ý nghĩa thống kê ở độ tin cậy 95%, vì Prob ($\hat{\beta}_1$) bằng 0.607 (lớn hơn 0.05). Hệ số hồi quy $\hat{\beta}_1$ có ý nghĩa thống kê ở độ tin cậy 95%, vì Prob($\hat{\beta}_2$) bằng 0.019 (nhỏ hơn 0.05). Đối với các hàm đa thức bậc 2, hoặc bậc 3, chỉ cần có ít nhất một hệ số độ dốc có ý nghĩa thống kê là được, đặc biệt là hệ số độ dốc đứng trước biến *t* có số mũ lớn nhất. Chúng ta có thể so sánh trị tuyệt đối của thống kê *t* tính toán với $t_{0.025,9}$ và cũng sẽ rút ra kết luận tương tự.

Có hai góp ý quan trọng cho những kiểm định độ dốc trong mô hình xu thế. Thứ nhất, chúng ta thường quan tâm đến mức độ phù hợp chung của mô hình dự báo và ít để ý đến sự đóng góp cá biệt của từng biến số độc lập, công việc này cần có sự tham gia của kiểm định *F*. Thứ hai, nếu những biến độc lập có hiện tượng cộng tuyến, sẽ tạo ra những sai số chuẩn của các hệ số hồi quy không chính xác; vì vậy, có thể sẽ không có những đóng góp của từng biến độc lập đến mô hình; tuy nhiên, nếu chỉ quan tâm đến nhiệm vụ dự báo thì hiện tượng cộng tuyến trong trường hợp này không thành vấn đề (Gaynor& Kirkpatrick, 1994, 215).

Đánh giá mức độ phù hợp của mô hình hồi quy

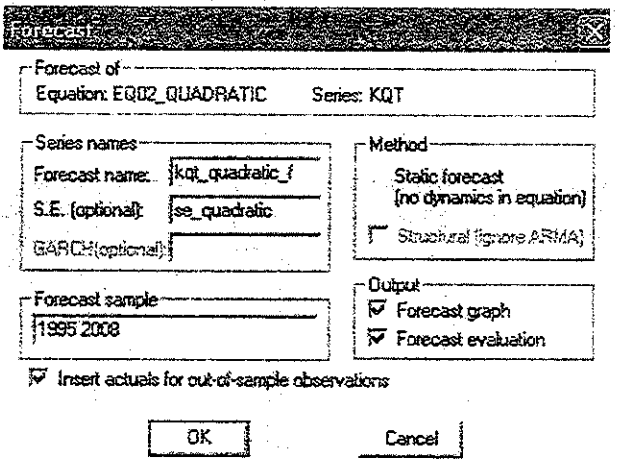
R^2=0.96. Thống kê F bằng 109.28 với Prob 0.000 (nhỏ hơn 0.05); điều này cho ta kết luận rằng mô hình phù hợp (hay thích hợp) với dữ liệu. Với hàm hồi quy bội, nên dựa vào Adjusted R^2 để nhận định mô hình giải thích được bao nhiêu phần trăm sự thay đổi của biến phụ thuộc. Trong trường hợp này, Adjusted R^2 bằng 0.952, nên nó cho biết 95% biến thiên của biến KQT được giải thích bởi mô hình. Nếu dựa vào Adjusted R^2, chúng ta cũng có thể kết luận rằng mô hình bậc hai tốt hơn so với mô hình tuyến tính bậc nhất trong việc giải thích biến thiên của biến KQT, thật vậy Adjusted R^2 của mô hình bậc nhất chỉ là 0.916 (nhỏ hơn so với Adjusted R^2 của mô hình bậc hai).

Thực hiện dự báo

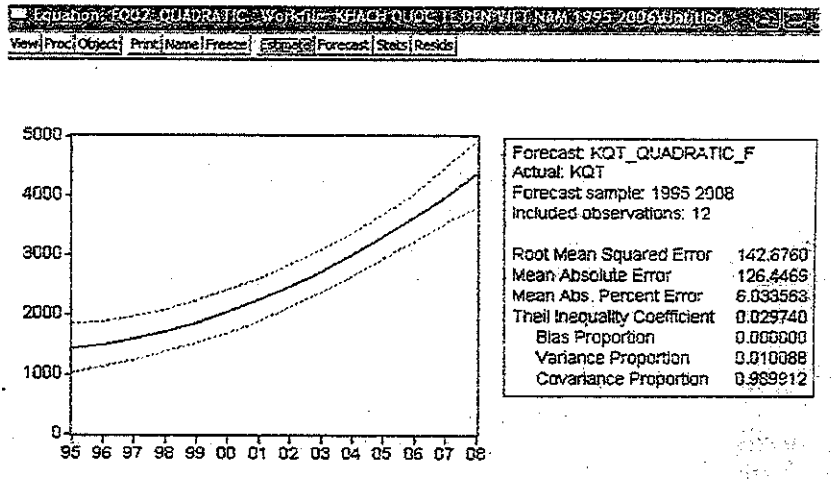
Kích cỡ Workfile bây giờ đã bao gồm cả năm 2007 và 2008, nên ta không cần phải mở rộng Workfile. Từ cửa sổ Equation của hàm bậc

hai (xem Hình 5.16) bấm nút Forecast và khai báo như Hình 5.17 rồi chọn nút **OK**. Trong hộp thoại Forecast, biến **kqt_quadratic_f** sẽ lưu giá trị dự báo điểm, còn biến **se_quadratic** sẽ lưu sai số chuẩn của giá trị dự báo. Sau khi nhấn OK, hai biến này sẽ xuất hiện trong cửa sổ Workfile, Đồ thị & các chỉ tiêu đo lường độ chính xác của mô hình bậc 2 sẽ xuất hiện như Hình 5.18.

■ HÌNH 5.17: **Thực hiện dự báo trên Eviews.**



■ HÌNH 5.18: **Kết quả ước lượng hàm xu thế bậc hai trên Eviews.**



Tính cận dưới, cận trên của khoảng dự báo của hàm bậc hai (bậc ba, hàm nghịch đảo, hàm logarithmic) cũng tương tự như đối với hàm bậc nhất. Vì vậy, chúng ta hãy lần lượt gõ hai lệnh sau vào cửa sổ lệnh của Eviews để tính toán (Hình 5.19)

GENR LO_QUADRATIC=kqt_quadratic_f-@qtdist(0.975,9)*se_quadratic

GENR UP_QUADRATIC=kqt_quadratic_f+@qtdist(0.975,9)*se_quadratic

Bây giờ, mở cửa sổ Group cho biến **kqt_quadratic_f, lo_quadratic, up_quadratic,** biến *t*, và KQT để quan sát kết quả dự báo điểm và dự báo khoảng.

■ HÌNH 5.19: **Kết quả dự báo khoảng trên Eviews.**

Group: GROUP02 QUADRATIC_F Workfile: KHACH QUOC TE DEN VIET NAM 1995 2006

View | Proc | Object | Print | Name | Freeze | Default | Sort | Transpose | Edit+/- | Smpl+/- | InsDel | Title | Sample

obs	KQT	T	KQT_QUADRATIC_F	LO_QUADRATIC	UP_QUADRATIC
1995	1351.3	1	1428.6	965.1	1892.1
1996	1607.2	2	1499.4	1078.3	1920.6
1997	1715.6	3	1596.0	1193.1	1999.0
1998	1520.1	4	1718.5	1318.6	2118.3
1999	1781.8	5	1866.7	1453.7	2269.6
2000	2140.1	6	2040.7	1634.6	2446.8
2001	2330.8	7	2240.5	1834.5	2646.6
2002	2628.2	8	2466.2	2063.2	2869.1
2003	2429.6	9	2717.6	2317.7	3117.4
2004	2927.9	10	2994.8	2591.9	3397.8
2005	3477.5	11	3297.9	2876.7	3719.0
2006	3583.5	12	3626.7	3163.2	4090.2
2007	NA	13	3981.4	3445.4	4517.3
2008	NA	14	4361.8	3721.8	5001.8

Nếu sử dụng hàm xu thế bậc hai để dự báo cho năm 2007, và 2008; bằng phương pháp dự báo điểm, khách quốc tế đến Việt Nam ở năm 2008 sẽ là 4361.8 nghìn lượt người; bằng phương pháp dự báo khoảng, ở độ tin cậy 95%, vào năm 2008, khách quốc tế đến Việt Nam có thể nằm trong khoảng từ 3721.8 đến 5001.8 nghìn lượt người.

Kiểm định chẩn đoán

Kiểm định BG về tương quan chuỗi bậc 1, hoặc bậc 2 cho thấy Prob của F-statistic bằng 0.13 (lớn hơn so với 0.05); nên ở độ tin cậy 95%,

mô hình không bị tương quan chuỗi (bậc 1, hoặc bậc 2). Chúng ta có thể sử dụng cách khác để kiểm định hiện tượng tương quan chuỗi; hệ số DW=2.249, rất gần con số 2 (nếu 1<DW<3 thì mô hình không bị hiện tượng tương quan chuỗi, đây là một quy tắc kinh nghiệm) nên mô hình không bị vi phạm hiện tượng tự tương quan. Nếu cẩn thận hơn, tra bảng Durbin-Watson tương ứng mức ý nghĩa 5% ở k'=2 (k' là số biến độc lập trong mô hình), và n =12 ta được d_L=0.812, d_U=1.579. DW nằm trong khoảng từ d_U đến 4- d_U nên mô hình không vi phạm hiện tượng tương quan chuỗi bậc 1. Qua dạng hàm bậc hai như đã thực hiện chúng ta nhận thấy rằng chỉ số DW đã được cải thiện hơn so với chỉ số DW của mô hình xu thế tuyến tính bậc một, do vậy dạng hàm dự báo xu thế bậc 2 này phù hợp hơn so với dạng hàm dự báo xu thế bậc 1.

■ HÌNH 5.20: **Kiểm định LM của Breusch – Godfrey.**

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	2.764777	Probability	0.130337
Obs*R-squared	5.295851	Probability	0.070798

Kiểm định White cho thấy Prob của F-statistic bằng 0.621 (lớn hơn 0.05), nên mô hình có phương sai đồng nhất.

■ HÌNH 5.21: **Kiểm định phương sai thay đổi.**

White Heteroskedasticity Test:

F-statistic	0.621010	Probability	0.620925
Obs*R-squared	2.266683	Probability	0.518934

Kiểm định Jarque-Bera cho thấy Prob của thống kê này bằng 0.635 (lớn hơn 0.05), nên ở độ tin cậy 95%, sai số dự báo của mô hình xấp xỉ phân phối chuẩn.

SO SÁNH VÀ LỰA CHỌN MÔ HÌNH PHÙ HỢP

Nếu chúng ta chỉ được chọn một trong hai mô hình thì chúng ta sẽ làm thế nào. Trường hợp đơn giản nhất, có một mô hình tốt (thỏa mãn các

kiểm định), và một mô hình không tốt (một số kiểm định không được thoả mãn) thì dĩ nhiên chúng ta sẽ chọn mô hình tốt. Trong ví dụ này, cả hai mô hình đều có độ chính xác tốt tốt, hệ số Theil's U của mô hình tuyến tính bậc nhất và mô hình bậc hai đều nhỏ hơn 0.55, chúng ta sẽ chọn mô hình nào?

Có hai cách thường được sử dụng. Cách một, vẽ đồ thị biểu diễn giá trị thực tế, và giá trị dự báo của cả hai mô hình lên cùng một đồ thị; sau đó ta sẽ chọn mô hình nào có đường biểu diễn của giá trị dự báo bám sát đường biểu diễn giá trị thực tế hơn. Bằng đồ thị, rất trực quan và dễ hiểu; tuy nhiên, cũng có khi chúng ta không phân biệt được đường nào bám sát đường giá trị thực tế hơn. Do vậy, cách thứ hai là so sánh các chỉ tiêu đo lường độ chính xác của mỗi mô hình, chúng ta sẽ chọn mô hình nào có độ chính xác tốt hơn.

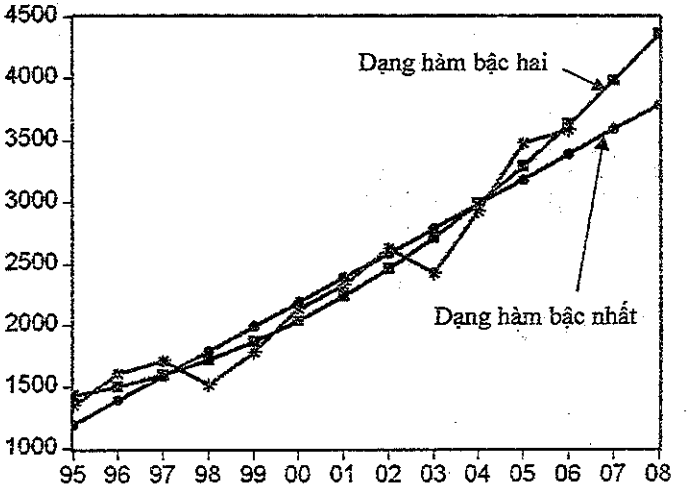
Bảng 5.5 trình bày các chỉ tiêu đo lường độ chính xác của hai mô hình. Các chỉ số đo lường độ chính xác của mô hình bậc hai nhỏ hơn so với các chỉ tiêu này của mô hình tuyến tính bậc nhất; vì vậy sử dụng mô hình bậc hai sẽ chính xác hơn.

Về mặt sư phạm là như vậy, trong thực tế một mô hình phù hợp chưa chắc đã là một mô hình có độ chính xác nhỏ nhất, bởi vì các chỉ tiêu này chỉ đo lường độ chính xác giữa giá trị dự báo và giá trị thực tế ở trong quá khứ, còn tương lai thì chưa biết được. Chúng ta cần phân tích bối cảnh của môi trường dự báo trong tương lai; nếu như nhiều thông tin cho chúng ta nhận định rằng Việt Nam sẽ có nhiều điểm du lịch hấp dẫn hơn, hình ảnh của Việt Nam trong mắt du khách hấp dẫn hơn rất nhiều khi Việt Nam đã gia nhập WTO, chất lượng phục vụ của ngành du lịch sẽ được cải tiến đáng kể so với trước đây v.v..., thì chúng ta càng có cơ sở vững hơn khi chọn hàm bậc hai để dự báo, vì hàm bậc hai cho ra kết quả dự báo lớn hơn so với kết quả dự báo từ hàm tuyến tính bậc nhất (đồ thị, và số liệu thể hiện điều này). Cũng có khi, chúng ta phải điều chỉnh tăng/giảm kết quả dự báo mà chúng ta đã chọn dựa trên ý kiến các phòng ban trong công ty, các chuyên gia; cũng có khi chúng ta sẽ kết hợp các kết quả dự báo mà chúng ta cho là tốt. Dự báo vừa là khoa học, vừa là nghệ thuật! Có nghĩa là dựa vào kết quả dự báo từ mô hình chúng ta phải tiến hành hiệu chỉnh kết quả dự báo thông qua ý kiến của các chuyên gia liên quan.

■ BẢNG 5.5: **So sánh các tiêu chí đo lường độ chính xác của các mô hình..**

Tiêu chí	Mô hình tuyến tính bậc nhất	Mô hình hàm bậc hai
ESS	466439.40	244277.30
RMSE	197.15	142.67
MAE	169.36	126.45
MAPE	8.25	6.03
Theil's U	0.04	0.03

■ HÌNH 5.21: **So sánh bằng cách so sánh các đồ thị.**



VÍ DỤ DẠNG HÀM TĂNG TRƯỞNG MŨ

Anh Dũng là chuyên viên về kế hoạch – chiến lược của một Tập đoàn đa quốc gia. Đầu mùa khô năm 2007, anh thực hiện dự báo GDP bình quân đầu người của Việt Nam (GDPPC) cho 3 năm tiếp theo (2007, 2008, 2009) nhằm phân tích môi trường bên ngoài, đồng thời làm dữ liệu đầu vào để dự báo doanh số của tập đoàn tại Việt Nam trong

tương lai, cũng như để so sánh với các thị trường khác (như Trung Quốc, Thái Lan, Ấn Độ...). Sau khi vào trang Web của Cơ quan thống kê liên hiệp quốc (http://unstats.un.org/unsd/snaama/dnllist.asp), thu thập được số liệu về GDP bình quân đầu người của Việt Nam từ năm 1970 đến nay (đơn vị tính là USD); nhưng chỉ chọn số liệu từ 1994-2006 để thực hiện mô hình dự báo, vì anh cho rằng thời kỳ này là thời kỳ mà Việt Nam mới thực sự khởi sắc so với thời kỳ của cơ chế kế hoạch hóa tập trung, và thời kỳ này ngành thống kê của Việt Nam mới áp dụng hệ thống SNA, từ đó cơ quan thống kê Liên hiệp quốc mới có được số liệu đáng tin cậy (**DATA5-2**).

Đầu tiên, một đồ thị Scatter thể hiện GDPPC theo T (là biến xu thế) được vẽ ra. Nhìn vào đồ thị này, dạng hàm có thể là hàm bậc hai, hàm bậc ba, cũng có thể là dạng hàm tăng trưởng mũ. Cách làm đối với dạng hàm bậc hai, bậc ba thật dễ dàng (vì tương tự Ví dụ ở phần 5.2), anh Dũng suy nghĩ rằng: thử ước lượng dạng hàm tăng trưởng mũ, thực hiện kiểm định và dự báo xem thế nào.

■ BẢNG 5.6: **Dữ liệu về GDP bình quân đầu người.**

Workfile:

View | Proc | Obje

Range: 1994

Sample: 1994

c
gdppc
resid
t

Untitled

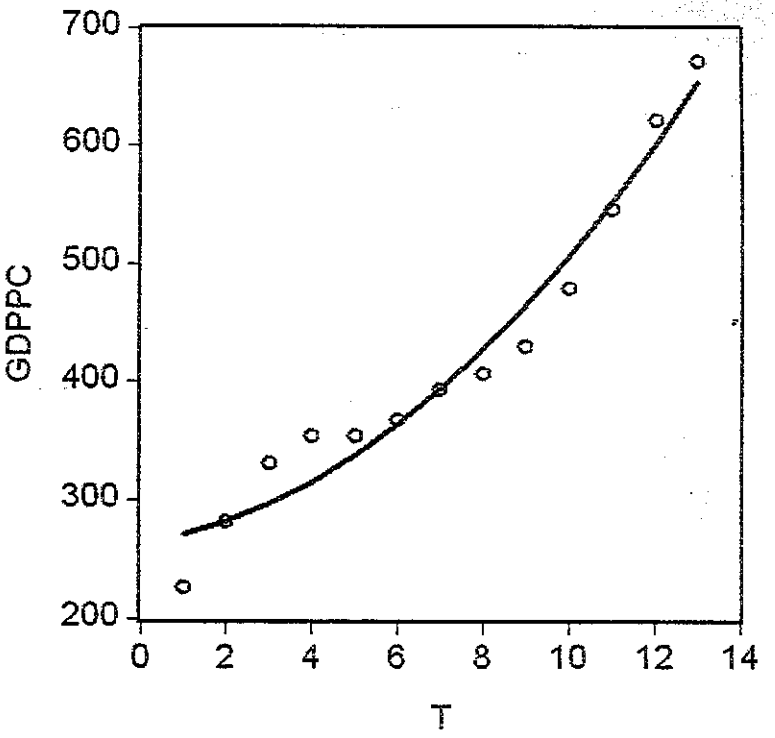
Group: UNTITLED Workfile: GDP BINH

View | Proc | Object | Print | Name | Freeze | Default

10

obs	obs	GDPPC	t
1994	1994	226.14	1
1995	1995	282.78	2
1996	1996	330.64	3
1997	1997	364.41	4
1998	1998	364.03	5
1999	1999	367.91	6
2000	2000	394.12	7
2001	2001	407.26	8
2002	2002	430.55	9
2003	2003	478.61	10
2004	2004	545.37	11
2005	2005	621.34	12
2006	2006	672.61	13
2007	2007	NA	14
2008	2008	NA	15
2009	2009	NA	16

■ HÌNH 5.22: **GDP bình quân đầu người ở Việt Nam.**



Trong ví dụ này, để ước lượng làm tăng trưởng mũ, cần phải ước lượng gián tiếp thông qua mô hình Log-tuyến tính. Tại cửa sổ lệnh chính của Eviews, chúng ta gõ lệnh:

LS LOG(GDPPC) C T

■ HÌNH 5.23: **Kết quả ước lượng GDP bình quân đầu người.**

Equation: EQ01_GROWTH Workfile: GDP BINH QUAN ...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LOG(GDPPC)
Method: Least Squares
Date: 10/16/07 Time: 06:04
Sample (adjusted): 1994 2006
Included observations: 13 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.465005	0.042958	127.2173	0.0000
T	0.076220	0.005412	14.08305	0.0000

R-squared	0.947452	Mean dependent var	5.998548
Adjusted R-squared	0.942675	S.D. dependent var	0.304955
S.E. of regression	0.073015	Akaike info criterion	-2.255675
Sum squared resid	0.058643	Schwarz criterion	-2.168760
Log likelihood	16.66189	F-statistic	198.3324
Durbin-Watson stat	0.762406	Prob(F-statistic)	0.000000

Kết quả ước lượng mô hình Log-tuyến tính sẽ như sau

$$\widehat{LN(KQT)} = 5.465 + 0.076T$$

(t-Stat) (127.22) (14.08)

$R^2 = 0.947$, Adj $R^2 = 0.943$, ESS = 0.059

DW = 0.76, n = 13

Kiểm định ý nghĩa thống kê của hệ số hồi quy

Hệ số hồi quy $\hat{\beta}_1$ có ý nghĩa thống kê ở độ tin cậy 95%, vì Prob ($\hat{\beta}_1$) bằng 0.000 (nhỏ hơn 0.05). Nếu không thích sử dụng phương pháp P-value, chúng ta có thể so sánh $|t\text{-stat}(\hat{\beta}_1)|$ với giá trị tra bảng $t_{0.025,11}$, và kết luận cũng rút ra tương tự. $|t\text{-stat}(\hat{\beta}_1)|$=14.08, lớn hơn so với 2.20 nên hệ số $\hat{\beta}_1$ có ý nghĩa thống kê.

Đánh giá mức độ phù hợp chung của mô hình hồi quy

R^2=0.947 cho thấy 97.7% biến thiên của biến LN(KQT) được giải thích bởi mô hình. Prob (F-statistic) = 0.000 nên mô hình phù hợp với dữ liệu.

Thực hiện dự báo

Từ phần lý thuyết, ta biết rằng chuyển từ hàm log-tuyến tính sang hàm tăng trưởng mũ cần theo công thức (4.14) để thu được giá trị dự báo đúng của Y.

$$\hat{Y}_t = e^{\left[\ln(\hat{Y}_t)+\frac{\hat{\sigma}^2}{2}\right]} \qquad (5.20)$$

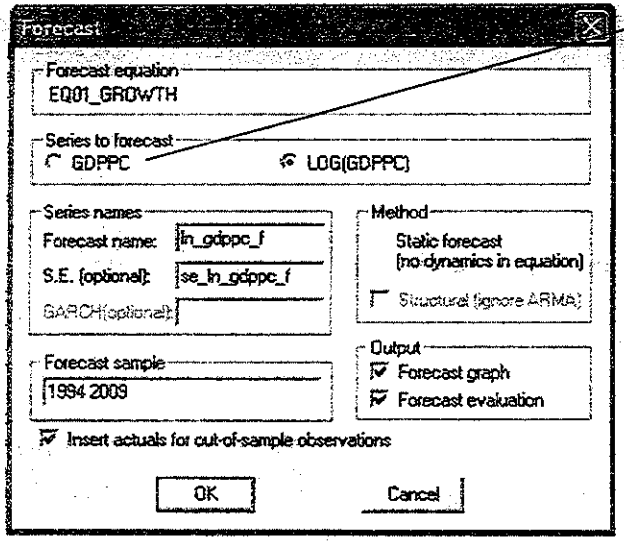
$$\hat{Y}_t = \exp[\ln(Y_t) + (\hat{\sigma}^2/2)] \qquad (5.21)$$

$$\widehat{GDPPC}=EXP[5.465 + 0.076t + 0.073^2/2]$$

Khoảng tin cậy trong việc dự báo Y là $\exp[\widehat{\ln(Y_t)} \pm t^* s_t + \hat{\sigma}^2/2]$. Trong đó t^* là giá trị tra bảng phân phối Student $t_{\frac{\alpha}{2},n-k}$ và s_t là giá sai số chuẩn của giá trị dự báo (giá trị cá biệt) khi thực hiện dự báo $\ln(Y_t)$.

Từ cửa sổ Equation của mô hình hàm Log-tuyến tính, bấm nút **forecast**, Hộp thoại Forecast xuất hiện. Khai báo như Hình 5.24.

■ HÌNH 5.24: **Dự báo GDP bình quân đầu người trên Eviews.**



Cách khác: Chúng ta có thể chọn mục này để lưu giá trị dự báo điểm gần đúng.

Khi đó, tên biến khai báo trong SE. (Optional) sẽ cho $SE_{\hat{u}_t}$ gần đúng để ta tính toán dự báo khoảng

Trong hộp thoại Forecast, đánh dấu chọn LOG(GDPPC).

Chú ý: Nếu đánh dấu chọn GDPPC thì kết quả sẽ cho ra giá trị dự báo điểm của GDPPC một cách gần đúng. Giá trị này bị thiên lệch, không nhất quán, và không hiệu quả. Tuy nhiên, cách làm này tạo thuận lợi cho người sử dụng nên có thể sử dụng. Chúng ta cũng có thể chọn mục GDPPC để có được giá trị dự báo điểm của GDPPC, và có thể sử dụng tên biến lưu sai số chuẩn của sai số dự báo trong ô **SE (Optional)** để tiếp tục dự báo khoảng.

$\widehat{Ln(gdppc)}$ sẽ được lưu trong biến ln_gdppc_f.

Sai số chuẩn của giá trị dự báo Ln(gdppc) sẽ lưu trong biến se_ln_gdppc_f.

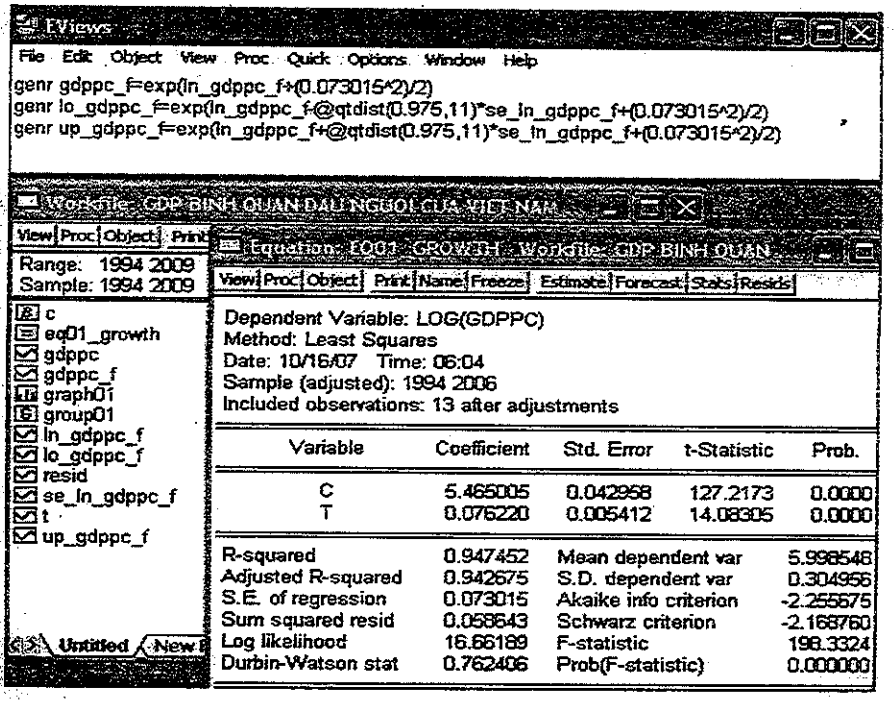
$\hat{\sigma}^2/2 = (0.073015^2)/2$

Gõ lệnh genr gdppc_f=exp(ln_gdppc_f+(0.073015^2)/2) vào cửa sổ lệnh của Eviews, sau đó nhấn Enter để tạo ra biến gdppcp_f, biến này lưu giá trị dự báo điểm đúng của gdppc.

Lần lượt gõ từng lệnh, sau đó nhấn Enter để tính cận dưới, cận trên của khoảng dự báo ở độ tin cậy 95%.

```
genr lo_gdppc_f=exp(ln_gdppc_f-qtdist(0.975,11)*se_ln_gdppc_f+(0.073015^2)/2)
genr p_gdppc_f=exp(ln_gdppc_f+@qtdist(0.975,11)*se_ln_gdppc_f+(0.073015^2)/2)
```

■ HÌNH 5.25: **Dự báo khoảng của GDP bình quân đầu người trên Eviews.**



Mở Group cho các biến gdppc_f, lo_gdppc_f, up_gdppc_f và một số biến khác để xem kết quả dự báo.

Kiểm định chẩn đoán

Chúng ta ước lượng gián tiếp mô hình tăng trưởng mũ thông qua mô hình log-tuyến tính. Vì mô hình log-tuyến tính cũng là mô hình hồi quy tuyến tính theo tham số, vì vậy nó cũng cần được kiểm định xem có bị vi phạm các các giả định của phương pháp OLS không. Kiểm

định BG, Kiểm định White, Kiểm định JB cho thấy mô hình log-tuyến tính không bị vi phạm hiện tượng tự tương quan, có phương sai của sai số đồng nhất, và sai số dự báo xấp xỉ phân phối chuẩn.

■ HÌNH 5.26: **Kiểm định LM của Breusch – Godfrey.**

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	2.629054	Probability	0.126128
Obs*R-squared	4.794143	Probability	0.090984

■ HÌNH 5.27: **Kiểm định phương sai thay đổi.**

White Heteroskedasticity Test:

F-statistic	2.417356	Probability	0.139189
Obs*R-squared	4.236770	Probability	0.120226

■ HÌNH 5.28: **Kết quả dự báo khoảng trên Eviews.**

Group: UNTITLED Workfile: GDP BINH QUAN DAU NGUOI CUA VIE...

View | Proc | Object | Print | Name | Freeze | Default | Sort | Transpose | Edit+/- | Smpl+/- | InsDel | I

obs	GDPPC	T	GDPPC_F	LO_GDPPC_F	UP_GDPPC_F
1994	226.14	1	255.67	213.25	306.54
1995	282.78	2	275.92	231.14	329.38
1996	330.64	3	297.77	250.36	354.17
1997	354.41	4	321.36	270.97	381.12
1998	354.03	5	346.81	293.04	410.44
1999	367.91	6	374.28	316.65	442.39
2000	394.12	7	403.92	341.87	477.22
2001	407.26	8	435.91	368.79	515.24
2002	430.55	9	470.43	397.50	555.75
2003	478.61	10	507.69	428.08	602.10
2004	545.37	11	547.90	460.66	651.67
2005	621.34	12	591.30	495.33	705.85
2006	672.61	13	638.13	532.24	765.08
2007	NA	14	688.67	571.52	829.82
2008	NA	15	743.21	613.33	900.59
2009	NA	16	802.07	657.82	977.95

Nếu áp dụng mô hình tăng trưởng mũ; với phương pháp dự báo GDP bình quân đầu người của Việt Nam vào năm 2009 sẽ là 802.07 USD; với phương pháp dự báo khoảng, ở độ tin cậy 95%, GDP bình quân đầu người của Việt Nam vào năm 2009 có khả năng nằm trong khoảng từ 657.82 USD đến 977.95 USD.

Đánh giá độ chính xác của mô hình dự báo

Với cách tính trên, Eviews không trực tiếp tính trực tiếp độ chính xác của mô hình tăng trưởng mũ. Vì vậy chúng ta cần tính sai số dự báo bằng cách lấy giá trị thực tế (biến **gdppc**) trừ cho giá trị dự báo (biến **gdppc_f**), sau đó tính toán các chỉ tiêu đo lường độ chính xác của mô hình. Giả sử, chúng ta cần tính chỉ tiêu ESS, MSE của mô hình dự báo, cần làm như sau.

- Tạo biến **resid_sq** để lưu giá trị bình phương của phần dư bằng cách gõ lệnh sau:

  genr resid_sq=(gdppc-gdppc_f)^2

  Hoặc là

  genr resid_sq= resid^2

  Điều này là bởi vì sau khi hồi quy thì thành phần mặc định của file Eviews sẽ chứa hiệu số của **gdppc-gdppc_f.**

- Tính các thống kê mô tả cho biến **resid_sq.**

■ BẢNG 5.7: **Mô tả phần dư.**

	RESID_SQ
Mean	664.3260
Median	845.5522
Maximum	1590.714
Minimum	6.406167
Std. Dev.	543.7713
Skewness	-0.028650
Kurtosis	1.663931
Jarque-Bera	0.968696
Probability	0.616099
Sum	8636.238
Sum Sq. Dev.	3548247.
Observations	13

MSE của mô hình tăng trưởng mũ là 664.32, còn ESS của mô hình tăng trưởng mũ là 8636.23. Các con số này có thể so sánh được với ESS, hay MSE của mô hình xu thế bậc hai, hoặc bậc 3; nếu chúng ta ước lượng các mô hình này.

Chúng ta cũng có thể vẽ giá trị thực tế (biến **gdppc**), giá trị dự báo (biến **gdppc_f**) theo thời gian lên cùng một đồ thị để đánh giá độ chính xác của dự báo.

Chúng ta cũng có thể tính hệ số tương quan của biến **gdppc và gdppc_f** (r=0.978399). Sau đó bình phương hệ số tương quan này (r^2=0.957265). Kết quả tính toán này chính là R^2 của mô hình tăng trưởng mũ. Con số này có thể được sử dụng để so sánh với R^2 của mô hình bậc hai, hay R^2 của mô hình bậc ba.

Lựa chọn các mô hình dự báo

Khi không xét đến các kiểm định khác như tương quan chuỗi, hiện tượng phương sai của sai số thay đổi, phân phối chuẩn của phần dư (bạn đọc hãy tự thực hành kiểm định các hiện tượng này thử xem); kết quả kiểm định hệ số hồi quy đối với hàm log-tuyến tính, hàm bậc hai, và hàm bậc ba đều thỏa mãn, các mô hình đều phù hợp với dữ liệu. Giả sử chúng ta chỉ căn cứ vào R-squared và ESS, mô hình nào trong

ba mô hình: tăng trưởng mũ, hàm bậc hai, hàm bậc ba sẽ có độ chính xác tốt nhất? Mô hình hàm bậc ba có R^2 lớn nhất, và có ESS nhỏ nhất, vì vậy nó nên được lựa chọn.

■ HÌNH 5.29: **Kết quả ước lượng hàm bậc hai.**

Dependent Variable: GDPPC
Method: Least Squares
Date: 10/16/07 Time: 09:52
Sample: 1994 2009 IF T<14
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	265.0938	28.02618	9.458793	0.0000
T	4.821876	9.207140	0.523711	0.6119
T^2	1.930126	0.639940	3.016102	0.0130

R-squared	0.959049	Mean dependent var	420.4449
Adjusted R-squared	0.950859	S.D. dependent var	129.1666
S.E. of regression	28.63331	Akaike info criterion	9.746193
Sum squared resid	8198.666	Schwarz criterion	9.876566
Log likelihood	-60.35025	F-statistic	117.0979
Durbin-Watson stat	0.705788	Prob(F-statistic)	0.000000

■ HÌNH 5.30: **Kết quả ước lượng hàm bậc ba.**

Dependent Variable: GDPPC
Method: Least Squares
Date: 10/16/07 Time: 09:58
Sample: 1994 2009 IF T<14
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	170.3181	20.63177	8.255138	0.0000
T	73.64708	12.28925	5.992805	0.0002
T^2	-9.916837	2.001515	-4.954664	0.0008
T^3	0.564141	0.094218	5.987625	0.0002

R-squared	0.991783	Mean dependent var	420.4449
Adjusted R-squared	0.989044	S.D. dependent var	129.1666
S.E. of regression	13.52018	Akaike info criterion	8.293903
Sum squared resid	1645.157	Schwarz criterion	8.467734
Log likelihood	-49.91037	F-statistic	362.0861
Durbin-Watson stat	1.620655	Prob(F-statistic)	0.000000

■ BẢNG 5.8: **So sánh các tiêu chí đo lường độ chính xác của các mô hình.**

	Hàm tăng trưởng mũ	Hàm bậc hai	Hàm bậc ba
R^2	0.957	0.959	0.992
ESS	8636.24	8198.67	1645.16

TÓM TẮT CHƯƠNG 5

Dự báo bằng các mô hình xu thế được sử dụng rộng rãi trong lĩnh kinh doanh, thiết lập & quản trị dự án, quản trị vận hành[1], quản trị chuỗi cung ứng và logistics[2], v.v..., vì tính đơn giản và dễ thực hiện của nó trong thực tế. Phương pháp này cũng không đòi hỏi nhiều về số quan sát; Wilson & Keating (2007) cho rằng cần ít nhất 10 quan sát khi áp dụng mô hình xu thế để dự báo. Phương pháp bự báo bằng mô hình xu thế cũng được sử dụng trong các phương pháp dự báo khác như phương pháp phân tích và dự báo bằng mô hình nhân quả mà chúng ta sẽ trình bày ở các chương tiếp theo. Trong phương pháp phân tích, chúng ta sẽ phải tách yếu tố xu thế của chuỗi thời gian bằng một hàm xu thế để dự báo yếu tố xu thế kết hợp với các yếu tố khác trong dữ liệu. Ngoài ra phương pháp dự báo bằng mô hình nhân quả, người ta có thể áp dụng phương pháp dự báo bằng các mô hình xu thế để dự báo các biến độc lập trong tương lai có trong mô hình nhân quả, các giá trị này sẽ được sử dụng làm kết quả đầu vào để dự báo biến số phụ thuộc trong tương lai. Sau cùng là: các dạng hàm toán học được sử dụng trong chương này sẽ giúp chúng ta nhận diện được xu hướng vận động của dữ liệu khi thực hiện dự báo bằng các mô hình phức tạp hơn như mô hình san bằng hàm mũ, hay mô hình ARIMA. Tuy vậy, nhược điểm của phương pháp dự báo bằng hàm xu thế là dựa trên giả định

[1] Stevenson (2005).
[2] Bowersox (2007).

rằng: quy luật vận động của dữ liệu trong quá khứ sẽ vẫn còn tiếp tục trong tương lai. Mà trong tương lai, thì bất kỳ một biến số kinh tế xã hội hay quản trị nào đều hàm chứa yếu tố rủi ro và bất định. Do vậy, để khắc phục một phần nào đó nhược điểm này, chúng ta cần phân tích môi trường dự báo để điều chỉnh kết quả dự báo theo một tỷ lệ nào đó, và sau đó cân nhắc lựa chọn dạng hàm xu thế phù hợp, và quan trọng hơn hết là các mô hình dự báo xu thế chỉ thích hợp cho các dự báo ngắn hạn.

CÂU HỎI VÀ BÀI TẬP

1. Anh/Chị cho biết biến xu thế là gì? Và làm sao để tạo các biến xu thế theo tháng, quý, và năm trên Eviews?

2. Là chuyên viên ở Phòng Tổng hợp của Sở Kế Hoạch & Đầu Tư TPHCM, anh Kiên tham gia vào việc điều chỉnh kế hoạch phát triển kinh tế - xã hội của thành phố đến 2010 và định hướng đến 2020. Công việc mà anh Kiên phải hoàn thành là dự báo về tăng trưởng Kinh tế TPHCM, cũng như xây dựng các kịch bản tăng trưởng. Có rất nhiều phương pháp dự báo mà anh đang cân nhắc. Trong các phương pháp dự báo đó, anh Kiên muốn đưa ra một kết quả dự báo GDP cho các năm 2009, 2010, 2011 bằng phương pháp xu thế (tập tin **"GDP.xls"**). Theo Anh/Chị, kết quả dự báo điểm, dự báo khoảng ở độ tin cậy 95% về GDP (theo giá so sánh 1994) của TPHCM sẽ như thế nào?

3. Tiêu dùng cuối cùng của dân cư TPHCM (tỷ đồng) được cho trong tập tin **"CONSUMPTION.xls"**. Giả sử hiện tại là năm 2006, Anh/Chị hãy thực hiện các yêu cầu sau:
   a. Tạo tập tin Eviews có thông tin của biến tieudung.
   b. Tạo biến xu thế T, với T=1 ở năm 1990, ..., T=16 ở năm 2005, T=17 ở năm 2006, T=18 ở năm 2007.
   c. Vẽ đồ thị biểu diễn tieudung theo T.

d. Ước lượng hàm xu thế tuyến tính thể hiện quan hệ của tieudung theo T với số liệu từ 1990 đến 2005. Viết phương trình hồi quy mẫu, nhận xét kết quả hàm hồi quy.

e. Thực hiện dự báo tiêu dùng cuối cùng của dân cư TPHCM trong năm 2006, 2007 bằng hàm xu thế tuyến tính (dự báo điểm, dự báo khoảng ở độ tin cậy 95%).

f. Đánh giá mức độ phù hợp của mô hình.

4. Giá trị sản xuất công nghiệp của Việt Nam theo giá hiện hành (tỷ đồng, biến GTSXCN) từ năm 1996 đến 2007 được cho trong tập tin "**IP.xls**". Giả sử, thời điểm hiện tại là đầu năm 2007; bạn hãy thực hiện các yêu cầu sau:

   a. Ước lượng mô hình xu thế dạng bậc hai và dự báo Giá trị sản xuất công nghiệp vào năm 2008, 2009?

   b. Ước lượng mô hình xu thế dạng tăng trưởng mũ và dự báo giá trị sản xuất công nghiệp của Việt Nam năm 2008, 2009?

   c. Theo Anh/Chị, trong hai mô hình xu thế dạng bậc hai và dạng tăng trưởng mũ, mô hình nào có độ chính xác hơn?

5. Diện tích trồng cây ăn quả của Việt Nam (nghìn ha) được lưu trong biến DT_CAQ như bảng trên. Dữ liệu này có sẵn trong tập tin "**AREA.xls**". Anh/Chị hãy áp dụng mô hình xu thế với dạng hàm logarithmic (mô hình tuyến tính-log) để dự báo diện tích trồng cây ăn quả của Việt Nam ở năm 2009?

6. Anh/Chị hãy vào trang Web sau của cơ quan thống kê Liên Hiệp Quốc[1]. Trang này cung cấp rất nhiều chỉ tiêu vĩ mô của các quốc gia trên thế giới từ năm 1970 đến nay. Anh/Chị hãy lựa chọn một chỉ tiêu của một quốc gia mà Anh/Chị quan tâm. Thu thập dữ liệu trong quá khứ cho chỉ tiêu đó theo từng năm (ít nhất là 10 năm gần đây), sau đó áp dụng mô hình xu thế phù hợp để dự báo chỉ tiêu này cho 3 năm tiếp theo trong tương lai.

[1] http://unstats.un.org/unsd/snaama/dnllist.asp.

7. Sử dụng tập tin **"CEC.xls"**, Anh/Chị hãy trả lời các câu hỏi sau đây:
   a. Lựa chọn mô hình hàm xu thế phù hợp để dự báo doanh thu cho các quý còn lại của năm 2002?
   b. Đánh giá kết quả dự báo và Anh/Chị cho biết kết quả dự báo bằng các mô hình hàm xu thế có tốt hơn các mô hình dự báo giản đơn ở chương 4 hay không?

8. Sử dụng tập tin **"REVENUE.xls"**, Anh/Chị hãy trả lời các câu hỏi sau đây:
   a. Lựa chọn mô hình hàm xu thế phù hợp để dự báo doanh thu của các công ty trong tập tin này cho bốn quý sau?
   b. Đánh giá kết quả dự báo và Anh/Chị cho biết kết quả dự báo bằng các mô hình hàm xu thế có tốt hơn các mô hình dự báo giản đơn ở chương 4 hay không?

9. Sử dụng dữ liệu **"GAS.xls"**, Anh/Chị hãy trả lời các câu hỏi sau đây:
   a. Lựa chọn mô hình hàm xu thế phù hợp để dự báo giá CP tháng 6/2009?
   b. Đánh giá kết quả dự báo và Anh/Chị cho biết kết quả dự báo bằng các mô hình hàm xu thế có tốt hơn các mô hình dự báo giản đơn ở chương 4 hay không?

10. Sử dụng dữ liệu **"GAP.xls"**, Anh/Chị hãy trả lời những câu hỏi sau đây:
    a. Lựa chọn mô hình hàm xu thế phù hợp để dự báo doanh số của GAP trong năm 2004?
    b. Đánh giá kết quả dự báo và Anh/Chị cho biết kết quả dự báo bằng các mô hình hàm xu thế có tốt hơn các mô hình dự báo giản đơn ở chương 4 hay không?

11. Tiếp tục sử dụng tập tin **"CCC.xls"**, Anh/Chị hãy trả lời các câu hỏi sau đây:
    a. Lựa chọn mô hình hàm xu thế phù hợp để dự báo lượng khách hàng mới của CCC trong năm 1996?

b. Đánh giá kết quả dự báo và Anh/Chị cho biết kết quả dự báo bằng các mô hình hàm xu thế có tốt hơn các mô hình dự báo giản đơn ở chương 4 hay không?

12. Tiếp tục sử dụng tập tin "**MURPHY.xls**", Anh/Chị hãy trả lời các câu hỏi sau đây:

a. Lựa chọn mô hình hàm xu thế phù hợp để dự báo doanh số bán lẻ toàn quốc trong năm 1996?

b. Lựa chọn mô hình hàm xu thế phù hợp để dự báo doanh số của công ty Murphy Brothers trong năm 1996?

c. Đánh giá kết quả dự báo và Anh/Chị cho biết kết quả dự báo bằng các mô hình hàm xu thế có tốt hơn các mô hình dự báo giản đơn hay không?

CHƯƠNG 6 DỰ BÁO BẰNG PHƯƠNG PHÁP PHÂN TÍCH

Không phải lúc nào dự báo bằng các mô hình xu thế cũng được áp dụng đơn giản như chương 5, vì các chuỗi dữ liệu được ví dụ ở trong chương 5 được giả định rằng yếu tố xu thế là nổi trội. Trong thực tế, có lúc, chúng ta muốn dự báo một chỉ tiêu nào đó cho một vài tháng, một vài quý trong tương lai. Khi đó, số liệu mà chúng ta thu thập trong quá khứ thường theo tháng, hoặc theo quý. Khi biểu diễn chuỗi dữ liệu thời gian loại này trên đồ thị, chúng ta có khả năng nhận ra rằng dữ liệu không chỉ tăng/giảm theo thời gian (yếu tố xu thế); mà nó còn có quy luật vận động lặp đi lặp lại sau mỗi năm (yếu tố mùa, hay người ta vẫn gọi là những dao động mang tính mùa vụ). Ví dụ, ở nước ta, doanh số của ngành vật liệu xây dựng thường tăng cao vào những tháng/hay những quý thuộc mùa khô, và giảm sút vào mùa mưa; hay doanh số của các doanh nghiệp thuộc lĩnh vực du lịch ở nhiều quốc gia thường tăng cao vào những tháng/quý thuộc mùa hè, v.v... Trong trường hợp như vậy, người ta nói rằng dữ liệu của chúng ta bao gồm trong đó yếu tố mùa. Ngoài ra, dữ liệu còn có hai thành phần khác là yếu tố ngẫu nhiên (bất thường), và yếu tố chu kỳ *(khi chúng ta xét một chuỗi dữ liệu dài khoảng vài chục năm)*.

MỤC TIÊU HỌC TẬP

Sau khi học xong chương này, chúng ta kỳ vọng sẽ hiểu và thực hiện được các kỹ thuật dự báo dựa trên phân tích các thành phần trong một chuỗi thời gian. Cụ thể, chúng ta có thể:

- Phân biệt được các thành phần của chuỗi thời gian.

- Phân biệt và trình bày được được mô hình cộng tính và mô hình nhân tính trong dự báo đặc biệt có yếu tố mùa nổi trội.
- Sử dụng được EViews để thực hiện dự báo bằng các phương pháp phân tích.
- Sử dụng được Excel để thực hiện dự báo bằng các phương pháp phân tích.
- Sử dụng được kiểm định Kruskal-Wallis để kiểm định yếu tố mùa bằng phần mềm Eviews.

CÁC THÀNH PHẦN CỦA CHUỖI THỜI GIAN

Các phương pháp phân tích *(Decomposition methods)* hay các mô hình phân tích chuỗi thời gian *(Time-series decomposition models)* được sử dụng cả trong dự báo ngắn hạn và dài hạn. Phương pháp này là một trong những phương pháp ra đời sớm nhất trong lịch sử của các kỹ thuật dự báo, và hiện nay vẫn còn được sử dụng phổ biến ở các nước phát triển (Wilson & Keating, 2007). Trong chương này, chúng ta chỉ chú trọng vào các dự báo ngắn hạn. Cũng có nhiều khác biệt trong các phương pháp phân tích chuỗi thời gian, và cách thức mà chúng ta bàn tới trong chương này là phân tích chuỗi thời gian cổ điển *(Classical times-series decomposition)* – cách thức thực hiện chủ yếu dựa trên nền tảng của các phương pháp trung bình di động và dự báo theo hàm xu thế mà chúng ta vừa đề cập ở các chương 4 và chương 5. Tuy nhiên, các phương pháp có tính đến sự kết hợp cộng tính hay kết hợp nhân tính với yếu tố mùa vụ mà chúng ta đã đề cập trong mô hình san bằng hàm mũ Winters.

BỐN THÀNH PHẦN CỦA CHUỖI THỜI GIAN

Như đã giới thiệu ở chương 3, một chuỗi thời gian thường bao gồm 4 thành phần khác biệt về bản chất. Đó là thành phần xu thế *(Trend component)*, thành phần chu kỳ *(Cyclical component)*, thành phần mùa *(Seasonal component)* và thành phần bất thường/ngẫu nhiên *(Irregular/Random component)*.

Xu thế (Trend). Xu thế là thành phần thể hiện sự tăng (hoặc giảm) ẩn bên trong của một chuỗi thời gian. Xu thế có thể được tạo ra do sự thay đổi dân số liên tục, lạm phát, thay đổi công nghệ, tăng năng suất. Thành phần này thường được ký hiệu là **Tr**, hay **T**.

Chu kỳ (Cyclical). Thành phần chu kỳ là một chuỗi những sự dao động giống như hình sóng và sự dao động này sẽ lặp lại sau một thời kỳ thường dài hơn một năm. Nói chung, chu kỳ được tạo ra do sự thay đổi của các điều kiện kinh tế, ví dụ sau 10 năm thì suy thoái nền kinh tế sẽ lập lại. Người ta thường ký hiệu thành phần chu kỳ là **Cl**, hay **C**.

Trong thực tế, Chu kỳ thường khó xác định và thường được xem như là một phần của yếu tố xu thế. Trong trường hợp này, thành phần thể hiện sự tăng (hoặc giảm) ẩn bên trong của một chuỗi dữ liệu được gọi là thành phần ***Xu thế - Chu kỳ*** (Trend-Cycle) và cũng được ký hiệu là Tr hay T (Hanke & Wichern, 2005). Khi đó, một chuỗi thời gian sẽ bao gồm 3 thành phần là Tr, Sn, Ir.

Mùa (Seasonal). Những dao động mùa vụ rất thường được tìm thấy với dữ liệu theo quý, theo tháng, hoặc thậm chí theo tuần. Nếu chỉ có dữ liệu theo năm thì không có biến động mùa. Sự dao động mùa vụ liên quan đến kiểu thay đổi khá ổn định xuất hiện hàng năm và kiểu thay đổi đó lại được lặp lại ở năm sau, và các năm sau nữa. Yếu tố mùa xảy ra do ảnh hưởng của thời tiết, các sự kiện trong năm liên quan đến lịch như nghỉ hè, ngày lễ. Thành phần này thường được ký hiệu là **Sn**, hay **S**. Mùa và chu kỳ đều là quy luật dao động của dữ liệu có tính chất lặp lại. Nếu như mùa là quy luật diễn ra giữa các thời điểm trong năm thì chu kỳ là quy luật diễn ra trong khoảng thời gian dài vài năm đến chục năm, với tần suất quan sát là năm, và chuỗi thời gian phải đủ dài thì mới có thể phát hiện ra quy luật chu kỳ.

Ngẫu nhiên/bất thường (Irregular). Thành phần ngẫu nhiên bao gồm những thay đổi ngẫu nhiên, hay không dự đoán được. Những sự thay đổi bất thường là kết quả của vô số những sự kiện mà nếu xét riêng lẻ thì không quan trọng gì, còn nếu kết hợp các sự kiện riêng lẻ đó lại thì có thể tạo ra một ảnh hưởng lớn. Thành phần bất thường này xuất hiện có thể do ảnh hưởng của tin đồn, thiên tai, động đất, nội chiến, khủng bố, v.v... Người ta thường ký hiệu thành phần ngẫu nhiên là **Ir**, hay **I**.

Trong bốn thành phần của chuỗi thời gian nói trên thì các mô hình dự báo chỉ có thể tập trung tìm ra các thành phần xu thế, mùa vụ. Thành phần chu kỳ cần có một chuỗi dữ liệu lưu trữ ít nhất là trên 30 năm, còn các dao động khác thường thì không thể nào dự báo được. Do vậy, phương pháp dự báo phân tích chỉ chủ yếu đề cập hai thành phần là xu thế và mùa vụ và cố gắng tìm ra những cách thức kết hợp của hai thành phần này nhằm phục vụ cho nhu cầu dự báo chuỗi thời gian.

Khi nghiên cứu các thành phần của một chuỗi thời gian. Nhà phân tích phải xem xét các thành phần này liên quan như thế nào với chuỗi dữ liệu gốc (biến Y). Có 2 mô hình thể hiện mối quan hệ này: (1) Mô hình nhân tính (*Multiplicative Components Model)* xem các giá trị của một chuỗi thời gian (biến Y) được tạo thành bởi tích số của từng thành phần Tr, Cl, Sn, Ir; và (2) Mô hình cộng tính *(Additive Components Model)* xem các giá trị của một chuỗi thời gian (biến Y) được tạo thành bởi tổng của các thành phần Tr, Cl, Sn, Ir.

Mô hình nhân tính: $Y_t = Tr_t . Cl_t . Sn_t . Ir_t$

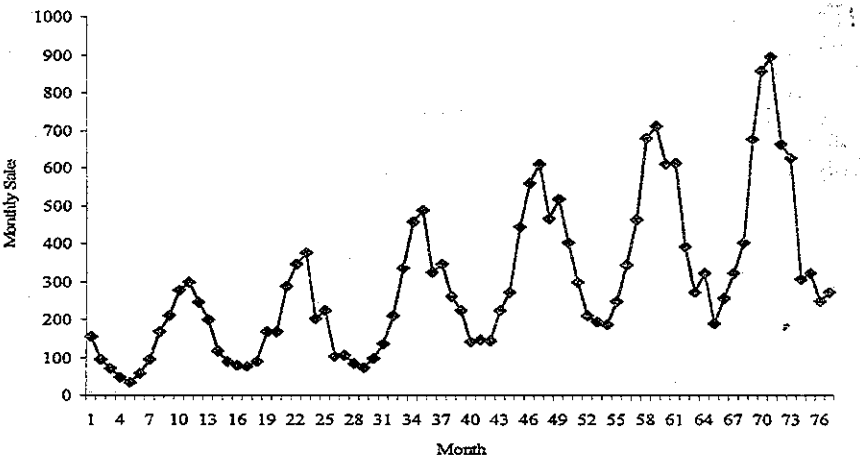
Mô hình cộng tính: $Y_t = Tr_t + Cl_t + Sn_t + Ir_t$

Mô hình nhân tính sẽ phù hợp khi sự biến thiên của chuỗi thời gian tăng dần theo thứ tự của thời gian. Điều này có nghĩa là các giá trị của chuỗi trải rộng ra khi xu thế tăng dần, và tập hợp các quan sát có dạng hình cái loa *(megaphone)*, hay hình cái phễu *(funnel)*.

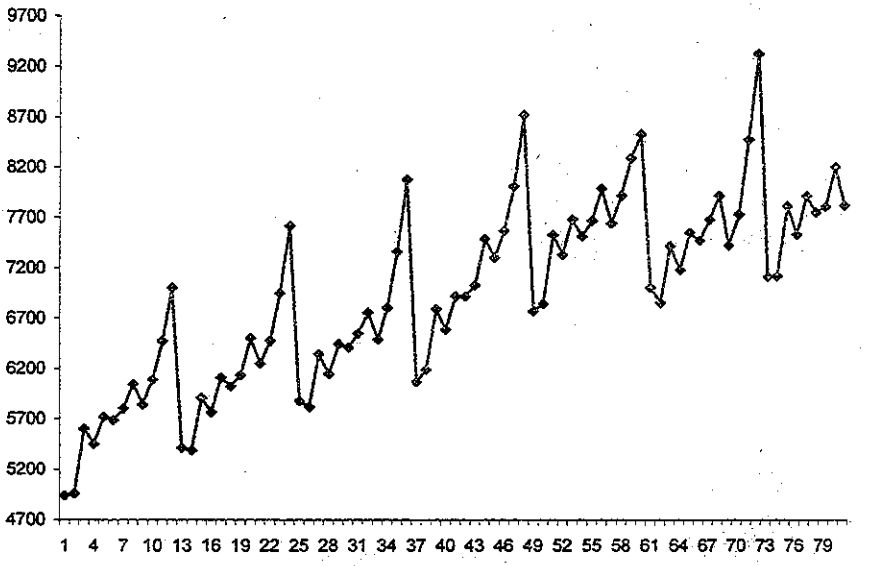
Mô hình cộng tính có hiệu quả khi chuỗi dữ liệu đang được phân tích có sự biến thiên xấp xỉ đều nhau suốt độ dài của chuỗi thời gian. Điều này có nghĩa là các giá trị của chuỗi thời gian về cơ bản nằm trong một dải giá trị có độ rộng là một hằng số và trung tâm của dải này là đường xu thế.

Hình 6.1 và Hình 6.2 thể hiện một chuỗi thời gian có sự biến thiên tăng dần theo thứ tự thời gian và một chuỗi thời gian có sự biến thiên không đổi (mô hình cộng tính). Cả hai hình này đều thể hiện dữ liệu theo từng tháng với xu thế tăng dần và có yếu tố mùa.

■ HÌNH 6.1: **Mô hình nhân tính.**



■ HÌNH 6.2: **Mô hình cộng tính.**



DỮ LIỆU ĐƯỢC ĐIỀU CHỈNH YẾU TỐ MÙA

Thành phần xu thế là sự vận động trong một thời gian dài mà có thể được mô tả bởi một đường thẳng, hay đường cong.

Nếu xu thế là xấp xỉ tuyến tính, tức là vận động tăng hoặc giảm theo dạng đường thẳng, thì thành phần xu thế có thể được thể hiện bởi phương trình sau:

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 T \qquad (6.1)$$

Với T = 1, 2, 3... tương ứng với từng thời kỳ của các quan sát trong chuỗi dữ liệu.

Xu thế cũng có thể dưới các dạng bậc hai, bậc ba, tăng trưởng mũ, v.v..., (xem chương 5).

Khi dữ liệu có yếu tố mùa, trước tiên ta cần tách yếu tố mùa ra khỏi chuỗi dữ liệu, sau đó mới sử dụng chuỗi dữ liệu được điều chỉnh yếu tố mùa để thực hiện dự báo xu thế. Có rất nhiều phương pháp để tách yếu tố mùa ra khỏi dữ liệu ban đầu: Census X12, X11 (Historical), Tramo/Seats, và các phương pháp trung bình di động *(Moving Avearage Methods)*. Dễ nhất là sử dụng các phương pháp trung bình động, nhưng bản thân trong phương pháp này cũng có rất nhiều cách khác nhau. Chúng ta sẽ tìm hiểu cách thức mà phần mềm Eviews sử dụng để thuận tiện trong việc thực hành[1]. Với mô hình nhân tính, chúng ta sử dụng Tỷ lệ trung bình di động *(Ratio to Moving Average)*. Với mô hình cộng tính, chúng ta sử dụng chênh lệch so với trung bình di động *(Difference from moving average)*.

Tỷ lệ trung bình di động – mô hình nhân tính

Các bước thực hiện như sau:

(1) Đầu tiên là tính toán trung bình trung tâm (CMA_*Centered Moving Average*) của Y_t:

- $CMA_t = (0.5Y_{t+6} + \ldots + Y_t + \ldots + 0.5Y_{t-6})/12$ nếu số liệu theo tháng.

[1] Xem Eviews 6 User's Guide (2007).

- $CMA_t = (0.5Y_{t+2} + Y_{t+1} + Y_t + Y_{t-1} + 0.5Y_{t-1})/4$ nếu số liệu theo quý.

Về mặt ý nghĩa, CMA_t bao gồm yếu tố Xu thế và chu kỳ kết hợp lại. Và như vậy trong mô hình dự báo nhân tính thì $CMA_t = Tr_t.Cl_t$

(2) Tính tỷ lệ $\tau_t = Y_t/CMA_t$

Ta biết rằng, trong mô hình nhân tính $Y_t = Tr_t.\ Cl_t.\ Sn_t.\ Ir_t$ nên

$$\tau_t = Sn_t.\ Ir_t$$

(3) Tính toán các chỉ số mùa vụ (The Seasonal Indices):

- Ở chuỗi dữ liệu theo tháng, chỉ số mùa i_m cho tháng m bằng trung bình của τ_t với các quan sát chỉ cho tháng m (mỗi năm có một tháng m).
- Ở chuỗi dữ liệu theo quý, chỉ số mùa i_q cho quý q bằng trung bình của τ_t với các quan sát chỉ cho quý q (mỗi năm có một quý q).

(4) Sau đó, chúng ta cần điều chỉnh các chỉ số mùa để tích của chúng bằng một. Điều này được thực hiện bằng cách tính các nhân tố mùa *(The Seasonal Factors)*. Nhân tố mùa Sn là tỷ số của chỉ số mùa và trung bình nhân của các chỉ số:

- $Sn = \dfrac{i_m}{\sqrt[12]{i_1 i_2 i_3 \cdots i_{12}}}$ nếu dữ liệu theo tháng
- $Sn = \dfrac{i_q}{\sqrt[4]{i_1 i_2 i_3 i_4}}$ nếu dữ liệu theo quý

Các Sn được Eviews báo là các ***Scaling Factors*** trong cửa sổ series và được lưu lại thành một biến nếu ta đặt tên cho biến này trong hộp chọn. Sn ở thời điểm t nào đó sẽ cho biết: Ở giai đoạn t, chuỗi Y_t cao hơn Sn_t % so với chuỗi dữ liệu đã hiệu chỉnh yếu tố mùa.

(5) Chuỗi dữ liệu đã hiệu chỉnh yếu tố mùa (The Seasonally Adjusted Series) có được bằng cách chia Y_t cho nhân tố mùa Sn_t.

$$Y_t/Sn_t=Tr_t.Cl_t.Ir_t \quad (6.2)$$

Để thuận tiện, ta giả định rằng không có yếu tố chu kỳ, và yếu tố ngẫu nhiên bị triệt tiêu khi chúng ta tính trung bình nhằm tìm ra chỉ số mùa τ_t ở bước 3. Khi đó $Cl_t=1$, và $Ir_t=1$. Lúc đó Chuỗi dữ liệu đã hiệu chỉnh yếu tố mùa chỉ còn lại yếu tố xu thế Tr_t. Người ta sẽ sử dụng chuỗi Y_t/Sn_t để dự đoán thành phần xu thế trong tương lai.

Chênh lệch so với trung bình di động – mô hình cộng tính

Trong mô hình cộng tính, các bước tính toán để có được chuỗi dữ liệu điều chỉnh yếu tố mùa cũng tương tự như với mô hình nhân tính.

(1) Tính toán trung bình trung tâm (CMA_Centered Moving Average) của Y_t:

- $CMA_t = (0.5Y_{t+6} +...+ Y_t +...+0.5Y_{t-6})/12$ nếu số liệu theo tháng.
- $CMA_t = (0.5Y_{t+2} + Y_{t+1} + Y_t + Y_{t-1} +0.5Y_{t-1})/4$ nếu số liệu theo quý.

(2) Tính sự khác biệt $d_t=Y_t-CMA_t$

(3) Tính toán các chỉ số mùa *(The Seasonal Indices)*.

- Ở chuỗi dữ liệu theo tháng, chỉ số mùa i_m cho tháng m bằng trung bình của d_t với các quan sát chỉ cho những tháng m (mỗi năm có một tháng m).
- Ở chuỗi dữ liệu theo quý, chỉ số mùa i_q cho quý q bằng trung bình của d_t với các quan sát chỉ cho những quý q (mỗi năm có một quý q).

(4) Chúng ta điều chỉnh các chỉ số mùa để tổng của chúng bằng không. Điều này được thực hiện bởi thiết lập $Sn_t=i_t-\bar{i}$; với $\bar{i}$ là

trung bình của tất cả các chỉ số mùa. Các Sn_t được Eviews báo cáo là các ***Scaling Factors***. Sn_t cho biết, ở thời đoạn *t*, Y cao hơn (hay thấp hơn) một lượng Sn_t so với chuỗi dữ liệu đã điều chỉnh yếu tố mùa.

(5) Chuỗi dữ liệu đã điều chỉnh yếu tố mùa có được bằng cách lấy Y_t-Sn_t

Trong mô hình cộng tính, $Y_t = Tr_t + Cl_t + Sn_t + Ir_t$, nên chuỗi dữ liệu đã được điều chỉnh yếu tố mùa được tính bởi công thức:

$$Y_t - Sn_t = Tr_t + Cl_t + Ir_t \qquad (6.3)$$

Để đơn giản, trong dự báo ngắn hạn ta giả định rằng không có yếu tố chu kỳ, và yếu tố ngẫu nhiên đã bị triệt tiêu khi tính toán trung bình nhằm tìm ra chỉ số mùa vụ ở bước 3. Khi đó $Cl_t = 0$, và $Ir_t = 0$. Khi đó chuỗi Y_t-Sn_t chỉ còn lại yếu tố xu thế.

DỰ BÁO VỚI MÔ HÌNH NHÂN TÍNH

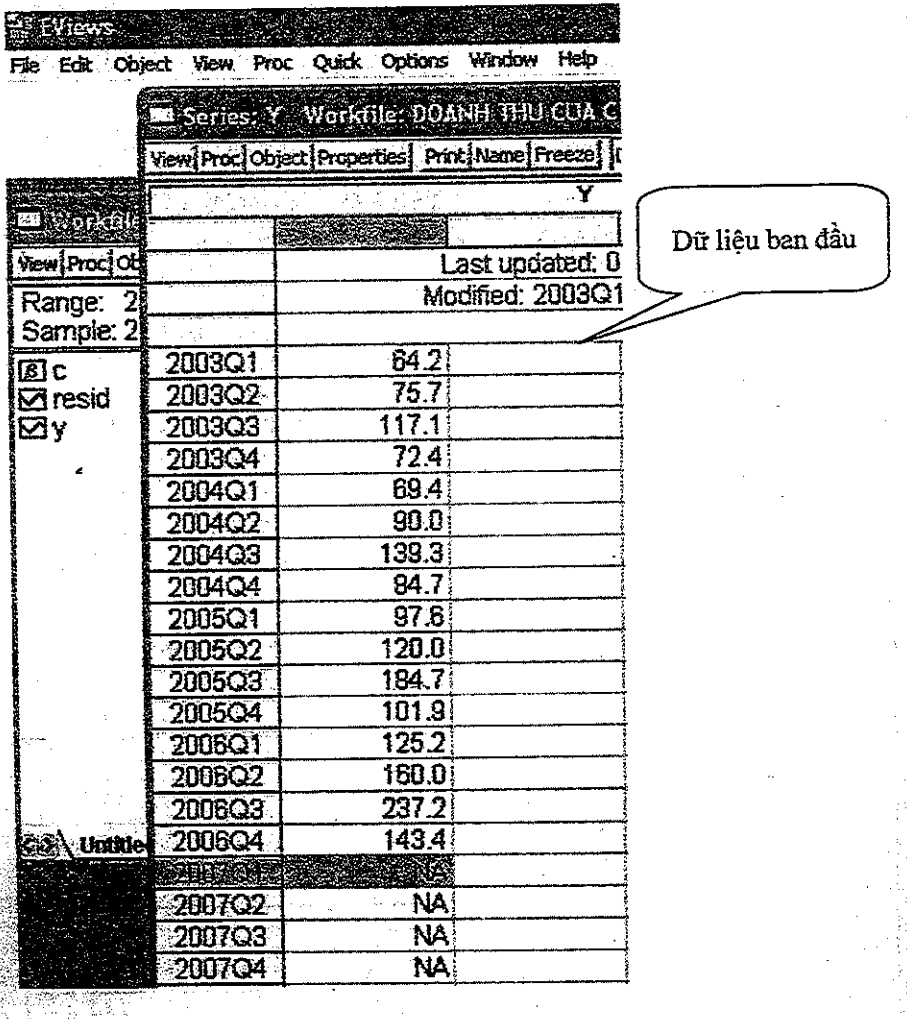
Chị Oanh đang làm việc tại phòng Nghiên cứu & Phát triển của một công ty du lịch lữ hành của Việt Nam, có trụ sở tại TPHCM. Hiện nay, là thời điểm đầu năm 2007. Chị Oanh muốn dự báo doanh thu của công ty ở các quý năm 2007 để lập kế hoạch kinh doanh. Dữ liệu được thu thập từ quý 1 năm 2003 đến quý 4 năm 2006 như Bảng 6.1 (**DATA6-1**). Trong bảng này, Y là doanh thu của công ty (tỷ đồng).

■ BẢNG 6.1: **Doanh thu của công ty.**

Năm	Quý	Y	Năm	Quý	Y
2003	1	64.2	2005	1	97.6
2003	2	75.7	2005	2	120.0
2003	3	117.1	2005	3	184.7
2003	4	72.4	2005	4	101.9
2004	1	69.4	2006	1	125.2
2004	2	90.0	2006	2	160.0
2004	3	139.3	2006	3	237.2
2004	4	84.7	2006	4	143.4

Khi phân tích môi trường kinh doanh, Oanh thấy rằng năm 2007 có khả năng môi trường kinh doanh không có biến động lớn, và thậm chí có nhiều dấu hiệu khởi sắc hơn năm 2006 do Việt Nam mới gia nhập WTO. Với dữ liệu như vậy, theo bạn chị Oanh có thể áp dụng mô hình nhân tính để dự báo doanh thu hay không? Kết quả dự báo doanh thu ở các quý năm 2007 sẽ như thế nào?

■ HÌNH 6.3: **Nhập dữ liệu vào Eviews.**

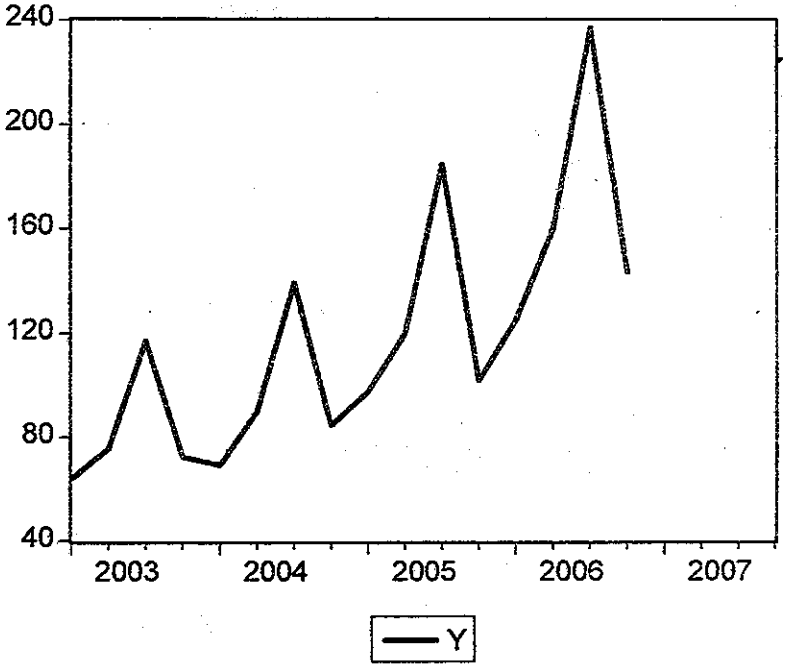


Dữ liệu ban đầu được nhập như Hình 6.3, biến Y là doanh thu của công ty.

Bước 1. Nhận dạng

Việc đầu tiên, chúng ta sẽ vẽ đồ thị của Y theo thời gian để xem chuỗi dữ liệu này có yếu tố xu thế, có yếu tố mùa v.v..., hay không? Tại cửa sổ Series: Y, chọn View\Graph\Line sẽ được đồ thị như Hình 6.4.

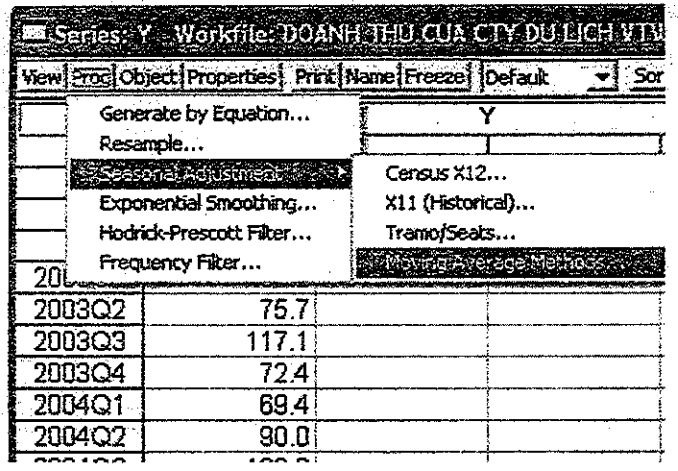
■ HÌNH 6.4: **Đồ thị doanh số theo quý.**



Hình 6.4 cho thấy, doanh thu của công ty có xu hướng tăng dần, có thể là xu thế tuyến tính; có yếu tố mùa: trong mỗi năm, doanh thu của công ty thường cao vào quý 3. Và yếu tố mùa ảnh hưởng càng mạnh theo thời gian. Như vậy, bằng trực quan, ta thấy mô hình nhân tính sẽ phù hợp hơn so với mô hình cộng tính.

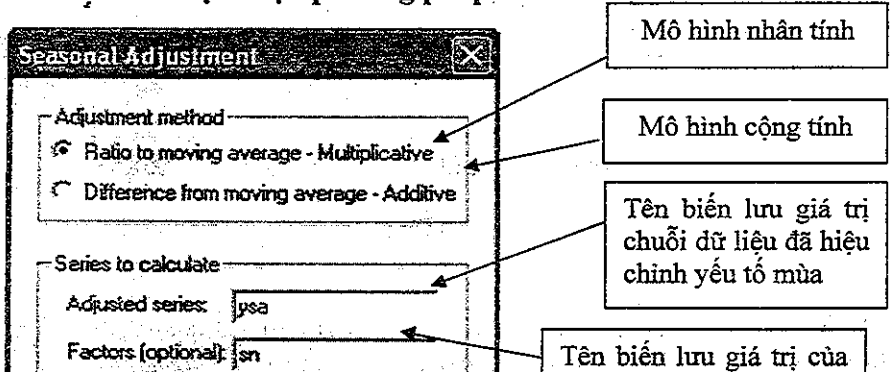
Bước 2. Tách yếu tố mùa

■ HÌNH 6.5 **Tách yếu tố mùa vụ.**



- Tại cửa sổ Series của biến Y. Chọn **Proc\Seasonal Adjustment\ Moving Average Methods** (Hình 6.5)

■ HÌNH 6.6: **Lựa chọn phương pháp.**



- Khi cửa sổ như Hình 6.6 xuất hiện, nhập tên chuỗi dữ liệu được hiệu chỉnh yếu tố mùa (được loại ra thành phần sai số) trong mục ***Adjusted series***. Nhập tên biến lưu ***nhân tố mùa*** trong mục ***Factors (optional).***
- Kết quả sẽ có 2 biến mới được tạo ra. Các Scaling Factors cũng được liệt kê như Hình 6.8. Chúng ta sẽ thấy nhân tố mùa của từng quý có tích bằng 1. (0.840x1.000x1.454x0.817=1).

■ HÌNH 6.7: **Mô hình nhân tính.**

Series: Y Workfile: DOANH THU CUA CTY DU

View | Proc | Object | Properties | Print | Name | Freeze | Sample |

Date: 06/12/09 Time: 17:03
Sample: 2003Q1 2007Q4
Included observations: 16
Ratio to Moving Average
Original Series: Y
Adjusted Series: YSA

Scaling Factors:	
1	0.840735
2	1.000564
3	1.454990
4	0.817026

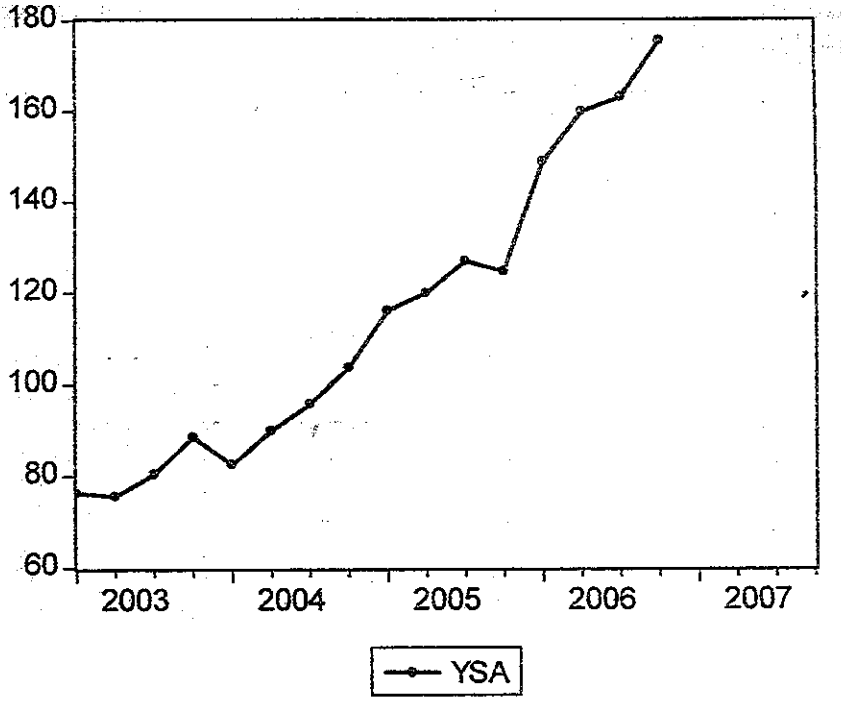
Mở cửa sổ Group cho các biến Y, SN, YSA ta sẽ có kết quả như Bảng 6.2. Cột SN cho thấy nhân tố mùa. Cột YSA chính là chuỗi dữ liệu đã được hiệu chỉnh yếu tố mùa. YSA=Y/SN. Chúng ta sẽ sử dụng cột YSA để dự báo xu thế trong tương lai. Bởi vì yếu tố chu kỳ xét trong ngắn hạn xem như không có: Cl=1. Yếu tố ngẫu nhiên bị triệt tiêu sau khi lấy trung bình khi tính ra chỉ số mùa vụ, do vậy: Ir=1. Nếu có yếu tố ngẫu nhiên thì trong chuỗi YSA ta có thể loại quan sát bất thường, rồi thay bằng trung bình cộng của hai quan sát liền kề nhau. Hay nói cách khác YSA=Tr. Đây là một chuỗi dữ liệu thực tế, ta cần tìm ra xu thế trong chuỗi này để từ đó dự báo cho tương lai.

■ BẢNG 6.2: **Kết quả dự báo mô hình nhân tính.**

Quý	Y	SN	YSA
2003Q1	64.20	0.84	76.36
2003Q2	75.70	1.00	75.66
2003Q3	117.10	1.45	80.48
2003Q4	72.40	0.82	88.61
2004Q1	69.40	0.84	82.55
2004Q2	90.00	1.00	89.95
2004Q3	139.30	1.45	95.74
2004Q4	84.70	0.82	103.67
2005Q1	97.60	0.84	116.09
2005Q2	120.00	1.00	119.93
2005Q3	184.70	1.45	126.94
2005Q4	101.90	0.82	124.72
2006Q1	125.20	0.84	148.92
2006Q2	160.00	1.00	159.91
2006Q3	237.20	1.45	163.03
2006Q4	143.40	0.82	175.51
2007Q1		0.84	
2007Q2		1.00	
2007Q3		1.45	
2007Q4		0.82	

Bước 3. Ước lượng hàm xu thế và dự báo

■ HÌNH 6.8: **Đồ thị ước lượng yếu tố xu thế.**



- Nhìn đồ thị, ta thấy chuỗi dữ liệu đã loại yếu tố mùa có thể là dạng hàm tuyến tính, bậc 2, hoặc hàm tăng trưởng mũ. Giả sử ta chọn hàm tăng trưởng mũ (ta có thể thử nhiều hàm và chọn ra hàm phù hợp nhất).

$$\hat{Y}_t = e^{\hat{\beta}_0 + \hat{\beta}_1 T} \qquad (6.4)$$

- Tạo biến thứ tự thời gian t, bằng cách gõ vào cửa sổ lệnh:

  genr t=@trend(2002:4)

- Gõ vào cửa sổ lệnh dòng LS LOG(YSA) C T

Kết quả hồi quy như sau:

■ HÌNH 6.9: **Kết quả ước lượng yếu tố xu thế.**

Equation: UNTITLED Workfile: DU LIEU CTY DU LICH VT 2003 200...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LOG(YSA)
Method: Least Squares
Date: 08/12/09 Time: 23:36
Sample (adjusted): 2003Q1 2006Q4
Included observations: 16 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.198907	0.025540	164.4053	0.0000
T	0.058953	0.002641	22.32002	0.0000

R-squared	0.972666	Mean dependent var	4.700011
Adjusted R-squared	0.970714	S.D. dependent var	0.284591
S.E. of regression	0.048703	Akaike info criterion	-3.089692
Sum squared resid	0.033207	Schwarz criterion	-2.993118
Log likelihood	26.71753	F-statistic	498.1832
Durbin-Watson stat	1.492009	Prob(F-statistic)	0.000000

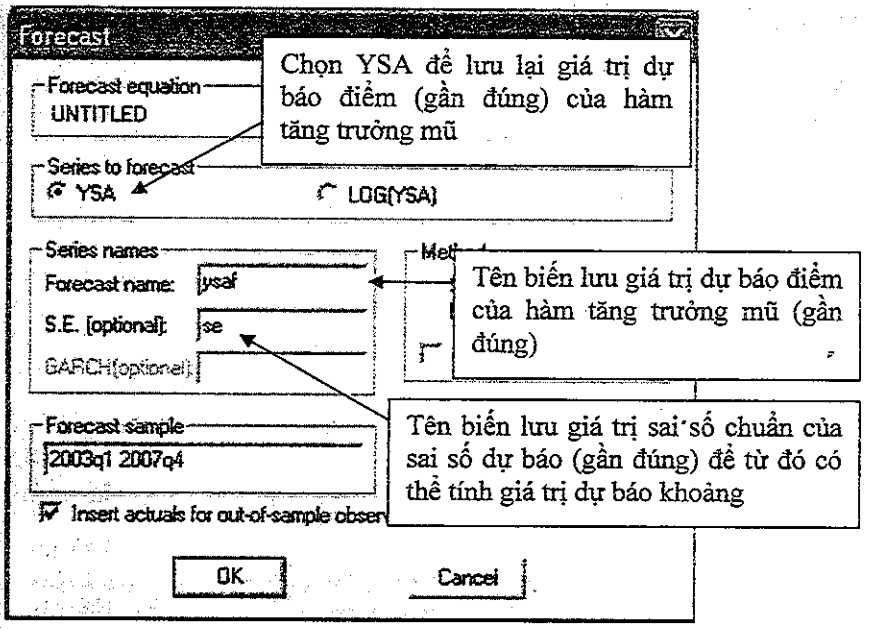
Kiểm định T có ý nghĩa thống kê ở độ tin cậy 95%, R^2 rất cao, và phương trình hồi quy là:

$$\widehat{Ln(YSA)} = 4.20 + 0.06T \text{ hoặc}$$

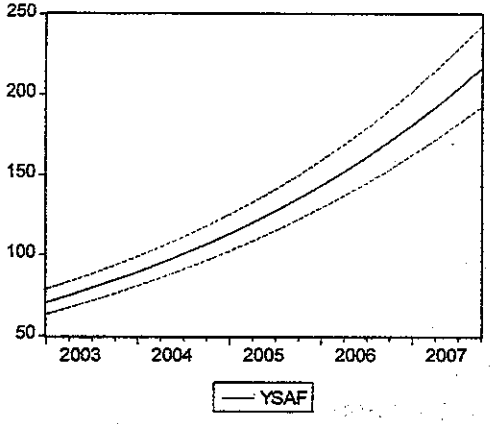
$$\hat{YSA} = e^{4.20+0.06T}$$

- Muốn dự báo điểm cho YSA ở các quý năm 2007, ta thế T bằng 17, 18, 19, 20 vào hàm tăng trưởng mũ đã được ước lượng. Về mặt thao tác trên Eviews, ta chỉ cần bấm nút **Forecast** trong cửa sổ Equation ở Hình 6.19, sau đó khai báo như Hình 6.10.

■ HÌNH 6.10: **Thực hiện dự báo trên Eviews.**



■ HÌNH 6.11: **Kết quả dự báo trên Eviews.**



Forecast: YSAF
Actual: YSA
Forecast sample: 2003Q1 2007Q4
Included observations: 16

Root Mean Squared Error	4.938460
Mean Absolute Error	4.062936
Mean Abs. Percent Error	3.748212
Theil Inequality Coefficient	0.020854
Bias Proportion	0.001655
Variance Proportion	0.074397
Covariance Proportion	0.923948

- Dò tìm sự vi phạm các giả định của mô hình:

Kết quả kiểm định hiện tượng tự tương quan bậc 1 hoặc bậc 2 cho thấy mô hình không bị vi phạm hiện tượng tự tương quan. Prob của F=0.86 (>0.05).

■ HÌNH 6.12: **Kiểm định LM của Breusch – Godfrey.**

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.154610	Probability	0.858428
Obs*R-squared	0.401937	Probability	0.817938

Kết quả kiểm định White về hiện tượng phương sai thay đổi cho thấy Prob của thống kê F = 0.64 (>0.05) nên mô hình không bị vi phạm hiện tượng phương sai thay đổi.

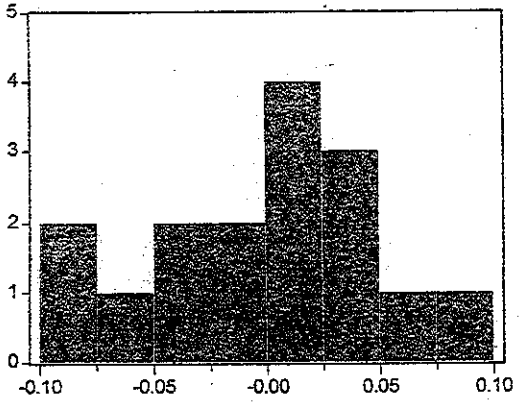
■ HÌNH 6.13: **Kiểm định phương sai thay đổi.**

White Heteroskedasticity Test:

F-statistic	0.463138	Probability	0.639299
Obs*R-squared	1.064206	Probability	0.587369

Kết quả kiểm định phân phối chuẩn của sai số cho thấy Prob(JB)=0.68 (>0.05) nên sai số của mô hình có phân phối chuẩn.

■ HÌNH 6.14: **Kiểm định Jarque – Bera.**



Series: Residuals
Sample 2003Q1 2006Q4
Observations 16

Mean	-4.58e-16
Median	0.010048
Maximum	0.077621
Minimum	-0.080309
Std. Dev.	0.047051
Skewness	-0.349569
Kurtosis	2.175363
Jarque-Bera	0.779213
Probability	0.677323

Bước 4. Kết hợp yếu tố xu thế, yếu tố mùa vụ để đưa ra kết quả dự báo cuối cùng

- Hãy gõ lệnh **genr yf=ysaf*sn** vào cửa sổ lệnh để tính giá trị dự báo điểm. Cột YF trong Hình 6.16 cho ta kết quả dự báo điểm của doanh thu:

■ HÌNH 6.15: **Kết quả dự báo cuối cùng.**

EViews

File Edit Object View Proc Quick Options Window Help

LS Log(ysa) c t

genr yf=ysaf*sn

Group: UNTITLED Workfile: DỮ LIỆU CTY DU LỊCH VT 2003-2007::Untitled

View | Proc | Object | Print | Name | Freeze | Default | Sort | Transpose | Edit+/- | Smpl+/- | InsDe

Range: 2004Q1 2007Q4
Sample 2004Q1 2007Q4

c
resid
se
sn
t
y
yf
ysa
ysaf

obs	T	Y	YSA	YSAF	SN	YF	SE
2004Q1	5.00	69.40	82.55	89.45	0.84	75.20	4.57
2004Q2	6.00	90.00	89.95	94.88	1.00	94.93	4.80
2004Q3	7.00	139.30	95.74	100.64	1.45	146.43	5.07
2004Q4	8.00	84.70	103.67	106.75	0.82	87.22	5.36
2005Q1	9.00	97.60	116.09	113.24	0.84	95.20	5.69
2005Q2	10.00	120.00	119.93	120.11	1.00	120.18	6.05
2005Q3	11.00	184.70	126.94	127.41	1.45	185.38	6.45
2005Q4	12.00	101.90	124.72	135.14	0.82	110.42	6.90
2006Q1	13.00	125.20	148.92	143.35	0.84	120.52	7.40
2006Q2	14.00	160.00	159.91	152.06	1.00	152.14	7.95
2006Q3	15.00	237.20	163.03	161.28	1.45	234.68	8.56
2006Q4	16.00	143.40	175.51	171.09	0.82	139.78	9.23
2007Q1	17.00	NA	NA	181.47	0.84	152.57	9.98
2007Q2	18.00	NA	NA	192.49	1.00	192.60	10.80
2007Q3	19.00	NA	NA	204.18	1.45	297.09	11.71
2007Q4	20.00	NA	NA	216.58	0.82	176.95	12.71

Do giả định CI=1, Ir=1 nên $\hat{Y} = \hat{YSA}.Sn$

Quý 1.2007: $\hat{Y} = \hat{YSA}.Sn = \hat{Tr}.Sn = 181.47x0.84=152.57$ (tỷ đồng)

Quý 2.2007: $\hat{Y} = \hat{YSA}.Sn = \hat{Tr}.Sn = 192.49x1 = 192.6$ (tỷ đồng)

(do thể hiện tròn số thập phân)

Quý 3.2007: $\hat{Y} = \hat{YSA}.Sn = \hat{Tr}.Sn = 204.18x1.45=297.09$ (tỷ đồng)

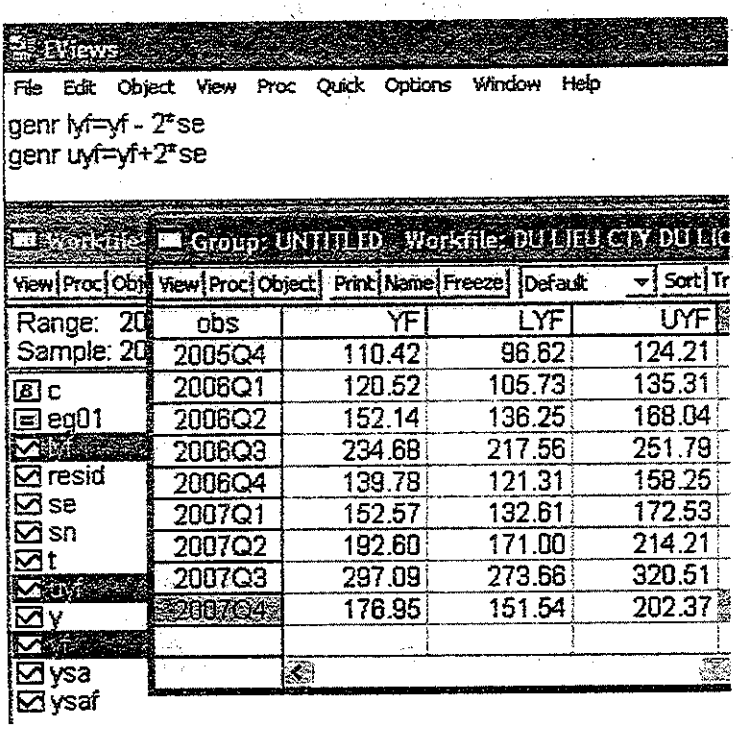
Quý 4.2007: $\hat{Y} = \hat{YSA}.Sn = \hat{Tr}.Sn = 216.58x0.82=176.95$ (tỷ đồng)

- Trong Hình 6.16 có cột SE. Nếu chúng ta muốn dự báo khoảng, ở độ tin cậy 95%, khoảng tin cậy (gần đúng) sẽ là:

  YF ± 2.SE

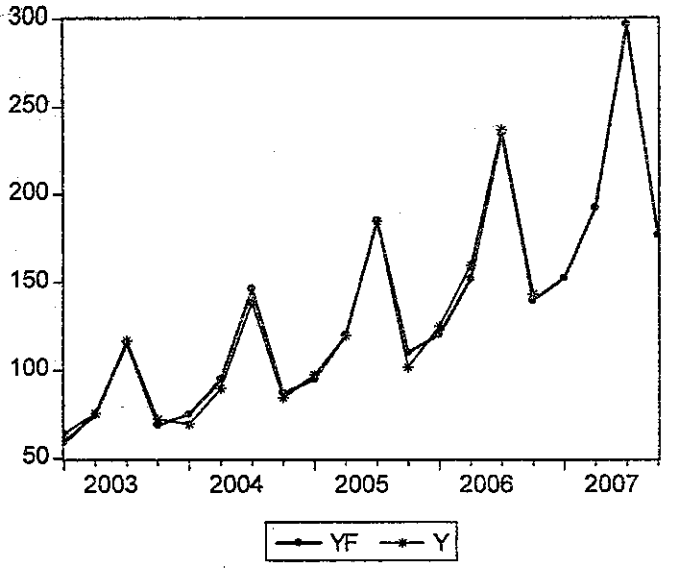
Kết quả dự báo điểm, dự báo khoảng như Hình 6.16. Như vậy, vào quý 1 năm 2007, Ở độ tin cậy 95%, doanh thu của công ty có khả năng nằm trong khoảng từ 132.61 đến 172.53 tỷ đồng.

■ HÌNH 6.16: **Kết quả dự báo khoảng.**



Chúng ta có thể đánh giá độ chính xác của mô hình dự báo thông qua các chỉ tiêu đo lường độ chính xác như Hình 6.11 trong việc dự báo giá trị tương lai của chuỗi dữ liệu khi đã loại trừ yếu tố mùa. Khi ấy MAPE=3.75%; Theil's U = 0.02 (<0.55) nên độ chính xác tốt.

■ HÌNH 6.17: **Đánh giá dự báo bằng đồ thị.**



Ngoài ra, có nhiều cách khác để đánh giá độ chính xác. Chúng ta có thể vẽ chuỗi Yf và chuỗi Y lên cùng một đồ thị. Hình 6.17 cho thấy, trong quá khứ, các giá trị thực tế (Y) và giá trị dự báo (Yf) bằng mô hình nhân tính bám rất sát nhau. Điều này thể hiện sự phù hợp của mô hình nhân tính trong việc dự báo doanh thu của công ty du lịch mà chị Oanh đang làm việc. Hoặc có thể sử dụng chuỗi Yf, Y để tính toán các chỉ tiêu đo lường độ chính xác cho mô hình dự báo.

DỰ BÁO VỚI MÔ HÌNH CỘNG

Chị Julie Murphy đang làm việc ở công ty kinh doanh hàng trang trí nội thất Murphy Brothers. Nhà máy ở Dallas bắt đầu sản xuất dòng sản phẩm gia dụng từ tháng 10 năm 1995. Doanh số hàng tháng của Công ty từ tháng 1 năm 1996 đến tháng 9 năm 2002 được thể hiện như Bảng 6.3. Đồ thị biểu diễn doanh số như Hình 6.18 (**DATA6-2**).

■ BẢNG 6.3: **Doanh số của công ty Murphy Brothers.**

Đvt: ngàn USD

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1996	4946	4968	5601	5454	5721	5690	5804	6040	5843	6087	6469	7002
1997	5416	5393	5907	5768	6107	6016	6131	6499	6249	6472	6946	7615
1998	5876	5818	6342	6143	6442	6407	6545	6758	6485	6805	7361	8079
1999	6061	6187	6792	6587	6918	6920	7030	7491	7305	7571	8013	8727
2000	6776	6847	7531	7333	7685	7518	7672	7992	7645	7923	8297	8537
2001	7005	6855	7420	7183	7554	7475	7687	7922	7426	7736	8483	9329
2002	7120	7124	7817	7538	7921	7757	7816	8208	7828			

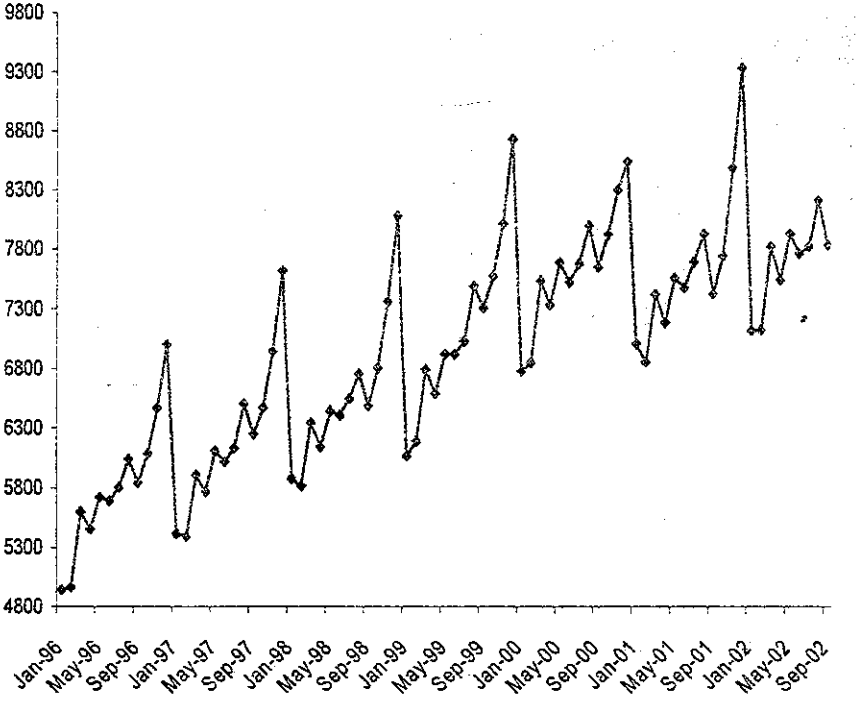
Nguồn: Hanke & Wichern (2005).

Hàng năm, chính sách chung của công ty là thuê lao động làm việc hai ca vào suốt mùa hè và đầu mùa thu và rồi làm việc một ca vào các tháng còn lại trong năm. Vì vậy hàng tồn kho được phát triển nhiều hơn vào các tháng cuối mùa hè và mùa thu cho tới khi nhu cầu bắt đầu tăng cao vào tháng 11 và tháng 12. Do yêu cầu cần thiết trong sản xuất, Julie rất băn khoăn trong việc thực hiện dự báo ngắn hạn cho công ty dựa trên nguồn thông tin liên quan đến nhu cầu sản phẩm sẵn có. Theo bạn, Julie có nên thực hiện dự báo bằng mô hình cộng tính hay không? Kết quả dự báo sẽ như thế nào?

Bước 1. Nhận dạng

Từ đầu năm 1996 đến nay (tháng 10 năm 2002), doanh số tăng dần. Hình 6.18 cũng biểu thị rằng nhu cầu (đo lường bằng doanh số bán hàng) của công ty có yếu tố mùa, và có yếu tố xu thế tuyến tính tăng dần. Vì vậy, sử dụng phương pháp phân tích là một trong các phương án hợp lý. Sự biến động của chuỗi dữ liệu có vẻ nằm trong một dải giới hạn bởi 2 đường thẳng song song, nên mô hình cộng tính là phù hợp hơn so với mô hình nhân tính. Julie quyết định điều chỉnh yếu tố mùa ra khỏi dữ liệu để có thể áp dụng phương pháp dự báo bằng mô hình xu thế dựa trên chuỗi dữ liệu đã loại trừ yếu tố mùa.

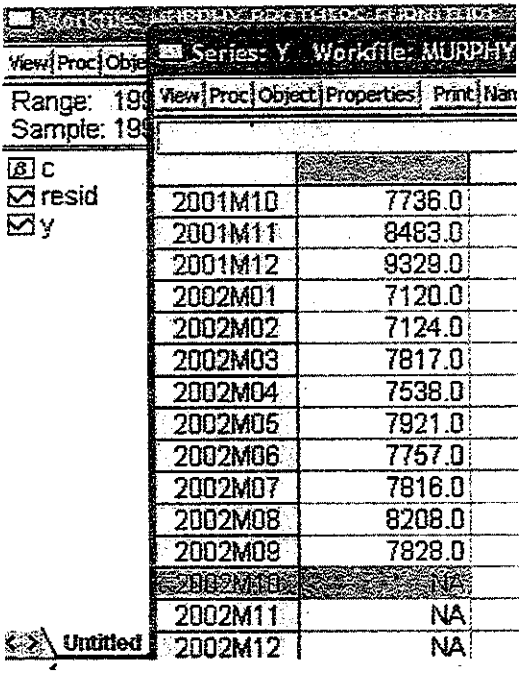
■ HÌNH 6.18: **Đồ thị doanh số của công ty Murphy Brothers.**



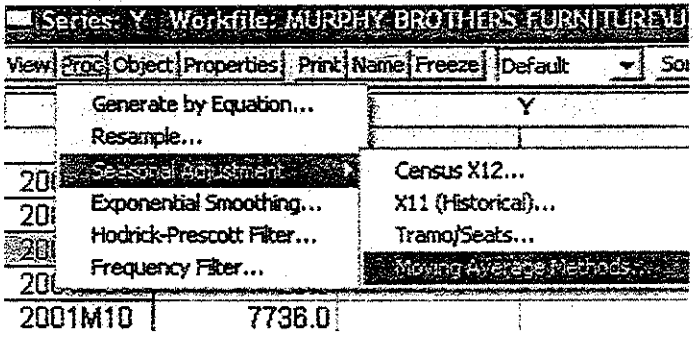
Bước 2. Tách yếu tố mùa

- Dữ liệu ban đầu (biến Y) đã được nhập vào Eviews như Hình 6.19. Chúng ta sẽ tách yếu tố mùa ra khỏi chuỗi này bằng cách: Tại cửa sổ Series của biến Y, chọn **Proc\ Seasonal Adjustment\Moving Avearage Methods** (Hình 6.20). Sau đó khai báo như Hình 6.21.

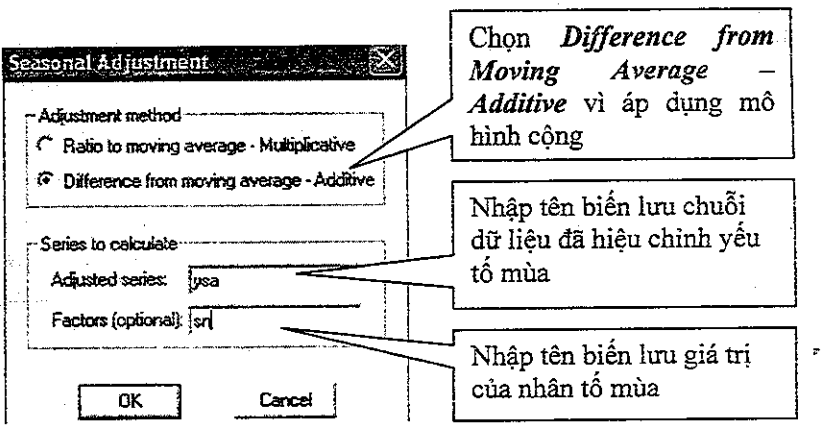
■ HÌNH 6.19: **Tập tin Eviews.**



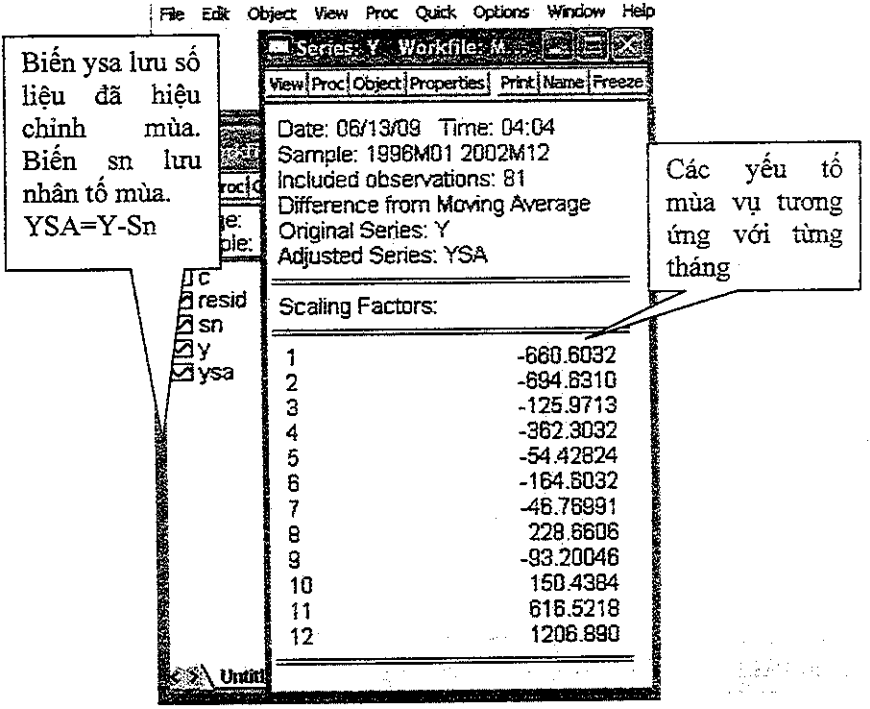
■ HÌNH 6.20: **Điều chỉnh yếu tố mùa.**



■ HÌNH 6.21: **Lựa chọn phương pháp.**



■ HÌNH 6.22: **Kết quả ước lượng trên Eviews .**



- Chúng ta hãy mở Group cho các biến Y, Sn, YSA. (Hình 6.23). Kết quả sẽ tương tự như Bảng 6.4. Trong bảng này, cột Y là dữ liệu gốc. Cột Sn là nhân tố mùa. Cột YSA là dữ liệu đã điều chỉnh yếu tố mùa (đã loại yếu tố mùa).

Trong mô hình cộng tính, $YSA_t = Y_t - Sn_t = Tr_t + Cl_t + Ir_t$

Để đơn giản, ta giả định rằng không có yếu tố chu kỳ nên $Cl_t = 0$, và yếu tố ngẫu nhiên đã bị triệt tiêu trong quá trình tính chỉ số mùa bằng cách tính giá trị trung bình nên $Ir_t = 0$. Vì vậy, bây giờ, chuỗi dữ liệu sau khi loại yếu tố mùa thì YSA_t chỉ còn yếu tố Xu thế. Hay nói cách khác $YSA_t = Tr_t$.

Ta có thể áp dụng mô hình xu thế để dự đoán xem doanh số của công ty (đã hiệu chỉnh yếu tố mùa) trong tương lai sẽ là bao nhiêu. Sau khi có được các giá trị này, chúng ta sẽ cộng thêm vào nhân tố mùa để đưa ra kết quả dự báo cuối cùng về doanh số trong tương lai.

BẢNG 6.4: **Tách yếu tố xu thế.**

Quý	Y	Sn	YSA
1996M01	4946	-660.6	5606.6
1996M02	4968	-694.6	5662.6
1996M03	5601	-126.0	5727.0
1996M04	5454	-362.3	5816.3
1996M05	5721	-54.4	5775.4
1996M06	5690	-164.6	5854.6
1996M07	5804	-46.8	5850.8
1996M08	6040	228.7	5811.3
1996M09	5843	-93.2	5936.2
1996M10	6087	150.4	5936.6
1996M11	6469	616.5	5852.5

Quý	Y	Sn	YSA
1996M12	7002	1206.9	5795.1
...	...	...	...
2002M01	7120	-660.6	7780.6
2002M02	7124	-694.6	7818.6
2002M03	7817	-126.0	7943.0
2002M04	7538	-362.3	7900.3
2002M05	7921	-54.4	7975.4
2002M06	7757	-164.6	7921.6
2002M07	7816	-46.8	7862.8
2002M08	8208	228.7	7979.3
2002M09	7828	-93.2	7921.2
2002M10		150.4	
2002M11		616.5	
2002M12		1206.9	

Bước 3. Dự báo YSA (doanh số đã loại yếu tố mùa) bằng mô hình xu thế

- Trước tiên, ta tạo ra biến thứ tự thời gian *t*, với t=1 ở tháng 1 năm 1996, t=2 ở tháng 2 năm 1996... và t=81 ở tháng 9 năm 2002 bằng lệnh:

```
genr t=@trend(1995:12)
```

Đồ thị biểu diễn YSA theo thứ tự thời gian như Hình 6.23. Với Hình này, YSA có vẻ như xu thế với dạng đường thẳng, hoặc đường cong bậc 2... Chúng ta thử ước lượng mô hình bậc 1 cho đơn giản nhằm mục tiêu minh hoạ cho phương pháp phân tích cho mô hình cộng tính

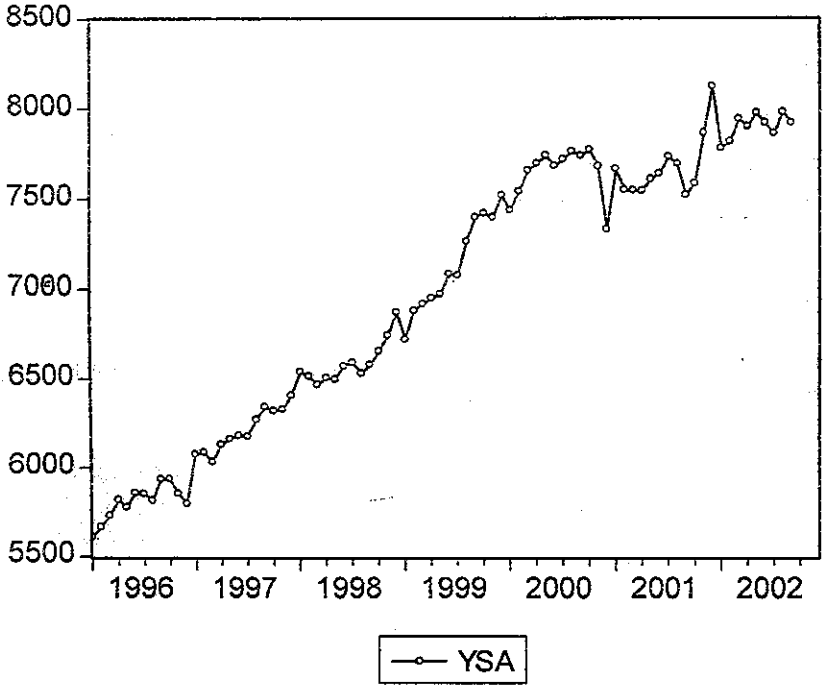
đang làm (chúng ta có thể ước lượng các mô hình xu thế khác tốt hơn theo quan điểm riêng của mình!).

- Gõ lệnh **LS YSA C T** vào cửa sổ lệnh, ta sẽ có được phương trình:

$$\hat{YSA} = 5670.01 + 31.45t$$

(SE) 39.75 0.84
(t-stat) 142.65 37.35
$R^2 = 0.946$ F = 1395.3 n = 81

HÌNH 6.23: **Đồ thị dự báo.**



■ HÌNH 6.24: **Kết quả ước lượng trên Eviews.**

Equation: UNTITLED Workfile: MURPHY BROTHERS FURNITURE::...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

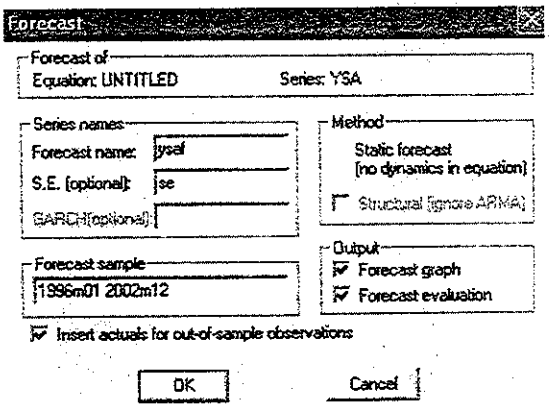
Dependent Variable: YSA
Method: Least Squares
Date: 06/13/09 Time: 05:00
Sample (adjusted): 1996M01 2002M09
Included observations: 81 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5670.012	39.74682	142.6532	0.0000
T	31.45737	0.842126	37.35468	0.0000

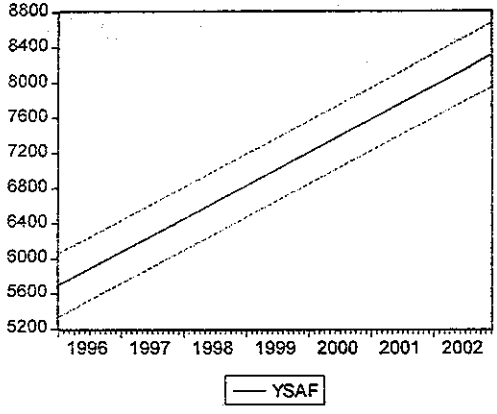
R-squared	0.946418	Mean dependent var	6959.784
Adjusted R-squared	0.945740	S.D. dependent var	760.7460
S.E. of regression	177.2071	Akaike info criterion	13.21680
Sum squared resid	2480786.	Schwarz criterion	13.27602
Log likelihood	-533.2843	F-statistic	1395.372
Durbin-Watson stat	0.384163	Prob(F-statistic)	0.000000

- Để lưu lại giá trị dự báo điểm cho chuỗi YSA, bấm nút **Forecast**. Khai báo các thông tin tương tự như Hình 6.25 sẽ có được giá trị dự báo điểm (Biến YSAF), và sai số chuẩn của sai số dự báo (SE).

■ HÌNH 6.25: **Dự báo khoảng trên Eviews.**



■ HÌNH 6.26: **Đồ thị dự báo khoảng trên Eviews.**



Forecast: YSAF
Actual: YSA
Forecast sample: 1996M01 2002M12
Included observations: 81

Root Mean Squared Error	175.0057
Mean Absolute Error	135.0975
Mean Abs. Percent Error	1.859262
Theil Inequality Coefficient	0.012501
Bias Proportion	0.000000
Variance Proportion	0.013767
Covariance Proportion	0.986233

Bước 4. Kết hợp yếu tố xu thế và yếu tố mùa để đưa ra kết quả dự báo

Kết quả dự báo điểm của doanh số sẽ là:

$$\hat{Y}_t = Y\hat{S}A_t + Sn_t = \hat{T}r_t + Sn_t \text{ (do } Cl_t=0, \text{ và } Ir_t=0)$$

- Tại cửa sổ lệnh, gõ lệnh **genr yf=ysaf+sn** để lưu kết quả dự báo điểm của doanh số

  Ở tháng 10 năm 2002: $\hat{Y}$ = 8249.5+150.4= 8400.0 (ngàn USD)

  Ở tháng 11 năm 2002: $\hat{Y}$ = 8281.0+616.5= 8897.5 (ngàn USD)

  Ở tháng 12 năm 2002:

  $$\hat{Y} = 8312.4+1206.9= 9519.3 \text{ (ngàn USD)}$$

- Kết quả dự báo điểm ở độ tin cậy 95% cũng tính gần đúng bằng cách: **YF±2.SE.** Và chúng ta có thể dễ dàng tính được chỉ tiêu này.

■ HÌNH 6.27: **Kết quả dự báo khoảng trên Eviews.**

TITLED Workfile: MURPHY BROTHERS FURNITURE\Untitled

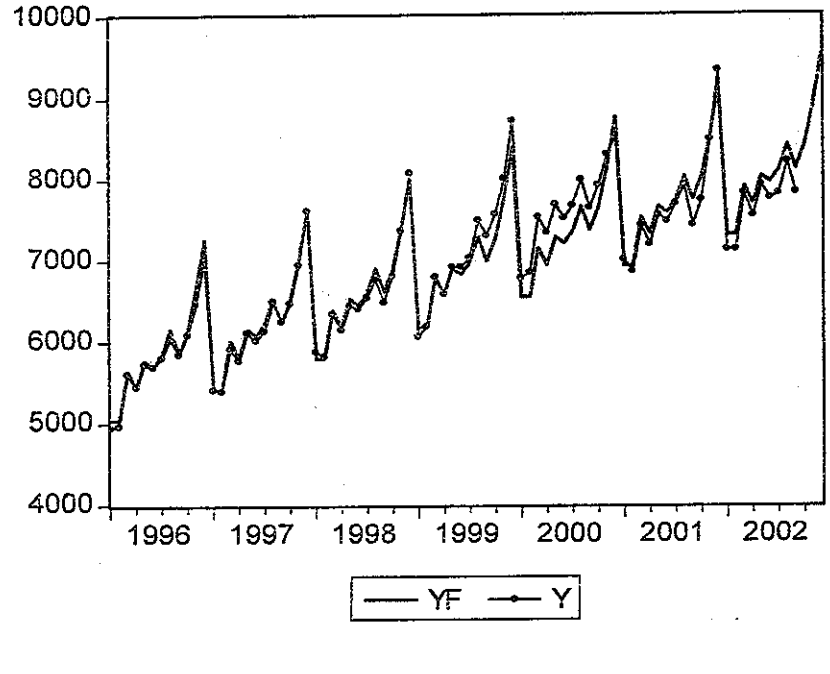
ct | Print | Name | Freeze | Default | Sort | Transpose | Edit+/- | Smpl+/- | InsDel | Title | Sample

obs	T	Y	YSA	SN	YSAF	YF	SE
2001M12	72.0	9329.0	8122.1	1206.9	7934.9	9141.8	180.2
2002M01	73.0	7120.0	7780.6	-660.6	7966.4	7305.8	180.3
2002M02	74.0	7124.0	7818.6	-694.6	7997.9	7303.2	180.5
2002M03	75.0	7817.0	7943.0	-126.0	8029.3	7903.3	180.6
2002M04	76.0	7538.0	7900.3	-362.3	8060.8	7698.5	180.7
2002M05	77.0	7921.0	7975.4	-54.4	8092.2	8037.8	180.9
2002M06	78.0	7757.0	7921.6	-164.6	8123.7	7959.1	181.0
2002M07	79.0	7816.0	7862.8	-46.8	8155.1	8108.4	181.1
2002M08	80.0	8208.0	7979.3	228.7	8186.6	8415.3	181.3
2002M09	81.0	7626.0	7921.2	-93.2	8218.1	8124.9	181.5
2002M10	82.0	NA	NA	150.4	8249.5	8400.0	181.6
2002M11	83.0	NA	NA	616.5	8281.0	8897.5	181.8
2002M12	84.0	NA	NA	1206.3	8312.4	9519.3	181.9

Độ chính xác của mô hình xu thế có thể được đánh giá thông qua các chỉ tiêu ở Hình 6.27. MAPE = 1.89 % cũng là rất nhỏ, đồng thời Theil's U = 0.012 (<0.55).

Ta có thể dựa vào các biến Y, YF để tính toán các chỉ tiêu đo lường độ chính xác cho cả mô hình phân tích. Tuy nhiên, để đơn giản, chúng ta vẽ Y, và YF lên cùng 1 đồ thị. Đường Y và YF trên đồ thị bám rất sát nhau, nên sử dụng mô hình cộng tính cũng phù hợp.

HÌNH 6.28: **Đồ thị dự báo theo mô hình cộng tính.**



KIỂM ĐỊNH TÍNH MÙA VỤ

Kiểm định Kruskal-Wallis

Anh Thắng đang làm việc ở một Công ty du lịch và đã thu thập số liệu về Khách quốc tế đến Việt Nam từ Quý 1 năm 2003 đến Quý 4 năm 2007 (biến Y, đơn vị tính là nghìn lượt khách) từ trang web của tổng cục du lịch: http://www.vietnamtourism.gov.vn (Hình 6.29, **DATA6-3**).

Dự báo được lượt khách quốc tế đến Việt Nam năm 2008 sẽ giúp Công ty anh xây dựng các kế hoạch kinh doanh hợp lý hơn. Đường biểu diễn Y theo thời gian như Hình 6.30. anh Thắng đang phân vân không biết dữ liệu này có yếu tố mùa vụ không?

Trong thực tế, chúng ta sẽ gặp nhiều trường hợp như trường hợp của anh Thắng. Bằng đồ thị, yếu tố mùa thể hiện chưa rõ ràng, và việc

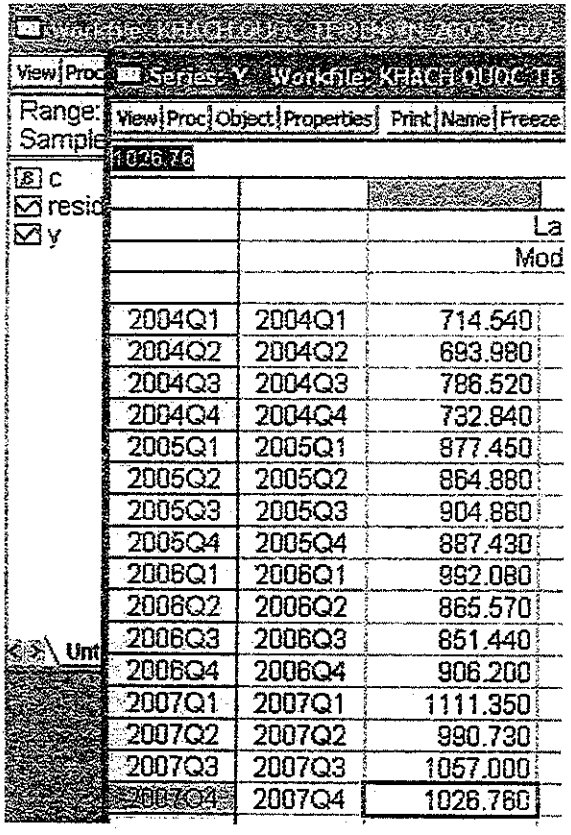
nhận định cũng khá chủ quan qua quan sát đồ thị. Phần này sẽ trình bày một phương pháp kiểm định yếu tố mùa vụ rất đơn giản nhưng thật sự hữu ích: Kruskal-Wallis *(ở các chương sau, chúng ta sẽ biết được một số phương pháp khác phức tạp hơn)*. Chúng ta nên kiểm định yếu tố mùa vụ trước khi thực hiện dự báo bằng phương pháp phân tích thì sẽ hiệu quả hơn.

Gaynor & Kirkpatrick (1994) cho rằng để kiểm định tính mùa, có thể sử dụng kiểm định Kruskal-Wallis cho chuỗi dữ liệu $Sn_t.Ir_t$ với một biến phân nhóm (mã hóa quý/ hoặc tháng). Chuỗi $Sn_t.Ir_t$ có được bằng cách lấy chuỗi dữ liệu ban đầu (biến Y_t) chia (với mô hình nhân tính) hoặc trừ (với mô hình cộng tính) cho CMA_t của chuỗi ban đầu (biến Y_t). Nếu không có yếu tố mùa cụ thể thì dữ liệu tính được ($Sn_t.Ir_t$) chỉ bao hàm yếu tố ngẫu nhiên (Ir_t) và vì thế hạng và phân phối của yếu tố mùa đều giống nhau cho các chuỗi thời gian ở mỗi quý hay (mỗi tháng).

Kruskal-Wallis là một trong các kiểm định phi tham số. Kiểm định này khá tương tự như kiểm định ANOVA một chiều vì nó cũng có thể đưa ra kết luận xem trung bình giữa các nhóm có khác biệt hay không. Tuy vậy, không như ANOVA một chiều, Kiểm định Kruskal-Wallis không đòi hỏi về phân phối chuẩn của dữ liệu trong từng nhóm, cũng không yêu cầu về số lượng quan sát trong mỗi nhóm phải lớn; nó dựa trên sự xếp hạng của $Sn_t.Ir_t$, và xem xét hạng trung bình của $Sn_t.Ir_t$ có khác biệt giữa các mùa (các quý/các tháng) hay không, hay phân phối của $Sn_t.Ir_t$ có khác biệt giữa các mùa (các quý/các tháng) hay không. Nếu có sự khác biệt, thì chuỗi dữ liệu gốc ban đầu có tồn tại yếu tố mùa.

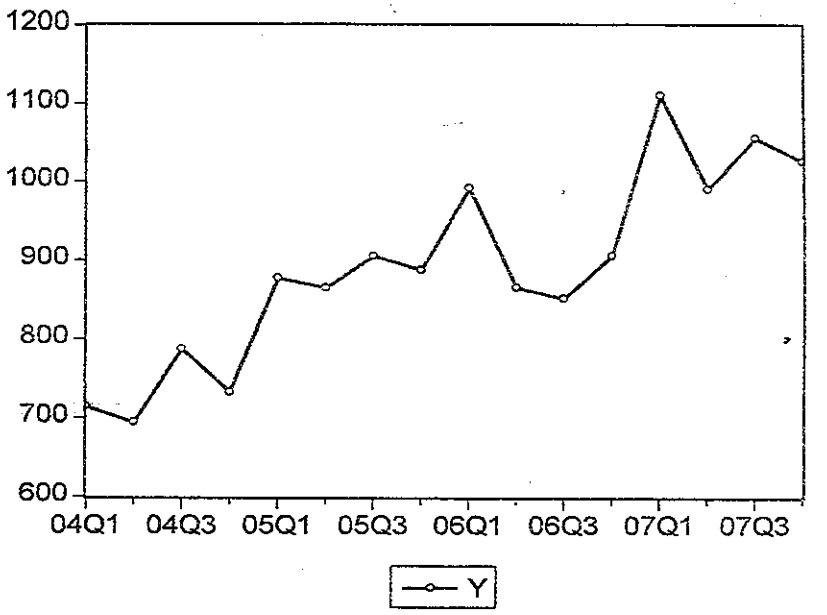
Chúng ta sẽ thực hiện kiểm định này với tình huống về lượt khách du lịch đến Việt Nam mà anh Thắng đang quan tâm.

■ HÌNH 6.29: **Lượng du khách theo quý.**



		La
		Mod
2004Q1	2004Q1	714.540
2004Q2	2004Q2	693.980
2004Q3	2004Q3	786.520
2004Q4	2004Q4	732.840
2005Q1	2005Q1	877.450
2005Q2	2005Q2	864.880
2005Q3	2005Q3	904.880
2005Q4	2005Q4	887.430
2006Q1	2006Q1	992.080
2006Q2	2006Q2	865.570
2006Q3	2006Q3	851.440
2006Q4	2006Q4	906.200
2007Q1	2007Q1	1111.350
2007Q2	2007Q2	990.730
2007Q3	2007Q3	1057.000
2007Q4	2007Q4	1026.760

■ HÌNH 6.30: **Đồ thị du khách theo thời gian.**



Bước 1. Tính CMA và chuỗi Sn.Ir

Chúng ta cũng chưa biết được là mô hình cộng tính, hay mô hình nhân tính. Vì vậy ta sẽ tính $Sn_t.Ir_t$ cho trường hợp cả mô hình cộng tính và mô hình nhân tính. Trong phương pháp phân tích, phần mềm Eviews sử dụng cách tính CMA (Centered Moving Average) của chuỗi Y_t như sau:

- $CMA_t = (0.5Y_{t+6} + \ldots + Y_t + \ldots + 0.5Y_{t-6})/12$ nếu số liệu theo tháng.
- $CMA_t = (0.5Y_{t+2} + Y_{t+1} + Y_t + Y_{t-1} + 0.5Y_{t-1})/4$ nếu số liệu theo quý.

Có nhiều cách tính CMA, tuy nhiên, do chúng ta sử dụng Eviews trong thực hành là chính, nên chúng ta sẽ sử dụng cách tính CMA theo hướng dẫn của **Eviews 6 User's Guide (2007)** cho thống nhất.

- Hãy gõ các lệnh sau vào cửa sổ lệnh của Eviews:

  genr cma=(0.5*y(-1)+y(-1)+y+y(1)+0.5*y(2))/4

  (để tính CMA_t theo với dữ liệu quý)

  genr snir_add=y-cma

  (để tính chuỗi $Sn_t.Ir_t$ nếu là mô hình cộng tính)

  genr snir_mul=y/cma

  (để tính chuỗi $Sn_t.Ir_t$ nếu là mô hình nhân tính)

  genr quarter=@quarter

  (để tạo biến quarter lưu mã các quý)

Kết quả giá trị của các biến này như Hình 6.31. Khi tính CMA_t theo công thức trên với khoảng trượt bằng 4, có 3 quan sát không có dữ liệu và Eviews ghi chú là chữ NA.

■ HÌNH 6.31: **Kết quả tính toán CMA và chuỗi Sn.Ir.**

File Edit Object View Proc Quick Options Window Help

```
genr cma=(0.5*y(-1)+y(-1)+y+y(1)+0.5*y(2))/4
genr snir_add=y-cma
genr snir_mul=y/cma
genr quarter=@quarter
```

Group: UNTITLED Workfile: UNTITLED\Untitled

View | Proc | Object | Print | Name | Freeze | Default | Sort | Transpose | Edit+/- | Smpl+/-

Range: 20
Sample: 20

- c
- cma
- quarter
- resid
- snir_add
- snir_mul
- y

obs	QUARTER	Y	CMA	SNIR_ADD	SNIR_MUL
2004Q1	1	714.54	NA	NA	NA
2004Q2	2	693.98	729.68	-35.70	0.95
2004Q3	3	786.52	749.76	36.76	1.05
2004Q4	4	732.84	805.63	-72.79	0.91
2005Q1	1	877.45	823.51	53.94	1.07
2005Q2	2	864.88	882.41	-17.53	0.98
2005Q3	3	904.88	896.42	8.46	1.01
2005Q4	4	887.43	917.40	-29.97	0.97
2006Q1	1	992.08	903.63	88.45	1.10
2006Q2	2	865.57	914.56	-48.99	0.95
2006Q3	3	851.44	902.92	-51.48	0.94
2006Q4	4	906.20	947.52	-41.32	0.96
2007Q1	1	1111.35	997.47	113.88	1.11
2007Q2	2	990.73	1057.03	-66.30	0.94
2007Q3	3	1057.00	NA	NA	NA
2007Q4	4	1026.76	NA	NA	NA

Bước 2. Thực hiện kiểm định Kruskal-Wallis

a. Nếu chọn mô hình nhân tính

- Double click vào biến **snir_mul** để mở **cửa sổ Series** của biến này.
- Tại cửa sổ Series của biến **snir_mul**, **Chọn View\Test for Descriptive Stats\Equality Tests by Classification** → Hộp thoại **Tests by Classification** như Hình 6.32 sẽ xuất hiện.
- Trong hộp thoại **Test By Classification**; ở khung **Test equality of**, chọn mục **Median**[1] để thực hiện kiểm định Kruskal-Wallis và một số kiểm định phi tham số khác; ở khung **Series/Group for classify** nhập tên biến phân nhóm (biến mã các quý nếu dữ liệu theo quý/các tháng nếu dữ liệu theo tháng) sau đó **OK**. Kết quả tính toán như Bảng 6.5.

Trong ví dụ này, kiểm định Kruskal-Wallis, giả thuyết H_o và H_1 được phát biểu như sau:

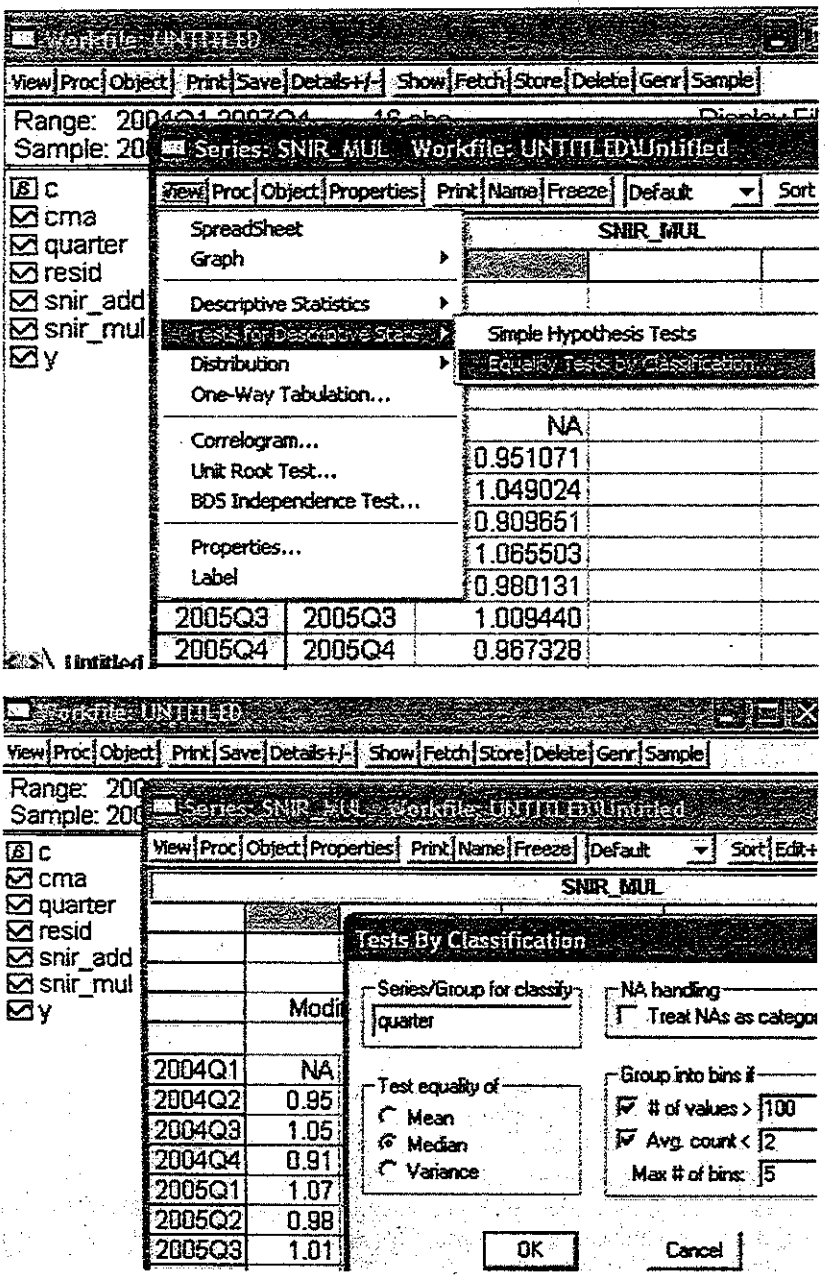
H_o: Không có yếu tố mùa (qua các năm, phân phối của Sn.Ir tương tự ở các quý).

H_1: Có yếu tố mùa (phân phối của Sn.Ir khác nhau giữa các quý).

Nếu P-value của thống kê **Kruskal-Wallis** < α thì ở mức ý nghĩa α, chuỗi dữ liệu (biến Y) có yếu tố mùa. Ngược lại, chuỗi dữ liệu (biến Y) không có yếu tố mùa.

[1] Phần mềm Eviews quy ước loại này để kiểm định sự tương tự về phân phối giữa các nhóm, gồm một số kiểm định phi tham số khác nhau như kiểm định dấu và hạng Wilcoxon, kiểm định Chi-square về trung vị, kiểm định Kruskal-Wallis, kiểm định van der Waerden.

■ HÌNH 6.32: **Kiểm định ngang bằng trên Eviews.**



■ BẢNG 6.5: **Kiểm định Kruskal-Wallis của mô hình nhân tính.**

Test for Equality of Medians of SNIR_MUL
Categorized by values of QUARTER
Date: 06/14/09 Time: 11:52
Sample (adjusted): 2004Q2 2007Q2
Included observations: 13 after adjustments

Method	df	Value	Probability
Med. Chi-square	3	7.299603	0.0629
Adj. Med. Chi-square	3	2.856647	0.4143
Kruskal-Wallis	3	7.379121	**0.0607**
Kruskal-Wallis (tie-adj.)	3	7.379121	0.0607
van der Waerden	3	7.391614	0.0604

Category Statistics

QUARTER	Count	Median	> Overall Median	Mean Rank	Mean Score
1	3	1.097885	3	**12.00000**	1.108148
2	4	0.948754	1	4.750000	-0.454903
3	3	1.009440	2	**7.333333**	0.046806
4	3	0.956393	0	4.666667	-0.548415
All	13	0.967328	6	7.000000	-1.71E-17

Bảng 6.5 cho thấy, Prob của thống kê **Kruskal-Wallis** bằng 0.0607 (nhỏ hơn 0.1) nên ở độ tin cậy 90% (hay ở mức ý nghĩa 10%) có yếu tố mùa tồn tại trong chuỗi dữ liệu (nếu thực hiện phân tích bằng mô hình nhân tính).

b. Nếu chọn mô hình cộng tính

Chúng ta hãy thao tác tương tự cho biến **snir_add**, và kết quả như Bảng 6.6. Với kết quả ở Bảng 6.6, Prob của thống kê **Kruskal-Wallis** bằng 0.0594 (nhỏ hơn 0.1) nên ở độ tin cậy 90% (hay ở mức ý nghĩa 10%) có yếu tố mùa tồn tại trong chuỗi dữ liệu (nếu thực hiện phân tích bằng mô hình cộng tính).

Việc lựa chọn mức ý nghĩa α bằng bao nhiêu (α =10%, α =5% hay α =1%) là tùy vào người làm dự báo. Tùy vào việc chúng ta quan tâm đến mức rủi ro khi ra các quyết định như thế nào, và tùy thuộc vào số quan sát trong dữ liệu mà chúng ta có được.

Trong trường hợp này, cỡ mẫu hơi nhỏ chỉ với số quan sát bằng 16 (4 quý trong 4 năm) nên chọn mức ý nghĩa α =10% cũng hợp lý. Cả 2 mô hình đều chỉ ra rằng có yếu tố mùa vụ trong chuỗi dữ liệu. Mô hình cộng tính tốt hơn mô hình nhân tính một chút (Prob của thống kê **Kruskal-Wallis của** mô hình cộng tính nhỏ hơn so với mô hình nhân tính) trong việc chỉ ra rằng có yếu tố mùa trong dữ liệu, nên chúng ta có thể chọn mô hình cộng tính để thực hiện dự báo bằng phương pháp phân tích. Việc chọn mô hình cộng tính, hay mô hình nhân tính cũng có thể căn cứ trên độ chính xác của từng mô hình mà ta chọn để dự báo, dựa trên những phân tích của mình về môi trường dự báo trong tương lai.

■ BẢNG 6.6: **Kiểm định Kruskal-Wallis của mô hình cộng tính.**

Test for Equality of Medians of SNIR_ADD
Categorized by values of QUARTER
Date: 06/14/09 Time: 12:15
Sample (adjusted): 2004Q2 2007Q2
Included observations: 13 after adjustments

Method	df	Value	Probability
Med. Chi-square	3	7.299603	0.0629
Adj. Med. Chi-square	3	2.856647	0.4143
Kruskal-Wallis	3	7.428571	**0.0594**
Kruskal-Wallis (tie-adj.)	3	7.428571	0.0594
van der Waerden	3	7.466621	0.0584

Category Statistics

	> Overall

QUARTER	Count	Median	Median	Mean Rank	Mean Score
1	3	88.45125	3	**12.00000**	1.108148
2	4	-42.34500	1	5.000000	-0.408380
3	3	8.462500	2	**7.333333**	0.046806
4	3	-41.31875	0	4.333333	-0.610447
All	13	-29.97375	6	7.000000	-3.42E-17

Chúng ta hãy xem kết quả kiểm định tính mùa vụ cho chuỗi dữ liệu doanh số theo từng tháng của Công ty Murphy Brothers.

- Chúng ta lần lượt gõ các lệnh sau để tính CMA:

  Genr cma=(0.5*y(-6)+y(-6)+y(-5)+ y(-4)+ y(-3)+ y(-2)+ y(-1)+y+y(1)+ y(2)+ y(3)+ y(4)+ y(5)+0.5*y(6))/12

  Từ hình vẽ, ta thấy mô hình cộng tính là hợp lý nên có thể gõ tiếp các lệnh:

  genr snir_add=y-cma (để tạo chuỗi $Sn_t.Ir_t$)

  genr month=@month (để tạo biến mã các tháng)

Chọn biến snir_add, sau đó thực hiện Kiểm định Kruskal-Wallis. Kết quả như Bảng 6.7. Trong bảng này, Prob(Kruskal-Wallis)=0.0000 (nhỏ hơn 0.05) nên ở độ tin cậy 95%, chuỗi số liệu về doanh số của công ty Murphy Brothers có yếu tố mùa.

BẢNG 6.7: **Kiểm định Kruskal-Wallis với dữ liệu theo tháng.**

Test for Equality of Medians of SNIR_ADD
Categorized by values of MONTH
Date: 06/14/09 Time: 14:55
Sample (adjusted): 1996M07 2002M03
Included observations: 69 after adjustments

Method	Df	Value	Probability
Med. Chi-square	11	54.86370	0.0000
Adj. Med. Chi-square	11	37.22666	0.0001
Kruskal-Wallis	11	64.37300	**0.0000**
Kruskal-Wallis (tie-adj.)	11	64.37300	0.0000
van der Waerden	11	63.61735	0.0000

Category Statistics

MONTH	Count	Median	> Overall Median	Mean Rank	Mean Score
1	6	-1280.458	0	7.666667	-1.279200
2	6	-1287.917	0	5.666667	-1.512596
3	6	-673.7917	0	29.00000	-0.218824
4	5	-921.7500	0	15.60000	-0.772214
5	5	-669.2917	2	31.40000	-0.131672
6	5	-859.7500	0	20.80000	-0.535033
7	6	-548.9792	5	40.50000	0.199894
8	6	-244.7500	6	53.33333	0.720051
9	6	-594.5417	3	34.66667	-0.020393
10	6	-341.4375	6	48.16667	0.499627
11	6	59.22917	6	60.50000	1.106857
12	6	703.9792	6	66.50000	1.703683
All	69	-638.3750	34	35.00000	0.000000

Kiểm định Kruskal-Wallis cho chuỗi Sn.Ir với công thức tính CMA của Gaynor & Kirkpatrick

Eviews hỗ trợ việc dự báo bằng phương pháp phân tích số liệu theo quý, theo tháng. Bây giờ, chúng ta chắc cũng đã quen thuộc với phương pháp chính mà chương này muốn chuyển tải. Tuy nhiên, việc tách yếu tố mùa vụ dựa trên trung bình trượt trung tâm có rất nhiều cách thức, phần mềm SPSS cũng cho kết quả khác với Eviews; và hiện nay, nhiều tài liệu hướng dẫn chúng ta tính CMA với khoảng trượt L là số chẵn (L = 4 nếu số liệu theo quý, L = 12 nếu số liệu theo tháng) theo công thức phổ biến được đề cập trong Gaynor & Kirkpatrick (1994). Và từ đó kết quả tính $Sn_t.Ir_t = Y_t/CMA_t$, và Sn_t, cũng như chuỗi Y_t đã điều chỉnh yếu tố mùa $=Y_t/Sn_t$ cũng hơi khác so với kết quả mà chúng ta đã thực hiện ở các phần trên của chương này. Chúng ta cũng nên biết qua cách thức này, và thực hiện nó trên Excel và Eviews để bạn đọc có thể liên hệ lại với kiến thức thống kê ứng dụng mà chúng ta đã quen thuộc ở giai đoạn 1 của quá trình học đại học, cũng như dễ dàng tham khảo các tài liệu khác nhau.

Từ số liệu về doanh thu của công ty du lịch ở Mục 6.2; bằng mô hình nhân tính, ta có kết quả tính toán chuỗi nhân tố mùa vụ (Sn_t), chuỗi dữ liệu đã điều chỉnh mùa (Tr.Cl.Ir) theo công thức của Gaynor&ctg như Bảng 6.8.

■ BẢNG 6.8: **Tính nhân tố mùa, chuỗi dữ liệu điều chỉnh mùa.**

T	năm	quý	Y_t	MA_t	CMA_t	$Sn_t.Ir_t$	Sn_t	$Tr_t.Cl_t.Ir_t$
1	2004	1	64.2				0.818	78.52
2	2004	2	75.7	82.4			0.973	77.80
3	2004	3	117.1	83.7	83.0	**1.411**	1.415	82.76
4	2004	4	72.4	87.2	85.4	0.847	0.795	91.12
5	2005	1	69.4	92.8	90.0	0.771	0.818	84.89
6	2005	2	90.0	95.9	94.3	0.954	0.973	92.50
7	2005	3	139.3	102.9	99.4	**1.402**	1.415	98.45
8	2005	4	84.7	110.4	106.7	0.794	0.795	106.61
9	2006	1	97.6	121.8	116.1	0.841	0.818	119.38
10	2006	2	120.0	126.1	123.9	0.969	0.973	123.33
11	2006	3	184.7	133.0	129.5	**1.426**	1.415	130.54
12	2006	4	101.9	143.0	138.0	0.739	0.795	128.25
13	2007	1	125.2	156.1	149.5	0.837	0.818	153.14
14	2007	2	160.0	166.5	161.3	0.992	0.973	164.44
15	2007	3	237.2				1.415	167.64
16	2007	4	143.4				0.795	180.49

Bảng $Sn_t.Ir_t$ theo từng quý:	Quý 1	Quý 2	**Quý 3**	Quý 4	
	0.771	0.954	**1.411**	0.847	
	0.841	0.969	**1.402**	0.794	
	0.837	0.992	**1.426**	0.739	
Trung bình	***0.816***	***0.972***	***1.413***	***0.793***	**3.994**
Sn_q	0.818	0.973	1.415	0.795	4.000

Sắp xếp lại theo từng quý

Để có được kết quả này trải qua một số bước sau:

Bước 1. Tính MA_t

Ở quan sát 2: $MA_2=(64.2 + 75.7 + 117.1 + 72.4)/4=82.4$

Ở quan sát 3: $MA_3=(75.7 + 117.1 + 72.4 + 69.4)/4=83.7$

Ở quan sát thứ 4: $MA_4=(117.1 + 72.4 + 69.4 + 90.0)/4=87.2$

...

Ở quan sát thứ t: $\mathbf{MA_t=(Y_{t-1} + Y_t + Y_{t+1} + Y_{t+2})/4}$

Với L = 4 ta mất đi 1 quan sát đầu và 2 quan sát cuối.

Bước 2. Tính CMA_t

Ở quan sát thứ 3: $CMA_3=(MA_3+ MA_2)/2 = (82.4+83.7)/2 = 83.0$

Ở quan sát thứ 4: $CMA_4=(MA_4+ MA_3)/2 = (87.2+83.7+)/2 = 85.4$

Ở quan sát thứ t: $CMA_t=(MA_t+ MA_{t-1})/2$

Bước3. Tính $Sn_t.Ir_t = Y_t/CMA_t$

(Nếu là mô hình cộng tính thì $Sn_t.Ir_t = Y_t- CMA_t$)

Bước 4. Sắp xếp $Sn_t.Ir_t$ theo từng quý

(Nếu số liệu theo tháng thì sắp xếp theo từng tháng)

Bước 5. Tính trung bình Sn.Ir theo từng quý

(nếu số liệu theo tháng thì tính trung bình Sn.Ir theo từng tháng)

Với quý 1. Trung bình Sn.Ir = (0.771+ 0.841 + 0.837)/3 = **0.816**

Với quý 2. Trung bình Sn.Ir = (0.954+0.9690.992)/3 = **0.972**

Với quý 3. Trung bình Sn.Ir = (**1.411+1.402+1.426**)/3 = **1.413**

Với quý 4. Trung bình Sn.Ir = (0.847+0.794+0.739)/3 = **0.793**

Bước 6. Điều chỉnh trung bình Sn.Ir để tính toán nhân tố mùa Sn

Theo nguyên tắc, tổng các trung bình Sn.Ir theo từng quý phải bằng 4 (bằng khoảng trượt L).

Nhưng hiện nay: **0.816+0.972+1.413+0.793=3.994**

Nên, chúng ta sẽ điều chỉnh như sau:

Ở quý 1 : Sn = 0.816 x 4/ 3.994 = 0.818

Ở quý 2 : Sn = 0.972 x 4/ 3.994 = 0.973

Ở quý 3 : Sn = 1.413 x 4/ 3.994 = 1.415

Ở quý 4 : Sn = 0.793 x 4/ 3.994 = 0.795

Ta thấy: 0.818+0.973+1.415+0.795=4

Với cách phân tích dựa trên CMA thì Sn trong từng quý không thay đổi theo các năm. Từ đó, chúng ta có được cột Sn_t.

Bước 7. Tính chuỗi dữ liệu đã hiệu chỉnh yếu tố mùa

Tính toán cột dữ liệu đã hiệu chỉnh yếu tố mùa bằng cách:

- Y_t/Sn_t nếu là mô hình nhân tính, kết quả như cột $Tr_t.Cl_t.Ir_t$
- Y_t-Sn_t nếu là mô hình cộng tính

Tách yếu tố mùa vụ ra khỏi chuỗi gốc ban đầu theo cách thức của Gaynor& Kirkpatrick (1994) là như vậy.

Wilson & Keating (2007) cũng hướng dẫn phân tích chuỗi thời gian tương tự như Gaynor& Kirkpatrick (1994) nhưng có khác ở cách tính MA_t và CMA_t; theo các tác giả này, với số liệu quý, MA_t= $(Y_{t-2}+Y_{t-1}+Y_t+Y_{t+1})/4$ và $CMA_t=(MA_t+Ma_{t+1})/2$. Chuỗi $Sn_t.Ir_t=Y_t/CMA_t$, tính $Sn_t.Ir_t$ trung bình tại mỗi quý sau đó điều chỉnh chỉ số này để ra được nhân tố mùa Sn_t. Ngoài ra, Wilson & Keating (2007) đề nghị sử dụng chuỗi CMA_t để dự báo yếu tố xu thế dài hạn theo thứ tự thời gian t. Khi ấy, giá trị dự báo của CMA_t gọi là $CMAT_t$ *(Centered Moving-Average Trend)*. Thành phần Chu kỳ Cl_t sẽ được tính bằng cách $Cl_t=CMA_t/CMAT_t$.

Hanke & Wichern (2005) cũng giới thiệu một cách thức khác để phân tích chuỗi thời gian mà nền tảng cũng hơi khác trong tính toán CMA_t, cách thức điều chỉnh chỉ số mùa *(dựa vào trung vị của Sn.Ir ở từng tháng/hay quý trong năm)*.

Bây giờ, chúng ta sẽ tiếp tục sử dụng Eviews để kiểm định yếu tố mùa với kiểm định Kruskal-Wallis theo cách thức tính CMA của Gaynor& Kirkpatrick.

Dữ liệu đã có sẵn biến Y, biến quarter như Hình 6.33.

- Tính MA, CMA, SnIr bằng cách gõ các dòng lệnh sau:

  genr ma_var=(Y(-1)+Y +Y(1)+Y(2))/4

  (không dùng tên biến là MA vì tên này trung tên với một lệnh của Eviews)

  genr cma_var=(ma_var+ma_var(-1))/2

  genr snir_mul=y/cma_var

Công thức chung được đề cập trong Gaynor & Kirkpatrick (1994) để tính MA, CMA với khoảng trượt L chẵn (L = 4 nếu dữ liệu theo quý, L = 12 nếu dữ liệu theo tháng):

$$MA_t = \frac{Y_{t-(L/2)+1} + \ldots + Y_t + \ldots + Y_{t+(L/2)}}{L} \quad (6.5)$$

và

$$CMA_t = \frac{MA_{t-1} + MA_t}{2} \quad (6.6)$$

■ HÌNH 6.33: **Tính CMA của Gaynor& Kirkpatrick.**

EViews

File Edit Object View Proc Quick Options Window Help

genr y=na
genr quarter=@quarter

Series: Y Workfile: UNTITLED\Untitled

View | Proc | Object | Properties | Print | Name | Freeze | Default

Y

Range: 2
Sample: 2

Last upda
Modified: 2C

c
quarter
resid
y

2004Q1	2004Q1	64.2
2004Q2	2004Q2	75.7
2004Q3	2004Q3	117.1
2004Q4	2004Q4	72.4
2005Q1	2005Q1	68.4
2005Q2	2005Q2	90.0
2005Q3	2005Q3	139.3
2005Q4	2005Q4	84.7
2006Q1	2006Q1	97.6
2006Q2	2006Q2	120.0
2006Q3	2006Q3	184.7
2006Q4	2006Q4	101.9
2007Q1	2007Q1	125.2
2007Q2	2007Q2	160.0
2007Q3	2007Q3	237.2
2007Q4	2007Q4	143.4

Bảng 6.9 là kết quả của MA, CMA, Sn.Ir từ tính toán của Eviews cho mô hình nhân tính.

■ BẢNG 6.9: **Tính nhân tố mùa, chuỗi dữ liệu điều chỉnh mùa.**

obs	QUARTER	Y	MA_VAR	CMA_VAR	SNIR_MUL
2004Q1	1	64.2			
2004Q2	2	75.7	82.4		
2004Q3	3	117.1	83.7	83.0	1.41
2004Q4	4	72.4	87.2	85.4	0.85
2006Q1	1	97.6	121.8	116.1	0.84
...	...	...	...	...	...
2006Q3	3	184.7	133.0	129.5	1.43
2006Q4	4	101.9	143.0	138.0	0.74
2007Q1	1	125.2	156.1	149.5	0.84
2007Q2	2	160	166.5	161.3	0.99
2007Q3	3	237.2			
2007Q4	4	143.4			

- Chọn chuỗi **Snir_mul** và thực hiện kiểm định Kruskal-Wallis, ta được kết quả như Bảng 6.10 (rút gọn). Bảng này cho thấy Prob(Kruskal-Wallis)=0.0249 (<0.05) nên ở độ tin cậy 95% chuỗi dữ liệu có tồn tại yếu tố mùa.

■ BẢNG 6.10: **Kiểm định Kruskal-Wallis với chuỗi Snir-mul.**

Test for Equality of Medians of SNIR_MUL
Categorized by values of QUARTER
Date: 06/14/09 Time: 17:47
Sample (adjusted): 2004Q3 2007Q2
Included observations: 12 after adjustments

Method	df	Value	Probability
Med. Chi-square	3	12.00000	0.0074
Adj. Med. Chi-square	3	5.333333	0.1490
Kruskal-Wallis	3	9.358974	**0.0249**
Kruskal-Wallis (tie-adj.)	3	9.358974	0.0249
Van Der Waerden	3	8.937055	0.0301

Nếu như, chúng ta không tính ra chuỗi **snir_mul** theo cách của Gaynor, hay cách thức theo công thức của Eviews mà thực hiện kiểm định Kruskal-Wallis trên ngay **chuỗi dữ liệu ban đầu (biến Y)** thì chúng sẽ nhận được kết quả như bảng 6.11. Bảng này cho thấy Prob(Kruskal-Wallis)=0.126. Con số này nói lên điều gì? Con số này cho biết khi kiểm định tính mùa vụ, chúng ta cần chạy trên chuỗi Sn.Ir. Dù rằng chuỗi Sn.Ir được tính theo công thức nào đi nữa!

■ BẢNG 6.11: **Kiểm định Kruskal-Wallis với chuỗi Y.**

Test for Equality of Medians of Y
Categorized by values of QUARTER
Date: 06/14/09 Time: 17:53
Sample: 2004Q1 2007Q4
Included observations: 16

Method	df	Value	Probability
Med. Chi-square	3	6.000000	0.1116
Adj. Med. Chi-square	3	3.000000	0.3916
Kruskal-Wallis	**3**	**5.713235**	**0.1264**
Kruskal-Wallis (tie-adj.)	3	5.713235	0.1264
Van Der Waerden	3	6.137469	0.1051

TÓM TẮT CHƯƠNG 6

Khi nói đến các kỹ thuật dự báo, người ta không thể bỏ qua phương pháp phân tích, bởi vì điều đơn giản là phương pháp này đã giúp cho những nhà chuyên môn hoặc thậm chí là những người ít am hiểu dự báo biết rằng một chuỗi thời gian chúng ta có thể quan sát chúng qua bốn thành phần cơ bản là: mùa vụ, xu thế, chu kỳ, và các dao động ngẫu nhiên. Hiện nay, phương pháp này vẫn còn phổ biến và không ngừng được các tác giả khác phát triển. Wilson & Keating (2007) đề cập đến ba lý do chính khiến phương pháp này mặc dù đã ra đời từ rất lâu nhưng nó vẫn được sử dụng phổ biến hiện nay. Thứ nhất, trong nhiều trường hợp, các mô hình phân tích chuỗi thời gian cung cấp những kết quả dự báo thật tuyệt vời vì tính phân tích cụ thể của nó. Thứ hai, các mô hình này tạo sự dễ dàng trong việc hiểu và giải thích kết quả dự báo với người sử dụng kết quả dự báo; nó làm tăng sự hợp lý trong việc giải thích kết quả dự báo cũng như sử dụng kết quả dự báo. Thứ ba, thông tin được cung cấp bởi phân tích chuỗi thời gian phù hợp với cách thức mà nhà quản lý xem xét dữ liệu, và giúp họ có được những hiểu biết sâu hơn về sự vận động của dữ liệu bởi nó cung cấp những đo lường cụ thể cho các thành phần mà không định lượng được bởi các phương pháp khác. Sau cùng, theo nhóm tác giả của cuốn sách này, thì phương pháp này có thể không cần tới phần mềm chuyên biệt như Eviews hoặc SPSS, mà chúng có thể diễn đạt trực tiếp trên phần mềm thông dụng Excel.

CÂU HỎI VÀ BÀI TẬP

1. Anh/Chị cho biết phương pháp phân tích các thành phần chuỗi thời gian là gì? Phương pháp này được sử dụng chủ yếu cho các loại dữ liệu gì?
2. Vào đầu năm 2008, một hãng du lịch muốn tìm hiểu nhu cầu của khách quốc tế từ Úc đến Việt Nam để xây dựng kế hoạch kinh doanh của mình. Họ đã thu thập được số liệu về số lượt khách quốc tế từ Úc đến Việt Nam (đơn vị tính là lượt khách, lưu trong biến KQT_UC) từ quý 1 năm 2004 đến quý 4 năm 2007. Dữ liệu về du khách Úc được cho trong tập tin **"TOURIST_A.xls"**.
   a. Anh/Chị hãy vẽ đồ thị thể hiện sự biến động của KQT_UC theo thời gian bằng Excel. Theo Anh/Chị, chuỗi dữ liệu KQT_UC biến động theo mô hình nhân hay mô hình cộng? Tại sao?
   b. Giả sử chuỗi dữ liệu KQT_UC biến động theo mô hình nhân. Bằng Excel, Anh/Chị hãy áp dụng công thức của **Gaynor & Kirkpatrick** để tính toán các cột MA, CMA, Sn.Ir, Sn, và Tr.Cl.Ir.
   c. Anh/Chị hãy Copy các cột dữ liệu Sn.Ir sang Eviews, sử dụng cột dữ liệu này để kiểm định tính mùa vụ của chuỗi dữ liệu KQT_UC.
   d. Chuỗi dữ liệu KQT_UC sau khi đã được loại trừ yếu tố mùa vụ có yếu tố xu thế (Chuỗi Tr.Cl.Ir). Bằng Excel, Anh/Chị hãy ước lượng hàm xu thế tuyến tính để mô tả sự biến động chuỗi dữ liệu KQT_UC (sau khi loại yếu tố mùa) theo thứ tự của thời gian.
   e. Nếu áp dụng mô hình nhân tính, và yếu tố xu thế của chuỗi dữ liệu KQT_UC (sau khi loại trừ yếu tố mùa) là tuyến tính. Bằng Excel, Anh/Chị sẽ dự báo Khách quốc tế Úc đến Việt Nam vào các quý năm 2008 là bao nhiêu?
3. Với dữ liệu đã cho về Khách quốc tế Úc đến Việt Nam. Bằng Eviews, Anh/Chị hãy thực hiện các công việc sau:
   a. Vẽ đồ thị biến động của KQT_UC đến Việt Nam theo thời gian và nhận xét?

b. Áp dụng mô hình nhân, tách yếu tố mùa vụ ra khỏi chuỗi dữ liệu, lưu lại chuỗi nhân tố mùa vụ, chuỗi dữ liệu sau khi đã loại yếu tố mùa vụ?

c. Đưa ra kết quả dự báo với hàm xu thế có dạng tuyến tính?

d. Đưa ra kết quả dự báo với hàm xu thế có dạng bậc hai?

e. Đưa ra kết quả dự báo với hàm xu thế có dạng tăng trưởng mũ?

f. Theo Anh/Chị, nên áp dụng mô hình xu thế nào để dự báo Khách quốc tế Úc đến Việt Nam cho năm 2008 (giả định rằng thị trường du lịch năm 2008 không có sự biến động lớn), và kết quả dự báo sẽ như thế nào?

g. Nếu thị trường du lịch năm 2008 có sự biến động không nhỏ do chịu ảnh hưởng của suy thoái kinh tế toàn cầu, và theo ý kiến của các chuyên gia, mặc dù ngành du lịch sẽ thực hiện rất nhiều biện pháp để kích cầu du lịch nhưng giá trị dự báo về KQT_UC ở các quý năm 2009 nên giảm đi 15% so với giá trị dự báo mà bạn đưa ra ở phần f, thì kết quả dự báo sẽ như thế nào?

4. Vào đầu năm 2009, một Công ty hàng không muốn dự báo khách Nhật Bản đến đến Việt Nam du lịch, công việc, thăm thân nhân cũng như các mục đích khác để xây dựng kế hoạch kinh doanh của mình. Công ty này đã thu thập được số liệu về số lượt khách quốc tế từ Nhật Bản đến Việt Nam (đơn vị tính là lượt khách, lưu trong biến KQT_NB) từ quý 1 năm 2004 đến quý 4 năm 2008. Dữ liệu về du khách Nhật Bản đến Việt Nam được cho trong tập tin **"TOURIST_J.xls"**.

a. Theo Anh/Chị, chuỗi dữ liệu KQT_NB có tồn tại yếu tố mùa vụ hay không?

b. Giả định rằng môi trường dự báo trong năm 2008 ít có sự biến động lớn, theo bạn nên áp dụng mô hình cộng hay mô hình nhân tính để dự báo.

c. Theo Anh/Chị kết quả dự báo điểm, dự báo khoảng ở độ tin cậy 95% ở các quý năm 2009 là bao nhiêu?

5. Có số liệu về tổng giá trị xuất khẩu hàng hóa của Việt Nam (biến XK_VN, có đơn vị là triệu USD) từ quý 1 năm 2004, đến quý 4 năm 2007. Dữ liệu này có sẵn trong tập tin **"EXPORT.xls"**.
   a. Bằng kiểm định Kruskal-Wallis, Anh/Chị hãy kiểm định xem chuỗi dữ liệu KX_VN có tồn tại yếu tố mùa không?
   b. Giả sử thời điểm hiện tại là đầu năm 2008, Anh/Chị hãy đưa ra kết quả dự báo về tổng giá trị xuất khẩu hàng hóa của Việt Nam ở các quý năm 2008?
6. Trong quá trình nghiên cứu thị trường thủy sản xuất khẩu, một công ty muốn dự báo tổng giá trị xuất khẩu thủy sản của Việt Nam, và họ đã thu thập được dữ liệu về tổng giá trị xuất khẩu thủy sản trong quá khứ (biến XK_thuysan, đơn vị tính là triệu USD). Dữ liệu lượng thủy sản xuất khẩu được cho trong tập tin **"FISH_EXPORT.xls"**.
   a. Vẽ đồ thị biểu diễn XK_thuysan theo thời gian.
   b. Hãy áp dụng mô hình cộng để dự báo giá trị xuất khẩu thủy sản của Việt Nam cho các quý năm 2008.
7. Hiệp hội dệt may Việt Nam dự định đưa ra kết quả dự báo về giá trị xuất khẩu sản phẩm dệt may cho tương lai (năm 2008). Số liệu về giá trị xuất khẩu sản phẩm dệt may (biến XK_detmay, đơn vị tính là triệu USD) theo quý trong quá khứ được cho trong tập tin **"TEXTILE.xls"**.
   a. Nếu áp dụng dự báo bằng phương pháp phân tích các thành phần của chuỗi thời gian cho chuỗi dữ liệu XK_detmay, theo Anh/Chị, chuỗi dữ liệu trên có yếu tố mùa không?
   b. Nếu sử dụng mô hình cộng để dự báo, kết quả dự báo cho các quý năm 2008 sẽ như thế nào?
8. Có số liệu về sản lượng dầu thô xuất khẩu của Việt Nam như bảng trên (biến XK_dautho, nghìn tấn), tập tin **"VN_OIL.xls"**.
   a. Theo Anh/Chị, chuỗi dữ liệu XK_dautho có yếu tố mùa không?
   b. Giả sử thời điểm hiện tại là đầu năm 2008, Anh/Chị sẽ dự báo sản lượng xuất khẩu dầu thô của Việt Nam ở các quý năm 2008 là bao nhiêu?

9. Bạn hãy vào trang Web của Tổng cục Thống kê: http://www.gso.gov.vn. Từ trang này, Anh/Chị có thể có được số liệu theo từng tháng của rất nhiều chỉ tiêu ở nhiều năm trong quá khứ.
   a. Hãy chọn một chỉ tiêu mà Anh/Chị quan tâm, sau đó thu thập dữ liệu cho chỉ tiêu này theo từng tháng trong quá khứ (trong vòng khoảng 4 đến 5 năm gần đây).
   b. Hãy dự báo cho chỉ tiêu ấy trong 6 tháng tiếp theo với phương pháp phân tích các thành phần của chuỗi thời gian phù hợp.

10. Bạn hãy vào trang Web của một công ty chứng khoán, thu thập số liệu theo quý về doanh thu của một công ty đã được niêm yết trên thị trường chứng khoán trong khoảng 4 đến 5 năm gần đây.
   a. Dùng đồ thị, kiểm định Kruskal-Wallis để xem xét xem chuỗi dữ liệu về doanh thu của công ty đó có yếu tố mùa hay không
   b. Anh/Chị hãy áp dụng mô hình dự báo bằng phương pháp phân tích phù hợp để dự báo doanh thu của công ty đó cho 6 tháng tiếp theo trong tương lai.

11. Tập tin **"GASOLINE.xls"** chứa chuỗi dữ liệu về nhu cầu sử dụng dầu theo tháng của một công ty dầu khí. Anh/Chị hãy trả lời những câu hỏi sau đây:
   a. Vẽ đồ thị nhu cầu sử dụng dầu theo thời gian và cho biết dữ liệu này có thể phù hợp với mô hình cộng tính hay nhân tính? Tại sao?
   b. Anh/Chị hãy thực hiện phân tích thành phần chuỗi thời gian theo mô hình được chọn từ câu a?
   c. Anh/Chị giải thích các chỉ số mùa vụ như thế nào?
   d. Anh/Chị hãy dự báo nhu cầu sử dụng dầu của công ty trong ba tháng còn lại của năm 2008?

12. Anh/Chị hãy áp dụng phương pháp phân tích thành phần chuỗi thời gian để dự báo số lượng khách hàng mới của công ty CCC trong năm 1996? Anh/Chị cho biết kết quả dự báo này có tốt hơn so với các mô hình trước đây ở chương 4 và chương 5 hay không? Tại sao?

13. Anh/Chị hãy áp dụng phương pháp phân tích thành phần chuỗi thời gian để dự báo doanh số của công ty Murphy Brothers trong năm 1996? Anh/Chị cho biết kết quả dự báo này có tốt hơn so với các mô hình trước đây ở chương 4 và chương 5 hay không? Tại sao?

14. Anh/Chị hãy áp dụng phương pháp phân tích thành phần chuỗi thời gian để dự báo doanh số của GAP trong năm 2004? Anh/Chị cho biết kết quả dự báo này có tốt hơn so với các mô hình trước đây ở chương 4 và chương 5 hay không? Tại sao?

15. Anh/Chị hãy áp dụng phương pháp phân tích thành phần chuỗi thời gian để dự báo giá CP tháng 6/2009? Anh/Chị cho biết kết quả dự báo này có tốt hơn so với các mô hình trước đây ở chương 4 và chương 5 hay không? Tại sao?

CHƯƠNG 7 DỰ BÁO BẰNG PHÂN TÍCH HỒI QUY

Chúng ta vừa khảo sát một số mô hình dự báo giản đơn thuộc nhóm các mô hình dự báo chuỗi thời gian. Như chúng tôi đã đề cập ở chương 1; mô hình dự báo chuỗi thời gian sẽ giúp dự báo các giá trị tương lai về một đối tượng dự báo nào đó trên nền tảng xu hướng vận động của chính chuỗi dữ liệu đó trong quá khứ và hiện tại. Tuy nhiên, các biến kinh tế thường có các mối quan hệ với nhau, và dựa trên các mối quan hệ đó mà chúng ta có thể suy luận được hành vi của một biến số nào đó khi đã có thông tin từ các biến số khác có liên quan. Chẳng hạn, các nhà hoạch định chính sách vĩ mô có thể dự báo được tốc độ tăng trưởng kinh tế trên cơ sở dự đoán được các thông tin tương lai về cung tiền, lãi suất, hay chi tiêu công. Hoặc các nhà nghiên cứu có thể dự đoán được mức độ chi tiêu của dân cư cho một nhóm hàng hóa nào đó trên cơ sở dự đoán xu hướng gia tăng thu nhập và trình độ học vấn. Hoặc giám đốc kinh doanh của một doanh nghiệp có thể dự đoán được doanh số trong tương lai trên cơ sở dự trù các khoản chi tiêu cho quảng cáo và chi tiêu cho nghiên cứu thị trường. Để có thể làm được như vậy, các phương pháp phân tích hồi quy trở thành một trong những công cụ vô cùng hữu ích. Ngoài ra, phân tích hồi quy còn giúp những người nghiên cứu kiểm chứng nhiều giả thiết kinh tế quan trọng nhằm có thêm thông tin chắc chắn cho việc ra quyết định về chính sách hay giải pháp nào đó. Hơn nữa, chúng ta sẽ tiếp tục tìm hiểu một số mô hình dự báo chuỗi thời gian phức tạp ở các chương sau, và các mô hình đó sẽ không thể nào thực hiện được nếu người phân tích không được trang bị một nền tảng tương đối về phân tích hồi quy.

MỤC TIÊU HỌC TẬP

Chương này giúp chúng ta hiểu được các vấn đề cơ bản nhất về phân tích hồi quy và các ứng dụng của phân tích hồi quy trong dự báo với các nội dung sau đây:

- Các vấn đề cơ bản về phân tích hồi quy.
- Giải thích ý nghĩa thống kê của các kết quả hồi quy.
- Thực hiện các kiểm định giả thiết quan trọng.
- Giải thích ý nghĩa kinh tế của các kết quả hồi quy.
- Nhận biết và khắc phục một số vấn đề thường gặp trong phân tích hồi quy.
- Một số ứng dụng của phân tích hồi quy trong việc ra quyết định về chính sách và dự báo.

MÔ HÌNH HỒI QUY ĐƠN

MỤC ĐÍCH CỦA PHÂN TÍCH HỒI QUY

Theo Gujarati (2003), phân tích hồi quy có thể giúp người phân tích:

- Ước lượng giá trị trung bình của biến phụ thuộc khi cho trước giá trị một hoặc các biến giải thích.
- Kiểm định các giả thiết về bản chất của sự phụ thuộc giữa biến độc lập và biến phụ thuộc.
- Dự báo giá trị trung bình của biến phụ thuộc khi cho trước các giá trị của các biến giải thích.
- Dự báo tác động biên hoặc độ co giãn của một biến độc lập lên biến phụ thuộc thong qua hệ số hồi quy.

MÔ HÌNH HỒI QUY TUYẾN TÍNH CỔ ĐIỂN

Mô hình hồi quy tuyến tính cổ điển là một cách xem xét bản chất và hình thức của mối quan hệ giữa hai hay nhiều biến số. Trong phần

này, chúng ta chỉ tập trung xem xét trường hợp mô hình hai biến. Trong đó Y là biến phụ thuộc và X là biến độc lập (hay còn gọi là biến giải thích). Như vậy, chúng ta muốn giải thích/dự báo giá trị của Y theo các giá trị khác nhau của X. Giả sử, X và Y có mối quan hệ tuyến tính như sau:

$$E(Y_t) = \beta_1 + \beta_2 X_t \quad (7.1)$$

Trong đó, $E(Y_t)$ là giá trị trung bình có điều kiện của Y_t theo X_t, và β_1, β_2 là các tham số chưa biết của tổng thể (t ký hiệu theo thông lệ dữ liệu chuỗi thời gian cho quan sát vào thời điểm t của biến quan sát). Phương trình (7.1) được gọi là phương trình hồi quy tổng thể. Giá trị thực Y_t sẽ không phải luôn luôn bằng giá trị kỳ vọng $E(Y_t)$, vì vậy Y_t có thể được thể hiện như sau:

$$Y_t = E(Y_t) + u_t$$

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad (7.2)$$

Trong đó, u_t được gọi là hạng nhiễu ngẫu nhiên. Và u_t luôn tồn tại do các nguyên nhân như bỏ sót biết giải thích, sai dạng mô hình do bỏ qua các tác động trễ, sai dạng hàm, lỗi đo lường, hoặc do đơn giản hóa mô hình bằng cách tổng hợp một số biến khác nhau thành một biến giải thích duy nhất.

PHƯƠNG PHÁP BÌNH PHƯƠNG BÉ NHẤT

Phương pháp được sử dụng phổ biến nhất nhằm ước lượng các hệ số hồi quy là **phương pháp bình phương bé nhất thông thường** (OLS)[1]. Theo Gujarati (2003), dưới các giả định của mô hình hồi quy tuyến tính cổ điển (sẽ trình bày ở phần sau), thì phương pháp OLS có nhiều tính chất thống kê rất hấp dẫn làm cho nó trở thành một phương pháp mạnh và phổ biến nhất trong phân tích hồi quy. Phương pháp OLS được cho là của nhà toán học nổi tiếng người Đức Carl Friedrich Gauss.

[1] Ordinary least squares.

Nhắc lại hàm hồi quy tổng thể ở phương trình (7.2):

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad (7.2)$$

Do hàm hồi quy tổng thể này không thể quan sát trực tiếp được, nên ta ước lượng nó từ hàm hồi quy mẫu từ phương trình (7.3):

$$\begin{aligned} Y_t &= \hat{\beta}_1 + \hat{\beta}_2 X_t + \hat{u}_t \quad (7.3) \\ &= \hat{Y}_t + \hat{u}_t \end{aligned}$$

Trong đó, Y_t là giá trị quan sát thực tế, $\hat{Y}_t$ là giá trị ước lượng hay trung bình có điều kiện của Y_t. Ta có

$$\begin{aligned} \hat{u}_t &= Y_t - \hat{Y}_t \\ &= Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_t \quad (7.4) \end{aligned}$$

Phương trình này cho biết phần dư $\hat{u}_t$ là hiệu số của giá trị Y thực tế và giá trị Y ước lượng vào thời điểm *t*, giá trị này có từ phương trình (7.3).

Xây dựng các hệ số của hàm hồi quy mẫu với điều kiện bình phương tổng phần dư $\sum \hat{u}_t = \sum (Y_t - \hat{Y}_t)$ là tối thiểu nhất. Nghĩa là, nghĩa là xác định $\hat{\beta}_1$ và $\hat{\beta}_2$ sao cho tổng bình phương phần dư $\sum \hat{u}_t^2$ (được gọi là RSS) là tối thiểu. RSS được định nghĩa như sau:

$$RSS = \sum_{t=1}^{n} \hat{u}_t^2 = \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{n} (Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_t)^2 \quad (7.5)$$

Để tối thiểu hóa (7.5), ta lấy đạo hàm bậc một của RSS theo $\hat{\beta}_1$ và $\hat{\beta}_2$ và cho các đạo hàm này bằng không.

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2\sum (Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_t) = 0 \quad (7.6)$$

$$\frac{\partial RSS}{\partial \hat{\beta}_2} = -2\sum (Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_t) X_t = 0 \quad (7.7)$$

Hai phương trình (7.6) và (7.7) có thể được viết lại như sau:

$$\sum Y_t = n\hat{\beta}_1 - \hat{\beta}_2 \sum X_t \tag{7.8}$$

$$\sum X_t Y_t = \hat{\beta}_1 \sum X_t + \hat{\beta}_2 \sum X_t^2 \tag{7.9}$$

Trong đó n là số quan sát trong mẫu. Hệ hai phương trình (7.8) và (7.9) có thể được biểu diễn dưới hình thức ma trận như sau:

$$\underbrace{\begin{bmatrix} n & \sum X_t \\ \sum X_t & \sum X_t^2 \end{bmatrix}}_{A_{2,2}} \underbrace{\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}}_{B_{2,1}} = \underbrace{\begin{bmatrix} \sum Y_t \\ \sum Y_t X_t \end{bmatrix}}_{C_{2,1}} \tag{7.10}$$

Có thể giải nhanh hệ phương trình (7.10) theo quy tắc Cramer để có $\hat{\beta}_1$ và $\hat{\beta}_2$ như sau:

$$\hat{\beta}_1 = \frac{\sum X_t^2 \sum Y_t - \sum X_t \sum Y_t X_t}{n\sum X_t^2 - (\sum X_t)^2} \tag{7.11}$$

$$\hat{\beta}_2 = \frac{n\sum Y_t X_t - \sum X_t \sum Y_t}{n\sum X_t^2 - (\sum X_t)^2} \tag{7.12}$$

Tuy nhiên, các công thức ước tính $\hat{\beta}_1$ và $\hat{\beta}_2$ như trên có vẻ hơi phức tạp nên rất dễ làm người đọc (nhất là sinh viên năm 2 và năm 3 các ngành kinh tế) ngao ngán vì tính phức tạp của nó. Từ phương trình (7.8) ta có:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} \tag{7.13}$$

Thế $\hat{\beta}_1$ ở phương trình (7.13) vào phương trình (7.9) để tìm $\hat{\beta}_2$ như sau:

$$\Sigma Y_t X_t = (\overline{Y} - \hat{\beta}_2 \overline{X}) \Sigma X_t + \hat{\beta}_2 \Sigma X_t^2$$

$$\Sigma Y_t X_t = \overline{Y} \sum X_t - \hat{\beta}_2 \overline{X} \sum X_t + \hat{\beta}_2 \Sigma X_t^2$$

Do $\sum X_t = n\overline{X}$, nên ta có:

$$\Sigma Y_t X_t = n\overline{YX} - n\hat{\beta}_2 \overline{X}^2 + \hat{\beta}_2 \Sigma X_t^2$$

$$\Sigma Y_t X_t - n\overline{YX} = \hat{\beta}_2 \left(\Sigma X_t^2 - n\overline{X}^2 \right) \qquad (7.14)$$

Ta lại có :

$$\begin{aligned} \Sigma (X_t - \overline{X})(Y_t - \overline{Y}) &= \Sigma (X_t Y_t - X_t \overline{Y} - \overline{X} Yt + \overline{XY}) \\ &= \Sigma X_t Y_t - \overline{Y} \Sigma X - \overline{X} \Sigma Y_t + \Sigma \overline{XY} \\ &= \Sigma X_t Y_t - n\overline{XY} - n\overline{XY} + n\overline{XY} \\ &= \Sigma X_t Y_t - n\overline{XY} \end{aligned} \qquad (7.15)$$

Và

$$\begin{aligned} \Sigma (X_t - \overline{X})^2 &= \Sigma (X_t^2 - 2X_t \overline{X} + \overline{X}^2) \\ &= \Sigma X_t^2 - 2\overline{X} \Sigma X_t + \Sigma \overline{X}^2 \\ &= \Sigma X_t^2 - 2n\overline{XX} + n\overline{X}^2 \\ &= \Sigma X_t^2 - n\overline{X}^2 \end{aligned} \qquad (7.16)$$

Thế phương trình (7.15) và (7.16) vào phương trình (7.14) ta có:

$$\Sigma (X_t - \overline{X})(Y_t - \overline{Y}) = \hat{\beta}_2 \Sigma (X_t - \overline{X})^2$$

$$\hat{\beta}_2 = \frac{\Sigma (X_t - \overline{X})(Y_t - \overline{Y})}{\Sigma (X_t - \overline{X})^2} \qquad (7.17)$$

$$= \frac{\Sigma x_t y_t}{\Sigma x_t^2}$$

Trong đó, $x_t = (X_t - \overline{X})$ và $y_t = (Y_t - \overline{Y})$. Như vậy, qua một vài bước biến đổi nhỏ ta có công thức ước tính $\hat{\beta}_2$ cực kỳ đơn giản và rất ý nghĩa. Tưởng tượng rằng, lấy cả tử và mẫu của (7.17) chia cho (n-1), ta có:

$$\hat{\beta}_2 = \frac{\text{Cov}(X_t, Y_t)}{\text{Var}(X_t)} \tag{7.18}$$

Ngoài ra, $\hat{\beta}_2$ ở phương trình (7.17) còn có thể được thể hiện một cách khác như sau:

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum x_t y_t}{\sum x_t^2} \\ &= \frac{\sum x_t (Y_t - \overline{Y})}{\sum (X_t - \overline{X})^2} = \frac{\sum x_t Y_t - \overline{Y}\sum x_t)}{\sum X_t^2 - n\overline{X}^2} \\ &= \frac{\sum x_t Y_t - \overline{Y}\sum (X_t - \overline{X})}{\sum X_t^2 - n\overline{X}^2} = \frac{\sum x_t Y_t}{\sum X_t^2 - n\overline{X}^2} \\ &= \frac{\sum x_t Y_t}{\sum X_t^2 - n\overline{X}^2} = \frac{\sum x_t Y_t}{\sum x_t^2}\end{aligned} \tag{7.19}$$

Các công thức ở phương trình (7.17) và (7.19) mách cho chúng ta một điều rất thú vị rằng, $\hat{\beta}_1$ là một hàm tuyến tính theo $\hat{\beta}_2$, $\hat{\beta}_2$ là một hàm tuyến tính theo Y_t, nên cả $\hat{\beta}_1$ và $\hat{\beta}_2$ đều là các hàm tuyến tính theo Y_t. Và Y_t là một hàm tuyến tính theo u_t, vậy $\hat{\beta}_1$ và $\hat{\beta}_2$ là các hàm tuyến tính theo u_t. Cho nên, nếu u_t có phân phối chuẩn thì $\hat{\beta}_1$ và $\hat{\beta}_2$ cũng sẽ có phân phối chuẩn.

CÁC GIẢ ĐỊNH CỦA HỒI QUY TUYẾN TÍNH CỔ ĐIỂN

Theo Gujarati (2003), nếu mục tiêu của ta chỉ là ước lượng các hệ số β_1 và β_2, thì chỉ cần phương pháp OLS là đủ. Nhưng, như ta đã biết, các mục tiêu của phân tích hồi quy không chỉ dừng lại ở việc có được các giá trị ước lượng $\hat{\beta}_1$ và $\hat{\beta}_2$, mà còn phải suy diễn (dự báo khoảng) về các giá trị thực β_1 và β_2 thực sự có ý nghĩa thống kê hay không. Chính vì vậy, chúng ta cần biết cụ thể về bản chất của hàm hồi quy tổng thể. Cụ thể, chúng ta không chỉ xác định dạng hàm của mô hình

hồi quy, mà còn đưa ra các giả định về cách mà Y_t được tạo ra như thế nào. Phương trình (7.2) cho thấy Y_t phụ thuộc vào cả X_t và u_t. Cho nên, nếu ta không biết X_t và u_t được tạo ra như thế nào, thì ta sẽ không có cách nào suy diễn được Y_t cũng như các hệ số β_1 và β_2. Chính vì thế, các giả định về biến giải thích X_t và số hạng nhiễu u_t có ý nghĩa rất quan trọng cho việc giải thích các giá trị ước lượng của hồi quy. Ta đã biết, các hạng nhiễu u_t (không thể quan sát được) là các hạng nhiễu ngẫu nhiên. Do hạng nhiễu u_t cộng với một số hạng phi ngẫu nhiên X_t để tạo ra Y_t, vậy Y_t sẽ là một biến ngẫu nhiên. Dưới đây là tóm tắt các giả định trong mô hình hồi quy tuyến tính cổ điển.

■ BẢNG 7.1: **Giả định của mô hình hồi quy tuyến tính cổ điển,**

Giả định		Biểu diễn dạng toán	Không thỏa mãn do
(1)	Mô hình tuyến tính	$Y_t = \beta_1 + \beta_2 X_t + u_t$	Sai dạng mô hình
(2)	Mô hình được xác định đúng		
(3)	X_t có thể biến thiên	$Var(X_t) \neq 0$	Sai dạng mô hình
(4)	X_t và u_t không tương quan	$Cov(X_t,u_t) = 0$	Tự hồi quy
(5)	Giá trị kỳ vọng của hạng nhiễu bằng không	$E(u_t) = 0$	Sai dạng mô hình
(6)	Không có đa cộng tuyến	$\sum(\delta_i X_{it} + \delta_j X_{jt}) \neq 0, i \neq j$	Đa cộng tuyến
(7)	Phương sai không đổi	$Var(u_t) = \sigma^2$	Phương sai thay đổi
(8)	Không có tương quan chuỗi	$Cov(u_t,u_s) = 0, t \neq s$	Tự tương quan
(9)	Hạng nhiễu phân phối chuẩn	$u_t \sim N(\mu,\sigma^2)$	Outliers

ĐẶC ĐIỂM CỦA CÁC ƯỚC LƯỢNG OLS

Các ước lượng của OLS khi tuân thủ các giả định sẽ đạt được tiêu chuẩn BLUE[2], có nghĩa là ước lượng không chệch, tuyến tính, và tốt

[2] Best linear unbiased estimator.

nhất. Ước lượng là tuyến tính do giá trị ước lượng hệ số hồi quy được biểu diễn tuyến tính theo Y (phương trình 7.20). Giá trị các ước lượng của các hệ số hồi quy là không chệch do kỳ vọng của ước lượng hệ số hồi quy trong hàm hồi quy mẫu bằng với giá trị của hệ số hồi quy trong hàm hồi quy tổng thể (phương trình 7.25 và 7.26), và ước lượng của các hệ số hồi quy là tốt nhất vì phương sai của các hệ số hồi quy của hàm hồi quy mẫu là nhỏ nhất (phương trình 7.29 và 7.30).

Công thức ở phương trình (7.19) có thể được viết lại như sau:

$$\hat{\beta}_2 = \frac{\sum x_t Y_t}{\sum x_t^2} = \sum k_t Y_t \qquad (7.20)$$

trong đó,

$$k_t = \frac{x_t}{\sum x_t^2} \qquad (7.21)$$

Phương trình (7.20) cho thấy $\hat{\beta}_2$ là một ước lượng tuyến tính bởi vì nó là một hàm tuyến tính của Y_t. Nói cách khác, các ước lượng OLS là một trung bình có trọng số của Y_t, với k_t đóng vai vai trò như các trọng số. Tương tự, $\hat{\beta}_2$ cũng là một ước lượng tuyến tính theo Y_t.

$$\begin{aligned}\hat{\beta}_1 &= \overline{Y} - \hat{\beta}_2 \overline{X} \\ &= \overline{Y} - \overline{X}\sum k_t Y_t \end{aligned} \qquad (7.22)$$

Tính chất của k_t

1. Do X_t được giả định là phi ngẫu nhiên (cố định), nên k_t cũng phi ngẫu nhiên.
2. $\sum k_t = 0$ (do $\sum x_t = 0$)
3. $\sum k_t^2 = \frac{1}{\sum x_t^2}$ (do $\sum k_t^2 = \frac{\sum x_t^2}{\sum x_t^2} \cdot \frac{1}{\sum x_t^2}$)
4. $\sum k_t x_t = \sum k_t X_t = 1$

   (do $\sum k_t x_t = \sum k_t (X_t - \overline{X}) = \sum k_t X_t - \overline{X}\sum k_t = \sum k_t x_t$)

Dựa vào các tính chất của k_t ta suy ra các công thức của $\hat{\beta}_1$ và $\hat{\beta}_2$ như sau. Thế công thức $Y_t = \beta_1 + \beta_2 X_i + u_t$ vào công thức (7.20), ta có

$$\begin{aligned}\hat{\beta}_2 &= \sum k_t(\beta_1 + \beta_2 X_t + u_t) \\ &= \beta_1 \sum k_t + \beta_2 \sum k_t X_t + \sum k_t u_t \\ &= \beta_2 + \sum k_t u_t \end{aligned} \quad (7.23)$$

Thế các công thức $\overline{Y} = \beta_1 + \beta_2 \overline{X}$ và công thức $Y_t = \beta_1 + \beta_2 X_t + u_t$ vào công thức (7.22), ta có:

$$\hat{\beta}_1 = \beta_1 - \overline{X} \sum k_t u_t \quad (7.24)$$

Như vậy, $\hat{\beta}_1$ và $\hat{\beta}_2$ là các hàm tuyến tính theo các hạng nhiễu ngẫu nhiên u_t. Chính vì thế $\hat{\beta}_1$ và $\hat{\beta}_2$ sẽ có phân phối theo u_t.

Trung bình của các ước lượng OLS

Từ hai phương trình (7.23) và (7.24), ta thấy rằng nếu lấy giá trị trung bình của các ước lượng $\hat{\beta}_1$ và $\hat{\beta}_2$ ta sẽ có:

$$E(\hat{\beta}_1) = E(\beta_1 - \overline{X} \sum k_t u_t) = \beta_1 \quad (7.25)$$

$$E(\hat{\beta}_2) = E(\beta_2 + \sum k_t u_t) = \beta_2 \quad (7.26)$$

Như vậy, các ước lượng OLS có một tính chất rất quan trọng là có giá trị trung bình đúng bằng giá trị thực của tổng thể. Chính nhờ điều này mà người ta gọi các ước lượng OLS là các ước lượng không chệch.

Phương sai của các ước lượng OLS

Từ định nghĩa về phương sai ta có:

$$\begin{aligned}\text{Var}(\hat{\beta}_2) &= E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2 \\ &= E(\hat{\beta}_2 - \beta_2)^2 \end{aligned} \quad (7.27)$$

Thế công thức (7.26) vào (7.27), ta có:

$$\text{Var}(\hat{\beta}_2) = E(\beta_2 + \sum_{t=1}^{n} k_t u_t - \beta_2)^2$$

$$= E\left(\sum_{t=1}^{n} k_t u_t\right)^2$$

$$= E\left(k_1^2 u_1^2 + k_2^2 u_2^2 + \ldots + k_n^2 u_n^2 + 2k_1 k_2 u_1 u_2 + \ldots + 2k_{n-1} k_n u_{n-1} u_n\right)$$

Do ta giả định phương sai nhiễu không đổi, nên $E(u_t^2) = \sigma^2$ tại mỗi giá trị t và không có tự tương quan nên $E(u_t u_s) = 0$, với $t \neq s$, nên ta có:

$$\text{Var}(\hat{\beta}_2) = k_1^2\sigma^2 + k_2^2\sigma^2 + \ldots + k_n^2\sigma^2$$

$$= \sigma^2 \sum k_t^2 \qquad (7.28)$$

Thế tính chất số (3) của k_t vào (7.28) ta có:

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_t^2} \qquad (7.29)$$

Thực hiện tương tự, ta có:

$$\text{Var}(\hat{\beta}_1) = \frac{\sum X_t^2}{n\sum x_t^2}\sigma^2 \qquad (7.30)$$

Lấy căn bậc hai các phương trình (7.29) và (7.30) ta có các sai số chuẩn của các hệ số hồi quy $\hat{\beta}_1$ và $\hat{\beta}_2$ như sau:

$$\text{se}(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_t^2}} \qquad (7.31)$$

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\sum X_t^2}{n\sum x_t^2}}\sigma \qquad (7.32)$$

Trong đó, σ^2 là một hằng số do ta giả định phương sai nhiễu không đổi. Với một dữ liệu mẫu nhất định thì ta có thể dễ dàng tính được $\sum X_t^2$ và $\sum x_t^2$, trừ σ^2. Nếu có được một giá trị phương sai nhất

định thì các sai số chuẩn của các hệ số hồi quy sẽ có một giá trị xác định. Trên thực tế, ta chỉ có ước lượng của σ^2 được tính theo công thức sau đây:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_t^2}{n-2} \tag{7.33}$$

Ở đây, $\hat{\sigma}^2$ cũng là một ước lượng không chệch của phương sai nhiễu σ^2. Ở công thức (7.33), (n-2) là bậc tự do, ký hiệu là d.f., và $\sum \hat{u}_t^2$ là tổng bình phương phần dư, ký hiệu là RSS. Chắc chắn chúng ta sẽ thắc mắc tại sao bậc tự do của RSS là (n-2), hay bằng số quan sát trong mẫu trừ số hệ số ước lượng trong mô hình hồi quy. Có nhiều cách giải thích số bậc tự do, như ta có thể giải thích đơn giản như sau. Ta thấy rằng, trước khi có thể tính được RSS như ở công thức (7.5), trước tiên ta phải có các hệ số $\hat{\beta}_1$ và $\hat{\beta}_2$ vì các giá trị của Y_t và X_t đã có sẵn từ dữ liệu mẫu. Để ước lượng được $\hat{\beta}_1$ và $\hat{\beta}_2$, ta cần ít nhất hai cặp quan sát (Y_t,X_t) bất kỳ (nghĩa là xác định phương trình đường thẳng qua hai điểm). Như vậy, hai giá trị ước lượng này là hai ràng buộc lên RSS. Nói cách khác, trong tập hợp tất cả các cặp quan sát (Y_t,X_t) trong miền giá trị của mẫu dữ liệu sẽ có ít nhất hai cặp quan sát nào đó nằm trên (hoặc rất gần với) đường hồi quy mẫu. Chính vì thế, phần dư tương ứng sẽ bằng không hoặc rất nhỏ. Như vậy, thực sự giá trị của RSS chỉ do (n-2) giá trị $\hat{u}_t^2$ tạo thành. Như vậy, (n-2) chính là số nguồn thông tin để tính RSS.

Lấy căn bậc hai của công thức (7.33) ta sẽ có sai số chuẩn của giá trị ước lượng hay sai số chuẩn của hồi quy ($\hat{\sigma}$) như sau:

$$\hat{\sigma} = \sqrt{\frac{RSS}{n-2}} \tag{7.34}$$

Đây chính là độ lệch chuẩn của các giá trị Y quanh đường hồi quy mẫu và được sử dụng như một thước đo "mức độ phù hợp" của đường hồi quy so với các giá trị thực tế từ mẫu dữ liệu.

HỆ SỐ XÁC ĐỊNH r^2

Cho đến đây chúng ta đã xem xét xong vấn đề ước lượng các hệ số hồi quy, các sai số chuẩn, và tính chất của các ước lượng OLS. Bây giờ chúng ta sẽ xem xét mức độ phù hợp của đường hồi quy mẫu với dữ liệu thực tế; nghĩa là, ta sẽ xem đường hồi quy mẫu phù hợp với dữ liệu mẫu như thế nào. Hệ số xác định r^2 (cho trường hợp mô hình hồi quy đơn) và R^2 (cho trường hợp mô hình hồi quy bội) là một thước đo chung cho biết một đường hồi quy nhất định sẽ phù hợp với dữ liệu mẫu như thế nào.

Để có thước đo độ phù hợp, trước hết ta cần phân tích giá trị thực Y_t theo các các trị ước lượng và phần dư như ở phương trình (7.3):

$$Y_t = \hat{Y}_t + \hat{u}_t \tag{7.3}$$

Cả trừ cả hai vế của phương trình (7.3) cho $\overline{Y}$, ta có:

$$Y_t - \overline{Y} = \hat{Y}_t - \overline{Y} + \hat{u}_t \tag{7.35}$$

Do chúng ta cần một thước đo về tổng biến thiên của Y_t quanh giá trị trung bình $\overline{Y}$, nên phương trình (7.35) được viết lại như sau:

$$\sum(Y_t - \overline{Y}) = \sum(\hat{Y}_t - \overline{Y} + \hat{u}_t) \tag{7.36}$$

Lấy bình phương hai vế của (7.36), ta có:

$$\sum(Y_t - \overline{Y})^2 = \sum(\hat{Y}_t - \overline{Y} + \hat{u}_t)^2 \tag{7.37}$$

Tương đương với:

$$\sum y_t^2 = \sum(\hat{y}_t^2 + \hat{u}_t)^2 \tag{7.38}$$

$$= \sum \hat{y}_t^2 + \sum \hat{u}_t^2 + 2\sum \hat{y}_t \hat{u}_t \tag{7.39}$$

Do $\sum \hat{y}_t \hat{u}_t = 0$ và $\hat{y}_t = \hat{\beta}_2 x_t$, nên phương trình (7.39) có thể được viết lại như sau:

$$\sum y_t^2 = \sum \hat{y}_t^2 + \sum \hat{u}_t^2$$

$$= \hat{\beta}_2^2 \sum x_t^2 + \sum \hat{u}_t^2 \tag{7.40}$$

Trong đó, $\sum y_t^2 = \sum (Y_t - \overline{Y})^2$ là tổng biến thiên của giá trị Y thực tế quanh giá trị trung bình mẫu và được gọi là **tổng bình phương** (TSS). $\sum \hat{y}_t^2 = \sum (\hat{Y}_t - \overline{\hat{Y}})^2 = \sum (\hat{Y}_t - \overline{Y})^2 = \hat{\beta}_2^2 \sum x_t^2$ là tổng biến thiên của giá trị Y ước lượng quanh giá trị ước lượng trung bình ($\overline{\hat{Y}} = \overline{Y}$) và được gọi là tổng bình phương được giải thích bởi hàm hồi quy, hay đơn giản hơn là **tổng bình phương phần được giải thích** (ESS). $\sum \hat{u}_t^2$ là tổng biến thiên phần dư hay phần không được giải thích của các giá trị Y quanh đường hồi quy, hay đơn giản là tổng bình phương phần dư (RSS). Như vậy, phương trình (7.40) được viết lại như sau:

$$TSS = ESS + RSS \tag{7.41}$$

Điều này có nghĩa rằng biến thiên trong các giá trị Y quan sát quanh giá trị trung bình mẫu có thể được chia thành hai phần, một đại diện cho đường hồi quy và một đại diện cho các yếu tố ngẫu nhiên bởi vì không phải tất cả các quan sát thực của Y đều nằm trên đường hồi quy. Ta có thể biểu diễn minh họa một giá trị Y quan sát bất kỳ như Hình (7.1).

Chia cả hai vế của phương trình (7.41) cho TSS, ta có:

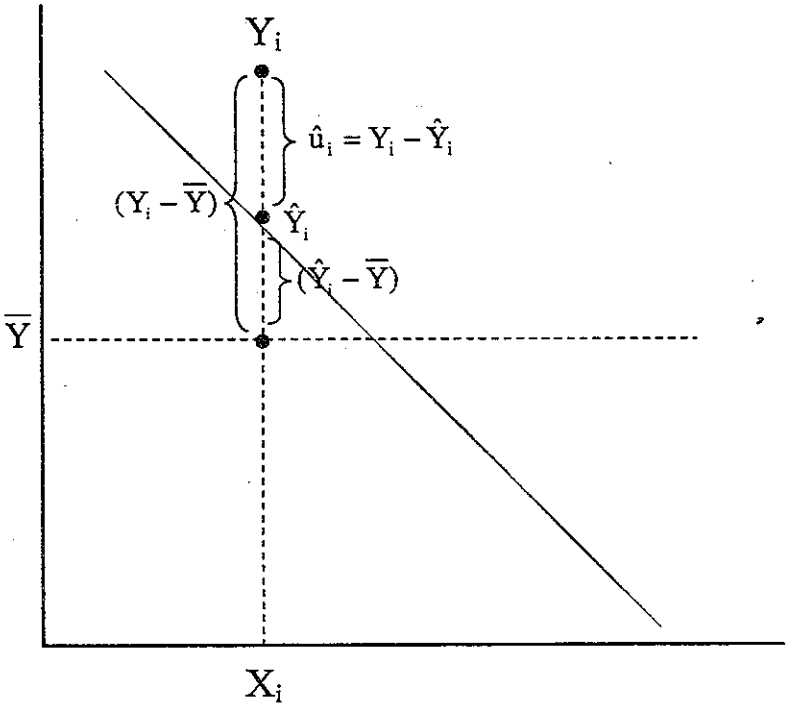
$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} \tag{7.42}$$

Bây giờ ta định nghĩa r^2 như sau:

$$r^2 = \frac{ESS}{TSS}$$

$$= 1 - \frac{RSS}{TSS} \tag{7.43}$$

■ HÌNH 7.1: **Các thành phần trong biến thiên của Y.**



Tóm lại, r^2 được biết như hệ số xác định và là thước đo được sử dụng phổ biến nhất về mức độ phù hợp của hàm hồi quy mẫu với dữ liệu quan sát.

Đặc điểm của hệ số xác định

1. r^2 là một đại lượng không âm.
2. $0 \le r^2 \le 1$. Nếu $r^2 = 1$, thì đường hồi quy phù hợp hoàn toàn; nghĩa là, $\hat{Y}_t = Y_t$ với mỗi *t*. Ngược lại, nếu $r^2 = 0$, thì không có mối quan hệ nào giữa biến giải thích và biến phụ thuộc.

Hệ số xác định r^2 còn được tính theo công thức sau đây:

$$r^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}_t^2}{\sum y_t^2} = \frac{\hat{\beta}_2^2 \sum x_t^2}{\sum y_t^2} = \hat{\beta}_2 \left(\frac{\sum x_t^2}{y_t^2} \right) \quad (7.44)$$

Nếu ta chia cả tử và mẫu của phương trình (6.98) cho (n-1), thì ta có:

$$r^2 = \hat{\beta}_2^2 \left(\frac{Var(X_t)}{Var(Y_t)} \right) = \hat{\beta}_2^2 \left(\frac{S_x^2}{S_y^2} \right) \quad (7.45)$$

Với S_x^2 và S_y^2 là các phương sai mẫu của X_t và Y_t trong mẫu dữ liệu có sẵn. Ngoài ra, ta biết rằng $\hat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2}$, nên phương trình (7.45) có thể được biến đổi như sau:

$$r^2 = \frac{(\sum x_t y_t)^2}{(\sum x_t^2)^2} \frac{\sum x_t^2}{\sum y_t^2} = \frac{(\sum x_t y_t)^2}{\sum x_t^2 \sum y_t^2} = \left(\frac{\sum x_t y_t}{\sqrt{\sum x_t^2 \sum y_t^2}} \right)^2 = (r_{XY})^2 \quad (7.46)$$

Trong đó r_{xy} là hệ số tương quan của biến phụ thuộc Y và biến độc lập X.

Một số vấn đề cần lưu ý khi sử dụng hệ số xác định

1. *Vấn đề hồi quy giả mạo*[3]. Trong trường hợp hai hoặc nhiều biến thực sự không có mối tương quan gì, nhưng bản thân chúng có thể tồn tại yếu tố xu thế mạnh (thường ở dữ liệu chuỗi thời gian), nên các giá trị r^2 (R^2) rất cao (đôi khi cao hơn 0.9). Nếu điều này xảy ra, chúng ta có thể bị ngộ nhận về mối quan hệ thực sự giữa các biến là quan trọng.

2. *Tương quan mạnh giữa các biến giải thích (hồi quy bội).* Trong trường hợp hồi quy bội, nếu các biến giải thích có tương quan với nhau (được gọi là hiện tượng đa cộng tuyến), thì giá trị R^2 thường rất cao. Điều này có thể dẫn đến sự nhầm lẫn trong việc cho rằng đường hồi quy rất phù hợp với dữ liệu.

[3] Spurious regression.

3. *Tương quan không nhất thiết hàm ý quan hệ nhân quả.* Cho dù giá trị R^2 cao bao nhiêu đi nữa, thì nó cũng không thể nói lên có mối quan hệ nhân quả giữa Y_t và X_t vì R^2 là một thước đo mối quan hệ giữa giá trị Y_t quan sát với giá trị Y_t ước lượng.

4. *Phương trình dữ liệu chuỗi thời gian với phương trình dữ liệu chéo.* Các phương trình dữ liệu chuỗi thời gian luôn có các giá trị R^2 cao hơn so với các phương trình dữ liệu chéo. Điều này bởi vì trong dữ liệu chéo chứa đựng rất nhiều sự biến thiên ngẫu nhiên nên làm cho ESS nhỏ tương đối so với TSS. Ngược lại, thậm chí các phương trình chuỗi thời gian được xác định không phù hợp lắm vẫn có thể có R^2 rất cao (có thể 0.999) do hiện tượng hồi quy giả mạo, hoặc do các biến có mối quan hệ tự tương quan.

5. *R^2 thấp không có nghĩa chọn lựa sai biến giải thích X_t.* Giá trị R^2 thấp không nhất thiết do kết quả của việc sử dụng một biến giải thích sai. Dạng hàm được sử dụng có thể không phù hợp (ví dụ tuyến tính chứ không phải bậc hai) hoặc trong trường hợp dữ liệu thời gian thì việc chọn giai đoạn thời gian có thể không chính xác và cũng có thể cần đưa vào mô hình các hạng trễ.

6. *Các giá trị R^2 từ các phương trình với biến phụ thuộc có dạng khác nhau không thể so sánh được.* Ví dụ ta ước lượng hai phương trình hồi quy sau đây:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{7.47}$$

$$\ln Y_t = \beta_1 + \beta_2 \ln X_t + u_t \tag{7.48}$$

Nếu so sánh r^2 của hai phương trình này là không chính xác. Điều này là do cách định nghĩa r^2. Giá trị r^2 của phương trình (7.47) cho biết phần trăm biến thiên trong Y_t được giải thích bởi X_t, trong khi đó r^2 của phương trình (7.48) cho biết phần trăm biến thiên trong logarith tự nhiên của Y_t được giải thích bởi logarith tự nhiên của X_t. Nói chung, bất kỳ khi nào biến phụ thuộc được biến đổi theo các hình thức khác nhau, thì chúng ta không nên sử dụng r^2 để so sánh giữa các mô hình.

KIỂM ĐỊNH GIẢ THIẾT VÀ CÁC KHOẢNG TIN CẬY

Với các giả định hồi quy CLRM thì hạng nhiễu u_t theo phân phối chuẩn, nên các ước lượng OLS cũng theo phân phối. Cụ thể, các ước lượng OLS có thể được biểu hiện như sau:

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}1}) \tag{7.49}$$

$$Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1) \tag{7.50}$$

$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}2}) \tag{7.51}$$

$$Z_2 = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}2}} \sim N(0,1) \tag{7.52}$$

Tuy nhiên, chúng ta thường không biết giá trị của $\sigma_{\hat{\beta}_1}$ và $\sigma_{\hat{\beta}_2}$. Theo lý thuyết thống kê, nếu $\sigma_{\hat{\beta}_1}$ và $\sigma_{\hat{\beta}_2}$ được thay bằng các ước lượng của chúng là $se(\hat{\beta}_1)$ và $se(\hat{\beta}_2)$, thì các biến $t_1 = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$ và $t_2 = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$ sẽ theo phân phối *t* với n-2 bậc tự do (trong trường hợp hồi quy đơn). Như vậy, chúng ta sẽ sử dụng thống kê *t* để kiểm định các giả thiết về các hệ số hồi quy.

Các bước kiểm định ý nghĩa của các hệ số hồi quy OLS

Bước 1: Xác định giả thiết không (H_0) và giả thiết khác (H_1 hoặc H_a). Thông thường, H_0: $\beta_2 = 0$; H_1: $\beta_2 \neq 0$ (kiểm định hai đuôi), hoặc nếu biết trước thông tin về dấu của hệ số ước lượng (ví dụ dấu dương), thì H_0: $\beta_2 = 0$; H_1: $\beta_2 > 0$ (kiểm định một đuôi).

Bước 2: Tính giá trị thống kê *t* tính toán (*t-stat*): $t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$, trong đó dưới giả thiết H_0: $\beta_2 = 0$, nên $t = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)}$. Giá trị này thường được báo cáo sẵn trong các kết quả ước lượng trên Eviews.

Bước 3: Tính giá trị thống kê *t* tra bảng (*t-crit*) theo công thức sau: =TINV(α,d.f.) trong excels.

Bước 4: Nếu $|t_{stat}| > |t_{crit}|$, ta bác bỏ giả thiết H_0.

Lưu ý, nếu ta muốn kiểm định một giả thiết nào khác (ví dụ, $\beta_2 = 1$), thì ta thay đổi giả thiết H_0 và H_1 ở bước 1, rồi tính một cách thủ công giá trị *t-stat* ở bước 2. Trong trường hợp này, chúng ta không thể sử dụng giá trị *t-stat* được báo cáo trong kết quả Eviews.

Trong thống kê, khi ta "bác bỏ" giả thiết không, nghĩa là ta nói rằng kết quả nghiên cứu của ta là **có ý nghĩa thống kê**. Ngược lại, khi ta 'không bác bỏ' giả thiết không, nghĩa là ta nói rằng kết quả nghiên cứu của ta là **không có ý nghĩa thống kê**. Thông thường, ta hay sử dụng ba mức ý nghĩa là 1%, 5%, và 10%. Tuy nhiên, sau này ta thấy rằng giá trị xác suất *p* (p-value hay prob của hệ số hồi quy) sẽ rất hữu ích vì chỉ cần nhìn vào giá trị xác suất *p*, ta có thể kết luận một hệ số ước lượng có ý nghĩa thống kê ở mức ý nghĩa là bao nhiêu. Giá trị xác xuất *p* sẽ được tính toán tự động khi chúng ta thực hiện hồi quy bằng phần mềm Eviews hay phần mềm khác.

Ý nghĩa của việc "chấp nhận" hay "bác bỏ" một giả thiết

Nếu trên cơ sở của một kiểm định ý nghĩa, ví dụ kiểm định *t*, ta quyết định "chấp nhận" giả thiết không (H_0), thì có nghĩa ta đang nói rằng với dữ liệu mẫu sẵn có ta chưa đủ cơ sở bác bỏ giả thiết đó, chứ ta không nói rằng giả thiết H_0 là đúng mà không có bất cứ hoài nghi nào. Tại sao? Để trả lời câu hỏi này, ta giả sử rằng H_0: $\beta_2 = -2.5$. Với hệ số ước lượng từ dữ liệu mẫu $\hat{\beta}_2 = -2.909$ và $se(\hat{\beta}_2) = 0.25$, thì giá trị *t* tính toán sẽ là $(-2.909 - (-2.5))/0.25 = 1.636$, ta kết luận hệ số ước

lượng không có ý nghĩa thống kê ở mức ý nghĩa $\alpha = 5\%$. Vì thế, ta "chấp nhận" H_0. Nhưng bây giờ giả sử ta giả định H_0: $\beta_2 = -3$, và tính được giá trị *t* tính toán là $(-2.909 - (-3))/0.25 = 0.364$. Với giá trị *t* tính toán này thì hệ số ước lượng vẫn không có ý nghĩa thống kê. Và bây giờ ta cũng "chấp nhận" H_0. Như vậy, trong hai giả thiết H_0 thì giả thiết nào thực sự là giả thiết "đúng"? Ta thực sự "không biết". Vì thế, khi "chấp nhận" một giả thiết H_0 ta luôn luôn nên hiểu rằng có một giả thiết khác có thể sẽ cũng tương thích với dữ liệu mẫu. Cho nên, tốt nhất là ta nên nói "có thể chấp nhận" giả thiết H_0, hơn là chỉ nói "chấp nhận" giả thiết H_0.

Giả thiết không "$\beta_i = 0$" và nguyên tắc t = 2

Một giả thiết H_0 được sử dụng phổ biến nhất trong các nghiên cứu thực nghiệm là H_0: $\beta_i = 0$; nghĩa là, hệ số độ dốc bằng không. Mục đích của loại giả thiết này là nhằm xem có mối quan hệ nào giữa biến phụ thuộc (Y) và một biến giải thích (X) nào đó hay không. Nếu kết quả cho thấy không có mối quan hệ nào giữa Y và X, thì việc kiểm định một giả thiết, ví dụ H_0: $\beta_i = -2$, là vô nghĩa.

Dependent Variable: Y
Method: Least Squares
Date: 09/20/08 Time: 23:59
Sample: 1 10
Included observati... 10

H_0: $\beta_i = 0$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	54.8000	1.5544	35.255	0.0000
X	-2.9091	0.2505	-11.613	0.0000

R-squared	0.9440	Mean dependent var	38.8000
Adjusted R-squared	0.9370	S.D. dependent var	9.0652
S.E. of regression	2.2754	Akaike info criterion	4.6590
Sum squared resid	41.4182	Schwarz criterion	4.7195
Log likelihood	-21.2951	F-statistic	134.8551
Durbin-Watson stat	1.0949	Prob(F-statistic)	0.0000

Giả thiết H_0 này có thể được kiểm định một cách dễ dàng bằng phương pháp khoảng tin cậy hay kiểm định mức ý nghĩa như đã trình bày ở trên. Nhưng thông thường người ta có thể kiểm định "nhanh" bằng cách áp dụng nguyên tắc "t=2" như sau:

■ BẢNG 7.2: **Nguyên tắc "t = 2".**

Nguyên tắc "t=2". Nếu số bậc tự do là 20 hoặc cao hơn và nếu mức ý nghĩa được chọn là $\alpha = 5\%$, thì giả thiết H_0: $\beta_i = 0$ có thể bị bác bỏ nếu giá trị tuyệt đối của giá trị t tính toán $(b_2/se(b_2))$ lớn hơn 2.

Nguồn: Gujarati, 2003, trang 134.

Tất cả các phần mềm kinh tế lượng đều có báo cáo giá trị *t* tính toán cho loại giả thiết này. Cho nên, ta chỉ cần so sánh giá trị *t* tính toán đó với giá trị *t* tra bảng ở một mức ý nghĩa xác định, hoặc đơn giản với t = 2.

Lưu ý rằng, chúng ta cần thiết phải kiểm định một hệ số hồi quy có ý nghĩa thống kê hay không vì đó là cơ sở quan trọng cho việc có thể sử dụng kết quả ước lượng cho các mục đích dự báo hệ số co giãn hoặc phân tích chính sách đối với các mô hình nhân quả. Ngoài ra, điều này cũng đúng đối với các mô hình dự báo bằng hồi quy hàm xu thế (ở chương 5).

ƯỚC LƯỢNG HỒI QUY ĐƠN TRÊN EVIEWS

Giả sử ta bắt đầu từ việc nhập dữ liệu vào Eviews rồi mới thực hiện ước lượng hồi.

Bước 1: Khởi động Eviews

Bước 2: Chọn File/New/Workfile để mở một tập tin Eviews mới

Bước 3: Chọn loại tần suất của dữ liệu. Trong trường hợp dữ liệu thời gian, chọn Dated-Regular Frequency, rồi chọn tần suất là Annual nếu dữ liệu theo năm, Quarterly nếu dữ liệu theo quý, Monthly nếu dữ liệu theo tháng, sau đó nhập thời điểm bắt

đầu (ví dụ 1990 nếu là năm, 2000Q1 nếu là quý, và 2000M1 nếu là tháng), và thời điểm kết thúc (ví dụ 2008 nếu là năm, 2008Q4 nếu là quý, và 2008M12 nếu là tháng). Trong trường hợp dữ liệu chéo (như ví dụ ta đang xét), chọn Unstructured/Undated, rồi nhập số quan sát của mẫu dữ liệu vào (ví dụ đang xét là 10). Sau khi chọn OK, ta sẽ có một cửa sổ mới với các thông tin mặc định bao gồm một hằng số (c) và một phần dư (resid).

Bước 4: Trong cửa sổ này ta chọn "genr" để tạo các biến Y và X như sau:

y=na (nhấn 'enter')

x=na (nhấn 'enter')

Như thế đã tạo được hai biến mới Y và X chưa có giá trị nào ở mỗi quan sát tương ứng (na = not available). Sau đó, ta chọn hai biến Y và X, rồi mở dưới dạng nhóm bằng cách nhấp đúp chuột vào hai biến đó.

Bước 5: Sau đó ta chọn Edit+/- để nhập dữ liệu vào hoặc có thể copy và paste từ bảng tính Excel. Sau khi đã nhập hoặc paste xong, ta lại chọn Edit+/- để kết thúc việc nhập dữ liệu từ bàn phím. Lưu ý, thông thường chúng ta chuyển trực tiếp một tập tin Excel (hoặc bất kỳ tập tin dạng nào khác) sang tập tin Eviews, chứ không cần thiết phải nhập một cách thủ công như vậy.

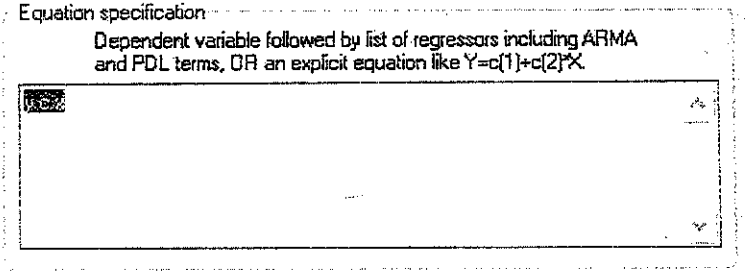
Bước 6: Sau khi đã nhập xong dữ liệu vào Eviews, ta có thể tiến hành ước lượng phương trình hồi quy bằng một trong hai cách sau đây:

Cách 1: Trên màn hình lệnh ta nhập vào như sau:

ls y c x (rồi nhấn 'enter')

Cách 2: Chọn Quick/Estimate Equation, rồi nhập vào hộp thoại 'equation specification' như sau:

y c x (nhấn 'enter')



Sau khi chọn "OK" chúng ta sẽ thấy xuất hiện một biểu tượng kết quả phương trình hồi quy như sau:

Tên biến phụ thuộc
Phương pháp ước lượng được sử dụng
Sai số chuẩn (se) của $\hat{\beta}_1$ và $\hat{\beta}_2$
Số quan sát
Hệ số $\hat{\beta}_1$
pr(|t|>35.56)
Dependent Variable: Y
Method: Least Squares
Date: 09/20/08 Time: 23:59
Sample: 1 10
Included observati... 10
Variable Coefficient Std. Error t-Statistic Prob.
pr(|t|>11.61)
Giá trị Y trung bình
Độ lệch chuẩn của Y
Giá trị thống kê F
Hằng số
Tên biến giải thích
Sai số chuẩn của ước lượng
$\hat{\sigma} = \sqrt{\frac{RSS}{n-2}}$
R^2
Hệ số $\hat{\beta}_2$
RSS
Giá trị thống kê của $\hat{\beta}_2$
Thống kê d
Durbin-Watson
$t_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = \frac{-2.9091}{0.2505}$
pr(|F|>134.85)
R-squared 0.9440 Mean dependent var 38.8000
Adjusted R-squared 0.9370 S.D. dependent var 9.0652
S.E. of regression 2.2754 Akaike info criterion 4.6590
Sum squared resid 41.4182 Schwarz criterion 4.7195
Log likelihood -21.2951 F-statistic 134.8551
Durbin-Watson stat 1.0949 Prob(F-statistic) 0.0000

MÔ HÌNH HỒI QUY BỘI

Thông thường trong các mối quan hệ kinh tế hay quản trị, biến phụ thuộc, Y, phụ thuộc vào nhiều biến giải thích khác nhau. Cho nên, chúng ta cần phải mở rộng phân tích hồi quy cho trường hợp tổng quát hơn. Hàm hồi quy tổng thể ngẫu nhiên với *k* biến có thể được biểu diễn như sau:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u_t \quad t = 1, 2, 3, \ldots, n \tag{7.53}$$

Trong đó, β_1 là hệ số cắt, β_2, …, β_k là các hệ số hồi quy riêng, u_t là hạng nhiễu ngẫu nhiên, và *t* là quan sát thứ *t*, *n* được xem là quy mô toàn bộ của tổng thể. Phương trình (7.53) cũng được chia thành hai thành phần (1) Thành phần xác định $E(Y_t/X_{2t}, X_{3t}, \ldots, X_{kt})$, nghĩa là giá trị trung bình có điều kiện của Y theo các giá trị cho trước của các X, và (2) Thành phần ngẫu nhiên u_t đại diện cho tất cả các yếu tố khác ngoài các biến $X_{2t}, \ldots, X_{kt}$ có ảnh hưởng lên Y_t.

ƯỚC LƯỢNG MÔ HÌNH HỒI QUY BỘI

Trong phạm vi cuốn sách này, chúng tôi chỉ trình bày minh họa trường hợp mô hình hồi quy ba biến. Cho nên, chúng ta có thể tham khảo trường hợp mô hình *k* biến ở các giáo trình chuyên về kinh tế lượng. Để ước lượng các hệ số hồi quy riêng ta vẫn sử dụng phương pháp tổng bình phương bé nhất thông thường (OLS) như đã giới thiệu trên. Giả sử ta có hàm hồi quy mẫu như sau:

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \hat{\beta}_3 X_{3t} + \hat{u}_t \tag{7.54}$$

Cũng theo phương pháp OLS, ta sẽ tìm các giá trị của $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$ sao cho tối thiểu hóa tổng bình phương phần dư (RSS). Ý tưởng này được thể hiện như sau:

$$RSS = \sum_{t=1}^{n} \hat{u}_t^2 = \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{n} (Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{2t} - \hat{\beta}_3 X_{3t})^2 \tag{7.55}$$

Để tối thiểu hóa (7.55), ta lấy đạo hàm bậc một của RSS theo $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$ và cho các đạo hàm này bằng không.

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2\sum(Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{2t} - \hat{\beta}_3 X_{3t}) = 0 \qquad (7.56)$$

$$\frac{\partial RSS}{\partial \hat{\beta}_2} = -2\sum(Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{2t} - \hat{\beta}_3 X_{3t}) X_{2t} = 0 \qquad (7.57)$$

$$\frac{\partial RSS}{\partial \hat{\beta}_3} = -2\sum(Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{2t} - \hat{\beta}_3 X_{3t}) X_{3t} = 0 \qquad (7.58)$$

Sắp xếp các phương trình (7.56), (7.57), và (7.58) ta có các phương trình tương đương như sau:

$$\Sigma Y_t = \Sigma \hat{\beta}_1 + \Sigma \hat{\beta}_2 X_{2t} + \Sigma \hat{\beta}_3 X_{3t} \qquad (7.59)$$

$$\Sigma Y_t X_{2t} = \hat{\beta}_1 \Sigma X_{2t} + \hat{\beta}_2 \Sigma X_{2t}^2 + \hat{\beta}_3 \Sigma X_{2t} X_{3t} \qquad (7.60)$$

$$\Sigma Y_t X_{3t} = \hat{\beta}_1 \Sigma X_{3t} + \hat{\beta}_2 \Sigma X_{2t} X_{3t} + \hat{\beta}_3 \sum X_{3t}^2 \qquad (7.61)$$

Có nhiều cách để có thể giải hệ gồm (7.59), (7.60) và (7.61) để tìm các nghiệm $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$. Thứ nhất, ta có thể giải ma trận 3 dòng 3 cột, như sau:

$$\begin{bmatrix} \sum Y_t \\ \sum Y_t X_{2t} \\ \sum Y_t X_{3t} \end{bmatrix} = \begin{bmatrix} n & \sum X_{2t} & \sum X_{2t} \\ \sum X_{2t} & \sum X_{2t}^2 & \sum X_{2t} X_{3t} \\ \sum X_{2t} & \sum X_{2t} X_{3t} & \sum X_{2t}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} \qquad (7.62)$$

Giải phương trình (7.62), ta có kết quả như sau:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3 \qquad (7.63)$$

$$\hat{\beta}_2 = \frac{(\sum y_t x_{2t})(\sum x_{3t}^2) - (\sum y_t x_{3t})(\sum x_{2t} x_{3t})}{(\sum x_{2t}^2)(\sum x_{3t}^2) - (\sum x_{2t} x_{3t})^2} \qquad (7.64)$$

$$\hat{\beta}_3 = \frac{(\sum y_t x_{3t})(\sum x_{2t}^2) - (\sum y_t x_{2t})(\sum x_{2t} x_{3t})}{(\sum x_{2t}^2)(\sum x_{3t}^2) - (\sum x_{2t} x_{3t})^2} \qquad (7.65)$$

GIẢI THÍCH CÁC HỆ SỐ HỒI QUY RIÊNG

Giả sử ta vẫn xét mô hình hồi quy ba biến như sau:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

Ở đây, β_2 đo lường ảnh hưởng của X_{2t} lên Y_t, với điều kiện giữ nguyên ảnh hưởng của X_3. Khái niệm này được áp dụng như thế nào khi chúng ta có các giá trị ước lượng OLS của β_2 (và β_3)? Để trả lời câu hỏi này, chúng ta thực hiện hai phương trình hồi quy đơn (và cũng có thể khái quát hóa cho mô hình *k* biến). Phương trình hồi quy thứ nhất điều chỉnh biến X_{2t} theo ý nghĩa "giữ nguyên X_{3t}"; và phương trình hồi quy thứ hai ước lượng ảnh hưởng của riêng biến được điều chỉnh này lên Y_t. Quy trình này được thực hiện theo hai bước sau đây:

Bước 1: Hồi quy X_{2t} theo X_{3t}. Sau khi ước lượng phương trình này, chúng ta tính các giá trị ước lượng của X_{2t} và phần dư $\hat{u}_t$. Để đơn giản, chúng ta sử dụng dữ liệu dưới dạng độ lệch ($x_t = X_t - \overline{X}_t$), và mô hình sẽ như sau:

$$x_{2t} = \hat{\alpha} x_{3t} + \hat{u}_t$$

Hoặc

$$x_{2t} = \hat{x}_{2t} + \hat{u}_t$$

Trong đó, $\hat{x}_{2t} = \hat{\alpha} x_{3t}$, $\hat{u}_t = x_{2t} - \hat{\alpha} x_{3t} = x_{2t} - \hat{x}_{2t}$ và $\hat{\alpha} = \dfrac{\sum x_{2t} x_{3t}}{\sum x_{3t}^2}$.

Mối quan tâm của chúng ta nằm ở $\hat{u}_t$, đại diện cho thành phần của X_{2t} không có liên quan gì đến X_{3t}. Cho nên, khái niệm "giữ nguyên X_{3t}" có nghĩa là chúng ta loại bỏ khỏi X_{2t} thành phần có liên quan đến X_{3t}.

Bước 2: Hồi quy y_t theo $\hat{u}_t$

$$y_t = \hat{\gamma} \hat{u}_t + v_t$$

$$\hat{\gamma} = \frac{\sum y_t \hat{u}_t}{\sum \hat{u}_t^2}$$

$\hat{\gamma}$ là ảnh hưởng của biến "X_{2t} điều chỉnh" lên Y_t, và đó chính là thước đo ảnh hưởng của riêng X_{2t} lên Y_t, khi X_{3t} được giữ nguyên[1]. Và $\hat{\gamma}$ sẽ đúng bằng $\hat{\beta}_2$. Chúng ta có thể làm tương tự cho X_{3t} và có thể mở rộng cho mô hình hồi quy k biến.

ĐẶC ĐIỂM CỦA CÁC ƯỚC LƯỢNG OLS

Dựa trên các giả định của CLRM, thì các hệ số hồi quy của mô hình hồi quy bội vẫn hội đủ các tính chất quan trọng như tuyến tính, không chệch, hiệu quả và nhất quán. Ngoài ra, các ước lượng OLS cũng theo phân phối chuẩn (không chứng minh), với giá trị trung bình và phương sai như sau:

- Giá trị trung bình của $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$

$$E(\hat{\beta}_1) = \beta_1 \quad (7.66)$$

$$E(\hat{\beta}_2) = \beta_2 \quad (7.67)$$

$$E(\hat{\beta}_3) = \beta_3 \quad (7.68)$$

- Phương sai của $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$

$$Var(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3t}^2 + \bar{X}_3^2 \sum x_{2t}^2 - 2\bar{X}_2\bar{X}_3 \sum x_{2t}x_{3t}}{(\sum x_{2t}^2)(\sum x_{3t}^2) - (\sum x_{2t}x_{3t})^2}\right].\sigma^2 \quad (7.69)$$

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2t}^2 (1 - r_{23}^2)} \quad (7.70)$$

[1] Xem chứng minh ở Pindyck & Rubinfeld, (1998), Econometric Models and Economic Forecasts, 4th Edition, McGraw-Hill.

$$\text{Var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3t}^2 (1 - r_{23}^2)} \tag{7.71}$$

Như vậy, phương sai của các hệ số hồi quy $\hat{\beta}_2$ và $\hat{\beta}_3$ không chỉ phụ thuộc vào phương sai hạng nhiễu và cỡ mẫu, mà còn phụ thuộc vào mối tương quan giữa các biến giải thích trong mô hình. Chỉ khi nào X_{2t} và X_{3t} hoàn toàn độc lập, nghĩa là hệ số tương quan $r_{23} = 0$, thì công thức phương sai của các hệ số $\hat{\beta}_2$, và $\hat{\beta}_3$ sẽ giống với công thức phương sai của hệ số hồi quy trong mô hình hồi quy đơn. Đây là một vấn đề quan trọng trong phân tích hồi quy, và sẽ được đề cập lại ở phần phân tích chẩn đoán.

Lấy căn bậc hai của các công thức (8.59), (8.61), và (8.71), ta sẽ có các sai số chuẩn của các hệ số $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$ như sau:

$$\text{se}(\hat{\beta}_1) = \sqrt{\text{Var}(\hat{\beta}_1)} \tag{7.72}$$

$$\text{se}(\hat{\beta}_2) = \sqrt{\text{Var}(\hat{\beta}_2)} \tag{7.73}$$

$$\text{se}(\hat{\beta}_3) = \sqrt{\text{Var}(\hat{\beta}_3)} \tag{7.74}$$

Sai số chuẩn của $\hat{\beta}_1$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.652419	**0.606198**	-2.725873	0.0144
LOG(K)	0.845997	**0.093352**	9.062488	0.0000
LOG(L)	0.339732	**0.185692**	1.829548	0.0849

Sai số chuẩn của $\hat{\beta}_3$ *Sai số chuẩn của* $\hat{\beta}_2$

Tương tự hồi quy đơn, phương sai hạn nhiễu (σ^2) được ước lượng thông qua công thức sau đây:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_t^2}{n-3} \tag{7.75}$$

Và ta cũng có:

$$E(\hat{\sigma}^2) = \frac{1}{n-3} E\left(\sum \hat{u}_t^2\right) = \sigma^2 \tag{7.76}$$

Vậy rõ ràng, tương tự hồi quy đơn, $\hat{\sigma}^2$ cũng là một ước lượng không chệch của phương sai nhiễu σ^2. Ở công thức (7.75), (n-3) là số bậc tự do, ký hiệu là d.f., và $\sum \hat{u}_t^2$ là tổng bình phương phần dư, ký hiệu là RSS. Số bậc tự do của RSS ở đây sẽ là (n-3), hay bằng số quan sát trong mẫu trừ số hệ số ước lượng trong mô hình hồi quy. Nhắc lại rằng, để có thể tính được RSS, trước tiên ta phải có các hệ số $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$ vì các giá trị của Y_t, X_{2t} và X_{3t} đã có sẵn từ dữ liệu mẫu. Để ước lượng được $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$ ta cần ít nhất ba cặp quan sát (Y_t, X_{2t}, X_{3t}) bất kỳ (nghĩa là xác định phương trình mặt thẳng qua ba điểm). Như vậy, ba giá trị ước lượng này là ba ràng buộc lên RSS. Nói cách khác, trong tập hợp tất cả các cặp quan sát (Y_t, X_{2t}, X_{3t}) trong miền giá trị của mẫu dữ liệu sẽ có ít nhất ba cặp quan sát nào đó nằm trên (hoặc rất gần với) phương trình hồi quy mẫu. Chính vì thế, phần dư tương ứng sẽ bằng không hoặc rất nhỏ. Như vậy, thực sự giá trị của RSS chỉ do (n-3) giá trị $\hat{u}_t^2$ tạo thành. Như vậy, (n-3) chính là số nguồn thông tin của RSS.

Lấy căn bậc hai của công thức (7.75) ta sẽ có sai số chuẩn của giá trị ước lượng hay sai số chuẩn của hồi quy ($\hat{\sigma}$) như sau:

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_t^2}{n-3}} \tag{7.77}$$

Đây chính là độ lệch chuẩn của các giá trị Y quanh đường hồi quy mẫu và được sử dụng như một thước đo "mức độ phù hợp" của đường hồi quy so với các giá trị thực từ mẫu dữ liệu. Thước đo này chỉ có ý nghĩa khi so sánh giữa các mô hình có cùng dạng biến phụ thuộc.

Từ hai công thức (7.54) và (7.63), ta có thể viết lại $\hat{u}_t$ dưới dạng độ lệch như sau:

$$\hat{u}_t = y_t - \hat{\beta}_2 x_{2t} - \hat{\beta}_3 x_{3t} \tag{7.78}$$

Như vậy,

$$\begin{aligned}\sum \hat{u}_t^2 &= \sum \hat{u}_t \hat{u}_t \\ &= \sum \hat{u}_t (y_t - \hat{\beta}_2 x_{2t} - \hat{\beta}_3 x_{3t}) \\ &= \sum \hat{u}_t y_t \end{aligned} \tag{7.79}$$

Như vậy, phương trình (7.79) có thể được viết lại như sau:

$$\begin{aligned}\sum \hat{u}_t^2 &= \sum (y_t - \hat{\beta}_2 x_{2t} - \hat{\beta}_3 x_{3t}) y_t \\ &= \sum y_t^2 - \hat{\beta}_2 \sum y_t x_{2t} - \hat{\beta}_3 \sum y_t x_{3t} \end{aligned} \tag{7.80}$$

Đặc điểm của các phương sai và sai số chuẩn của các hệ số ước lượng

(1) Phương sai của $\hat{\beta}_2$ tỷ lệ thuận với phương sai số hạng nhiễu σ^2 và hệ số tương quan giữa X_{2t} và X_{3t} nhưng tỷ lệ nghịch với $\sum x_{2t}^2$. Điều này có nghĩa là, với giá trị σ^2 không đổi, các giá trị X_t càng biến thiên quanh giá trị trung bình, thì phương sai của $\hat{\beta}_2$ càng nhỏ và vì thế độ chính xác trong việc ước lượng giá trị thực của β_2 càng cao. Ngược lại, với giá trị $\sum x_{2t}^2$ không đổi, phương sai nhiễu σ^2 càng lớn, hoặc hệ số tương quan giữa các biến giải thích trong mô hình càng cao thì phương sai $\hat{\beta}_2$ càng lớn. Lưu ý rằng, khi cỡ mẫu tăng, số số hạng trong $\sum x_{2t}^2$ sẽ tăng, nên $\sum x_{2t}^2$ sẽ tăng. Như vậy, khi số quan sát tăng, thì độ chính xác trong việc ước lượng giá trị thực của β_2 càng cao.

(2) Phương sai của $\hat{\beta}_3$ tỷ lệ thuận với phương sai nhiễu σ^2 và hệ số tương quan giữa X_{2t} và X_{3t} nhưng tỷ lệ nghịch với $\sum x_{3t}^2$. Điều này có nghĩa là, với giá trị σ^2 không đổi, các giá trị X_t càng

biến thiên quanh giá trị trung bình, thì phương sai của b_3 càng nhỏ và vì thế độ chính xác trong việc ước lượng giá trị thực của β_3 càng cao. Ngược lại, với giá trị $\sum x_{3t}^2$ không đổi, phương sai nhiễu σ^2 càng lớn, hoặc hệ số tương quan giữa các biến giải thích trong mô hình càng cao thì phương sai $\hat{\beta}_3$ càng lớn. Lưu ý rằng, khi cỡ mẫu tăng, số số hạng trong $\sum x_{3t}^2$ sẽ tăng, nên $\sum x_{3t}^2$ sẽ tăng. Như vậy, khi cỡ mẫu tăng, thì độ chính xác trong việc ước lượng giá trị thực của β_2 càng cao.

(3) Phương sai của $\hat{\beta}_1$ tỷ lệ thuận với phương sai nhiễu σ^2 và hệ số tương quan giữa X_{2t} và X_{3t}, nhưng tỷ lệ nghịch với $\sum x_{2t}^2$, $\sum x_{3t}^2$ và cỡ mẫu.

Như vậy, khi đã có các sai số chuẩn của các ước lượng OLS, $se(\hat{\beta}_1)$, $se(\hat{\beta}_2)$ và $se(\hat{\beta}_3)$, ta có thể dễ dàng tính được các ước lượng khoảng của các ước lượng OLS.

HỆ SỐ XÁC ĐỊNH R^2 MÔ HÌNH HỒI QUY BỘI

Ta biết rằng, trong mô hình hồi quy đơn, r^2 là thước đo mức độ phù hợp của hàm hồi quy; nghĩa là, nó cho biết tỷ lệ hay phần trăm tổng biến thiên của biến phụ thuộc Y được giải thích bởi biến giải thích X. Tương tự, trong mô hình hồi quy bội, ta cũng muốn biết tỷ lệ phần trăm biến thiên trong Y được giải thích đồng thời bởi các biến giải thích, ví dụ, X_2 và X_3. Đại lượng cung cấp thông tin này được gọi là hệ số xác định đa biến và được ký hiệu bằng R^2. Ta có,

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \hat{\beta}_3 X_{3t} + \hat{u}_t$$

$$= \hat{Y}_t + \hat{u}_t \qquad (7.81)$$

Trong đó, $\hat{Y}_t$ là giá trị được ước lượng của Y_t từ đường hồi quy mẫu và là một ước lượng của giá trị thực $E(Y_t/X_{2t},X_{3t})$. Phương trình (7.81) có thể được viết lại dưới dạng độ lệch so với các giá trị trung bình như sau:

$$Y_t = \hat{\beta}_2 x_{2t} + \hat{\beta}_3 x_{3t} + \hat{u}_t$$
$$= \hat{y}_t + \hat{u}_t \quad (7.82)$$

Lấy bình phương hai vế của (7.82) và rồi tổng tất các giá trị mẫu lại, ta sẽ có được phương trình sau đây:

$$\sum y_t^2 = \sum \hat{y}_t^2 + \sum \hat{u}_t^2 + 2\sum \hat{y}_t \hat{u}_t$$
$$\sum y_t^2 = \sum \hat{y}_t^2 + \sum \hat{u}_t^2 \quad (7.83)$$

Phương trình (7.83) cho rằng tổng bình phương (TSS) bằng tổng bình phương phần được giải thích (ESS) cộng tổng bình phương phần dư (RSS). Bây giờ, ta thế phương trình (7.80) vào (7.83), ta có:

$$\sum y_t^2 = \sum \hat{y}_t^2 + \sum y_t^2 - \hat{\beta}_2 \sum y_t x_{2t} - \hat{\beta}_3 \sum y_t x_{3t} \quad (7.84)$$

Sắp xếp lại phương trình (7.84), ta có

$$ESS = \sum \hat{y}_t^2 = \hat{\beta}_2 \sum y_t x_{2t} + \hat{\beta}_3 \sum y_t x_{3t} \quad (7.85)$$

Từ định nghĩa hệ số xác định ở trên, ta có

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}_2 \sum y_t x_{2t} + \hat{\beta}_3 \sum y_t x_{3t}}{\sum y_t^2} \quad (7.86)$$

Tương tự r^2, hệ số xác định R^2 là một đại lượng nằm trong khoảng từ 0 đến 1. Nếu $R^2 = 1$, đường hồi quy mẫu giải thích 100% của biến thiên trong Y. Ngược lại, nếu $R^2 = 0$, thì mô hình không giải thích được gì cho biến thiên trong Y. Thông thường, R^2 nằm giữa hai giá trị này. R^2 càng gần 1 thì mô hình được cho là có độ phù hợp (với dữ liệu mẫu) càng cao, vì thế mô hình được cho là tốt hơn.

Gujarati (2003) cho rằng trong mô hình hồi quy bội (k biến) thì mối quan hệ giữa R^2 và phương sai của một hệ số hồi quy riêng bất kỳ sẽ được thể hiện như sau:

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1-R_j^2}\right) \quad (7.87)$$

Trong đó, $\hat{\beta}_j$ là hệ số hồi quy riêng của X_j và R^2_j là R^2 trong phương trình hồi quy của X_j theo (k-2) biến giải thích còn lại. Phương trình này rất có ý nghĩa khi ta phân tích vấn đề hiện tượng đa cộng tuyến.

R^2 VÀ R^2 ĐIỀU CHỈNH

Như đã trình bày ở trên, hệ số xác định R^2 vẫn là một thước đo mức độ phù hợp trong mô hình hồi quy bội. Tuy nhiên, R^2 không thể được sử dụng như một phương tiện để so sánh hai phương trình hồi quy khác nhau có số biến giải thích khác nhau. Điều này bởi vì khi các biến giải thích mới được đưa thêm vào mô hình, thì tỷ lệ biến thiên trong Y được giải thích bởi các biến giải thích X, tức R^2, sẽ luôn luôn tăng. Chính vì thế, chúng ta sẽ luôn luôn có một R^2 cao hơn bất kể biến giải thích được đưa thêm vào mô hình có quan trọng hay không. Gujarati (2003) cho rằng R^2 là một hàm không giảm của số biến giải thích trong mô hình. Điều này rất dễ nhận ra trong công thức sau đây:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_t^2}{\sum y_t^2} \qquad (7.88)$$

Ta biết rằng, $\sum y_t^2$ không phụ thuộc vào số biến giải thích trong mô hình bởi vì nó đơn giản chỉ là $\sum (Y_t - \overline{Y})^2$. Tuy nhiên, RSS, $\sum \hat{u}_t^2$ lại phụ thuộc vào số biến giải thích hiện có trong mô hình. Chỉ bằng trực giác ta cũng có thể nhận thấy rằng khi số biến X tăng lên, $\sum \hat{u}_t^2$ có thể sẽ giảm (hoặc ít nhất là không tăng), vì thế R^2 sẽ tăng. Cho nên, nếu so sánh hai mô hình có cùng biến phụ thuộc nhưng khác số biến giải thích, chúng ta có thể rất dễ bị nhầm lẫn vì sẽ chọn mô hình có R^2 cao hơn. Vì lẽ nsày, chúng ta cần một thước đo khác có tính đến số biến giải thích trong mỗi mô hình. Thước đo đó được gọi là R^2 điều chỉnh (*adjusted* R^2), thường được ký hiệu là $\overline{R}^2$ bởi vì nó đã điều chỉnh số

biến giải thích (hay nói đúng hơn là điều chỉnh số bậc tự do) trong mô hình.

$$\overline{R}^2 = 1 - \frac{\sum \hat{u}_i^2 /(n-k)}{\sum y_i^2 /(n-1)} \qquad (7.89)$$

Trong đó, k = số hệ số ước lượng trong mô hình (kể cả hệ số cắt $\hat{\beta}_1$). Trong mô hình hồi quy 3 biến, k = 3; mô hình hồi quy 4 biến, k = 4; ... Từ công thức (7.89) ta thấy rằng $\overline{R}^2$ đã điều chỉnh số bậc tự do tương ứng từng tổng bình phương trong công thức tính R^2. Như vậy, khi số biến giải thích tăng, *k* sẽ tăng (*n-k* sẽ giảm) và RSS cũng giảm. Khi đó, tử số của (7.89) đã được bù trừ, và chính vì thế $\overline{R}^2$ là một thước đo tương đối 'công bằng' hơn trong việc so sánh giữa các mô hình có số biến giải thích khác nhau. Công thức (7.89) cũng có thể được viết lại như sau:

$$\overline{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_Y^2} \qquad (7.90)$$

Trong đó, $\hat{\sigma}^2$ là phương sai của phần dư, một ước lượng không chệch của phương sai nhiễu, σ^2, và S_Y^2 là phương sai mẫu của Y.

Thế công thức (7.88) vào (7.89), ta dễ dàng nhận thấy mối quan hệ giữa R^2 và $\overline{R}^2$ sẽ như sau:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \qquad (7.91)$$

Như vậy, khi k = 1, $R^2 = \overline{R}^2$, khi k > 1, $R^2 > \overline{R}^2$, nghĩa là khi số biến giải thích tăng, $\overline{R}^2$ tăng ít hơn R^2. Ngoài ra, $\overline{R}^2$ có thể là một đại lượng âm (khi $R^2 = 0$ và k > 1), mặc dù R^2 là một đại lượng không âm.

R^2 ĐIỀU CHỈNH LOẠI TRỪ VIỆC GIA TĂNG CHỦ QUAN BIẾN ĐỘC LẬP

Gujarati (2003) cho rằng đôi khi nhiều người nghiên cứu chơi trò tối đa hóa R^2 điều chỉnh; nghĩa là, chọn mô hình có R^2 điều chỉnh cao

nhất. Tuy nhiên, trò chơi này có thể rất nguy hiểm, vì phân tích hồi quy không nhằm mục tiêu có được một giá R^2 điều chỉnh cao, mà mục đích chính là tìm ra được các giá trị ước lượng của các hệ số hồi quy thực của tổng thể và rút ra các suy luận thống kê về các giá trị thực này. Nhiều nghiên cứu thực tiễn có R^2 điều chỉnh rất cao nhưng có một số hệ số hồi quy không có ý nghĩa thống kê hoặc thậm chí có dấu trái với kỳ vọng. Chính vì vậy, chúng ta nên chú ý hơn đến sự phù hợp về mặt lý thuyết của các biến giải thích đối với biến phụ thuộc trong mô hình và mức ý nghĩa thống kê của các hệ số hồi quy. Ngoài ra, một mô hình tốt hay không còn phụ thuộc vào việc nó có thỏa mãn các giả định của mô hình hồi quy tuyến tính cổ điển hay không. Và các nội dung này sẽ được trình bày ở phần sau của chương này. Cũng theo Gujarati (2003), nếu chúng ta có cơ sở lý thuyết tốt, mô hình đã được xác định đúng, và có phân tích chẩn đoán cẩn thận, thì việc có được một giá trị R^2 điều chỉnh cao là một mô hình đáng mong muốn. Trái lại, nếu chúng ta có cơ sở lý thuyết tốt, mô hình đã được xác định đúng, và có phân tích chẩn đoán cẩn thận, thì việc có được một giá trị R^2 điều chỉnh thấp không có nghĩa đó là một mô hình tồi. Lưu ý rằng, khi chúng ta ước lượng mô hình với dữ liệu chéo, ví dụ sử dụng số liệu điều tra riêng hoặc VHLSS, thì giá trị R^2 điều chỉnh có thể tương đối thấp (trong khoảng 0.2 đến 0.55). Cho nên, người làm dự báo hãy yên tâm với kết quả nghiên cứu của mình, đừng vì một R^2 điều chỉnh thấp mà cố gắng biến hóa mô hình để thuyết phục người khác.

CÁC TIÊU CHÍ LỰA CHỌN MÔ HÌNH

Bên cạnh R^2 và $\overline{R}^2$, một số tiêu chí khác cũng thường được sử dụng để đánh giá mức độ phù hợp của một mô hình hồi quy như AIC, FPE, SBC, và HQC (có sẵn trong kết quả hồi quy trên Eviews).

Nhắc lại rằng, khi tăng số biến giải thích trong một mô hình hồi quy bội sẽ làm giảm RSS, và vì thế R^2 sẽ tăng. Tuy nhiên, cái giá của việc tăng R^2 là giảm số bậc tự do trong mô hình. Một phương pháp khác – ngoài $\overline{R}^2$, cho phép số biến giải thích thay đổi khi đánh giá mức độ phù hợp là sử dụng các tiêu chí khác cho việc so sánh giữa các mô hình, chẳng hạn như Akaike Information Criterion (AIC) của Akaike (1974):

$$AIC = \left(\frac{RSS}{n}\right)\hat{u}^{2k/n} \tag{7.92}$$

Các phần mềm kinh tế lượng thường sử dụng công thức biến đổi của công thức (7.92) như sau:

$$\ln(AIC) = \ln\left(\frac{RSS}{n}\right) + \frac{2k}{n} \tag{7.93}$$

Tiêu chí Schwarz Bayesian Criterion (SBC) của Schwarz (1978):

$$SBC = \left(\frac{RSS}{n}\right)\hat{u}^{k/n} \tag{7.94}$$

Các phần mềm kinh tế lượng thường sử dụng công thức biến đổi của công thức (7.94) như sau:

$$\ln(SBC) = \ln\left(\frac{RSS}{n}\right) + \frac{k}{n} \tag{7.95}$$

Tiêu chí Finite Prediction Error (FPE) của Akaike (1970):

$$FPE = \left(\frac{RSS}{n}\right)\frac{n+k}{n-k} \tag{7.96}$$

Và tiêu chí Hannan and Quin Criterion (HQC) của Quin (1979):

$$HQC = \left(\frac{RSS}{n}\right)(\ln(n))^{2k/n} \tag{7.97}$$

Asteriou (2007) cho rằng chúng ta nên chọn mô hình với các tiêu chí trên sao cho chúng có giá trị nhỏ nhất. Nói chung, thường thì các tiêu chí này có thể cho các kết quả trái ngược nhau, dẫn đến có thể có các kết luận khác nhau. Tuy nhiên, nguyên tắc chung là nên chọn mô hình nào có nhiều tiêu chí có giá trị nhỏ hơn so với các mô hình khác. AIC và SBC là hai tiêu chí được sử dụng phổ biến nhất trong phân tích chuỗi thời gian như mô hình ARIMA, ARCH, GARCH, VAR, hay ECM. Lưu ý rằng, dù sử dụng tiêu chí nào thì các mô hình đang xem xét phải có cùng biến phụ thuộc và có cùng dạng hàm.

ƯỚC LƯỢNG HỒI QUY BỘI TRÊN EVIEWS

Bước 1: Khởi động Eviews

Bước 2: Chọn File/New/Workfile để mở một tập tin Eviews mới

Bước 3: Chọn loại tần suất của dữ liệu. Trong trường hợp dữ liệu thời gian, chọn Dated-Regular Frequency, rồi chọn tần suất là Annual nếu dữ liệu theo năm, Quarterly nếu dữ liệu theo quý, Monthly nếu dữ liệu theo tháng, sau đó nhập thời điểm bắt đầu (ví dụ 1990 nếu là năm, 2000Q1 nếu là quý, và 2000M1 nếu là tháng), và thời điểm kết thúc (ví dụ 2008 nếu là năm, 2008Q4 nếu là quý, và 2008M12 nếu là tháng). Trong trường hợp dữ liệu chéo (như ví dụ ta đang xét), chọn Unstructured/Undated, rồi nhập số quan sát của mẫu dữ liệu vào. Sau khi chọn OK, ta sẽ có một cửa sổ mới với các thông tin mặc định bao gồm một hằng số (c) và một phần dư (resid).

Bước 4: Trong cửa sổ này ta chọn "genr" để tạo các biến Y, X_2, và X_3 như sau:

y=na (nhấn 'enter')

x2=na (nhấn 'enter')

x3=na (nhấn 'enter')

Như thế đã tạo được ba biến mới Y, X2 và X3 chưa có giá trị nào ở mỗi quan sát tương ứng (na = not available). Sau đó, ta chọn ba biến Y, X2 và X3, rồi mở dạng nhóm bằng cách nhấp đúp chuột vào ba biến đó. Lưu ý, chúng ta có thể đặt tên biến theo chữ tắt trong tiếng Anh và có chú thích tên nhãn.

Bước 5: Sau đó ta chọn Edit+/- để nhập dữ liệu vào hoặc có thể copy và paste từ bảng tính Excel. Sau khi đã nhập hoặc paste xong, ta lại chọn Edit+/- để kết thúc việc nhập dữ liệu từ bàn phím.

Bước 6: Sau khi đã nhập xong dữ liệu vào Eviews, ta có thể tiến hành ước lượng phương trình hồi quy bằng một trong hai cách sau đây:

Cách 1: Trên màn hình lệnh ta nhập vào như sau:

ls y c x2 x3 (rồi nhấn 'enter')

Cách 2: Chọn Quick/Estimate Equation, rồi nhập vào hộp thoại 'equation specification' như sau:

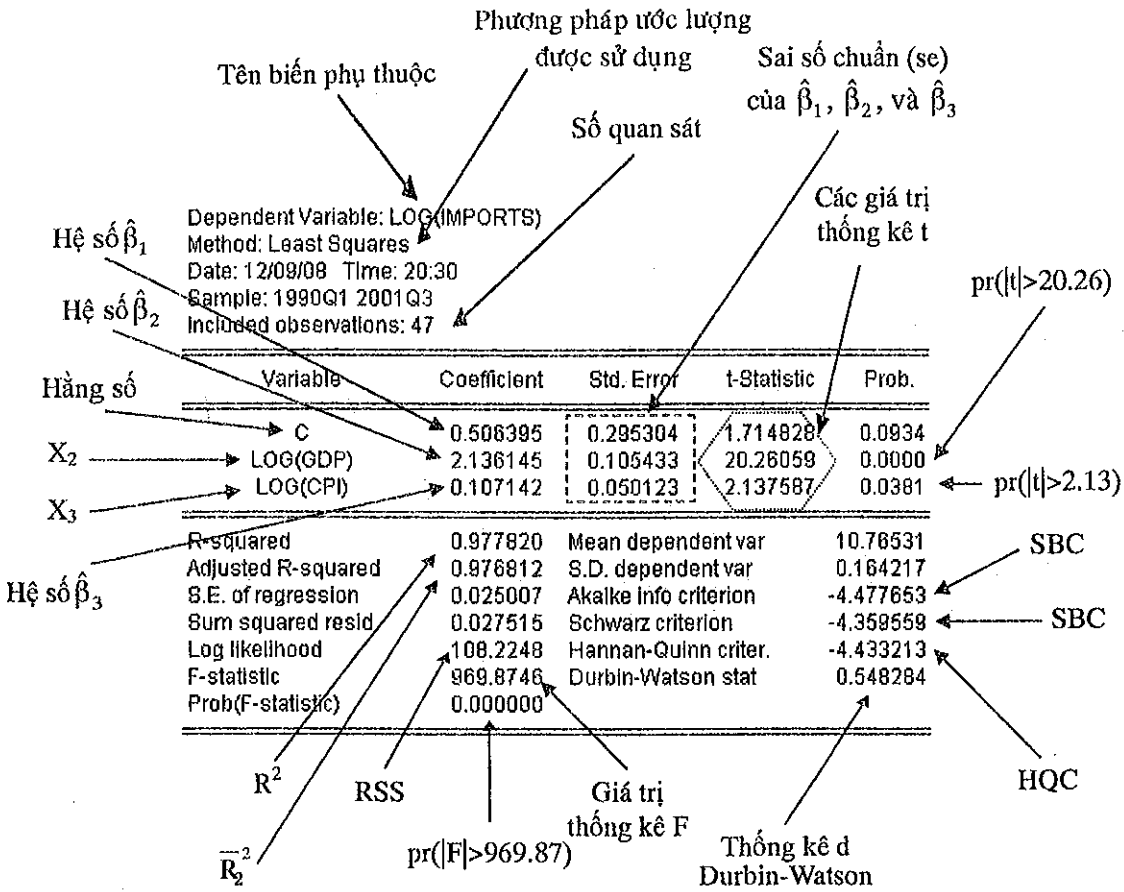
y c x2 x3 (nhấn 'enter')

Lưu ý, Eviews không phân biệt chữ thường với chữ hoa. Eviews sẽ mặc định chọn phương pháp ước lượng là ls (least squares), và số mẫu dùng để ước lượng sẽ là số quan sát tối đa hiện có trong mẫu dữ liệu.

Ví dụ, mở tập tin "**DATA7-1**", trong đó, IMPORTS, GDP, và CPI lần lượt là các biến giá trị nhập khẩu (triệu đôla), tổng sản phẩm nội địa (triệu đôla), và chỉ số giá tiêu dùng (%) từ quý I năm 1990 đến quý III năm 2001. Chọn Quick/Estimate Equation, rồi nhập vào hộp thoại 'equation specification' như sau:

log(imports) c log(gdp) log(cpi)

Sau khi chọn "OK" chúng ta sẽ thấy xuất hiện một biểu tượng kết quả phương trình hồi quy như sau:


Phương pháp ước lượng được sử dụng
Tên biến phụ thuộc
Sai số chuẩn (se) của $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$
Số quan sát
Các giá trị thống kê t
pr(|t|>20.26)
Hệ số $\hat{\beta}_1$
Hệ số $\hat{\beta}_2$
Hằng số
X_2
X_3
Hệ số $\hat{\beta}_3$
pr(|t|>2.13)
SBC
SBC
HQC
R^2
RSS
Giá trị thống kê F
$\bar{R}_2^2$
pr(|F|>969.87)
Thống kê d Durbin-Watson
Dependent Variable: LOG(IMPORTS)
Method: Least Squares
Date: 12/09/08 Time: 20:30
Sample: 1990Q1 2001Q3
Included observations: 47
Variable Coefficient Std. Error t-Statistic Prob.
C 0.506395 0.295304 1.714828 0.0934
LOG(GDP) 2.136145 0.105433 20.26059 0.0000
LOG(CPI) 0.107142 0.050123 2.137587 0.0381
R-squared 0.977820 Mean dependent var 10.76531
Adjusted R-squared 0.976812 S.D. dependent var 0.164217
S.E. of regression 0.025007 Akaike info criterion -4.477653
Sum squared resid 0.027515 Schwarz criterion -4.359559
Log likelihood 108.2248 Hannan-Quinn criter. -4.433213
F-statistic 969.8746 Durbin-Watson stat 0.548284
Prob(F-statistic) 0.000000

Thông thường, chúng ta sử dụng dữ liệu đã có sẵn hoặc chuyển dữ liệu từ các tập tin Excel, Stata, hay SPSS, v.v..., thay vì phải mất nhiều thời gian nhập dữ liệu như vừa hướng dẫn ở trên, có nghĩa là nếu đã có dữ liệu chứa sẵn trong Eviews thì chúng ta chỉ nên bắt đầu hồi quy từ bước 6.

KIỂM ĐỊNH GIẢ THIẾT

Kiểm định giả thiết về các hệ số hồi quy riêng

Cũng tương tự mô hình hồi quy đơn, với các giả định cho rằng hạng nhiễu $u_t \sim N(0, \sigma^2)$, thì chúng ta có thể sử dụng thống kê *t* để,kiểm định một giả thiết về bất kỳ một hệ số hồi quy riêng nào. Để minh họa cách thức thực hiện kiểm định, chúng ta hãy xem lại ví dụ về nhập khẩu như đã được minh họa ở bước 6, phần "Ước lượng mô hình hồi quy bội trên Eviews". Giả sử, chúng ta có giả thiết như sau:

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

Giả thiết không này cho rằng, với X_3 (logarith của chỉ số giá tiêu dùng) được giữ nguyên, thì X_2 (logarith của tổng sản phẩm quốc nội) không có ảnh hưởng (tuyến tính) lên Y (logarith của kim ngạch nhập khẩu). Để kiểm định giả thiết này, chúng ta sử dụng thống kê *t* như đã trình bày ở phần hồi quy đơn. Nguyên tắc quyết định chung sẽ như sau: nếu giá trị *t* tính toán lớn hơn giá trị *t* tra bảng ở mức ý nghĩa được chọn, thì chúng ta có thể bác bỏ giả thiết H_0. Ở ví dụ này, dưới giả thiết H_0: $\beta_2 = 0$, ta có:

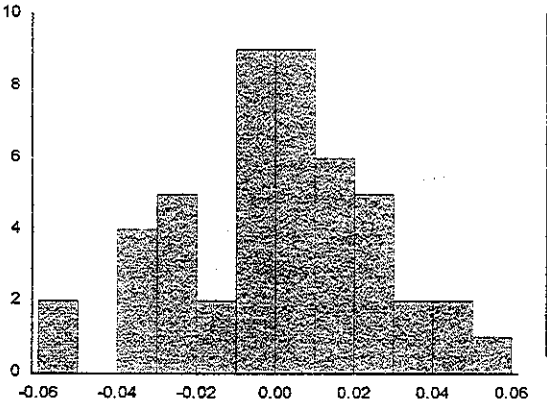
$$t = \frac{2.136 - 0}{0.1054} = 20.26$$

Chúng ta có thể dễ dàng nhận biết được giá trị *t* tính toán dưới giả thiết H_0: $\beta_k = 0$ ở cột *t-Statistic* trên bảng kết quả hồi quy Eviews. Với số quan sát n = 47, nên số bậc tự do sẽ là 44. Giả sử ta chọn mức ý nghĩa $\alpha = 5\%$, thì giá trị *t* tra bảng là 2 đối với kiểm định hai phía (=TINV(5%,44)) hoặc là 1.68 đối với kiểm định một phía

(=TINV(10%,44)). Trong ví dụ đang xét, giả thiết H_1 là loại giả thiết hai phía, nên chúng ta sử dụng giá trị *t* hai phía. Do giá trị *t* tính toán là 20.26 lớn hơn giá trị *t* tra bảng là 2, nên chúng ta có thể bác bỏ giả thiết H_0 cho rằng log(GDP) không có ảnh hưởng gì lên log(IMPORTS). Tương tự như vậy, chúng ta cũng bác bỏ giả thiết H_0 cho rằng log(CPI) không có ảnh hưởng gì lên log(IMPORTS) do giá trị *t* tính toán là 2.138.

Trên thực tế, chúng ta không cần phải giả định một mức ý nghĩa α cụ thể để thực hiện kiểm định giả thiết. Thông thường, chúng ta sử dụng giá trị xác suất *p*, ví dụ là 0.0381 đối với biến log(CPI). Giải thích giá trị xác suất p này như sau: Nếu giả thiết H_0 là đúng, thì xác suất để có được giá trị *t* bằng hoặc lớn hơn 2.138 là 0.0381 hay 0.381%, và đây là một xác suất tương đối nhỏ. Nói cách khác, xác suất để hệ số hồi quy của log(CPI) bằng 0 chỉ là 0.381% (hay là 0.0381) nhỏ hơn nhiều so với 5% hay 0.05.

Nên nhớ rằng, thủ tục kiểm định dựa vào giả thiết cho rằng hạng nhiễu u_t theo phân phối chuẩn. Mặc dù chúng ta không quan sát được u_t, nhưng chúng ta có thể quan sát đại diện của nó là $\hat{u}_t$, tức phần dư của phương trình hồi quy. Từ kết quả hồi quy mô hình về IMPORTS, ta có đồ thị phần dư như ở Hình 7.2. Đồ thị này cho thấy phần dư từ mô hình hồi quy có phân phối chuẩn. Chúng ta cũng tính được giá trị thống kê Jarque-Bera (JB) cho việc kiểm định tính chuẩn. Trong ví dụ này, giá trị JB là 0.295 với xác suất *p* là 0.863. Như vậy, hạng nhiễu trong mô hình của chúng ta có phân phối chuẩn. Dĩ nhiên, lưu ý rằng, kiểm định JB là loại kiểm định cho cỡ mẫu lớn và ví dụ của chúng ta với 47 quan sát có thể chưa phải là một mẫu lớn. Ngoài ra, ta có thể nhận thấy rằng các giá trị skewness và kurtosis là -0.18 và 2.86, gần bằng giá trị phân phối chuẩn là 0 và 3.



■ HÌNH 7.2: **Đồ thị phần dư của mô hình log(IMPORTS)**

Kiểm định ràng buộc tuyến tính

Trong phân tích và dự báo kinh tế, chúng ta thường hay kiểm định các giả thiết về các mối quan hệ nhất định giữa các hệ số hồi quy. Chẳng hạn, xét ví dụ về hàm sản xuất Cobb-Douglas có dạng như sau:

$$Q = AL^{\beta_1}K^{\beta_2} \qquad (7.98)$$

Trong đó, Q là sản lượng, L là lao động, K là vốn, và A là một tham số ngoại sinh đại diện cho yếu tố công nghệ, kỹ năng quản trị, và các yếu tố khác ngoài K và L. Nếu lấy logarithms hai vế của phương trình (7.98) và đưa thêm một hạng nhiễu ngẫu nhiên, ta có:

$$\ln Q = \beta_1 + \beta_2 \ln L + \beta_3 \ln K + u \qquad (7.99)$$

Trong đó, $\beta_1 = \ln A$, là một hằng số, β_2 và β_3 lần lượt là các hệ số co giãn của sản lượng theo lao động và vốn. Trong các nghiên cứu có sử dụng hàm sản xuất như thế này, chúng ta thường quan tâm đến kiểm định giả thiết H_0: $\beta_2 + \beta_3 = 1$, nghĩa là, tính kinh tế không đổi theo quy mô (tập tin **DATA7-2**). Với giả thiết này, thì phương trình (7.99) sẽ được viết lại như sau:

$$\ln Q = \beta_1 + (1 - \beta_3)\ln L + \beta_3 \ln K + u$$

$$\ln Q - \ln L = \beta_1 + \beta_3(\ln K - \ln L) + u$$

$$\ln\left(\frac{Q}{L}\right) = \beta_1 + \ln\left(\frac{K}{L}\right) + u \qquad (7.100)$$

Theo ngôn ngữ thống kê và kinh tế lượng, thì phương trình (7.99) được gọi là mô hình không ràng buộc (mô hình không giới hạn), và phương trình (7.100) được gọi là mô hình ràng buộc (mô hình giới hạn) (bởi giả thiết H_0). Nếu sau khi kiểm định, ta chấp nhập giả thiết H_0, điều này có nghĩa là chúng ta nên sử dụng mô hình giới hạn cho các mục đích phân tích chính sách và dự báo.

Đôi khi chúng ta đưa ra đồng thời nhiều ràng buộc chứ không chỉ có một ràng buộc duy nhất như trường hợp vừa xét. Ví dụ, giả sử ta có phương trình không giới hạn được cho như sau:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + u_t \qquad (7.101)$$

Và có hai ràng buộc đồng thời như sau:

$$H_0: \beta_3 + \beta_4 = 1 \text{ và } \beta_2 = \beta_5$$

Nếu thế các ràng buộc này vào phương trình (7.101), ta sẽ có phương trình sau đây:

$$Y_t = \beta_1 + \beta_5 X_{2t} + (1 - \beta_4) X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + u_t$$

$$Y_t = \beta_1 + \beta_5 X_{2t} + X_{3t} - \beta_4 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + u_t$$

$$Y_t - X_{3t} = \beta_1 + \beta_5(X_{2t} + X_{5t}) + \beta_4(X_{4t} - X_{3t}) + u_t$$

$$Y_t^* = \beta_1 + \beta_5 X_{1t}^* + \beta_4 X_{2t}^* + u_t \qquad (7.102)$$

Trong đó, $Y_t^* = Y_t - X_{3t}$, $X_{1t}^* = X_{2t} + X_{5t}$, và $X_{2t}^* = X_{4t} - X_{3t}$. Trong trường hợp này, phương trình (7.102) được gọi là mô hình giới hạn theo giả thiết H_0.

Có ba cách để thực hiện các kiểm định ràng buộc vừa nêu trên, đó là, Likelihood Ratio (*LR*), Wald, và Lagrane Multiplier (*LM*). Ý tưởng cơ bản của ba thủ tục kiểm định này là đánh giá *sự khác biệt* giữa mô

hình giới hạn và mô hình không giới hạn. Nếu (các) ràng buộc không ảnh hưởng nhiều đến mức độ phù hợp của mô hình, thì chúng ta có thể chấp nhận (các) ràng buộc là hợp lý. Ngược lại, nếu mô hình giới hạn không phù hợp bằng mô hình không giới hạn, thì chúng ta có thể bác bỏ giả thiết H_0 (bác bỏ mô hình giới hạn). Nếu mục đích chỉ nhằm kiểm định các ràng buộc tuyến tính giản đơn trong Eviews, thì nên sử dụng các thủ tục kiểm định Wald hoặc *LR*. Ngược lại, khi chúng ta muốn kiểm định các giả thiết phức tạp hơn, chẳng hạn như tương quan chuỗi hay ảnh hưởng ARCH, thì thủ tục kiểm định *LM* trở nên rất hữu ích (được trình bày ở phần phân tích tự tương quan và các mô hình ARCH). Ngoài ra, *LR* thường được sử dụng để kiểm định có nên đưa thêm hay bỏ bớt một hoặc một số biến giải thích vào hoặc ra khỏi mô hình hay không.

Kiểm định Wald

Bước 1: Xác định giả thiết H_0.

Bước 2: Ước lượng cả hai mô hình giới hạn và không giới hạn, và tính RSS_R và RSS_U. Trong đó, RSS_R và RSS_U lần lượt là RSS của mô hình giới hạn và mô hình không giới hạn.

Bước 3: Tính giá trị thống kê *F* theo công thức sau đây:

$$F_{stat} = \frac{\dfrac{(RSS_R - RSS_U)}{(k_U - k_R)}}{\dfrac{RSS_U}{(n - k_U)}} \qquad (7.103)$$

Trong đó, k_U và k_R là số biến giải thích trong mô hình không giới hạn và mô hình giới hạn, và *n* là số quan sát trong mẫu dữ liệu.

Bước 4: Tìm giá trị *F* tra bảng (F_{crit}) với số bậc tự do lần lượt là $(k_U - k_R)$ và $(n - k_U)$ theo hàm =FINV(α,k_U-k_R,n-k_U).

Bước 5: Nếu $F_{stat} > F_{crit}$ thì ta bác bỏ giả thiết H_0 cho rằng giả thiết về (các) ràng buộc là đúng.

Để thực hiện kiểm định Wald trên Eviews (ví dụ sử dụng tập tin **DATA7-2**), ta thực hiện như sau:

Bước 1: Ước lượng mô hình không giới hạn: ls log(Y) c log(L) log(K).

Bước 2: Từ cửa sổ kết quả hồi quy, ta chọn **View/Coefficient Tests/Wald-Coefficient Restrictions** ... rồi nhập điều kiện ràng buộc vào hộp thoại với quy ước về hệ số như sau: C(1) là hệ số cắt, C(2) là hệ số của biến giải thích thứ nhất, C(3) là hệ số của biến giải thích thứ hai, v.v... Ứng với giả thiết ở phương trình (7.99) và (7.100), ta nhập vào hộp thoại như sau: **C(2)+C(3)=1**. Sau khi chọn <OK>, ta có kết quả kiểm định như sau:

Test Statistic	Value	df	Probability
F-statistic	15.81852	(1, 150)	0.0001
Chi-square	15.81852	1	0.0001

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
-1 + C(2) + C(3)	-0.394508	0.099191

Bước 3: Vì giá trị F tính toán (15.82) lớn hơn giá trị F tra bảng ở mức ý nghĩa $\alpha = 5\%$ (3.9) hoặc giá trị xác suất p (0.01%) nhỏ hơn mức ý nghĩa $\alpha = 5\%$, nên ta bác bỏ giả thiết H_0: $\beta_2 + \beta_3 = 1$.

Kiểm định LR

Trong phân tích kinh tế lượng và dự báo, chúng ta thường gặp các vấn đề phải quyết định đưa thêm hay bỏ bớt một hoặc một số biến giải thích từ một mô hình vừa ước lượng. Khi chỉ xét một biến duy nhất, thì một tiêu chí an toàn nhất là kiểm tra tỷ số t, nhưng khi xét một nhóm các biến, thì chúng ta có lẽ nên đánh giá ảnh hưởng kết hợp của chúng lên mô hình. Xem xét mô hình sau đây:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + ut \qquad (7.104)$$

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + \beta_{k+1} X_{k+1t} + \ldots + \beta_m X_{mt} + u_t \qquad (7.105)$$

Trong trường hợp này, ta có mô hình giới hạn và mô hình không giới hạn với *m-k* biến giải thích cần đánh giá ảnh hưởng kết hợp để xem nên chọn mô hình (7.104) hay (7.105). Giả thiết ràng buộc ở đây sẽ là:

$$H_0: \beta_{k+1} = \beta_{k+2} = \ldots = \beta_m = 0$$

Như vậy, nếu mô hình lúc đầu đang xét là (7.105), thì ta có thể kiểm định xem có phải các biến X_{k+1t}, X_{k+2t}, ..., X_{mt} là những biến thừa trong mô hình (7.105) hay không. Ngược lại, nếu mô hình lúc đầu đang xét là (7.104), thì ta có thể kiểm định xem có phải các biến X_{k+1t}, X_{k+2t}, ..., X_{mt} là những biến quan trọng bị bỏ sót trong mô hình (7.104) hay không. Hai giả thiết này có thể được kiểm định bằng kiểm định *Wald* hoặc kiểm định *LR*. Thống kê *LR* được tính theo công thức sau đây:

$$LR = -2(L_R - L_U) \qquad (7.106)$$

Trong đó, L_R và L_U là các giá trị tối đa hóa của hàm log-likelihood của hai mô hình giới hạn và mô hình không giới hạn bởi giả thiết H_0. Thống kê *LR* theo phân phối χ^2 với số bậc tự do bằng số ràng buộc (hay số biến bị bỏ sót hoặc được đưa thêm).

Các bước thực hiện **kiểm định thừa biến** trên Eviews (sử dụng tập tin **DATA7-3**) sẽ như sau:

Bước 1: Ước lượng mô hình sau đây: ls log(wage) c educ exper tenure construc services trade

Bước 2: Từ cửa sổ kết quả hồi quy, ta chọn **View/Coefficient Tests/Redundant variables – Likelihood ratio**, rồi nhập tên các biến ở giả thiết muốn kiểm định (construc services trade).

Bước 3: Kết quả hồi quy (Bảng 7.2) cho thấy giá trị *F* tính toán (13.99) hoặc χ^2 tính toán (40.91) cao hơn giá trị *F* tra bảng (2.62) hoặc χ^2 tra bảng (7.82), hoặc giá trị xác suất *p* (0%) nhỏ hơn mức ý nghĩa $\alpha = 5\%$, ta có thể bác bỏ giả thiết cho rằng hệ số của các biến construc, services, và trade đồng thời bằng không, và vì thế các biến construc, services, và trade không phải là các biến thừa trong mô hình.

■ BẢNG 7.3: **Kiểm định thừa biến.**

Redundant Variables: CONSTRUC SERVICES TRADE

F-statistic	13.99346	Prob. F(3,519)	0.0000
Log likelihood ratio	40.91334	Prob. Chi-Square(3)	0.0000

Test Equation:
Dependent Variable: LOG(WAGE)
Method: Least Squares
Date: 06/21/09 Time: 14:37
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.284360	0.104190	2.729230	0.0066
EDUC	0.092029	0.007330	12.55525	0.0000
EXPER	0.004121	0.001723	2.391437	0.0171
TENURE	0.022067	0.003094	7.133070	0.0000

R-squared	0.316013	Mean dependent var	1.623268
Adjusted R-squared	0.312082	S.D. dependent var	0.531538
S.E. of regression	0.440862	Akaike info criterion	1.207406
Sum squared resid	101.4556	Schwarz criterion	1.239842
Log likelihood	-313.5478	Hannan-Quinn criter.	1.220106
F-statistic	80.39092	Durbin-Watson stat	1.768805
Prob(F-statistic)	0.000000		

Tương tự, các bước thực hiện **kiểm định thiếu biến** trên Eviews (sử dụng tập tin **DATA7-3**) sẽ như sau:

Bước 1: Ước lượng mô hình sau đây: ls log(wage) c educ exper tenure

Bước 2: Từ cửa sổ kết quả hồi quy, ta chọn **View/Coefficient Tests/Omitted variables – Likelihood ratio**, rồi nhập tên các biến ở giả thiết muốn kiểm định (construc services trade).

Bước 3: Kết quả hồi quy (Bảng 7.3) cho thấy giá trị F tính toán (13.99) hoặc χ^2 tính toán (40.91) cao hơn giá trị F tra bảng (2.62) hoặc χ^2 tra bảng (7.82), hoặc giá trị xác suất p (0%) nhỏ hơn mức ý nghĩa $\alpha = 5\%$, ta có thể nói rằng các biến construc, services, và trade thực sự là những biến đã bị bỏ

sót vì những biến này đóng một vai trò rất quan trọng trong việc xác định giá trị của log(wage).

■ BẢNG 7.4: **Kiểm định thiếu biến.**

Omitted Variables: CONSTRUC SERVICES TRADE

F-statistic	13.99346	Prob. F(3,519)	0.0000
Log likelihood ratio	40.91334	Prob. Chi-Square(3)	0.0000

Test Equation:
Dependent Variable: LOG(WAGE)
Method: Least Squares
Date: 06/21/09 Time: 14:44
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.434621	0.106601	4.077064	0.0001
EDUC	0.087859	0.007177	12.24091	0.0000
EXPER	0.004214	0.001678	2.510718	0.0124
TENURE	0.019514	0.003011	6.481313	0.0000
CONSTRUC	0.090136	0.091360	0.986604	0.3243
SERVICES	-0.311687	0.064095	-4.862920	0.0000
TRADE	-0.205976	0.043325	-4.754216	0.0000

R-squared	0.367199	Mean dependent var	1.623268
Adjusted R-squared	0.359883	S.D. dependent var	0.531538
S.E. of regression	0.425269	Akaike info criterion	1.141031
Sum squared resid	93.86325	Schwarz criterion	1.197793
Log likelihood	-293.0911	Hannan-Quinn criter.	1.163256
F-statistic	50.19378	Durbin-Watson stat	1.818683
Prob(F-statistic)	0.000000		

HIỆN TƯỢNG ĐA CỘNG TUYẾN

Để có thể sử dụng một mô hình hồi quy cho mục đích dự báo, điều quan trọng là mô hình hồi quy đó phải là một mô hình tốt. Bây giờ chúng ta sẽ lần lượt khảo sát (một cách ngắn gọn) hậu quả và cách thức khắc phục một số vấn đề thực tiễn thường hay gặp trong phân tích hồi quy. Trước hết, chúng ta sẽ xem xét hiện tượng đa cộng tuyến.

Giả định số 6 của hồi quy tuyến tính cổ điển cho rằng không có các mối quan hệ tuyến tính hoàn hảo giữa các giá trị mẫu của các biến giải thích. Trên thực tế, chúng ta thường gặp các mối quan hệ tuyến tính không hoàn hảo nhưng lại khá chặt chẽ, và vấn đề này luôn là một mối quan tâm của những người nghiên cứu và làm chính sách vì nó có thể tồn tại trong cả các mô hình hồi quy dữ liệu chéo và dữ liệu chuỗi thời gian. Trong phần này, chúng ta sẽ xem xét một cách ngắn gọn hậu quả của hiện tượng đa cộng tuyến hoàn hảo, không hoàn hảo, cách phát hiện đa cộng tuyến không hoàn hảo, và cách thức khắc phục.

HẬU QUẢ CỦA ĐA CỘNG TUYẾN HOÀN HẢO

Theo ngôn ngữ của toán ma trận, thì nếu có hiện tượng đa cộng tuyến hoàn hảo giữa X_{it} và X_{jt} ($\delta_i X_{it} + \delta_j X_{jt} = 0$) hoặc $Cov(X_{it}, X_{jt}) = 0$, thì chúng ta không thể nào xác định được giá trị của các định thức ở phương trình (7.62). Điều này có nghĩa, chúng ta không thể nào xác định được các nghiệm $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$ của phương trình này một cách duy nhất vì ma trận X trong phương trình (7.62) là một ma trận suy biến. Để làm rõ điều này, chúng ta hãy thực hiện một phân tích đơn giản sau đây.

BẢNG 7.5: Công thức tính các hệ số hồi quy.

Hồi quy đơn	Hồi quy bội
$Y = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{u}$ (7.3)	$Y = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{u}$ (7.54)
$\hat{\beta}_2 = \dfrac{\sum x_2 y}{\sum x_2^2}$ (7.17)	$\hat{\beta}_2 = \dfrac{(\sum yx_2)(\sum x_3^2) - (\sum yx_3)(\sum x_2 x_3)}{(\sum x_2^2)(\sum x_3^2) - (\sum x_2 x_3)^2}$ (7.64)
$\hat{\beta}_2 = \dfrac{Cov(X_2, Y)}{Var(X_2)}$ (7.18)	$\hat{\beta}_2 = \dfrac{Cov(X_2, Y)Var(X_3) - Cov(X_3, Y)Cov(X_2, X_3)}{Var(X_2)Var(X_3) - [Cov(X_2, X_3)]^2}$ (7.64a)

Nếu X_2 và X_3 có mối quan hệ tuyến tính hoàn hảo, thì chúng ta có công thức sau đây:

$$r_{2,3} = \frac{Cov(X_2, X_3)}{\sqrt{Var(X_2}\sqrt{Var(X_3}} = \pm 1 \qquad (7.107)$$

Như vậy, chuyển đổi, rồi thế công thức (7.107) vào công thức (7.64a), ta thấy rằng mẫu số của (7.64a) bằng không. Điều này có nghĩa, chúng ta không thể xác định được các ước lượng OLS nếu có hiện tượng đa cộng tuyến hoàn hảo. Hơn nữa, nếu X_2 và X_3 độc lập hoàn toàn, nghĩa là $Cov(X_2,X_3) = 0$, thì công thức tính $\hat{\beta}_2$ ở (7.64a) và (7.18) là như nhau.

Đa cộng tuyến hoàn hảo thực sự là một vấn đề hết sức nghiêm trọng. Tuy nhiên, điều này hiếm khi xảy ra đối với dữ liệu trên thực tế. Sự hiện diện của đa cộng tuyến hoàn hảo thường xảy ra đối với một số lỗi như bẫy biến giả.

HẬU QUẢ CỦA ĐA CỘNG TUYẾN KHÔNG HOÀN HẢO

Trong hồi quy đa biến, nhất là hồi quy chuỗi thời gian, thường có hiện tượng các biến giải thích có một mối quan hệ tuyến tính nhất định nào đó. Cho nên, vấn đề quan trọng là chúng ta cần nhận diện mức độ đa cộng tuyến có nghiêm trọng hay không để đảm bảo kết quả hồi quy là đáng mong muốn. Đa cộng tuyến không hoàn hảo có thể dẫn đến nhiều hậu quả nghiêm trọng, đáng chú ý nhất là các hậu quả sau đây:

(1) Các giá trị ước lượng của các hệ số hồi quy OLS có thể không chính xác do có sai số chuẩn, $se(\hat{\beta}_k)$, quá lớn, làm cho các khoảng tin cậy của các tham số thực của tổng thể rộng hơn. Nếu điều này xảy ra, thì khả năng chấp nhận giả thiết H_0 của các hệ số hồi quy riêng sẽ tăng. Chúng ta biết rằng, trong các mô hình hồi quy đơn và hồi quy bội (hai biến giải thích), thì phương sai của các hệ số hồi quy được cho bởi các công thức sau đây:

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_t^2} \qquad (7.29)$$

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2t}^2 (1 - r_{23}^2)} \qquad (7.70)$$

Nếu mở rộng cho trường hợp có hơn hai biến giải thích, thì phương sai của hệ số $\hat{\beta}_j$ sẽ được cho bởi công thức sau đây:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2 (1 - R_j^2)} \qquad (7.108)$$

Trong đó, R_j^2 là hệ số xác định của mô hình hồi quy phụ của biến X_j theo tất cả các biến giải thích khác. Theo các công thức (7.70) và (7.108), nếu r_{23}^2 hoặc R_j^2 bằng không (các biến giải thích độc lập nhau hoàn toàn), thì phương sai của các hệ số hồi quy riêng trong mô hình hồi quy bội sẽ đúng bằng phương sai của nó trong mô hình hồi quy đơn. Khi hệ số xác định tăng lên, thì phương sai của các hệ số hồi quy riêng sẽ tăng lên. Như thế, $\text{se}(\hat{\beta}_j)$ sẽ tăng, và làm cho khoảng tin cậy $\hat{\beta}_j \pm \text{se}(\hat{\beta}_j) t_{\alpha/2}$ sẽ rộng hơn so với trường hợp không có đa cộng tuyến.

(2) Các hệ số hồi quy bị ảnh hưởng bởi đa cộng tuyến có thể sẽ không có ý nghĩa thống kê bởi vì có các giá trị thống kê *t* thấp, và điều này làm cho người phân tích loại bỏ một cách nhầm lẫn các biến quan trọng ra khỏi mô hình. Theo định nghĩa ở các phần trên, tỷ số *t* tính toán được tính theo công thức $t_{stat} = \hat{\beta}_j \Big/ \text{se}(\hat{\beta}_j)$, nên khi $\text{se}(\hat{\beta}_j)$ tăng sẽ làm t_{stat} giảm.

(3) Dấu của các hệ số hồi quy có thể sai so với kỳ vọng (từ cơ sở lý thuyết). Chính vì thế, nếu người nghiên cứu và người ra quyết định bất cẩn, có thể ra những quyết định sai lầm từ các kết quả nghiên cứu bị hiện tượng đa cộng tuyến. Ở công thức (7.64a), thông thường dấu của hệ số $\hat{\beta}_2$ phụ thuộc vào mối tương quan giữa X_2 và Y, nhưng một khi mối quan hệ giữa X_2

và X_3 quá mạnh (giá trị của $Cov(X_2,X_3)$ quá lớn một cách tương đối) có thể làm thay đổi dấu của hệ số hồi quy.

(4) Kết quả hồi quy rất nhạy cảm với chỉ một vài thay đổi nhỏ trong bộ dữ liệu. Nghĩa là, các hệ số hồi quy sẽ thay đổi một cách đáng kể chỉ với việc bỏ bớt, thêm vào một vài quan sát, hoặc thay đổi giá trị của một vài quan sát. Điều này cũng rất nguy hiểm trong nghiên cứu và ra quyết định từ kết quả hồi quy.

PHÁT HIỆN ĐA CỘNG TUYẾN

Có nhiều cách giúp phát hiện đa cộng tuyến trước và sau khi thực hiện việc ước lượng mô hình.

(1) **Hệ số tương quan**. Hầu hết các nhà nghiên cứu kinh tế lượng cho rằng khi hệ số tương quan giữa hai biến giải thích nào đó bằng hoặc cao hơn 0.9, thì đó là một dấu hiệu quan trọng xảy ra hiện tượng đa cộng tuyến. Trong Eviews, ta có thể tạo ma trận hệ số tương quan như sau: **Quick/Group Statistics/Correlations**, rồi nhập tên các biến giải thích vào, chọn <OK>.

(2) **Quan sát kết quả hồi quy**. Sau khi đã thực hiện ước lượng phương trình, chúng ta có thể quan sát ba thông tin sau đây: dấu của các hệ số ước lượng, tỷ số *t* tính toán, và R^2. Ví dụ, nếu R^2 cao nhưng tỷ số *t* lại thấp thì nguy cơ là có đa cộng tuyến.

(3) **Hồi quy phụ**. Sau khi hồi quy, chúng ta có thể thực hiện các hồi quy phụ. Các "ứng viên" làm biến phụ thuộc trong các hồi quy phụ thường là các biến có dấu hiệu bất thường từ kết quả hồi quy ban đầu. Nếu có tồn tại đa cộng tuyến, thì kết quả hồi quy phụ có sai số chuẩn của ước lượng thấp, R^2 cao, và các tỷ số *t* tính toán cao.

KHẮC PHỤC ĐA CỘNG TUYẾN

Có nhiều cách khắc phục đa cộng tuyến, nhưng phổ biến nhất là các cách sau đây:

(1) **Chuyển đổi dạng biến.** Tạo một biến giải thích mới (X^*_{jt}) như sau:

$$X^*_{jt} = \frac{(X_{jt} - \overline{X}_j)}{\sqrt{Var(X_{jt}}} \qquad (7.109)$$

Rồi sử dụng biến X^*_{jt} thế cho biến X_{jt} có thể giúp giảm đáng kể hiện tượng đa cộng tuyến.

(2) **Nhận diện và loại bỏ một hoặc một số biến trong các biến thực sự có hệ số tương quan khá cao.** Như chúng ta sẽ biết ở phần sau, nếu loại bỏ một biến không cần thiết ra khỏi mô hình thì kết quả ước lượng không bị ảnh hưởng. Tuy nhiên, tránh trường hợp loại bỏ những biến quan trọng vì điều này dẫn đến một vấn đề còn nghiêm trọng hơn là "sai dạng mô hình".

(3) **Thu thập thêm dữ liệu.** Khi số quan sát tăng lên thì $\sum x_j^2$ sẽ tăng, và điều này có thể làm giảm phương sai của $\hat{\beta}_j$.

VÍ DỤ MINH HỌA

Sử dụng tập tin **DATA7-4**, trong đó chứa các thông tin theo quý của các biến kim ngạch nhập khẩu (IMP), sản lượng quốc nội (GDP), chỉ số giá tiêu dùng (CPI), và chỉ số giá sản xuất (PPI).

Bước 1: Xác định ma trận hệ số tương quan bằng cách chọn Quick/Group Statistics/Correlations, rồi nhập các biến IMP GDP CPI PPI.

	IMP	GDP	CPI	PPI
IMP	1	0.987	0.877	0.879
GDP	0.987	1	0.880	0.892
CPI	0.877	0.880	1	**0.991**
PPI	0.879	0.892	0.991	1

Bước 2: Ước lượng phương trình:

ls log(IMP) C log(GDP) log(CPI) log(PPI)

Dependent Variable: LOG(IMP)
Method: Least Squares
Date: 06/21/09 Time: 18:04
Sample: 1990Q1 2001Q3
Included observations: 47

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.394233	0.281260	1.401665	0.1682
LOG(GDP)	2.247337	0.108137	20.78236	0.0000
LOG(CPI)	0.553845	0.179087	3.092604	0.0036
LOG(PPI)	**-0.574464**	0.222176	-2.585628	0.0132

R-squared	0.980804	Mean dependent var	10.76531
Adjusted R-squared	0.979465	S.D. dependent var	0.164217
S.E. of regression	0.023532	Akaike info criterion	-4.579612
Sum squared resid	0.023812	Schwarz criterion	-4.422153
Log likelihood	111.6209	Hannan-Quinn criter.	-4.520359
F-statistic	732.3599	Durbin-Watson stat	0.612992

Bước 3: Ước lượng các phương trình sau đây:

ls log(IMP) c log(GDP) log(CPI)

ls log(IMP) c log(GDP) log(PPI)

Các hệ số hồi quy của biến log(CPI) và log(PPI) trong hai mô hình này đều có dấu dương, nhưng chỉ có hệ số của log(CPI) có ý nghĩa thống kê.

Bước 4: Hồi quy phụ:

ls log(PPI) c log(GDP) log(CPI)

Dependent Variable: LOG(PPI)
Method: Least Squares
Date: 06/21/09 Time: 18:01
Sample: 1990Q1 2001Q3
Included observations: 47

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.195247	0.188563	-1.035444	0.3061
LOG(GDP)	0.193559	0.067324	2.875060	0.0062
LOG(CPI)	0.777596	0.032006	24.29574	0.0000

R-squared	0.984785	Mean dependent var	4.494072
Adjusted R-squared	0.984094	S.D. dependent var	0.126606
S.E. of regression	0.015968	Akaike info criterion	-5.374795
Sum squared resid	0.011219	Schwarz criterion	-5.256700
Log likelihood	129.3077	Hannan-Quinn criter.	-5.330355
F-statistic	1423.951	Durbin-Watson stat	0.164358

Từ kết quả trên, chúng ta có các nhận xét như sau:

(1) Hệ số tương quan giữa các biến giải thích rất cao, điều này có thể cho biết có tồn tại đa cộng tuyến và mức độ rất nghiêm trọng giữa CPI và PPI. Tuy nhiên, theo lý thuyết thì việc chỉ nhìn vào hệ số tương quan giữa các biến giải thích chưa đủ cơ sở để kết luận chắc chắn có hiện tượng đa cộng tuyến.

(2) Các sai số chuẩn và tỷ số *t* của các hệ số hồi quy thay đổi từ ước lượng này qua ước lượng khác, điều này cho biết vấn đề đa cộng tuyến trong trường hợp này rất nghiêm trọng.

(3) Tính ổn định của các hệ số ước lượng cũng có vấn đề, chuyển từ dương sang âm cho cùng biến log(PPI).

(4) R^2 trong mô hình hồi quy phụ rất cao, điều này khẳng định rõ ràng rằng mức độ đa cộng tuyến rất nghiêm trọng và vì thế đã ảnh hưởng đến kết quả ước lượng.

(5) Giải pháp đề nghị đơn giản nhất là loại PPI hay CPI ra khỏi mô hình. Điều này tùy thuộc vào hệ số tương quan của GDP với hai biến này, hoặc còn phụ thuộc vào quan điểm của nhà

nghiên cứu là PPI hay CPI là biến được xem là quan trọng nhất khi giải thích GDP. Ngoài ra, điều này còn tùy thuộc vào sự sẵn có của dữ liệu.

Như vậy, nếu một mô hình hồi quy có hiện tượng đa cộng tuyến, thì chúng ta nên tìm cách khắc phục trước khi sử dụng cho các mục đích và phân tích chính sách.

HIỆN TƯỢNG TỰ TƯƠNG QUAN

Đáng lý ra chúng ta phải xem xét hiện tượng phương sai thay đổi trước khi chuyển sang vấn đề tương quan chuỗi[1]. Tuy nhiên, hiện tượng phương sai thay đổi thường được đề cập trong phân tích hồi quy dữ liệu chéo. Đối với dữ liệu chuỗi thời gian, hiện tượng phương sai thay đổi theo thời gian cũng là một vấn đề đáng quan tâm và đã được nhiều nhà kinh tế lượng tài chính quan tâm nghiên cứu từ thập niên 1970. Vì tính quan trọng của nó, nên chúng tôi quyết định dành một chương riêng trong giáo trình này bàn về vấn đề phương sai thay đổi theo thời gian. Đó là các mô hình ARCH. Các mô hình này sẽ được trình bày chi tiết ở chương 9. Bây giờ, chúng ta tập trung xem xét một vấn đề rất được quan tâm trong kinh tế lượng và dự báo chuỗi thời gian: tự tương quan hay tương quan chuỗi.

HIỆN TƯỢNG TỰ TƯƠNG QUAN

Giả định số 8 trong mô hình hồi quy tuyến tính cổ điển cho rằng hiệp phương sai và hệ số tương quan giữa các hạng nhiễu khác nhau là bằng không.

$$\text{Cov}(u_t, u_s) = 0 \text{ cho tất cả các } t \neq s \qquad (7.110)$$

Giả định này phát biểu rằng các hạng nhiễu u_t và u_s, được phân phối một cách độc lập, nghĩa là không có tương quan chuỗi. Tuy nhiên, trong kinh tế lượng chuỗi thời gian thường xảy ra hiện tượng một hạng nhiễu ở thời điểm *t* có thể có quan hệ với một hạng nhiễu ở thời điểm *s*. Tự tương quan thường xảy ra trong khung phân tích chuỗi thời

[1] Theo nhiều tài liệu về kinh tế lượng truyền thống.

gian. Khi dữ liệu được thu thập theo thứ tự thời gian, thì hạng nhiễu ở giai đoạn này có thể ảnh hưởng đến hạng nhiễu ở giai đoạn kế tiếp (hoặc một số giai đoạn kế tiếp nhau).

NGUYÊN NHÂN CỦA TỰ TƯƠNG QUAN

Có nhiều cách lý giải hiện tượng tự tương quan, nhưng thường có ba nhóm nguyên nhân sau đây. Nguyên nhân thứ nhất có thể dẫn đến hiện tượng tự tương quan là do *bỏ sót biến quan trọng*. Ví dụ, Y_t thực sự phụ thuộc vào X_{2t} và X_{3t}, vì một lý do nào đó mà người nghiên cứu không đưa X_{3t} vào mô hình. Như vậy, ảnh hưởng của X_{3t} sẽ được bao hàm trong hạng nhiễu u_t. Nếu X_{3t}, cũng như nhiều chỉ báo kinh tế khác có phụ thuộc vào $X_{3,t-1}$, $X_{3,t-2}$, v.v... Điều này sẽ dẫn đến một hệ quả không thể tránh khỏi là tồn tại mối tương quan giữa u_t và u_{t-1}, u_{t-2}, v.v... Như vậy, các biến bị bỏ sót là một nguyên nhân của tự tương quan.

Tự tương quan cũng có thể xảy ra do *lỗi sai dạng hàm*. Giả sử, Y_t phụ thuộc vào X_{2t} theo dạng hàm bậc hai, nghĩa là $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{2t}^2 + u_t$, nhưng người phân tích lại giả sử và ước lượng mô hình tuyến tính $Y_t = \beta_1 + \beta_2 X_{2t} + u_t$. Như vậy, hạng nhiễu từ mô hình tuyến tính sẽ phụ thuộc vào X_{2t}^2. Nếu X_{2t} là một hàm tăng hoặc giảm theo thời gian, thì u_t cũng sẽ là một hàm tăng hoặc giảm theo thời gian. Điều này chứng tỏ tự tương quan là do xác định sai dạng hàm.

Nguyên nhân thứ ba là do *lỗi sai sót hệ thống trong việc đo lường*. Giả sử một công ty cập nhật tồn kho của mình định kỳ theo thời gian, nếu một lỗi hệ thống xảy ra trong việc đo lường (ví dụ do ước lượng quá cao tồn kho ở một giai đoạn nào đó sẽ dẫn đến ước lượng quá cao ở các giai đoạn tiếp theo), thì lượng tồn kho tích lũy sẽ thể hiện các sai số do đo lường.

Trong kinh tế lượng chuỗi thời gian, người ta rất quan tâm đến việc phân loại hiện tượng tự tương quan do sai dạng mô hình với hiện tượng tự tương quan thuần túy. Hiện tượng tự tương quan do sai dạng mô hình có thể dễ dàng khắc phục bằng việc kiểm tra và xác định lại dạng mô hình thích hợp. Ngược lại, hiện tượng tự tương quan thuần

túy là do bản chất nội tại của các chuỗi thời gian, khi đó, dù đã chuyển đổi dạng mô hình nhưng vẫn tồn tại tự tương quan. Đây là vấn đề chúng ta quan tâm nhiều hơn trong quá trình phân tích.

Một điểm quan trọng nữa cần lưu ý khi phân tích hồi quy chuỗi thời gian là chúng ta nên để ý đến việc phân tích chẩn đoán phần dư. Vì nếu phần dư không ngẫu nhiên, không có phân phối chuẩn là một dấu hiệu của khả năng tự tương quan. Và nếu điều này xảy ra, thì chúng ta trước hết nên xem xét lại dạng mô hình. Chỉ khi nào mô hình được xác định đúng, không có tự tương quan (và không có phương sai thay đổi) thì chúng ta mới có thể sử dụng kết quả hồi quy cho các mục đích phân tích chính sách và dự báo.

HẬU QUẢ CỦA TỰ TƯƠNG QUAN

Giả sử chúng ta xét trường hợp đơn giản nhất và phổ biến nhất là tự tương quan bậc một. Giả sử ta có phương trình sau:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{7.111}$$

Trong đó, quan sát hiện tại của hạng nhiễu (u_t) là một hàm của quan sát trước đó (độ trễ) của hạng nhiễu (u_{t-1}):

$$u_t = \rho u_{t-1} + \varepsilon_t \tag{7.112}$$

Ở đây, ρ là hệ số tự tương quan bậc một và ε_t là một hạng nhiễu mới được giả định có phân phối chuẩn. Hệ số ρ có giá trị từ -1 đến +1. Lưu ý, ta giả định $|\rho| < 1$ nhằm tránh trường hợp "gia tăng đột biến" khi đó ρ có thể lớn hơn 1. Vấn đề này sẽ được đề cập ở chương 8 khi bàn về các mô hình ARIMA. Hệ số tự tương quan bậc một có thể được định nghĩa như sau:

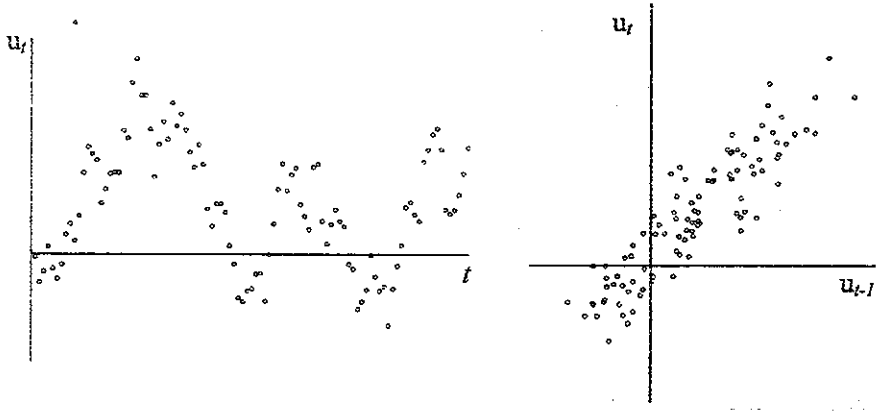
$$\rho = \frac{E\{[u_t - E(u_t)][u_{t-1} - E(u_{t-1})]\}}{\sqrt{\text{var}(u_t)}\sqrt{\text{var}(u_{t-1})}} \tag{7.113}$$

Giá trị của ρ sẽ cho biết mức độ của sự tương quan chuỗi, và chúng ta thường quan tâm đến ba trường hợp sau đây:

(1) Nếu $\rho = 0$, thì chúng ta có thể nói rằng không có tương quan chuỗi, bởi vì khi đó $u_t = \varepsilon_t$ và vì thế u_t là một hạng nhiễu có phân phối chuẩn.

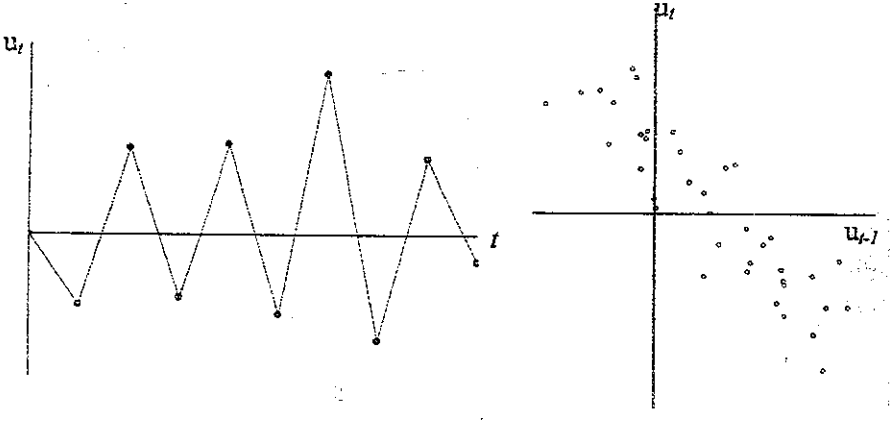
(2) Nếu ρ dần tới 1, thì giá trị quan sát trước đó của hạng nhiễu (u_{t-1}) trở nên quan trọng trong việc xác định xác định giá trị quan sát hiện hành của hạng nhiễu (u_t), và vì thế khả năng tồn tại hiện tượng tự tương quan dương càng lớn. Trong trường hợp này, quan sát hiện hành của hạng nhiễu có xu hướng mang cùng dấu với quan sát trước đó của hạng nhiễu.

■ HÌNH 7.3: **Tự tương quan dương.**



(3) Nếu ρ dần tới -1, mức độ tương quan chuỗi cũng rất mạnh. Tuy nhiên, trong trường hợp này chúng ta có hiện tượng tự tương quan âm. Nghĩa là, dấu các quan sát của hạng nhiễu sẽ thay đổi liên tục.

■ HÌNH 7.4: **Tự tương quan âm.**



Theo Pindyck và Rubinfied (1998), trong phân tích kinh tế lượng và dự báo, chúng ta thường quan tâm nhiều đến vấn đề tự tương quan dương hơn là tự tương quan âm. Tự tương quan dương thường xảy ra trong các nghiên cứu chuỗi thời gian do lỗi đo lường hoặc do bỏ sót biến giải thích.

Để phân tích hậu quả của tự tương quan khi sử dụng phương pháp hồi quy OLS, ta thực hiện như sau. Từ phương trình (7.112), ta lấy giá trị kỳ vọng cả hai vế, và có kết quả như sau:

$$E(u_t) = \rho E(u_{t-1}) + E(\varepsilon_t) = 0 \qquad (7.114)$$

Nên

$$Var(u_t) = \rho^2 Var(u_{t-1}) + Var(\varepsilon_t)$$

$$Var(u_t) = \frac{\sigma^2_{\varepsilon_t}}{1-\rho^2} \qquad (7.115)$$

$$\sigma^2_{OLS} = \frac{\sigma^2_{GLS}}{1-\rho^2} \qquad (7.116)$$

Quy trình ước lượng với việc kết hợp đồng thời giữa (7.111) và (7.112) được gọi là được gọi là ước lượng theo phương pháp bình phương bé nhất tổng quát (GLS). Như vậy, có sự khác biệt đáng kể giữa σ^2_{OLS} và σ^2_{GLS}. Cụ thể như sau:

(1) Nếu không có tự tương quan, $\rho = 0$, thì phương sai hạn nhiễu theo OLS và GLS bằng nhau.

(2) Nếu có tự tương quan, $\rho \neq 0$, thì phương sai hạn nhiễu theo OLS > phương sai hạn nhiễu theo GLS (do $(1-\rho^2) < 1$).

Vậy, khi có tự tương quan, nếu ước lượng theo OLS sẽ dẫn đến các hậu quả quan trọng sau đây:

(1) Các ước lượng OLS của các $\hat{\beta}_j$ vẫn là các ước lượng không chệch và nhất quán. Điều này bởi vì vấn đề không chệch và nhất quán không phụ thuộc vào giả định số 7.

(2) Các ước lượng OLS sẽ không còn là các ước lượng hiệu quả nữa nên không thỏa mãn tính chất BLUE.

(3) Trong trường hợp tự tương quan dương (vốn thường xảy ra nhất trong chuỗi thời gian), các giá trị ước lượng của sai số chuẩn theo OLS có xu hướng nhỏ hơn các sai số chuẩn thực sự của tổng thể. Nói cách khác, các ước lượng OLS vẫn không chệch, nhưng sai số chuẩn của hồi quy ($\hat{\sigma}$) sẽ bị chệch theo hướng thấp hơn. Và vì thế các sai số chuẩn của các hệ số hồi quy có xu hướng nhỏ hơn. Điều này dễ dẫn đến khả năng kết luận nhầm lẫn rằng các giá trị ước lượng OLS có độ chính xác cao. Chính vì thế, chúng ta có xu hướng bác bỏ giả thiết H_0 (khi nhìn vào kết quả ước lượng OLS), trong khi, thật sự chúng ta nên chấp nhận giả thiết H_0. Nói chung, các sai số chuẩn của các hệ số hồi quy OLS sẽ bị chệch và không nhất quán, và vì thế việc kiểm định thống kê sẽ không còn đáng tin cậy nữa. Trong hầu hết các trường hợp, R^2 sẽ luôn bị ước lượng quá mức (dễ nhầm lẫn mức độ phù hợp cao), và các tỷ số *t* tính toán có xu hướng cao hơn.

PHÁT HIỆN TỰ TƯƠNG QUAN

Phương pháp đồ thị

Một cách đơn giản nhất để phát hiện tự tương quan là xem xét đồ thị vẽ phần dư theo thời gian hoặc vẽ phần dư $\hat{u}_t$ với phần dư $\hat{u}_{t-1}$. Sử dụng tập tin **DATA7-5**, trong đó, LCONS là chi tiêu tiêu dùng (triệu đôla), LDISP là thu nhập khả dụng (triệu đôla), và LPRICE là chỉ số giá tương đối của lương thực. Mục tiêu của nghiên cứu này là muốn ước lượng hệ số co giãn của chi tiêu tiêu dùng theo thu nhập khả dụng và giá của lương thực. Kết quả của nghiên cứu có thể hữu ích cho cả các công ty kinh doanh hàng tiêu dùng hoặc các nhà hoạch định chính sách vĩ mô, hoặc chính sách ngành. Tất cả các biến này đã ở dạng logarithms.

Bước 1: Ước lượng phương trình hồi quy sau đây:

```
ls lcons c ldisp lprice
```

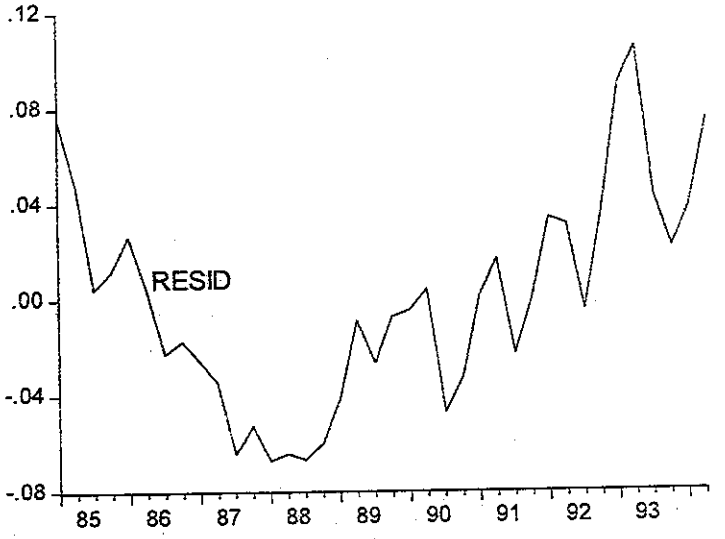
Dependent Variable: LCONS
Method: Least Squares
Date: 06/22/09 Time: 11:14
Sample: 1985Q1 1994Q2
Included observations: 38

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.485434	0.788349	3.152708	0.0033
LDISP	0.529285	0.292327	1.810589	0.0788
LPRICE	-0.064029	0.146506	-0.437040	0.6648

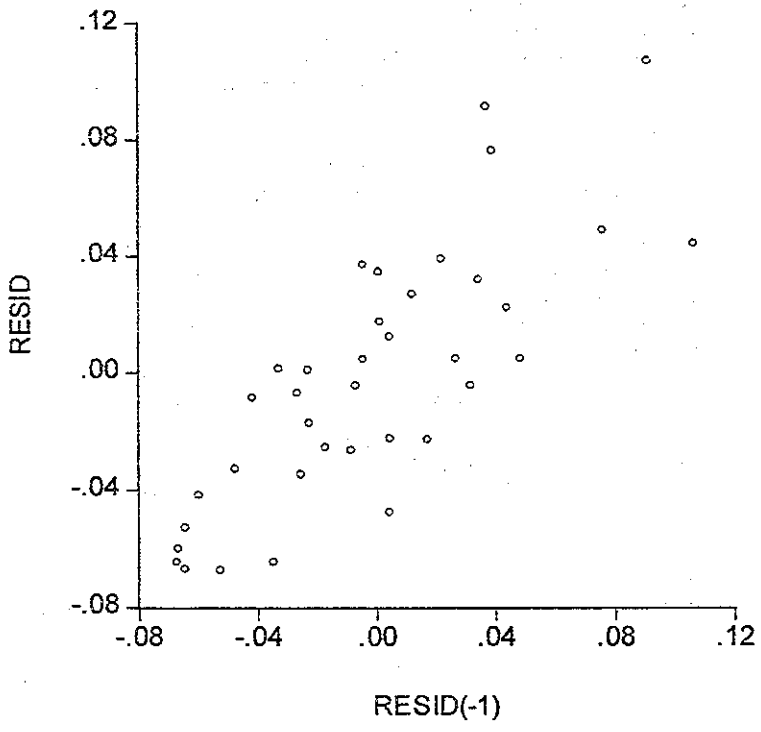
R-squared	0.234408	Mean dependent var	4.609274
Adjusted R-squared	0.190660	S.D. dependent var	0.051415
S.E. of regression	0.046255	Akaike info criterion	-3.233656
Sum squared resid	0.074882	Schwarz criterion	-3.104373
Log likelihood	64.43946	Hannan-Quinn criter.	-3.187658
F-statistic	5.358118	Durbin-Watson stat	0.370186

Bước 2: Từ kết quả ước lượng, ta vẽ đồ thị phần dư (RESID) theo thời gian, hoặc đồ thị phân tán RESID theo RESID(-1).

■ HÌNH 7.5: **Phần dư theo thời gian.**



■ HÌNH 7.6: **Phần dư RESID theo RESID(-1).**



Bước 3: Nhận xét.

Từ hai đồ thị trên, chúng ta dễ dàng nhận thấy rằng phần dư của mô hình hồi quy có tương quan chuỗi. Cụ thể, có tự tương quan dương. Lưu ý, ngoài các dạng đồ thị trên, chúng ta còn có thể sử dụng các dạng đồ thị khác như giản đồ tự tương quan, đồ thị tần suất, và đồ thị RESID(-1) và RESID theo thời gian.

Kiểm định Durbin-Watson

Kiểm định thống kê được sử dụng phổ biến nhất để phát hiện sự hiện diện của hiện tượng tự tương quan là kiểm định Durbin-Watson

(1950). Kiểm định Durbin-Watson (DW) chỉ có giá trị khi mô hình hồi quy thỏa mãn các giả định sau đây:

(1) Mô hình hồi quy phải có hệ số cắt;

(2) Tương quan chuỗi được giả định dưới dạng tự tương quan bậc một;

(3) Mô hình hồi quy không có các biến giải thích là biến trễ của biến phụ thuộc (các mô hình tự hồi quy);

(4) Không được "thiếu quan sát", nghĩa là, trật tự chuỗi dữ liệu phải được liên tục theo thời gian.

Giả sử, chúng ta có mô hình sau đây:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u_t \qquad t = 1, 2, 3, \ldots, n \tag{7.53}$$

Trong đó,

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad |\rho| < 1 \tag{7.112}$$

Với giả thiết H_0: $\rho = 0$ có nghĩa là mô hình hồi quy không có hiện tượng tự tương quan bậc 1, thì kiểm định Durbin-Watson được thực hiện như sau:

Bước 1: Ước lượng mô hình (7.53) theo OLS và lưu phần dư $\hat{u}_t$.

Bước 2: Tính giá trị tính toán của thống kê Durbin-Watson, ký hiệu là d theo công thức sau đây:

$$d = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2} \tag{7.117}$$

Giải thích thêm về giá trị giới hạn của thống kê d

Công thức này có thể được triển khai như sau:

$$d = \frac{\sum \hat{u}_t^2 + \sum \hat{u}_{t-1}^2 - 2\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2} \tag{7.118}$$

Do $\sum \hat{u}_t^2$ và $\sum \hat{u}_{t-1}^2$ chỉ khác nhau một quan sát, nên chúng được xem là xấp xỉ bằng nhau, vậy công thức (7.118) có thể được viết gọn lại như sau:

$$d = 2\left(1 - \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}\right) \qquad (7.119)$$

Nếu đặt $\hat{\rho} = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}$ (là ước lượng của ρ ở công thức (7.113)), vậy d ở công thức (7.119) được viết lại như sau:

$$d \approx 2(1 - \hat{\rho}) \qquad (7.120)$$

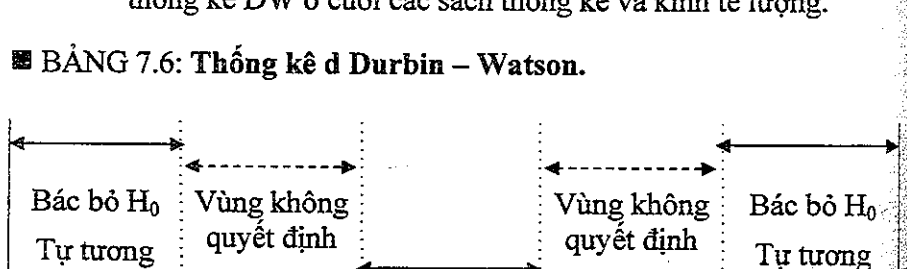
- Nếu $\hat{\rho} = 0$, thì $d = 2$, không có tự tương quan.
- Nếu $\hat{\rho} = 1$, thì $d = 0$, tự tương quan dương hoàn hảo.
- Nếu $\hat{\rho} = -1$, thì $d = 4$, tự tương quan âm hoàn hảo.

Như vậy, ta có khoảng giá trị của thống kê d như sau:

$$0 \le d \le 4 \qquad (7.121)$$

Bước 3: Lập bảng kiểm định Durbin-Watson (Bảng 7.6) với các giá trị tra bảng của d_U, d_L, $4 - d_U$, và $4 - d_L$ được lấy từ bảng thống kê DW ở cuối các sách thống kê và kinh tế lượng.

■ BẢNG 7.6: **Thống kê d Durbin – Watson.**



Bước 4a: Để kiểm định mô hình có hiện tượng tự tương quan dương hay không, chúng ta đặt giả thiết như sau:

H_0: $\rho = 0$ (không có tự tương quan)

H_1: $\rho > 0$ (tự tương quan dương)

Quyết định:

(1) Nếu $d \leq d_L$, chúng ta bác bỏ giả thiết H_0 và kết luận rằng mô hình có tự tương quan dương.

(2) Nếu $d \geq d_U$, chúng ta không thể bác bỏ H_0 và vì thế mô hình không có tự tương quan dương.

(3) Trong trường hợp, $d_L < d < d_U$, chúng ta không có quyết định gì về kết quả kiểm định.

Bước 4b: Để kiểm định mô hình có hiện tượng tự tương quan âm hay không, chúng ta đặt giả thiết như sau:

H_0: $\rho = 0$ (không có tự tương quan)

H_1: $\rho < 0$ (tự tương quan âm)

Quyết định:

(1) Nếu $d \geq 4 - d_L$, chúng ta bác bỏ giả thiết H_0 và kết luận rằng mô hình có tự tương quan âm.

(2) Nếu $d \leq 4 - d_U$, chúng ta không thể bác bỏ H_0 và vì thế mô hình không có tự tương quan âm.

(3) Trong trường hợp, $4 - d_U < d < 4 - d_L$, chúng ta không có quyết định gì về kết quả kiểm định.

Trường hợp không thể xác định về kiểm định DW thường do vấn đề phân phối của mẫu nhỏ đối với thống kê DW phụ thuộc vào các biến giải thích và rất khó xác định. Chính vì thế, thủ tục kiểm định LM (sẽ trình bày sau) là kiểm định nên dùng tiếp theo kiểm định DW.

Một quy tắc kinh nghiệm về kiểm định DW

(1) Do $\hat{\rho} = 0$, $d = 2$, nên nếu giá trị của d gần 2 cho biết mô hình không có tự tương quan.

(2) Do $\hat{\rho} \approx 1$, $d \approx 1$, nên nếu giá trị của d gần bằng 0 cho biết mô hình bị tự tương quan dương.

(3) Do $\hat{\rho} \approx -1$, $d \approx 4$, nên nếu giá trị của d gần bằng 4 cho biết mô hình bị tự tương quan âm.

Kiểm định DW trên Eviews

Eviews luôn báo cáo kết quả thống kê kiểm định DW trực tiếp trong phần các thống kê chẩn đoán của mỗi bảng kết quả hồi quy. Công việc duy nhất còn lại mà người phân tích phải làm là xây dựng Bảng (7.6) với đầy đủ các giá trị tra bảng và kết luận có hiện diện tự tương quan hay không. Tiếp tục với kết quả ước lượng từ Bước 1 của phương pháp đồ thị, ta thấy thống kê DW tính toán là 0.37. Trong bảng DW, với n = 38, k' = 2, thì $d_L = 1.17$ và $d_U = 1.38$ ở mức ý nghĩa 1%; hoặc $d_L = 1.37$ và $d_U = 1.59$ ở mức ý nghĩa 5%. Rõ ràng là d = 0.37 nhỏ hơn d_L rất nhiều, vậy đây là minh chứng rất rõ cho thấy mô hình bị tự tương quan dương.

Kiểm định LM của Breusch-Godfrey

Do kiểm định DW chỉ hạn chế khi kiểm định hiện tượng tự tương quan bậc 1, nên không thể áp dụng cho trường hợp tổng quát. Chẳng hạn, (a) DW có thể đưa đến khả năng "không biết quyết định thế nào"; (b) DW không thể áp dụng cho trường hợp mô hình có các biến trễ của biến phụ thuộc; và (c) DW không xét trường hợp tự tương quan bậc cao.

Chính vì các lý do này, Breusch (1978) và Godfrey (1978) phát triển kiểm định LM để có thể áp dụng cho tất cả các trường hợp vừa nêu trên. Xét mô hình sau đây:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u_t \qquad t = 1, 2, 3, \ldots, n \tag{7.53}$$

Trong đó,

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_p u_{t-p} + \varepsilon_t \tag{7.122}$$

Kiểm định LM của Breusch-Godfrey kết hợp hai mô hình (7.53) và (7.122) như sau:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_p u_{t-p} + \varepsilon_t \quad (7.123)$$

Và đặt giả thiết như sau:

H_0: $\rho_1 = \rho_2 = \rho_3 = \ldots = \rho_p = 0$ (không có tự tương quan)

H_1: Có ít nhất một hệ số ρ khác không, và vì thế có tự tương quan.

Quy trình kiểm định LM của Breusch-Godfrey

Bước 1: Ước lượng phương trình (7.53) và lưu phần dư $\hat{u}_t$

Bước 2: Ước lượng mô hình hồi quy sau đây với số độ trễ p của phần dư $\hat{u}_t$ (thường được xác định dựa vào xem xét PAC trong giản đồ tự tương quan của phần dư $\hat{u}_t$).

$$\hat{u}_t = \alpha_1 + \alpha_2 X_{2t} + \ldots + \alpha_R X_{Rt} + \alpha_{R+1}\hat{u}_{t-1} + \ldots + \alpha_{R+p}\hat{u}_{t-p} + v_t \quad (7.124)$$

Bước 3: Tính thống kê LM = $(n - p)R^2$ từ phương trình hồi quy (7.124). Thống kê LM này sẽ theo phân phối χ^2 với số bậc tự do là *p*. Nếu $(n - p)R^2 > \chi^2$ tra bảng ở mức ý nghĩa được chọn, ta bác bỏ giả thiết H_0 và kết luận rằng mô hình (7.53) có tự tương quan.

Hạn chế của kiểm định LM của Breusch-Godfrey là việc xác định số độ trễ tối ưu *p*. Thông thường người ta sử dụng các thống kê AIC, SIC, hoặc giản đồ tự tương quan để chọn số độ trễ.

■ BẢNG 7.7: **Kiểm định LM của Breusch-Godfrey.**

F-statistic	17.25931	Prob. F(4,31)	0.0000
Obs*R-squared	26.22439	Prob. Chi-Square(4)	0.0000

Test Equation:
Dependent Variable: RESID

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.483704	0.489336	-0.988491	0.3306
LDISP	0.178048	0.185788	0.958341	0.3453
LPRICE	-0.071428	0.093945	-0.760322	0.4528
RESID(-1)	0.840743	0.176658	4.759155	0.0000
RESID(-2)	-0.340727	0.233486	-1.459306	0.1545
RESID(-3)	0.256762	0.231219	1.110471	0.2753
RESID(-4)	0.196959	0.186608	1.055465	0.2994

R-squared	0.690115	Mean dependent var	-7.48E-16
Adjusted R-squared	0.630138	S.D. dependent var	0.044987
S.E. of regression	0.027359	Akaike info criterion	-4.194685
Sum squared resid	0.023205	Schwarz criterion	-3.893024
Log likelihood	86.69901	Hannan-Quinn criter.	-4.087356
F-statistic	11.50621	Durbin-Watson stat	1.554119

Kiểm định LM của Breusch-Godfrey trên Eviews

Tiếp tục với kết quả hồi quy ở trên, ta thực hiện hai bước sau đây. Bước 1, xác định độ trễ thích hợp của phần dư là 1 (bằng cách sử dụng PAC trên giản đồ tự tương quan: Quick/Series Statistics/Correlogram...). Bước 2, ta chọn **View/Residual Tests/Serial Correlation LM Test**, với các độ trễ lần lượt bằng 4 và 1. Các kết quả kiểm định LM (Bảng 7.7) cho thấy mô hình có hiện tượng tự tương quan, nhưng chỉ là tự tương quan bậc một vì hệ số của RESID(-1) là hệ số duy nhất có ý nghĩa thống kê.

KHẮC PHỤC HIỆN TƯỢNG TỰ TƯƠNG QUAN

Có nhiều cách khắc phục hiện tượng tự tương quan tùy vào việc có sẵn thông tin về ρ hay không.

Khi biết ρ

Giả sử, chúng ta có mô hình sau đây:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u_t \quad t = 1, 2, 3, \ldots, n \qquad (7.53)$$

Trong đó,

$$u_t = \rho u_{t-1} + \varepsilon_t \quad |\rho| < 1 \qquad (7.112)$$

Nếu (7.53) đúng cho giai đoạn *t*, thì nó cũng đúng cho giai đoạn *t -1*, nên:

$$Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + \ldots + \beta_k X_{kt-1} + u_{t-1} \qquad (7.125)$$

Nhân hai vế của (7.125) cho ρ, ta có:

$$\rho Y_{t-1} = \beta_1 \rho + \beta_2 \rho X_{2t-1} + \ldots + \beta_k \rho X_{kt-1} + \rho u_{t-1} \qquad (7.126)$$

Lấy (7.53) trừ (7.126), ta có:

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_{2t} - \rho X_{2t-1}) + \ldots + \beta_k(X_{kt} - \rho X_{kt-1}) + (u_t - \rho u_{t-1}) \qquad (7.127)$$

Hoặc có thể viết lại như sau:

$$Y_t^* = \beta_1^* + \beta_2^* X_{2t}^* + \beta_3^* X_{3t}^* + \ldots + \beta_k^* X_{kt}^* + \varepsilon_t \qquad (7.128)$$

Phương trình (7.128) sẽ được ước lượng theo phương pháp OLS. Và quy trình vừa nêu trên được gọi là thủ tục/phương pháp sai phân tổng quát. Tuy nhiên, với thủ tục này chúng ta bị mất đi một bậc tự do. Để tránh việc mất quan sát như vậy, các nhà kinh tế lượng thực nghiệm đề xuất nên chuyển đổi Y_1 và X_{i1} theo cách sau đây:

$$Y_1^* = Y_1\sqrt{1-\rho^2} \text{ và } X_{i1}^* = X_{i1}\sqrt{1-\rho^2} \qquad (7.129)$$

Thực hiện thủ tục sai phân tổng quát trên Eviews

Để áp dụng phương pháp sai phân tổng quát, trước hết chúng ta cần tìm một giá trị ước lượng của ρ. Một cách phổ biến nhất để có giá trị ước lượng của ρ là chúng ta lưu phần dư từ mô hình hồi quy ở Bước 1 (phương pháp đồ thị), đặt tên phần dư này là *res01* (nếu không tạo biến mới res01, mà vẫn dùng resid, thì Eviews sẽ không chấp nhận). Sau đó, chúng ta hồi quy *res01* theo *res01(-1)* không có hệ số cắt. Ta có kết quả như sau:

■ BẢNG 7.8: **Giá trị ước lượng của ρ.**

Dependent Variable: RES01
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RES01(-1)	**0.799544**	0.100105	7.987073	0.0000

R-squared	0.638443	Mean dependent var	-0.002048
Adjusted R-squared	0.638443	S.D. dependent var	0.043775
S.E. of regression	0.026322	Akaike info criterion	-4.410184
Sum squared resid	0.024942	Schwarz criterion	-4.366646
Log likelihood	82.58841	Hannan-Quinn criter.	-4.394835
Durbin-Watson stat	1.629360		

Sau khi đã có giá trị ước lượng của ρ, chúng ta lần lượt tạo các biến Y_1^*, X_{i1}^*, Y_t^*, β_1^*, X_{2t}^*, X_{3t}^*, ..., và X_{kt}^* như sau:

```
ls lcons c ldisp lprice
genr res01=resid
ls res01 res01(-1)
scalar rho=c(1)
smpl 1985:1 1985:1
genr lcons_star=((1-rho^2)^(0.5))*lcons
genr ldisp_star=((1-rho^2)^(0.5))*ldisp
genr lprice_star=((1-rho^2)^(0.5))*lprice
genr beta1_star=((1-rho^2)^(0.5))

smpl 1985:2 1994:2
genr lcons_star=lcons-rho*lcons(-1)
genr ldisp_star=ldisp-rho*ldisp(-1)
genr lprice_star=lprice-rho*lprice(-1)
```

```
genr beta1_star=1-rho
smpl 1985:1 1994:2
```

Cuối cùng, chúng ta ước lượng mô hình hồi quy với các biến chuyển hóa này (không có hệ số cắt) như sau:

```
ls lcons_star beta1_star ldisp_star lprice_star
```

■ BẢNG 7.9: **Kết quả ước lượng sai phân tổng quát.**

Dependent Variable: LCONS_STAR
Method: Least Squares
Included observations: 38

Variable	Coefficient	Std. Error	t-Statistic	Prob.
BETA1_STAR	4.089403	1.055839	3.873131	0.0004
LDISP_STAR	0.349452	0.231708	1.508155	0.1405
LPRICE_STAR	-0.235900	0.074854	-3.151460	0.0033

R-squared	0.995922	Mean dependent var	1.061130
Adjusted R-squared	0.995689	S.D. dependent var	0.306551
S.E. of regression	0.020129	Akaike info criterion	-4.897684
Sum squared resid	0.014191	Schwarz criterion	-4.768401
Log likelihood	96.05600	Durbin-Watson stat	1.686825

Trường hợp đặc biệt nhất của sai phân tổng quát là phương pháp sai phân bậc một. Trong trường hợp này, phương trình hồi quy cũng không có hệ số cắt. Trên Eviews, chúng ta thực hiện như sau:

```
ls d(lcons) d(ldisp) d(lprice)
```

Khi không biết ρ

Mặc dù phương pháp sai phân tổng quát có vẻ rất dễ áp dụng, nhưng trên thực tế thì chúng ta không biết giá trị của ρ là bao nhiêu. Vì thế, nhiều thủ tục khác đã được phát triển nhằm cung cấp cho chúng ta những giá trị ước lượng của ρ để ước lượng mô hình (7.128). Mặc dù, có rất nhiều thủ tục đã được giới thiệu trong nhiều tài liệu về kinh tế lượng, nhưng phổ biến nhất là hai thủ tục lặp: Thủ tục lặp của Cochrane-Orcutt và Thủ tục tìm kiếm của Hildreth-Lu. Trong phạm vi

cuốn giáo trình này, chúng tôi chỉ trình bày thủ tục lặp của Cochrane-Orcutt.

Thủ tục lặp của Cochrane-Orcutt

Cochrane và Orcutt (1949) phát triển một thủ tục lặp vốn trở nên khá phổ biến trong giới nghiên cứu kinh tế lượng. Thủ tục Cochrane-Orcutt được thực hiện theo các bước sau đây:

Bước 1: Ước lượng phương trình (7.53) theo OLS và lưu phần dư $\hat{u}_t$

Bước 2: Ước lượng hệ số tương quan chuỗi bậc một, $\hat{\rho}$, theo OLS từ phương trình sau đây:

$$\hat{u}_t = \hat{\rho}\hat{u}_{t-1} + \varepsilon_t$$

Bước 3: Chuyển hóa các biến gốc theo cách sau đây: $Y_t^* = Y_t - \hat{\rho}Y_{t-1}$, $\beta_1^* = \beta_1(1-\hat{\rho})$, và $X_{it}^* = Y_{it} - \hat{\rho}Y_{it-1}$ cho các quan sát từ t = 2 trở đi; và $Y_1^* = Y_1\sqrt{1-\hat{\rho}^2}$ và $X_{i1}^* = X_{i1}\sqrt{1-\hat{\rho}^2}$ cho quan sát t = 1.

Bước 4: Hồi quy phương trình (7.128) với các biến chuyển hóa và lưu phần dư của mô hình này. Do chúng ta không biết có phải $\hat{\rho}$ từ Bước 2 là giá trị ước lượng "tốt nhất" của ρ chưa, rồi quay lại bước 2, tiếp tục thực hiện quy trình này từ Bước 2 đến Bước 4 (bước lặp) một số lần cho đến khi nào giá trị ước lượng của ρ ở hai lần lặp liền kề khác nhau rất ít (ví dụ 0.001). Tuy nhiên, nếu chúng ta thực hiện thủ tục lặp này một cách thủ công (ước lượng, tính toán, rồi ước lượng lại, v.v...) sẽ tốn kém rất nhiều thời gian. Chính vì thế, các thủ tục lặp này luôn được lập trình trong hầu hết các phần mềm kinh tế lượng.

Trên thực tế, thủ tục này trở nên rất đơn giản với Eviews. Giả sử lúc đầu ta có phương trình:

```
ls lcons c ldisp lprice
```

Thì thủ tục lặp của Cochrane-Orcutt trên Eviews chỉ đơn giản là ước lượng phương trình sau đây:

ls lcons c ldisp lprice AR(1)

Trong kết quả hồi quy, hệ số ứng với AR(1) chính là giá trị $\hat{\rho}$ tối ưu sau một số bước lặp. Lưu ý rằng, nếu mô hình hồi quy có hiện tượng tự tương quan bậc 2, thì thủ tục lặp của Cochrane – Orcutt trên Eviews sẽ như sau:

ls lcons c ldisp lprice AR(1) AR(2)

Ngoài ra, nếu dữ liệu theo quý, thì chúng ta nên quan tâm đến độ trễ theo quý, và trên Eviews chúng ta sử dụng lệnh sau đây:

ls y c x AR(1) AR(4)

Trở lại ví dụ về chi tiêu tiêu dùng, sau khi thực hiện hồi quy, chúng ta có các kết quả ước lượng theo thủ tục lặp này như sau:

■ BẢNG 7.10: **Kết quả ước lượng theo thủ tục lặp.**

Dependent Variable: LCONS
Method: Least Squares
Date: 06/23/09 Time: 10:59
Sample (adjusted): 1985Q2 1994Q2
Included observations: 37 after adjustments
Convergence achieved after 13 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.762759	1.067582	9.144742	0.0000
LDISP	-0.180461	0.222169	-0.812269	0.4225
LPRICE	-0.850378	0.057714	-14.73431	0.0000
AR(1)	0.974505	0.013289	73.33297	0.0000

R-squared	0.962878	Mean dependent var	4.608665
Adjusted R-squared	0.959503	S.D. dependent var	0.051985
S.E. of regression	0.010461	Akaike info criterion	-6.180445
Sum squared resid	0.003612	Schwarz criterion	-6.006291
Log likelihood	118.3382	Hannan-Quinn criter.	-6.119047
F-statistic	285.3174	Durbin-Watson stat	2.254662
Prob(F-statistic)	0.000000		

Inverted AR Roots	.97

■ BẢNG 7.11: **Kết quả ước lượng theo thủ tục lặp.**

Dependent Variable: LCONS
Method: Least Squares
Date: 06/23/09 Time: 11:01
Sample (adjusted): 1986Q1 1994Q2
Included observations: 34 after adjustments
Convergence achieved after 13 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.21012	0.984907	10.36659	0.0000
LDISP	-0.308138	0.200041	-1.540374	0.1343
LPRICE	-0.820115	0.065877	-12.44914	0.0000
AR(1)	0.797677	0.123851	6.440606	0.0000
AR(4)	0.160974	0.115526	1.393406	0.1741

R-squared	0.967582	Mean dependent var	4.610894
Adjusted R-squared	0.963111	S.D. dependent var	0.053370
S.E. of regression	0.010251	Akaike info criterion	-6.187920
Sum squared resid	0.003047	Schwarz criterion	-5.963455
Log likelihood	110.1946	Hannan-Quinn criter.	-6.111371
F-statistic	216.3924	Durbin-Watson stat	**2.045798**
Prob(F-statistic)	0.000000		

Inverted AR Roots	.97	.16+.55i	.16-.55i	-.50

Trong hai mô hình này, thì giá trị ước lượng của ρ là 0.97. Kết quả cho thấy thống kê DW bây giờ gần bằng 2. Nếu mục đích ước lượng của chúng ta chỉ là xác định hệ số co giãn của chi tiêu tiêu dùng theo thu nhập khả dụng hoặc chỉ số giá lương thực, thì chúng ta không cần quan tâm đến các hệ số của AR(1) và AR(4).

SAI DẠNG MÔ HÌNH

Một trong vấn đề quan trọng nhất trong kinh tế lượng và dự báo theo các mô hình nhân quả là trên thực tế chúng ta không bao giờ biết chắc chắn về dạng đúng nhất của hàm hồi quy mà chúng ta muốn ước lượng. Ba trường hợp hay gặp phải trên thực tế là (1) Bỏ sót biến giải thích quan trọng hoặc thừa biến giải thích không cần thiết, (2) Sử dụng sai dạng hàm, và (3) Sai sót trong việc đo lường. Và đây là vấn

đề sẽ được trình bày một cách ngắn gọn trong phần này. Cũng lưu ý rằng, chúng ta cần đảm bảo dạng mô hình được xác định đúng trước khi sử dụng cho các mục đích dự báo và phân tích chính sách.

BỎ SÓT BIẾN GIẢI THÍCH QUAN TRỌNG HOẶC THỪA BIẾN GIẢI THÍCH KHÔNG CẦN THIẾT

Hậu quả của việc bỏ sót biến giải thích quan trọng

Bỏ sót biến giải thích quan trọng có thể làm cho các biến này trở thành "một bộ phận" của hạng nhiễu ngẫu nhiên trong hàm hồi quy tổng thể. Và như thế sẽ dẫn đến một hoặc một số giả định của mô hình hồi quy tuyến tính cổ điển bị phá vỡ. Để giải thích vấn đề này, chúng ta xem xét mô hình sau đây:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad (7.130)$$

Trong đó, $\beta_2 \neq 0$ và $\beta_3 \neq 0$, và giả sử dạng hàm ở phương trình (7.130) là đúng.

Tuy nhiên, vì một lý do nào đó, mà khi ước lượng chúng ta lại sử dụng mô hình sau đây:

$$Y = \beta_1 + \beta_2 X_2 + u^* \qquad (7.131)$$

Như vậy, X_3 bị bỏ sót một cách sai sót. Trong phương trình này, chúng ta đang "buộc" u^* chứa cả thông tin về X_3 và các yếu tố ngẫu nhiên thuần túy khác. Nghĩa là,

$$u^* = \beta_3 X_3 + u \qquad (7.132)$$

Dựa trên các giả định của mô hình hồi quy tuyến tính cổ điển, thì bây giờ giả định giá trị trung bình của hạng nhiễu không còn phù hợp:

$$E(u^*) = E(\beta_3 X_3 + u) = E(\beta_3 X_3) + E(u) = E(\beta_3 X_3) \neq 0 \qquad (7.133)$$

Hơn nữa, nếu biến X_3 có tương quan với X_2, thì hạng nhiễu (u^*) bây giờ không còn "độc lập" với X_2 nữa. Kết quả của hai vi phạm giả định

này sẽ làm cho các ước lượng của β_1 và β_2 sẽ bị chệch và không nhất quán (không chứng minh).

Hậu quả của việc thừa biến giải thích không cần thiết

So với việc bỏ sót biến giải thích quan trọng, thì nếu một mô hình hồi quy bao gồm các biến giải thích không có ảnh hưởng gì đến biến phụ thuộc, thì vấn đề không quá nghiêm trọng. Giả sử, mô hình đúng sẽ có dạng như sau:

$$Y = \beta_1 + \beta_2 X_2 + u \qquad (7.134)$$

Và bây giờ chúng ta lại ước lượng mô hình sau đây:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad (7.135)$$

Như vậy, X_3 bị đưa vào mô hình một cách dư thừa. Trong trường hợp này, vì X_3 không thuộc phương trình (7.135), nên hệ số hồi quy tổng thể của nó bằng không ($\beta_3 = 0$). Nếu $\beta_3 = 0$, thì không có một giả định nào của mô hình hồi quy tuyến tính cổ điển bị phá vỡ khi ta ước lượng phương trình (7.135) và vì thế các ước lượng OLS vẫn không chệch và nhất quán. Tuy nhiên, việc đưa vào mô hình một biến không cần thiết làm cho các ước lượng OLS của β_1 và β_2 không còn đảm bảo hiệu quả một cách hoàn toàn. Trong trường hợp X_3 có tương quan với X_2, thì chúng ta có thể gặp phải vấn đề đa cộng tuyến. Và hậu quả có thể làm cho sai số chuẩn của $\hat{\beta}_2$ cao hơn, và khả năng chấp nhận giả thiết H_0 cho rằng X_2 không ảnh hưởng lên Y (mặc dù thực chất là có). Chính vì vậy, chúng ta thường 'lo lắng' việc bỏ sót biến giải thích quan trọng hơn là việc đưa thừa biến giải thích không cần thiết khi thực hiện dự báo bằng các mô hình nhân quả.

Trên thực tế, nhiều khi chúng ta vừa bỏ sót biến giải thích quan trọng vừa đưa biến giải thích không cần thiết vào mô hình. Và, dĩ nhiên, hậu quả mà chúng ta gặp phải sẽ là hậu quả của cả hai trường hợp trên.

Ngoài ra, trong nhiều trường hợp chúng ta không có thông tin về một hoặc một số biến giải thích quan trọng (theo lý thuyết kinh tế), nhưng chúng ta không có sẵn thông tin hoặc khó thu thập thông tin một cách chính xác, thì chúng ta nên tìm các biến đại diện khác có thể thu thập được (ví dụ biến giả).

DẠNG HÀM

Một trong những mục đích quan trọng nhất của dự báo là ước lượng các hệ số co giãn hoặc xây dựng các hàm lợi ích/chi phí biên (ví dụ doanh thu biên, chi phí biên). Tuy nhiên, trên thực tế chúng ta thường gặp phải vấn đề chọn lựa sai dạng hàm (nhất là đối với các nhà nghiên cứu hoặc doanh nghiệp Việt Nam luôn khan hiếm nguồn tài liệu tham khảo các nghiên cứu trước đây). Ví dụ, dạng hàm đúng là phi tuyến nhưng ta lại ước lượng dạng hàm tuyến tính. Lỗi sai dạng hàm thường dẫn đến các vấn đề như tự tương quan hoặc phương sai thay đổi. Hơn nữa, nếu chọn sai dạng hàm sẽ dẫn đến khả năng dự báo sai các hệ số co giãn hoặc không xác định đúng đắn dạng hàm lợi ích/chi phí biên, và điều này có thể dẫn đến việc ra quyết định sai lầm. Một cách phát hiện sai dạng hàm là xem xét đồ thị phần dư. Nếu đồ thị phần dư biểu thị một phân tán theo một hệ thống nhất định, thì chúng ta có thể hoài nghi về khả năng sai dạng mô hình. Dưới đây là một số dạng hàm được sử dụng phổ biến trong phân tích kinh tế lượng và dự báo:

■ BẢNG 7.12: **Dạng hàm.**

Tên hàm	Dạng hàm	Ảnh hưởng biên (dY/dX)	Độ co giãn (X/Y)(dY/dX)
Tuyến tính	$Y = \beta_1 + \beta_2 X$	β_2	$\beta_2 X/Y$
Lin-Log	$Y = \beta_1 + \beta_2 \ln X$	β_2/X	β_2/Y
Nghịch đảo	$Y = \beta_1 + \beta_2(1/X)$	$-\beta_2/X^2$	$-\beta_2/(XY)$
Bậc hai	$Y = \beta_1 + \beta_2 X + \beta_3 X^2$	$\beta_2 + 2\beta_3 X$	$(\beta_2 + 2\beta_3 X)X/Y$
Tương tác	$Y = \beta_1 + \beta_2 X + \beta_3 XZ$	$\beta_2 + \beta_3 Z$	$(\beta_2 + \beta_3 Z)X/Y$
Log-Lin	$\ln Y = \beta_1 + \beta_2 X$	$\beta_2 Y$	$\beta_2 X$

Tên hàm	Dạng hàm	Ảnh hưởng biên (dY/dX)	Độ co giãn (X/Y)(dY/dX)
Log-Nghịch đảo	$\ln Y = \beta_1 + \beta_2(1/X)$	$-\beta_2 Y/X^2$	$-\beta_2/X$
Log-Bậc hai	$\ln Y = \beta_1 + \beta_2 X + \beta_3 X^2$	$Y(\beta_2 + 2\beta_3 X)$	$X(\beta_2 + 2\beta_3 X)$
Log kép	$\ln Y = \beta_1 + \beta_2 \ln X$	$\beta_2 Y/X$	β_2
Logistic	$\ln[Y/(1-Y)] = \beta_1 + \beta_2 X$	$\beta_2 Y(1-Y)$	$\beta_2(1-Y)X$

Việc lựa chọn dạng hàm (thông thường dựa trên cơ sở lý thuyết và khảo sát dữ liệu thực tế) đóng một vai trò quan trọng trong việc giải thích các hệ số hồi quy và tránh lỗi sai dạng hàm. Vì thế, chúng ta cần có một cách kiểm định chính thức để hướng dẫn chúng ta nên sử dụng dạng hàm nào cho một trường hợp cụ thể (đặc biệt trong những trường hợp chúng ta không biết chắc chắn về mối quan hệ tổng thể). Nếu các mô hình có biến phụ thuộc giống nhau, thì chúng ta có thể sử dụng tiêu chí R^2. Tuy nhiên, trong nhiều trường hợp chúng ta phải cân nhắc giữa các mô hình có biến phụ thuộc khác nhau, thì phương pháp chuyển hóa Box-Cox (1964) là một lựa chọn tối ưu.

Giả sử, chúng ta phải lựa chọn giữa hai mô hình sau đây:

$$Y_t = \beta_1 + \beta_2 X_t \qquad (7.136)$$

và

$$\ln Y_t = \beta_1 + \beta_2 \ln X_t \qquad (7.137)$$

Bước 1: Tính giá trị trung bình hình học của các giá trị Y_t mẫu:

$$\overline{Y} = (Y_1 Y_2 Y_3 ... Y_n)^{1/n} = \exp(\frac{1}{n}\sum \ln Y_t) \qquad (7.138)$$

Bước 2: Chuyển hóa giá trị Y_t bằng cách chia từng quan sát của Y_t cho $\overline{Y}$, và ta có:

$$Y_t^* = \frac{Y_t}{\overline{Y}} \qquad (7.139)$$

Bước 3: Ước lượng các phương trình (7.136) và (7.137) với Y_t^* được dùng thay cho Y_t. Bây giờ, RSS của hai mô hình có thể được so sánh trực tiếp, và phương trình nào có RSS bé hơn sẽ tốt hơn.

Bước 4: Nếu muốn kiểm định để biết phương trình nào tốt hơn một cách có ý nghĩa thống kê, thì chúng ta phải tính một thống kê kiểm định sau đây:

$$\left(\frac{1}{2}n\right)\ln\left(\frac{RSS_2}{RSS_1}\right) \qquad (7.140)$$

Trong đó, RSS_2 là RSS của phương trình có RSS cao hơn. Thống kê trên sẽ theo phân phối χ^2 với 1 bậc tự do. Nếu giá trị χ^2 tính toán lớn hơn χ^2 tra bảng, thì ta kết luận rằng mô hình với RSS thấp hơn là mô hình có dạng hàm phù hợp hơn một cách có ý nghĩa thống kê.

LỖI ĐO LƯỜNG

Đây là các lỗi liên quan đến việc xác định các biến số trong mô hình và thu thập dữ liệu. Lỗi đo lường có thể xảy ra ở biến phụ thuộc và ở biến giải thích.

Lỗi đo lường ở biến phụ thuộc

Giả sử phương trình đúng của tổng thể có dạng như sau:

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u \qquad (7.141)$$

Phương trình này thỏa mãn tất cả các giả định của mô hình hồi quy tuyến tính cổ điển, nhưng chúng ta không thể quan sát được các giá trị thực của Y. Do không có các thông tin chính xác về Y, nên chúng ta sử dụng các dữ liệu có sẵn của Y vốn có chứa các lỗi đo lường. Cụ thể, các giá trị Y^* quan sát có thể như sau:

$$Y^* = Y + w \qquad (7.142)$$

Trong đó, w thể hiện lỗi trong đo lường.

Như vậy, phương trình (7.141) sẽ được thể hiện như sau:

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + (u + w) \qquad (7.143)$$

Các hệ số OLS chỉ không bị ảnh hưởng chỉ với các điều kiện sau đây được thỏa mãn. Thứ nhất, nếu w có giá trị trung bình bằng không, thì chúng ta sẽ có ước lượng không chệch cho β_1. Ngược lại, nếu giá trị trung bình của w khác không, thì ước lượng OLS của β_1 bị chệch. Tuy nhiên, đây không phải là vấn đề quan trọng trong kinh tế lượng và dự báo. Thứ hai, nếu w không có tương quan gì đến các biến giải thích, thì các ước lượng OLS cho các hệ số độ dốc sẽ không chệch và nhất quán, và ngược lại.

Tuy nhiên, trong trường hợp *u* và *w* không tương quan, thì $\text{var}(u+w) = \sigma_u^2 + \sigma_w^2 > \sigma_u^2$. Như vậy, lỗi đo lường ở biến phụ thuộc có thể làm cho phương sai của phần dư lớn hơn, và vì thế làm cho sai số chuẩn của các hệ số ước lượng lớn hơn.

Lỗi đo lường ở biến giải thích

Giả sử phương trình đúng của tổng thể là:

$$Y = \beta_1 + \beta_2 X_2 + u \qquad (7.144)$$

Thỏa mãn các giả định của mô hình hồi quy tuyến tính cổ điển, nhưng chỉ có điều chúng ta không thể có được thông tin chính xác về X_2. Chẳng hạn, dữ liệu có sẵn về X_2 là:

$$X_2 = X_2^* - v \qquad (7.145)$$

Như vậy, (7.144) sẽ được viết lại như sau:

$$\begin{aligned} Y &= \beta_1 + \beta_2 (X_2^* - v) + u \\ &= \beta_1 + \beta_2 X_2^* + (u - \beta_2 v) \end{aligned} \qquad (7.146)$$

Nếu trường hợp u và v không tương quan với X_2^* và cả hai có giá trị trung bình bằng không, thì các ước lượng OLS vẫn là các ước lượng nhất quán cho cả β_1 và β_2. Do u và v không tương quan nhau, nên phương sai của phần dư là $\text{var}(u-\beta_2 v) = \sigma_u^2 + \beta_2\sigma_v^2 > \sigma_u^2$. Chỉ trường hợp β_2 bằng không thì lỗi đo lường mới không ảnh hưởng đến việc làm tăng phương sai hạng nhiễu, và vì thế không làm tăng sai số chuẩn của các hệ số β_1 và β_2.

KIỂM ĐỊNH SAI DẠNG MÔ HÌNH

Phân phối chuẩn của phần dư

Nếu phần dư không ngẫu nhiên, không có phân phối chuẩn là một thông tin quan trọng cho biến mô hình hồi quy chưa tốt do có thể bị các lỗi như bỏ sót biến quan trọng, sai dạng hàm, phương sai thay đổi, tự tương quan, v.v... Hơn nữa, một giả định quan trọng của mô hình hồi quy tuyến tính cổ điển là các hạng nhiễu (mà phần dư là đại diện trong hàm hồi quy mẫu) có trung bình bằng không và phương sai không đổi. Nếu giả định này không được thỏa mãn, thì các thống kê suy luận của một mô hình hồi quy (như t_{stat}, F_{stat}, v.v...) không có giá trị nữa. Chính vì thế, kiểm định tính chuẩn của phần dư là một công việc có ý nghĩa quan trọng trong phân tích hồi quy và dự báo. Như đã đề cập ở chương 2, kiểm định phần dư có phân phối chuẩn hay không, chúng ta sử dụng thống kê JB của Jarque-Berra (1990). Quy trình kiểm định JB như sau:

Bước 1: Tính các mô men thứ hai (μ_2), thứ ba (μ_3), và thứ tư (μ_4) của phần dư ($\hat{u}$) trong mô hình hồi quy.

$$\mu_2 = \frac{\sum \hat{u}^2}{n}; \quad \mu_3 = \frac{\sum \hat{u}^3}{n}; \quad \mu_4 = \frac{\sum \hat{u}^4}{n} \qquad (7.147)$$

Bước 2: Tính thống kê JB theo công thức sau đây (giống như công thức ở chương 2):

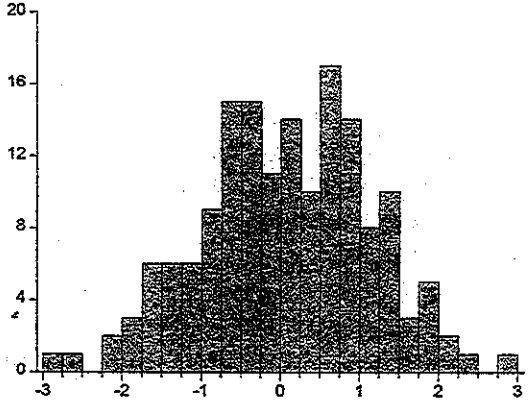
$$JB = n\left[\frac{\mu_3^2}{6} + \frac{(\mu_4 - 3)^2}{24}\right] \qquad (7.148)$$

Có phân phối χ^2 với hai bậc tự do. Lưu ý, ở các công thức trên, đôi khi chúng ta sử dụng (n-k) thay cho *n*. Trong đó, *k* là số hệ số ước lượng trong mô hình hồi quy.

Bước 3: Tìm giá trị χ^2 tra bảng theo hàm =CHIINV(α,2).

Bước 4: Nếu JB > χ^2 tra bảng, chúng ta bác bỏ giả thiết H_0 (phần dư có phân phối chuẩn). Hoặc, nếu giá trị xác suất p < α (có thể 5% hoặc 1%), chúng ta bác bỏ giả thiết H_0.

■ HÌNH 7.7: **Đồ thị tần suất của phần dư.**



Kiểm định JB trên Eviews

Bước 1: Ước lượng hàm hồi quy trên Eviews.

Bước 2: Quick/Series Statistics/Histogram and Stats, chọn RESID, <OK> (Hình 7.7).

Bước 3: Giống bước 3 ở trên.

Bước 4: Giống bước 4 ở trên.

Kiểm định RESET của Ramsey

Một trong những kiểm định phổ biến nhất để kiểm định sai dạng mô hình là kiểm định RESET của Ramsay (1969). Giả sử ta có mô hình 'đúng' của tổng thể như sau:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + u \qquad (7.149)$$

Nhưng khi ước lượng, chúng ta sử dụng mô hình sai (do bỏ sót biến quan trọng) như sau:

$$Y = \beta_1 + \beta_2 X_2 + u^* \qquad (7.150)$$

Kiểm định RESET sẽ dựa vào giá trị Y ước lượng từ phương trình hồi quy sau đây:

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 \qquad (7.151)$$

Kiểm định RESET sẽ đưa thêm một số lũy thừa của $\hat{Y}$ như là các đại diện cho X_2^2 để thể hiện các mối quan hệ phi tuyến có thể có. Trước khi thực hiện kiểm định, chúng ta cần xác định số số hạng sẽ đưa thêm vào mô hình mở rộng. Không có câu trả lời chính thức về số số hạng này, nhưng thông thường người ta đưa số hạng bình phương và lũy thừa ba trong hầu hết các ứng dụng thực tế. Vì thế, phương trình mở rộng sẽ như sau:

$$Y = \beta_1 + \beta_2 X_2 + \delta_1 \hat{Y}^2 + \delta_2 \hat{Y}^3 + \varepsilon \qquad (7.152)$$

Đây là loại kiểm định Wald thông thường (dựa trên thống kê F) cho việc đưa thêm các biến giải thích $\hat{Y}^2$ và $\hat{Y}^3$ vào mô hình. Nếu một hoặc một số hệ số có ý nghĩa thống kê, thì đó là dấu hiệu của việc sai dạng mô hình (tổng quát). Một hạn chế quan trọng của kiểm định RESET là nếu bác bỏ giả thiết cho rằng mô hình ban đầu là mô hình đúng, thì điều này chỉ có ý nghĩa mô hình bị xác định sai chứ không đề xuất các mô hình 'đúng'. Quy trình kiểm định RESET sẽ như sau:

Bước 1: Ước lượng phương trình mà ta cho rằng đúng, rồi lưu giá trị Y ước lượng ($\hat{Y}$).

Bước 2: Ước lượng lại mô hình ở bước 1, lần này đưa thêm các biến $\hat{Y}^2$ và $\hat{Y}^3$ vào mô hình.

Bước 3: Mô hình ở bước 1 là mô hình ràng buộc và mô hình ở bước 2 là mô hình không ràng buộc. Tính thống kê *F* cho hai mô hình này (Wald).

Bước 4: Tìm giá trị F tra bảng với số bậc tự do lần lượt là 2, n – k – 3 (với *k* là số biến giải thích ở mô hình bước 1).

Bước 5: Nếu *F* tính toán > *F* tra bảng thì chúng ta bác bỏ giả thiết H_0 (mô hình ở bước 1 là mô hình đúng). Hoặc giá trị xác suất *p* của F_{stat} nhỏ hơn mức ý nghĩa yêu cầu (α), ta bác bỏ H_0. Lưu ý, nếu sử dụng kiểm định *LM* thì ta so sánh với χ^2 với số bậc tự do bằng 2.

Kiểm định RESET trên Eviews

Sử dụng tập tin **DATA7-5**, ta ước lượng mô hình sau đây:

```
ls lcons c ldisp
```

Từ kết quả hồi quy, ta chọn View/Stability Tests/Ramsey RESET Test ..., sau đó nhập số số hạng đưa thêm vào ô <RESET Specification>. Kết quả như sau (bác bỏ giả thiết H_0):

■ BẢNG 7.13: **Kiểm định RESET.**

Ramsey RESET Test

F-statistic	21.75213	Prob. F(1,35)	0.0000
Log likelihood ratio	18.36711	Prob. Chi-Square(1)	0.0000

Test Equation:
Dependent Variable: LCONS

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-204.0132	44.32787	-4.602368	0.0001
LDISP	-204.4011	43.91501	-4.654469	0.0000
FITTED^2	53.74839	11.52430	4.663917	0.0000

R-squared	0.525270	Mean dependent var	4.609274
Adjusted R-squared	0.498142	S.D. dependent var	0.051415
S.E. of regression	0.036423	Akaike info criterion	-3.711559
Sum squared resid	0.046433	Schwarz criterion	-3.582275
Log likelihood	73.51961	Hannan-Quinn criter.	-3.665561
F-statistic	19.36302	Durbin-Watson stat	0.795597

Tóm lại, trước khi kiểm định một hệ số hồi quy nào đó có ý nghĩa thống kê hay không (để sử dụng cho mục đích dự báo), chúng ta cần thực hiện tất cả các kiểm định chuẩn đoán để đảm bảo đó là một mô hình tốt nhất.

BIẾN GIẢ

Trong giáo trình này chúng tôi không trình bày một cách chi tiết về bản chất, phân loại, và cách giải thích biến giả trong kinh tế lượng. Tuy nhiên, ngoài tầm quan trọng của biến trong việc giúp tránh rủi ro do bỏ sót biến giải thích, và thực hiện nhiều kiểm định giả thiết nghiên cứu, biến giả đặc biệt cần thiết đối với người nghiên cứu khi sử dụng các mô hình dự báo nhân quả với dữ liệu thời gian hoặc các mô hình dự báo chuỗi thời gian (ARIMA, ARCH). Chính vì thế, trong phần này chúng tôi sẽ giới thiệu một số vấn đề cơ bản như cách tạo các loại biến giả trên Eviews và một số ứng dụng thực tiễn trong dự báo.

TẠO BIẾN GIẢ THEO NHÓM

Trong kinh tế lượng, thỉnh thoảng chúng ta muốn tạo thêm và đưa vào mô hình các biến giả về một tính chất nào đó, ví dụ trình độ học vấn, nhóm tuổi, nhóm thu nhập, v.v..., từ một biến định lượng. Ví dụ, sử dụng tập tin **DATA7-3** ta tạo các biến giả sao đây. Lưu ý, biến "EDUC" có giá trị từ 0 đến 18. Bây giờ, ta tạo 4 biến giả với các thuộc tính như sau:

- D_1 = 1 nếu người lao động chưa tốt nghiệp cấp 3 (EDUC < 12)

  genr D1=(EDUC<12)

- D_2 = 1 nếu người lao động tốt nghiệp cấp 3 (EDUC = 12)

  genr D2=(EDUC<12)

- D_3 = 1 nếu người lao động tốt nghiệp đại học và cao đẳng (12 < EDUC < 16)

  genr D3= (EDUC<17)-(EDUC<13)

- D_4 = 1 nếu người lao động tốt nghiệp sau đại học (16 < EDUC)

  genr D3= (EDUC>16)

Lưu ý, ngoài lệnh "genr" ta có thể sử dụng lệnh "series".

TẠO BIẾN GIẢ THEO QUÝ/THÁNG

Trong phân tích các chuỗi thời gian theo tháng hoặc theo quý, chúng ta thường muốn kiểm định xem các chỉ báo kinh tế có khác nhau giữa các tháng hoặc quý hay không. Để làm như vậy, chúng ta cần phải tạo ra các biến giả theo tháng và theo quý.

Sử dụng tập tin **DATA7-5**, ta lần lượt tạo biến giả theo bốn quý với các lệnh sau đây:

- D_1 = 1, nếu là quý I

  genr D1=@quarter=1

- $D_2 = 1$, nếu là quý II

  genr D2=@quarter=2

- $D_3 = 1$, nếu là quý III

  genr D3=@quarter=3

- $D_4 = 1$, nếu là quý IV

  genr D4=@quarter=4

Tương tự, nếu dữ liệu theo tháng, thì thay vì dùng hàm @quarter, ta dùng hàm @month. Tuy nhiên, trên thực tế chúng ta không nên tạo ra quá nhiều biến như vậy vì sẽ phức tạp cho việc quản lý dữ liệu trong tập tin Eviews. Để đơn giản, chúng ta có thể sử dụng một cách trực tiếp các lệnh sau đây (trong mô hình hồi quy):

ls lcons ldisp lprice @expand(@quarter)

Nhiều nghiên cứu chuỗi thời gian muốn kiểm định xem dữ liệu có tính ổn định cấu trúc hay không để quyết định có nên sử dụng các mô hình khác nhau cho các giai đoạn khác nhau hay không. Để làm điều này, người ta thường sử dụng biến giả hơn là sử dụng kiểm định Chow. Như vậy, biến giả sẽ được tạo như thế nào? Ví dụ, chúng ta muốn kiểm tra xem giữa giai đoạn 1985Q1-1990Q4 có khác giai đoạn 1991Q1-1994Q2, ta tạo biến giả như sau:

smpl @first 1990:4 (hoặc 1985:1 1990:4)

genr dum=0

smpl 1991:1 @last

genr dum=1

ẢNH HƯỞNG THÁNG GIÊNG TRÊN THỊ TRƯỜNG CHỨNG KHOÁN VIỆT NAM

Kiểm định ảnh hưởng tháng Giêng là một nội dung rất được giới đầu tư tài chính quan tâm cho các chiến lược đầu tư của mình, nhất là đối

với các thị trường kém hiệu quả như Việt Nam. Để kiểm định ảnh hưởng tháng Giêng, hầu hết các nghiên cứu trước đây đều sử dụng kỹ thuật biến giả mùa vụ. Để làm như vậy, trước hết chúng ta nên tạo ra 12 biến giả đại diện cho các tháng trong năm.

D_{it} = 1 nếu suất sinh lợi tương ứng với tháng i trong năm

D_{it} = 0 nếu khác

Theo kinh tế tài chính, thì mô hình tổng thể cho kiểm định ảnh hưởng tháng Giêng được cho như sau:

$$R_{it} = \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + ... + \beta_{12} D_{12t} + u_t \qquad (7.153)$$

Trong đó, R_{it} (=ln(P_t/P_{t-1}) là suất sinh lợi của thị trường tại thời điểm *t*, β_i là suất sinh lợi trung bình của tháng *i*. Giả thiết H_0 cần kiểm định là tất cả các hệ số β_i đều bằng nhau. Nếu các hệ số này bằng nhau, sẽ không có ảnh hưởng mùa vụ, và ngược lại.

Vì thế, để kiểm định ảnh hưởng tháng Giêng, chúng ta thường điều chỉnh mô hình (7.153) theo cách như sau:

$$R_{it} = \beta_0 + \beta_2 D_{2t} + \beta_3 D_{3t} + ... + \beta_{12} D_{12t} + u_t \qquad (7.153)$$

Ở đây, β_0 thể hiện suất sinh lợi trung bình của tháng Giêng, và trong trường hợp này, các hệ số β còn lại thể hiện sự chênh lệch của suất sinh lợi ở tháng Giêng và các tháng khác trong năm. Giả thiết H_0 bây giờ sẽ là tất cả các hệ số hồi quy của các biến giả đều bằng không. Lưu ý, dấu âm của các hệ số hồi quy của các biến giả có thể là một dấu hiệu quan trọng cho biết có ảnh hưởng tháng Giêng.

Sử dụng tập tin **DATA7-6** (giai đoạn tháng 7/2000-6/2009), chúng ta nhận thấy có tồn tại ảnh hưởng tháng Giêng trên thị trường chứng khoán Việt Nam.

ls log(VNI/VNI(-1)) @expand(@month)

Dependent Variable: LOG(VNI/VNI(-1))
Included observations: 2060 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
@MONTH=1	**0.002658**	0.001431	**1.857248**	0.0634
@MONTH=2	-0.000270	0.001510	-0.178508	0.8583
@MONTH=3	0.000922	0.001304	0.707060	0.4796
@MONTH=4	0.002207	0.001363	1.619013	0.1056
@MONTH=5	0.001350	0.001352	0.999162	0.3178
@MONTH=6	0.001593	0.001371	1.161466	0.2456
@MONTH=7	-0.002754	0.001379	-1.996377	0.0460
@MONTH=8	-6.17E-05	0.001329	-0.046429	0.9630
@MONTH=9	0.000142	0.001400	0.101696	0.9190
@MONTH=10	-0.000324	0.001325	-0.244389	0.8070
@MONTH=11	**0.002822**	0.001355	**2.082136**	0.0375
@MONTH=12	0.001139	0.001336	0.852094	0.3943

```
genr d2=@month=2
genr d3=@month=3
genr d4=@month=4
...
genr d9=@month=9
genr d10=@month=10
genr d11=@month=11
genr d12=@month=12
```

ls log(vni/vni(-1)) c d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12

Dependent Variable: LOG(VNI/VNI(-1))
Method: Least Squares
Date: 06/24/09 Time: 12:26
Sample (adjusted): 7/31/2000 6/11/2009
Included observations: 2060 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	**0.002658**	0.001431	**1.857248**	0.0634
D2	-0.002927	0.002080	-1.407099	0.1595
D3	-0.001736	0.001936	-0.896373	0.3702
D4	-0.000451	0.001976	-0.228015	0.8197
D5	-0.001307	0.001968	-0.664156	0.5067
D6	-0.001065	0.001982	-0.537445	0.5910
D7	-0.005411	0.001987	-2.722651	0.0065
D8	-0.002719	0.001953	-1.392432	0.1639
D9	-0.002515	0.002002	-1.256393	0.2091
D10	-0.002982	0.001951	-1.528631	0.1265
D11	0.000164	0.001971	0.083421	0.9335
D12	-0.001519	0.001958	-0.775750	0.4380

R-squared	0.006813	Mean dependent var	0.000792
Adjusted R-squared	0.001478	S.D. dependent var	0.017944
S.E. of regression	0.017930	Akaike info criterion	-5.198833
Sum squared resid	0.658429	Schwarz criterion	-5.166034
Log likelihood	5366.798	Hannan-Quinn criter.	-5.186808
F-statistic	1.277144	Durbin-Watson stat	1.294074

Trong kết quả hồi quy trên thì chúng ta có thể kết luận rằng suất sinh lợi trên thị trường rất có ý nghĩa đối với tháng Giêng do hệ số cắt C có ý nghĩa thống kê ở mức xấp xỉ 6% và dương, các hệ số hồi quy của các tháng khác thì hầu hết âm và ít có ý nghĩa thống kê.

HỆ SỐ HỒI QUY CHUẨN HÓA VÀ DỰ BÁO

Trong phân tích chính sách và dự báo, thỉnh thoảng người sử dụng kết quả nghiên cứu muốn biết trong số các biến giải thích được chọn, thì những biến nào có ảnh hưởng nhiều đến biến phụ thuộc. Điều này có ý nghĩa quan trọng vì dựa trên cơ sở xếp thứ tự ưu tiên, người sử dụng sẽ có các chiến lược thích hợp. Để làm như vậy, chúng ta không thể dựa vào các hệ số hồi quy riêng theo cách ước lượng thông thường, mà phải dựa vào các hệ số hồi quy chuẩn hóa. Theo Pindyck &

Rubinfeld (1998), thì các hệ số hồi quy chuẩn hóa cho biết tầm quan trọng tương đối của các biến giải thích trong một mô hình hồi quy. Để ước lượng các hệ số hồi quy chuẩn hóa, chúng ta cần phải chuyển hóa mỗi biến (cả biến phụ thuộc) sang dạng biến chuẩn hóa, rồi sử dụng phương pháp ước lượng OLS thông thường. Như vậy, mô hình (7.53) sẽ được chuyển hóa như sau:

$$\frac{Y_t - \overline{Y}}{s_Y} = \beta_2^* \frac{X_{2t} - \overline{X}_2}{s_{X_2}} + \beta_3^* \frac{X_{3t} - \overline{X}_3}{s_{X3}} + ... + \beta_k^* \frac{X_{kt} - \overline{X}_k}{s_{Xk}} + u_t \quad (7.154)$$

Các hệ số hồi quy chuẩn hóa có mối quan hệ rất gần với các hệ số hồi quy riêng. Cụ thể,

$$\beta_j^* = \beta_j \frac{s_{Xj}}{s_Y} \quad (7.155)$$

Như vậy, nếu một hệ số hồi quy chuẩn hóa ($\hat{\beta}_j^*$) với giá trị bằng 0.7 nói lên rằng một sự thay đổi bằng 1 độ lệch chuẩn của biến giải thích X_j sẽ dẫn đến một sự thay đổi 0.7 độ lệch chuẩn trong biến phụ thuộc.

Để chuyển từ các hệ số hồi quy riêng sang các hệ số hồi quy chuẩn hóa trên Eviews, ta thực hiện như sau:

Bước 1: Ước lượng mô hình hồi quy theo OLS (giả sử đó là mô hình tốt nhất)

Bước 2: Tính các độ lệch chuẩn của biến phụ thuộc và biến giải thích theo hàm sau đây:

scalar sy=scalar(Y)

scalar sx=scalar(X)

Bước 3: Tính hệ số hồi quy chuẩn hóa theo công thức (7.155)

ỨNG DỤNG DỰ BÁO

Theo Gujarati (2003) có hai loại dự báo: (1) Dự báo giá trị trung bình có điều kiện của Y theo một giá trị X cho trước, ví dụ X_0; nghĩa là một điểm trên đường hồi quy tổng thể. Loại dự báo này được gọi là dự báo trung bình, và (2) Dự báo một giá trị Y cá biệt nào đó theo X_0; nghĩa là xung quanh giá trị E(Y) có thể có rất nhiều giá trị Y. Loại dự báo này được gọi là dự báo cá biệt. Và, sau cùng căn cứ vào dạng hàm khác nhau đã trình bày ở bảng 7.2, chúng ta có thể dự báo tác động biên của một biến độc lập lên biến phụ thuộc hoặc dự báo độ co giãn của biến phụ thuộc theo một biến độc lập.

DỰ BÁO TRUNG BÌNH

Giả sử, ta có phương trình hồi quy sau đây:

$$E(Y_t) = \beta_1 + \beta_2 X_t \tag{7.156}$$

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_t \tag{7.157}$$

$$= 54.8 - 2.9X_t$$

Giả sử ta có $X = X_0 = 12$ và ta muốn dự đoán giá trị trung bình của tổng thể tại $X = X_0$ sẽ là bao nhiêu, nghĩa là $E(Y|X_0=12)$. Kết quả ước lượng từ phương trình (7.157) cho thấy giá trị ước lượng điểm của dự đoán trung bình này là $\hat{Y}_0$ như sau:

$$\hat{Y}_0 = 54.8 - 2.9 *(12)$$

$$= 19.89$$

Trong đó: $\hat{Y}_0$ là ước lượng của $E(Y|X=X_0)$. Vì $\hat{Y}_0$ là một ước lượng, nên $\hat{Y}_0$ có thể khác giá trị thực của nó trên đường hồi quy tổng thể. Chênh lệch giữa hai giá trị này là sai số dự báo. Để đánh giá sai số dự báo này, chúng ta cần tìm phân phối mẫu của $\hat{Y}_0$. Cho $X_i = X_0$, giá trị dự đoán trung bình thực $E(Y_0|X_0)$ như sau:

$$E(Y_0|X_0) = \beta_1 + \beta_2 X_0 \tag{7.158}$$

Ta ước lượng (7.158) từ:

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_t \qquad (7.157)$$

Vậy giá trị kỳ vọng của $\hat{Y}_0$ khi $X_i = X_0$ sẽ là:

$$E(\hat{Y}_0) = E(\hat{\beta}_1) + X_0 E(\hat{\beta}_2)$$
$$= \beta_1 + \beta_2 X_0 \qquad (7.159)$$

Bởi vì $\hat{\beta}_1$ và $\hat{\beta}_2$ là các ước lượng không chệch nên $\hat{Y}_0$ là một ước lượng không chệch của $E(Y_0 | X_0)$.

Khoảng tin cậy cho giá trị dự báo trung bình

Ta có:

$$Var(\hat{Y}_0) = E\left(\hat{Y}_0 - E(\hat{Y}_0)\right)^2$$
$$= E[\hat{\beta}_1 + \hat{\beta}_2 X_0 - \beta_1 - \beta_2 X_0]^2$$
$$= E[(\hat{\beta}_1 - \beta_1) + X_0(\hat{\beta}_2 - \beta_2)]^2$$
$$= E[(\hat{\beta}_1 - \beta_1)^2 + X_0^2(\hat{\beta}_2 - \beta_2)^2 + 2X_0(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)]^2$$
$$= Var(\hat{\beta}_1) + X_0^2 Var(\hat{\beta}_2) + 2X_0 Cov(\hat{\beta}_1, \hat{\beta}_2) \qquad (7.160)$$

Trong đó:

$$var(\hat{\beta}_1) = \frac{\sum X_t^2}{n\sum x_t^2}\sigma$$

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_t^2}$$

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = E\{[\hat{\beta}_1 - E(\hat{\beta}_1)][\hat{\beta}_2 - E(\hat{\beta}_2)]\}$$
$$= E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)$$

(Do $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$ và $E(\hat{\beta}_1) = \overline{Y} - \beta_2 \overline{X}$ nên

$$\hat{\beta}_1 - E(\hat{\beta}_1) = -\overline{X}(\hat{\beta}_2 - \beta_2))$$
$$= -\overline{X}(\hat{\beta}_2 - \beta_2)^2$$
$$= -\overline{X}\,var(\hat{\beta}_2) \qquad (7.161)$$

Từ (7.160) và (7.161) ta có:

$$Var(\hat{Y}_0) = \sigma^2_{\hat{Y}_0} = \sigma^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum(X - \overline{X})^2}\right] \qquad (7.162)$$

Bằng cách thay σ^2 bằng $\hat{\sigma}^2$ ta có:

$$\hat{\sigma}^2_{\hat{Y}_0} = \hat{\sigma}^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum(X - \overline{X})^2}\right] \qquad (7.163)$$

$$t = \frac{\hat{Y}_0 - E(\hat{Y}_0)}{se(\hat{Y}_0)} = \frac{\hat{\beta}_1 + \hat{\beta}_2 X_0 - (\beta_1 + \beta_2 X_0)}{se(\hat{Y}_0)} \qquad (7.164)$$

theo phân phối *t* với n- 2 bậc tự do. Vì thế phân phối *t* có thể được sử dụng để suy ra các khoảng tin cậy cho giá trị kỳ vọng thực $E(Y_0|X_0)$.

$$Pr[\hat{Y}_0 - t_{\alpha/2}se(\hat{Y}_0) \le Y_0 \le \hat{Y}_0 + t_{\alpha/2}se(\hat{Y}_0)] = 0.95 \qquad (7.165)$$

Khi X_0 càng xa giá trị trung bình thì sai số dự báo càng lớn và khoảng tin cậy càng rộng. Điều này có nghĩa nếu dự báo được thực hiện quá xa phạm vi của mẫu, độ tin cậy của dự báo sẽ giảm. Nếu $X_0 = \overline{X}$, khoảng tin cậy sẽ hẹp nhất. Với $X_0 = 12$, $n = 10$, $\overline{X} = 5.5$, $\sum(X - \overline{X})^2 = 82.5$, thì:

$$Var(\hat{Y}_0) = 5.18*\left[\frac{1}{10} + \frac{(12_0 - \overline{5.5})^2}{82.5}\right] = 3.17, \text{ và } se(\hat{Y}_0) = 1.78$$

Vậy khoảng tin cậy 95% của giá trị $E(Y/X_0$ được tính như sau:

$$19.89 - 2.306*1.78 \le E(Y_0/X=12) \le 19.89 + 2.306*1.78$$
$$15.78 \le E(Y_0/X=12) \le 23.99$$

DỰ BÁO CÁ BIỆT

Nếu ta muốn dự báo giá trị Y cá biệt, ví dụ Y_0, tương ứng với một giá trị X cho trước, ví dụ X_0, thì Y_0 được xác định như sau:

$$Y_0 = \beta_1 + \beta_2 X_0 + u_0 \tag{7.166}$$

Ta dự đoán Y_0 khi

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 \tag{7.157}$$

Sai số dự báo, $Y_0 - \hat{Y}_0$ được xác định như sau:

$$\begin{aligned} Y_0 - \hat{Y}_0 &= \beta_1 + \beta_2 X_0 + u_0 - \hat{\beta}_1 - \hat{\beta}_2 X_0 \\ &= (\beta_1 - \hat{\beta}_1) + (\beta_2 - \hat{\beta}_2) X_0 + u_0 \end{aligned} \tag{7.167}$$

Vì thế, với $\hat{\beta}_1$ và $\hat{\beta}_2$ là các ước lượng không chệch, ta có

$$E(Y_0 - \hat{Y}_0) = E(\beta_1 - \hat{\beta}_1) + E(\beta_2 - \hat{\beta}_2) X_0 + E(u_0) = 0 \tag{7.168}$$

Khoảng tin cậy cho giá trị dự báo cá biệt

Lấy bình phương hai vế của phương trình (7.168) ta có:

$$E(Y_0 - \hat{Y}_0)^2 = [E(\beta_1 - \hat{\beta}_1) + E(\beta_2 - \hat{\beta}_2) X_0 + E(u_0)]^2 \tag{7.169}$$

Đặt $f = Y_0 - \hat{Y}_0$, ta có:

$$\begin{aligned} \mathrm{Var}(f) = \mathrm{Var}(\hat{\beta}_1) + X_0^2 \mathrm{Var}(\hat{\beta}_2) + 2X_0 \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) + \\ + \mathrm{Var}(u_0) \end{aligned} \tag{7.170}$$

$$\mathrm{Var}(f) = \sigma_f^2 = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X - \overline{X})^2} \right] > \mathrm{Var}(\hat{Y}_0) \tag{7.171}$$

$$\hat{\sigma}_f^2 = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X - \overline{X})^2} \right] \tag{7.172}$$

$$\hat{\sigma}_f^2 = \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{n} + (X_0 - \overline{X})^2 \hat{\sigma}_{\hat{\beta}2}^2 \tag{7.173}$$

Ta có:

$$t = \frac{Y_0 - \hat{Y}_0}{se(Y_0 - \hat{Y}_0)} \qquad (7.174)$$

cũng theo phân phối *t* với bậc tự do d.f. = n-2. Vì thế phân phối *t* có thể được sử dụng để rút ra các suy luận thống kê về giá trị thực Y_0. Với $X_0 = 12$, ta có khoảng tin cậy cho giá trị Y_0 tại $X_0 = 12$ như sau:

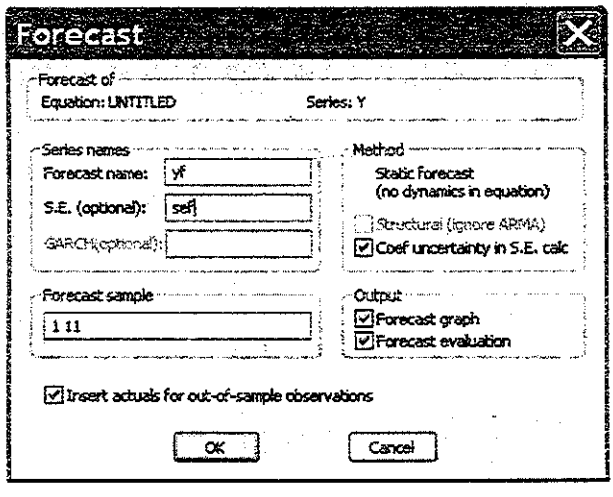
Với $X_0 = 12$, thì $Var(\hat{Y}_0) = 5.18*\left[1+\frac{1}{10}+\frac{(12_0-\overline{5.5})^2}{82.5}\right] = 8.35$, và $se(\hat{Y}_0) = 2.89$. Vậy khoảng tin cậy 95% của giá trị Y_0 được tính như sau:

$$19.89 - 2.306*2.89 \le Y_0 \le 19.89 + 2.306*2.89$$

$$13.23 \le E(Y_0/X=12) \le 26.55$$

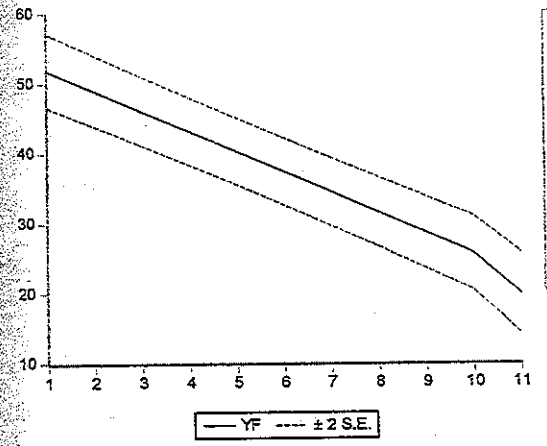
Các khoảng tin cậy của hai loại dự báo này được minh họa ở Hình 7.8, trong đó, khoảng tin cậy của dự báo cá biệt rộng hơn khoảng tin cậy của dự báo trung bình.

Thao tác thực hiện dự báo trên Eviews



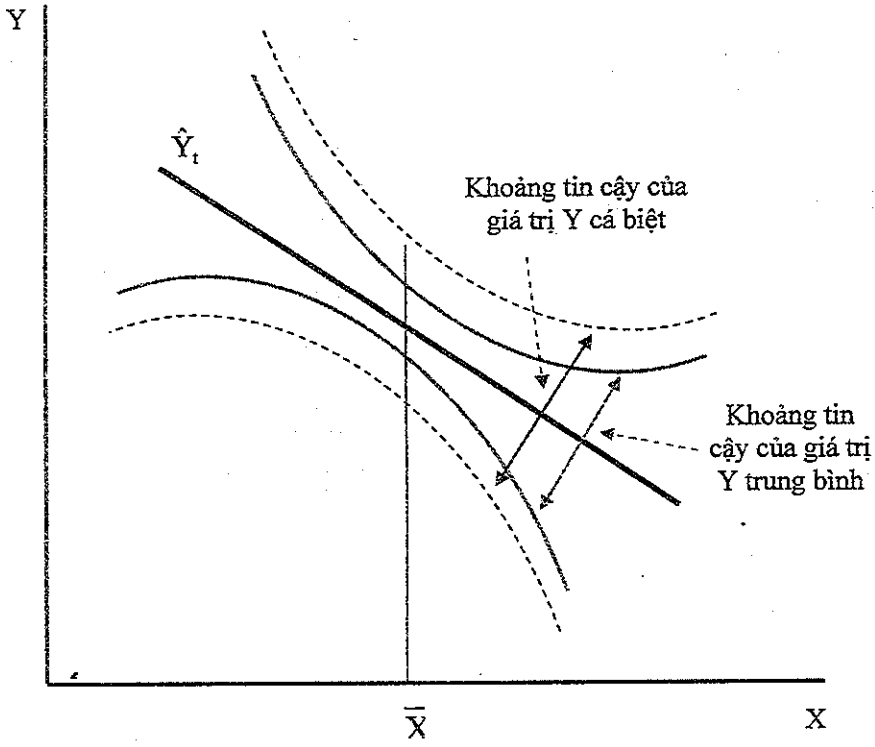
- Từ đối tượng tập tin Eviews, nhấp đúp vào "Range": Thay đổi số quan sát từ 10 lên 11.
- Nhấp đúp vào biến X, chọn Edit+/-: Nhập giá trị mới X = 12 vào quan sát thứ 11. Chọn Edit+/- để kết thúc biên tập.
- Quay lại cửa sổ kết quả ước lượng, nhấp vào "forecast".

■ H ÌNH 7.8: **Kết quả dự báo trên Eviews.**



Forecast: YF	
Actual: Y	
Forecast sample: 1 11	
Included observations: 10	
Root Mean Squared Error	2.035146
Mean Absolute Error	1.836364
Mean Abs. Percent Error	5.017189
Theil Inequality Coefficient	0.025621
Bias Proportion	0.000000
Variance Proportion	0.014406
Covariance Proportion	0.985594

HÌNH 7.8: **Khoảng tin cậy của dự báo.**



TÓM TẮT CHƯƠNG 7

Dựa trên lý thuyết kinh tế - quản trị, nghiên cứu trước, hoặc kinh nghiệm đã tồn tại thì chúng ta có thể mô hình hóa mối quan hệ của một biến mục tiêu (biến phụ thuộc) theo các nhân tố ảnh hưởng (biến độc lập hay biến giải thích). Khi có dữ liệu mẫu về các biến này thì các nhà nghiên cứu dự báo dễ dàng tiến hành hồi quy nhằm ước lượng các mối quan hệ thông qua sự hỗ trợ của phần mềm Eviews hoặc các phần mềm khác. Trước khi tiến hành các ứng dụng mang tính dự báo thì các nhà nghiên cứu cần tiến hành một số kiểm định cần thiết như đa cộng tuyến, dạng hàm, tự tương quan, phân phối chuẩn của phần dư hồi quy nhằm đảm bảo các hệ số hồi quy có tính chất BLUE. Ứng dụng thông thường từ mô hình hồi quy bội cho dự báo thường là dự báo biến mục tiêu thông qua các nhân tố ảnh hưởng khi chúng ta có dữ liệu về các nhân tố ảnh hưởng này trong tương lai. Mô hình hồi quy cũng tạo cơ sở dự báo các quá trình ra quyết định tác động hay không tác động của một nhân tố nào đó lên biến mục tiêu thông qua kiểm định *t* về giả thiết về hệ số hồi quy. Thêm vào đó, mô hình hồi quy còn có thể dự báo sự biến động của biến mục tiêu có bị tác động bởi yếu tố mùa hay không khi sử dụng công cụ biến giả. Hơn nữa, với các dạng hàm khác nhau chúng ta có thể dự báo mức độ tác động định lượng của một nhân tố lên biến mục tiêu và dự báo độ co giãn của biến mục tiêu với một nhân tố ảnh hưởng nào đó. Sau cùng là, việc trình bày các ước lượng, giả định và các kiểm định phương pháp OLS của hồi quy CLRM thông qua những trình bày chi tiết ở chương này không có mục đích giới thiệu môn học kinh tế lượng, mà chúng tôi chỉ muốn rằng các sinh viên đại học hoặc các nhà nghiên cứu hiểu rõ bản chất tính quan trọng của việc hình thành một mô hình hồi quy tốt nhằm phục vụ cho dự báo phải bắt đầu từ đâu, và điều này quả là không đơn giản.

Do chỉ gói gọn nội dung phân tích hồi quy trong chương 7, nên chúng tôi không thể đưa vào nhiều ví dụ ứng dụng thực tế để ta cảm nhận được tầm quan trọng của kinh tế lượng trong dự báo. Chính vì vậy, chúng ta nên tham khảo ở các sách kinh tế lượng ứng dụng khác.

CÂU HỎI VÀ BÀI TẬP

1. Một nghiên cứu mối quan hệ giữa đầu tư (I) và lãi suất (R) dựa trên dữ liệu theo năm được cho trong tập tin **"INVESTMENT.xls"**.

   a. Thực hiện hồi quy trên Eviews. Anh/Chị cho biết mối quan hệ giữa I và R có ý nghĩa thống kê ở mức ý nghĩa 5% hay không? Tại sao?

   b. Anh/Chị cho biết ý nghĩa của hệ số xác định r^2 trong kết quả hồi quy trên Eviews nói lên điều gì?

   c. Giả sử đây là một mô hình tốt, Anh/Chị cho biết nếu ba năm sau lãi suất là 4% thì lượng đầu tư sẽ là bao nhiêu? Anh/Chị cho biết với thời gian dự báo xa như vậy, liệu kết quả dự báo này có còn ý nghĩa hay không? Tại sao?

   d. Anh/Chị cho biết, với lãi suất là 4%, thì lượng đầu tư dự báo sẽ nằm trong khoảng nào với mức ý nghĩa là 5%?

2. Dữ liệu trong tập tin **"VALUATION.xls"** được thu thập cho một nghiên cứu định giá bất động sản. Dữ liệu này thu thập từ 30 căn nhà khác nhau trong một khu phố nhằm hỗ trợ công ty trong việc ước lượng giá trị thị trường của một căn nhà bất kỳ nằm trong khu phố này từ báo cáo thẩm định của các chuyên viên thẩm định bất động sản. Trong đó, Y là giá thị trường (ngàn đôla) và X là giá trị thẩm định (ngàn đôla).

   a. Anh/Chị hãy vẽ đồ thị phân tán (có đường hồi quy) giữa Y và X? Nhận xét?

   b. Trên Eviews, Anh/Chị hãy hồi quy hàm tuyến tính của Y theo X và giải thích kết quả?

   c. Giả sử có một căn nhà được định giá là 90.5 ngàn đôla, thì Anh/Chị cho biết giá thị trường của căn nhà này sẽ là bao nhiêu? Anh/Chị cho biết cách dự đoán như vậy có rủi ro gì không?

   d. Anh/Chị hãy khảo sát phần dư và cho biết nhận xét của Anh/Chị về kết quả ước lượng này?

3. Ông Khang, giám đốc nhân sự của công ty SONKIM đang quan tâm đến việc dự báo xem liệu một ứng viên cụ thể nộp đơn vào công ty có thể trở thành một nhân viên bán hàng giỏi hay không. Để làm điều này, ông Khang quyết định sử dụng doanh số bán hàng của tháng đầu tiên làm biến phụ thuộc (Y) và hồi quy theo các biến giải thích sau đây (tập tin **"SONKIM.xls"**):

   - X_2 = Điểm về kỹ năng bán hàng (/100)
   - X_3 = Tuổi
   - X_4 = Mức độ hăng hái trong công việc (/10)
   - X_5 = Số năm kinh nghiệm
   - X_6 = Điểm tốt nghiệp PTTH

   a. Anh/Chị hãy lập bảng so sánh giá trị trung bình của Y theo biến X_2? Anh/Chị rút ra nhận xét gì về mối quan hệ này?

   b. Anh/Chị hãy xác định mô hình dự báo phù hợp với dữ liệu trên? Giải thích phương pháp lựa chọn mô hình của Anh/Chị?

   c. Anh/Chị cho biết dấu của các hệ số hồi quy có đúng với kỳ vọng của mình hay không? Tại sao?

   d. Ước lượng mô hình dự báo Anh/Chị đã chọn trên Eviews, kiểm định giả thiết đồng thời và kiểm định ý nghĩa của các hệ số hồi quy riêng?

   e. Từ kết quả nghiên cứu này, Anh/Chị hãy tư vấn cho ông Khang biết nên sử dụng tiêu các chí để tuyển mộ nhân viên bán hàng trong tương lai?

4. Để nâng cao hiệu quả việc tư vấn lập kế hoạch ngân sách vốn đầu tư cho các doanh nghiệp kinh doanh trong lĩnh vực kinh doanh bất động sản, phòng nghiên cứu của công ty tư vấn thiết kết xây dựng "Ngôi nhà tương lai" đã tiến hành thu thập dữ liệu từ các báo cáo tài chính năm 2008 của 266 công ty chuyên khai thác kinh doanh căn hộ cho thuê ở Việt Nam. Dữ liệu tổng hợp cuối cùng được cho trong tập tin **"REC.xls"**. Trong đó:

   - Y = Doanh số năm 2008 (triệu đôla)
   - X_2 = Tổng số lao động (ngàn người)

- X_3 = Chi tiêu vốn hữu hình (triệu đôla)
- X_4 = Chi tiêu vốn vô hình (triệu đôla)
- X_5 = Giá vốn hàng bán (triệu đôla)
- X_6 = Chi phí quản lý (triệu đôla)
- X_7 = Chi phí quảng cáo và bán hàng (triệu đôla)
- X_8 = Chi phí nghiên cứu & phát triển (triệu đôla)

a. Anh/Chị hãy lập ma trận hệ số tương quan giữa các biến trên (kể cả biến Y)? Nhận xét?

b. Anh/Chị hãy vẽ đồ thị phân tán (có đường hồi quy) giữa Y và các biến giải thích có hệ số tương quan với Y cao hơn 0.7? Nhận xét?

c. Anh/Chị hãy xây dựng mô hình hồi quy thích hợp? Trình bày chiến lược lựa chọn mô hình của Anh/Chị?

d. Sau khi đã xác định mô hình hồi quy thích hợp, Anh/Chị hãy kiểm định xem có hiện tượng phương sai thay đổi trong mô hình đó không? Tại sao? Anh/Chị trình bày cách thức khắc phục nếu mô hình có hiện tượng phương sai thay đổi?

e. Kiểm định và giải thích ý nghĩa kinh tế của các hệ số hồi quy trong mô hình tốt nhất mà Anh/Chị chọn?

f. Theo Anh/Chị, công ty sẽ sử dụng kết quả hồi quy như thế nào trong việc tư vấn lập ngân sách vốn đầu tư cho các doanh nghiệp trong tương lai?

5. Cơ chế trả lương cho giám đốc dự án đang ngày càng được các công ty xây dựng ở Việt Nam quan tâm. Chính vì thế, trung tâm Tư vấn doanh nghiệp và phát triển vùng của Đại học Kinh tế TP.HCM đã tiến hành thu thập thông tin của 50 giám đốc dự án ở TP.HCM và tổng hợp trong tập tin **"CPM.xls"**. Trong đó:

   - Salary = Tiền lương theo hợp đồng (đôla/tháng)
   - Bonus = Tiền thưởng trung bình hàng tháng (đôla/tháng)
   - Othercom = Các khoản tiền thưởng khác (đôla/tháng)
   - Compens = Tổng tiền lương (đôla/tháng)
   - Age = Tuổi

- Edu = Trình độ học vấn (0 = tốt nghiệp PTTH, 1 = tốt nghiệp đại học, 2 = tốt nghiệp sau đại học)
- Prof = Trình độ chuyên nghiệp (được đo lường bằng số khóa đào tạo chuyên nghiệp đã từng tham gia)
- Tenure = Số năm làm việc cho công ty hiện tại
- Exper = Số năm kinh nghiệm trong lĩnh vực quản lý dự án
- Value = Giá trị thị trường năm 2008 của công ty đang làm việc (triệu đôla)
- Profit = Lợi nhuận năm 2008 của công ty hiện đang làm việc (triệu đôla)
- Sales = Doanh số năm 2008 của công ty hiện đang làm việc (triệu đôla)

a. Anh/Chị hãy lập bảng so sánh tổng tiền lương trung bình của giám đốc dự án theo trình độ học vấn? Nhận xét?

b. Anh/Chị hãy lập bảng so sánh tổng tiền lương trung bình của giám đốc dự án theo số khóa đào tạo chuyên nghiệp đã tham gia? Nhận xét?

c. Anh/Chị hãy xác định mô hình hồi quy phù hợp về các nhân tố ảnh hưởng đến tiền lương của giám đốc dự án? Anh/Chị hãy trình bày chiến lược xây dựng mô hình của mình là gì?

d. Anh/Chị hãy thực hiện các kiểm định cần thiết và giải thích ý nghĩa các hệ số hồi quy của mô hình được chọn?

e. Anh/Chị cho biết mô hình hồi quy trên có thể sử dụng như thế nào?

6. Dữ liệu "**NONFARM.xls**" được rút trích từ VHLSS2006 có chứa các biến sau đây:

- Y = Thu nhập phi nông nghiệp năm 2006
- X_2 = Trình độ học vấn (đo bằng số năm đi học của chủ hộ)
- X_3 = Tổng chi tiêu cho giáo dục của hộ gia đình năm 2006
- X_4 = Tổng chi tiêu cho thông tin (báo chí, internet, TV, điện thoại)
- X_5 = Diện tích đất nông nghiệp

- X_6 = Quy mô hộ gia đình
- X_7 = Tuổi của chủ hộ
- Gender = Giới tính của chủ hộ (1 = Nam, 0 = Nữ)
- City = Biến giả địa bàn sinh sống (1 = Thành thị, 0 = Nông thôn)
- Quint = Năm nhóm chi tiêu
- Reg = 8 vùng kinh tế của Việt Nam

a. Anh/Chị cho biết thu nhập phi nông nghiệp có khác nhau giữa năm nhóm chi tiêu hay không? Tại sao?

b. Anh/Chị cho biết thu nhập phi nông nghiệp có khác nhau giữa thành thị và nông thôn hay không? Tại sao?

c. Anh/Chị cho biết thu nhập phi nông nghiệp có khác nhau giữa 8 vùng kinh tế của Việt Nam hay không? Tại sao?

d. Anh/Chị hãy tạo ra 8 biến giả đại diện cho 8 vùng kinh tế ở Việt Nam trên Eviews?

e. Anh/Chị hãy xây dựng mô hình kinh tế lượng phù hợp nhất về nhân tố ảnh hưởng đến thu nhập phi nông nghiệp? Anh/Chị hãy trình bày chiến lược xây dựng mô hình của mình?

f. Anh/Chị hãy xác định và kiểm định giả thiết cho rằng thu nhập phi nông nghiệp không khác nhau giữa các vùng kinh tế của Việt Nam?

g. Anh/Chị hãy kiểm định xem có hiện tượng đa cộng tuyến trong mô hình này không? Tại sao?

h. Anh/Chị hãy kiểm định xem liệu có hiện tượng phương sai thay đổi trong mô hình này hay không?

i. Anh/Chị cho biết kết quả nghiên cứu này có thể được sử dụng như thế nào trong dự báo và phân tích chính sách?

7. Sử dụng tập tin "**TOTALINVESTMENT.xls**" bao gồm các biến: I = tổng vốn đầu tư, Y = GDP, và R = lãi suất từ quý III năm 2001 đến quý IV năm 2008 để trả lời các câu hỏi sau:

a. Sử dụng kiểm định nghiệm đơn vị để kiểm định xem các biến I, Y, và R có phải là các chuỗi dừng hay không? Nếu chúng không phải là các chuỗi dừng thì Anh/Chị có cảnh báo gì khi phân tích hồi quy với các biến này?

b. Ước lượng mô hình hồi quy (Mô hình 1) trong đó I là biến phụ thuộc, Y và R là các biến giải thích. Sử dụng tất cả các phương pháp kiểm định để kiểm định xem có hiện tượng tự tương quan hay không?

c. Ước lượng mô hình hồi quy (Mô hình 2) trong đó lnI là biến phụ thuộc, lnY và lnR là các biến giải thích. Sử dụng tất cả các phương pháp kiểm định để kiểm định xem có hiện tượng tự tương quan hay không?

d. Ước lượng mô hình hồi quy (Mô hình 3) trong đó I là biến phụ thuộc, Y, R, và biến xu thế là các biến giải thích. Sử dụng tất cả các phương pháp kiểm định để kiểm định xem có hiện tượng tự tương quan hay không?

e. Anh/Chị có rút ra nhận xét gì về bản chất của hiện tượng tự tương quan trong Mô hình 1?

f. Nếu kết luận có hiện tượng tự tương quan trong Mô hình 1, Anh/Chị hãy áp dụng thủ tục Cochrane-Orcutt (trên Eviews) để khắc phục hiện tượng tự tương quan?

g. Với các điều kiện khác không đổi, nếu biết rằng lãi suất và GDP quý I năm 2009 lần lượt là 16% và 38 thì đầu tư trung bình dự kiến sẽ là bao nhiêu?

8. Sử dụng tập tin "**PRODUCT.xls**" trong đó Q = lượng cà phê được sản xuất trong năm, P = giá bán cà phê trung bình trong năm, F = lượng phân bón sử dụng trong năm, R = lượng mưa trung bình trong năm từ năm 1978 đến 2008 để trả lời các câu hỏi sau đây:

a. Sử dụng kiểm định nghiệm đơn vị để kiểm định xem các biến Q, P, F và R có phải là các chuỗi dừng hay không? Nếu chúng không phải là các chuỗi dừng thì Anh/Chị có cảnh báo gì khi phân tích hồi quy với các biến này?

b. Ước lượng mô hình hồi quy (Mô hình 1) với mẫu dữ liệu từ 1978 đến 2007 trong đó Q là biến phụ thuộc, P, F và R là các biến giải thích. Sử dụng tất cả các phương pháp kiểm định để kiểm định xem có hiện tượng tự tương quan hay không?

c. Ước lượng mô hình hồi quy (Mô hình 2) với mẫu dữ liệu từ 1978 đến 2007 trong đó lnQ là biến phụ thuộc, lnP, lnF và lnR là các biến giải thích. Sử dụng tất cả các phương pháp kiểm định để kiểm định xem có hiện tượng tự tương quan hay không?

d. Nếu kết luận có hiện tượng tự tương quan trong hai mô hình trên, Anh/Chị hãy chuyển sang hồi quy sai phân bậc một để khắc phục hiện tượng tự tương quan?

e. Với các điều kiện khác không đổi, Anh/Chị hãy dự báo sản lượng cà phê cho năm 2008? Theo Anh/Chị mô hình nào tốt hơn? Tại sao?

9. Một chuyên viên phòng kinh doanh của công ty điện lực TP.HCM muốn dự báo doanh thu (Y) của công ty cho năm 2008 bằng cách sử dụng hồi quy hàm đa biến. Chuyên viên này quyết định chọn ba biến giải thích như sau: (1) Mức sử dụng/kWh (X_2), phí sử dụng điện/kWh (X_3), và số lượng khách hàng mua điện từ công ty (X_4). Dữ liệu được cho ở tập tin "**ELECTRICITY.xls**". Anh/Chị hãy trả lời các câu hỏi sau đây:

   a. Hồi quy dạng hàng logarith và cho biết mô hình này có hiện tượng tự tương quan hay không? Tại sao?

   b. Nếu có tự tương quan, Anh/Chị đề xuất chuyên viên này nên xử lý như thế nào?

   c. Anh/Chị dự đoán xem chuyên viên này sẽ sử dụng kết quả hồi quy này như thế nào trong việc đề xuất các kế hoạch kinh doanh của công ty?

10. Mặc dù các mô hình giản đơn, hàm xu thế, và phân tích thành phần chuỗi thời gian cũng giúp ích cho việc dự báo giá CP, nhưng Ban giám đốc công ty kinh doanh sản phẩm khí cho rằng giá CP thực

sự phụ thuộc rất nhiều vào giá dầu của Mỹ, tình hình kinh tế thế giới (và 'sức khỏe' của các nền kinh tế lớn như Mỹ, Châu Âu, Nhật, và Trung Quốc), giá CP trong một vài tháng trước đó, và yếu tố mùa vụ. Với ý tưởng này, cùng các dữ liệu trong tập tin "**GAS.xls**", Anh/Chị hãy trả lời các câu hỏi sau đây:

a. Khảo sát ma trận hệ số tương quan và nhận diện các biến có thể ảnh hưởng đến giá CP?

b. Thực hiện các mô hình hồi quy theo phương pháp của Hendry và lựa chọn mô hình phù hợp nhất cho giá CP (kể cả các biến trễ của giá CP) cho giai đoạn trước tháng 12/2008?

c. Từ kết quả phần b, Anh/Chị hãy tạo ra và đưa các biến giả theo tháng vào mô hình. Anh/Chị cho biết các biến giả nào nên được đưa vào mô hình và giải thích ý nghĩa kinh tế của chúng?

d. Anh/Chị cho biết giá trị dự báo giá CP tháng 12/2008 là bao nhiêu? Anh/Chị cho biết giá trị thực của giá CP có nằm trong khoảng dự báo từ kết quả hồi quy này hay không?

e. Nếu ban giám đốc muốn biết hệ số co giãn của giá CP theo giá dầu, thì Anh/Chị sẽ làm như thế nào?

11. Giá vàng Việt Nam được cho là phụ thuộc nhiều vào giá vàng thế giới, lãi suất SIBOR, giá dầu của Mỹ, và một số chỉ số giá chứng khoán trên các thị trường quan trọng. Từ tập tin "**PRICE.xls**", Anh/Chị hãy xây dựng mô hình hồi quy tốt nhất về các nhân tố ảnh hưởng đến giá vàng Việt Nam? Anh/Chị hãy lưu ý vấn đề hồi quy giả mạo.

12. Trong tập tin "**GAP.xls**" có các biến xu thế (T), xu thế bình phương (T2), các biến giả theo quý (Q2 = 1 nếu là quý 2, Q3 = 1 nếu là quý 3, Q4 = 1 nếu là quý 4), biến giả D911 (D911 = 0 nếu trước vụ kiện khủng bố ngày 9/11, = 1 nếu sau ngày 9/11), và biến chỉ số niềm tin người tiêu dùng (ICS). Anh/Chị hãy xây dựng hàm kinh tế lượng về các nhân tố ảnh hưởng đến doanh số của GAP và dự báo cho năm 2004? Anh/Chị cho biết điều gì xảy ra với mức độ chính xác của mô hình nếu chúng ta mở rộng dự báo cho năm 2005?

13. Mặc dù đã thử thực hiện dự báo lượng khách hàng mới theo các mô hình giản đơn, hàm xu thế, và phân tích thành phần chuỗi thời gian, nhưng dựa vào kinh nghiệm quá khứ, vị giám đốc điều hành của CCC cho rằng lượng khách hàng gặp khó khăn về tài chính thường là những người rơi hoàn cảnh khó khăn như thất nghiệp, phá sản, bệnh tật. Chính vì thế, Ông đề nghị cô giám đốc nhân sự thu thập thêm hai thông tin như sau: (1) Số người nhận tem phiếu lương thực từ chính phủ hàng tháng (STAMP), và (2) Chỉ số về tình hình hoạt động kinh tế địa phương hàng tháng (BAI). Ông đề nghị cô giám đốc nhân sự thực hiện hồi quy số lượng khách mới theo hai biến trên.

    a. Anh/Chị dự đoán xem kết quả hồi quy của cô giám đốc nhân sự sẽ như thế nào?

    b. Dựa vào kết quả hồi quy này, Anh/Chị hãy dự báo số khách hàng mới cho ba tháng đầu của năm 1993? Anh/Chị cho biết kết quả này có khác gì so với các kết quả dự báo trước đây hay không?

14. Khi mới về làm việc cho Eden Group, Ông Đức, giám đốc tài chính, gặp phải một vấn đề khó khăn về cơ cấu vốn của công ty. Ông cho rằng công ty cần thêm tiền để thanh toán các khoản nợ ngắn hạn sắp đáo hạn và để tiếp tục triển khai một dự án phát triển khu liên hợp quy mô lớn. Vấn đề quan tâm lớn nhất của ông Đức là ước lượng lãi suất thị trường các trái phiếu 10 hoặc 30 năm bởi vì công ty cần quyết định xem nên tài trợ bằng nguồn vốn chủ sở hữu (phát hành thêm cổ phiếu) hoặc bằng các khoản vay dài hạn (phát hành trái phiếu). Để thực hiện quyết định quan trọng này, ông Đức cho rằng công ty cần có một dự báo đáng tin cậy về lãi suất mà công ty sẽ trả lúc phát hành trái phiếu. Đầu tuần, ông Đức triệu tập cuộc họp toàn bộ phòng tài chính để thảo luận kỹ về thị trường trái phiếu. Trong cuộc họp, một thành viên tên là Tài, vừa tốt nghiệp MDE của Chương trình cao học Việt Nam – Hà Lan, cho rằng công ty nên áp dụng phương pháp hồi quy đa biến để dự báo lãi suất trái phiếu vì vấn đề này đã được nhiều công ty áp dụng trên thực tế. Ông Đức rất giỏi trong lĩnh vực tài chính nhưng không biết nhiều về kinh tế lượng, nên ông hướng cuộc họp qua chủ đề khác một cách khéo léo. Sau cuộc họp, ông Đức yêu cầu ông Tài thực hiện nghiên cứu này và báo cáo kết quả cho ông vào thứ Hai tuần sau.

Ông Tài biết rằng mấu chốt của việc xây dựng một mô hình dự báo tốt bằng phân tích hồi quy là phải nhận diện đúng các biến giải thích liên quan đến lãi suất mà công ty phải trả lúc phát hành trái phiếu. Sau khi nghiên cứu và thảo luận với một số thành viên trong nhóm nghiên cứu, ông Tài quyết định chọn các biến sau đây: (1) Xếp hạng trái phiếu trong ngành bất động sản theo xếp hạng tín dụng của Moody's, (2) Tỷ số thu nhập/chi phí cố định, (3) Lãi suất trái phiếu chính phủ, (4) Thời gian đáo hạn của trái phiếu, và (5) Lãi suất cho vay cơ bản tại thời điểm phát hành.

Ông Tài thu thập dữ liệu liên quan về các lãi suất trái phiếu của các trái phiếu ngành bất động sản được xếp hạng bằng hoặc cao hơn Eden Group phát hành trong vòng hai năm qua từ hãng tin Reuters. Cuối cùng, ông có được bộ dữ liệu của 93 trái phiếu cho việc nghiên cứu của nhóm. Bộ dữ liệu được cho trong tập tin "**BOND.xls**" với các định nghĩa như sau:

- Y = lãi suất phải trả của công ty bất động sản lúc phát hành trái phiếu
- X_2 = 1 nếu trái phiếu được xếp loại A
- X_3 = 1 nếu trái phiếu được xếp loại AA
- X_4 = Tỷ số thu nhập/chi phí cố định
- X_5 = Lãi suất trái phiếu chính phủ (10 hoặc 30 năm) lúc phát hành trái phiếu
- X_6 = Thời gian đáo hạn (10 hoặc 30 năm)
- X_7 = Lãi suất cơ bản lúc phát hành trái phiếu

Kết quả nghiên cứu được trình lên ông Đức với ba nội dung ngắn gọn như sau:

(1) Mô hình dự báo tốt nhất là: $Y = -1.28 - 0.929X_2 - 1.18X_3 + 1.23X_5 + 0.0615X_6$. Mô hình này giải thích được 90.6% cho biến thiên của lãi suất trái phiếu của các công ty bất động sản.

(2) Sai số chuẩn của ước lượng là 0.53. Vì thế, khoảng 95% các giá trị thực của Y sẽ nằm trong khoảng 2*0.53 = 1.06 của giá trị dự báo.

(3) Các hệ số hồi quy có ý nghĩa thống kê và tỏ ra đáng tin cậy.

a. Anh/Chị đoán xem ông Đức sẽ hỏi ông Tài điều gì sau khi đọc báo cáo này?

b. Anh/Chị cho biết tại sao bộ dữ liệu có 6 biến giải thích nhưng trong báo cáo ông Tài chỉ đưa kết quả của 4 biến giải thích?

c. Theo Anh/Chị, ông Đức sẽ sử dụng kết quả ước lượng này như thế nào trong quyết định tài trợ của Eden Group?

d. Anh/Chị hãy thực hiện lại kết quả này trên Eviews?

CHƯƠNG 8 CÁC MÔ HÌNH DỰ BÁO THEO PHƯƠNG PHÁP BOX-JENKINS

Trong chương này chúng ta thảo luận các kỹ thuật ước lượng phương trình dự báo theo một cách khác so với các chương trước đây. Như đã biết, trong các mô hình hồi quy ở chương 7, chúng ta cố gắng phân tích hành vi và sự biến thiên của một biến phụ thuộc bằng cách hồi quy biến phụ thuộc đó với một số biến giải thích khác nhau. Trái lại, trong khung kinh tế lượng chuỗi thời gian, thì xuất phát điểm quan trọng nhất là khai thác thông tin sẵn có trong bản thân một biến số nào đó để tìm hiểu sự biến thiên của chính bản thân biến số đó. Việc phân tích một chuỗi thời gian duy nhất được gọi là một chuỗi thời gian đơn biến và đó chính là chủ đề sẽ được trình bày chi tiết trong chương 8 và chương 9 trong giáo trình này. Trong chương này, chúng ta sẽ khảo sát các mô hình ARIMA, một phương pháp dự báo chuỗi thời gian được sử dụng khá phổ biến trong việc dự báo các chỉ báo kinh tế có độ nhạy cao như lãi suất, chỉ số giá chứng khoán, giá vàng, giá dầu, giá khí, và giá cả các mặt hàng cụ thể của thị trường trong nước và thế giới.

MỤC TIÊU HỌC TẬP

Sau khi học xong chương này, chúng ta kỳ vọng sẽ đạt được các nội dung sau đây:

- Áp dụng kinh tế lượng vào chuỗi thời gian

- Các mô hình ARIMA
- Khái niệm tính dừng
- Kiểm định tính dừng
- Các mô hình chuỗi thời gian tự hồi quy
- Các mô hình bình quân di động
- Các mô hình ARMA
- Các mô hình ARIMA
- Quy trình thực hiện dự báo bằng phương pháp Box-Jenkins
- Tiêu chí đánh giá các mô hình ARIMA

KINH TẾ LƯỢNG VỀ CHUỖI THỜI GIAN

Như chúng ta đã đề cập trước đây, các nhà kinh tế lượng truyền thống chỉ tập trung vào việc sử dụng lý thuyết kinh tế và nghiên cứu các mối quan hệ tại cùng một thời điểm để giải thích các mối quan hệ giữa các biến phụ thuộc và biến giải thích. Trong các mô hình đó, đôi khi người ta sử dụng các biến trễ nhưng không theo một cách có hệ thống nào, hoặc ít nhất cũng không nhằm phân tích các trạng thái động hoặc phân tích cấu trúc thời gian của dữ liệu. Có nhiều khía cạnh khác nhau trong phân tích chuỗi thời gian nhưng một chủ đề phổ biến nhất là khai thác một cách triệt để cấu trúc động của chúng vốn có trong dữ liệu. Nói như vậy có nghĩa rằng chúng ta sẽ cố gắng rút trích càng nhiều thông tin quá khứ chứa đựng trong dữ liệu càng tốt. Theo Asteriou (2007), hai loại phân tích chuỗi thời gian chủ yếu là dự báo chuỗi thời gian và mô hình hóa trạng thái động của chuỗi thời gian. Dự báo chuỗi thời gian khác với kinh tế lượng truyền thống ở chỗ nó không quan tâm nhiều đến việc xây dựng các mô hình cấu trúc, tìm hiểu sự vận động của nền kinh tế hay kiểm định giả thiết. Vấn đề chính trong dự báo chuỗi thời gian là cố gắng xây dựng các mô hình hiệu quả để dự báo tốt nhất xu hướng vận động của chuỗi thời gian. Nói cách khác, cùng một chuỗi thời gian, mô hình nào cho kết quả dự báo có sai số dự báo bé nhất được xem là mô hình tốt nhất. Chính vì thế, các mô hình dự báo chuỗi thời gian thường được thực hiện bằng

cách khai thác tối đa mối quan hệ nội tại ở trạng thái động vốn tồn tại qua thời gian áp dụng cho bất kỳ một biến số nào. Ngược lại, mặc dù mô hình hóa trạng thái động vẫn quan tâm đến việc tìm hiểu cấu trúc của nền kinh tế và kiểm định giả thiết, nhưng nó bắt đầu trên quan điểm cho rằng hầu hết các chuỗi số liệu kinh tế vốn chậm điều chỉnh theo các cú sốc trong nên kinh tế và vì thế để hiểu quy trình này chúng ta nhất định phải hiểu được một cách đầy đủ quy trình điều chỉnh vốn có thể phức tạp và mất nhiều thời gian.

GIỚI THIỆU TỔNG QUAN CÁC MÔ HÌNH ARIMA

Box và Jenkins (1976) là những người đầu tiên giới thiệu các mô hình ARIMA, trong đó:

AR = Autogressive (tự hồi quy)

I = Integrated (chuỗi dừng sau khi chuyển sang dạng sai phân)

MA = Moving average (bình quân di động)

Ở các phần sau của chương này chúng ta sẽ được giới thiệu một số dạng khác nhau của các mô hình ARIMA và khái niệm về chuỗi dừng. Sau khi chúng ta đã hiểu một cách chi tiết hơn (so với chương 3) thế nào là một chuỗi dừng, tầm quan trọng của chuỗi dừng trong kinh tế lượng và dự báo, và một số cách kiểm định tính dừng, chúng ta sẽ lần lượt được giới thiệu một số mô hình khác nhau từ mô hình tự hồi quy bậc một giản đơn nhất cho đến các mô hình ARIMA phức tạp. Cuối cùng, và cũng chính là nội dung quan trọng nhất của chương này, chúng ta sẽ tập trung phân tích cách tiếp cận Box-Jenkins trong việc lựa chọn mô hình và dự báo theo phương pháp ARIMA.

TÍNH DỪNG

CHUỖI THỜI GIAN 'DỪNG'

Một khái niệm quan trọng trong các quy trình phân tích chuỗi thời gian là tính dừng. Mặc dù khái niệm này đã được giới thiệu ở chương 3,

nhưng do tầm quan trọng đặc biệt của nó trong các mô hình ARIMA và các mô hình ARCH/GARCH, nên chúng tôi sẽ đề cập lại một cách chi tiết hơn ở chương này. Ngoài ra, sau khi chúng ta đã có một nền tảng nhất định về phân tích hồi quy, thì cách tiếp cận dưới đây sẽ giúp chúng ta hiểu rõ hơn về khái niệm 'dừng' trong phân tích chuỗi thời gian. Một chuỗi thời gian 'dừng' có các đặc điểm sau đây:

- Dữ liệu dao động xung quanh một giá trị trung bình cố định trong dài hạn.
- Dữ liệu có giá trị phương sai xác định không thay đổi theo thời gian.
- Dữ liệu có một giản đồ tự tương quan với các hệ số tự tương quan sẽ giảm dần khi độ trễ tăng lên.

Theo ngôn ngữ thống kê, các đặc điểm trên của một chuỗi thời gian Y_t được thể hiện như sau:

- $E(Y_t)$ là một hằng số cho tất cả các thời điểm t

$$E(Y_t) = \mu \tag{8.1}$$

- $Var(Y_t)$ là một hằng số cho tất cả các thời điểm t

$$Var(Y_t) = E(Y_t-\mu)^2 = \sigma^2 \tag{8.2}$$

- $Cov(Y_t, Y_{t+k})$ là một hằng số cho tất cả các thời điểm *t* và *k* khác không. Lưu ý, giá trị của hiệp phương sai giữa hai giai đoạn chỉ phụ thuộc vào khoảng cách giữa hai giai đoạn.

$$Cov(Y_t, Y_{t+k}) = \gamma_k = E[(Y_t-\mu)(Y_{t+k}-\mu)] \tag{8.3}$$

Trong đó, γ_k là hiệp phương sai ở độ trễ *k*, là hiệp phương sai giữa các giá trị Y_t và Y_{t+k} (hoặc Y_{t-k}); nghĩa là, giữa hai giá trị Y cách nhau *k* giai đoạn. Nếu k = 0, ta có γ_0, đó cũng chính là phương sai của Y (σ^2); nếu k = 1, γ_1 là hiệp phương sai giữa hai giá trị Y liền nhau.

Giả sử khi ta di chuyển giá trị gốc của Y từ Y_t sang Y_{t+k} (ví dụ, từ quý I năm 1975 sang quý I năm 1985). Nếu Y_t là một chuỗi dừng, thì giá trị trung bình, phương sai, và hiệp phương sai của Y_{t+m} phải bằng

giá trị đại lượng này của Y_t. Tóm lại, nếu một chuỗi dừng, thì giá trị trung bình, phương sai, và hiệp phương sai (ở các độ trễ khác nhau) sẽ giống nhau không cần biết ta đang đo lường chúng tại thời điểm nào; điều này có nghĩa là, các đại lượng này không thay đổi theo thời gian. Một chuỗi dữ liệu như vậy sẽ có xu hướng trở về giá trị trung bình và những dao động xung quanh giá trị trung bình (được đo bằng phương sai là những biến thiên của giá trị chuỗi thời gian xoay quanh giá trị trung bình) sẽ là như nhau. Trong khi đó, nếu một chuỗi thời gian không dừng theo cách ta vừa định nghĩa ở trên, thì ta gọi đó là chuỗi không dừng. Nói cách khác, một chuỗi thời gian không dừng sẽ có giá trị trung bình thay đổi theo thời gian, hoặc giá trị phương sai thay đổi theo thời gian, hoặc cả hai.

Tại sao chuỗi thời gian dừng lại quan trọng? Gujarati (2003) cho rằng nếu một chuỗi thời gian không dừng, chúng ta chỉ có thể nghiên cứu hành vi của nó chỉ trong khoảng thời gian đang được xem xét. Vì thế, mỗi một mẫu dữ liệu thời gian sẽ mang một tình tiết nhất định và chỉ thể hiện những hành vi cụ thể trong một khoang thời gian xem xét. Kết quả là, chúng ta không thể khái quát hóa cho các giai đoạn thời gian khác. Đối với mục đích dự báo, các chuỗi thời gian không dừng như vậy có thể sẽ không có giá trị thực tiễn. Vì như chúng ta đã biết, trong dự báo chuỗi thời gian, chúng ta luôn giả định rằng xu hướng vận động của dữ liệu trong quá khứ và hiện tại được duy trình cho các giai đoạn tương lai. Và như vậy, chúng ta không thể dự báo được điều gì cho tương lai nếu như bản thân dữ liệu luôn thay đổi. Hơn nữa, đối với phân tích hồi quy, nếu chuỗi thời gian không dừng thì tất cả các kết quả điển hình của một phân tích hồi quy tuyến tính cổ điển sẽ không có giá trị cho việc dự báo, và thường được gọi là hiện tượng "hồi quy giả mạo". Do vậy, điều kiện cơ bản nhất cho việc dự báo một chuỗi thời gian là nó phải có tính dừng.

CHUỖI KHÔNG DỪNG

Gujarati (2003) cho rằng, mặc dù mối quan tâm chính của ta là ở các chuỗi dừng, nhưng thông thường ta lại hay gặp phải các chuỗi không dừng do bản chất của chuỗi có yếu tố xu thế hoặc ngẫu nhiên. Và đó dường như là bản chất của các biến kinh tế. Ví dụ cổ điển của trường

hợp chuỗi không dừng là mô hình bước ngẫu nhiên. Người ta cho rằng giá tài sản như giá cổ phiếu hay tỷ giá thường theo mô hình bước ngẫu nhiên nếu thực sự là thị trường chứng khoán hoạt động thực sự hiệu quả và thông tin khá cân xứng; điều này có nghĩa, chúng là các chuỗi không dừng. Kinh tế lượng chuỗi thời gian thường chia bước ngẫu nhiên thành hai loại: (i) bước ngẫu nhiên không có hằng số và (ii) bước ngẫu nhiên có hằng số.

Bước ngẫu nhiên không có hằng số

Giả sử u_t là một hạng nhiễu trắng với trung bình bằng 0 và phương sai bằng σ^2. Thì chuỗi Y_t được gọi là một bước ngẫu nhiên nếu:

$$Y_t = Y_{t-1} + u_t \tag{8.4}$$

Như vậy, theo phương trình (8.4) thì giá trị của Y tại thời điểm *t* bằng giá trị của Y tại thời điểm *t-1* cộng thêm một sai số ngẫu nhiên; như vậy, đây được coi như một mô hình theo cơ chế tự hồi quy bậc một, AR(1). Cơ chế AR(1) đã được đề cập ở chương 7. Ta có thể xem phương trình (8.4) như một phương trình hồi quy của Y tại thời điểm *t* theo giá trị trễ một giai đoạn của chính nó.

Phương trình (8.4) có thể được viết lại theo cách khác như sau:

$$Y_1 = Y_0 + u_1$$

$$Y_2 = Y_1 + u_2 = Y_0 + u_1 + u_2$$

$$Y_3 = Y_2 + u_3 = Y_0 + u_1 + u_2 + u_3$$

Như vậy, nếu quy trình bắt đầu ở thời điểm 0 với giá trị Y_0, thì ta có công thức tổng quát như sau:

$$Y_t = Y_0 + \sum u_t \tag{8.5}$$

Vì thế,

$$E(Y_t) = E(Y_0 + \sum u_t) = Y_0 \tag{8.6}$$

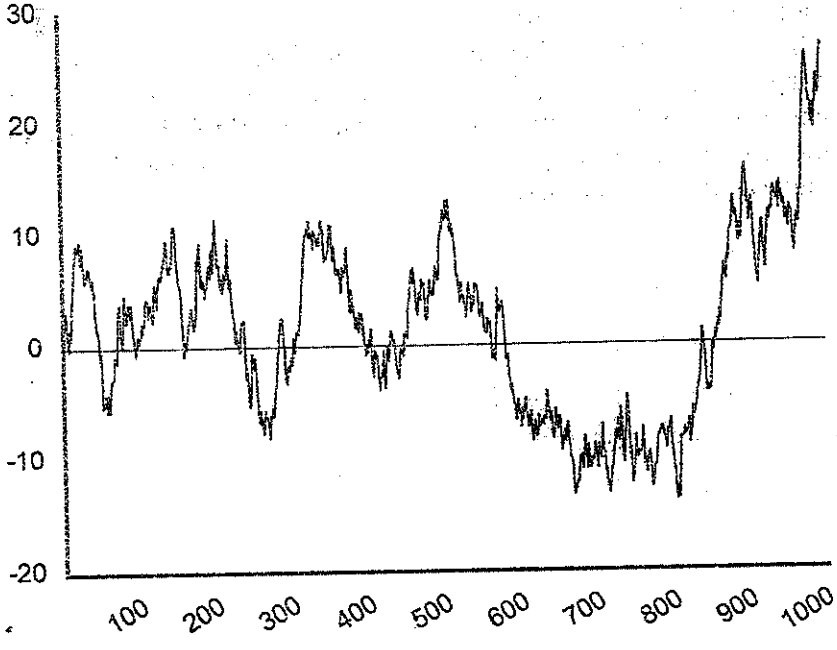
Và

$$Var(Y_t) = t\sigma^2 \tag{8.7}$$

Như vậy, rõ ràng rằng, giá trị trung bình của Y tại thời điểm *t* bằng giá trị trung bình của Y tại thời điểm ban đầu, và là một giá trị không đổi qua thời gian. Nhưng, khi *t* tăng lên, thì phương sai cũng sẽ tăng lên (phương trình (8.7)). Và điều này đã vi phạm điều kiện thứ hai của một chuỗi thời gian dừng. Tóm lại, một bước ngẫu nhiên không có hằng số là một chuỗi không dừng. Trên thực tế, Y_0 thường có giá trị bằng 0, và như thế $E(Y_t) = 0$. Để minh họa một bước ngẫu nhiên, chúng ta có thể thực hiện các bước sau đây trên Eviews. Sau khi mở một tập tin Eviews mới với số quan sát n = 500, từ cửa sổ lệnh, ta thực hiện các thao tác như sau:

```
Smpl 1 1
Genr yt=0
Smpl 2 1000
Genr yt=yt(-1)+nrnd
Smpl 1 1000
Plot yt
```

■ HÌNH 8.1: **Bước ngẫu nhiên không có hằng số.**



Phương trình (8.4) có thể được viết lại như sau:

$$(Y_t - Y_{t-1}) = \Delta Y_t = u_t \qquad (8.8)$$

Như vậy, nếu Y_t là một chuỗi không dừng, thì sai phân bậc một của nó có thể là một chuỗi dừng vì một chuỗi thời gian sau khi lấy sai phân bậc 1 thì nó đã loại trừ yếu tố xu thế hoặc ngẫu nhiên ra khỏi bản thân nó. Điều này rất có ý nghĩa trong việc phân tích và dự báo các chuỗi thời gian không dừng, cụ thể là nhà dự báo có thể biến một chuỗi không dừng thành một chuỗi dừng nhằm phục vụ cho quá trình dự báo theo một trình tự nhất định mà mô hình ARIMA là một trong những công cụ có khả năng thích nghi cho chuỗi không dừng sau khi lấy sai phân.

Bước ngẫu nhiên có hằng số

Nếu ta điều chỉnh phương trình (8.4) theo cách sau đây:

$$Y_t = \delta + Y_{t-1} + u_t \qquad (8.9)$$

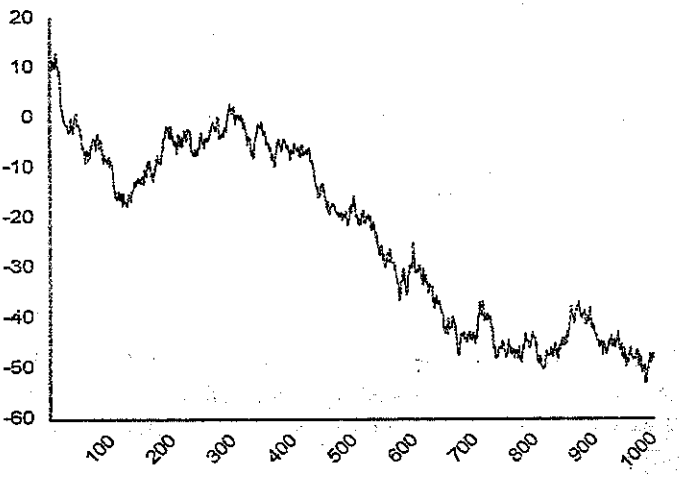
Trong đó, δ được gọi là một hằng số. Hằng số δ này có ý nghĩa như sau:

$$Y_t - Y_{t-1} = \Delta Y_t = \delta + u_t \qquad (8.10)$$

Như vậy, Y_t sẽ vận động lên hay xuống tùy thuộc vào δ dương hay âm. Để thấy rõ điều này, ta có thể tạo hai chuỗi Y_t như sau đây.

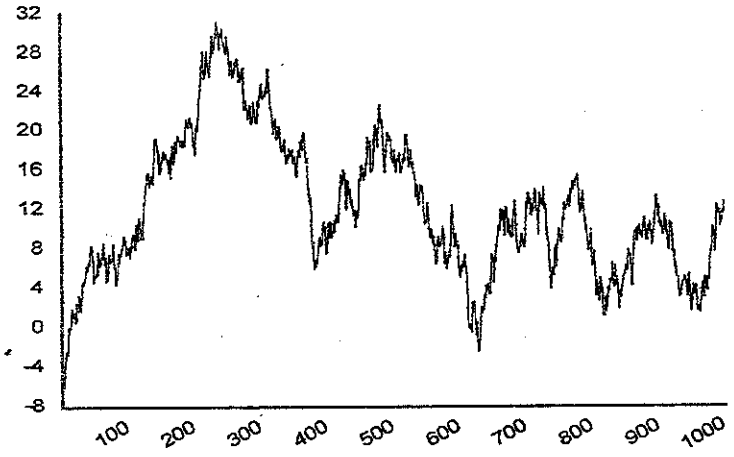
```
Smpl 1 1
Genr yt=10
Smpl 2 1000
Genr yt=yt(-1)+nrnd
Smpl 1 1000
Plot yt
```

■ HÌNH 8.2: **Bước ngẫu nhiên có hằng số.**



```
Smpl 1 1
Genr yt=-5
Smpl 2 1000
Genr yt=yt(-1)+nrnd
Smpl 1 1000
Plot yt
```

■ HÌNH 8.3: **Bước ngẫu nhiên có hằng số.**



Tương tự như đã thảo luận ở phần trên, một bước ngẫu nhiên có hằng số có các đặc điểm sau đây:

$$E(Y_t) = E(\delta + Y_0 + \sum u_t) = Y_0 + t\delta \qquad (8.11)$$

và

$$Var(Y_t) = t\sigma^2 \qquad (8.12)$$

Như vậy, một bước ngẫu nhiên có hằng số có giá trị trung bình và phương sai đều tăng theo thời gian (phương trình (8.11) và (8.12)). Và, điều này cho thấy nó vi phạm cả hai điều kiện của chuỗi dừng.

Nói cách khác, một bước ngẫu nhiên có hằng số bản thân nó là một chuỗi không dừng.

CHUỖI DỪNG SAI PHÂN

Như đã nói ở trên, một bước ngẫu nhiên là một chuỗi không dừng, nhưng sai phân bậc một của nó là một chuỗi dừng. Như vậy, ta có thể nói một bước ngẫu nhiên là một chuỗi dừng sai phân bậc một, và được ký hiệu là I(1). Lập luận tương tự, nếu một chuỗi thời gian không dừng ở sai phân bậc một, nhưng dừng ở sai phân bậc hai (lấy sai phân của sai phân bậc một), thì ta gọi đó là chuỗi thời gian đó là một chuỗi dừng ở sai phân bậc hai, và được ký hiệu là I(2). Một cách tổng quát, nếu một chuỗi dừng ở sai phân bậc *d*, thì ta ký hiệu là I(d). Và một chuỗi dừng cũng có thể gọi là một chuỗi dừng ở sai phân bậc 0, ký hiệu là I(0). Tuy vậy, theo kinh nghiệm của các nhà dự báo thì d thông thường cao nhất thường không vượt quá 2.

Đặc điểm của các chuỗi dừng sai phân

Theo Gujarati (2003), một chuỗi dừng sai phân có các đặc điểm quan trọng như sau:

- Nếu $X_t \sim I(0)$ và $Y_t \sim I(1)$, thì $Z_t = (X_t + Y_t) = I(1)$.
- Nếu $X_t \sim I(d)$ và $Z_t = (a + bX_t) = I(d)$.
- Nếu $X_t \sim I(d_1)$ và $Y_t \sim I(d_2)$, thì $Z_t = (aX_t + bY_t) \sim I(d_2)$, với $d_1 < d_2$.
- Nếu $X_t \sim I(d)$và $Y_t \sim I(d)$, thì $Z_t = (aX_t + bY_t) \sim I(d^*)$, với d^* có thể bằng d hoặc có khi $d^* < d$ do hiện tượng đồng liên kết.

Các đặc điểm này rất quan trọng. Nhớ rằng, trong mô hình hồi quy đơn ở chương 7, ta có:

$$\hat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2} \qquad (8.13)$$

Nếu $Y_t \sim I(0)$ và $X_t \sim I(1)$, thì có thể rất khó xác định phân phối xác suất của $\hat{\beta}_2$ (Lưu ý, nếu Y và X là các dữ liệu chéo, thì bản thân

chúng đã là các chuỗi dừng. Cho nên, ước lượng $\hat{\beta}_2$ sẽ theo phân phối chuẩn dưới các giả định của mô hình hồi quy tuyến tính cổ điển). Như vậy, chúng ta sẽ rất khó suy diễn thống kê cho các ước lượng OLS.

KIỂM ĐỊNH TÍNH DỪNG

Hai phương pháp kiểm định tính dừng thường được sử dụng là giản đồ tự tương quan (dựa vào thống kê t và thống kê Q) và kiểm định nghiệm đơn vị (dựa vào thống kê tau (τ) của Dickey-Fuller).

GIẢN ĐỒ TỰ TƯƠNG QUAN

Biểu đồ tự tương quan là một đồ thị biểu diễn mối quan hệ giữa hệ số tự tương quan bậc k với độ trễ k tương ứng. Hệ số tự tương quan bậc k (ký hiệu là r_k) được xác định theo công thức sau đây:

$$\rho_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2} \qquad (8.14)$$

Nếu ta chia cả tử và mẫu của phương trình (8.14) cho n, thì hệ số tự tương quan trên có thể được viết lại như sau:

$$\rho_k = \frac{Cov(Y_t, Y_{t-k})}{Var(Y_t)} \qquad (8.15)$$

Các phương trình (8.14) và (8.15) được gọi là hàm tự tương quan, ký hiệu là ACF.

Do thực tế chúng ta chỉ có dữ liệu mẫu, nên ta chỉ có thể ước lượng được hệ số tự tương quan mẫu theo công thức sau đây:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2} \qquad (8.16)$$

Trong đó, $\overline{Y}$ là giá trị trung bình mẫu của chuỗi Y_t, k là độ trễ, n là số quan sát của mẫu. Có hai phương pháp kiểm định xem hệ số tự tương quan có ý nghĩa thống kê hay không: Thống kê t, và Thống kê Q.

Thống kê t

Gọi ρ_k là hệ số tự tương quan tổng thể (r_k là ước lượng không chệch của ρ_k), ta có các giả thiết sau đây:

$$H_0: \rho_k = 0$$

$$H_1: \rho_k \neq 0$$

Nếu một chuỗi thời gian ngẫu nhiên thì các hệ số tự tương quan là một biến ngẫu nhiên có phân phối chuẩn với trung bình là 0 và phương sai là 1/N. Như vậy, với sai số chuẩn của hệ số tự tương quan $se(r_k)$ là $\sqrt{1/N}$, ta có thể xây dựng khoảng tin cậy cho ρ_k hoặc tìm được giá trị thống kê t tính toán ở một mức ý nghĩa xác định. Nếu ρ_k nằm ngoài khoảng tin cậy đó hoặc giá trị t tính toán lớn hơn giá trị t quan sát ta bác bỏ giả thiết H_0.

Thống kê Q

Hai cột cuối trong biểu đồ tự tương quan là thống kê Q của Ljung-Box và giá trị xác suất tương ứng. Thống kê Q kiểm định giả thiết đồng thời là tất cả các hệ số ρ_k cho tới một độ trễ đồng thời bằng không. Giá trị thống kê Q tính toán theo công thức sau đây:

$$Q = n \sum_{k=1}^{m} \rho_k^2 \qquad (8.17)$$

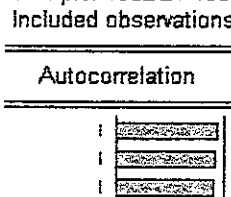
Với cỡ mẫu lớn, Q có phân phối theo χ^2 với bậc tự do bằng số độ trễ. Nếu giá trị thống kê Q tính toán lớn hơn giá trị thống kê Q tra bảng (χ^2 tra bảng, =CHIINV) ở một mức ý nghĩa xác định, ta bác bỏ giả thiết H_0.

Trong Eviews, ta lập biểu đồ tự tương quan bằng cách chọn **View/Correlogram** ... , xác định biểu đồ tự tương quan của chuỗi gốc (level) hay chuỗi sai phân bậc một (1st difference) và bậc hai 2nd

difference), và cuối cùng là xác định độ trễ k (lag). Ví dụ, chuỗi GDP trong có biểu đồ tự tương quan như sau (tập tin **DATA8-1**):

■ HÌNH 8.4: **Giản đồ tự tương quan của GDP.**

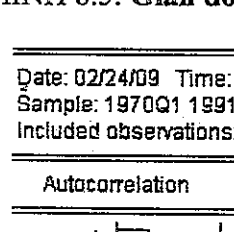
Date: 02/12/08 Time: 22:51
Sample: 1952Q1 1996Q4
Included observations: 180



		AC	PAC	Q-Stat	Prob
1		0.982	0.982	176.45	0.000
2		0.964	-0.005	347.46	0.000
3		0.946	-0.012	513.05	0.000
4		0.928	0.000	673.40	0.000
5		0.910	-0.009	828.56	0.000
6		0.893	-0.009	978.59	0.000
7		0.875	-0.008	1123.6	0.000
8		0.857	-0.017	1263.5	0.000

■ HÌNH 8.5: **Giản đồ tự tương quan của ΔGDP**

Date: 02/24/09 Time: 12:41
Sample: 1970Q1 1991Q4
Included observations: 87



		AC	PAC	Q-Stat	Prob
1		0.316	0.316	9.0136	0.003
2		0.186	0.095	12.165	0.002
3		0.049	-0.038	12.389	0.006
4		0.051	0.033	12.631	0.013
5		-0.007	-0.032	12.636	0.027
6		-0.019	-0.020	12.672	0.049
7		-0.073	-0.062	13.188	0.068
8		-0.289	-0.280	21.380	0.006
9		-0.067	0.128	21.820	0.009
10		0.019	0.100	21.855	0.016

Dựa vào biểu đồ tự tương quan để xác định một chuỗi thời gian dừng hay không với ý tưởng chính như sau. Nếu hệ số tự tương quan đầu

tiên khác không nhưng các hệ số tự tương quan tiếp theo bằng không có ý nghĩa thống kê, thì đó là một chuỗi dừng. Nếu một số hệ số tự tương quan khác không một cách có ý nghĩa thống kê thì đó là một chuỗi không dừng. Như vậy, bản thân chuỗi GDP là một chuỗi không dừng, như sai phân bậc 1 của GDP là một chuỗi dừng. Nói cách khác, GDP là một chuỗi dừng sai phân bậc 1.

KIỂM ĐỊNH NGHIỆM ĐƠN VỊ

Kiểm định nghiệm đơn vị là một kiểm định được sử dụng khá phổ biến để kiểm định một chuỗi thời gian dừng hay không dừng. Lưu ý, khi chúng ta thực hiện một nghiên cứu khoa học với dữ liệu chuỗi thời gian, thì chúng ta nên áp dụng kiểm định nghiệm đơn vị (thay vì giản đồ tự tương quan), vì loại kiểm định này có tính học thuật và chuyên nghiệp cao hơn. Giả sử ta có phương trình tự hồi quy như sau:

$$Y_t = \rho Y_{t-1} + u_t \quad (-1 \le \rho \le 1) \tag{8.18}$$

Ta có các giả thiết:

H_0: $\rho = 1$ (Y_t là chuỗi không dừng)

H_1: $\rho < 1$ (Y_t là chuỗi dừng)

Phương trình (8.18) tương đương với phương trình (8.19) sau đây:

$$\begin{aligned} Y_t - Y_{t-1} &= \rho Y_{t-1} - Y_{t-1} + u_t \\ &= (\rho - 1)Y_{t-1} + u_t \\ \Delta Y_t &= \delta Y_{t-1} + u_t \end{aligned} \tag{8.19}$$

Như vậy các giả thiết ở trên có thể được viết lại như sau:

H_0: $\delta = 0$ (Y_t là chuỗi không dừng)

H_1: $\delta < 0$ (Y_t là chuỗi dừng)

Dickey và Fuller cho rằng giá trị *t* ước lượng của hệ số Y_{t-1} sẽ theo phân phối xác suất τ (tau statistic, τ = giá trị δ ước lượng/sai số của hệ số δ). Kiểm định thống kê τ còn được gọi là kiểm định Dickey – Fuller (DF). Kiểm định DF được ước lượng với 3 hình thức:

- Khi Y_t là một bước ngẫu nhiên không có hằng số:

$$\Delta Y_t = \delta Y_{t-1} + u_t \qquad (8.20)$$

- Khi Y_t là một bước ngẫu nhiên có hằng số:

$$\Delta Y_t = \beta_1 + \delta Y_{t-1} + u_t \qquad (8.21)$$

- Khi Y_t là một bước ngẫu nhiên với hằng số xoay quanh một đường xu thế ngẫu nhiên:

$$\Delta Y_t = \beta_1 + \beta_2 TIME + \delta Y_{t-1} + u_t \qquad (8.22)$$

Để kiểm định H_0 ta so sánh giá trị thống kê τ tính toán với giá trị thống kê τ tra bảng DF (các phần mềm kinh tế lượng đều cung cấp giá trị thống kê τ). Nếu giá trị tuyệt đối của thống kê τ lớn hơn giá trị giá trị τ tra bảng, ta bác bỏ giả thiết H_0, nghĩa là Y_t là một chuỗi dừng. Ngược lại, nếu giá trị tuyệt đối của thống kê τ nhỏ hơn giá trị giá trị τ tra bảng, ta không bác bỏ giả thiết H_0, nghĩa là Y_t là một chuỗi không dừng. Sử dụng tập tin **DATA8-1** cho ba mô hình (8.20), (8.21), và (8.22), với giả thiết H_0 cho rằng chuỗi GDP có nghiệm đơn vị, hoặc nói cách khác, chuỗi GDP là một chuỗi không dừng. Theo ngôn ngữ thống kê, ta có thể viết như sau:

H_0: $\delta = 0$ (GDP là chuỗi không dừng)

H_1: $\delta < 0$ (GDP là chuỗi dừng)

Sau đây là các kết quả kiểm định trên Eviews (sẽ được hướng dẫn cụ thể sau ví dụ minh họa này):

■ HÌNH 8.6: **Kiểm định nghiệm đơn vị với Mô hình (8.20).**

Null Hypothesis: GDP has a unit root

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		5.798077	1.0000
Test critical values:	1% level	-2.591813	
	5% level	-1.944574	
	10% level	-1.614315	

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(GDP)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GDP(-1)	**0.005765**	0.000994	5.798077	0.0000

■ HÌNH 8.7: **Kiểm định nghiệm đơn vị với Mô hình (8.21).**

Null Hypothesis: GDP has a unit root

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-0.219165	0.9310
Test critical values:	1% level	-3.507394	
	5% level	-2.895109	
	10% level	-2.584738	

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(GDP)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GDP(-1)	-0.001368	0.006242	-0.219165	0.8270
C	28.20542	24.36532	1.157605	0.2503

■ HÌNH 8.8: **Kiểm định nghiệm đơn vị với Mô hình (8.22).**

Null Hypothesis: GDP has a unit root

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.625296	0.7750
Test critical values:	1% level	-4.066981	
	5% level	-3.462292	
	10% level	-3.157475	

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(GDP)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GDP(-1)	-0.060317	0.037111	-1.625296	0.1078
C	190.3837	103.5257	1.838999	0.0694
@TREND(1970Q1)	1.477641	0.917244	1.610958	0.1109

Chúng ta có nhận xét gì về kết quả kiểm định của ba mô hình này? Vấn đề quan tâm chính của chúng ta ở đây là giá trị t (=τ) của hệ số theo biến GDP_{t-1}. Lưu ý rằng, các giá trị τ tính toán và τ phê phán (tra bảng) của ba mô hình trên là khác nhau. Trước khi quyết định mô hình nào trong ba mô hình trên là thích hợp nhất để sử dụng, ta phải kiểm tra và so sánh kết quả giữa chúng. Ta nên loại bỏ mô hình (8.20) vì hệ số của biết GDP_{t-1} là dương. Bởi vì $\delta = (\rho-1)$, nên δ dương hàm ý rằng $\rho > 1$. Mặc dù, về mặt lý thuyết thì điều này có thể xảy ra, nhưng ta phải loại trường hợp này do GDP có thể là một chuỗi gia tăng đột biến (không ổn định hoặc tính xu thế rất rõ rệt). Như vậy, ta phải so sánh hai mô hình (8.21) và (8.22). Cả hai trường hợp đều phù hợp vì có δ âm, nghĩa là $\rho < 1$. Nếu ta hồi quy GDP_t theo GDP_{t-1}, và GDP_t theo GDP_{t-1} và @trend(1969Q4), ta sẽ có các giá trị $\hat{\rho}$ lần lượt là 0.9986 và 0.9397. Điều này có nghĩa $\hat{\rho} \sim 1$. Vậy, có thể GDP là một bước ngẫu nhiên. Để làm rõ vấn đề này, ta cần kiểm định xem các giá trị này có nhỏ hơn 1 có ý nghĩa thống kê hay không trước khi có kết luận GDP là một chuỗi không dừng?

. Đối với mô hình (8.21), thì giá trị tuyệt đối của τ tính toán là 0.219, nhỏ hơn giá trị τ tra bảng ở mức ý nghĩa 10%. Tương tự, giá trị

tuyệt đối của τ tính toán trong mô hình (8.22) là 1.625, nhỏ hơn giá trị τ tra bảng ở mức ý nghĩa 10%. Như vậy, rõ ràng GDP là một chuỗi không dừng. Kết luận này hoàn toàn phù hợp với giản đồ tự tương quan của GDP như ở phần trên.

Tuy nhiên, do có thể có hiện tượng tương quan chuỗi giữa các u_t do thiếu biến, nên người ta thường sử dụng kiểm định DF mở rộng là ADF (Augmented Dickey – Fuller Test). Kiểm định này được thực hiện bằng cách đưa thêm vào phương trình (8.22) các biến trễ của sai phân biến phụ thuộc ΔY_t:

$$\Delta Y_t = \beta_1 + \beta_2 TIME + \delta Y_{t-1} + \alpha_i \Sigma \Delta Y_{t-i} + \varepsilon_t \qquad (8.23)$$

Áp dụng cho chuỗi GDP ta có kết quả kiểm định như sau:

■ HÌNH 8.9: **Kiểm định nghiệm đơn vị với Mô hình (8.23).**

Null Hypothesis: GDP has a unit root

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.215287	0.4749
Test critical values:	1% level	-4.068290	
	5% level	-3.462912	
	10% level	-3.157836	

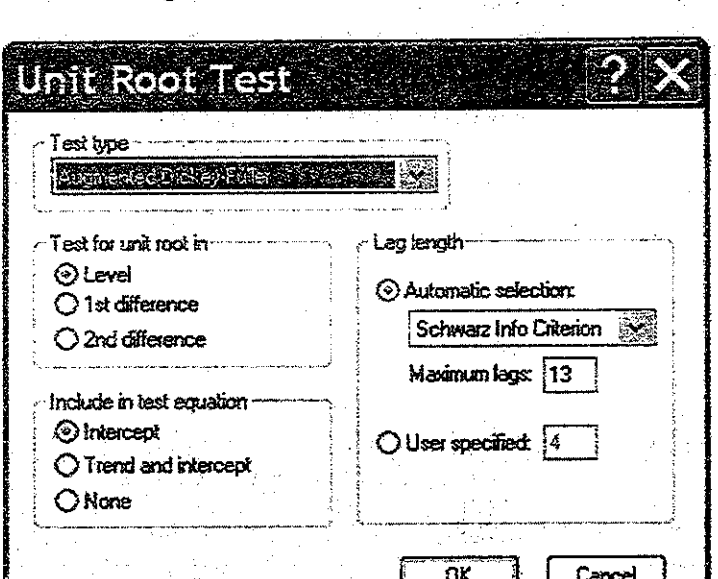
Augmented Dickey-Fuller Test Equation
Dependent Variable: D(GDP)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GDP(-1)	-0.078661	0.035508	-2.215287	0.0295
D(GDP(-1))	0.355794	0.102691	3.464708	0.0008
C	234.9729	98.58764	2.383391	0.0195
@TREND(1970Q1)	1.892199	0.879168	2.152260	0.0343

Như vậy, giá trị t ($=\tau$) tuyệt đối của biến GDP_{t-1} là 2.215 nhỏ hơn giá trị τ tra bảng ở mức ý nghĩa 10% một lần nữa khẳng định rằng GDP là một chuỗi không dừng.

Để tiến hành kiểm định nghiệm đơn vị trên Eviews ta chọn **View/Unit Root Test ...**, sẽ xuất hiện hộp thoại **Unit Root Test.**

HÌNH 8.10: **Hướng dẫn kiểm định nghiệm đơn vị trên Eviews.**



Ở lựa chọn **Test for unit root in**, chọn level nếu muốn kiểm định chuỗi gốc có phải là một chuỗi dừng hay không, chọn <1st difference> nếu muốn kiểm định chuỗi sai phân bậc một có phải là một chuỗi dừng hay không. Ở lựa chọn **Include in test equation**, chọn <intercept> nếu dùng phương trình (8.21), chọn <trend and intercept> nếu dùng phương trình (8.22), chọn <none> nếu dùng phương trình (8.20), chọn <trend and intercept> và xác định độ trễ ở lựa chọn <lag length> nếu dùng phương trình (8.23).

CÁC MÔ HÌNH TỰ HỒI QUY

MÔ HÌNH AR(1)

Như đã trình bày ở chương 7, AR nghĩa là cơ chế tự hồi quy. Nói cách khác, biến phụ thuộc được hồi quy theo các biến trễ của nó. Đơn giản nhất là mô hình AR(1) có dạng như sau:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + u_t \qquad (8.24)$$

Để đơn giản ta không đưa vào mô hình giá trị hằng số, $-1 < \phi_1 < 1$, và u_t là số hạng đảm bảo tính nhiễu trắng. Các giả định về u_t cũng giống như các giả định trong các mô hình hồi quy tuyến tính cổ điển.

Hàm ý của AR(1) là hành vi của chuỗi thời gian Y_t phần lớn được xác định bởi giá trị trước đó của chính chuỗi thời gian đó. Ví dụ, chỉ số giá tiêu dùng của tháng 2 năm 2009 có thể lớn phụ thuộc vào chỉ số giá tiêu dùng của tháng 1 năm 2009, hoặc chỉ số giá chứng khoán hôm nay có thể phụ thuộc vào chỉ số giá chứng khoán của ngày hôm qua.

Các mô hình tự hồi quy chỉ phù hợp với các chuỗi dừng và hệ số ϕ_0 thể hiện mức trung bình của chuỗi. Nếu dữ liệu dao động xung quanh giá trị 0 hoặc dạng sai phân thì không cần hệ số ϕ_0 trong mô hình (8.24).

Trong phương trình (8.24), ta có ràng buộc $-1 < \phi_1 < 1$ để đảm bảo tính dừng của chuỗi thời gian Y_t. Nếu giá trị tuyệt đối của $\phi > 1$, thì Y_t sẽ có xu hướng càng ngày càng lớn hơn và vì thế có thể trở thành một chuỗi gia tăng đột biến. Xem ví dụ sau đây (lưu ý, các ví dụ này khác với các ví dụ ở phần bước ngẫu nhiên, vì bước ngẫu nhiên có $\phi = 1$):

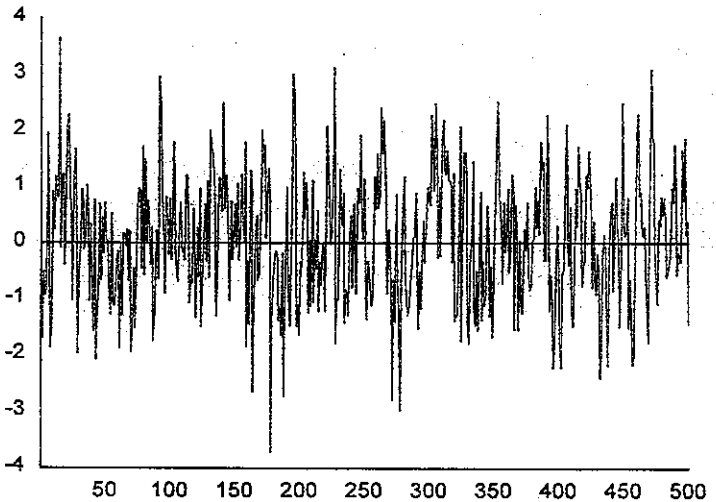
- **Chuỗi dừng:**

Mở Eviews và tạo một tập tin mới, có thể dạng 'undated', với quan sát thứ nhất là 1 và quan sát cuối cùng là 500. Để tạo một chuỗi dừng ta thực hiện các lệnh sau đây:

```
Smpl 1 1
Genr yt=0
Smpl 2 500
Genr yt=0.4*yt(-1)+nrnd
Smpl 1 500
Plot yt
```

Lưu ý, việc tạo giá trị đầu tiên của chuỗi Y_t bằng 0 có nghĩa rằng chuỗi Y_t sẽ xoay quanh giá trị trung bình là 0. Quan sát đồ thị ta thấy đây là một chuỗi dừng vì giá trị trung bình và phương sai là hằng số. Nếu ta xem giản đồ tự tương quan sẽ nhận thấy rằng các hệ số tự tương quan sẽ giảm xuống một cách nhanh chóng (=0) sau vài độ trễ.

■ HÌNH 8.11: **Đồ thị một chuỗi dừng.**



■ HÌNH 8.12: **Giản đồ tự tương quan của một chuỗi dừng.**

Date: 01/13/09 Time: 11:00
Sample: 1 500
Included observations: 500

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.377	0.377	71.356	0.000
		2	0.144	0.002	81.799	0.000
		3	0.105	0.059	87.393	0.000
		4	0.078	0.023	90.502	0.000
		5	0.045	0.002	91.524	0.000
		6	0.005	-0.023	91.538	0.000
		7	-0.029	-0.034	91.957	0.000
		8	-0.013	0.008	92.048	0.000
		9	-0.075	-0.081	94.915	0.000
		10	-0.096	-0.045	99.618	0.000
		11	-0.089	-0.033	103.67	0.000
		12	-0.088	-0.035	107.63	0.000

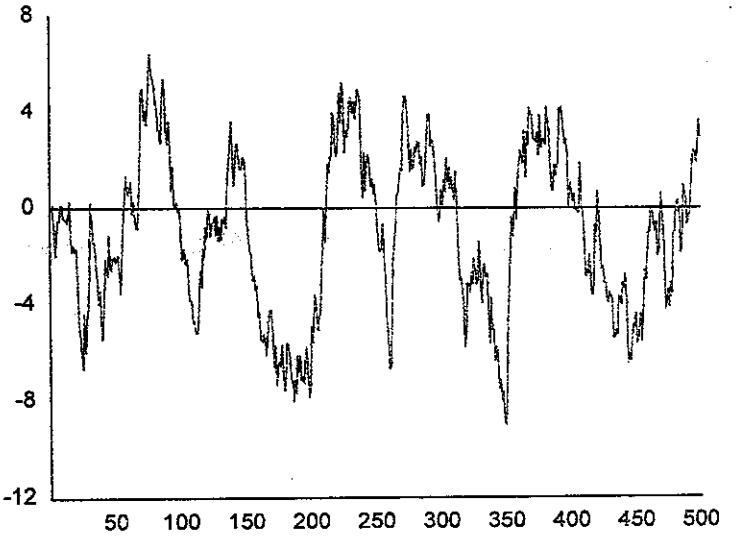
- **Chuỗi không dừng:**

Để tạo một chuỗi không dừng, trước hết chúng ta mở một tập tin Eviews mới với n=500, rồi thực hiện các lệnh sau đây:

```
Smpl 1 1
Genr yt=0
Smpl 2 500
Genr yt=0.98*yt(-1)+nrnd
Smpl 1 500
Plot yt
```

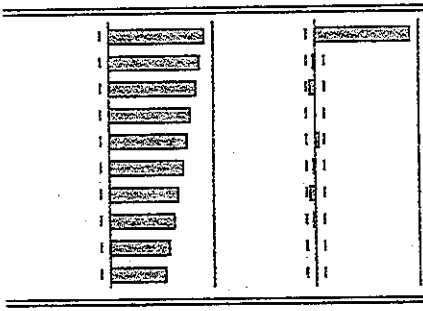
Quan sát đồ thị (Hình 8.13) ta thấy đây là một chuỗi không dừng vì giá trị trung bình không đổi (bằng 0) nhưng phương sai tăng lên khi *t* tăng lên. Như vậy, chuỗi thời gian trong Hình 8.13 chỉ thỏa mãn được điều kiện thứ nhất của một chuỗi dừng. Nếu ta xem giản đồ tự tương quan (Hình 8.14), thì ta sẽ nhận thấy rằng các hệ số tự tương quan cao và giảm dần khi độ trễ tăng lên. Như vậy, chuỗi thời gian trong Hình 8.13 đúng là một chuỗi không dừng (đã đề cập ở chương 3).

■ HÌNH 8.13: **Đồ thị một chuỗi không dừng (quanh 0).**



■ HÌNH 8.14: **Giản đồ tự tương quan của một chuỗi không dừng.**

Date: 02/24/09 Time: 20:22
Sample: 1 500
Included observations: 500



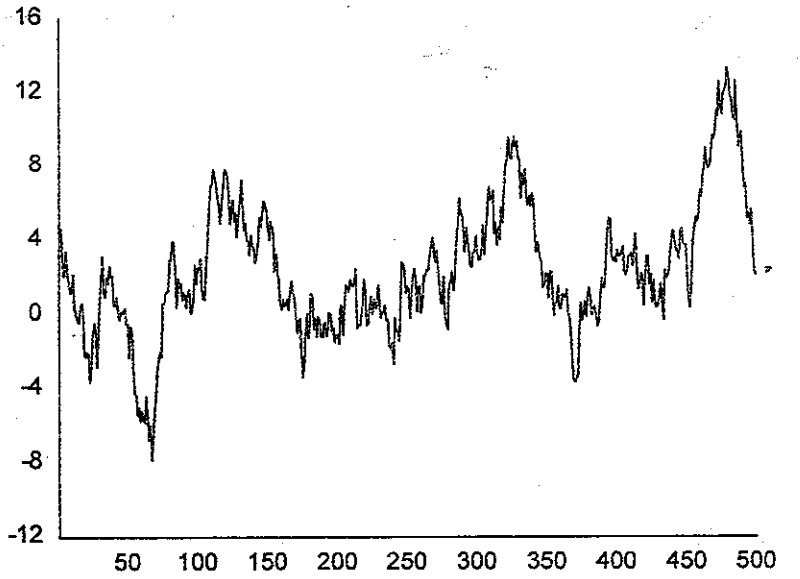
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.954	0.954	458.15	0.000
		2	0.909	-0.021	874.58	0.000
		3	0.861	-0.057	1248.6	0.000
		4	0.814	-0.006	1584.0	0.000
		5	0.772	0.028	1886.4	0.000
		6	0.731	-0.018	2157.8	0.000
		7	0.687	-0.058	2398.0	0.000
		8	0.644	-0.011	2609.7	0.000
		9	0.603	-0.003	2795.5	0.000
		10	0.564	0.003	2958.6	0.000

Nếu tạo một chuỗi không dừng xoay quanh giá trị trung bình là 4, thì ta làm như sau:

```
Smpl 1 1
Genr yt=4
Smpl 2 500
Genr yt=0.97*yt(-1)+nrnd
Plot yt
```

Ta sẽ thấy trên đồ thị Hình 8.15, Y_t là một chuỗi không dừng và xoay quanh giá trị trung bình bằng 4.

■ HÌNH 8.15: **Đồ thị một chuỗi không dừng (quanh 4) .**



- **Chuỗi gia tăng đột biến**

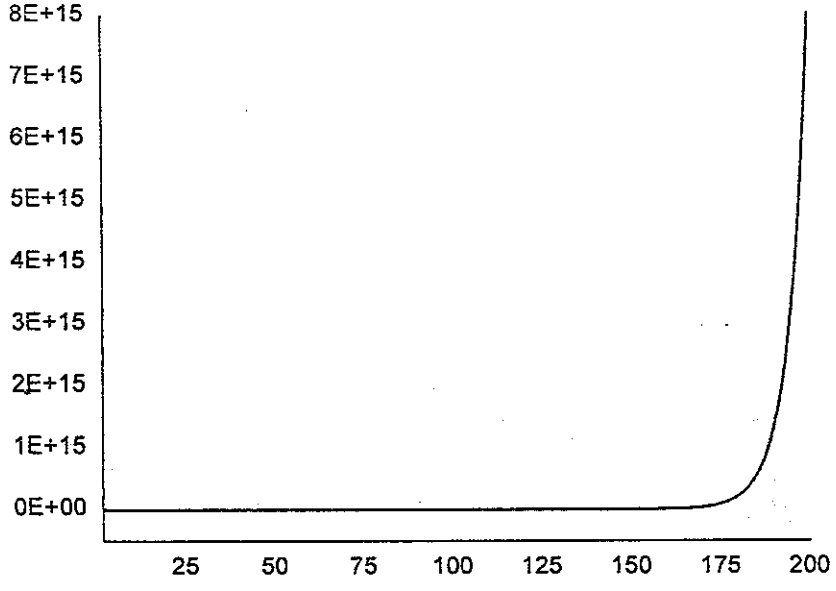
Bây giờ, ta tạo thêm một chuỗi X_t (trên tập tin Eviews đã có sẵn) có giá trị tuyệt đối của $\phi > 1$, ví 1.2 với các lệnh như sau:

```
Smpl 1 1
Genr xt=1
Smpl 2 500
Gen xt=1.2*xt(-1)+nrnd
Smpl 1 200
Plot xt
```

Hình 8.16 cho thấy chuỗi X sẽ gia tăng một cách rất nhanh chóng sau một vài giai đoạn. Và một chuỗi gia tăng đột biến hiển nhiên là một

chuỗi không dừng. Trong thực tế hiếm khi chúng ta gặp những chỉ báo kinh tế vận động theo cách này, cho nên giáo trình này sẽ không đề cập đến các chuỗi thời gian như vậy. Tuy nhiên, một số chuỗi thời gian, nhất là chứng khoán, nếu chúng ta chỉ xét một giai đoạn ngắn, thì khả năng xảy ra hiện tượng này là không nhỏ. Chính vì vậy, khi thiếu dữ liệu quá khứ, thì chúng ta không nên sử dụng các mô hình ARIMA.

■ HÌNH 8.16: **Đồ thị một chuỗi gia tăng đột biến.**



■ HÌNH 8.17: **Giản đồ tự tương quan của chuỗi gia tăng đột biến.**

Date: 01/13/09 Time: 11:10
Sample: 1 200
Included observations: 200

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.833	0.833	140.88	0.000
		2	0.694	-0.000	239.12	0.000
		3	0.578	-0.000	307.59	0.000
		4	0.481	-0.000	355.30	0.000
		5	0.400	-0.000	388.52	0.000
		6	0.333	-0.000	411.63	0.000
		7	0.277	-0.000	427.70	0.000
		8	0.230	-0.000	438.85	0.000
		9	0.191	-0.001	446.58	0.000
		10	0.159	-0.001	451.93	0.000
		11	0.131	-0.001	455.62	0.000
		12	0.109	-0.001	458.16	0.000

Như vậy, về mặt lý thuyết, một chuỗi gia tăng đột biến hiển nhiên là một chuỗi không dừng.

MÔ HÌNH AR(p)

Mô hình AR(2) có dạng như sau:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + u_t \qquad (8.25)$$

Mô hình AR(p) có dạng như sau:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_2 Y_{t-p} + u_t \qquad 8.26)$$

Hay dạng rút gọn như sau:

$$Y_t = \phi_0 + \sum_{i=1}^{p} \phi_i Y_{t-i} + u_t \qquad (8.27)$$

Hệ số ϕ_0 cũng được giải thích như đã được trình bày trong cơ chế AR(1).

Điều kiện để một chuỗi trong mô hình AR(p) là chuỗi dừng phải thỏa điều kiện sau:

$$\sum_{i=1}^{p} \phi_i < 1 \qquad (8.28)$$

Để xác định số độ trễ p ta sử dụng giản đồ tự tương quan theo cách như sau: ACF sẽ có xu hướng bằng không ngay lập tức, trong khi đó, hệ số tự tương quan riêng, ký hiệu là PACF, sẽ có xu hướng khác không một cách có ý nghĩa thống kê cho đến độ trễ p và sẽ bằng không ngay sau độ trễ p đó. $PACF_k$ là gì?

Trong phân tích hồi quy bội, nếu biến phụ thuộc được hồi quy theo các biến giải thích X_2 và X_3, thì điều mà chúng ta quan tâm là muốn biết X_2 ảnh hưởng lên Y như thế nào khi đã loại trừ ảnh hưởng của X_3 lên Y. Điều này có nghĩa hồi quy Y theo X_3, lưu phần dư, rồi hồi quy phần dư theo X_2. Trong phân tích chuỗi thời gian, thì khái niệm PACF cũng tương tự như vậy. $PACF_k$ được sử dụng để đo lường mức độ quan hệ giữa Y_t và Y_{t-k}, khi các ảnh hưởng của các độ trễ từ 1 đến k-1 đã được loại trừ. Mục đích phổ biến của việc xác định $PACF_k$ là để xác định mô hình ARIMA thích hợp. Thực vậy, $PACF_k$ hầu như chỉ được sử dụng cho mỗi mục đích này trong phân tích chuỗi thời gian.

Hệ số tự tương quan riêng bậc m được định nghĩa như hệ số tự hồi quy cuối cùng của mô hình AR(m). Ví dụ, các phương trình (8.29), đến (8.31) được sử dụng để xác định các mô hình AR(1), AR(2), ..., AR(m). Hệ số cuối cùng của Y trong mỗi phương trình này, $\hat{\phi}_1$, $\hat{\phi}_2$, ..., $\hat{\phi}_m$, là hệ số tự tương quan riêng.

$$Y_t = \hat{\phi}_1 Y_{t-1} + e_t \qquad (8.29)$$

$$Y_t = \phi_1 Y_{t-1} + \hat{\phi}_2 Y_{t-1} + e_t \qquad (8.30)$$

...

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \hat{\phi}_m Y_{t-m} + e_t \qquad (8.31)$$

Trên lý thuyết, ta có thể giải hệ các phương trình này để tìm các hệ số $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_m$, nhưng đòi hỏi rất nhiều thời gian. Trong phạm vi bài giảng này, chúng ta chỉ cần hiểu sẽ sử dụng các hệ số tự tương quan riêng như thế nào trong việc xác định mô hình AR(p) thì hợp là được, vì các phần mềm kinh tế lượng như Eviews hoặc các phần mềm khác đều có cung cấp các giá trị này.

Sau khi đã hiểu sơ qua hệ số tự tương quan riêng là gì và chúng được ước tính như thế nào, bây giờ chúng ta tìm hiểu chi tiết hơn cách sử dụng chúng trong việc xác định một mô hình ARIMA phù hợp. Nếu quá trình tạo ra một chuỗi theo mô hình AR(1), thì chỉ có hệ số $\hat{\phi}_1$ có ý nghĩa thống kê, trong khi các hệ số $\hat{\phi}_2, \hat{\phi}_3, \ldots, \hat{\phi}_m$ đều không có ý nghĩa thống kê. Nếu quá trình tạo ra một chuỗi theo mô hình AR(2), thì chỉ có $\hat{\phi}_1$ và $\hat{\phi}_2$ có ý nghĩa thống kê, trong khi các hệ số $\hat{\phi}_3, \ldots, \hat{\phi}_m$ đều không có ý nghĩa thống kê. Lập luận tương tự cho các mô hình AR(3), AR(4), ..., AR(p). Lưu ý, nếu quá trình tạo ra một chuỗi theo cơ chế MA (sẽ được trình bày ở phần sau), chứ không phải cơ chế AR, thì các hệ số tự tương quan riêng không cho chúng ta biết bậc của quá trình MA, vì chúng chỉ được thiết lập để giúp xác định một quá trình AR phù hợp. Trong trường hợp này, ACF là thống kê thích hợp nhất để xác định số bậc thích hợp cho quá trình MA.

VÍ DỤ MINH HỌA

Sử dụng tập tin **DATA8-2**, trong đó Y là doanh số theo tháng của một công ty, và thực hiện các bước sau đây:

Bước 1: Vẽ giản đồ tự tương quan

■ HÌNH 8.18: **Giản đồ tự tương quan của doanh số Y.**

Date: 02/24/09 Time: 17:35
Sample: 1 75
Included observations: 75

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.528	-0.528	21.796	0.000
		2	0.282	0.003	28.068	0.000
		3	-0.038	0.155	28.186	0.000
		4	0.008	0.065	28.191	0.000
		5	0.144	0.189	29.908	0.000
		6	-0.137	0.002	31.488	0.000
		7	0.147	0.026	33.316	0.000
		8	-0.036	0.060	33.428	0.000
		9	0.068	0.084	33.831	0.000
		10	-0.150	-0.184	35.841	0.000

Giản đồ tự tương quan này cho thấy chỉ có hệ số tự tương quan riêng (PAC_1) bậc một có ý nghĩa thống kê, vậy có thể thích hợp với mô hình AR(1).

Bước 2: Ước lượng mô hình AR(1)

■ HÌNH 8.19: **Kết quả ước lượng mô hình AR(1).**

Dependent Variable: Y
Method: Least Squares
Date: 02/24/09 Time: 17:42
Sample (adjusted): 2 75
Included observations: 74 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	115.2651	7.570277	15.22601	0.0000
Y(-1)	-0.528921	0.098905	-5.347738	0.0000

R-squared	0.284282	Mean dependent var	75.44595
Adjusted R-squared	0.274342	S.D. dependent var	13.79636
S.E. of regression	11.75251	Akaike info criterion	7.792666
Sum squared resid	9944.745	Schwarz criterion	7.854938
Log likelihood	-286.3286	Hannan-Quinn criter.	7.817507
F-statistic	28.59830	Durbin-Watson stat	1.990986
Prob(F-statistic)	0.000001		

Như vậy, phương trình ước lượng sẽ được viết như sau:

$$\hat{Y}_t = 115.2651 - 0.5289Y_{t-1}$$

Với giá trị quan sát thứ 75 (tức Y_{t-1}) là 72, thì giá trị dự báo cho quan sát thứ 76 (Y_t) sẽ là 77.19.

Bước 3: Đánh giá mô hình? Giả sử vì một lý do nào đó, ta ước lượng mô hình AR(2), thì làm sao biết mô hình nào tốt hơn? Kết quả dự báo hai mô hình AR(1) và AR(2) trên Eviews (rồi chọn <forecast> để xác định sai số dự báo) như sau:

■ HÌNH 8.20: **So sánh mô hình AR(1) và mô hình AR(2).**

AR(1)		AR(2)	
Root Mean Squared Error	13.65862	Root Mean Squared Error	13.75470
Mean Absolute Error	10.47346	Mean Absolute Error	10.59118
Mean Absolute Percentage Error	15.94804	Mean Absolute Percentage Error	16.14147
Theil Inequality Coefficient	0.089771	Theil Inequality Coefficient	0.090487
Bias Proportion	0.000001	Bias Proportion	0.000002
Variance Proportion	0.849667	Variance Proportion	0.843315
Covariance Proportion	0.150332	Covariance Proportion	0.156683

Các tiêu chí đánh giá độ chính xác của dự báo ở mô hình AR(1) nhỏ hơn ở mô hình AR(2). Như vậy, AR(1) có thể là mô hình thích hợp trong trường hợp này (bởi vì có thể có các mô hình khác tốt hơn nữa). Ngoài ra, chúng ta có thể vẽ trên cùng một đồ thị theo thời gian các giá trị dự báo của mô hình AR(1) và AR(2) với giá trị doanh số thực của công ty để kiểm tra xem mô hình nào dự báo đúng xu hướng vận động của dữ liệu doanh số trong quá khứ hơn.

MÔ HÌNH BÌNH QUÂN DI ĐỘNG

MÔ HÌNH MA(1)

Mô hình MA(1) có dạng như sau:

$$Y_t = \mu + u_t + \theta_1 u_{t-1} \qquad (8.29)$$

Trong đó, μ là giá trị trung bình của quá trình, u_t là số hạng nhiễu ngẫu nhiên tương tự như ở các mô hình hồi quy trước đây, θ_1 là hệ số ước lượng, và u_{t-1} là sai số ở giai đoạn *t-1*.

Hàm ý của mô hình MA(1) là Y_t phụ thuộc vào giá trị của sai số hiện tại và sai số quá khứ, tức tại thời điểm *t* và *t-1*. Điều này có nghĩa, Y_t phụ thuộc vài giá trị sai số trước đó chứ không phải giá trị trễ của Y_t như trong các mô hình AR. Ví dụ, khi xem giá giá cổ phiếu tại thời điểm *t*, thì các sai số này có thể đại diện cho ảnh hưởng của của các thông tin thị trường tại thời điểm *t–1* ngoài yếu tố giá của cổ phiếu trước đó.

MÔ HÌNH MA(q)

Mô hình MA(q) có dạng như sau:

$$Y_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \ldots + \theta_q u_{t-q} \qquad (8.30)$$

Mô hình này có thể được viết gọn như sau:

$$Y_t = \mu + u_t + \sum_{j=1}^{q} \theta_j u_{t-j} \qquad (8.31)$$

Điều này có nghĩa, giá trị Y tại thời điểm *t* không chỉ phụ thuộc vào các thông tin hiện tại mà còn phụ thuộc vào các thông tin trong quá khứ. Tuy nhiên, các thông tin gần nhất có ý nghĩa nhiều hơn so với các thông tin trước đó. Như vậy, các mô hình MA cung cấp giá trị dự báo của Y_t trên cơ sở một kết hợp tuyến tính của các giá trị sai số quá khứ, trong khi đó, các mô hình AR dự báo Y_t như một hàm tuyến tính của các giá trị quá khứ của bản thân Y_t.

Các phương trình (8.30) và (8.31) có thể được viết lại theo một cách khác như sau:

$$Y_t - \mu = u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \ldots - \theta_q u_{t-q}$$

$$Y_{t+1} - \mu = u_{t+1} - \theta_1 u_t - \theta_2 u_{t-1} - \ldots - \theta_q u_{t-q+1}$$

Nói cách khác, độ lệch của Y_t là một hàm tuyến tính của các sai số hiện tại và quá khứ.

Để xác định độ trễ q ta sử dụng giản đồ tự tương quan theo cách sau đây: ACF sẽ có xu hướng khác không một cách có ý nghĩa thống kê cho đến độ trễ q và sẽ bằng không ngay sau độ trễ q đó. Điều này có nghĩa rằng, nếu chuỗi thời gian Y_t là một chuỗi theo MA(2) thì các hệ số ACF_1 và ACF_2 có ý nghĩa thống kê, và các hệ số khác không có ý nghĩa thống kê. Trong khi đó, PACF sẽ có xu hướng bằng không ngay lập tức.

Thông thường, ít có chuỗi thời gian nào thỏa mãn các điều kiện của mô hình AR(p) hoặc MA(q), mà thường là kết hợp của hai mô hình này, có nghĩa là một chuỗi dừng thì có thể tuân theo mô hình tổng quát là ARMA(p,q).

VÍ DỤ MINH HỌA

Cũng sử dụng tập tin **DATA8-2** và thực hiện các bước sau:

Bước 1: Vẽ giản đồ tự tương quan (như ở Hình 8.18)

■ HÌNH 8.18: **Giản đồ tự tương quan của doanh số Y.**

Date: 02/24/09 Time: 17:35
Sample: 1 75
Included observations: 75

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.528	-0.528	21.796	0.000
		2	0.282	0.003	28.068	0.000
		3	-0.038	0.155	28.186	0.000
		4	0.008	0.065	28.191	0.000
		5	0.144	0.189	29.906	0.000
		6	-0.137	0.002	31.488	0.000
		7	0.147	0.026	33.316	0.000
		8	-0.036	0.060	33.428	0.000
		9	0.068	0.084	33.831	0.000
		10	-0.150	-0.184	35.841	0.000

Giản đồ tự tương quan cho thấy có thể hai hệ số tự tương quan AC_1 và AC_2 khác không một cách có ý nghĩa thống kê. Tuy nhiên, để biết mô hình MA(2) hay MA(1) phù hợp hơn, ta nên thực hiện cả hai, rồi so sánh kết quả.

Bước 2: Ước lượng mô hình MA(1)

HÌNH 8.21: Kết quả ước lượng mô hình MA(1).

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	75.28428	0.857488	87.79629	0.0000
MA(1)	-0.404556	0.109535	-3.693390	0.0004
R-squared	0.213939	Mean dependent var		75.24000
Adjusted R-squared	0.203171	S.D. dependent var		13.81841
S.E. of regression	12.33504	Akaike info criterion		7.889070
Sum squared resid	11107.19	Schwarz criterion		7.950870
Log likelihood	-293.8401	Hannan-Quinn criter.		7.913746
F-statistic	19.86807	Durbin-Watson stat		2.234567
Prob(F-statistic)	0.000029			
Inverted MA Roots	.40			

Bước 3: Ước lượng mô hình MA(2)

HÌNH 8.22: Kết quả ước lượng mô hình MA(2).

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	75.42991	1.062526	70.99110	0.0000
MA(1)	-0.571434	0.109267	-5.229707	0.0000
MA(2)	0.363663	0.113213	3.212211	0.0020
R-squared	0.311809	Mean dependent var		75.24000
Adjusted R-squared	0.292693	S.D. dependent var		13.81841
S.E. of regression	11.62150	Akaike info criterion		7.782768
Sum squared resid	9724.262	Schwarz criterion		7.875468
Log likelihood	-288.8538	Hannan-Quinn criter.		7.819782
F-statistic	16.31106	Durbin-Watson stat		1.939571
Prob(F-statistic)	0.000001			
Inverted MA Roots	.29+.53i	.29-.53i		

Bước 4: So sánh kết quả

■ HÌNH 8.23: **So sánh mô hình MA(1) và mô hình MA(2).**

MA(1)		MA(2)	
Root Mean Squared Error	13.69846	Root Mean Squared Error	13.69269
Mean Absolute Error	10.56956	Mean Absolute Error	10.55208
Mean Absolute Percentage Error	16.05280	Mean Absolute Percentage Error	16.06173
Theil Inequality Coefficient	0.090168	Theil Inequality Coefficient	0.090038
Bias Proportion	0.000226	Bias Proportion	0.000014
Variance Proportion	0.946022	Variance Proportion	0.942522
Covariance Proportion	0.053752	Covariance Proportion	0.057465

Như vậy, mô hình MA(2) có vẻ tốt hơn mô hình MA(1) vì các sai số dự báo nhỏ hơn một chút. Điều này phù hợp với thông điệp mà các độ trễ AC_1 và AC_2 trên đưa ra.

Và phương trình dự báo được viết lại như sau:

$$\hat{Y}_t = 75.43 - 0.57e_{t-1} + 0.36e_{t-2}$$

Kết quả dự báo như sau:

■ BẢNG 8.1: **Giá trị dự báo mô hình MA(2).**

Ký hiệu	T	Y_t	$\hat{Y}_t$	e_t
t-6	70	73.50	69.20	4.30
t-5	71	90.00	76.12	13.88
t-4	72	78.00	69.06	8.94
t-3	73	87.00	75.37	11.63
t-2	74	99.00	72.03	26.97
t-1	75	72.00	64.25	7.75
T	**76**		**80.72**	

MÔ HÌNH ARMA

Nếu kết hợp mô hình AR(p) với mô hình MA(q) ta có mô hình ARMA(p,q) có dạng như sau:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_2 Y_{t-p} + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \ldots + \theta_q u_{t-q} \quad (8.32)$$

Dạng rút gọn của ARMA(p,q) như sau:

$$Y_t = \phi_0 + \sum_{i=i}^{p} \phi_i Y_{t-i} + u_t + \sum_{j=1}^{q} \theta_j u_{t-j} \quad (8.33)$$

Tương tự như các mô hình AR(p) và MA(q), các mô hình ARMA(p,q) chỉ thích hợp cho các chuỗi dừng. Trong trường hợp này, ta cần phải xác định độ trễ *p* và *q* thích hợp theo cách như đã trình bày ở các phần trên.

MÔ HÌNH ARIMA

Các mô hình ARMA chỉ có thể được thực hiện khi chuỗi Y_t là chuỗi dừng. Tuy nhiên, hầu hết các chuỗi thời gian liệu kinh tế và tài chính đều là các chuỗi có yếu tố xu thế, nghĩa là, giá trị trung bình của Y_t trong năm này có thể khác năm kia. Nói cách khác, các chuỗi thời gian trong kinh tế và tài chính thường là các chuỗi không dừng. Chính vì thế, để suy ra các chuỗi dừng chúng ta phải khử yếu tố xu thế trong các chuỗi dữ liệu gốc thông qua quy trình lấy sai phân. Nếu lấy sai phân bậc 1 ta có được chuỗi dừng thì được gọi là dừng sai phân bậc 1, và ký hiệu là I(1). Mở rộng ra, nếu một chuỗi dừng ở sai phân bậc *d*, ta có ký hiệu là I(d). Như vậy mô hình ARIMA được ký hiệu chung là ARIMA(p,d,q).

QUY TRÌNH LỰA CHỌN MÔ HÌNH ARIMA(p,d,q)

Thông thường quy trình lựa chọn mô hình ARIMA thông qua ba bước sau: Nhận dạng, ước lượng, và kiểm tra chẩn đoán.

Nhận dạng

Bước 1: Thống kê mô tả để kiểm tra xem dữ liệu có những yếu tố bất thường outliers hay không, thiếu dữ liệu hay có thay đổi cấu trúc hay không.

Bước 2: Kiểm tra xem dữ liệu có dừng hay không (giản đồ tự tương quan hay kiểm định nghiệm đơn vị)? Nếu không dừng thì cách thường sử dụng là lấy sai phân bậc 1.

Bước 3: Khi đã chuyển sang chuỗi dừng, bước quan trọng tiếp theo là xác định p và q.

- Đối với mô hình MA(q) thuần túy, ACF sẽ có xu hướng khác không một cách có ý nghĩa thống kê cho đến độ trễ q và sẽ bằng không ngay sau độ trễ q đó. Trong khi đó, PACF sẽ có xu hướng bằng không ngay lập tức.
- Đối với mô hình AR(p) thuần túy, ACF sẽ có xu hướng bằng không ngay lập tức, trong khi đó, PACF sẽ có xu hướng khác không một cách có ý nghĩa thống kê cho đến độ trễ p và sẽ bằng không ngay sau độ trễ p đó.

Nếu cả p và q đều khác không, ta phải sử dụng mô hình kết hợp (ARMA) cho dữ liệu đã chuyển đổi sang chuỗi dừng. Trong trường hợp này, ta khó xác định chính xác số bậc của AR và MA, nên ta phải sử dụng nhiều mô hình khác nhau (trên cơ sở AR và MA thuần túy) và sẽ tiến hành so sánh lựa chọn.

Dạng ACF và PACF cho các mô hình ARMA(p,q) có thể có được tóm tắt như sau:

■ BẢNG 8.2: **Lựa chọn độ trễ phù hợp.**

Mô hình	ACF	PACF
MA(1)	Có ý nghĩa ở độ trễ thứ nhất	Bằng không ngay lập tức
AR(1)	Bằng không ngay lập tức	Có ý nghĩa ở độ trễ thứ nhất

ARMA(1,1)	Bằng không sau độ trễ thứ nhất	Bằng không sau độ trễ thứ nhất
ARMA(p,q)	Bằng không sau độ trễ thứ q	Bằng không sau độ trễ thứ p

Ước lượng

Ước lượng từng mô hình có thể có:

- Sử dụng các tiêu chí AIC và SBC để so sánh giữa các mô hình.
- Kiểm tra dấu và thống kê t của từng hệ số.

Phân tích chuẩn đoán

- Vẽ đồ thị phần dư theo thời gian hoặc đồ thị tần suất
- Kiểm tra tính ngẫu nhiên của phần dư bằng giản đồ tự tương quan
- Quan sát và so sánh đồ thị giá trị dự báo với giá trị thực tế
- Các kiểm định thống kê khác
- Kiểm tra sai số dự báo

Quy trình sáu bước của Box – Jenkins

Bước 1: Tính ACF và PACF của dữ liệu gốc, kiểm tra xem chuỗi gốc có dừng hay không. Nếu dừng, chuyển tới bước 3.

Bước 2: Lấy log rồi lấy sai phân bậc một của dữ liệu gốc, rồi tính ACF, PACF của dữ liệu chuyển đổi này. Tuy nhiên, nếu dữ liệu gốc ít biến động, ta có thể lấy sai phân trực tiếp mà không nhất thiết phải chuyển sang dạng logarith. Lưu ý, việc lấy logarith nhằm mục đích hỗ trợ nhận dạng các độ trễ p và q dễ dàng hơn. Sau khi đã xác định độ trễ thích hợp, chúng ta có thể ước lượng dạng sai phân bậc một của dữ liệu gốc mà không nhất thiết phải ước lượng theo sai phân bậc một của logarith của dữ liệu gốc.

Bước 3: Phân tích giản đồ tự tương quan để xác định các mô hình có thể có. Lưu ý rằng, dữ liệu thời gian phải đủ lớn để có thể giúp người phân tích nhìn nhận đồ thị một cách chính xác. Theo kinh nghiệp, để áp dụng các mô hình ARIMA, đòi hỏi dữ liệu phải có ít nhất 50 quan sát. Tuy nhiên, có trường hợp cần đến 80 hoặc 90 quan sát. Ngoài ra, chúng ta không nhất thiết phải có thật nhiều quan sát vì những quan sát rất xa hiện tại có thể không có ý nghĩa nhiều trong phân tích. Nếu gặp những trường hợp như vậy, chúng ta nên khảo sát từng 'đoạn' dữ liệu trước khi xác định một mẫu thích hợp nhất. Đối với một số chuỗi thời gian như doanh số, tiêu dùng, sản lượng công nghiệp, GDP, v.v..., thường chịu ảnh hưởng mùa vụ, chúng ta thường sử dụng các mô hình ARIMA điều chỉnh mùa vụ, được gọi tắt là các mô hình SARIMA.

Bước 4: Ước lượng các mô hình dự kiến.

Bước 5: Đối với mỗi mô hình được ước lượng:

- Kiểm tra hệ số của độ trễ cao nhất xem có ý nghĩa thống kê hay không. Nếu không, giảm bớt độ trễ của p hoặc q.
- Kiểm tra ACF và PACF đối với phần dư. Nếu mô hình đúng, thì các ACF và PACF của phần dư không có ý nghĩa thống kê.
- Kiểm tra AIC, SBC, và R^2 điều chỉnh để xem mô hình nào phù hợp nhất.
- So sánh các sai số dự báo.
- Phân tích đồ thị phần dư (đồ thị tần suất, giản đồ tự tương quan)
- Phân tích đồ thị giá trị dự báo và giá trị thực tế (lưu ý đến các bước ngoặt quan trọng trong dữ liệu, đặc biệt là giai đoạn gần hiện tại).
- Khi đánh giá các mô hình ARIMA, người phân tích nên so sánh giữa các mô hình với nhau, chứ không nên phân tích một cách riêng lẻ.

Bước 6: Nếu có thay đổi gì trong mô hình gốc, hãy quay lại bước 4.

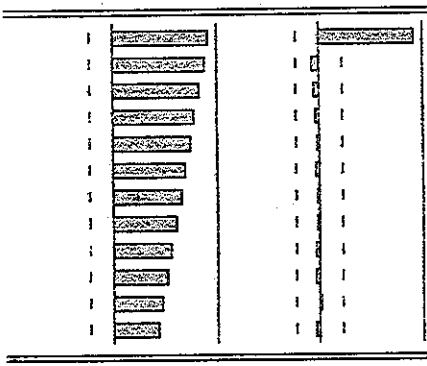
VÍ DỤ MINH HỌA

Sử dụng tập tin **DATA8-3**, trong đó có chứa các biến GDP và CPI, và thực hiện các bước sau đây:

Bước 1: Khảo sát chuỗi dữ liệu gốc

■ HÌNH 8.24: **Giản đồ tự tương quan của GDP.**

Date: 01/13/09 Time: 12:43
Sample: 1980Q3 1998Q2
Included observations: 72



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.958	0.958	68.932	0.000
		2	0.913	-0.067	132.39	0.000
		3	0.865	-0.050	190.23	0.000
		4	0.817	-0.030	242.57	0.000
		5	0.770	-0.013	289.73	0.000
		6	0.723	-0.032	331.88	0.000
		7	0.675	-0.024	369.26	0.000
		8	0.629	-0.022	402.15	0.000
		9	0.582	-0.038	430.77	0.000
		10	0.534	-0.035	455.31	0.000
		11	0.490	0.009	476.31	0.000
		12	0.446	-0.033	494.00	0.000

Như vậy, chuỗi GDP là chuỗi không dừng.

Bước 2: Lấy log và sai phân bậc 1 của log

Genr lgdp=log(gdp)

Genr dlgdp=lgdp-lgdp(-1)

■ HÌNH 8.25: **Giản đồ tự tương quan của d[log(GDP)].**

Date: 01/13/09 Time: 12:47
Sample: 1980Q3 1998Q2
Included observations: 71

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.420	0.420	13.080	0.000
		2	0.177	0.001	15.436	0.000
		3	0.283	0.252	21.520	0.000
		4	0.286	0.103	27.840	0.000
		5	0.201	0.046	31.023	0.000
		6	0.216	0.095	34.740	0.000
		7	0.039	-0.192	34.866	0.000
		8	-0.089	-0.147	35.512	0.000
		9	-0.050	-0.073	35.718	0.000
		10	0.006	0.015	35.721	0.000
		11	-0.127	-0.111	37.121	0.000
		12	-0.259	-0.163	43.011	0.000

Như vậy, sai phân bậc 1 của GDP là một chuỗi dừng. Và bây giờ chúng ta sẽ xác định mô hình ARIMA phù hợp để dự báo sai phân của GDP, rồi sau đó sẽ dự báo GDP từ giá trị dự báo của sai phân GDP.

Bước 3: Xác định p và q

p = 1 và q = 3

Bước 4: Ước lượng 3 mô hình có thể có

Ls dlgdp c ar(1) ma(1) ma(2) ma(3)

■ HÌNH 8.26: **Kết quả ước lượng mô hình ARMA(1,3).**

Dependent Variable: DLGDP
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.006822	0.001488	4.585603	0.0000
AR(1)	0.676880	0.155586	4.350514	0.0000
MA(1)	-0.361690	0.185546	-1.949334	0.0556
MA(2)	-0.212155	0.134013	-1.583088	0.1183
MA(3)	0.277806	0.121230	2.291551	0.0252

R-squared	0.269960	Mean dependent var	0.006335
Adjusted R-squared	0.225034	S.D. dependent var	0.006323
S.E. of regression	0.005567	Akaike info criterion	-7.475335
Sum squared resid	0.002014	Schwarz criterion	-7.314728
Log likelihood	266.6367	Hannan-Quinn criter.	-7.411540
F-statistic	6.009048	Durbin-Watson stat	1.905971
Prob(F-statistic)	0.000354		

Ls dlgdp c ar(1) ma(1) ma(2)

■ HÌNH 8.27: **Kết quả ước lượng mô hình ARMA(1,2).**

Dependent Variable: DLGDP

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.006795	0.001367	4.972132	0.0000
AR(1)	0.692728	0.202898	3.414165	0.0011
MA(1)	-0.254315	0.243568	-1.044124	0.3002
MA(2)	-0.162587	0.160778	-1.011253	0.3156

R-squared	0.218823	Mean dependent var	0.006335
Adjusted R-squared	0.183315	S.D. dependent var	0.006323
S.E. of regression	0.005714	Akaike info criterion	-7.436204
Sum squared resid	0.002155	Schwarz criterion	-7.307719
Log likelihood	264.2671	Hannan-Quinn criter.	-7.385168
F-statistic	6.162620	Durbin-Watson stat	2.073474
Prob(F-statistic)	0.000930		

Ls dlgdp c ar(1) ma(1)

■ HÌNH 8.28: **Kết quả ước lượng mô hình ARMA(1,1).**

Dependent Variable: DLGDP

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.006726	0.001321	5.092872	0.0000
AR(1)	0.631409	0.161374	3.912704	0.0002
MA(1)	-0.304829	0.216558	-1.407604	0.1639

R-squared	0.204555	Mean dependent var	0.006335
Adjusted R-squared	0.180810	S.D. dependent var	0.006323
S.E. of regression	0.005723	Akaike info criterion	-7.446676
Sum squared resid	0.002195	Schwarz criterion	-7.350312
Log likelihood	263.6337	Hannan-Quinn criter.	-7.408399
F-statistic	8.614790	Durbin-Watson stat	1.930923
Prob(F-statistic)	0.000468		

Bước 5: Phân tích chẩn đoán và lựa chọn mô hình

Việc chẩn đoán mô hình nhằm chứng minh phần dư của mô hình tuân thủ tính chất nhiễu trắng. Lựa chọn mô hình thì chúng ta sẽ căn cứ vào các tiêu thức đánh giá sai số dự báo như ý nghĩa các hệ số, AIC, RMSE, đồ thị phần dư, đồ thị giá trị dự báo và giá trị thực tế, v.v... Kết quả cuối cùng cho thấy, mô hình ARMA(1,1) là mô hình tốt nhất để dự báo sai phân của GDP. Nói cách khác, chúng ta đã dự báo GDP theo mô hình ARIMA(1,1,1).

CÁC TIÊU CHÍ LỰA CHỌN MÔ HÌNH ARIMA

Sử dụng tập tin **DATA8-1**, khảo sát mô hình ARIMA phù hợp cho việc dự báo biến GDP theo các bước như sau:

Bước 1: Kiểm tra tính dừng của biến GDP bằng giản đồ tự tương quan và kiểm định nghiệm đơn vị. Kết quả khảo sát cho thấy chuỗi GDP là một chuỗi không dừng.

Bước 2: Tạo biến sai phân bậc một của GDP và đặt tên là dGDP. Kết quả kiểm định cho thấy dGDP là một chuỗi dừng.

Bước 3: Xác định các độ trễ p và q cho mô hình ARMA với chuỗi dGDP. Kết quả khả sát cho thấy các độ trễ khả dĩ của p và q như sau:

p = 1, 8, và 12

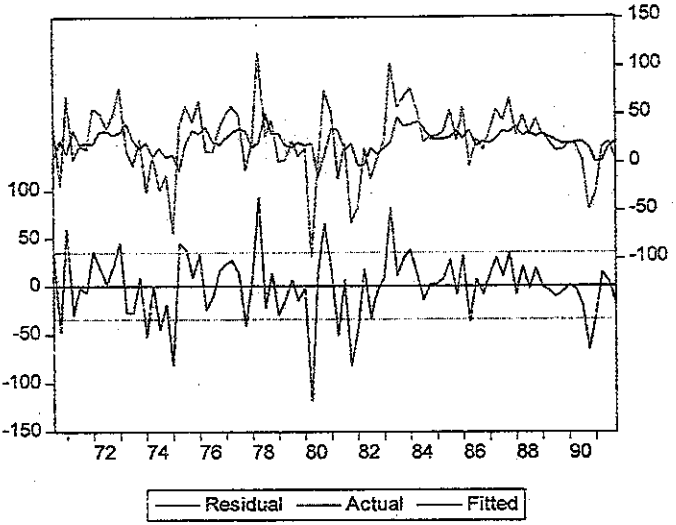
q = 1, 8, và 12

Bước 4: Giả sử ta thực hiện và so sánh kết quả ước lượng hai mô hình ARMA sau đây:

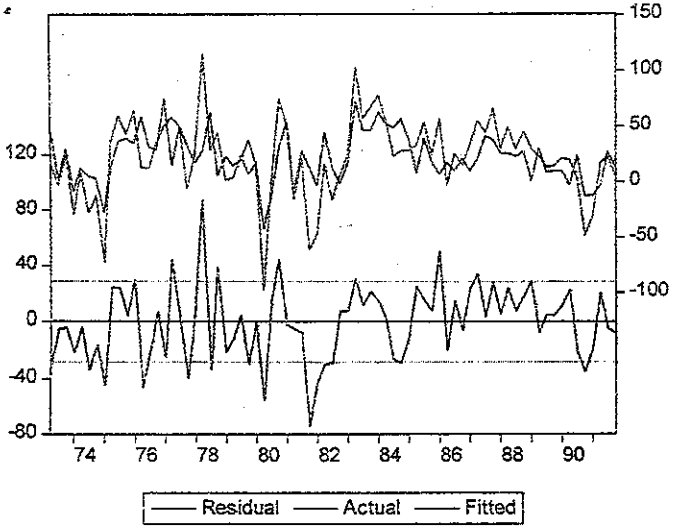
Mô hình 1: dGDP c AR(1) MA(1)

Mô hình 2: dGDP c AR(1) AR(8) AR(12) MA(1) MA(8) MA(12)

■ HÌNH 8.29: **Đồ thị giá trị dự báo của mô hình 1.**



■ HÌNH 8.30: **Đồ thị giá trị dự báo của mô hình 2.**



Sử dụng các tiêu chí sau đây để đánh giá lựa chọn hai mô hình ARMA vừa ước lượng:

(1) Phần dư của mô hình dự báo phải là một chuỗi ngẫu nhiên

Sau khi ước lượng mô hình ARIMA, chúng ta cần kiểm tra phần dư (RESID, sai số dự báo) có phải là một chuỗi ngẫu nhiên hay không bằng cách sử dụng giản đồ tự tương quan (Quick/Series Statistics/Correlogram). Kết quả khảo sát giản đồ tự tương quan cho thấy phần dư của mô hình 1 chưa phải là một chuỗi ngẫu nhiên, ngược lại, phần dư của mô hình 2 là một chuỗi ngẫu nhiên.

(2) Tiêu chí AIC/SBC/HQ

Mô hình nào có giá trị AIC/SBC/HQ nhỏ nhất là mô hình tốt nhất. Lưu ý, các mô hình có biến phụ thuộc dạng logarith thì giá trị AIC/SBC/HQ âm. Cho nên, chúng ta cần chuyển sang dạng exp(AIC/SBC/HQ) để kiểm tra mô hình nào có giá trị AIC/SBC/HQ bé nhất. Mô hình 1 có AIC = 9.95 và mô hình 2 có AIC = 9.62.

(3) Sai số dự báo càng nhỏ càng tốt

Trong Eviews ta hay sử dụng tiêu chí RMSE (bằng cách chọn Forecast/Dynamic hoặc Forecast/Static)ngay sau kết quả hồi quy trên Eviews. Mô hình 1 có RMSE = 35.78 và mô hình 2 có RMSE = 34.55.

(4) So sánh giá trị dự báo với giá trị thực tế

Nếu mô hình có giá trị dự báo càng gần với giá trị thực tế thì đó là mô hình dự báo tốt. Ở đây, có hai điểm cần lưu ý: (i) Xem các bước ngoặc, (ii) Xem xu hướng của giá trị dự báo và giá trị thực ở các giai đoạn gần hiện tại nhất. Sau khi dự báo, chúng ta vào View/Actual,Fitted,Residual/Actual,Fitted,Residual Graph. Nhìn vào giá trị dự báo ở Hình 8.29 và 8.30 ta thấy đường dự báo (fitted) ở mô hình 2 gần với giá trị thực hơn.

(5) Hệ số hồi quy có ý nghĩa thống kê hay không

Mô hình nào có tất cả các hệ số hồi quy (AR, MA)có ý nghĩa thống kê ở mức ý nghĩa được chọn thì mô hình đó tốt hơn. Mặc dù mô hình 2 có vẽ tốt hơn mô hình 1, nhưng vẫn còn một số hệ số ước lượng không có ý nghĩa thống kê. Theo chiến lược lựa chọn mô hình của Hendry, chúng ta có thể loại bỏ dần các độ trễ không có ý nghĩa thống kê ra khỏi mô hình 2.

Bước 5: Giả sử mô hình 2 là mô hình tốt nhất, chúng ta sẽ sử dụng mô hình này cho mục đích dự báo GDP ở giai đoạn t+1 như sau:

(1) Dự báo giá trị dGDP ở giai đoạn t+1 ($dGDP_{t+1}$) bằng các bước như sau: (a) Nhấp vào <Range>, mở rộng dữ liệu thêm một quan sát, (b) Từ kết quả ước lượng, vào <Forecast>, chọn <Dynamic>, <OK>. Như vậy, Eviews sẽ tạo ta một biến dGDPf, trong đó có chứa giá trị dGDP của giai đoạn t+1.

(2) Dự báo GDP giai đoạn t+1 (GDP_{t+1}) theo công thức sau: $GDP_{t+1} = GDP_t + dGDP_{t+1}$.

ƯỚC LƯỢNG MÔ HÌNH ARIMA TRÊN THỰC TẾ

Theo kinh nghiệm, chúng ta nên chọn các độ trễ p, q, sao cho các giá trị AC hoặc PAC nằm ngoài đường diền trong giản đồ tự tương quan thì đó là mô hình ARIMA tốt nhất. Nếu dữ liệu có yếu tố mùa (ví dụ dữ liệu theo quý), thì chúng ta có thể sử dụng mô hình SARIMA trong đó có tính các độ trễ theo quý (xem tập tin **DATA8-4**).

TÓM TẮT CHƯƠNG 8

Phân tích dữ liệu thường có hai mục đích chính là đưa ra các gợi ý chính sách hoặc là dự báo. Mục đích đầu tiên thì dựa vào chủ yếu các mô hình kinh tế lượng nhân quả hoặc hành vi, ví dụ GDP chịu ảnh hưởng bởi chính sách FDI hoặc môi trường đầu tư. Nếu các biến chính sách FDI hoặc môi trường đầu tư tác động có ý nghĩa thống kê đến GDP thì các gợi ý chính sách sẽ được hoạch định nhằm tác động đến GDP. Mục đích thứ hai là đưa ra các kết quả dự báo dựa trên việc phân tích chính cấu trúc của một chuỗi thời gian nào đó và tìm ra mô hình diễn đạt thích hợp nhất với sai số dự báo thấp nhất. Đối với mục đích thứ hai này rất quan trọng khi áp dụng cho các biến kinh tế có độ nhạy cao như chỉ số giá chứng khoán hoặc là biến động của giá dầu thế giới và nhà kinh tế thường muốn mô hình hóa sự biến động của chúng trong tương lai dựa vào những thông tin vốn có của chúng được lưu trữ trong quá khứ. Mô hình ARIMA là mô hình khá thích hợp khi sử dụng cho các chuỗi thời gian có độ nhạy như vậy trong mục đích dự báo. Bước cơ bản đầu tiên của mô hình này là tìm ra một chuỗi thời gian 'dừng' từ chuỗi thời gian ban đầu bằng các kiểm định thống kê, hoặc bằng các bước chuyển sai phân hoặc log. Sau đó xác định dạng thích hợp tự hồi quy của nó (AR) kết hợp với trung bình di động (MA) qua phân tích ACF và PACF để tìm ra mô hình dự báo ARMA dựa trên chuỗi thỏa điều kiện dừng. Mô hình ARIMA có ưu điểm là mô hình hóa gần như tất cả các dao động của chuỗi thời gian ban đầu, có nghĩa là nó dự báo tốt hơn nhiều so với các mô hình kinh tế lượng nhân quả hoặc xu thế đã thảo luận trước đây. Tuy vậy, mô hình ARIMA vẫn còn một nhược điểm là nó chưa tính đến yếu tố mùa của bản thân chuỗi thời gian, một mô hình nâng cao xử lý yếu tố mùa áp dụng dựa trên nền tảng lý thuyết chuỗi dừng là mô hình SARIMA sẽ được nghiên cứu sau này.

CÂU HỎI VÀ BÀI TẬP

1. Dữ liệu trong tập tin **"ARIMA1.xls"** chứa dữ liệu của một chuỗi thời gian có 126 quan sát. Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo cho quan sát 127, 128, và 129?

2. Dữ liệu trong tập tin **"ARIMA2.xls"** chứa dữ liệu của một chuỗi thời gian có 80 quan sát. Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo cho quan sát 81, 82, và 83?

3. Dữ liệu trong tập tin **"ARIMA3.xls"** chứa dữ liệu của một chuỗi thời gian có 80 quan sát. Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo cho quan sát 81, 82, và 83?

4. Dữ liệu trong tập tin **"ARIMA4.xls"** chứa dữ liệu theo tháng của một chuỗi thời gian có 96 quan sát. Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo cho 12 quan sát của năm tiếp theo?

5. Dữ liệu trong tập tin **"ARIMA5.xls"** chứa dữ liệu theo tháng của một chuỗi thời gian có 120 quan sát. Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo cho 12 tháng của năm tiếp theo?

6. Dữ liệu trong tập tin **"ARIMA6.xls"** chứa các dữ liệu về chỉ số giá chứng khoán của thị trường chứng khoán TP.HCM (VNI) và thị trường chứng khoán Hà Nội (HNX). Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo chỉ số VNI và HNX cho 6 ngày tiếp theo?

7. Dữ liệu trong tập tin **"ARIMA7.xls"** chứa các dữ liệu về giá chứng khoán của một số cổ phiếu Blue-Chip trên thị trường chứng khoán TP.HCM. Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo giá của các cổ phiếu này cho 6 ngày tiếp theo?

8. Dữ liệu trong tập tin **"ARIMA8.xls"** chứa dữ liệu theo tuần về doanh số của một nhà hàng. Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo cho 4 tuần tiếp theo?

9. Dữ liệu trong tập tin **"ARIMA9.xls"** chứa dữ liệu theo tháng về sản lượng bia của một hãng bia. Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo sản lượng bia cho 12 tháng của năm tiếp theo?

10. Dữ liệu trong tập tin **"ARIMA10.xls"** chứa dữ liệu về một số loại lãi suất. Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo từng loại lãi suất cho 3 giai đoạn tiếp theo?

11. Sử dụng tập tin **"PRICE.xls"**, Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo giá vàng thế giới, giá vàng Việt Nam, giá dầu thế giới, giá cao su, giá cà phê, và tỷ giá hối đoái VND/USD cho 3 giai đoạn tiếp theo? Anh/Chị cho biết kết quả này có tốt hơn các kết quả trước đây hay không? Tại sao?

12. Sử dụng tập tin **"VIETNAM.xls"**, Anh/Chị hãy dự báo số lượng du khách, CPI, doanh số bán lẻ, sản lượng công nghiệp, xuất khẩu, nhập khẩu, nhập khẩu, và FDI của Việt Nam 6 tháng đầu năm 2009?

13. Sử dụng tập tin **"GAS.xls"**, Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo giá CP cho tháng 6/2009? Anh/Chị cho biết kết quả này có tốt hơn các kết quả trước đây hay không? Tại sao?

14. Sử dụng tập tin **"GAP.xls"**, Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo doanh số của GAP cho các tháng trong năm 2004? Anh/Chị cho biết kết quả này có tốt hơn các kết quả trước đây hay không? Tại sao?

15. Sử dụng tập tin **"MURPHY.xls"**, Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo doanh số bán lẻ toàn quốc trong năm 1996? Anh/Chị cho biết kết quả này có tốt hơn các kết quả trước đây hay không? Tại sao?

16. Sử dụng tập tin **"CCC.xls"**, Anh/Chị hãy xây dựng mô hình ARIMA phù hợp để dự báo lượng khách hàng mới cho tháng sau? Anh/Chị cho biết kết quả này có tốt hơn các kết quả trước đây hay không? Tại sao?

CHƯƠNG 9 CÁC MÔ HÌNH ARCH/GARCH VÀ DỰ BÁO RỦI RO

Sự phát triển ứng dụng công cụ kinh tế lượng trong lĩnh vực tài chính đã giới thiệu nhiều mô hình và kỹ thuật phân tích giúp chúng ta không những có thể dự báo hành vi của những nhà đầu tư qua suất sinh lợi kỳ vọng, mà còn dự báo rủi ro bằng các chỉ báo phương sai hay độ lệch chuẩn. Nhiều mô hình định giá tài sản đã nỗ lực ước lượng suất sinh lợi kỳ vọng của một tài sản cụ thể (ví dụ cổ phiếu của một công ty), và ứng với mỗi suất sinh lợi kỳ vọng đều bao hàm yếu tối rủi ro hệ thống và rủi ro phi hệ thống. Với thực tiễn như vậy, các mô hình kinh tế lượng và dự báo đòi hỏi phải có khả năng dự báo mức độ dao động của các chuỗi thời gian. Các mô hình dự báo như vậy thuộc nhóm các mô hình ARCH (Autogressive Conditional Heteroskedasticity) và chúng sẽ được đề cập trong chương này. Trong những năm gần đây, các mô hình ARCH đã được nhiều nhà nghiên cứu sử dụng để ước lượng các nhân tố ảnh hưởng đến rủi ro của các tài sản tài chính trên thị trường chứng khoán, thị trường vàng, thị trường dầu, thị trường bất động sản, và nhiều thị trường cao cấp khác nhằm cung cấp thông tin cho các quyết định kinh doanh, và đặc biệt là trong quản trị rủi ro.

MỤC TIÊU HỌC TẬP

Chương này sau khi nghiên cứu chúng ta sẽ có thể dự báo rủi ro các biến số kinh tế và tài chính có độ dao động cao. Các mô hình dự báo không còn đơn thuần là dự báo giá trị trung bình nữa mà còn tiến tới dự báo rủi ro cho các biến số này. Các mô hình dự báo rủi ro với sự hỗ trợ của phần mềm Eviews trong chương này bao gồm:

- Mô hình ARCH
- Mô hình ARCH(1)
- Mô hình ARCH(q)
- Mô hình GARCH(p,q)
- Mô hình GARCH-M
- Mô hình TGARCH
- Ý tưởng về các mô hình ARCH mở rộng (mô hình hóa nhân tố ảnh hưởng đến rủi ro)

GIỚI THIỆU Ý TƯỞNG CỦA CÁC MÔ HÌNH ARCH

Chúng ta đã biết rằng, phân tích kinh tế lượng cổ điển đều giả định phương sai của sai số là không đổi theo thời gian. Tuy nhiên, các chuỗi dữ liệu về tài chính và kinh tế thường có xu hướng dao động cao vào một số giai đoạn theo sau một số giai đoạn tương đối ít biến động. Trong tài chính, người ta cho rằng có sự dao động như vậy là do bất kỳ một chuỗi thời gian nào đều chịu ảnh hưởng ít nhiều của các tin tức tốt và xấu có liên quan và các nhà đầu tư trên thị trường đều ứng xử theo kiểu hành vi đám đông. Cho nên, giả định phương sai không đổi theo thời gian thường không còn phù hợp đối với các dữ liệu chuỗi thời gian.

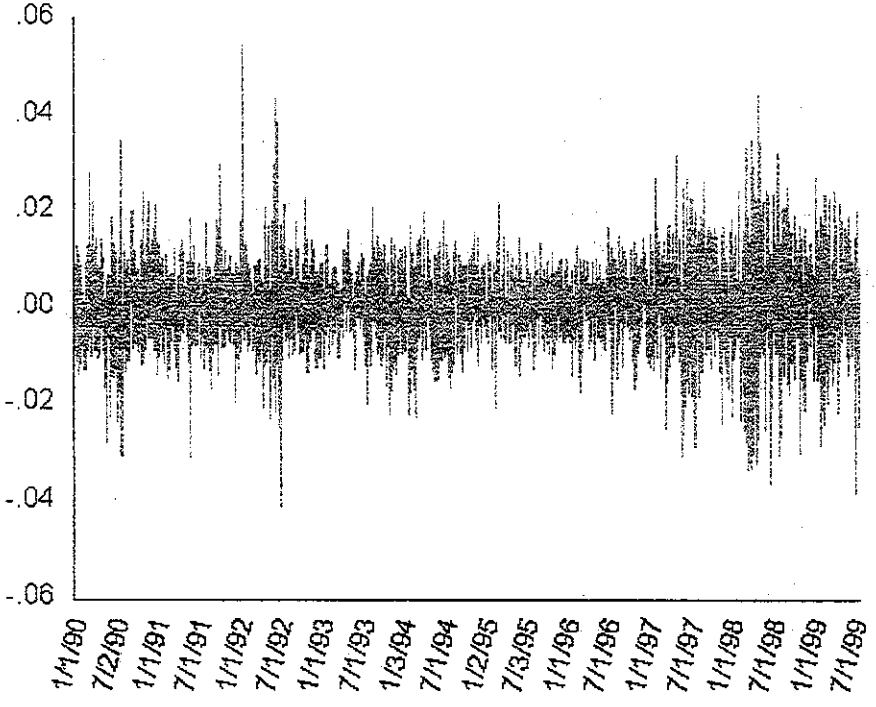
Hình 9.1 minh họa xu hướng vận động của một chuỗi dữ liệu tài chính (suất sinh lợi hàng ngày của cổ phiếu ABC giai đoạn từ ngày 1 tháng 1 năm 1990 đến ngày 31 tháng 12 năm 1999, **DATA9-1**). Ở đồ thị này, chúng ta nhận thấy rằng trong một số giai đoạn suất sinh lợi của cổ phiếu ABC biến động cao hơn (và vì thế rủi ro sẽ cao hơn) so với các giai đoạn khác. Điều này có nghĩa rằng giá trị kỳ vọng của độ lớn các hạng nhiễu ở các giai đoạn này lớn hơn các giai đoạn khác. Hơn nữa, các giai đoạn có rủi ro cao và thấp dường như có tính tập trung, chứ không kéo dài mãi mãi. Nói cách khác, các thay đổi lớn

trong suất sinh lợi của cổ phiếu ABC dường như được theo sau bởi những thay đổi lớn khác trước khi có xu hướng giảm xuống. Và một khi đã giảm xuống thì xu hướng này có vẽ được tiếp tục ổn định trong một thời gian nhất định.

Như vậy, trong các trường hợp như Hình 9.1 thì rõ ràng rằng giả định phương sai không đổi có vẽ không còn phù hợp. Và điều này nảy sinh ý tưởng cần phải xem xét các dạng dữ liệu trong đó cho phép phương sai của nó phụ thuộc vào các giá trị phương sai trong quá khứ (chứ không chỉ riêng giá trị trung bình như đã đề cập trong các mô hình ARIMA). Nói cách khác, tốt hơn là chúng ta nên xem xét không chỉ trường hợp phương sai không có điều kiện, mà còn trường hợp phương sai có điều kiện. Điều này có nghĩa rằng, phương sai của thời điểm t có thể phụ thuộc vào phương sai tại các thời điểm trước đó, hay còn gọi là các phương sai trễ, và hiện tượng này trong kinh tế lượng gọi là tự tương quan (Autocorrelation).

Để hiểu rõ hơn về vấn đề này, chúng ta hãy xem một nhà đầu tư dự định mua cổ phiếu ABC tại thời điểm t và bán tại thời điểm $t+1$. Đối với nhà đầu tư này, chỉ dự báo suất sinh lợi kỳ vọng của cổ phiếu ABC sẽ là chưa đủ. Thực tế, nhà đầu tư này có thể sẽ quan tâm và thực sự có quan tâm đến phương sai của suất sinh lợi sẽ như thế nào trong giai đoạn nắm dữ cổ phiếu ABC. Điều này có nghĩa, nhà đầu tư không chỉ quan tâm đến suất sinh lợi kỳ vọng, mà còn quan tâm đến mức độ rủi ro của cổ phiếu ABC. Như vậy, nhà đầu tư có thể muốn xem xét hành vi của phương sai có điều kiện của chuỗi dữ liệu cổ phiếu ABC để ước lượng mức độ rủi ro của cổ phiếu ABC trong một giai đoạn nhất định nào đó.

■ HÌNH 9.1: **Biến động của suất sinh lợi cổ phiếu ABC theo thời gian.**



CÁC MÔ HÌNH ARCH

Mô hình ARCH do Engle phát triển năm 1982. Mô hình này cho rằng phương sai của các số hạng nhiễu[1] tại thời điểm *t* phụ thuộc vào các số hạng nhiễu bình phương ở các giai đoạn trước. Engle cho rằng tốt nhất chúng ta nên mô hình hóa đồng thời giá trị trung bình và phương sai của chuỗi dữ liệu khi nghi ngờ rằng giá trị phương sai thay đổi theo thời gian. Hãy xem mô hình đơn giản sau:

$$Y_t = [\beta_1] + \beta_2 [X_t] + u_t \qquad (9.1)$$

[1] Khi ước lượng với mẫu thì chúng ta thay khái niệm hạng nhiễu bằng khái niệm phần dư.

Trong đó, $[X_t]$ là một véctơ $k \times 1$ các biến giải thích và $[\beta_2]$ là một véctơ $k \times 1$ các hệ số. Thông thường, u_t được giả định tuân theo phân phối chuẩn với trung bình bằng 0 và phương sai không đổi là σ^2. Giả định này được viết như sau:

$$u_t \sim N(0, \sigma^2) \tag{9.2}$$

Ý tưởng của Engle bắt đầu từ sự thật rằng ông cho phép phương sai của các hạng nhiễu phụ thuộc vào các giá trị quá khứ, hay phương sai thay đổi qua thời gian. Một cách để mô hình hóa ý tưởng này là cho phương sai phụ thuộc vào các biến trễ của các hạng nhiễu bình phương. Điều này có thể được minh họa như sau:

$$\sigma_t^2 = \gamma_0 + \gamma_1 u_{t-1}^2 \tag{9.3}$$

Phương trình (9.3) được gọi là quy trình ARCH(1), và ý tưởng này cũng tương tự như trong các mô hình ARIMA.

MÔ HÌNH ARCH(1)

Mô hình ARCH(1) sẽ mô hình hóa đồng thời giá trị trung bình và phương sai của một chuỗi thời gian theo cách được xác định sau đây:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{9.4}$$

$$u_t \sim N(0, h_t)$$

$$h_t = \gamma_0 + \gamma_1 u_{t-1}^2 \tag{9.5}$$

Ở đây, phương trình (9.4) được gọi là phương trình ước lượng giá trị trung bình (ví dụ suất sinh lợi kỳ vọng của cổ phiếu ABC) và phương trình (9.5) được gọi là phương trình ước lượng giá trị phương sai (ví dụ rủi ro của cổ phiếu ABC). Lưu ý, để đơn giản trong việc thể hiện công thức của phương trình phương sai, từ đây về sau chúng ta sử dụng ký hiệu h_t thay cho σ_t^2.

Mô hình ARCH(1) cho rằng khi có một cú sốc lớn xảy ra ở giai đoạn *t-1*, thì giá trị u_t (giá trị tuyệt đối hoặc bình phương) sẽ cũng lớn hơn. Nghĩa là, khi u_{t-1}^2 lớn/nhỏ, thì phương sai của u_t cũng sẽ lớn/nhỏ. Hệ số ước lượng γ_1 phải có dấu dương vì phương sai luôn dương.

MÔ HÌNH ARCH(q)

Thực tế, phương sai có điều kiện có thể phụ thuộc không chỉ một độ trễ mà còn nhiều độ trễ trước đó nữa, vì mỗi trường hợp có thể tạo ra một quy trình ARCH khác nhau.

Mô hình ARCH(2) sẽ được thể hiện như sau:

$$h_t = \gamma_0 + \gamma_1 u_{t-1}^2 + \gamma_2 u_{t-2}^2 \tag{9.6}$$

Và mô hình ARCH(3) sẽ là:

$$h_t = \gamma_0 + \gamma_1 u_{t-1}^2 + \gamma_2 u_{t-2}^2 + \gamma_3 u_{t-3}^2 \tag{9.7}$$

Và trường hợp tổng quát sẽ là ARCH(q) được thể hiện như sau:

$$h_t = \gamma_0 + \gamma_1 u_{t-1}^2 + \gamma_2 u_{t-2}^2 + ... + \gamma_q u_{t-q}^2 \tag{9.8}$$

$$= \gamma_0 + \sum_{j=1}^{q} \gamma_j u_{t-j}^2$$

Mô hình ARCH(q) sẽ mô hình hóa đồng thời giá trị trung bình và phương sai của một chuỗi theo cách như được xác định sau đây:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{9.9}$$

$$u_t \sim N(0, h_t)$$

$$h_t = \gamma_0 + \sum_{j=1}^{q} \gamma_j u_{t-j}^2 \tag{9.10}$$

Các hệ số ước lượng γ_j phải có dấu dương vì phương sai luôn dương.

KIỂM ĐỊNH ẢNH HƯỞNG ARCH

Trước khi ước lượng các mô hình ARCH(q), điều quan trọng là chúng ta cần kiểm tra xem có tồn tại các ảnh hưởng ARCH hay không để biết các mô hình nào cần ước lượng theo phương pháp ước lượng ARCH thay vì theo phương pháp ước lượng OLS. Kiểm định ảnh hưởng ARCH sẽ được thực hiện theo quy trình như sau:

Bước 1: Ước lượng phương trình trung bình (9.11) theo phương pháp OLS

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad (9.11)$$

Lưu ý, các biến giải thích có thể bao gồm các biến trễ của biến phụ thuộc và các biến giải thích khác có ảnh hưởng đến Y_t. Ngoài ra, khi thực hiện với dữ liệu mẫu, thì hạng nhiễu u_t trong mô hình (9.11), được đổi thành phần dư e_t (ở đây e_t được dùng để thay cho ký hiệu $\hat{u}_t$).

Bước 2: Ước lượng phương trình hồi quy phụ sau đây:

$$e_t^2 = \gamma_0 + \gamma_1 e_{t-1}^2 + \gamma_2 e_{t-2}^2 + ... + \gamma_q e_{t-q}^2 + w_t \qquad (9.12)$$

Xác định hệ số xác định của mô hình hồi quy phụ, đặt tên là R_{aux}^2.

Bước 3: Xác định giả thiết H_0 như sau:

$$H_0: \gamma_1 = \gamma_2 = ... = \gamma_q \qquad (9.13)$$

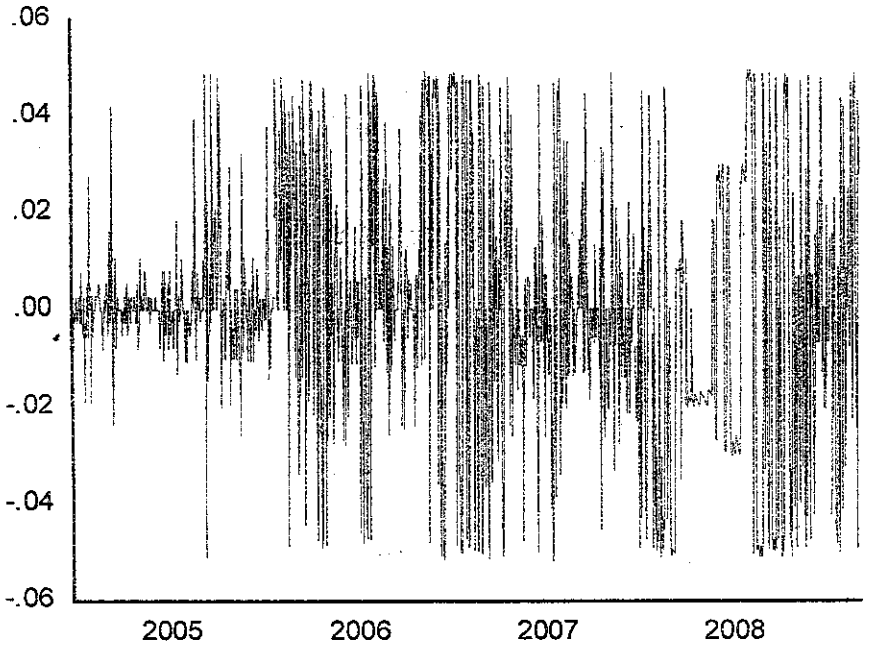
Từ kết quả hồi quy phụ (9.12), ta tính $R_{aux}^2 * T$, với T là số quan sát của chuỗi dữ liệu đang được xem xét. Thống kê này sẽ theo phân phối chi χ^2 với số bậc tự do là số độ trễ q (do e_t^2 trong phương trình (9.12) là một tổng của q thành phần lấy bình phương). Nếu giá trị thống kê χ^2 tính toán ($R^2 * T$) lớn hơn giá trị χ^2 tra bảng (theo hàm CHIINV(α,d.f.) trong excel), thì chúng ta bác bỏ giả thiết H_0, và ngược lại. Nếu bác bỏ giả thiết H_0, thì ta có thể kết luận rằng chuỗi dữ liệu đang xét có ảnh hưởng ARCH.

ƯỚC LƯỢNG CÁC MÔ HÌNH ARCH TRÊN EVIEWS

Sử dụng tập tin **DATA9-2** chứa dữ liệu theo ngày giá cổ phiếu SAM và suất sinh lợi được tính theo công thức R=log(SAM/SAM(-1)) trong giai đoạn từ ngày 28/7/2000 đến ngày 26/3/2009. Trước hết ta xem xét dạng dữ liệu của suất sinh lợi R của cổ phiếu SAM để chọn dạng mô hình phù hợp cho phương trình suất sinh lợi trung bình.

Bước 1: Vẽ đồ thị của R theo thời gian.

■ HÌNH 9.2: **Biến động của suất sinh lợi cổ phiếu SAM.**



Như vậy, suất sinh lợi R của cổ phiếu SAM có thể là một chuỗi dừng và có thể có ảnh hưởng ARCH vì các dao động của R quanh giá trị 0 không đều nhau.

Bước 2: Kiểm định tính dừng.

■ HÌNH 9.3: **Giản đồ tự tương quan của R.**

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.329	0.329	217.27	0.000
		2	0.084	-0.027	231.47	0.000
		3	0.083	0.071	245.35	0.000
		4	0.122	0.084	275.39	0.000
		5	0.104	0.041	297.16	0.000
		6	0.099	0.054	317.03	0.000

Như vậy, R là một chuỗi dừng tại các độ trễ 1 cho AR và MA. Ta có thể ước lượng thử ba mô hình sau đây để xem mô hình nào phù hợp nhất cho việc ước lượng suất sinh lợi trung bình: ARMA(1,0), ARMA(0,1) và ARMA(1,1).

Bước 3: Lựa chọn mô hình phù hợp cho suất sinh lợi trung bình.

Kết quả ước lượng cho thấy mô hình ARMA(0,1) không có hệ số truc tung có thể là mô hình phù hợp nhất cho suất sinh lợi trung bình vì sai số . Kết quả như sau:

■ HÌNH 15.4: **Kết quả mô hình ARMA(0,1).**

Dependent Variable: R
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
MA(1)	0.322559	0.021154	15.24833	0.0000

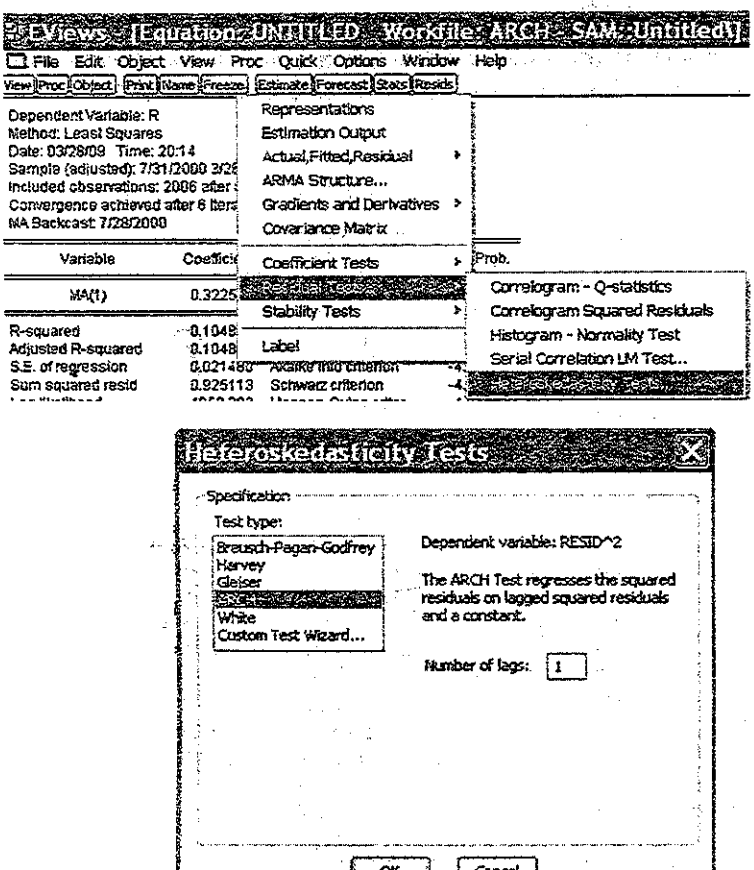
R-squared	0.104800	Mean dependent var	0.000540
Adjusted R-squared	0.104800	S.D. dependent var	0.022703
S.E. of regression	0.021480	Akaike info criterion	-4.842863
Sum squared resid	0.925113	Schwarz criterion	-4.840069
Log likelihood	4858.392	Hannan-Quinn criter.	-4.841837
Durbin-Watson stat	1.956272		

Bước 4: Kiểm tra có tồn tại các ảnh hưởng ARCH hay không.

Hình 9.2 cho thấy có vẻ như phương sai của hạng nhiễu tại thời điểm t phụ thuộc vào phương sai của hạng nhiễu ở các giai đoạn trước đó vì các dao động cao được tiếp theo bởi các dao động cao khác và ngược lại. Để kiểm tra các ảnh hưởng ARCH trên Eviews, ta thực hiện như sau:

- Ước lượng mô hình ARMA(0,1) như ở Hình 9.4
- Vào View/Residuals Tests/Heteroskedasticity Tests …

■ HÌNH 9.5: **Kiểm định ảnh hưởng ARCH trên Eviews/**



Xác định độ trễ bằng 1, rồi chọn OK, ta sẽ có kết quả hồi quy phụ như sau:

■ HÌNH 9.6: **Kiểm định ảnh hưởng ARCH(1).**

Heteroskedasticity Test: ARCH

F-statistic	486.2532	Prob. F(1,2003)	0.0000
Obs*R-squared	391.6587	Prob. Chi-Square(1)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000258	2.01E-05	12.83127	0.0000
RESID^2(-1)	0.441987	0.020044	22.05115	0.0000

R-squared	0.195341	Mean dependent var	0.000461
Adjusted R-squared	0.194939	S.D. dependent var	0.000890
S.E. of regression	0.000798	Akaike info criterion	-11.42756
Sum squared resid	0.001276	Schwarz criterion	-11.42197
Log likelihood	11458.12	Hannan-Quinn criter.	-11.42550
F-statistic	486.2532	Durbin-Watson stat	2.277987
Prob(F-statistic)	0.000000		

Giá trị Chi bình phương tính toán bằng 486,25 là quá cao so với giá trị Chi bình phương tra bảng ở mức ý nghĩa 1% với 1 bậc tự do (là 6,6349), nên ta bác bỏ giả thiết H_0. Nghĩa là, có ảnh hưởng ARCH. Tiếp tục tăng số độ trễ lên 2, 3, 4, 5, 6, và 7, ta nhận thấy rằng có thể độ trễ bậc 5 là độ trễ tối ưu, vì các hệ số ước lượng trong mô hình hồi quy phụ đều có ý nghĩa ở mức 1%, và các thống kê khác như R^2 điều chỉnh, AIC, SBC, v.v., không có khác biệt lớn so với độ trễ bằng 6. Ngoài ra, với độ trễ là 6 thì hệ số của độ trễ bậc 5 có dấu hiệu không có ý nghĩa, và khi độ trễ là 7 thì hệ số của độ trễ bậc 5 trở nên không có ý nghĩa. Lưu ý rằng, nếu ta chọn độ trễ không thích hợp thì trong kết quả ước lượng của mô hình ARCH sẽ có nhiều hệ số không có ý nghĩa thống kê. Chính vì thế, chúng ta sẽ so sánh kết quả mô hình ARCH(1) và ARCH(5) để xem nên chọn mô hình nào cho mục đích dự báo trung bình và phương sai của suất sinh lợi R. Tuy nhiên, như chúng ta sẽ biết ở phần sau, việc sử dụng quá nhiều độ trễ không phải

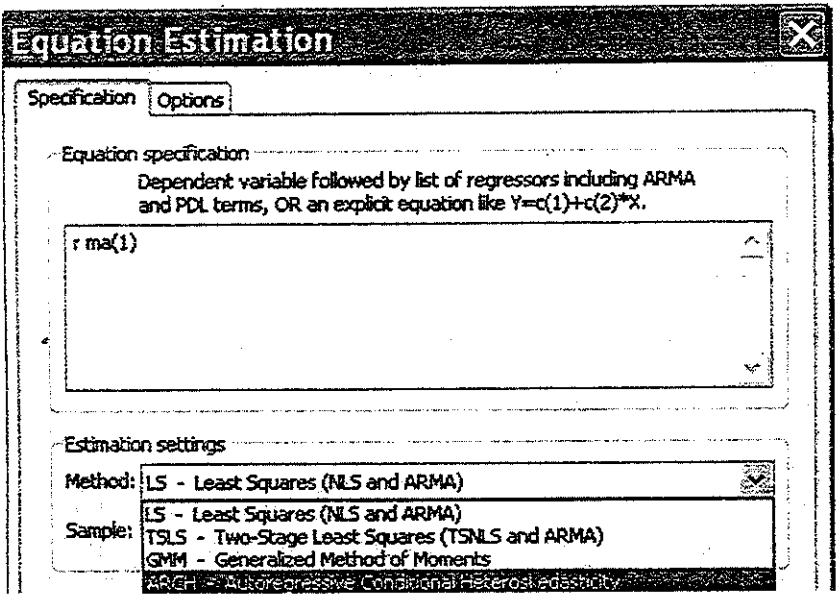
luôn là giải pháp tối ưu, nên trong những trường hợp như vậy người ta thích sử dụng mô hình GARCH(p,q) hơn.

Bước 5: Ước lượng mô hình ARCH(1)

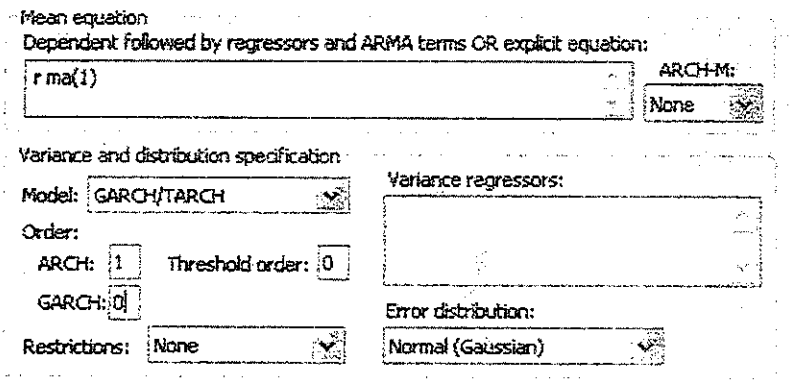
Sau khi đã biết có ảnh hưởng ARCH, nên tốt nhất chúng ta sử dụng phương pháp ước lượng ARCH thay vì phương pháp OLS như ở Hình 9.4.

Tương tự như phương pháp OLS, chúng ta vào **Quick/Estimate Equation**, và thấy xuất hiện hộp thoại như sau:

■ HÌNH 9.7: **Ước lượng mô hình ARCH trên Eviews.**



Sau khi chọn OK, ta có kết quả ước lượng mô hình ARCH(1) như sau:



Nếu ước lượng mô hình ARCH(1), thì ta nhập vào ô "ARCH" số 1, và để trống ô "GARCH". Nếu chọn OK, chúng ta sẽ có kết quả ước lượng như sau:

$$R_t = 0.271e_{t-1} + e_t \tag{9.14}$$

$$u_t \sim N(0,h_t)$$

$$h_t = 0.000134 + 0.9096\, e_{t-1}^2 \tag{9.15}$$

■ HÌNH 9.8: **Kết quả ước lượng ARCH(1).**

Dependent Variable: R
Method: ML - ARCH (Marquardt) - Normal distribution
Date: 03/28/09 Time: 20:19
Sample (adjusted): 7/31/2000 3/26/2009
Included observations: 2006 after adjustments
Convergence achieved after 14 iterations
MA Backcast: 7/28/2000
Presample variance: backcast (parameter = 0.7)
GARCH = C(2) + C(3)*RESID(-1)^2

Variable	Coefficient	Std. Error	z-Statistic	Prob.
MA(1)	0.270536	0.010786	25.08127	0.0000
Variance Equation				
C	0.000134	4.17E-06	32.25845	0.0000
RESID(-1)^2	0.909573	0.062379	14.58147	0.0000

R-squared	0.101856	Mean dependent var	0.000540
Adjusted R-squared	0.100959	S.D. dependent var	0.022703
S.E. of regression	0.021526	Akaike info criterion	-5.211620
Sum squared resid	0.928156	Schwarz criterion	-5.203539
Log likelihood	5230.556	Hannan-Quinn criter.	-5.208643
Durbin-Watson stat	1.847725		

Giá trị ước lượng của hệ số γ_1 có dấu dương và rất có ý nghĩa thống kê, điều này cho thấy kết quả ước lượng phù hợp với kết luận ở phần kiểm định ảnh hưởng ARCH. Giá trị ước lượng của hệ số $\hat{\beta}_2$ từ mô hình OLS cũng có thay đổi một chút và trở nên có ý nghĩa cáo hơn (z-statistic thay đổi từ 9,24 lên 25,08).

Để ước lượng một mô hình ARCH bậc cao hơn, ví dụ ARCH(5), chúng ta cũng thực hiện tương tự như ở mô hình ARCH(1), nhưng thay vì nhập số 1 vào ô 'ARCH', bây giờ ta nhập số 5. Kết quả ước lượng mô hình ARCH(5) được trình bày ở Hình 9.9.

Bây giờ, tất cả các hệ số γ_s đều có dấu dương và có ý nghĩa thống kê, điều này cũng phù hợp với kết quả kiểm định ảnh hưởng ARCH được thảo luận ở phần trên. Sau khi ước lượng mô hình ARCH(5), chúng ta có thể tạo và vẽ đồ thị của chuỗi số liệu về độ lệch chuẩn có điều kiện của suất sinh lợi R bằng cách chọn **View/Garch Graph/Conditional Standard Deviation** (xem Hình 9.10).

■ HÌNH 9.9: **Kết quả ước lượng ARCH(6).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
MA(1)	0.264317	0.020879	12.65941	0.0000
	Variance Equation			
C	4.68E-05	2.09E-06	22.41985	0.0000
RESID(-1)^2	0.436807	0.043932	9.942683	0.0000
RESID(-2)^2	0.206318	0.033354	6.185750	0.0000
RESID(-3)^2	0.164925	0.029148	5.658257	0.0000
RESID(-4)^2	0.133041	0.022342	5.954690	0.0000
RESID(-5)^2	0.085465	0.016613	5.144375	0.0000

R-squared	0.101123	Mean dependent var	0.000540
Adjusted R-squared	0.098425	S.D. dependent var	0.022703
S.E. of regression	0.021557	Akaike info criterion	-5.429348
Sum squared resid	0.928913	Schwarz criterion	-5.409793
Log likelihood	5452.636	Hannan-Quinn criter.	-5.422169
Durbin-Watson stat	1.835032		

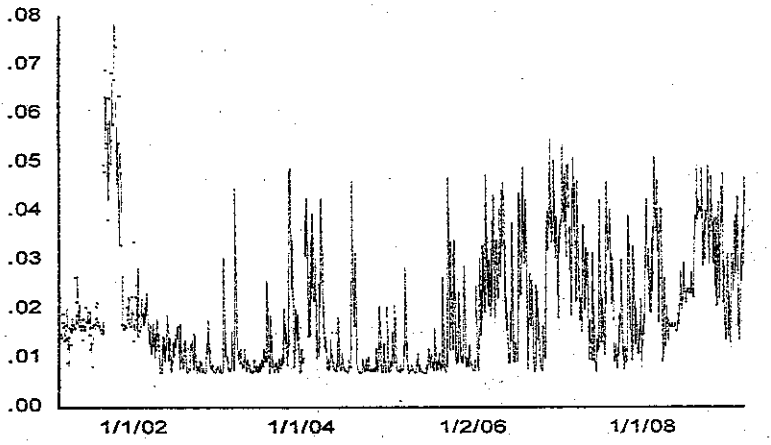
Eviews cũng cho phép chúng ta tạo ra chuỗi dữ liệu về phương sai của suất sinh lợi R bằng cách chọn **Proc/Make GARCH Variance**

Series. Eviews sẽ tự động đặt tên các chuỗi này là GARCH01, GARCH02, v.v. Chúng ta cần phải đặt tên lại các phương sai này với các tên phù hợp với mô hình vừa được ước lượng, ví dụ ARCH1, ARCH5 để thuận lợi trong việc quản lý dữ liệu trong tập tin Eviews. Sau khi đã tạo các chuỗi này, chúng ta có thể vẽ trên cùng đồ thị để dễ dàng phân tích, so sánh giữa các mô hình (xem Hình 9.11). Ở Hình 9.11, có vẽ như mô hình ARCH(5) cho ta ước lượng phương sai nhẵn hơn và rõ ràng hơn so với mô hình ARCH(1). Điều này phần nào chứng tỏ mô hình ARCH(5) phù hợp với dữ liệu suất sinh lợi của cổ phiếu SAM hơn so với mô hình ARCH(1). Để tạo chuỗi độ lệch chuẩn có điều kiện, chúng ta cũng có thể thực hiện như sau:

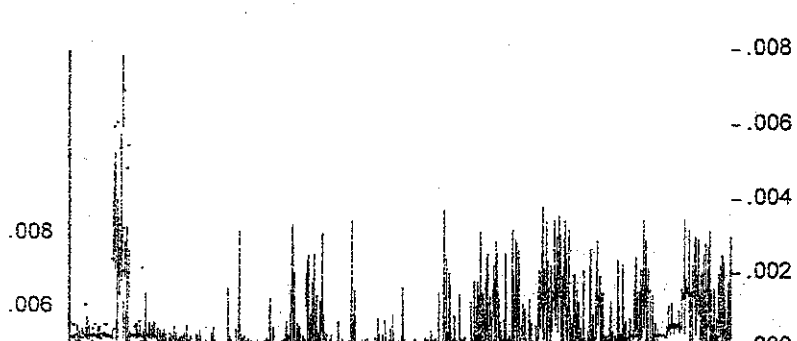
Genr sd_arch5=arch5^1/2

Như vậy, để dự báo suất sinh lợi trung bình và rủi ro của cổ phiếu SAM, chúng ta sẽ sử dụng mô hình ARCH(5). Các thông tin cần cho việc dự báo giai đoạn t+1 sẽ là e_t, e_t^2, e_{t-1}^2, e_{t-2}^2, e_{t-3}^2, và e_{t-4}^2. Các giá trị phần dư có thể được tạo ra bằng hai cách: (1) genr e_5=resid, hoặc (2) **Proc/Make Residual Series**, rồi đặt lại với tên e_5. Sau khi đã có e5, chúng ta dễ dàng tạo các giá trị bình phương bằng cách genr e_5^2=e_5^2.

■ HÌNH 9.10: **Độ lệch chuẩn của mô hình ARCH(6).**



■ HÌNH 9.11: **Phương sai của mô hình ARCH(1).**



Bước 6: Ứng dụng dự báo

■ BẢNG 9.1: **Kết quả dự báo mô hình ARCH(5).**

Ngày	e_t	e_t^2	$\hat{r}_t$ (suất sinh lợi dự báo)	$\hat{h}_t$ (phương sai dự báo)
20/3/2009	0.00815	0.0000664		
23/3/2009	-0.05129	0.0026303		
24/3/2009	0.055816	0.0031154		
25/3/2009	0.012459	0.0001552		
26/3/2009	0.029713	0.0008829		
27/3/2009			0.00785	0.00123

Theo kết quả mô hình ARCH(5), vào ngày 27/3/2009 suất sinh lợi kỳ vọng của cổ phiếu SAM sẽ tăng lên khoảng 0,79% (là 0,00785*100%), với độ lệch chuẩn dự kiến sẽ là 3.5% (là 0,00123^(1/2)*100%).

MÔ HÌNH GARCH

Theo Engle (1995), một trong những hạn chế của mô hình ARCH là nó có vẽ giống dạng mô hình trung bình di động hơn là dạng mô hình tự hồi quy (AR). Vì vậy, một ý tưởng mới được đề xuất là chúng ta nên đưa thêm các biến trễ của phương sai có điều kiện vào phương trình phương sai theo dạng tự hồi quy. Ý tưởng này do Tim Bollerslev đề xuất lần đầu tiên vào năm 1986 trên tạp chí *Journal of Econometrics* với tên gọi "Generalised Autogressive Conditional Heteroskedasticity", và viết tắt là mô hình GARCH. Ngoài ra, nếu các ảnh hưởng ARCH có quá nhiều độ trễ sẽ có thể ảnh hưởng đến kết quả ước lượng do giảm đáng kể số bậc tự do trong mô hình, và điều này càng nghiêm trọng đối với các chuỗi thời gian ngắn, ví dụ giá của các cổ phiếu mới lưu hành trên thị trường. Chính vì vậy, mô hình GARCH có có xu hướng được các nhà dự báo sử dụng phổ biến hơn.

MÔ HÌNH GARCH(p,q)

Mô hình GARCH(p,q) có dạng sau đây:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad (9.16)$$

$$u_t \sim N(0, h_t)$$

$$h_t = \gamma_0 + \sum_{i=1}^{p} \delta_i h_{t-i} + \sum_{j=i}^{q} \gamma_j u_{t-j}^2 \qquad (9.17)$$

Phương trình (9.17) nói lên rằng phương sai h_t bây giờ phụ thuộc vào cả giá trị quá khứ của những cú sốc, đại diện bởi các biến trễ của hạng nhiễu bình phương, và các giá trị quá khứ của bản thân h_t, đại diện bởi các biến h_{t-i}. Nếu p = 0, có nghĩa là bậc của AR bằng 0 thì thì mô hình GARCH (0,q) đơn giản là mô hình ARCH(q). Dạng đơn giản nhất của

mô hình GARCH(p,q) là mô hình GARCH(1,1). Phương trình phương sai của mô hình GARCH(1,1) được thể hiện như sau:

$$h_t = \gamma_0 + \delta_1 h_{t-1} + \gamma_1 u_{t-1}^2 \qquad (9.18)$$

MÔ HÌNH GARCH(1,1) VÀ ARCH(q) VÔ TẬN

Để nhận thấy mô hình GARCH(1,1) là một cách biểu diễn thu gọn của mô hình ARCH(q), với q kéo dài vô tận, chúng ta cần một vài biến đổi. Phương trình (9.18) có thể được viết lại như sau:

$$\begin{aligned}
h_t &= \gamma_0 + \delta h_{t-1} + \gamma_1 u_{t-1}^2 \\
&= \gamma_0 + \delta(\gamma_0 + \delta h_{t-2} + \gamma_1 u_{t-2}^2) + \gamma_1 u_{t-1}^2 \\
&= \gamma_0 + \gamma_1 u_{t-1}^2 + \delta\gamma_0 + \delta^2 h_{t-2} + \delta\gamma_1 u_{t-2}^2 \\
&= \gamma_0 + \gamma_1 u_{t-1}^2 + \delta\gamma_0 + \delta^2(\gamma_0 + \delta h_{t-3} + \gamma_1 u_{t-3}^2) + \delta\gamma_1 u_{t-2}^2 \\
&= \gamma_0 + \gamma_1 u_{t-1}^2 + \delta\gamma_0 + \delta^2\gamma_0 + \delta^3 h_{t-3} + \delta^2\gamma_1 u_{t-3}^2 + \delta\gamma_1 u_{t-2}^2 \\
&\dots \\
&= \gamma_0 + \delta\gamma_0 + \delta^2\gamma_0 + \dots + \gamma_1 u_{t-1}^2 + \delta\gamma_1 u_{t-2}^2 + \delta^2\gamma_1 u_{t-3}^2 + \dots \qquad (9.19)
\end{aligned}$$

Đặt $A = \gamma_0 + \delta\gamma_0 + \delta^2\gamma_0 + \dots + \delta^\infty\gamma_0$ (9.20)

Nếu nhân hai vế của phương trình (9.20) cho δ ta sẽ có:

$$\delta A = \delta\gamma_0 + \delta^2\gamma_0 + \dots + \delta^\infty\gamma_0 \qquad (9.21)$$

Lấy (9.20) trừ (9.21), rồi sắp xếp lại, ta sẽ có công thức A thu gọn như sau:

$$A = \frac{\gamma_0}{1-\delta} \qquad (9.22)$$

Thế công thức (9.22) vào phương trình (9.19) ta sẽ có:

$$h_t = \frac{\gamma_0}{1-\delta} + \gamma_1\left(u_{t-1}^2 + \delta u_{t-2}^2 + \delta^2 u_{t-3}^2 + \dots\right)$$

$$h_t = \frac{\gamma_0}{1-\delta} + \gamma_1 \sum_{j=1}^{\infty} \delta^{j-1} u_{t-j}^2 \qquad (9.23)$$

Như vậy, phương trình (9.23) cho thấy mô hình GARCH(1,1) tương đương với mô hình ARCH bậc vô cùng với các hệ số có xu hướng giảm dần. Vì lý do này, chúng ta nên sử dụng mô hình GARCH(1,1) thay cho các mô hình ARCH bậc cao bởi vì với mô hình GARCH(1,1), chúng ta có ít số hệ số cần ước lượng hơn và vì thế sẽ giúp hạn chế khả năng mất đi một số bậc tự do trong mô hình.

ƯỚC LƯỢNG MÔ HÌNH GARCH TRÊN EVIEWS

Tiếp tục sử dụng tập tin **DATA9-2**, và ước lượng mô hình GARCH(1,1) như sau. Giả sử phương trình trung bình vẫn có dạng MA(1), ta vào **Quick/Estimate Equation**, nhập vào hộp thoại "Equation Specification" r ma(1), rồi chọn phương pháp ước lượng ARCH – Autogressive Conditional Heteroskedasticity. Ở đây, hộp thoại ở trên dành cho việc xác định dạng phương trình trung bình, và hộp thoại ở dưới dành cho việc xác định dạng phương trình phương sai. Ở hộp thoại này, chúng ta nhập bậc của q và p vào các ô 'ARCH' và 'GARCH'.

Nếu mô hình GARCH(1,1) thì ta nhập số 1 ở ô 'ARCH' và số 1 ở ô 'GARCH'. Nếu mô hình GARCH(2,4) thì ta nhập số 4 vào ô 'ARCH' và số 2 vào ô 'GARCH'. Sau khi đã xác định số độ trễ q và p, ví dụ 1 và 1 cho mô hình GARCH(1,1), ta chọn 'OK', và có kết quả như ở Hình 9.12. Kết quả này có thể được viết như sau:

$$R_t = 0.26535 e_{t-1} + e_t \qquad (9.24)$$

$$u_t \sim N(0, h_t)$$

$$h_t = 0.000012 + 0.684 h_{t-1} + 0.319 e_{t-1}^2 \qquad (9.25)$$

■ HÌNH 9.12: **Kết quả ước lượng GARCH(1,1).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
MA(1)	0.265351	0.022220	11.94191	0.0000
		Variance Equation		
C	1.18E-05	8.09E-07	14.59229	0.0000
RESID(-1)^2	0.318982	0.022715	14.04294	0.0000
GARCH(-1)	0.683746	0.014290	47.84678	0.0000

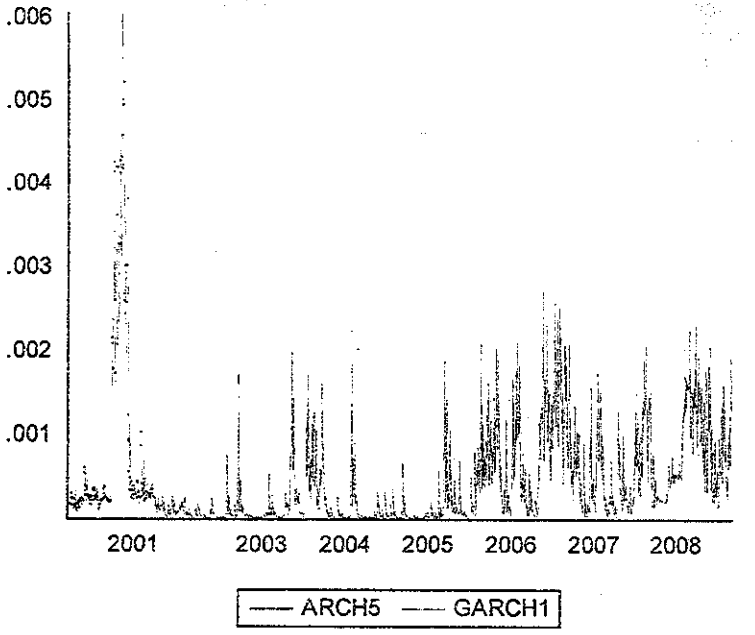
R-squared	0.101251	Mean dependent var	0.000540
Adjusted R-squared	0.099904	S.D. dependent var	0.022703
S.E. of regression	0.021539	Akaike info criterion	-5.450539
Sum squared resid	0.928781	Schwarz criterion	-5.439365
Log likelihood	5470.891	Hannan-Quinn criter.	-5.446437
Durbin-Watson stat	1.837138		

Trong bảng kết quả ta nhận thấy rằng các hệ số δ và γ_0 đều có ý nghĩa thống kê rất cao. Để tạo ra chuỗi dữ liệu về phương sai của mô hình GARCH(1,1), ta vào **Proc/Make GARCH Variance Series**, và vẽ đồ thị chuỗi này như ở Hình 9.13.

Quan sát đồ thị trên Hình 9.13 ta nhận thấy có vẽ như hai mô hình ARCH(5) và GARCH(1,1) rất giống nhau. Hơn nữa, nếu quan sát kỹ, thì có vẽ mô hình GARCH(1,1) có giá trị phương sai ước lượng tương đối nhẵn hơn so với mô hình ARCH(5). Như vậy, tốt nhất ta nên sử dụng mô hình GARCH thay cho các mô hình ARCH bậc cao vì như thế ta vừa tiết kiệm được số bậc tự do (nhất là khi số quan sát ít) vừa thuận tiện hơn trong việc dự báo (giảm việc tính toán).

Nếu thay đổi giá trị trong các ô 'ARCH' và 'GARCH' thành 1 và 3, ta có kết quả ước lượng của mô hình GARCH(2,1) như ở Hình 9.14. Tương tự, ta có kết quả ước lượng mô hình GARCH(3,1) ở Hình 9.15 và mô hình GARCH(2,2) ở Hình 9.16. Ngoài ra, chúng ta cũng có thể thử các mô hình khác như GARCH(4,1), GARCH(3,2), GARCH(3,3), v.v. Tuy nhiên, điều này là không cần thiết vì chỉ có mô hình GARCH(2,1) là mô hình thích hợp nhất cho dữ liệu suất sinh lợi R của cổ phiếu SAM.

■ HÌNH 9.13: **So sánh phương sai của ARCH(5) và GARCH(1,1).**



■ HÌNH 9.14: **Kết quả ước lượng GARCH(2,1).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
MA(1)	0.252263	0.022206	11.35997	0.0000
	Variance Equation			
C	1.28E-05	1.01E-06	12.73792	0.0000
RESID(-1)^2	0.370828	0.032412	11.44125	0.0000
GARCH(-1)	0.365090	0.080241	4.549907	0.0000
GARCH(-2)	0.267493	0.063125	4.237498	0.0000

R-squared	0.099480	Mean dependent var	0.000540
Adjusted R-squared	0.097680	S.D. dependent var	0.022703
S.E. of regression	0.021566	Akaike info criterion	-5.453670
Sum squared resid	0.930611	Schwarz criterion	-5.439702
Log likelihood	5475.031	Hannan-Quinn criter.	-5.448542
Durbin-Watson stat	1.810591		

HÌNH 9.15: Kết quả ước lượng GARCH(3,1).

Variable	Coefficient	Std. Error	z-Statistic	Prob.
MA(1)	0.253351	0.022636	11.19255	0.0000
Variance Equation				
C	1.31E-05	1.15E-06	11.34041	0.0000
RESID(-1)^2	0.381658	0.037211	10.25655	0.0000
GARCH(-1)	0.350745	0.085382	4.107940	0.0000
GARCH(-2)	0.181249	0.088625	2.045130	0.0408
GARCH(-3)	0.088285	0.058726	1.503341	0.1328
R-squared	0.099641	Mean dependent var		0.000540
Adjusted R-squared	0.097390	S.D. dependent var		0.022703
S.E. of regression	0.021569	Akaike info criterion		-5.453633
Sum squared resid	0.930445	Schwarz criterion		-5.436871
Log likelihood	5475.994	Hannan-Quinn criter.		-5.447479
Durbin-Watson stat	1.812789			

HÌNH 9.16: Kết quả ước lượng GARCH(2,2).

Variable	Coefficient	Std. Error	z-Statistic	Prob.
MA(1)	0.268350	0.021920	12.24219	0.0000
Variance Equation				
C	9.40E-07	2.01E-07	4.684152	0.0000
RESID(-1)^2	0.411010	0.039871	10.30853	0.0000
RESID(-2)^2	-0.376240	0.035694	-10.54062	0.0000
GARCH(-1)	1.389593	0.054604	25.44852	0.0000
GARCH(-2)	-0.424226	0.048533	-8.740943	0.0000
R-squared	0.101607	Mean dependent var		0.000540
Adjusted R-squared	0.099361	S.D. dependent var		0.022703
S.E. of regression	0.021545	Akaike info criterion		-5.467849
Sum squared resid	0.928413	Schwarz criterion		-5.451087
Log likelihood	5490.252	Hannan-Quinn criter.		-5.461695
Durbin-Watson stat	1.843255			

Từ các kết quả trên ta nhận thấy rằng, mô hình GARCH(2,2) không phù hợp do độ trễ ARCH bậc hai có dấu âm. Đối với mô hình GARCH(3,1), thì hệ số của độ trễ bậc ba của GARCH không có ý

nghĩa thống kê. Lập luận tương tự cho các mô hình GARCH(4,1), GARCH(3,2), GARCH(3,3), v.v. Như vậy, để dự báo suất sinh lợi trung bình và rủi ro của R, ta có thể sử dụng mô hình GARCH(2,1).

Kết quả mô hình GARCH(2,1) có thể được viết lại như sau:

$$R_t = 0.25226e_{t-1} + e_t \qquad (9.26)$$

$$u_t \sim N(0,h_t)$$

$$h_t = 0.0000128 + 0.365h_{t-1} + 0.268h_{t-2} + 0.371\, e^2_{t-1} \qquad (9.27)$$

Để dự báo cho giai đoạn $t+1$, ta cần hai thông tin sau đây: (i) giá trị trễ của phần dư và (ii) giá trị trễ của phương sai (xem Bảng 9.2).

■ BẢNG 9.2: **Kết quả dự báo mô hình ARCH(5).**

Ngày	e_t	e_t^2	$\hat{r}_t$	$\hat{h}_t$
25/3/2009	0.013302	0.000177		0.00191
26/3/2009	0.029651	0.000879		0.001189
27/3/2009			0.00748	0.00135

Theo kết quả mô hình ARCH(5), vào ngày 27/3/2009 suất sinh lợi kỳ vọng của cổ phiếu SAM sẽ tăng lên khoảng 0.75%, với độ lệch chuẩn dự kiến sẽ là 3.68%. Kết quả này không có nhiều khác biệt với mô hình ARCH đã trình bày ở phần trên. Lưu ý, khi đã có giá trị trung bình và độ lệch chuẩn (cùng với giả định phân phối chuẩn), thì chúng ta có thể dễ dàng dự báo được xác suất để $\hat{r}_{t+1}$ dương hoặc trong một khoảng nào đó bằng hàm NORMDIST hoặc trên Excel.

MÔ HÌNH GARCH Ở GIÁ TRỊ TRUNG BÌNH

Các mô hình GARCH ở giá trị trung bình (GARCH-M) cho phép giá trị trung bình có điều kiện phụ thuộc vào phương sai có điều kiện của chính nó. Ví dụ, xem xét hành vi các nhà đầu tư thuộc dạng 'sợ' rủi ro, và vì thế, họ có xu hướng đòi hỏi thêm một mức phí bù rủi ro như

một phần đền bù để quết định nắm giữ một tài sản rủi ro. Như vậy, phí bù rủi ro là một hàm đồng biến với rủi ro; nghĩa là, rủi ro càng cao thì phí bù rủi ro phải càng nhiều. Nếu rủi ro được đo lường bằng mức dao động hay bằng phương sai có điều kiện (như được trình bày ở trên), thì phương sai có điều kiện có thể là một phần trong phương trình trung bình của biến Y_t. Theo cách này, mô hình GARCH-M(p,q) sẽ có dạng sau:

$$Y_t = \beta_1 + \beta_2 X_t + \theta h_t + u_t \tag{9.28}$$

$$u_t \sim N(0, h_t)$$

$$h_t = \gamma_0 + \sum_{i=1}^{p} \delta_i h_{t-i} + \sum_{j=1}^{q} \gamma_j u_{t-j}^2 \tag{9.29}$$

Một dạng khác của mô hình GARCH-M(p,q) là, thay vì sử dụng chuỗi phương sai trong phương trình trung bình, người ta sử dụng độ lệch chuẩn của chuỗi phương sai có điều kiện như sau:

$$Y_t = \beta_1 + \beta_2 X_t + \theta \sqrt{h_t} + u_t \tag{9.30}$$

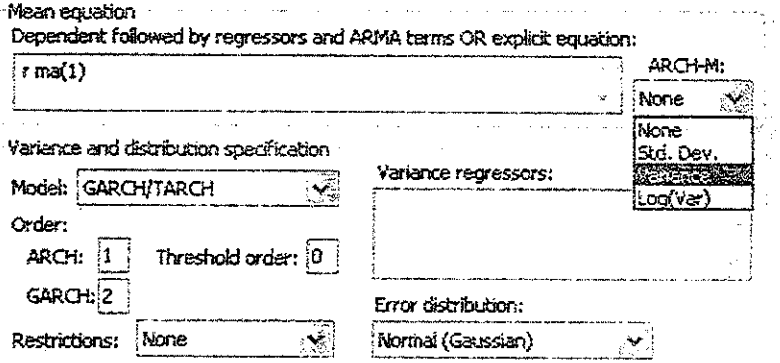
$$u_t \sim N(0, h_t)$$

$$h_t = \gamma_0 + \sum_{i=1}^{p} \delta_i h_{t-i} + \sum_{j=1}^{q} \gamma_j u_{t-j}^2 \tag{9.31}$$

ƯỚC LƯỢNG CÁC MÔ HÌNH GARCH-M TRÊN EVIEWS

Để ước lượng mô hình GARCH-M trên Eviews, trước hết ta cũng vào **Quick/Estimate Equation** để mở cửa sổ ước lượng. Sau đó chúng ta chọn chọn phương pháp ước lượng **ARCH-Autogressive Conditional Heteroskedasticity**. Ví dụ, ước lượng mô hình GARCH-M(2,1), ta thực hiện như hướng dẫn ở Hình 9.17:

■ HÌNH 9.17: **Ước lượng GARCH-M(2,1).**



Tất cả các lựa chọn khác đều giống với cách ước lượng mô hình GARCH đã được trình bày ở trên, nhưng để ước lượng mô hình GARCH-M, ta bổ sung thêm phần ở gốc trên với tên gọi là 'ARCH-M'. Ở đây, chúng ta có bốn sự lựa chọn (i) None, (ii) Std.Dev, (iii) Variance, và (iv) log(Var). Nếu chọn Variance, ta sẽ có kết quả ước lượng như trong Hình 9.18.

■ HÌNH 9.18: **Kết quả ước lượng GARCH-M(2,1).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
GARCH	1.305456	1.109211	1.176924	0.2392
MA(1)	0.251810	0.022169	11.35844	0.0000
	Variance Equation			
C	1.29E-05	1.01E-06	12.82056	0.0000
RESID(-1)^2	0.375033	0.032518	11.53326	0.0000
GARCH(-1)	0.351948	0.078741	4.469720	0.0000
GARCH(-2)	0.276484	0.062025	4.457590	0.0000

R-squared	0.096336	Mean dependent var	0.000540
Adjusted R-squared	0.094077	S.D. dependent var	0.022703
S.E. of regression	0.021609	Akaike info criterion	-5.453445
Sum squared resid	0.933860	Schwarz criterion	-5.436683
Log likelihood	5475.805	Hannan-Quinn criter.	-5.447291
Durbin-Watson stat	1.799222		

Kết quả ước lượng mô hình GARCH-M(2,1) cho thấy hệ số của phương sai trong phương trình trung bình không có ý nghĩa thống kê. Điều này có thể nói lên rằng mô hình GARCH-M không phù hợp trong trường hợp này.

MÔ HÌNH TGARCH

Hạn chế lớn nhất trong các mô hình ARCH và GARCH là chúng được giả định có tính chất đối xứng. Điều này có nghĩa các mô hình này chỉ quan tâm đến giá trị tuyệt đối của các cú sốc chứ không quan tâm đến 'dấu' của chúng (bởi vì hạng nhiễu/phần dư được xử lý dưới dạng bình phương). Vì thế, trong các mô hình ARCH/GARCH, một cú sốc mạnh có giá trị dương có ảnh hưởng lên sự dao động của chuỗi dữ liệu hoàn toàn giống với một cú sốc mạnh có giá trị âm. Tuy nhiên, kinh nghiệm cho thấy rằng, trong tài chính, các cú sốc âm (hoặc tin tức xấu) trên thị trường thường có tác động mạnh và dai dẳng hơn so với các cú sốc dương (hoặc tin tức tốt) vì nó làm cho các nhà đầu tư bị tê liệt và trở nên bi quan chán nản và thậm chí chờ đợi một cách thụ động các dấu hiệu thị trường. Chính vì vậy, người ta cố gắng mô hình hóa sự khác biệt trong ảnh hưởng này. Và, các mô hình TGARCH đã được phát triển.

Mô hình TGARCH được phát triển bởi Zakoian (1990), và Glosten, Jaganathan, và Runkle (1993). Mục đích chính của mô hình này là nhằm xem xét tính chất bất cân xứng giữa các cú sốc âm và dương. Và đây cũng là một cách kiểm định tính hiệu quả của thị trường. Để làm như vậy, các học giả này đề xuất nên đưa vào phương trình phương sai một biến giả tương tác giữa hạng nhiễu bình phương và biến giả d_t, trong đó d_t có giá trị bằng 1 nếu $u_t < 0$, và bằng 0 nếu $u_t > 0$. Nếu hệ số của biến tương tác này có ý nghĩa thống kê sẽ chứng tỏ có sự khác biệt trong các cú sốc khác nhau.

Từ ý tưởng này, phương trình phương sai trong mô hình TGARCH(1,1) sẽ có dạng như sau:

$$h_t = \gamma_0 + \delta_1 h_{t-1} + \gamma_1 u_{t-1}^2 + \upsilon_1 u_{t-1}^2 d_{t-1} \qquad (9.32)$$

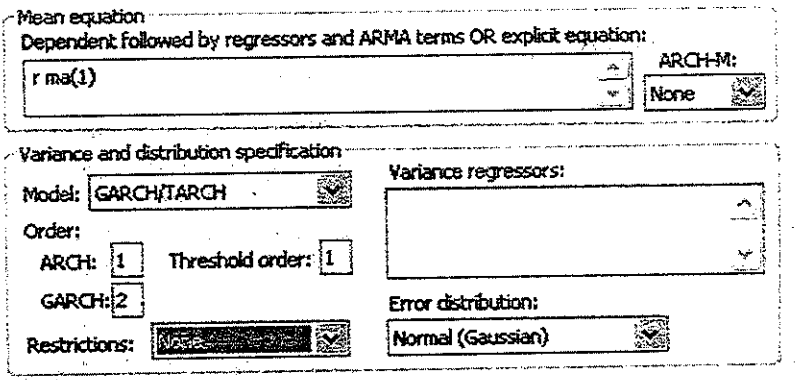
Nếu hệ số υ_1 có ý nghĩa thống kê, thì các tin tức tốt và tin tức xấu sẽ có ảnh hưởng khác nhau lên phương sai. Cụ thể, tin tức tốt chỉ có ảnh hưởng γ_1, trong khi đó, tin tức xấu có ảnh hưởng $(\gamma_1+\upsilon_1)$. Nếu $\upsilon_1 > 0$, thì chúng ta có thể nói rằng có sự bất cân xứng trong tác động giữa tin tức tốt và tin tức xấu. Ngược lại, nếu $\upsilon_1 = 0$, thì tác động của tin tức có tính chất cân xứng. TGARCH bậc cao có thể được thể hiện như sau:

$$h_t = \gamma_0 + \sum_{i=1}^{p}\delta_i h_{t-i} + \sum_{j=1}^{q}(\gamma_j + \upsilon_j d_{t-j})u_{t-j}^2 \qquad (9.33)$$

ƯỚC LƯỢNG MÔ HÌNH TGARCH TRÊN EVIEWS

Để ước lượng mô hình TGARCH trên Eviews, trước hết ta cũng vào **Quick/Estimate Equation** để mở cửa sổ ước lượng. Sau đó chúng ta chọn chọn phương pháp ước lượng **ARCH-Autogressive Conditional Heteroskedasticity**. Ví dụ, ước lượng mô hình TGARCH(2,1), ta thực hiện như hướng dẫn ở Hình 9.19:

■ HÌNH 9.19: **Ước lượng TGARCH(2,1).**



Tất cả các lựa chọn khác đều giống với cách ước lượng mô hình GARCH đã được trình bày ở trên, nhưng để ước lượng mô hình TGARCH, ta bổ sung thêm phần ở ô 'Threshold Order' số độ trễ của biến giả d_t. Nếu chọn độ trễ của $d_t = 1$, ta sẽ có kết quả ước lượng như trong Hình 9.20.

■ HÌNH 9.20: **Kết quả ước lượng TGARCH(2,1).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
MA(1)	0.251915	0.022294	11.29992	0.0000
	Variance Equation			
C	1.28E-05	1.01E-06	12.69081	0.0000
RESID(-1)^2	0.377644	0.035266	10.70844	0.0000
RESID(-1)^2*(RESID(-1)<0)	-0.014365	0.041456	-0.346515	0.7290
GARCH(-1)	0.366132	0.080121	4.569751	0.0000
GARCH(-2)	0.266588	0.062827	4.243211	0.0000

R-squared	0.099429	Mean dependent var	0.000540
Adjusted R-squared	0.097177	S.D. dependent var	0.022703
S.E. of regression	0.021572	Akaike info criterion	-5.452728
Sum squared resid	0.930664	Schwarz criterion	-5.435966
Log likelihood	5475.086	Hannan-Quinn criter.	-5.446574
Durbin-Watson stat	1.809889		

Kết quả ước lượng cho thấy hệ số υ_1 không có ý nghĩa thống kê. Như vậy, trong trường hợp này không có sự khác biệt giữa tin tức tốt và tin tức xấu. Nói cách khác, ảnh hưởng của tin tức có tính chất cân xứng.

VÍ DỤ MINH HỌA

Trong ví dụ này, chúng ta sẽ thực hiện một phân tích hoàn chỉnh nhằm lựa chọn mô hình dự báo thích hợp cho giá trị trung bình và phương sai cho suất sinh lợi của cổ phiếu giả định ABC (tập tin **DATA9-1**). Từ đồ thị Hình 9.1, ta nhận thấy rằng suất sinh lợi của cổ phiếu ABC có thể có ảnh hưởng ARCH. Để lựa chọn mô hình thích hợp, chúng ta thực hiện lần lược các bước sau đây.

BƯỚC 1: LỰA CHỌN DẠNG MÔ HÌNH PHÙ HỢP CHO PHƯƠNG TRÌNH SUẤT SINH LỢI TRUNG BÌNH

Từ đồ thị giản đồ tự tương quan ta nhận thấy suất sinh lợi của cổ phiếu ABC là một chuỗi dừng. Như vậy, các mô hình AR(1), MA(1), và ARMA(1,1) có thể phù hợp với dữ liệu này. Để lựa chọn mô hình phù hợp nhất, chúng ta lần lượt ước lượng ba mô hình này, rồi so sánh, đánh

giá, lựa chọn mô hình tốt nhất trên cơ sở các tiêu chí AIC, SBC, RMSE, và các hệ số hồi quy. Hình 9.22, 9.23, và 9.24 cung cấp kết quả ước lượng các mô hình này.

■ HÌNH 9.21: **Giản đồ tự tương quan suất sinh lợi ABC.**

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.071	0.071	12.997	0.000
		2	-0.039	-0.044	16.933	0.000
		3	-0.039	-0.033	20.892	0.000
		4	0.008	0.011	21.048	0.000
		5	-0.023	-0.027	22.404	0.000
		6	-0.051	-0.049	29.300	0.000
		7	-0.066	-0.061	40.697	0.000
		8	0.034	0.037	43.655	0.000
		9	0.019	0.006	44.586	0.000
		10	0.033	0.030	47.424	0.000
		11	0.030	0.029	49.824	0.000
		12	0.006	-0.001	49.917	0.000

■ HÌNH 9.22: **Mô hình ARMA(1,1) cho suất sinh lợi ABC.**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000390	0.000195	1.998513	0.0458
AR(1)	-0.188319	0.242011	-0.778143	0.4366
MA(1)	0.263883	0.237678	1.110251	0.2670

R-squared	0.005874	Mean dependent var	0.000390
Adjusted R-squared	0.005111	S.D. dependent var	0.009400
S.E. of regression	0.009376	Akaike info criterion	-6.500241
Sum squared resid	0.229077	Schwarz criterion	-6.493495
Log likelihood	8482.564	Hannan-Quinn criter.	-6.497797
F-statistic	7.699594	Durbin-Watson stat	2.004385
Prob(F-statistic)	0.000463		

RMSE = 0.009398

■ HÌNH 9.23: **Mô hình AR(1) cho suất sinh lợi ABC.**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000390	0.000198	1.972608	0.0486
AR(1)	0.070574	0.019543	3.611253	0.0003

R-squared	0.004977	Mean dependent var	0.000390
Adjusted R-squared	0.004596	S.D. dependent var	0.009400
S.E. of regression	0.009378	Akaike info criterion	-6.500106
Sum squared resid	0.229284	Schwarz criterion	-6.495608
Log likelihood	8481.388	Hannan-Quinn criter.	-6.498477
F-statistic	13.04115	Durbin-Watson stat	1.993207
Prob(F-statistic)	0.000310		

RMSE = 0.009398

■ HÌNH 9.24: **Mô hình MA(1) cho suất sinh lợi ABC.**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000391	0.000197	1.977977	0.0480
MA(1)	0.076201	0.019530	3.901851	0.0001

R-squared	0.005392	Mean dependent var	0.000391
Adjusted R-squared	0.005011	S.D. dependent var	0.009398
S.E. of regression	0.009374	Akaike info criterion	-6.500888
Sum squared resid	0.229193	Schwarz criterion	-6.496393
Log likelihood	8485.659	Hannan-Quinn criter.	-6.499260
F-statistic	14.13957	Durbin-Watson stat	2.004414
Prob(F-statistic)	0.000173		

RMSE = 0.009395

Như vậy, cả hai mô hình AR(1) và MA(1) đều phù hợp với dữ liệu của cổ phiếu ABC, trừ mô hình ARMA(1,1). Để minh họa cho ví dụ này, giả sử mô hình AR(1) là mô hình thích hợp nhất.

BƯỚC 2: KIỂM ĐỊNH ẢNH HƯỞNG ARCH VÀ LỰA CHỌN MÔ HÌNH THÍCH HỢP

Kết quả kiểm định ảnh hưởng ARCH(1) cho thấy giá trị χ^2 tính toán bằng 42.3652 lớn hơn giá trị χ^2 phê phán ở mức ý nghĩa 1% (=CHIINV(1%,1) = 6.64). Như vậy, chúng ta có thể kết luận rằng có ảnh hưởng ARCH(1) trong mô hình suất sinh lợi cổ phiếu ABC.

■ HÌNH 9.25: **Kiểm định ảnh hưởng ARCH(1).**

Heteroskedasticity Test: ARCH

F-statistic	43.03148	Prob. F(1,2607)	0.0000
Obs*R-squared	42.36520	Prob. Chi-Square(1)	0.0000

■ HÌNH 9.26: **Kiểm định ảnh hưởng ARCH(8).**

Heteroskedasticity Test: ARCH

F-statistic	31.12802	Prob. F(8,2593)	0.0000
Obs*R-squared	227.9927	Prob. Chi-Square(8)	0.0000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.73E-05	4.59E-06	8.115440	0.0000
RESID^2(-1)	0.055674	0.019592	2.841689	0.0045
RESID^2(-2)	0.114882	0.019575	5.868757	0.0000
RESID^2(-3)	0.088667	0.019654	4.511473	0.0000
RESID^2(-4)	0.050330	0.019700	2.554862	0.0107
RESID^2(-5)	0.059928	0.019695	3.042660	0.0024
RESID^2(-6)	0.073261	0.019650	3.728360	0.0002
RESID^2(-7)	0.067999	0.019576	3.473655	0.0005
RESID^2(-8)	0.064947	0.019077	3.404550	0.0007

■ HÌNH 9.27: **Kiểm định ảnh hưởng ARCH(9).**

Heteroskedasticity Test: ARCH

F-statistic	27.74813	Prob. F(9,2591)	0.0000
Obs*R-squared	228.6578	Prob. Chi-Square(9)	0.0000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.75E-05	4.65E-06	8.055732	0.0000
RESID^2(-1)	0.055907	0.019643	2.846142	0.0045
RESID^2(-2)	0.115023	0.019627	5.860443	0.0000
RESID^2(-3)	0.088895	0.019710	4.510250	0.0000
RESID^2(-4)	0.050707	0.019736	2.569287	0.0102
RESID^2(-5)	0.059957	0.019729	3.038988	0.0024
RESID^2(-6)	0.073453	0.019734	3.722096	0.0002
RESID^2(-7)	0.068487	0.019707	3.475360	0.0005
RESID^2(-8)	0.069059	0.019626	3.518810	0.0004
RESID^2(-9)	-0.007513	0.019124	-0.392854	0.6945

Như vậy, cho đến độ trễ bằng 8 thì các hệ số hồi quy trong mô hình hồi quy phụ đều có ý nghĩa thống kê. Tuy nhiên, khi độ trễ bằng 9, thì hệ số hồi quy của phần dư bình phương với độ trễ bằng 9 không có ý nghĩa thống kê và có dấu âm. Cho nên, chúng ta có thể kết luận rằng, ảnh hưởng trễ có thể phù hợp nhất ở độ trễ bằng 8. Như vậy, mô hình ARCH(8) có thể phù hợp nhất cho phương trình phương sai. Theo phân tích ở trên, tốt nhất chúng ta nên chọn mô hình GARCH(1,1), thay cho mô hình ARCH(8). Hình 9.28 và 9.29 cung cấp kết quả ước lượng hai mô hình ARCH(8) và GARCH(1,1).

■ HÌNH 9.28: **Kết quả ước lượng mô hình ARCH(8).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.000417	0.000171	2.445897	0.0144
AR(1)	0.067551	0.019891	3.396080	0.0007
Variance Equation				
C	3.06E-05	2.56E-06	11.96909	0.0000
RESID(-1)^2	0.064985	0.015119	4.298207	0.0000
RESID(-2)^2	0.116105	0.024308	4.776372	0.0000
RESID(-3)^2	0.085490	0.024586	3.477186	0.0005
RESID(-4)^2	0.061910	0.021041	2.942351	0.0033
RESID(-5)^2	0.062521	0.022807	2.741267	0.0061
RESID(-6)^2	0.098949	0.023569	4.198279	0.0000
RESID(-7)^2	0.092608	0.022027	4.204243	0.0000
RESID(-8)^2	0.078813	0.025109	3.138871	0.0017

■ HÌNH 9.29: **Kết quả ước lượng mô hình GARCH(1,1).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.000464	0.000169	2.739206	0.0062
AR(1)	0.062791	0.020700	3.033430	0.0024
	Variance Equation			
C	7.67E-07	2.31E-07	3.323046	0.0009
RESID(-1)^2	0.049277	0.006446	7.644347	0.0000
GARCH(-1)	0.942418	0.007671	122.8522	0.0000

Các hệ số của phương trình trung bình ở mô hình ARCH(8) và GARCH(1,1) đều có ý nghĩa thống kê cao hơn so với mô hình AR(1) ở Hình 9.21. Điều này chứng tỏ các mô hình ARCH/GARCH có thể phù hợp hơn các mô hình ARIMA.

BƯỚC 3: GARCH(1,1) HAY GARCH(p,q)

Để xác định xem mô hình GARCH(1,1) hay GARCH(p,q) phù hợp hơn với dữ liệu của cổ phiếu ABC, chúng ta lần lượt thực hiện các mô hình GARCH bậc cao vào so sánh với mô hình GARCH(1,1). Hình 9.30, 9.31, và 9.32 cung cấp kết quả ước lượng các mô hình GARCH(2,1), GARCH(1,2), và GARCH(1,8).

■ HÌNH 9.30: **Kết quả ước lượng mô hình GARCH(2,1).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.000465	0.000169	2.743483	0.0061
AR(1)	0.062862	0.020841	3.016312	0.0026
	Variance Equation			
C	6.88E-07	2.79E-07	2.470111	0.0135
RESID(-1)^2	0.043295	0.013104	3.304059	0.0010
GARCH(-1)	1.094227	0.274933	3.979977	0.0001
GARCH(-2)	-0.144994	0.260506	-0.556584	0.5778

■ HÌNH 9.31: **Kết quả ước lượng mô hình GARCH(1,2).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.000466	0.000169	2.755727	0.0059
AR(1)	0.063211	0.020586	3.070623	0.0021
Variance Equation				
C	8.64E-07	2.53E-07	3.412790	0.0006
RESID(-1)^2	0.033254	0.012222	2.720783	0.0065
RESID(-2)^2	0.019890	0.013285	1.497123	0.1344
GARCH(-1)	0.937453	0.008442	111.0424	0.0000

■ HÌNH 9.32: **Kết quả ước lượng mô hình GARCH(1,8).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.000483	0.000171	2.830555	0.0046
AR(1)	0.063705	0.020731	3.072985	0.0021
Variance Equation				
C	3.82E-07	1.43E-07	2.675559	0.0075
RESID(-1)^2	0.034037	0.015426	2.206481	0.0274
RESID(-2)^2	0.053174	0.026067	2.039902	0.0414
RESID(-3)^2	-0.019638	0.032664	-0.601200	0.5477
RESID(-4)^2	-0.060870	0.028493	-2.136317	0.0327
RESID(-5)^2	0.023132	0.027853	0.830488	0.4063
RESID(-6)^2	0.008795	0.029842	0.294728	0.7682
RESID(-7)^2	0.045332	0.030715	1.475902	0.1400
RESID(-8)^2	-0.052934	0.023396	-2.262478	0.0237
GARCH(-1)	0.965005	0.007110	135.7332	0.0000

Các kết quả trên cho thấy chỉ có mô hình GARCH(1,1) là phù hợp nhất với dữ liệu của cổ phiếu ABC. Ở các mô hình GARCH(2,1) và GARCH(1,2) đều có hệ số không có ý nghĩa thống kê, trong khi đó, ở mô hình GARCH(1,8) có ba hệ số của biến trễ phần dư bình phương mang dấu âm và không có ý nghĩa thống kê.

BƯỚC 4: GARCH(1,1) HAY TGARCH(1,1)

Cuối cùng, chúng ta kiểm định xem giữa mô hình GARCH(1,1) và TGARCH(1,1) thì mô hình nào phù hợp hơn với dữ liệu của cổ phiếu ABC.

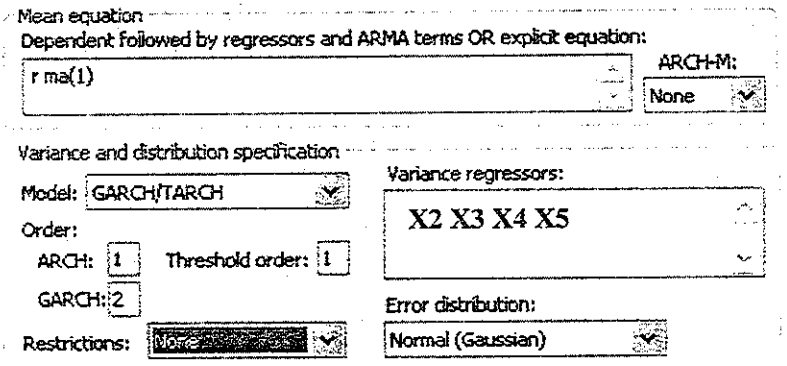
■ HÌNH 9.33: **Kết quả ước lượng mô hình TGARCH(1,1).**

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.000340	0.000169	2.014060	0.0440
AR(1)	0.058246	0.020600	2.827514	0.0047
Variance Equation				
C	6.05E-07	1.66E-07	3.636696	0.0003
RESID(-1)^2	0.013223	0.006336	2.087150	0.0369
RESID(-1)^2*(RESID(-1)<0)	0.049787	0.008948	5.563886	0.0000
GARCH(-1)	0.955519	0.006066	157.5123	0.0000

Hệ số hồi quy của biến tương tác có ý nghĩa rất cao chứng tỏ có sự khác biệt đáng kể giữa ảnh hưởng của các tin tức xấu và tin tức tốt lên cổ phiếu ABC. Điều này cũng cho thấy rằng mô hình TGARCH(1,1) là mô hình phù hợp nhất đối với việc dự báo suất sinh lợi trung bình và phương sai của cổ phiếu ABC.

Để phân tích các nhân tố ảnh hưởng rủi ro, chúng ta có thể đưa các biến giải thích (kể cả các biến giả) vào hộp thoại <Variance regressors>. Chẳng hạn, nếu ta muốn xem có phải các biến như giá vàng thế giới, giá dầu thế giới, hay một số chỉ số giá chứng khoán thế giới, v.v., ảnh hưởng đến phương sai của suất sinh lợi thị trường chứng khoán Việt Nam hay không. Tuy nhiên, việc lựa chọn các biến giải thích phù hợp cần phải dựa vào một cơ sở lý thuyết kinh tế phù hợp. Chúng ta có thể tham khảo thêm các nghiên cứu thực nghiệm để hiểu rõ hơn về các ứng dụng thú vị này.

■ HÌNH 9.34: **Mô hình hóa nhân tố ảnh hưởng rủi ro.**



TÓM TẮT CHƯƠNG 9

Chương này đã cung cấp cho chúng ta các công cụ tối thiểu nhằm dự báo rủi ro hoặc mức độ dao động của các chuỗi thời gian có độ dao động cao. Khác với các mô hình dự báo ứng dụng từ phương pháp kinh tế lượng cổ điển bằng phương pháp bình phương tối thiểu, các mô hình dự báo rủi ro thừa nhận rằng phương sai của số hạng nhiễu hoặc sai số có mối quan hệ với nhau, và như vậy các mô hình dự báo rủi ro xác định thông qua một tiếp cận ước lượng các hệ số theo các thành phần của phương sai theo thời gian và các độ trễ của nó như ARCH, GARCH, GARCH-M, và TGARCH dựa trên phương pháp maximum likelihood. Với các mô hình này cho chúng ta có thể ứng dụng dự báo rủi ro cho các chuỗi thời gian kinh tế và tài chính có độ nhạy cao như suất sinh lợi chứng khoán, tỷ giá hối đoái, lãi suất, giá dầu, giá vàng trên thị trường trong nước và quốc tế. Dĩ nhiên các mô hình này, nếu không có sự hỗ trợ của phần mềm phân tích dữ liệu thì khó lòng tìm ra kết quả dự báo mong đợi kịp thời trong một khoảng thời gian quá eo hẹp khi quan sát các biến số này.

Các mô hình ARCH là một bước phát triển có tính bổ sung quan trọng cho các mô hình ARIMA. Nói chung, sau khi thực hiện các mô hình ARIMA cho các chuỗi thời gian có tính biến động cao, nếu phương sai của phần dư thay đổi theo thời gian thì sẽ là hạn chế nếu chúng ta không mô hình hóa các ảnh hưởng ARCH. Hơn nữa, đối với những ai quan tâm đến việc quản lý rủi ro, đặc biệt là trong các công ty tài chính, thì các mô hình ARCH là một công cụ rất hữu ích cho việc nhận dạng các yếu tố rủi ro để các nhà quản lý có thể tìm ra các cơ chế chia sẻ rủi ro hợp lý.

CÂU HỎI VÀ BÀI TẬP

1. Dữ liệu trong tập tin **"ARCH1.xls"** chứa các dữ liệu về chỉ số giá chứng khoán của thị trường chứng khoán TP.HCM (VNI) và thị trường chứng khoán Hà Nội (HNX).
   a. Anh/Chị hãy xây dựng mô hình ARCH phù hợp để dự báo chỉ số VNI và HNX cho 6 ngày tiếp theo?
   b. Anh/Chị hãy dự báo xác suất để suất sinh lợi thị trường 'dương' tương ứng với các thời điểm dự báo trên?
   c. Anh/Chị hãy kiểm chứng rằng không tồn tại vấn đề bất cân xứng giữa tin tức tốt và tin tức xấu trên thị trường chứng khoán Việt Nam?
2. Dữ liệu trong tập tin **"ARCH2.xls"** chứa các dữ liệu về giá chứng khoán của một số cổ phiếu Blue-Chip trên thị trường chứng khoán TP.HCM.
   a. Anh/Chị hãy xây dựng mô hình ARCH phù hợp để dự báo giá của các cổ phiếu này cho 6 ngày tiếp theo?
   b. Anh/Chị hãy kiểm chứng rằng không tồn tại vấn đề bất cân xứng giữa tin tức tốt và tin tức xấu trên thị trường chứng khoán Việt Nam?
   c. Anh/Chị hãy cho biết loại cổ phiếu nào đúng với với giả thiết cho rằng "rủi ro càng cao, lợi nhuận càng nhiều"?
3. Dữ liệu trong tập tin **"ARCH3.xls"** chứa dữ liệu về một số loại lãi suất. Anh/Chị hãy xây dựng mô hình ARCH phù hợp để dự báo từng loại lãi suất cho 3 giai đoạn tiếp theo?
4. Sử dụng tập tin **"ARCH4.xls"**, Anh/Chị hãy xây dựng mô hình ARCH phù hợp để dự báo giá vàng thế giới, giá vàng Việt Nam, giá dầu thế giới, giá cao su, giá cà phê, và tỷ giá hối đoái VND/USD cho 3 giai đoạn tiếp theo?
5. Sử dụng tập tin **"ARCH5.xls"**, Anh/Chị hãy xây dựng mô hình ARCH phù hợp về nhân tố ảnh hưởng đến suất sinh lợi và rủi ro của VNI?

6. Sử dụng tập tin **"ARCH6.xls"**, Anh/Chị hãy xây dựng mô hình ARCH phù hợp về nhân tố ảnh hưởng đến suất sinh lợi và rủi ro của VNI?
7. Sử dụng tập tin **"ARCH7.xls"**, Anh/Chị hãy xây dựng mô hình ARCH phù hợp về nhân tố ảnh hưởng đến rủi ro ro của chỉ số FTSE?
8. Sử dụng tập tin **"ARCH8.xls"**, Anh/Chị hãy xây dựng mô hình ARCH phù hợp cho biết các cổ phiếu nào có ảnh hưởng đến rủi ro của cổ phiếu ACB?
9. Sử dụng tập tin **"ARCH9.xls"**, Anh/Chị hãy sử dụng mô hình TGARCH phù hợp để kiểm chứng có dấu hiệu ảnh hưởng bất cân xứng ở ba cổ phiếu (ở thị trường LONDON) này hay không?
10. Sử dụng tập tin **"ARCH10.xls"**, Anh/Chị hãy xây dựng mô hình ARCH phù hợp để dự báo các chỉ số giá chứng khoán sau đây: DJI, FCHI, MSCI, Kospi, Nikkei, HSI, và SSEC.
11. Sử dụng tập tin **"GAS.xls"**, Anh/Chị hãy xây dựng mô hình ARCH phù hợp để dự báo giá CP cho tháng 6/2009? Anh/Chị cho biết kết quả này có tốt hơn các kết quả trước đây hay không? Tại sao?
12. Sử dụng tập tin **"GAP.xls"**, Anh/Chị hãy xây dựng mô hình ARCH phù hợp để dự báo doanh số của GAP cho các tháng trong năm 2004? Anh/Chị cho biết kết quả này có tốt hơn các kết quả trước đây hay không? Tại sao?
13. Sử dụng tập tin **"MURPHY.xls"**, Anh/Chị hãy xây dựng mô hình ARCH phù hợp để dự báo doanh số bán lẻ toàn quốc trong năm 1996? Anh/Chị cho biết kết quả này có tốt hơn các kết quả trước đây hay không? Tại sao?
14. Sử dụng tập tin **"CCC.xls"**, Anh/Chị hãy xây dựng mô hình ARCH phù hợp để dự báo lượng khách hàng mới cho tháng sau? Anh/Chị cho biết kết quả này có tốt hơn các kết quả trước đây hay không? Tại sao?

CHƯƠNG 10 KIỂM SOÁT VÀ QUẢN LÝ QUY TRÌNH DỰ BÁO

Cho đến bây giờ chúng ta đã hiểu khá đầy đủ các phương pháp dự báo quan trọng đã và đang được giảng dạy và áp dụng phổ biến trên thế giới. Dựa trên cơ sở đó, chương này sẽ thảo luận một số vấn đề thực tế mà những người làm dự báo cũng như người sử dụng kết quả dự báo nên quan tâm khi áp dụng dự báo vào việc lập kế hoạch kinh doanh, chiến lược đầu tư, phân tích tài chính hoặc hoạch định chính sách kinh tế xã hội. Thứ nhất, chương này sẽ thảo luận một số hướng dẫn quan trọng để có các dự báo tốt. Thứ hai, quy trình dự báo sẽ được đánh giá lại trên cơ sở người học đã có hiểu biết nhất định về dự báo. Thứ ba, chúng tôi trình bày một số kinh nghiệm trong việc lựa chọn phương pháp dự báo thích hợp. Thứ tư, chương này sẽ trình bày các vấn đề thực tế trong quá trình thực hiện dự báo như giám sát quy trình dự báo, trách nhiệm dự báo và chi phí thực hiện dự báo. Thứ năm, chúng tôi sẽ phân tích tại sao nhiều tổ chức vẫn chưa áp dụng các phương pháp dự báo chính thức trong quá trình ra quyết định. Sau cùng, chúng tôi đề cập một cách ngắn gọn về vai trò của dự báo định tính trong toàn bộ quy trình thực hiện và sử dụng kết quả dự báo.

MỤC TIÊU HỌC TẬP

Sau khi học xong chương này, chúng ta kỳ vọng sẽ đạt được các nội dung sau đây:

- Các nhân tố then chốt quyết định kết quả dự báo
- Quy trình dự báo chuẩn mực
- Lựa chọn được các phương pháp dự báo thích hợp

- Cách thức giám sát tốt quy trình dự báo
- Xây dựng khung quản lý quy trình dự báo
- Trách nhiệm thực hiện dự báo
- Chi phí dự báo bao gồm những thành phần nào
- Tại sao nhiều tổ chức vẫn chưa áp dụng các kỹ thuật dự báo chính thức
- Vai trò của dự báo định tính trong giai đoạn tiền dự báo, thực hiện dự báo, và sử dụng kết quả dự báo.

NHÂN TỐ QUYẾT ĐỊNH KẾT QUẢ DỰ BÁO

Để có một kết quả dự báo tốt thì điều quan trọng là cần phải có sự trao đổi thường xuyên giữa người làm dự báo và người sử dụng kết quả dự báo. Tuy nhiên, đó chưa phải là tất cả. Trong bài nghiên cứu của Mentzer và cộng sự (1998) đã đưa ra bảy nhân tố quyết định sự thành công của dự báo (trong kinh doanh). Cho đến nay, bảy nhân tố này vẫn được các chuyên gia dự báo và những nhà quản lý doanh nghiệp đồng ý.

Nhân tố 1: Xác định đối tượng dự báo

Có hai vấn đề quan trọng mà các tổ chức phải quan tâm: (i) dự báo là một quy trình quản trị và (ii) việc phân biệt điều gì thì nên tiến hành dự báo.

Thứ nhất, dự báo nên được hiểu là một quy trình quản trị chứ không phải là một lập trình máy tính. Trong doanh nghiệp, dự báo có ý nghĩa hết sức quan trọng đối với các bộ phận sản xuất và tác nghiệp. Sẽ không thể có kế hoạch tốt về các vấn đề như quản lý tồn kho, nguyên vật liệu, nhân công, và các dịch vụ hậu cần khác nếu không có sự dự báo tương đối chính xác về doanh số (hoặc giá cả). Tuy nhiên, nhiều doanh nghiệp vẫn xem việc lựa chọn và phát triển các phần mềm máy tính phục vụ cho dự báo là quan trọng hơn cả. Họ luôn lạc quan tin tưởng rằng "nếu chúng ta có phầm mềm tốt, chúng ta sẽ có

dự báo tốt".Thực tế cho thấy nhiều doanh nghiệp trang bị các phần mềm phức tạp và tốn kém nhưng vẫn không thể có các dự báo chính xác vì không giám sát và quản lý tốt quy trình dự báo. Ngược lại, số khác lại rất thành công nhờ họ nhận biết tầm quan trọng của dự báo như một bộ phận của quy trình quản trị. Các công ty này có một nhóm độc lập hoặc thậm chí một phòng chuyên phụ trách toàn bộ quy trình dự báo. Nhóm dự báo có nhiệm vụ hướng dẫn các phương pháp và quy trình dự báo cho toàn công ty, tạo điều kiện khuyến khích trao đổi giữa nhiều bộ phận khác nhau như tài chính, nhân sự, marketing, và sản xuất. Như thế sẽ cải thiện được sự hiệu quả của dự báo.

Thứ hai, nhiều tổ chức chưa phân biệt giữa điều gì cần dự báo và điều gì không cần dự báo vì không hiểu đầy đủ mối quan hệ giữa dự báo, kế hoạch và thiết lập mục tiêu. Chẳng hạn, dự báo doanh số là giá trị ước lượng của doanh số tương lai với các điều kiện nhất định cho trước về môi trường kinh doanh. Kế hoạch doanh số là một quyết định hoặc cam kết của ban quản trị mà công ty sẽ thực hiện suốt trong giai đoạn kế hoạch. Mục tiêu doanh số là một kết quả mà toàn thể công ty phấn đấu đạt tới. Mỗi khía cạnh có mục đích riêng của nó. Mục đích chính của dự báo doanh số là giúp ban quản trị xây dựng các kế hoạch kinh doanh của doanh nghiệp. Mục đích của kế hoạch doanh số là hướng đến các quyết định quản trị tác nghiệp và chiến lược (mua nguyên vật liệu, tồn kho, kế hoạch nhân sự, kế hoạch hậu cần, v.v...) trong các điều kiện cụ thể của công ty. Mục tiêu doanh số chỉ nhằm cung cấp động cơ cho toàn thể thành viên của doanh nghiệp phấn đấu. Do vậy dự báo và kế hoạch có liên quan mật thiết với nhau nhưng dự báo là những tiên liệu có cơ sở nhằm phục vụ cho việc xây dựng kế hoạch có căn cứ, và khi kế hoạch có căn cứ rồi thì việc phấn đấu thực hiện kế hoạch sẽ khả thi hơn.

Nhân tố 2: Dự báo sự khác biệt giữa nhu cầu và khả năng

Một lỗi thường gặp là nhiều công ty thực hiện dự báo dựa trên khả năng họ có thể cung cấp hàng hóa và dịch vụ hơn là nhu cầu thực tế của khách hàng. Cách làm này có thể dẫn đến sự nhầm tưởng vào mức độ chính xác của dự báo. Ví dụ, nếu chỉ dựa vào số liệu quá khứ để dự báo theo khả năng cung ứng trước đây của công ty, thì giá trị dự báo là

1000 đơn vị, nhưng nhu cầu thực sự lên đến 1500 đơn vị. Dĩ nhiên, sản lượng sẽ được bán hết và dựa vào đó công ty đánh giá cao kết quả dự báo. Như vậy, cách dự báo này sẽ tạo ra một vòng luẩn quẩn của cái được dự báo sai trước đó. Và Mentzer (1998) cho rằng dự báo dựa vào nhu cầu sẽ cho biết khoảng cách giữa nhu cầu và khả năng cung ứng để doanh nghiệp có kế hoạch mở rộng hay thu hẹp sản xuất. Dự báo nhu cầu đúng hướng sẽ giúp công ty đưa ra các quyết định đúng đắn, dài hạn, gia tăng khả năng cạnh tranh và ảnh hưởng rất lớn đến thị phần của công ty. Chính vì thế, công ty nên kết hợp chặt chẽ giữa dự báo khả năng cung cấp và dự báo nhu cầu để có các quyết định tốt nhất cho tương lai.

Nhân tố 3: Dự báo cần trao đổi, hợp tác, và cộng tác

Tương tác thông tin từ các thành viên khác nhau trong các bộ phận chức năng khác nhau sẽ góp phần cải thiện đáng kể kết quả dự báo chung của toàn công ty. Để làm được điều này, cần có sự tương tác qua lại ở tất cả các bộ phận trong công ty. Việc này cũng cần phân biệt rõ ba cấp độ: trao đổi (communication), hợp tác (cooperation), và cộng tác (collaboration). Những công ty với cơ cấu ít phức tạp hơn thì chỉ cần trao đổi bằng các báo cáo một chiều (bộ phận chịu trách nhiệm dự báo chỉ cần thông báo kết quả dự báo). Với cấp độ hợp tác, đại diện các bộ phận chức năng phải gặp gỡ để thảo luận quá trình dự báo. Tuy nhiên, thực tế cho thấy rằng bộ phận nào chủ trì dự báo thường có xu hướng lấn át trong các cuộc thảo luận và thuyết phục các bộ phận khác chấp nhận kết quả họ đã dự báo. Hợp tác ưu việt hơn hình thức trao đổi vì nó mở ra cơ hội cho thảo luận giữa các bộ phận chức năng, tuy nhiên không hiệu quả bằng hình thức cộng tác. Trong hình thức cộng tác, các bộ phận chức năng có quyền xem xét, bàn bạc như nhau. Kết quả dự báo từ sự cộng tác tích cực sẽ có sự nhất trí cao, nên việc triển khai dự báo và thực hiện kế hoạch sẽ trở nên dễ dàng hơn. Để tăng cường khả năng cộng tác cần đảm bảo hai vấn đề. Thứ nhất, cần tạo niềm tin giữa các bộ phận chức năng. Thứ hai, thiết lập một cơ chế để kết nối các bộ phận chức năng này lại với nhau. Để làm được điều này cần phải có một bộ phận chuyên trách dự báo riêng và cần sự cam kết mạnh mẽ của ban quản trị cấp cao trong doanh nghiệp.

Nhân tố 4: Dự báo cần loại bỏ những "ốc đảo"

Những "ốc đảo" là những bộ phận khác nhau trong cùng công ty nhưng lại thực hiện các chức năng gần như chồng chéo nhau hoặc tương tự nhau. Mỗi bộ phận chức năng có một quy trình riêng biệt, vì thế thực hiện nhiều công việc thừa thãi và cũng thường có các trách nhiệm giống nhau. Bởi vì thông thường, những ốc đảo được hỗ trợ bởi hệ thống máy tính độc lập, nên thông tin ở các ốc đảo khác nhau không được chia sẻ và sử dụng hiệu quả dưới các hình thức phối hợp. Chẳng hạn, các bộ phận tiến hành các dự báo mang tính độc lập cùng tồn tại ở bộ phận dịch vụ hậu cần, kế hoạch sản xuất, tài chính và marketing, và như vậy kết quả dự báo sẽ mang tính "ốc đảo". Có thể dễ dàng nhận thấy nguyên nhân của vấn đề này là do thiếu sự cộng tác, dẫn đến thiếu tin tưởng giữa các bộ phận với nhau, và mỗi bộ phận tự có xu hướng thực hiện các dự báo cho riêng mình.

Tồn tại các ốc đảo gây ảnh hưởng xấu đến thành quả chung của doanh nghiệp, rất tốt kém, khó phát huy khả năng cộng tác, và thiếu sự tin cậy vào quy trình dự báo. Các dự báo được xây dựng theo cách này cũng thường không chính xác và có xu hướng mâu thuẫn nhau. Do mỗi bộ phận có quy trình dự báo riêng và có một hệ thống máy tính riêng, nên nếu có chia sẻ dữ liệu thì dữ liệu chỉ được chuyển qua lại dưới dạng thủ công, vì thế khả năng sai sót rất cao. Chưa kể đến giả định về dự báo cũng sẽ khác nhau và dự báo ở mỗi bộ phận có sự sai lệch riêng, khiến các kết quả có thể mâu thuẫn nhau và những bộ phận khác không sử dụng được. Hơn nữa, nếu vấn đề này tiếp tục tồn tại, ban quản trị cao cấp sẽ không tin tưởng vào kết quả dự báo và vì thế dự báo khó có cơ hội tồn tại trong quy trình ra quyết định của doanh nghiệp.

Để giải quyết vấn đề ốc đảo, cần phải thiết lập một quy trình dự báo duy nhất dưới sự hỗ trợ của "hạ tầng dự báo" và cơ sở dữ liệu thống nhất. Quy trình này nên có phần mềm kết nối với các hệ thống thông tin trong nội bộ công ty. Một khi đã có hạ tầng dự báo hoàn chỉnh, công ty cần tổ chức công tác đào tạo để mọi người hiểu đầy đủ quy trình và hệ thống dự báo. Nhờ loại bỏ các ốc đảo mà doanh

nghiệp có thể giảm chi phí đáng kể, kết quả dự báo chính xác hơn, và được tin cậy cũng như đồng thuận hơn.

Nhân tố 5: Sử dụng các phương pháp dự báo hiệu quả

Nhiều doanh nghiệp có khuynh hướng chỉ dựa vào hoặc các phương pháp định tính hoặc các phương pháp định lượng để đưa ra kết quả dự báo cuối cùng. Thực tế cho thấy kết hợp hai phương pháp lại có thể cho kết quả dự báo tốt hơn. Tuy nhiên, để mang lại hiệu quả cao hơn, các phương pháp phải được hiểu và sử dụng một cách sáng suốt phù hợp với mỗi điều kiện môi trường kinh doanh đặc thù. Thông thường, dự báo định tính đóng vai trò quan trọng trong giai đoạn tiền dự báo để xác định mục tiêu dự báo, cơ sở dữ liệu cho dự báo định lượng, và trong giai đoạn thảo luận chuyên sâu các kết quả dự báo định lượng. Kết quả dự báo định lượng đóng vai trò quan trọng trong việc cung cấp thông tin cho các phán xét dựa vào ý kiến chuyên gia để đưa ra các quyết định thích hợp.

Sử dụng các công cụ một cách tốt nhất, đòi hỏi cần phải biết khi nào nên dùng loại nào, và biết tận dụng ưu điểm của từng phương pháp. Thông thường quy trình dự báo nên thực hiện như sau: sử dụng các mô hình chuỗi thời gian để dự báo xu thế và mùa vụ, sử dụng hồi quy để dự báo các mối quan hệ nhân quả hoặc dự báo hệ số co giãn, và ý kiến định tính từ các bộ phận chức năng để điều chỉnh kết quả dự báo định lượng ban đầu.

Nhân tố 6: Làm cho dự báo trở nên quan trọng

Hầu hết các doanh nghiệp điều biết rằng dự báo doanh số là một chức năng quan trọng, nhưng hiếm khi họ có chính sách thể hiện dự báo thực sự quan trọng cho thành công của doanh nghiệp vì ta biết rằng, luôn tồn tại một khoảng cách lớn giữa nói và làm của những người quản lý. Nhiều công ty đã nói với những người làm dự báo rằng "dự báo rất quan trọng" nhưng không hề khen tặng mỗi khi họ có thành tích tốt hoặc cũng không có hình phạt nào mỗi khi họ dự báo kém cỏi.

Một cách làm cho dự báo trở nên quan trọng trong công ty là làm cho cả người sử dụng lẫn người thực hiện quen với toàn bộ quy trình dự báo. Nếu không hiểu hết quy trình, những người làm dự báo trong công ty sẽ không lường hết được tác hại của kết quả dự báo không chính xác mà họ đưa ra. Và người sử dụng cũng sẽ đánh giá thấp những nỗ lực của người dự báo, ít tin tưởng vào kết quả của họ.

Để giải quyết vấn đề trên, doanh nghiệp cần thực hiện hai việc sau đây. Thứ nhất, đào tạo những ai có liên quan đến quy trình dự báo để họ hiểu hơn, có trách nhiệm hơn, và nhận thức được tầm quan trọng của các dự báo. Thứ hai, cần phải đưa thêm tiêu chí đánh giá kết quả dự báo vào quá trình đánh giá mức độ hoàn thành công việc của cả cấp quản lý và nhân viên các bộ phận chức năng. Ngoài ra, công ty cần phải thiết lập một khung quản lý quy trình dự báo thống nhất. Khung dự báo này sẽ được trình bày một cách khái quát ở phần sau của chương này.

Nhân tố 7: Quy trình và kết quả dự báo cần đo lường và đánh giá

Trước hết, doanh nghiệp cần có hệ thống các tiêu chí và thang đo thành quả của dự báo để có thể đánh giá mức độ hoàn thành công việc của từng cá nhân. Nếu không thể đo lường và theo dõi kết quả dự báo, thì không thể nhận biết được những thay đổi trong việc xây dựng và ứng dụng các dự báo có đóng góp như thế nào vào sự thành công của doanh nghiệp. Nhiều kết quả nghiên cứu về dự báo cho rằng đa số các doanh nghiệp không thực hiện việc giám sát, theo dõi kết quả dự báo. Việc đánh giá kết quả dự báo sẽ giúp doanh nghiệp nhận biết được lý do tại sao một dự báo thành công hoặc thất bại, từ đó có những biện pháp cải thiện quy trình dự báo cho các dự báo trong tương lai. Bảng 10.1 tóm tắt các nhân tố then chốt để có một dự báo tốt.

■ BẢNG 10.1: **Bảy nhân tố then chốt để có một dự báo tốt hơn.**

Nhân tố	Nguyên nhân	Giải pháp	Kết quả
Xác định đối tượng dự báo	• Xem hệ thống máy tính quan trọng hơn quy trình quản trị và kiểm soát • Không phân biệt giữa dự báo, kế hoạch và mục tiêu	• Thiết lập nhóm dự báo • Thực hiện các hệ thống kiểm soát quản trị trước khi chọn phần mềm dự báo • Lập kế hoạch từ các dự báo • Phân biệt dự báo và mục tiêu	• Một môi trường trong đó dự báo được xem như một chức năng đặc biệt quan trọng • Đề cao độ chính xác
Dự báo sự khác biệt giữa nhu cầu và khả năng	• Dự báo "khả năng cung ứng" *chứ không phải* nhu cầu thực sự • Ảo tưởng rằng các kết quả dự báo của mình "quá chính xác"	• Nhận dạng nguồn thông tin • Xây dựng hệ thống thu thập dữ liệu về nhu cầu	• Kế hoạch vốn và các dịch vụ *khách hàng* được cải thiện
Dự báo cần trao đổi, hợp tác, và cộng tác	• Tăng cường các nỗ lực dự báo • Hoài nghi dự báo "chính thức" • Không hiểu tác động xuyên suốt toàn công ty	• Thiết lập cách tiếp cận dự báo có sự kết hợp giữa các bộ phận chức năng • Thành lập nhóm dự báo độc lập để hỗ trợ sự cộng tác giữa các bộ phận chức năng	• Tất cả các thông tin thích hợp được sử dụng cho dự báo • Những người sử dụng tin tưởng các kết quả dự báo • Loại bỏ các ốc đảo Các dự báo thích hợp và chính xác hơn

Nhân tố	Nguyên nhân	Giải pháp	Kết quả
Dự báo cần loại bỏ những "ốc đảo"	• Thông tin không đáng tin và không thích hợp, những người cần dự báo tự tạo ra kết quả dự báo riêng cho mình	• Xây dựng một cơ sở hạ tầng dự báo duy nhất • Đào tạo cho cả người sử dụng lẫn người làm dự báo	• Các dự báo tin cậy, thích hợp và chính xác hơn • Đầu tư tối ưu vào hệ thống thông tin, liên lạc
Sử dụng các công cụ dự báo một cách sáng suốt	• Chỉ dựa vào các phương pháp định lượng hoặc định tính • Chi phí/Lợi ích của thông tin có thêm	• Kết hợp các phương pháp định tính và định lượng • Nhận dạng các nguồn cải thiện độ chính xác và tăng sai sót • Đưa ra các chỉ dẫn	• Cải thiện quy trình
Làm cho dự báo trở nên quan trọng	• Không có trách nhiệm giải trình các kết quả dự báo kém • Những người làm dự báo không hiểu dự báo dùng làm gì	• Đào tạo người làm dự báo để họ hiểu hậu quả của những dự báo kém • Đưa tiêu chí dự báo vào kế hoạch thực hiện cá nhân và hệ thống khen thưởng	• Những người làm dự báo thực hiện dự báo nghiêm túc hơn • Cố gắng cải thiện độ chính xác • Kết quả chính xác và tin cậy hơn
Quy trình và kết quả dự báo cần đo lường và đánh giá	• Không biết doanh nghiệp có đang tốt lên hay không	• Thiết lập các thước đo đa phương diện	• Kết quả dự báo có thể được tính trong các kế hoạch thực hiện của cá nhân

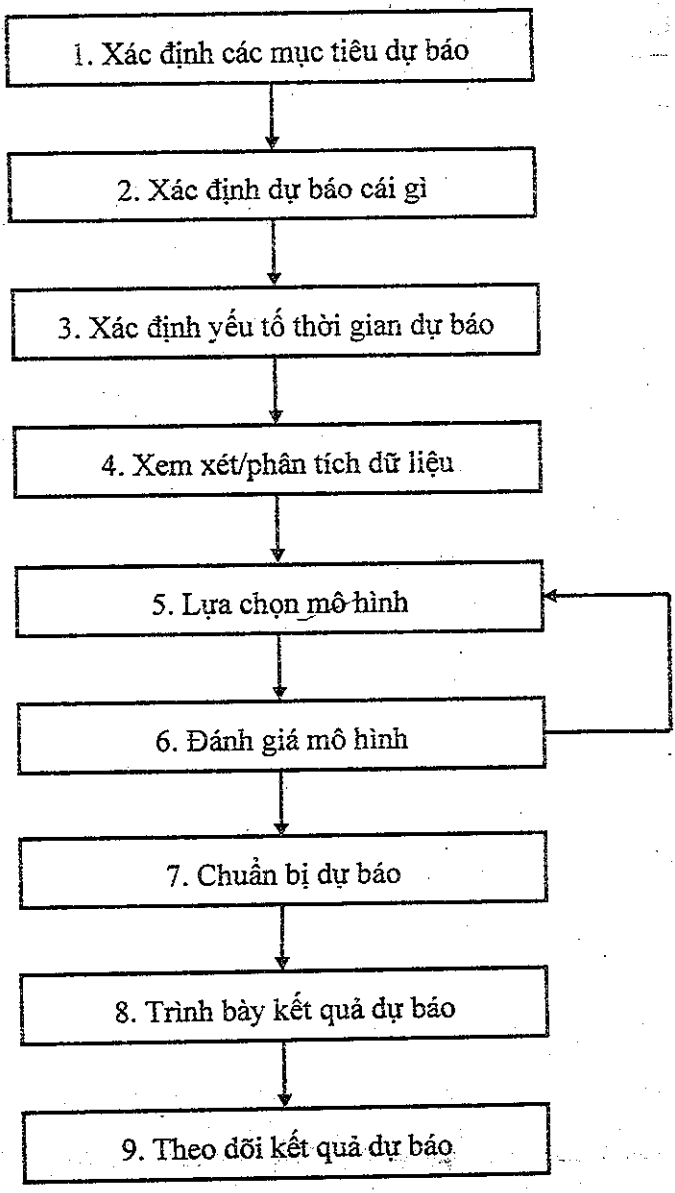
Nhân tố	Nguyên nhân	Giải pháp	Kết quả
	• Mức độ chính xác không được đo lường ở các mức độ kết hợp thích hợp	• Đưa vào các thước đo đa cấp độ	• Các nguồn gây sai số dự báo có thể được tách riêng và làm mục tiêu cải thiện
	• Không thể tách riêng các nguồn gây ra sai số dự báo	• Đo lường độ chính xác bất cứ khi nào và ở đâu các dự báo được điều chỉnh	• Gia tăng sự tin cậy trong quá trình dự báo

Nguồn: Mentzer, 1998.

ĐÁNH GIÁ LẠI QUY TRÌNH DỰ BÁO

Thông thường một quy trình dự báo có thể bao gồm chín bước như đã trình bày ở chương 1, bao gồm (1) Xác định các mục tiêu dự báo, (2) Xác định đối tượng dự báo, (3) Xác định yếu tố thời gian dự báo, (4) Thu thập, khảo sát dữ liệu, (5) Lựa chọn mô hình dự báo, (6) Đánh giá kết quả dự báo của từng mô hình để chọn ra mô hình tốt nhất, (7) Chuẩn bị dự báo (với mô hình tốt nhất), (8) Trình bày kết quả dự báo, và (9) Theo dõi kết quả dự báo. Hình 10.1 thể hiện quy trình dự báo đã mô tả ở đây.

HÌNH 10.1: **Quy trình thực hiện dự báo.**



Nguồn: Wilson, 2007, pp.430.

Trong suốt quy trình đó đòi hỏi phải có sự trao đổi qua lại liên tục giữa những người làm dự báo và những người sử dụng kết quả dự báo. Cũng cần lưu ý thêm rằng, một quy trình dự báo có thể được thực hiện ở hai giai đoạn khác nhau - một diễn ra chủ yếu ở cấp chiến lược và một diễn ra ở cấp tác nghiệp. Ở cấp độ chiến lược, các quyết định về dự báo cái gì, dự báo được sử dụng như thế nào, và ai chịu trách nhiệm thực hiện dự báo được đưa ra. Trong giai đoạn này, các thông tin định tính có ý nghĩa rất quan trọng. Ở cấp độ tác nghiệp, những người có liên quan sẽ thu thập dữ liệu, thực hiện dự báo, báo cáo kết quả, và theo dõi đánh giá dự báo. Quy trình dự báo cũng giống các quy trình khác ở chỗ luôn luôn phải được quản lý và giám sát chặt chẽ. Dĩ nhiên, hai cấp độ này không thể tách rời nhau vì cả quy trình dự báo phải được đảm bảo có sự trao đổi, hợp tác và cộng tác trong toàn bộ doanh nghiệp. Trong suốt quy trình thực hiện dự báo cần phải kết hợp hài hòa giữa các phương pháp dự báo định lượng và các phương pháp dự báo định tính. Ý kiến đánh giá của các chuyên gia luôn cần thiết trong việc xem xét và lựa chọn các phương pháp dự báo thích hợp. Ngoài ra, nên đưa ra nhiều câu hỏi để xem việc quản lý quy trình dự báo có được thực hiện tốt không. Dưới đây là một số câu hỏi điển hình:

- Tại sao cần dự báo?
- Ai sử dụng kết quả dự báo, và cụ thể là họ cần gì?
- Dự báo chi tiết hay tổng hợp đến mức nào, và độ dài thời gian của dự báo là bao nhiêu?
- Dữ liệu sẵn có là gì, và dữ liệu đó có đủ để thực hiện dự báo mong muốn không?
- Chi phí dự báo là bao nhiêu?
- Mức độ chính xác mong muốn của dự báo là bao nhiêu?
- Kết quả dự báo có kịp cho quá trình ra quyết định hay không?

- Người làm dự báo có hiểu rõ kết quả dự báo sẽ được sử dụng như thế nào trong tổ chức hay không?

- Đã có sẵn quy trình phản hồi để đánh giá dự báo sau khi được thực hiện và để điều chỉnh quy trình dự báo thích hợp hơn hay chưa?

LỰA CHỌN CÁC PHƯƠNG PHÁP DỰ BÁO THÍCH HỢP

Sau khi đã hiểu các phương pháp dự báo khác nhau, các chuyên gia hay tổ chức dự báo cần một khung tổng quát giúp xác định khi nào sử dụng phương pháp nào. Khoa học dự báo đã đưa ra một số qui tắc giúp chúng ta có thể dễ dàng xác định các phương pháp dự báo thích hợp. Nếu chúng ta hiểu cách sử dụng các phương pháp đã được trình bày trong giáo trình này, thì đã có một khởi đầu rất tốt cho việc xác định nên sử dụng phương pháp dự báo nào trong những trường hợp cụ thể. Ví dụ, nếu chúng ta chuẩn bị dự báo doanh số của một sản phẩm theo quí có tính mùa vụ, thì chúng ta có thể sẽ sử dụng các phương pháp chuyên xử lý các dao động có tính mùa vụ như phương pháp Winters, phương pháp phân tích, các mô hình SARIMA. Có nhiều khía cạnh quyết định việc lựa chọn phương pháp dự báo, nhưng ta thường chỉ tập trung vào ba khía cạnh dữ liệu, thời gian và nhân sự.

Đối với dữ liệu, ta xem xét loại và số lượng dữ liệu sẵn có cũng như dạng dữ liệu thể hiện (xu thế, chu kỳ, mùa vụ, v.v...). Phần lớn các mô hình dự báo chuỗi thời gian quan trọng đều đòi hỏi dữ liệu phải có tính dừng. Chính vì thế, người làm dự báo cần dành thời gian để hiểu thế nào là một chuỗi dừng, phương pháp kiểm định tính dừng như đã được trình bày ở các chương 3 và chương 8. Nếu các mục đích dự báo của doanh nghiệp liên quan đến việc xác định hệ số co giãn hoặc nhóm các nhân tố ảnh hưởng đến biến số cần dự báo thì các mô hình kinh tế lượng là sự lựa chọn tốt nhất. Việc xác định nhóm các biến nào cần đưa vào cơ sở dữ liệu cần phải có sự nhất trí cao của các chuyên gia (có thể dựa trên kinh nghiệm thực tế hoặc cơ sở lý thuyết kinh tế nền tảng). Chẳng hạn, một công ty kinh doanh các sản phẩm

khí có thể cần xác định các nhân tố ảnh hưởng đến giá khí như giá dầu, giá vàng, chỉ số giá chứng khoán, lãi suất, chỉ số giá của một số ngành công nghiệp quan trọng, và các biến đại diện cho yếu tố chính trị. Dự báo bằng các mô hình hồi quy được trình bày ở chương 7, trong đó người làm dự báo cần đảm bảo một mô hình tốt trước khi có thể sử dụng cho các mục đích dự đoán của mình.

Đối với khía cạnh thời gian, ta quan tâm đến độ dài dự báo. Đối với khía cạnh nhân sự, nên xem xét kiến thức cơ bản cần thiết của người làm và người sử dụng dự báo. Bảng 10.2 là một bảng tổng hợp hướng dẫn lựa chọn một phương pháp dự báo thích hợp cho từng trường hợp cụ thể. Khía cạnh độ dài dự báo có liên quan đến mục đích sử dụng và cấp quản lý khác nhau trong doanh nghiệp. Những người quản lý cấp thấp trong doanh nghiệp có lẽ quan tâm đến việc phân tích chuỗi thời gian về doanh số bán theo tháng với dữ liệu được thu thập trong khoảng bốn năm qua. Từ kết quả phân tích, kết hợp với phán đoán về xu hướng vận động của các thành phần chuỗi thời gian, dự báo về doanh số theo tháng cho năm sau có thể được đưa ra làm cơ sở cho việc lập kế hoạch sản xuất hằng tháng của doanh nghiệp. Những người quản lý cấp trung có thể sử dụng cùng chuỗi dữ liệu thời gian như thế để phân tích doanh số cho khoảng tám năm qua và dự báo doanh số cho năm năm tới nhằm dùng cho mục đích dự trù các nhu cầu đầu tư vốn cho doanh nghiệp. Đồng thời, nhóm những người quản lý cấp cao có thể sử dụng các dự báo định tính như phương pháp chuyên gia, kết hợp với xây dựng kịch bản để đánh giá vị trí hiện tại của doanh nghiệp trên thị trường và xem xét khả năng những thay đổi công nghệ hoặc xã hội trong tương lai có mang lại những cơ hội quý báu hoặc đe dọa thị phần trong hai mươi năm tới hay không.

■ BẢNG 10.2: **Lựa chọn phương pháp dự báo thích hợp.**

Phương pháp dự báo	Dạng dữ liệu	Lượng dữ liệu	Độ dài dự báo
Phương pháp định tính			
• Tổng hợp lực lượng bán hàng	Bất kỳ	Ít	Ngắn, trung hạn
• Khảo sát khách hàng	Bất kỳ	Không	Trung, dài hạn
• Ý kiến ban quản lý	Bất kỳ	Ít	Bất kỳ
• Ý kiến chuyên gia	Bất kỳ	Ít	Dài hạn
Dự báo thô	Dừng*	1 hoặc 2	Rất ngắn
Bình quân di động	Dừng*	Ít nhất bằng hệ số trượt	Rất ngắn
San mũ			
• Giản đơn	Dừng*	5 – 10	Ngắn hạn
• Holt	Xu thế tuyến tính	10 – 15	Ngắn, trung hạn
• Winter	Xu thế và mùa vụ	Ít nhất là 4 – 5 quan sát/mùa vụ	Ngắn, trung hạn
Hồi quy			
Xu thế	Tuyến tính/phi tuyến có hoặc không có yếu tố mùa	Tối thiểu 10/4 hoặc 5 quan sát/mùa nếu có yếu tố mùa	Ngắn, trung hạn
Nhân quả	Tất cả các dạng dữ liệu	Tối thiểu 10 quan sát/biến giải thích	Ngắn, trung và dài hạn
Phân tích chuỗi thời gian	Xu thế, mùa vụ, và chu kỳ	Đủ lớn để có thể phát hiện các đỉnh và đáy trong chu kỳ	Ngắn, trung, và dài hạn
ARIMA/SARIMA	Dừng*	Tối thiểu 50	Ngắn, trung và dài hạn

* Kể cả các chuỗi sau khi đã biến đổi thành chuỗi dừng .

Nguồn: Holton Wilson & Barry Keating, 2007, tr.436.

Các phương pháp phân tích dữ liệu và dự báo quan trọng đã được trình bày trong giáo trình này có thể được tóm tắt như trong Bảng 10.3. Trong bảng này, chúng ta sẽ đánh giá từng nhóm phương pháp dự báo dựa trên các khía cạnh như phạm vi áp dụng, chi phí ứng dụng, và vai trò của máy tính và các phần mềm chuyên dụng.

■ BẢNG 10.3: **Đánh giá các phương pháp dự báo thích hợp.**

Phương pháp dự báo	Ứng dụng	Chi phí	Máy tính
Bình quân di động	Dự báo ngắn hạn cho các hoạt động tác nghiệp như tồn kho, lên chương trình, kiểm soát, định giá, lên lịch các khuyến mãi đặc biệt, giá chứng khoán hoặc các chỉ tiêu kinh tế có độ nhạy cao như tỉ giá hối đoái, lãi suất, hoặc giá dầu.	Thấp	Không
San mũ	Dự báo ngắn hạn cho các hoạt động tác nghiệp như tồn kho, lên chương trình, kiểm soát, định giá, lên lịch các khuyến mãi đặc biệt, giá chứng khoán hoặc các chỉ tiêu kinh tế có độ nhạy cao như tỉ giá hối đoái, lãi suất, hoặc giá dầu.	Thấp	Cần
Phương pháp phân tích	Dự báo trung hạn cho việc hoạch định đầu tư nhà xưởng, thiết bị, tài trợ, phát triển sản phẩm mới, các phương pháp mới của dây chuyền sản xuất; dự báo ngắn hạn về nhân sự, quảng cáo, tồn kho, tài trợ, kế hoạch sản xuất.	Thấp	Không
Các mô hình tự hồi quy	Dự báo độ co giãn, dự báo ngắn đến trung hạn về nhu cầu sản phẩm dịch vụ trong nghiên cứu thị trường, chiến lược tiếp thị, sản xuất, nhân sự, hoạch định đầu tư.	Thấp đến trung bình	Cần

Phương pháp dự báo	Ứng dụng	Chi phí	Máy tính
ARIMA	Dự báo ngắn đến trung hạn cho các biến kinh tế theo thời gian, giá cả, tồn kho, sản xuất, chứng khoán và doanh số, hoặc các chỉ tiêu kinh tế có độ nhạy cao như tỉ giá hối đoái, lãi suất, hoặc giá dầu.	Cao	Cần

Nguồn: Hanke, 2005, pp.487-488.

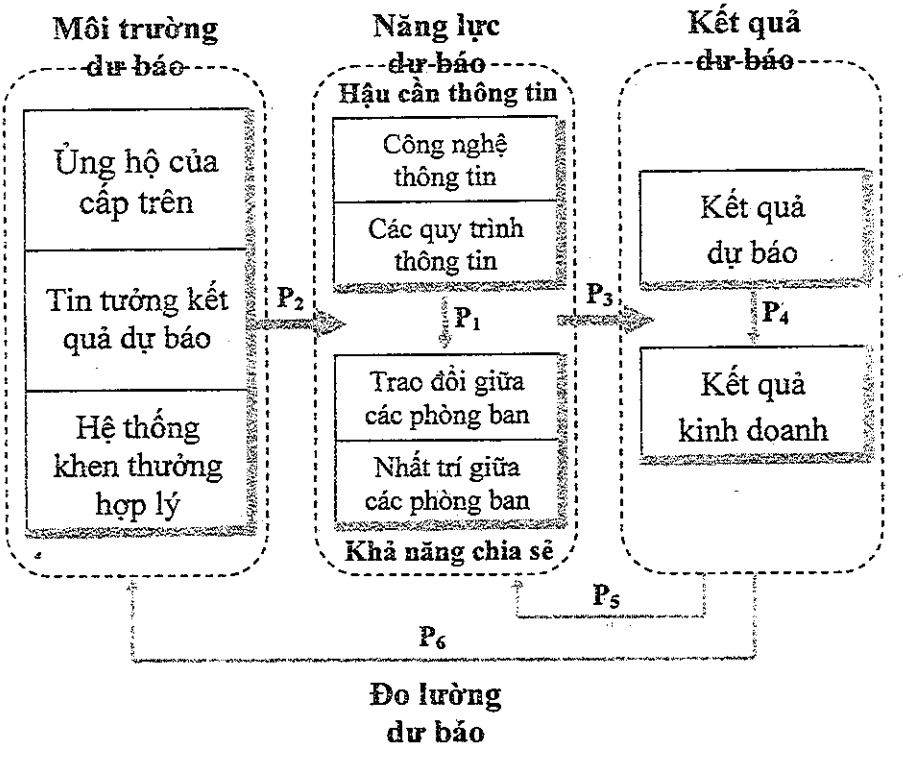
XÂY DỰNG KHUNG QUẢN LÝ QUY TRÌNH DỰ BÁO

Davis và Mentzer (2007) đã đưa ra một khung quản lý quy trình dự báo toàn diện cho việc quản lý quy trình dự báo doanh số. Mặc dù mục đích chính của khung quản lý này chỉ giới hạn trong việc dự báo doanh số, nhưng nó vẫn phù hợp cho các loại dự báo khác như dự báo giá cả, dự báo sản xuất, hoặc dự báo nhu cầu. Bốn thành phần chủ yếu trong quy trình này bao gồm: (1) Thiết lập một môi trường dự báo, (2) Phát triển năng lực dự báo, (3) Đánh giá các kết quả dự báo, và (4) Đo lường và giám sát kết quả dự báo. Toàn bộ quy trình này được trình bày trong Hình 10.2.

Năng lực dự báo bao gồm hai thành phần cơ bản là hậu cần thông tin (khả năng về cơ sở dữ liệu và phần mềm phân tích dữ liệu) và khả năng hiểu biết và cộng tác giữa các bộ phận chức năng (khả năng chia sẻ). Một hậu cần thông tin tốt phải dựa trên một hệ thống công nghệ thông tin hoàn chỉnh (phần mềm, phần cứng, và con người) và các quá trình thu thập, chuyển hóa dữ liệu (các nguồn thông tin nội bộ và các nguồn thông tin bên ngoài, kể cả các khảo sát khách hàng). Ngoài ra, các quá trình dự báo còn được chia thành các phân tích dự báo nội bộ và các phân tích dự báo bên ngoài (các chuyên gia phân tích ở các trường đại học, các viện nghiên cứu). Khả năng hiểu biết và cộng tác được thể hiện qua khả năng

trao đổi và nhất trí giữa các bộ phận về dữ liệu đầu vào, phương pháp dự báo, và kết quả dự báo. Giữa hai thành phần này có mối quan hệ chặt chẽ với nhau thông qua giả thiết P_1 "hậu cần thông tin có mối quan hệ tích cực với khả năng chia sẻ giữa các bộ phận chức năng".

■ HÌNH 10.2: **Khung quản lý quy trình dự báo.**



Nguồn: Davis và Mentzer, 2007.

Môi trường dự báo bao gồm sự ủng hộ của cấp trên, sự tin tưởng vào kết quả dự báo trong việc ra quyết định, và hệ thống khen thưởng hợp lý dành cho những bộ phận tham gia vào quá trình dự báo. Nói chung, "môi trường dự báo càng tích cực thì năng lực dự báo của doanh nghiệp càng được phát huy" (P_2).

Kết quả dự báo bao gồm việc đánh giá các kết quả dự báo và các kết quả kinh doanh của doanh nghiệp. Việc đánh giá kết quả dự báo cần dựa trên cả các tiêu chí đánh giá độ chính xác dự báo nội bộ (các tiêu chí thống kê) và mức độ đáp ứng nhu cầu khách hàng mục tiêu (mức độ hàng lòng của khách hàng về chất lượng dịch vụ). Như vậy, kết quả dự báo không chỉ nhằm gia tăng giá trị cho các khách hàng nội bộ (người sử dụng, các phòng ban, v.v...) mà còn gia tăng giá trị cho các khách hàng bên ngoài (nhà cung cấp, khách hàng, các chuyên gia, v.v...). Chúng ta kỳ vọng rằng "năng lực dự báo có ảnh hưởng tích cực đến kết quả dự báo" (P_3). Kết quả dự báo tốt (sai số dự báo thấp) giúp giảm chi phí tồn khi, tăng lợi nhuận, cải thiện chuỗi cung ứng, đáp ứng cao nhất nhu cầu khách hàng. Nhiều nghiên cứu thực nghiệm đã cho thấy rằng "mức độ chính xác của dự báo có mối quan hệ tích cực tới kết quả hoạt động của doanh nghiệp" (P_4).

Cuối cùng, kết quả dự báo (và kết quả hoạt động kinh doanh) có ảnh hưởng tích cực lên năng lực dự báo và môi trường dự báo (P_5, P_6).

Tóm lại, để có kết quả dự báo tốt, doanh nghiệp cần tạo dựng một môi trường tích cực trong đó sự ủng hộ và tin cậy của ban quản trị cấp trung/cao là hết sức cần thiết. Các nhà quản trị trung/cao hiểu về tầm quan trọng của dự báo là một điều kiện thuận lợi để đưa dự báo ngày càng gần hơn với các hoạt động của doanh nghiệp. Trên cơ sở đó, các doanh nghiệp phải xây dựng một năng lực dự báo nhất định, trong đó cần quan tâm đến cơ sở dữ liệu, phần mềm hỗ trợ, chuyên viên phân tích, và sự công tác tích cực, xuyên suốt trong toàn công ty. Một khi đã có môi trường và năng lực dự báo tốt, thì kết quả dự báo trở nên hữu ích hơn cho việc ra quyết định của doanh nghiệp.

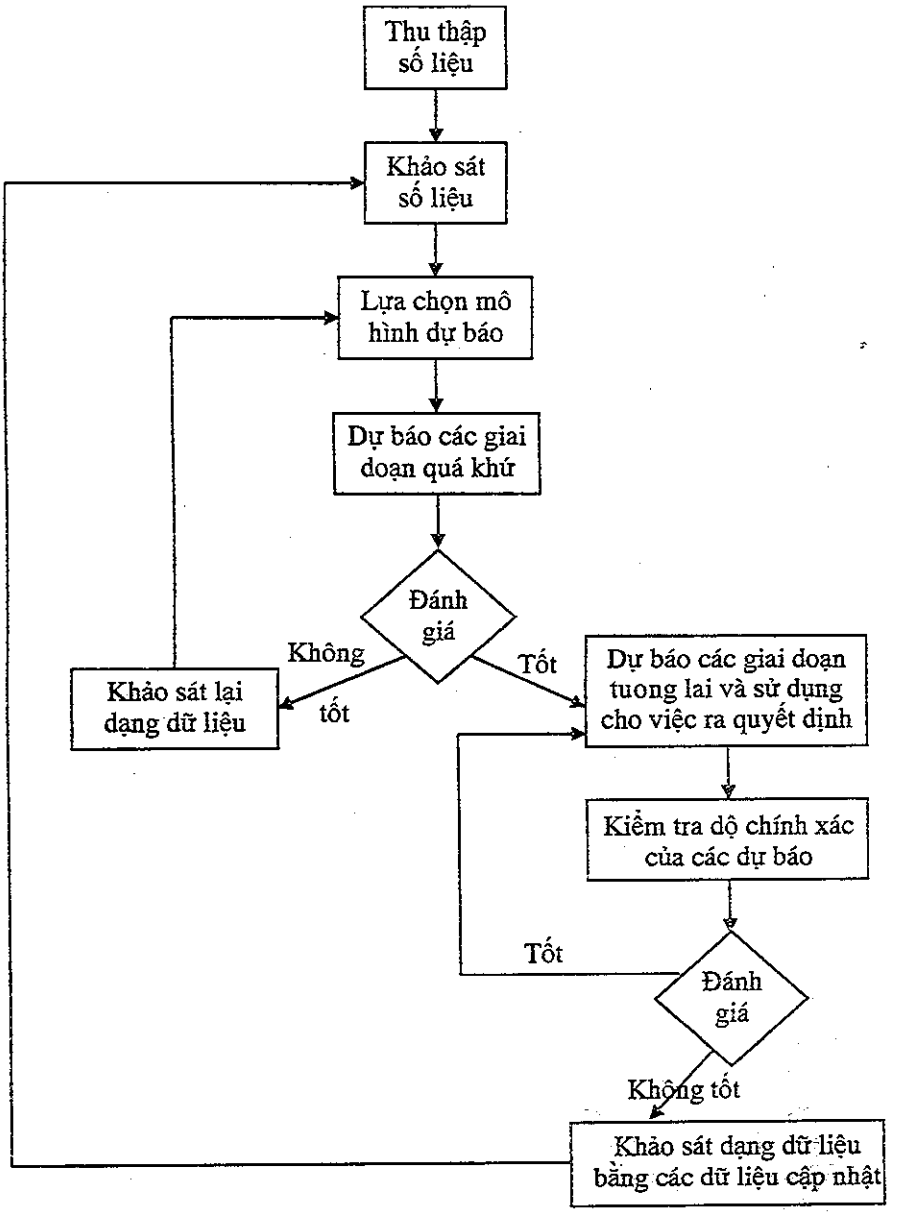
GIÁM SÁT KẾT QUẢ DỰ BÁO

Thu thập dữ liệu và lựa chọn mô hình dự báo thích hợp chỉ là một trong những bước đầu tiên trong quy trình dự báo. Giáo trình này đã trình bày một số bước trong giai đoạn tác nghiệp của quy trình dự báo, trong đó nhấn mạnh các phương pháp đang được áp dụng rộng rãi trên thực tế. Các bước trong giai đoạn tác nghiệp được tóm tắt lại ở Hình 10.3.

Việc thu thập, phân tích các dữ liệu quá khứ, và lựa chọn mô hình/phương pháp dự báo thích hợp được trình bày ở chương 3. Bước tiếp theo thường là dự báo một số giai đoạn quá khứ khi các giá trị thực đã biết. Để so sánh các mô hình khác nhau ta thường dựa vào các tiêu chí thống kê, đồ thị hoặc các phân tích chẩn đoán phần dư như đã trình bày ở chương 1, chương 3 (và chương 8).

Thông thường người làm dự báo cần thiết phải tuân theo một quy trình dự báo và xem xét lại các quy trình đã được sử dụng. Các bước này có thể được tóm tắt như sau:

■ **HÌNH 10.3: Quy trình dự báo.**



1. Các dữ liệu cũ nên được loại ra và cập nhật dữ liệu gần đây vào cơ sở dữ liệu nhằm phục vụ cho dự báo.

2. Khi dữ liệu cập nhật, các tham số được sử dụng trong mô hình dự báo cần được tính toán lại, hay nói khác đi là chúng ta cũng cần cập nhật mô hình dự báo khi cơ sở dữ liệu thay đổi.

3. Mô hình dự báo với các tham số mới được xem xét để kiểm tra độ chính xác. Nếu độ chính xác thích hợp, thì mô hình lại tiếp tục được sử dụng cho đến lần cập nhật dữ liệu mới. Nếu độ chính xác không phù hợp thì nên xem xét lại dạng dữ liệu và chọn mô hình khác thích hợp hơn.

TRÁCH NHIỆM DỰ BÁO

Vị trí của quy trình dự báo trong doanh nghiệp sẽ khác nhau tùy thuộc vào qui mô doanh nghiệp, tầm quan trọng của dự báo trong việc ra quyết định, và bản chất cách thức quản lý của doanh nghiệp. Dự báo có thể là trách nhiệm của một phòng dự báo riêng hoặc một đơn vị quản lý nhỏ.

Các nhóm dự báo chuyên biệt thường phổ biến trong các tổ chức lớn. Các công ty lớn có khả năng thuê chuyên gia để thực hiện các dự báo phức tạp và có thể trang bị hệ thống máy tính và phần mềm hiện đại. Nỗ lực này có ưu điểm là ý kiến chuyên gia luôn có sẵn cho tất cả các đơn vị của doanh nghiệp. Tuy nhiên, nhược điểm lớn nhất là rất khó có được sự hợp tác giữa nhóm dự báo và các giám đốc bộ phận. Nhóm dự báo có thể mất nhiều thời gian cho việc thương lượng với những người sử dụng và giải thích vai trò của mình hơn là thực hiện việc dự báo. Ngược lại, ở thái cực đối lập là một quy trình dự báo thuộc từng các bộ phận chức năng của doanh nghiệp mà không hề có sự hợp tác giữa các bộ phận chức năng khác nhau này. Ưu điểm quy trình này là không có sự hiểu nhầm giữa người làm dự báo và người sử dụng dự báo vì họ đều nằm trong cùng một bộ phận chức năng cụ thể. Các dự báo thế này có xu hướng dễ được chấp nhận đưa vào quá trình ra quyết định. Nhược điểm là không thể thực hiện được dự báo phức tạp và thiếu chính xác vì sự không sẵn có nguồn lực như hệ

thống máy tính, phần mềm và các chuyên gia kỹ thuật. Điều này cũng khó thuyết phục quản lý cấp cao trang bị cơ sở vật chất thích hợp cho việc dự báo.

Trong một số trường hợp, phòng tiếp thị thực hiện và sở hữu các dự báo ví dụ như các dự báo thị trường. Tiếp thị được đo lường bằng sự sẵn có của sản phẩm, được xác định bằng phần trăm thời gian một sản phẩm cụ thể sẵn có cho việc mua hàng hóa của một khách hàng. Tuy nhiên, bộ phận tiếp thị không chịu hình phạt nào cho sự dư thừa tồn kho phát sinh từ sự phụ thuộc của sản xuất theo kết quả dự báo của bộ phận tiếp thị. Mặc dù các dự báo này dựa trên các điều kiện thị trường, nhưng chúng bị ảnh hưởng rất nhiều vào kỳ vọng doanh thu chủ quan của nhà quản trị. Hậu quả là phòng tiếp thị thường đưa ra các dự báo nhu cầu rất lạc quan để đảm bảo rằng sự sẵn có của một sản phẩm để có thể đáp ứng các mục tiêu doanh số. Ngược lại, đơn vị kế hoạch sản xuất sẽ sản xuất đáp ứng tho dự báo lạc quan về sản phẩm và hệ qủa là dư tồn kho doanh nghiệp cao.

Nhiều tổ chức cố gắng qui trách nhiệm dự báo cụ thể nhằm hạn chế các nhược điểm vừa được trình bày. Một nhóm nhỏ các nhân viên dự báo có thể được phân công xuống các đơn vị trong công ty để phục vụ nhu cầu của nhiều lĩnh chức năng của công ty. Công việc của nhân viên dự báo bao gồm việc kết hợp với khách hàng cũng như tạo ra các dự báo chính xác. Đôi khi trách nhiệm dự báo được kết hợp với các chức năng khác như hỗ trợ với thống kê, hay máy tính.

CHI PHÍ DỰ BÁO

Các phần cứng, phần mềm, nhân viên dự báo là những khoản chi phí hiển nhiên liên quan đến việc thực hiện dự báo. Nhưng có nhiều khoản chi phí khác có thể không nhìn thấy được như chi phí thời gian của những người liên quan trong việc thu thập dữ liệu cho quy trình dự báo, giám sát quy trình, giải thích kết quả thông qua hội thảo và các cuộc họp trong tổ chức. Những chi phí này phải được cân đối với lợi ích nhận được từ dự báo nếu các quyết định hợp lý có sử dụng các kết

quả dự báo. Ngoài ra, có thể có các khoản chi phí thuê tư vấn bên ngoài.

LÀM GÌ ĐỂ PHÁT TRIỂN DỰ BÁO ĐỊNH LƯỢNG?

Nhiều thập kỷ qua, các doanh nghiệp thường sử dụng các phương pháp dự báo định tính như phương pháp chuyên gia và ý kiến ban quản lý hoặc chỉ sử dụng các phương pháp dự báo định lượng giản đơn như bình quân di động và san mũ cho các quyết định của mình. Thực tế này đã thúc giục nhiều nghiên cứu thực tiễn nhằm tìm ra giải pháp cho tương lai của dự báo định lượng. Nói chung, đa số các kết quả nghiên cứu đều nhất trí bốn nhóm nguyên do sau đây.

Thiếu sự ủng hộ của quản lý cấp cao. Các kết quả nghiên cứu cho thấy rằng hầu hết quản lý cấp cao của doanh nghiệp không ủng hộ chức năng dự báo, đặc biệt là ở khâu đào tạo, không ủng hộ hoặc không dành thời gian cho việc học và nghiên cứu dự báo.

Thiếu các liên kết trong tổ chức. Các kênh liên lạc trao đổi thông tin dự báo trong doanh nghiệp không có hoặc hoạt động rất không hiệu quả. Các phòng ban không hợp tác với các bộ phận khác dẫn đến sự lưu chuyển thông tin rất kém và chất lượng thông tin không đáng tin cậy.

Không có trang bị các nguồn lực cho dự báo. Đa số các doanh nghiệp chưa trang bị cơ sở vật chất (máy tính, phần mềm), cơ sở dữ liệu, và nhân sự phục vụ mục đích dự báo.

Lựa chọn các kỹ thuật dự báo. Ngoài việc dựa vào các kỹ thuật dự báo định tính, phần lớn các doanh nghiệp phụ thuộc nhiều vào các phương pháp dự báo chuỗi thời gian. Tuy nhiên, rất nhiều lĩnh vực cần phải áp dụng các phương pháp khác nữa.

Chính vì thế, để khắc phục các khó khăn nói trên cần phải quan tâm các vấn đề sau đây:

- Ban quản lý cấp cao nên hiểu rõ và ủng hộ hơn chức năng của dự báo trong quá trình hoạch định.
- Trang bị kiến thức dự báo cho các cấp quản lý và tiến tới việc thành lập nhóm dự báo độc lập.
- Đầu tư cơ sở vật chất phục vụ dự báo như máy tính, phần mềm chuyên dụng.
- Xây dựng được một cơ sở dữ liệu chính xác, tin cậy, phù hợp, nhất quán, và kịp thời.
- Đào tạo và nâng cao kiến thức dự báo, đặc biệt là các phương pháp dự báo mới cho những người làm dự báo.
- Tăng cường sự trao đổi qua lại với những người thực hiện dự báo và những người sử dụng kết quả dự báo.
- Tăng cường sự hợp tác, cộng tác giữa các bộ phận chức năng trong cùng tổ chức.
- Đưa thêm tiêu chí đánh giá kết quả dự báo vào quá trình đánh giá mức độ hoàn thành công việc của cả cấp quản lý và nhân viên các bộ phận chức năng.
- Thiết lập một khung quản lý quy trình dự báo thống nhất.

ĐỪNG QUÊN VAI TRÒ CỦA DỰ BÁO ĐỊNH TÍNH

Hầu hết các phương pháp đã được trình bày trong cuốn giáo trình này đều thuộc các mô hình dự báo định lượng. Tuy nhiên, xuyên suốt quá trình thực hiện dự báo, khả năng phán đoán của những người làm dự báo là một yếu tố có ý nghĩa quan trọng để có các quả dự báo tốt. Khả năng phán đoán cần thiết nhất là trong việc đánh giá chất lượng dữ liệu và giải thích các kết quả của quá trình phân tích dữ liệu. Không phải kết quả dự báo định lượng hiển nhiên được sử dụng trong việc ra quyết định, mà nó

chỉ đóng vai trò 'cung cấp thông tin' cho việc ra quyết định. Người sử dụng kết quả dự báo nên xem kết quả dự báo định lượng như một thông tin 'khởi đầu' và cần phải cân nhắc kỹ lưỡng dựa trên kinh nghiệm, thực tế hoạt động và định hướng của tổ chức, và những phán đoán khác về môi trường kinh doanh. Nếu kết quả dự báo dựa vào các thông tin quá khứ có thể thay thế cho mọi quyết định của con người, thì cuộc sống có thể trở nên vô vị. Nên nhớ rằng, các mô hình dự báo định lượng đều dựa trên một giả định rằng xu hướng vận động của đối tượng dự báo trong quá khứ và hiện tại sẽ được tiếp tục duy trì trong tương lai. Và tương lai là không chắc chắn, nhất là trong kỷ nguyên công nghệ thông tin phát triển nhanh chóng như hiện nay. Chính vì vậy, cho dù dữ liệu quá khứ có tốt và đầy đủ như thế nào đi nữa, thì những phán đoán hợp lý vẫn rất cần thiết cho việc ra quyết định. Chẳng hạn, sử dụng doanh số quá khứ để dự báo doanh số tương lai của GAP chỉ có ý nghĩa nếu chúng ta giả định rằng công nghệ sản xuất sẽ không thay đổi, hành vi của người tiêu dùng sẽ không thay đổi, hành vi của đối thủ cạnh tranh sẽ không thay đổi, chiến lược công ty sẽ không thay đổi, và chính sách của chính phủ sẽ vẫn được giữ nguyên. Hoặc sử dụng giá CP và giá dầu của Mỹ trong quá khứ để dự báo giá CP trong tương lai chỉ có thể đáng tin cậy nếu hành vi của các nhà sản xuất dầu khí ở Trung Đông, Algeria, và Nga sẽ không đổi, trữ lượng dự trữ ở Mỹ sẽ không đổi, nhu cầu tiêu dùng ở Trung Quốc, Mỹ, và Châu Âu sẽ không đổi, hành vi của những nhà đầu tư trên thị trường thế giới sẽ không đổi, và công nghệ sản xuất sẽ không đổi. Như vậy, người sử dụng kết quả dự báo định lượng phải là người duy lý vì thông tin dự báo định lượng, mặc dù quan trọng, nhưng chỉ đóng góp phần nào cho các quyết định của mình.

Ngoài các phương pháp định tính đã được trình bày ở chương 1 như tổng hợp lực lượng bán hàng, ý kiến ban quản trị, khảo sát thị trường, kiểm chứng thị trường, và phương pháp Delphi, người ta còn sử dụng nhiều phương pháp khác như phân tích kịch bản, sở đồ cây quyết định, và đặc biệt là phương pháp điều khiển mạng nơ-ron nhân tạo[1]. Trong phần này, chúng tôi xin giới thiệu khái quát về phương pháp điều khiển mạng nơ-ron nhân tạo, một phương pháp đang được quan tâm nhiều

[1] Neutral Networks.

trong thời gian gần đây, để những bạn đọc nào yêu thích lĩnh vực dự báo có thể khám phá sâu hơn trong tương lai.

Các phương pháp dự báo truyền thống, như đã đề cập trong cuốn giáo trình này, chủ yếu dựa vào dữ liệu quá khứ để phát triển một mô hình và sử dụng mô hình đó để ước tính giá trị của biến được quan tâm trong một hoặc một số giai đoạn tương lai. Các giá trị ước tính này sẽ trở thành các giá trị dự báo có thể được sử dụng cho việc lập các kế hoạch kinh doanh của doanh nghiệp. Trong các mô hình này, chúng ta giả định rất mạnh rằng tương lai sẽ giống trong quá khứ. Các phương pháp dự báo truyền thống đôi khi còn đưa ra các giả định về dạng phân phối xác suất của đối tượng cần dự báo.

Lĩnh vực đang phát triển của cơ chế 'thông minh nhân tạo' đã nỗ lực bắt chước các quá trình vận động của hệ thần kinh và bộ não con người nhờ sự phát triển của công nghệ máy tính. Mặc dù bắt nguồn từ sinh học và tâm lý học, nhưng cơ chế thông minh nhân tạo đã phát triển nhanh chóng qua nhiều lĩnh vực khác, kể cả kinh doanh và kinh tế. Ba ứng dụng chính của cơ chế thông minh nhân tạo là xử lý ngôn ngữ, chế tạo người máy, và điều khiển mạng nơ-ron nhân tạo. Lĩnh vực điều khiển mạng nơ-ron nhân tạo có nhiều ứng dụng thương mại nhất, kể cả dự báo.

Trong điều khiển mạng nơ-ron nhân tạo, rất nhiều ví dụ 'đã' được lập trình sẵn vào các phần mềm máy tính, các ví dụ này thể hiện toàn bộ các mối quan hệ khác nhau trong quá khứ giữa các biến có thể ảnh hưởng đến kết quả của các biến dự báo. Sau đó, chương trình điều khiển mạng nơ-ron nhân tạo sẽ 'học tập và đồng hóa' các ví dụ này một cách hệ thống và cố gắng xây dựng các mối quan hệ nền tảng điển hình thông qua quá trình học tập. Nhờ quá trình học tập này mà 'kinh nghiệm' của chương trình điều khiển mạng nơ-ron nhân tạo không ngừng được cải thiện. Cho nên, khi gặp một tình huống mới, hệ thống xử lý sẽ tự động xem xét và tìm ra kết quả dự báo tốt nhất. Ưu điểm của chương trình điều khiển mạng nơ-ron nhân tạo trong dự báo là chúng ta không cần xác định trước các mối quan hệ giữa các biến bởi vì chương trình có thể 'tự suy đoán' ra các mối quan hệ khả dĩ nhất nhờ có một 'vốn kiến thức' phong phú từ các mối quan hệ được đúc kết qua rất nhiều ví dụ trước đây. Ngoài ra, điều khiển mạng nơ-ron nhân tạo cũng không đòi hỏi bất

cứ giả định nào về phân phối xác suất. Phần mềm về điều khiển mạng nơ-ron nhân tạo có thể được cài đặt trên nền Excel, hoặc có sẵn trong SPSS, v.v...

Một vấn đề quan trọng khác chúng ta cần lưu ý khi thực hiện dự báo là đôi khi 'phải' kết hợp các dự báo với nhau để có một kết quả dự báo có khả năng được sử dụng vào quá trình ra quyết định cao hơn. Theo Amstrong (1989)[2], kết quả từ 200 nghiên cứu khác nhau cho thấy việc kết hợp các dự báo sẽ tạo tạo ra một kết quả vẫn nhất quán nhưng có độ chính xác cao hơn. Lợi ích của việc kết hợp các dự báo còn làm giảm thiểu các ảnh hưởng thiên lệch và tiết kiệm chi phí cho doanh nghiệp. Một chiến lược kết hợp dự báo là lấy trung bình giản đơn của các kết quả từ nhiều phương pháp dự báo khác nhau (kể cả định lượng và định tính). Ví dụ, gọi $\hat{Y}_{11}$, $\hat{Y}_{12}$, $\hat{Y}_{13}$, ..., $\hat{Y}_{1m}$ là các giá trị doanh số dự báo cho tháng tiếp theo từ *m* phương pháp dự báo, thì giá trị dự báo kết hợp $\hat{Y}_{1C}$ sẽ được tính như sau:

$$\hat{Y}_{1C} = \frac{\hat{Y}_{11} + \hat{Y}_{12} + ... + \hat{Y}_{1m}}{m}$$

Cách khác, chúng ta có thể gán các trọng số khác nhau cho các giá trị dự báo theo những phương pháp khác nhau, sao cho tổng các trọng số phải bằng 1.

$$\hat{Y}_{1C} = w_1\hat{Y}_{11} + w_2\hat{Y}_{12} + ... + w_m\hat{Y}_{1m}$$

Có nhiều cách khác nhau để xác định trọng số của từng giá trị dự báo trong giá trị dự báo kết hợp, như dựa vào tỷ lệ của chúng trong tổng sai số bình phương hoặc dựa vào phân tích hồi quy đa biến.

[2] Hanke, 2005, Business Forecasting, 8th Edition, pp.468.

TÓM TẮT CHƯƠNG 10

Nhu cầu dự báo của một tổ chức hay một doanh nghiệp trong bối cảnh thế giới thay đổi rất nhanh hiện nay là có thực và rất cần thiết. Tính không chắc chắn và mức độ rủi ro của môi trường trong và ngoài nước mà một tổ chức hay một doanh nghiệp đang hoạt động hiện nay là rất cao. Nếu tổ chức hoặc doanh nghiệp ít có khả năng thực hiện các dự báo hoặc không thể hiểu được các kết quả dự báo của những tổ chức khác thì chắc chắn là sẽ bị động trong quá trình ra quyết định của mình. Như vậy, doanh nghiệp hoặc tổ chức sẽ phải đối diện trước hai lựa chọn cho việc tìm ra những kết quả dự báo cần thiết phục vụ cho quá trình ra quyết định của mình là tự thực hiện dự báo hoặc là nhờ tư vấn từ các chuyên gia dự báo khác. Hai cách tiếp cận tự thực hiện hoặc nhờ tư vấn đều yêu cầu cấp quyết định cao nhất của tổ chức và doanh nghiệp phải nắm được là việc lựa chọn phương pháp dự báo phù hợp, giám sát quy trình dự báo cũng như phân công trách nhiệm dự báo và chi phí thực hiện dự báo cụ thể nhằm trả lời một câu hỏi quan trọng là kết quả dự báo phải đáng tin cậy và sử dụng chúng một cách hiệu quả cho qua trình ra quyết định. Ngoài ra, để tránh trường hợp doanh nghiệp phải tốn nhiều nguồn lực cho việc thực hiện các dự báo định lượng, nhưng cuối cùng lại vẫn sử dụng kết quả dự báo định tính, chúng tôi cho rằng doanh nghiệp nên biết cách kết hợp các kết quả dự báo lại với nhau.

CÂU HỎI VÀ BÀI TẬP

1. Anh/Chị cho biết các nhân tố then chốt quyết định kết quả dự báo là gì? Ngoài các nhân tố này, Anh/Chị cho biết có thể còn các nhân tố nào khác hay không?

2. Anh/Chị hãy cho biết quy trình thực hiện một dự báo bất kỳ bao gồm những bước nào? Trong đó, Anh/Chị cho biết những bước nào có ý nghĩa quan trọng nhất? Tại sao?

3. Anh/Chị cho biết làm sao có thể lựa chọn được một phương pháp dự báo thích hợp?

4. Anh/Chị cho biết làm sao tổ chức có thể giám sát tốt quy trình dự báo của mình?

5. Anh/Chị hãy giải thích khung quản lý quy trình dự báo do Davis và Mentzer đề xuất?

6. Anh/Chị cho biết ai là người chịu trách nhiệm chủ yếu trong việc thực hiện dự báo của một tổ chức?

7. Anh/Chị cho biết chi phí dự báo bao gồm những thành phần nào? Làm sao có thể giảm thiểu chi phí dự báo cho doanh nghiệp nhưng vẫn đảm bảo việc ra quyết định của doanh nghiệp?

8. Theo Anh/Chị, tại sao nhiều tổ chức ở Việt Nam chưa áp dụng các kỹ thuật dự báo định lượng?

9. Anh/Chị cho biết tại sao dự báo định tính lại có vai trò quan trọng trong nhiều quyết định của các tổ chức?

10. Anh/Chị cho biết tại sao trên thực tế người ta có xu hướng kết hợp nhiều kỹ thuật dự báo khác nhau?

11. Từ kết quả các mô hình dự báo trước đây, Anh/Chị cho biết công ty kinh doanh sản phẩm khí nên dự báo giá CP như thế nào?

12. Từ kết quả các mô hình dự báo trước đây, Anh/Chị cho biết công ty GAP nên dự báo doanh số như thế nào?

13. Từ kết quả các mô hình dự báo trước đây, Anh/Chị cho biết công ty CCC nên dự báo lượng khách hàng mới như thế nào?

14. Từ kết quả các mô hình dự báo trước đây, Anh/Chị cho biết công ty Murphy Brothers nên dự báo doanh số như thế nào?

15. Tổng hợp lại, Anh/Chị cho biết các kỹ thuật dự báo có thể có ích trong những lĩnh vực nào? Cho ví dụ cụ thể?

TÀI LIỆU THAM KHẢO

Akihiro Amono (1987), "A Small Forecasting Model of the World Oil Market", *Journal of Policy Modeling*, 9(4), 615-635.

Andrew Fight (2005), *Cash Flow Forecasting*, 1st Edition, Butterworth-Heinemann.

Antonino Parisi, Franco Parisi, and David Diaz (2008), "Forecasting Gold Price Changes: Rolling and Recursive Neutral Network Models", *Journal of Multinational Financial Management*, Vol.18, pp.477-487.

Boaz Nandwa and Samuel K. Andoh (2008), *Economic Liberalization and Conditional Volatility of Exchange Rate in Sub-Saharan Africa: Asymmetric GARCH Analysis*, Blackwell Publishing Ltd.

Bollersley, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, Vol.31, pp.307-327.

Bruce L. Bowerman, Richard O'Connell, and Anne Koehler (2004), *Forecasting, Time Series, & Regression*, 4th Edition, South-Western College.

Chaman L. Jain (2006), "Benchmarking Forecasting Practices in Corporate America", *The Journal of Business Forecasting*, Vol.24, No.4.

Chaman L. Jain (2007), "Benchmarking Forecasting Models", *The Journal of Business Forecasting*, Vol.25, No.4, pp.14-17.

Chaman L. Jain and Jack Malehorn (2005), *Practical Guide to Business Forecasting*, 2nd Edition, Graceway Publishing Company.

Charles F. Roos (1955), "Survey of Economic Forecasting Techniques: A Survey Article", *Econometrica*, Vol.23, No.4, pp.363-395.

Charles W. Chase (1997), "Selecting the Approriate Forecasting Method", *The Journal of Business Forecasting Methods & Systems*, Vol.16, No.3, pp.2-29.

Choo Wei Chong, Muhammad Idrees Ahmad and Mat Yusoff Abdullah (1999), "Performance of GARCH Models in Forecasting Stock Market Volatility", *Journal of Forecasting*, Vol.18, pp.333-433.

Christain Haefke and Christain Helmenstein (2002), "Index Forecasting and Model Selection", *International Journal of Intelligent Systems in Accounting, Finance & management*, Vol.11, pp.119-135.

Christopher Chatfield (2001), *Time Seris Forecasting*, CRC Press.

Colin Robinson (1965), "Some Principles of Forecasting in Business", *The Journal of Industrial Economics*, Vol.14, No.1, pp.1-13.

Czinkota, M.R., and Ronkainen, I.A. (2005), "A Forecast of Globalization, International Business and Trade: Report from a Delphi Study", *Journal of World Business*, Vol.40, pp.111-123.

David G. Loomis, James E., & Cox, Jr. (2000), "A Course in Economic Forecasting: Rationale and Content", *The Journal of Economic Education*, Vol.31, No.4, pp.349-357.

David G. McMillan and Alan E. H. Speight (2004), "Daily Volatiliity Forecasts: Reassessing the Performance of the GARCH Models", *Journal of Forecasting*, Vol.23, pp.449-460.

Demetriades, P.O. and K.A. Hussen (1996), "Does Financial Development Cause Economic Growth? Time-Series Evidence from 16 Countries", *Journal of Development Economics*, Vol.51, pp.387-411.

Dianne Waddell & Amrik S. Sohal (1994), "Forecasting: The Key to Managerial Decision Making", *Management Decision*, Vol.32 No.1, pp.41-49.

Dimitrios Asteriou and Costas Siriopoulos (2000), "The Role of Political Instability in Stock Market Development and Economic Growth: The Case of Greece", *Economic Notes*, Vol.29, No.3, pp.355-374.

Dimitrios Asteriou and Simon Price (2001), "Political Instability and Economic Growth: UK Times Series Evidence", *Scottish Journal of Political Economy*, Vol.48, No.4, pp.383-399.

Dimitrios Asteriou and Stephen G. Hall (2007), *Applied Econometrics: A Modern Approach Using Eviews and Microfit*, Revised Edition, Palgrave Macmillan.

Domodar Gujarati (2006), *Essentials of Econometrics*, 3nd Edition, McGraw-Hill.

Domodar Gujarati (2009), *Basic Econometrics*, 5th Edition, McGraw-Hill.

Dayananda, D. *et al* (2002), *Capital Budgeting: Financial Appraisal of Investment Projects*, Cambridge.

Donald J. Bowersox, David J. Colss, M. Bixby Cooper (2007), *Supply Chain Logistics Management*, Second Edition, McGraw-Hill.

Donna F. Davis and John T. Mentzer (2007), "Organizational Factors in Sales Forecasting Management", *International Journal of Forecasting*, Vol.23, pp.475-495.

Edel Tully and Brian M. Lucey (2007), "A Power of GARCH Examination of the Gold Market", *Research in International Business and Finance*, Vol.20, pp.316-325.

Engle, R.F. (1981), "Autoregressive Conditional Heteroskedasticity with Estimates of Variance of U.K. Inflation", *Econometria*, Vol.50, pp.987-1008.

Eviews 6 User's Guide (2007), Quantitative Micro Software, United States of America.

Francis X. Diebold (2004), *Elements of Forecasting*, 3rd Edition, Thomson: South - Western.

Frank A.Friday (1952a), "Business Forecasting", *The Incorporated Statistician*, Vol.3, No.2, pp.25-37.

Frank A.Friday (1952b) "The Problem of Business Forecasting", *The Journal of Industrial Economics*, Vol.1, No.1, pp.55-71.

Gary Levee (1993), "The Key to Understanding the Forecasting Process", *The Journal of Business Forecasting Methods & Systems*, Vol.11, No.4, pp.12-16.

George C.S. Wang (2004), "Forecasting Practices in Electric and Gas Utility Companies, *The Journal of Business and Forecasting Methods & Systems*, Vol.23, pp.11-16.

Gillian Rice (1997), "Forecasting in US Firms: A Role of TQM", *International Journal of Operations and Production Management*, Vol.17, No.2, pp.211-220.

Granger, C.W.J. (1988), "Some Recent Developments in the Concept of Causality", *Journal of Econometrics*, Vol.39, pp.199-211.

Henry C. Smith III, Paul Herbig, John Milewicz & James E. Golden (1996), "Differences in forecasting behaviour between large and small firms", *Journal of Marketing Practice: Applied Marketing Science*, Vol.2, No.1, pp.35-51.

Hoàng Trọng (2007), *Thống kê ứng dụng*, NXB Thống Kê.

Ioannis T. Lazaridis (2002), "Cashflow Estimation and Forecasting Practices of Large Firms in Cyprus: Survey Findings", *Journal of Financial Management and Analysis*, Vol.15, No.2, pp.62-68.

Wilson, J.Holton & Barry Keating (2007), *Business Forecasting With Accompanying Excel-Based ForecastXTM Software*, 5th Edition.

Jeffrey M. Wooldridge (2008), *Essentials of Econometrics*, 1st Edition, South-Western College.

Jeffrey M.Wooldridge (2003), *Introductory Econometrics: A Modern Approach*, 2nd Edition, US: Thomson, South-Western.

John C. Brocklebank and David A. Dickey (2003), *SAS for Forecasting Time Series*, 2nd Edition, Wiley-SAS.

John C. Pickett, David P. Reilley, and Robert M. McIntyre (2005), "How to Select a Most Efficient 'OLS' Model for a Time Series Data", *The Journal of Business Forecasting*, Vol.24, No.2, pp.28-32.

John Clare Brocklebank, David A. Dickey (2003), *SAS for forecasting time series*, SAS Publishing.

Hanke, J.E. & Wichern, D.W. (2005), *Business Forecasting*, 8th Edition.

John G.Wacker & Rhonda Lummus (2002), "Sales Forecasting for Strategic Resource Planning", *International Journal of Operations and Production Management*, Vol.22, No.9, pp.1014-1031.

John Hanke (1989), "Forecasting in Business Schools: A Follow-Up Survey", *International Journal of Forecasting*, Vol.5, pp.259-262.

John McCormick (2008), "25% of Data is Bad; What Should We Do?", www.BASELINEMAG.com.

Kam Fong Chan (2005), "Modelling Conditional Heteroscedasticity and Jumps in Australian Short-Term Interest Rates", *Accounting & Finance*, Vol.45, pp.537-551.

Kenneth B. Kahn (2002), "An Exploratory Investigation of New Product Forecasting Practices", *The Journal of Product Innovation Management*, Vol.19, pp.133-143.

Kenneth Lawrence, Ronald K Klimberg, and Sheila M Lawrence (2009), *Fundamentals of Forecasting Using Excel*, 1st Edition, Industrial Press Inc.

Larry Lapide (2007), "Questions to Ask When Reviewing the Benchmarking Data", *The Journal of Business Forecasting*, Vol.25, No.4, pp.4-7.

Lon-Mu Liu and Maw-Wen Lin (1991), "Forecasting Residential Consumption of Natural Gas Using Monthly and Quarterly Time Series", *International Journal of Forecasting*, Vol.7, pp.3-16.

M. Pilar Munoz, M. Dolores Marquez and Lesly M. Acosta (2007), "Forecasting Volatility by Means of Threshold Models", *Journal of Forecasting*, Vol.26, pp.343-363.

M.C. Hughes (2001), "Forecasting Practice: Organisational Issues", *The Jourbal of the Operational Research Society*, Vol.52, No.2, pp.143-149.

Makridakis, S., Wheelwright, S.C., & Hyndman, R.J. (1998), *Forecasting Methods and Applications*, John Wiley & Sons.

Mark A. Moon, John T. Mentzer, Carlo D. Smith, and Michael S. Garver (1998), "Seven Keys to Better Forecasting", Business Horizons.

Michael Clements and David Hendry (2008), *Forecasting Economic Time Series*, 2nd Edition, Cambridge University Press.

Michael K. Evans (2002), *Practical Business Forecasting*, Willey & Son.

Michael R. Donihue (1995), "Teaching Economic Forecasting to Undergraduates", *The Journal of Economic Education*, Vol. 26, No. 2, pp.113-121.

Moira C. Watson (1996), "Forecasting in the Scottish Electronics Industry", *International Journal of Forecasting*, Vol.12, pp.361-371.

Mullen, M.P. (2003), "Delphi: Myths and Reality", *Journal of Health Organization and Management*, Vol.17, No.1, pp.37-52.

Nada R.Sanders & Larry P.Ritzman (2004), "Integrating Judgemental and Quantitative Forecasts: Methodologies for Pooling Marketing and Operations Information", *International Journal of Operations and Production Management*, Vol.24, No.5, pp.514-529.

Nguyễn Trọng Hoài (2003), *Mô hình hóa chuỗi thời gian trong kinh doanh và kinh tế*, Ấn bán lần 2, Nhà xuất bản Đại học Quốc Gia.

Nigel Meade (2000), "Evidence for the Selection of Forecasting Methods", *Journal of Forecasting*, Vol.19, pp.515-535.

P. W. Abelson, Roselyne Joyeux (2000), *Economic forecasting*, Allen & Unwin.

Patricia E. Gaynor, Rickey C. Kirkpatrick (1994) *Introduction to time-series modeling and forecasting in business and economics*, McGrow-Hill

Penelope M. Mullen (2003), "Delphi: Myths and Reality", *Journal of Health Organization and management*, Vol.17, No.1, pp.37-52.

Peter J. Brockwell and Richard A. Davis (2003), *Introduction to Time Series and Forecasting*, 2nd Edition, Springer.

R. Carter Hill, William E. Griffiths, and Guay C. Lim (2008), *Using Eviews for Principles of Econometrics*, 3rd Edition, John Wiley & Sons.

Ramanathan, R. (1998), *Introductory Econometrics with Application*, The Dryden Press – Harcourt Brace College Publishers.

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, fifth edition, Harcourt College Publisher.

Richard Barrett and Jeremy Hope (2006), "Re-Forecasting Practice in the UK", *Measuring Business Excellence*, Vol.10, No.2, pp.28-40.

Rob J. Hyndman, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder (2008), *Forecasting with Exponential Smoothing: The State Space Approach*, 1st Edition, Springer.

Robert Altabet (1998), "The Forecaster as a Key Member of the Strategic Planning Team", *The Journal of Business Forecasting Methods & Systems*, Vol.17, No.3, pp.3-6.

Robert D. Klassen and Benito E. Flores (2001), "Forecasting Practices of Canadian Firms: Survey Results and Comparisons", *International Journal of Production Economics*, Vol.70, pp.163-174.

Robert Fildes & Robert Hastings (1994), "The Organization and Improvement of Market Forecasting", *The Journal of the Operational Research Society*, Vol.45, No.1, pp.1-16.

Robert Loo & Karran Thorpe (2003), "A Delphi Study Forecasting Management Training and Development for First-Line Nurse managers", *Journal of Management Development*, Vol.22, No.9, pp.824-834.

Pindyck, R.S. and Rubinfeld, D.L. (1998), *Econometric Models and Economic Forecasts*, 4th Edition, McGraw-Hill.

SAS Institute (1999), *SAS/Ets User's Guide, Version 8*, 2nd Edition, SAS Publishing.

Scott Amstrong (2001), *Principles of Forecasting*, 1st Edition, Springer.

Seyed-Mahmoud Aghazadeh (2007), "Revenue Forecasting Models for Hotel Management", *The Journal of Business Forecasting*, Vol.24, No.2, pp.28-32.

Stephen Satchell (2007), *Forecasting Expected Returns in the Financial Markets*, 1 st Edition, Academic Press.

Studenmund, A.H. (2001), *Using Econometrics: A Practical Guide*, 4th Edition, Addison Wesley Longman.

Suleyman Gokcan (2000), "Forecasting Volatility of Emerging Stock Markets: Linear vs Non-Linear GARCH Models", *Journal of Forecasting*, Vol.19, pp.499-504.

Taufiq Choudhry and Hao Wu (2008), "Forecasting Ability of GARCH vs Kalman Filter Method: Evidence from Daily UK Time-Varying Beta", *Journal of Forecasting*, Vol.26, pp.670-689.

Thomas Cook (1995), "Understand Your Customer Before Preparing Forecasts", *The Journal of Business Forecasting Methods & Systems*, Vol.13, No.4, pp.27-29.

William J. Stevenson (2005), *Operations Management*, 8th edition, McGraw-Hill.

Xiao-Ming Li (2003), "China: Further Evidence on the Evolution of Stock Markets in Transition Economies", *Scottish Journal of Political Economy*, Vol.50, No.3, pp.341-358.

http://www.forecasters.org

http://www.forecastingprinciples.com

http://unstats.un.org

http://www.gso.gov.vn

http://www.vdf.org.vn

http://www.reuters.com

Tổng cục Thống kê Việt Nam, 2006, Điều tra mức sống hộ gia đình Việt Nam (VHLSS2004, VHLSS2006).

Quỹ tiền tệ Quốc tế, CD-ROM.