TRƯỜNG ĐẠI HỌC VINH
TRUNG TÂM ĐÀO TẠO TỪ XA VÀ QUAN HỆ DOANH NGHIỆP
----- o0o -----

ThS.NGUYỄN HOÀI NAM

Giáo trình

Kinh tế lượng

Vinh - 2011

TR□□NG □□I H□C VINH

TRUNG TÂM ĐÀO TẠO TỪ XA VÀ QUAN HỆ DOANH NGHIỆP

Chủ biên: ThS. Nguyễn Hoài Nam

GIÁO TRÌNH

Kinh tế lương

(Giáo trình đào tạo từ xa)

Vinh - 2011

LỜI NÓI ĐẦU

Kinh tế lượng là môn khoa học định lượng, ngày càng được áp dung rộng rãi và phổ biến trong nghiên cứu kinh tế xã hội, quản trị kinh doanh và đã được đưa vào giảng dạy trong chương trình đào tạo đại học và sau đại học khối ngành kinh tế trong các trường Đại học. Việc tìm hiểu, ứng dụng, nghiên cứu môn học này trở thành tất yếu trong phân tích kinh tế.

Nhằm đáp ứng yếu cầu tìm hiểu, áp dụng kinh tế lượng trong các lĩnh vực kinh tế và quản trị kinh doanh, cuốn sách Kinh tế lượng được viết nhằm trả lời những vấn đề cụ thể về cả lý thuyết và thực hành, dành cho đối tượng đang theo học khóa học từ xa cũng như tự nghiên cứu về khối ngành kinh tế.

Chúng tôi đã cố gắng biên soạn theo tinh thần đơn giản dễ hiểu để đông đảo độc giả và học viên có thể dễ dàng sử dụng, đặc biệt cho các học viên Đại học từ xa có thể tự học, tự nghiên cứu. Bởi vậy, phần lý thuyết không quá đi sâu vào các chứng minh phức tạp mà chú ý tới các khái niệm, các phương pháp thực hành. Cuối mỗi chương có câu hỏi ôn tập.

Sách này kèm với đĩa CD-ROM (sách điện tử) để anh chị học viên tiện học tập và tra cứu.

Trong quá trình biên soạn, chắc chắn không tránh khỏi những khiếm khuyết.Chúng tôi chân thành cảm ơn các độc giả đóng góp ý kiến để cuốn sách ngày càng hoàn thiện hơn.

TÁC GIẢ

CHƯƠNG 1. MỞ ĐẦU

Chương này nhằm giúp bạn đọc khái quát về Kinh tế lượng, phương pháp luận và một số đặc điểm khi vận dụng Kinh tế lượng trong thực tiễn.

Chương học bao gồm các nội dung sau:

O Kinh tế lượng là gì

O Phương pháp luận kinh tế lượng

O Nội dung của kinh tế lượng

1.1. Kinh tế lượng là gì?

Cho đến nay chưa có một định nghĩa nào về kinh tế lượng được mọi người thống nhất. Kinh tế lượng có nghĩa là đo lường kinh tế. Mặc dù đo lường kinh tế là một nội dung quan trọng của kinh tế lượng nhưng phạm vi của kinh tế lượng rộng hơn nhiều. Điều đó được thể hiện thông qua một số định nghĩa sau đây:

Kinh tế lượng có thể được định nghĩa như một môn khoa học xã hội trong đó người ta dùng các công cụ của lý thuyết kinh tế, toán kinh tế và thống kê kinh tế để phân tích các hiện tượng kinh tế.

Kinh tế lượng bao gồm việc áp dụng thống kê toán cho các số liệu kinh tế để củng cố về mặt thực nghiệm cho các mô hình do các nhà kinh tế toán đề xuất và tìm ra lời giải bằng số.

Kinh tế lượng là một môn khoa học phân tích định lượng một cách tổng hợp. Nó khắc phục được nhược điểm của các môn khoa học như lý thuyết kinh tế, thống kê, toán kinh tế.

1.2. Phương pháp luận kinh tế lượng

Phân tích kinh tế lượng được thực hiện theo các bước sau đây:

- Nêu ra các giả thiết về các mối quan hệ giữa các biến kinh tế. Chẳng hạn kinh tế vĩ mô khẳng định rằng mức tiêu dùng của các hộ gia đình phụ thuộc theo quan hệ cùng chiều với thu nhập khả dụng của họ.

- Thiết lập các mô hình toán học để mô tả mối quan hệ giữa các biến số này. Chẳng hạn: $Y = \beta_1 + \beta_2 X + u$

Trong đó: Y: chi tiêu cho tiêu dùng của một hộ gia đình

X: thu nhập khả dụng của hộ gia dùng

β_1: hệ số chặn; β_2: hệ số góc; u: yếu tố ngẫu nhiên.

Sự tồn tại của yếu tố ngẫu nhiên bắt nguồn từ mối quan hệ giữa các biến kinh tế nói chung là không chính xác.

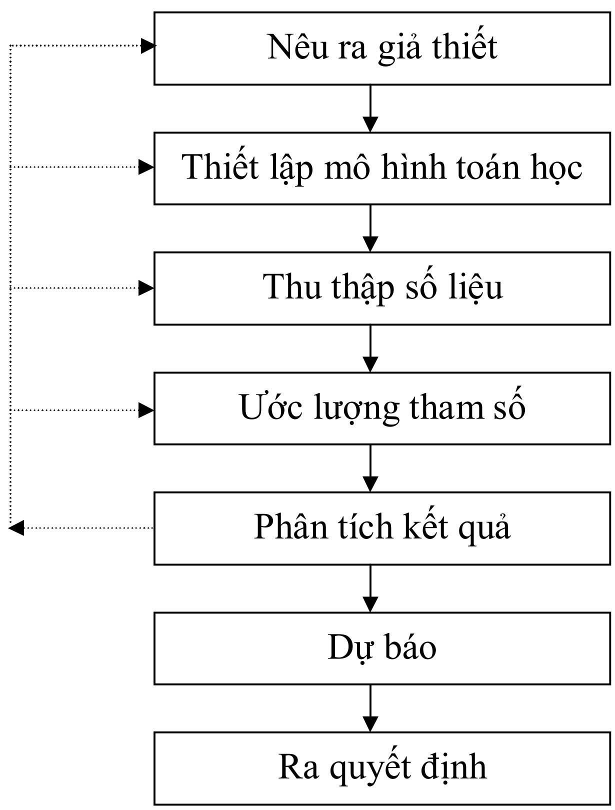
- Thu thập số liệu: Để ước lượng các tham số của mô hình, cần phải thu thập số liệu. Kinh tế lượng đòi hỏi kích thước mẫu khá lớn.

- Ước lượng các tham số của mô hình nhằm nhận được số đo về mức ảnh hưởng của các biến với các số liệu hiện có. Các ước lượng này là các kiểm định thực nghiệm cho lý thuyết kinh tế.

- Phân tích kết quả dựa trên lý thuyết kinh tế để phân tích và đánh giá kết quả nhận được. Xét xem các kết quả nhận được có phù hợp với lý thuyết kinh tế không, kiểm định các giả thiết thống kê về các ước lượng nhận được. Trong mô hình $Y = \beta_1 + \beta_2 X + u$, nếu ước lượng của β_2 là số dương và nhỏ hơn 1 thì ước lượng này là hợp lý về mặt kinh tế. Trong trường hợp ngược lại (< 0 hoặc > 1) thì không phù hợp về mặt kinh tế. Khi đó cần phải tìm ra một mô hình đúng.

- Dự báo: Nếu như mô hình phù hợp với lý thuyết kinh tế thì có thể sử dụng mô hình để dự báo. Dự báo giá trị trung bình hoặc giá trị cá biệt.

- Sử dụng mô hình để đề ra chính sách.



Các bước trên đây có nhiệm vụ khác nhau trong quá trình phân tích một vấn đề kinh tế và chúng được thực hiện theo một trình tự nhất định. Tìm ra bản chất một vấn đề kinh tế là một việc không đơn giản. Vì vậy, quá trình trên đây phải được thực hiện nhiều lần như là các phép lặp cho đến khi chúng ta thu được một mô hình đúng.

1.3. Nội dung của kinh tế lượng.

Nội dung của môn học gồm 7 chương:

Chương 1. Mở Đầu

Chương 2. Mô hình hồi quy đơn

Chương 3. Mô hình hồi quy bội

Chương 4. Hồi quy với biến giả

Chương 5. Đa cộng tuyến và tự tương quan

Chương 6. Phương sai và sai số thay đổi

Chương 7. Chọn mô hình và kiểm định việc chọn mô hình

Câu hỏi chương 1

1. Trình bày các bước phân tích Kinh tế lượng?
2. Mục tiêu của nghiên cứu Kinh tế lượng là gì?
3. Đối tượng nghiên cứu của Kinh tế lượng là gì? Kinh tế lượng thường sử dụng những công cụ nào trong nghiên cứu?
4. Những môn học nào cần biết để nghiên cứu Kinh tế lượng?
5. Cho mô hình $Y = \beta_1 + \beta_2 X + u$

Dựa vào các kiến thức về kinh tế/ xã hội mà anh chị đã biết. Các anh chị có kỳ vọng gì về dấu của hệ số góc trong mô hình hồi quy trên, ứng với các tình huống cụ thể sau:

a) Chi tiêu cho tiêu dùng (Y) và thu nhập (X) của một người.

b) Mức cầu của một loại hàng hóa (Y) và giá bán (X).

c) Doanh số bán hàng của một loại hàng (Y) và giá bán (X).

d) Tỷ lệ (lượng tiền mặt lưu giữ/thu nhập) của cá nhân (Y) và mức lạm phát (X).

e) Mức cầu của một sản phẩm (Y) và chi phí quảng cáo (X).

f) giá bán của một căn nhà (Y) và diện tích của một căn nhà (X).

CHƯƠNG 2. MÔ HÌNH HỒI QUY ĐƠN

Chương này trình bày các vấn đề cơ bản trong Kinh tế lượng, vận dụng trong trường hợp đơn giản nhất, đó là mô hình hồi quy tuyến tính hai biến, bao gồm việc ước lượng các tham số hồi quy trong mô hình Kinh tế lượng, đánh giá ý nghĩa thống kê của các tham số ước lượng, cũng như đánh giá sự phù hợp của mô hình hồi quy. Mặc dù trong thực tế quan hệ giữa các biến số kinh tế thường phức tạp, gồm nhiều yếu tố tác động, nên mô hình hai biến thường ít có ý nghĩa, nhưng nó đặc biệt có ích về mặt lý thuyết. Hiểu được các vấn đề cơ bản trong mô hình hồi quy hai biến đơn giản giúp ta vân dụng được những tình huống phức tạp hơn.

Nội dung cơ bản của chương bao gồm:

O Một vài khái niệm cơ bản

- Phân tích hồi quy

- Quan hệ thống kê và quan hệ hàm số

- Hàm hồi quy và quan hệ nhân quả

- Hồi quy và tương quan

- Bản chất và nguồn số liệu cho phân tích hồi quy

O Mô hình hồi quy tổng thể

- Sai số ngẫu nhiên và bản chất của nó

- Hàm hồi quy mẫu

O Ước lượng và kiểm định giả thiết trong mô hình hồi quy hai biến

- Phương pháp bình phương nhỏ nhất

- Các giả thiết cơ bản của phương pháp bình phương nhỏ nhất

- Độ chính xác của các ước lượng bình phương nhỏ nhất

- Hệ số r^2 đo độ phù hợp của hàm hồi quy mẫu

- Khoảng tin cậy và kiểm định giả thiết về các hệ số hồi quy

- Phân tích hồi quy và dự báo

2.1. MỘT VÀI KHÁI NIỆM CƠ BẢN

2.1.1. Phân tích hồi qui

Phân tích hồi qui nghiên cứu mối liên hệ phụ thuộc của một biến (gọi là biến phụ thuộc hay biến được giải thích) với một hay nhiều biến khác (được gọi là các biến độc lập hay biến giải thích) nhằm ước lượng và dự báo giá trị trung bình của biến phụ thuộc với các giá trị đã biết của các biến độc lập.

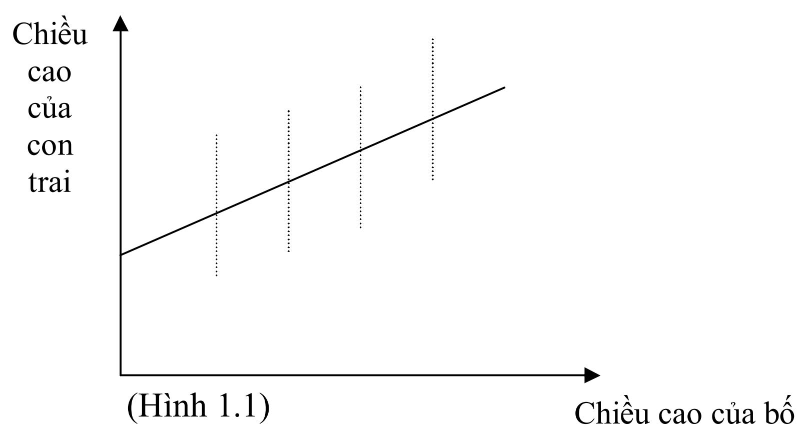
Ta xem xét thí dụ:

<u>Thí dụ 1.1</u>

a. Luật Galton Pearson nghiên cứu sự phụ thuộc chiều cao của các cháu trai vào chiều cao của bố những đứa trẻ này. Ông đã xây dựng được đồ thị chỉ ra phân bố chiều cao của các cháu trai ứng với chiều cao của người cha. Qua mô hình này có thể thấy:

Thứ nhất, với chiều cao đã biết của người cha thì chiều cao của các cháu trai sẽ là một khoảng dao động quanh giá trị trung bình.

Thứ hai, chiều cao của cha tăng thì chiều cao của các cháu trai cũng tăng. Mô hình này giải thích được điều mà Galton đặt ra và còn được dùng trong dự báo.



(Hình 1.1)

Tiếp tục nghiên cứu vấn đề trên, Karl Pearson đã phát hiện ra rằng: chiều cao trung bình của các cháu trai của nhóm bố cao nhỏ hơn chiều cao của bố và chiều cao trung bình của các cháu trai của nhóm bố thấp lớn hơn chiều cao của bố. Điều này được thể hiện: hệ số góc của đường thẳng trên hình nhỏ hơn 1.

Trong thí dụ này, chiều cao của các cháu trai là biến phụ thuộc, chiều cao của người bố là biến độc lập.

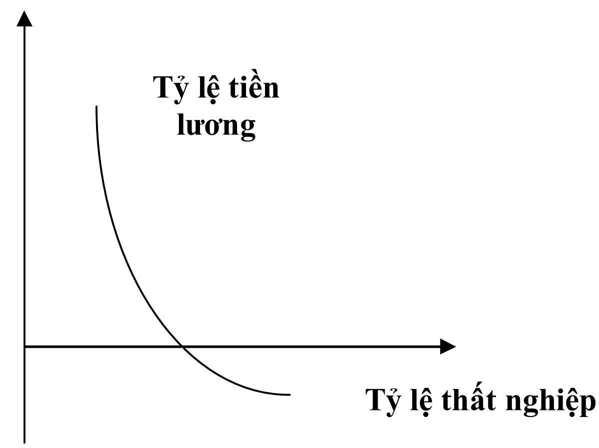
b. Một nhà nghiên cứu sự phụ thuộc của nhu cầu về một loại hàng hoá phụ thuộc vào giá bản thân hàng hoá, thu nhập của người tiêu dùng và giá của những hàng hoá khác cạnh tranh với hàng hoá này.

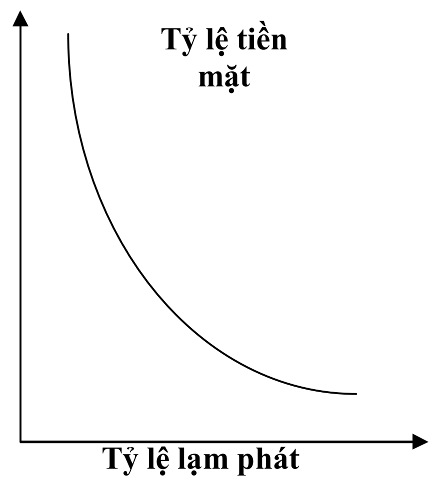
Trong trường hợp này, nhu cầu là biến phụ thuộc, giá của bản thân hàng hoá, của các hàng hoá cạnh tranh và thu nhập của người tiêu dùng là các biến độc lập.

c. Một nhà kinh tế lao động đã nghiên cứu tỷ lệ thay đổi của tiền lương trong quan hệ với tỷ lệ thất nghiệp đã đưa ra đồ thị ở *hình 1.2*. Đường cong trên hình 1.2 được gọi là đường cong Philip, trong đó: tỷ lệ thay đổi của tiền lương là biến phụ thuộc, tỷ lệ thất nghiệp - biến độc lập. Mô hình cho phép dự đoán được sự thay đổi trung bình của tỷ lệ tiền lương với một tỷ lệ thất nghiệp nhất định.

d. Trong điều kiện các yếu tố khác không đổi, tỷ lệ lạm phát càng cao thì tỷ lệ thu nhập của nhân dân được giữ dưới dạng tiền mặt càng ít.

Có thể minh hoạ điều đó bằng đồ thị ở hình 1.3





Hình 1.2: Mối quan hệ tiền lương và thất nghiệp Hình 1.3: Mối quan hệ giữa tiền mặt và lạm phát

Ta có thể đưa ra được rất nhiều ví dụ về sự phụ thuộc của một biến vào một hoặc nhiều biến khác. Kỹ thuật phân tích hồi quy giúp ta nghiên cứu mối quan hệ như vậy giữa các biến.

Các ký hiệu: Y- biến phụ thuộc

X_i - biến độc lập

Trong đó, biến phụ thuộc Y là đại lượng ngẫu nhiên, có quy luật phân bố xác suất, các biến độc lập X_i không phải là biến ngẫu nhiên, giá trị của chúng đã được cho trước.

Phân tích hồi qui giải quyết các vấn đề sau:

1. Ước lượng giá trị trung bình của biến phụ thuộc với giá trị đã cho của biến độc lập.

2. Kiểm định giả thiết về bản chất của sự phụ thuộc.

3. Dự đoán giá trị trung bình của biến phụ thuộc khi biết giá trị của các biến độc lập.

4. Kết hợp các vấn đề trên.

Trong phân tích hồi quy chúng ta phân biệt các quan hệ sau đây:

2.1.2. Quan hệ thống kê và quan hệ hàm số

Vấn đề mấu chốt trong phân tích hồi quy là sự phụ thuộc thống kê của biến phụ thuộc vào một hay nhiều biến giải thích. Biến phụ thuộc là đại lượng ngẫu nhiên, có phân bố xác suất. Các biến giải thích thì giá trị của chúng đã biết. Biến phụ thuộc là ngẫu nhiên vì có vô vàn nhân tố tác động đến nó mà trong mô hình ta không đề cập đến được, ứng với mỗi giá trị đã biết của biến độc lập có thể có nhiều giá trị khác nhau của biến phụ thuộc. Trong quan hệ hàm số các biến không phải là ngẫu nhiên, ứng với mỗi giá trị của biến độc lập có một giá trị của biến phụ thuộc, phân tích hồi quy không quan tâm đến các quan hệ hàm số.

Thí dụ 1.2: Sự phụ thuộc của năng suất một loại lúa trên một hécta vào nhiệt độ, lượng mưa, độ chiếu sáng, phân bón... là quan hệ thống kê. Các biến: nhiệt độ, lượng mưa, độ chiếu sáng, phân bón là các biến độc lập. Năng suất tính trên 1 hécta là biến phụ thuộc, là đại lượng ngẫu nhiên không thể dự báo một cách chính xác năng suất của giống lúa này trên một héc ta vì:

- Có sai số trong phép đo các biến này.
- Còn rất nhiều nhân tố khác cũng ảnh hưởng đến năng suất mà ta không liệt kê ra và nếu có cũng không thể tách được ảnh hưởng của riêng từng nhân tố đến năng suất dù rằng chúng ta có đưa thêm bao nhiêu biến giải thích khác.

Trong vật lý khi xét một động tử chuyển động đều, người ta có công thức:

$$S = v.t$$

Trong đó: S là độ dài quãng đường đi được; v là vận tốc trong một đơn vị thời gian; t là thời gian.

Đây là quan hệ hàm số, ứng với mỗi giá trị của vận tốc và thời gian ta chỉ có một giá trị duy nhất của độ dài quãng đường, phân tích hồi quy không xét các quan hệ này.

2.1.3. Hàm hồi quy và quan hệ nhân quả

Phân tích hồi quy nghiên cứu quan hệ giữa một biến phụ thuộc với một hoặc nhiều biến độc lập khác. Điều này không đòi hỏi giữa biến độc lập và các biến phụ thuộc có mối quan hệ nhân quả. Nếu như quan hệ nhân quả tồn tại thì nó phải xác lập dựa trên các lý thuyết kinh tế khác. Thí dụ, luật cầu nói rằng trong điều kiện các biến khác không đổi thì nhu cầu về một loại hàng hoá tỷ lệ nghịch với giá của chính hàng hoá này hay trong thí dụ 1.2 chúng ta có thể dự đoán sản lượng dựa vào lượng mưa và các biến khác, nhưng không thể chấp nhận được việc dự báo lượng mưa bằng việc thay đổi sản lượng.

2.1.4. Hồi quy và tương quan

Hồi quy và tương quan khác nhau về: mục đích, kỹ thuật. Phân tích tương quan trước hết là đo mức độ kết hợp tuyến tính giữa hai biến. Ví dụ: Mức độ quan hệ giữa nghiện thuốc lá và ung thư phổi, giữa kết quả thi môn thống kê và môn toán. Nhưng phân tích hồi quy lại ước lượng hoặc dự báo một biến trên cơ sở giá trị đã cho của các biến khác. Về kỹ thuật, trong phân tích hồi quy các biến không có tính chất đối xứng. Biến phụ thuộc là đại lượng ngẫu nhiên. Các biến giải thích giá trị của chúng đã được xác định. Trong phân tích tương quan, không có sự phân biệt giữa các biến, chúng có tính chất đối xứng:

$$r(X,Y) = r(Y,X)$$

2.1.5. Bản chất và nguồn số liệu cho phân tích hồi quy

Thành công của bất kỳ một sự phân tích kinh tế nào đều phụ thuộc việc sử dụng các số liệu thích hợp và phụ thuộc vào phương pháp xử lý các số liệu đó. Do vậy ở đây sẽ trình bày đôi nét về bản chất, nguồn gốc và những hạn chế của số liệu mà chúng ta sẽ gặp phải trong phân tích kinh tế nói chung và phân tích hồi quy nói riêng.

2.1.5.1. Các loại số liệu

Có ba loại số liệu: các số liệu theo thời gian (chuỗi thời gian), các số liệu chéo và các số liệu hỗn hợp của hai loại trên.

Các số liệu theo thời gian là các số liệu được thu thập trong một thời kỳ nhất định, ví dụ như các số liệu về GNP, số người thất nghiệp, lượng cung về tiền... Có số liệu được thu thập hàng năm như lượng cung về tiền, có số liệu thu thập hàng tháng, quý, năm... Các số liệu này có thể đo được bằng những con số như giá cả, thu thập nhưng cũng có những số liệu không đo được bằng con số, chúng là những chỉ tiêu chất lượng như: nam hoặc nữ, có gia đình hay chưa có gia đình, có việc làm hay chưa có việc làm... Người ta gọi các biến loại này là biến giả. Chúng cũng quan trọng như những biến số lượng khác.

Các số liệu chéo là các số liệu về một hay nhiều biến được thu thập tại một thời điểm ở nhiều địa phương, đơn vị khác nhau: ví dụ như các số liệu về điều tra dân số vào 0 giờ ngày 01/4/2009; các số liệu điều tra về vốn cơ bản của các xí nghiệp Than ngày 01/01/2011 ở Việt Nam.

Các số liệu hỗn hợp theo thời gian và không gian: các số liệu về giá vàng hàng ngày ở thành phố Hà Nội, Hải Phòng, Vinh.

2.1.5.2. Nguồn gốc số liệu

Các số liệu có thể do cơ quan nhà nước, các tổ chức quốc tế, các công ty tư nhân hay các cá nhân thu thập. Chúng có thể là các số liệu thực nghiệm hoặc không phải thực nghiệm. Các số liệu thực nghiệm thường được thu thập trong khoa học tự nhiên, một điều tra viên muốn thu thập các số liệu ảnh hưởng của một số nhân tố đến đối tượng nghiên cứu, anh ta đã giữ nguyên các yếu tố khác. Thí dụ, một kỹ sư nông nghiệp nghiên cứu khả năng chịu bệnh của một giống lúa mới. Anh ta tiến hành thí nghiệm bằng cách trồng hai giống lúa mới và cũ trên hai khu ruộng có độ màu mỡ như nhau, thực hiện chế độ chăm sóc hai khu ruộng như nhau và theo dõi sự phát triển của sâu bệnh trên hai khu ruộng này bằng cách gây cùng một loại bệnh trên hai khu ruộng. Các số liệu thu được sẽ là các số liệu thực nghiệm.

Trong khoa học xã hội, các số liệu nói chung là các số liệu không phải do thực nghiệm mà có. Các số liệu về GNP, số người thất nghiệp, giá cổ phiếu... không nằm dưới sự kiểm soát của kỹ thuật viên. Điều này thường gây ra những vấn đề đặc biệt trong việc tìm ra những nguyên nhân chính xác ảnh hưởng đến một tình huống riêng biệt. Thí dụ: có phải lượng cung về tiền ảnh hưởng đến GDP hay còn có nguyên nhân khác...?

2.1.5.3. Nhược điểm của các số liệu

Chất lượng của các số liệu thu được thường không tốt. Điều đó do các nguyên nhân sau đây:

- Hầu hết các số liệu trong khoa học xã hội đều là các số liệu phi thực nghiệm. Do vậy, có thể có sai số quan sát hoặc bỏ sót quan sát hoặc cả hai.
- Ngay với các số liệu được thu thập bằng thực nghiệm cũng có sai số của phép đo.
- Trong các cuộc điều tra bằng câu hỏi, vấn đề không nhận được câu trả lời hoặc có trả lời nhưng không trả lời hết các câu hỏi.
- Các mẫu được thu thập trong các cuộc điều tra rất khác nhau về kích cỡ cho nên rất khó khăn trong việc so sánh các kết quả giữa các đợt điều tra.
- Các số liệu kinh tế thường có sẵn ở mức tổng hợp cao, không cho phép đi sâu vào các đơn vị nhỏ.
- Ngoài ra còn có những số liệu thuộc bí mật quốc gia mà không phải ai cũng có thể sử dụng được.

Do vậy, kết quả nghiên cứu sẽ phụ thuộc vào chất lượng của các số liệu được sử dụng và phụ thuộc vào mô hình được lựa chọn.

2.2. MÔ HÌNH HỒI QUY TỔNG THỂ

Phần trên chúng ta nói phân tích hồi quy đặc biệt quan tâm đến ước lượng hoặc dự báo giá trị trung bình của biến phụ thuộc trên cơ sở biết các giá trị của biến độc lập. Ta xét các thí dụ sau đây:

Thí dụ 1.3:

Y: Chi tiêu của một gia đình trong một tuần tính bằng $.

X: Thu nhập sau khi đã trừ thuế của một gia đình tính bằng $.

Giả sử rằng ở một địa phương chỉ có cả thảy 60 gia đình, 60 gia đình này được chia thành 10 nhóm, chênh lệch về thu nhập của các nhóm gia đình này sang nhóm gia đình tiếp theo đều bằng nhau. Ta có bảng số liệu sau đây:

Bảng 2.1. Thu nhập và chi tiêu trong một tuần của tổng thể

Y \ X	**80**	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**
	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	.	88	.	113	125	140	.	160	189	185
	.	.	.	115	.	.	.	162	.	191
Tổng	**325**	**462**	**445**	**707**	**678**	**750**	**685**	**1043**	**966**	**1211**

Các số liệu ở bảng trên có nghĩa là: với thu thập trong một tuần chẳng hạn X =$100 thì có 6 gia đình mà chi tiêu trong tuần nằm giữa 65 và 88. Hay nói khác đi ở mỗi cột của bảng cho ta phân bố của số chi tiêu trong tuần Y với mức thu nhập đã cho X, đó chính là phân bố có điều kiện của Y với giá trị X đã cho.

Vì bảng 2.1 là tổng thể nên ta dễ dàng tìm P(Y/X). Chẳng hạn, P(Y=85,X=100)=1/6. Chúng ta có bảng xác suất có điều kiện sau đây;

Bảng 1.2. Xác suất có điều kiện P(Y/X)

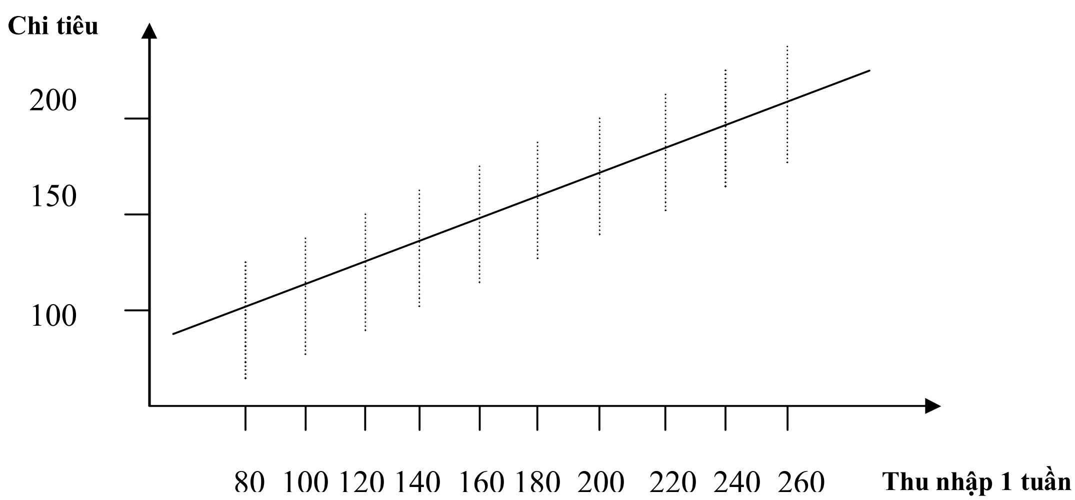
Y \ X	**80**	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	.	1/6	.	1/7	1/6	1/6	.	1/7	1/6	1/7
	.	.	.	1/7	.	.	.	1/7	.	1/7
Tổng	**65**	**77**	**89**	**101**	**113**	**125**	**137**	**149**	**161**	**173**

Trong đó: $E(Y/X_i) = \sum_{j=1}^{n} Y_j P(Y = Y_j / X = X_i)$

Chẳng hạn $E(Y/100) = \sum_{j=1}^{n} Y_j P(Y = Y_j / X = 100)$

$$= 65*1/6 + 70*1/6 + 74*1/6 + 80*1/6 + 85*1/6 + 88*1/6 = 77$$

Biểu diễn các điểm của bảng 1.1 và các trung bình $E(Y/X_i)$: i=1,2,..10 lên hệ toạ độ, nối các điểm có toạ độ (X_i, $E(Y/X_i)$) ta được đồ thị sau đây:



Theo hình 1.4, ta thấy rằng trung bình có điều kiện của mức chi tiêu trong tuần nằm trên đường thẳng có hệ số góc tương đương. Khi thu nhập tăng thì mức chi tiêu cũng tăng.

Một cách tổng quát, $E(Y/X_i)$ là một hàm của X_i:

$$E(Y/X_i) = f(X_i) \qquad (1.1)$$

Trong đó $f(X_i)$ là một hàm nào đó của biến giải thích X_i, với ví dụ trên $f(X_i)$ là hàm tuyến tính.

Phương trình (1.1) gọi là hàm hồi quy tổng thể (PRF). Nếu như hàm hồi quy tổng thể có một biến độc lập thì gọi là hàm hồi quy đơn, có hơn một biến độc lập thì gọi là hàm hồi quy bội.

Hàm hồi quy tổng thể cho chúng ta biết giá trị trung bình của biến Y sẽ thay đổi như thế nào theo X.

Hàm $f(X_i)$ có dạng như thế nào - tuyến tính hay phi tuyến - chúng ta chưa biết được bởi lẽ trong thực tế chúng ta chưa có sẵn tổng thể để kiểm tra. Xác định đúng hàm hồi quy là vấn đề thực nghiệm (chúng ta sẽ đề cập đến vấn đề này các chương sau...).

Giả sử rằng PRF $E(Y/X_i)$ là hàm tuyến tính:

$$E(Y/X_i) = \beta_1 + \beta_2 X_i \qquad (1.2)$$

Trong đó β_1, β_2 là các tham số chưa biết nhưng cố định, và được gọi là các hệ số hồi quy.

β_1 là hệ số tự do (hệ số chặn).

β_2 là hệ số góc.

Phương trình (1.2) gọi là phương trình hồi quy tuyến tính đơn.

Trong phân tích hồi quy chúng ta phải ước lượng giá trị trung bình của biến Y tức là ước lượng hàm hồi quy chẳng hạn dạng (1.2). ở phương trình (1.2), giá trị của các X_i ta đã biết, do vậy việc ước lượng (1.2) trở thành việc ước lượng các tham số chưa biết β_1, β_2, trên cơ sở những quan sát của Y và X.

Thuật ngữ "tuyến tính" ở đây được hiểu theo hai nghĩa: tuyến tính đối với tham số và tuyến tính đối với các biến. Thí dụ $E(Y/X_i) = \beta_1 + \beta_2 X_i^2$ là hàm tuyến tính đối với tham số nhưng không tuyến tính (phi tuyến) đối với biến; $E(Y/X_i) = \beta_1 + \sqrt{\beta_2} X_i$ là hàm tuyến tính đối với biến nhưng phi tuyến đối với tham số.

Hàm hồi quy tuyến tính luôn luôn được hiểu là hồi quy tuyến tính đối với các tham số, nó có thể hoặc không phải là tuyến tính đối với biến.

2.2.1. Sai số ngẫu nhiên và bản chất của nó

Giả sử chúng ta đã có hàm hồi quy tổng thể $E(Y/X_i)$; vì $E(Y/X_i)$ là giá trị trung bình của biến Y với giá trị X_i đã biết, cho nên các giá trị cá biệt Y_i không phải bao giờ cũng trùng với $E(Y/X_i)$ mà chúng xoay quanh nó.

Kí hiệu U_i là chênh lệch giữa giá trị cá biệt Y_i và $E(Y/X_i)$:

$$U_i = Y_i - E(Y/X_i)$$

Hay: $Y_i = E(Y/X_i) + U_i$ (1.3)

U_i là biến ngẫu nhiên, người ta gọi U_i là yếu tố ngẫu nhiên (hoặc nhiễu) và (1.3) được gọi là PRF ngẫu nhiên.

Nếu như $E(Y/X_i)$ là tuyến tính đối với X_i thì $Y_i = \beta_1 + \beta_2 X_i + U_i$

Với thí dụ (1.3) và với X=$100 ta có:

$$Y_1 = 65 = \beta_1 + 100\beta_2 + U_1$$

$$Y_2 = 70 = \beta_1 + 100\beta_2 + U_2$$

$$Y_3 = 74 = \beta_1 + 100\beta_2 + U_3$$

$$Y_4 = 80 = \beta_1 + 100\beta_2 + U_4$$

$$Y_5 = 85 = \beta_1 + 100\beta_2 + U_5$$

$$Y_6 = 88 = \beta_1 + 100\beta_2 + U_6$$

Từ (1.3) ta có

$$E(Y_i/X_i) = E(E(Y/X_i) + E(U_i/X_i))$$

$$E(Y_i/X_i) = E(Y/X_i) + E(U_i/X_i) \quad (1.4)$$

$$E(U_i/X_i) = 0$$

Như vậy, nếu đường hồi quy của tổng thể đi qua các trung bình có điều kiện của Y thì $E(U_i/X_i) = 0$, trong trường hợp này (1.2) và (1.3) là như nhau. Nhưng (1.3) chỉ ra rằng ngoài các biến giải thích đã có trong mô hình còn có các yếu tố khác ảnh hưởng đến biến phụ thuộc Y. Nhưng trung bình ảnh hưởng của các yếu tố này đến biến phụ thuộc bằng 0 và do vậy không cần đưa các yếu tố này vào mô hình.

2.2.2. Hàm hồi quy mẫu

Ở phần phụ lục, chương II có trình bày sơ lược về tổng thể và mẫu, vì sao phải nghiên cứu mẫu. Vấn đề này ở đây cũng tương tự như vậy. Chúng ta không có tổng thể, hoặc có nhưng không thể nghiên cứu được toàn bộ tổng thể. Chúng ta chỉ có mẫu ngẫu nhiên được lấy từ tổng thể. Chúng ta muốn ước lượng PRF từ những thông tin thu được trên mẫu ngẫu nhiên của các giá trị Y đối với các giá trị của X đã biết. Một điều chắc chắn rằng chúng ta không thể ước lượng một cách chính xác PRF dựa trên mẫu ngẫu nhiên.

Hàm hồi quy được xây dựng trên cơ sở một mẫu ngẫu nhiên được gọi là hàm hồi quy mẫu (SRF) hoặc hồi quy mẫu.

Bảng 2.3 và 2.4 cho 2 mẫu ngẫu nhiên từ tổng thể trong thí dụ 1.2

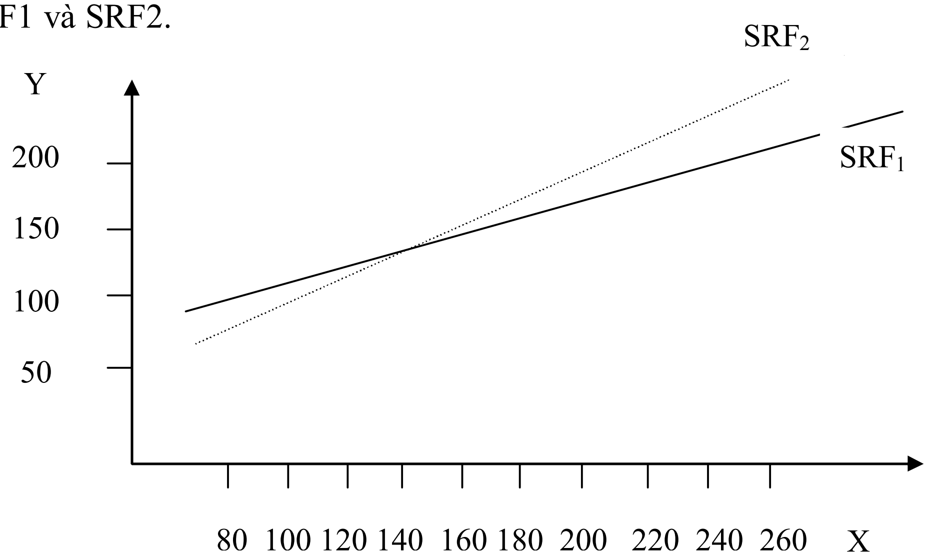
Bảng 2.3. Mẫu thứ nhất

Y	70	65	90	95	110	1145	120	140	155	150
X	80	100	120	140	160	180	200	220	240	260

Bảng 2.4. Mẫu thứ hai

Y	55	88	90	80	118	120	145	175
X	80	100	120	140	160	180	200	220

Với hai mẫu ngẫu nhiên ta xây dựng được hai hàm hồi quy mẫu ký hiệu SRF1 và SRF2.



Vậy đường hồi quy nào sẽ được coi là thích hợp với PRF. Câu hỏi này không trả lời được bởi lẽ PRF chưa biết. Cũng giống như ước lượng một tham số, ta sẽ ước lượng PRF bằng SRF mà SRF này có tính chất: tuyến tính, không chệch, có phương sai nhỏ nhất.

Giả sử rằng đường hồi quy mẫu có dạng: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Trong đó: $\hat{Y}_i$ là ước lượng của E(Y/X$_i$)

$\hat{\beta}_1, \hat{\beta}_2$ là ước lượng của β_1, β_2.

Mặt khác theo (1.3): $Y_i = E(Y/X_i) + U_i$

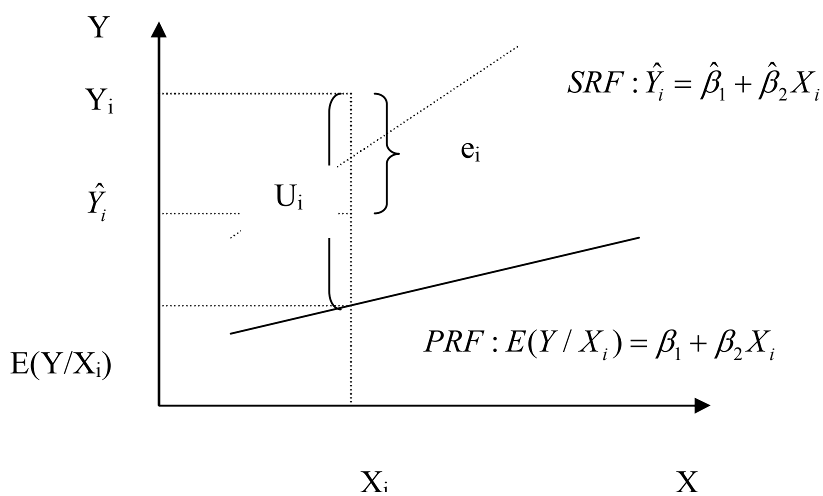
Do đó: $$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

Trong đó, e_i được gọi là phần dư hay chính là ước lượng của U_i. Sự tồn tại của e_i được giải thích như sự tồn tại của U_i.

Trên mẫu, với $X = X_i$ ta có $Y = Y_i$

$$Y_i = \hat{Y}_i + e_i$$

$$Y_i = E(Y/X_i) + U_i$$



Vấn đề đặt ra là ta có thể đưa ra một phương pháp và một số điều kiện mà nhờ nó SRF là ước lượng tuyến tính, không chệch có phương sai nhỏ nhất của PRF hay nói khác đi $\hat{\beta}_1$ và $\hat{\beta}_2$ gần nhất với giá trị thực β_1 và β_2 có thể được dù rằng chúng ta không bao giờ biết giá trị thực của β_1 và β_2.

2.3. ƯỚC LƯỢNG VÀ KIỂM ĐỊNH TRONG MÔ HÌNH HỒI QUY HAI BIẾN

2.3.1. Phương pháp bình phương nhỏ nhất và các giả thiết của nó

2.3.1.1. Nội dung phương pháp bình phương nhỏ nhất

Giả sử $E(Y/X_i) = \beta_1 + \beta_2 X_i$ là PRF

Khi đó giá trị quan sát Y_i:

$Y_i = E(Y/X_i) + U_i = \beta_1 + \beta_2 X_i + U_i$

$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ là SRF

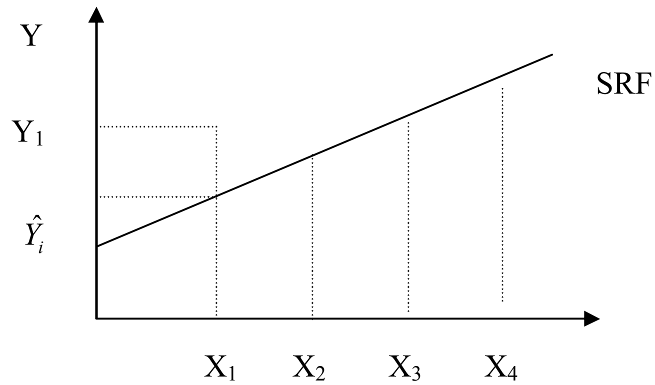
$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

Vấn đề là phải tìm $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Giả sử rằng chúng ta có n cặp quan sát của Y và X, cặp quan sát thứ i có giá trị tương ứng (Y_i,X_i): i =1..n. Ta phải tìm $\hat{Y}_i$ sao cho nó càng gần với giá trị thực của Y_i có thể được, tức là phần dư:

$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_2$ càng nhỏ càng tốt.

Ta xem đồ thị sau:



Do e_i: i = 1..n có thể dương, có thể âm do vậy cần phải tìm $\hat{Y}_i$ sao cho tổng bình phương của các phần dư đạt cực tiểu. Tức là:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2 \Rightarrow \min$$

Do X_i, Y_i: i = 1..n đã biết, nên $\sum_{i=1}^{n} e_1^2$ là hàm của:

$$f(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{n} e_i^2 = \sum_{i=n}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2 \Rightarrow \min$$

$\hat{\beta}_1, \hat{\beta}_2$ là nghiệm của hệ thống phương trình sau:

$$\frac{\partial f(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} = \sum 2\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)(-1) = 0$$

hay $\quad n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$

$$\frac{\partial f(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)(-X_i) = 0$$

hay $\quad \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i$

$\hat{\beta}_1, \hat{\beta}_2$ được tìm từ hệ phương trình sau:

$$n\hat{\beta}_1 + \hat{\beta}_2\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

$$\hat{\beta}_1\sum_{i=1}^{n} X_i + \hat{\beta}_2\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i \qquad (2.1)$$

Hệ phương trình (2.1) gọi là hệ phương trình chuẩn. Giải hệ phương trình trên ta được:

$$\hat{\beta}_2 = \frac{n\sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left[\sum_{i=1}^{n} X_i\right]^2}$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2\overline{X}$$

Đặt $x_i = X_i - \overline{X}$

$y_i = Y_i - \overline{Y}$

Khi đó $\hat{\beta}_2 = \dfrac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$

$\hat{\beta}_1, \hat{\beta}_2$ là các ước lượng của β_1 và β_2 được tính bằng phương pháp bình phương nhỏ nhất - được gọi là các ước lượng bình phương nhỏ nhất.

2.3.1.2. Các tính chất của các ước lượng bình phương nhỏ nhất

- $\hat{\beta}_1, \hat{\beta}_2$ được xác định một cách duy nhất ứng với n cặp quan sát (X_i,Y_i).
- $\hat{\beta}_1, \hat{\beta}_2$ là các ước lượng điểm của β_1, β_2 và là các đại lượng ngẫu nhiên, với các mẫu khác nhau chúng có giá trị khác nhau.

$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i - SRF$ có các tính chất sau đây:

- SRF đi qua trung bình mẫu $(\overline{X}, \overline{Y})$, nghĩa là:

$$\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2\overline{X}$$

- Giá trị trung bình của $\hat{Y}_i$ bằng giá trị trung bình của các quan sát: $\overline{\hat{Y}} = \overline{Y}$
- Giá trị trung bình của các phần dư: $\sum_{i=1}^{n} e_i = 0$
- Các phần dư e_i không tương quan với $\hat{Y}_i$ tức là $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$
- Các phần dư e_i không tương quan với X_i tức là $\sum_{i=1}^{n} e_i X_i = 0$.

2.3.1.3. Các giả thiết cơ bản của phương pháp bình phương nhỏ nhất

Trong phân tích hồi quy, mục đích của chúng ta là ước lượng, dự báo về tổng thể, tức là ước lượng E(Y/X_i) hay trong mô hình hồi quy tuyến tính đơn là ước lượng $E(Y/X_i) = \beta_1 + \beta_2 X_i$. $\hat{\beta}_1$ và $\hat{\beta}_2$ tìm được bằng phương pháp bình phương nhỏ nhất là các ước lượng điểm của β_1 và β_2. Chúng ta không biết được chất lượng của các ước lượng này như thế nào. Chất lượng của các ước lượng phụ thuộc vào:

- Dạng hàm của mô hình được lựa chọn.
- Phụ thuộc vào các X_i và U_i.
- Phụ thuộc vào kích thước mẫu.

Giả thiết 1: Các biến giải thích là phi ngẫu nhiên, tức là các giá trị của chúng là các số đã được xác định.

Giả thiết này không có gì mới, vì phân tích hồi quy được đề cập là phân tích hồi quy có điều kiện, phụ thuộc vào các giá trị X đã cho.

Giả thiết 2: Kỳ vọng của yếu tố ngẫu nhiên bằng không, tức là:

$$E(U_i/X_i) = 0$$

Giả thiết này có nghĩa là các yếu tố không có trong mô hình, U_i đại diện cho chúng, không có ảnh hưởng hệ thống đến giá trị trung bình của Y.

Giả thiết 3: Phương sai bằng nhau (phương sai thuần nhất) của các U_i.

$$Var\left(U_i / X_i\right) = Var\left(U_i / X_i\right) = \sigma^2$$

Điều này có nghĩa là phân bố có điều kiện của Y với giá trị đã cho của X có phương sai bằng nhau, các giá trị cá biệt của Y xoay quanh giá trị trung bình với phương sai như nhau. Giả thiết 3 kéo theo $Var(Y_i / X_i) = \sigma^2$.

Giả thiết 4: Không có sự tương quan giữa các U_i:

$$Cov(U_i, U_j) = 0$$

Giả thiết này có nghĩa là U_i là ngẫu nhiên. Về mặt hình học, có nghĩa là nếu như có một giá trị U nào đó lớn hơn (nhỏ hơn) giá trị trung bình thì không có nghĩa giá trị khác cũng lớn hơn (nhỏ hơn) giá trị trung bình.

Giả thiết 5: U_i và X_i không tương quan với nhau:

$$Cov(U_i, X_i) = 0$$

2.3.2. Độ chính xác của phương pháp bình phương nhỏ nhất

Theo phương pháp bình phương nhỏ nhất, các ước lượng $\hat{\beta}_1, \hat{\beta}_2$ được xác định theo công thức:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

$$\hat{\beta}_2 = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$$

Các ước lượng này là hàm của mẫu, là đại lượng ngẫu nhiên, với các mẫu khác nhau ta có ước lượng khác nhau. Vì phương sai hay độ lệch chuẩn đặc trưng cho độ phân tán của đại lượng ngẫu nhiên, nên ta dùng chúng làm thước đo cho chất lượng của ước lượng.

Với các giả thiết của phương pháp bình phương nhỏ nhất, phương sai và độ lệch chuẩn của các ước lượng được cho bởi công thức sau:

$$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}; \qquad se\left(\hat{\beta}_2\right) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

$$\operatorname{var}\left(\hat{\beta}_1\right) = \frac{\sum_{i=4}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\sigma^2; \qquad se\left(\hat{\beta}_1\right) = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}}$$

Trong đó, $\sigma^2 = Var(U_i)$

se: sai số tiêu chuẩn

$x_i = X_i - \overline{X}$

Trong các công thức trên σ^2 chưa biết. σ^2 được ước lượng bằng ước lượng không chệch của nó là $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$; $\hat{\sigma} = \sqrt{\sum_{i=1}^{n} e_i^2 / (n-2)}$ là sai số tiêu chuẩn của đường hồi quy. Nó chính là độ lệch tiêu chuẩn các giá trị Y quanh đường hồi quy mẫu.

Các tính chất của các ước lượng bình phương nhỏ nhất được thể hiện qua định lý sau đây:

Định lý Gauss – Markov: Với các giả thiết 1-5 của phương pháp bình phương nhỏ nhất, các ước lượng bình phương nhỏ nhất là các ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất trong lớp các ước lượng tuyến tính không chệch.

2.3.3. Hệ số r^2 đo độ phù hợp của hàm hồi quy mẫu

Ta có $Y_i = \hat{Y}_i + e_i$

$Y_i - \overline{Y} = \hat{Y}_i - \overline{Y} + e_i = \hat{Y}_i - \overline{\hat{Y}} + e_i$ hay $y_i = \hat{y}_i + e_i$

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 + 0$$

Vì $\hat{y}_i = \hat{\beta}_2 x_i$ nên: $\sum_{i=1}^{n} y_i^2 = \hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2$

Ký hiệu: $TSS = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$

- TSS: là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát Y_i với giá trị trung bình của chúng.

$$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{\hat{Y}}\right) = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \hat{y}_i^2 = \hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2$$

- ESS: là tổng bình phương của tất cả các sai lệch giữa các giá trị của biến phụ thuộc Y nhận được từ hàm hồi quy mẫu với giá trị trung bình của chúng. Phần này đo độ chính xác của hàm hồi quy.

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

- RSS: là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát Y và các giá trị nhận được từ hàm hồi quy.

TSS được chia thành hai phần: một phần ESS do đường hồi quy mẫu gây ra và phần của RSS do các yếu tố ngẫu nhiên gây ra.

Từ TSS = ESS + RSS ta chia cả hai vế cho TSS, ta có:

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} + \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}$$

$$r^2 = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \hat{\beta}_2^2 \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$$

$$r^2 = \hat{\beta}_2^2 \frac{\sum_{i=1}^{n} x_i^2 \Big/ n-1}{\sum_{i=1}^{n} y_i^2 \Big/ n-1} = \hat{\beta}_2^2 \frac{S_X^2}{S_Y^2}$$

trong đó: S_X^2 và S_Y^2 là phương sai mẫu của X và Y.

Mặt khác: $\hat{\beta}_2 = \dfrac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \Rightarrow r^2 = \dfrac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)}} = \frac{\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}}$$

Từ định nghĩa r^2 chúng ta thấy r^2 đo tỷ lệ hay số phần trăm của toàn bộ sai lệch của Y với giá trị trung bình của chúng được giải thích bằng mô hình (hay biến độc lập). r^2 được sử dụng để đo độ thích hợp của hàm hồi quy. Dễ dàng thấy được $0 \le r^2 \le 1$. Nếu lấy căn bậc hai của r^2 ta được r. r chính là hệ số tương quan mẫu, tuy nhiên dấu của r tuỳ thuộc vào quan hệ cùng chiều hay ngược chiều giữa Y và X.

Các tính chất của hệ số tương quan r:

1. r có thể âm hoặc dương, dấu của r phụ thuộc vào dấu của tỷ số, đó chính là dấu của Cov(X,Y), hay là dấu của hệ số góc.
2. $-1 \le r \le 1$
3. r có tính chất đối xứng r(X,Y)=r(Y,X)
4. Nếu $X^* = aX + c$; $Y^* = bY + a$; a, b, c, d là các hằng số; a, b > 0 thì $r(X^*,Y^*) = r(X,Y)$.
5. Nếu X, Y độc lập với nhau thì r(X,Y) = 0. Điều ngược lại không đúng.
6. r đo sự phụ thuộc tuyến tính. Nhưng không có ý nghĩa trong việc định rõ tính chất các quan hệ phi tuyến.
7. r đo độ phụ thuộc tuyến tính giữa X và Y, vậy không đòi hỏi X, Y có mối quan hệ nhân quả.
8. r^2 cũng có thể tính bằng công thức:

$$r = \frac{\left[\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(\hat{Y}_i - \bar{Y}\right)\right]^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 \sum\left(\hat{Y}_i - \bar{Y}\right)^2} = \frac{\left(\sum_{i=1}^{n} y_i \hat{y}_i\right)^2}{\sum_{i=1}^{n} y_i^2 \sum_{i=1}^{n} \hat{y}_i^2}$$

2.3.4. Phân phối xác suất của U_i

Với các giả thiết cơ bản: $E(U_i) = 0$; var $(U_i) = \sigma^2$; cov$(U_i,U_j) = 0$ thì $\hat{\beta}_1, \hat{\beta}_2$ là các ước lượng tuyến tính không chệch có phương sai nhỏ nhất của β_1 và β_2. Mục đích của phân tích hồi quy không phải chỉ là sự suy đoán về β_1 và β_2 hay PRF mà còn phải kiểm tra bản chất của sự phụ thuộc, còn phải thực hiện các dự đoán khác. Do vậy cần phải biết phân bố xác suất của $\hat{\beta}_1$ và $\hat{\beta}_2$. Các phân bố này phụ thuộc vào phân bố của các U_i.

Bây giờ chúng ta đưa thêm giả thiết.

Giả thiết 6. U_i có phân bố $N(0,\sigma^2)$.

Với các giả thiết trên, các ước lượng bình phương nhỏ nhất $\hat{\beta}_1, \hat{\beta}_2$ và $\hat{\sigma}^2$ có các tính chất sau đây:

1. Chúng là các ước lượng không chệch.
2. Có phương sai cực tiểu.
3. Khi số quan sát đủ lớn thì các ước lượng này xấp xỉ với giá trị thực của phân bố.
4. $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$. Từ tính chất này suy ra $Z = \dfrac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1)$.
5. $\hat{\beta}_2 \sim N\left(\beta_2, \sigma^2_{\hat{\beta}_2}\right)$. Từ tính chất này suy ra $Z = \dfrac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \sim N(0,1)$.
6. $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim X^2(n-2)$
7. Trong các ước lượng không chệch của β_1, β_2 bất kể là tuyến tính hay phi tuyến tính thì $\hat{\beta}_1, \hat{\beta}_2$ có phương sai nhỏ nhất.
8. $Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$.

Với các tính chất trên chúng ta có thể tìm khoảng tin cậy và kiểm định giả thiết về các tham số hồi quy.

2.3.5. Khoảng tin cậy và kiểm định giả thiết về các hệ số hồi quy

2.3.5.1. Khoảng tin cậy của β_1

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim T(n-2)$$

Với hệ số tin cậy 1 - α ta tìm được $t_{\alpha/2}(n-2)$ thoả mãn:

$$P(-t_{\alpha/2}(n-2) \le \frac{\hat{\beta}_1 - \beta_1}{se\left(\hat{\beta}_1\right)} \le t_{\alpha/2}(n-2)) = 1-\alpha$$

Khoảng tin cậy (1 - α) của β_1 là:

$$\left(\hat{\beta}_1 - t_{\alpha/2}(n-2)se(\hat{\beta}_1); \hat{\beta}_1 + t_{\alpha/2}(n-2)se(\hat{\beta}_1)\right)$$

2.3.5.2. Kiểm định giả thiết với β_2

Có thể đưa ra giả thiết nào đó về β_1, chẳng hạn $\beta_1 = \beta_1^*$. Nếu giả thiết này đúng thì:

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim T(n-2)$$

Ta có bảng sau đây:

Bảng 2.4. Kiểm định giả thiết về β_1

Loại giả thiết	Giả thiết H_0	Giả thiết đối H_1	Miền bác bỏ
Hai phía	$\beta_1 = \beta_1^*$	$\beta_1 \neq \beta_1^*$	$\|t\| > t_{\alpha/2}(n-2)$
Phía phải	$\beta_1 \leq \beta_1^*$	$\beta_1 > \beta_1^*$	$t > t_{\alpha}(n-2)$
Phía trái	$\beta_1 \geq \beta_1^*$	$\beta_1 < \beta_1^*$	$t < -t_{\alpha}(n-2)$

α thường nhỏ hơn 0,1.

$t_\alpha(n-2)$ được xác định bởi $P(t > t_\alpha(n-2)) = \alpha$.

2.3.5.3. Khoảng tin cậy của β_2

Dựa vào: $t = \dfrac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \sim T(n-2)$

Do đó với hệ số tin cậy 1 - α, khoảng tin cậy của β_2 được xác định bởi:

$$P(-t_{\alpha/2}(n-2) \leq \frac{\hat{\beta}_2 - \beta_2}{se\left(\hat{\beta}_2\right)} \leq t_{\alpha/2}(n-2)) = 1-\alpha$$

$$\Leftrightarrow P\left(\hat{\beta}_2 - t_{\alpha/2}(n-2)se(\hat{\beta}_2); \hat{\beta}_2 + t_{\alpha/2}(n-2)se(\hat{\beta}_2)\right) = 1-\alpha$$

2.3.5.4. Kiểm định giả thiết đối với β_2

Có thể đưa ra giả thiết về giá trị thực của β_2, chẳng hạn $\beta_2 = \beta_2^*$.

Nếu giả thiết này đúng thì $t = \dfrac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \sim T(n-2)$

Bảng 2.3. Kiểm định giả thiết về β_2

Loại giả thiết	Giả thiết H_0	Giả thiết đối H_1	Miền bác bỏ
Hai phía	$\beta_2 = \beta_2^*$	$\beta_2 \neq \beta_2^*$	$\|t\| > t_{\alpha/2}(n-2)$
Phía phải	$\beta_2 \leq \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_{\alpha}(n-2)$
Phía trái	$\beta_2 \geq \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_{\alpha}(n-2)$

Nếu như đưa ra giả thiết $\beta_2 = \beta_2^* = 0$ thì điều này cónghĩa là đưa ra giả thiết biến độc lập X không ảnh hưởng đến biến phụ thuộc Y, khi các t_i được tính bằng công thức:

$$t = \frac{\hat{\beta}_i - 0}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}, i = 1, 2$$

2.3.5.5. Khoảng tin cậy đối với σ^2.

$$\chi^2 = (n-2)\frac{\hat{\sigma}_2}{\sigma^2} \sim \chi^2(n-2)$$

Do đó, khoảng tin cậy (1 - α) của σ^2 được xác định từ:

$$P\left(\chi^2_{1-\alpha/2}(n-2) \le (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \le \chi^2_{\alpha/2}(n-2)\right) = 1-\alpha$$

hay $P\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2}(n-2)} \le \sigma^2 \le \frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}(n-2)}\right) = 1-\alpha$

2.3.5.6. Kiểm định giả thiết đối với σ^2

Loại giả thiết	**Giả thiết H_0**	**Giả thiết đối H_1**	**Miền bác bỏ**
Hai phía	$\sigma^2 = \sigma_0^2$	$\sigma^2 \ne \sigma_0^2$	$(n-2)\frac{\hat{\sigma}^2}{\sigma_0^2} > \chi_{\alpha/2}^2(n-2)$ hoặc $(n-2)\frac{\hat{\sigma}^2}{\sigma_0^2} < \chi^2_{1-\alpha/2}(n-2)$
Phía phải	$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$(n-2)\frac{\hat{\sigma}^2}{\sigma_0^2} > \chi^2_{\alpha}(n-2)$
Phía trái	$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$(n-2)\frac{\hat{\sigma}^2}{\sigma_0^2} < \chi^2_{1-\alpha}(n-2)$

2.3.5.7. Kiểm định sự phù hợp của hàm hồi quy

Phần này sẽ trình bày việc phân tích hồi quy theo quan điểm của phân tích phương sai, nó cung cấp cho chúng ta một cách khác, hữu ích trong việc giải quyết các vấn đề phán đoán thống kê. Chúng ta có:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 = \hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2$$

hay TSS = ESS + RSS

Mặt khác:

r^2 = ESS/TSS, nên ESS = r^2TSS và RSS = $(1-r^2)$TSS

Do β_2 có phân bố $N(\beta_2, \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2})$

Nên $\dfrac{\hat{\beta}_2 - \beta_2}{\sigma}\sqrt{\sum_{i=1}^{n} x_i^2} \sim N(0,1)$

Và $\dfrac{(\hat{\beta}_2 - \beta_2)^2}{\sigma^2}\sum_{i=1}^{n} x_i^2 \sim \chi^2(1)$

Do $(n-2)\dfrac{\hat{\sigma}^2}{\sigma^2} = \dfrac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2(n-2)$

Nên: $F = \dfrac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum_{i=1}^{n} x_i^2}{\sigma^2} / 1 : \dfrac{\sum_{i=1}^{n} e_i^2}{\sigma^2} / (n-2) = \dfrac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum x_i^2}{\sum_{i=1}^{n} e_i^2 \;/(n-2)}$

Hay: $F = \dfrac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum_{i=1}^{n} x_i^2}{\hat{\sigma}^2}$ có phân bố F(1,n-2).

Chúng ta kiểm định giả thiết:

H_0: $\beta_2 = 0$

H_1: $\beta_2 \neq 0$

Nếu $F = \dfrac{\hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2}{\hat{\sigma}^2} > F_\alpha(1, n-2)$ thì bác bỏ giả thiết H_0, tức là bác bỏ giả thiết X không ảnh hưởng đến Y. Trong đó α thường nhỏ hơn 0,1.

Mặt khác:

$$F = \frac{\hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2}{\hat{\sigma}^2} = \frac{ESS/1}{RSS/(n-2)} = \frac{TSSr^2/1}{(1-r^2)TSS/(n-2)}$$

$$F = \frac{r^2}{1-r^2}\frac{n-1}{1}$$

Cho nên quá trình phân tích phương sai cho phép chúng ta đưa ra các phán đoán thống kê về độ thích hợp của hàm hồi quy. Ta có thể đưa ra quá trình phân tích phương sai một cách ngắn gọn bằng bảng sau đây:

Bảng 2.5. Bảng phân tích phương sai cho mô hình hồi quy hai biến

Nguồn biến thiên	Tổng bình phương	Bậc tự do	Phương sai
Từ hàm hồi quy (ESS)	$\sum_{i=1}^{n}\hat{y}_i^2 = \hat{\beta}_2^2\sum_{i=1}^{n}x_i^2$	1	$\hat{\beta}_2^2\sum_{i=1}^{n}x_i^2$
Từ các yếu tố ngẫu nhiên (RSS)	$\sum_{i=1}^{n}e_i^2$	n-2	$\frac{\sum_{i=1}^{n}e_i^2}{(n-2)} = \hat{\sigma}^2$
TSS	$\sum_{i=1}^{n}y_i^2$	n-1	

2.3.6. Phân tích hồi quy và dự báo

Các phần trên đã trình bày phương pháp xây dựng một hàm hồi quy, các đánh giá và phán xét về các hệ số của hàm hồi quy. Tuy nhiên mục đích của chúng ta không chỉ dừng lại ở đó. Có thể sử dụng hàm hồi quy để dự báo. Có hai loại dự báo:

- Dự báo trung bình có điều kiện của Y với một giá trị X = X_0.
- Dự báo giá trị cá biệt của Y với X = X_0.

Giả sử X = X_0, ta muốn dự báo E(Y/X_0). Đường hồi quy mẫu cho ước lượng điểm của E(Y/X_0): $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$.

$\hat{Y}_0$ là ước lượng không chệch có phương sai nhỏ nhất của E(Y/X_0). Tuy nhiên $\hat{Y}_0$ vẫn khác so với giá trị thực của nó.

$\hat{Y}_0$ có phân bố chuẩn với kỳ vọng $\beta_1 + \beta_2 X_0$, nên

$$Var(\hat{Y}_0) = E(\hat{\beta}_1 + \hat{\beta}_2 X_0 - \beta_1 - \beta_2 X_0)^2$$
$$= E[(\hat{\beta}_1 - \beta_1) + X_0(\hat{\beta}_2 - \beta_2)]^2$$
$$= E[(\hat{\beta}_1 - \beta_1)^2 + 2(\hat{\beta}_1 - \beta_1)X_0(\hat{\beta}_2 - \beta_2) + X_0^{\ 2}(\hat{\beta}_2 - \beta_2)^2]$$
$$Var(\hat{Y}_0) = E[(\hat{\beta}_1 - \beta_1)^2] + E[X_0^{\ 2}(\hat{\beta}_2 - \beta_2)^2] + 2X_0 E[(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)]$$
$$= \mathrm{var}(\hat{\beta}_1) + X_0^2\,\mathrm{var}(\hat{\beta}_2) + 2X_0 Cov(\hat{\beta}_1, \hat{\beta}_2)$$
$$Var(\hat{Y}_0) = \mathrm{var}\left(\overline{Y} - \hat{\beta}_2\overline{X}\right) + X_0^2\,\mathrm{var}(\hat{\beta}_2) + 2X_0 Cov(\hat{\beta}_1, \hat{\beta}_2)$$
$$Var(\hat{Y}_0) = \frac{\sigma^2}{n} + \overline{X}^2\,\mathrm{var}(\hat{\beta}_2) + X_0^2\,\mathrm{var}(\hat{\beta}_2) + 2X_0 Cov(\hat{\beta}_1, \hat{\beta}_2)$$

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = E[(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)]$$

$$\hat{\beta}_1 - \beta_1 = \bar{Y} - \hat{\beta}_2\bar{X} - E(\hat{\beta}_1) = \bar{Y} - \hat{\beta}_2\bar{X} - (\bar{Y} - \beta_2\bar{X})$$

$$\hat{\beta}_1 - \beta_1 = -\bar{X}\left(\hat{\beta}_2 - \beta_2\right)$$

$$Var(\hat{Y}_0) = \frac{\sigma^2}{n} + \bar{X}^2\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} + X_0^2\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} - 2X_0\bar{X}\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

$$Var(Y_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{\sum_{i=1}^{n} x_i^2}\right]$$

Những σ^2 chưa biết, nên ta sử dụng ước lượng không chệch của σ^2 là $\hat{\sigma}^2$, khi đó:

$$t = \frac{\hat{Y}_0 - \left(\beta_1 + \beta_2 X_0\right)}{se(\hat{Y}_0)} \sim T(n-2)$$

Do đó khoảng tin cậy 1- α của E(Y/X_0):

$$P\left[\hat{\beta}_1 + \hat{\beta}_2 X_0 - t_{\alpha/2}(n-2)se(\hat{Y}_0) \le \beta_1 + \beta_2 X_0 \le \hat{\beta}_1 + \hat{\beta}_2 X_0 + t_{\alpha/2}(n-2)se(\hat{Y}_0)\right] = 1 - \alpha$$

$$\hat{Y}_0 - t_{\alpha/2}(n-2)se(\hat{Y}_0) \le E(Y/X_0) \le \hat{Y}_0 + t_{\alpha/2}(n-2)se(\hat{Y}_0)$$

Câu hỏi ôn tập chương 2.

1. Anh (chị) cho biết sự khác nhau giữa:

a) Hàm hồi quy tổng thể và hàm hồi quy mẫu.

b) Hàm hồi quy dạng ngẫu nhiên và hàm hồi quy dạng xác định.

2. Xác định dạng của hàm hồi quy tổng thể như thế nào? Thế nào là một hàm hồi quy bị định dạng sai?

3. Với số liệu được thu thập sau đây, anh chị hãy phân loại nó (số liệu chuỗi thời gian, chéo, hỗn hợp).

a) Thu thập số liệu về nhà ở của người dân ở một thành phố trong 10 năm. Số liệu được thu thập trong từng năm.

b) Thu thập số liệu về nhu cầu nhà ở của các thành phó khác nhau trong cùng một năm.

c) Thu thập số liệu về nhu cầu nhà ở của người dân các thành phố trong thời gian 10 năm.

d) Doanh số bán hàng của công ty trong 1 năm theo từng quý.

e) Doanh số bán hàng của các công ty cùng ngành nghề trong cùng một quý.

f) Doanh số bán của các công ty cùng ngành nghề trong cùng một năm.

4. Giả sử có số liệu thống kê về lãi suất ngân hàng (X-%/ năm), tổng vốn đầu tư (Y - tỷ đồng) trên địa bàn tỉnh A qua 10 năm kế tiếp như sau:

X	7,0	6,5	6,5	6,0	6,0	6,0	5,5	5,5	5,0	4,5
Y	28	32	30	34	32	35	40	42	48	50

Yêu cầu:

a) Lập mô hình hồi quy tuyến tính mô tả quan hệ giữa tổng vốn đầu tư và lãi suất ngân hàng. Nêu ý nghĩa kinh tế của hệ số hồi quy ước lượng được.

b) Kiểm định giả thiết hệ số hồi quy của X trong hàm hồi quy tổng thể bằng 0 với mức ý nghĩa 5% và nêu ý nghĩa của kết quả.

c) Với mức ý nghĩa 5%, có thể xem hệ số góc của mô hình bằng -11 được không?

d) Xác định khoảng tin cậy của hệ số góc, hệ số chặn với độ tin cậy 90%.

e) Đánh giá độ phù hợp của hàm hồi quy với mức ý nghĩa 5%.

f) Nếu lãi suất ngân hàng là 10%/ năm, thì tổng vốn đầu tư trung bình và cá biệt ở tính A là bao nhiêu?

5. Bảng số liệu về giá bán và diện tích sinh hoạt của 13 căn nhà dành cho một gia đình ở tại cộng đồng Thành phố đại học của San Diego vào năm 1990.

Y: giá bán (tính bằng nghìn USD).

X: Diện tích sinh hoạt tính bằng (feet vuông).

Y	19	23	23	28	24	29	28	36	29	29	38	50	42
X	106	125	130	157	160	175	180	187	193	194	225	260	280

Giả sử ta có mô hình:

$Y = \beta_1 + \beta_2 X + u$

Kết quả ước lượng như sau:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,924660686
R Square	0,854997384
Adjusted R Square	0,841815329
Standard Error	3,454095344
Observations	13

ANOVA

	Df	*SS*	*MS*	*F*	*Significance F*
Regression	1	773,838402	773,8384	64,8607	6,12991E-06
Residual	11	131,2385211	11,93077		
Total	12	905,0769231			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1,65398274	3,72149279	0,444441	0,665341	-6,53696766	9,844933
X Variable 1	0,158726064	0,019708675	8,053614	6,13E-06	0,115347563	0,202105

Yêu cầu:
a) Xác định hàm hồi quy mẫu.
b) Ý nghĩa kinh tế của các hệ số nhận được.
c) Kiểm định sự phù hợp của hàm hồi quy.
d) Kiểm định ảnh hưởng của X đến Y
e) Có nhận định rằng: khi diện tích tăng lên 1 (feet vuông) giá trung bình tăng lên 0,5 (nghìn USD). Bạn có tin điều đó không, với mức ý nghĩa 5%.

CHƯƠNG 3. MÔ HÌNH HỒI QUY BỘI

Nội dung cơ bản của chương III là trình bày lại những nội dung cơ bản trong Kinh tế lượng, nhưng tiếp cận dưới dạng ma trận cho mô hình hồi quy nhiều biến. Ngoài ra chương này đề cập đến một số vấn đề đặc trưng liên quan tới mô hình nhiều biến như hiện tượng Đa cộng tuyến, Hệ số xác định bội đã hiệu chỉnh, Hệ số tương quan riêng phần.

Nội dung cơ bản của chương này bao gồm:

O Mô hình hồi quy ba biến

- Các giả thiết của mô hình

- Ước lượng các tham số của mô hình hồi quy ba biến

- Phương sai và độ lệch chuẩn của mô hình hồi quy ba biến

- Khoảng tin cậy của các hệ số hồi quy

- Kiểm định giả thiết các hệ số hồi quy

O Mô hình hồi quy k biến – Phương pháp ma trận

- Hàm hồi quy tổng thể

- Ước lượng các tham số

- Ma trận hiệp phương sai

- Kiểm định giả thiết

- Dự báo

Mô hình hồi quy hai biến được trình bày ở chương II thường là không phù hợp trong thực tiễn. Có nhiều biến tác động đến biến phụ thuộc Y. Thí dụ như khi nghiên cứu nhu cầu về một loại hàng hoá nào đó (Y), thì nhu cầu này phụ thuộc vào nhiều yếu tố, trước hết là: Thu nhập của người tiêu dùng, giá của bản thân hàng hoá, giá của các loại hàng hoá thay thế hàng hoá này... Do đó cần phải mở rộng mô hình hai biến thành mô hình có nhiều biến hơn.

3.1. MÔ HÌNH HỒI QUY 3 BIẾN

Giống như trong mô hình hai biến, hàm hồi quy 3 biến của tổng thể PRF có dạng: $E(Y/X_2, X_3) = \beta_1 + \beta_2 X_2 + \beta_3 X_3$

PRF là kỳ vọng có điều kiện của biến Y với giá trị đã cho của các biến X_2 và X_3. Trong đó Y: biến phụ thuộc; X_2, X_3: Các biến độc lập.

β_1 : Hệ số tự do (hệ số chặn), nó chính là giá trị trung bình của biến Y khi $X_2 = X_3 = 0$.

β_2, β_3 : Gọi là các hệ số hồi qui riêng.

Y_i là giá trị của biến Y ở quan sát thứ i, khi đó:

$Y_i = E(Y / X_{2i}, X_{3i}) + U_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$

U_i là yếu tố ngẫu nhiên, sự tồn tại của U_i đã được giải thích ở chương 2.

3.1.1. Các giả thiết của mô hình

- Các U_i có kỳ vọng bằng 0: $E(U_i / X_{2i}, X_{3i}) = 0 (\forall i)$
- Không có sự tương quan giữa các U_i: $Cov(U_i, U_j) = 0$.
- Các U_i thuần nhất: $Var(U_i) = \sigma^2$
- Giữa các biến giải thích X_2, X_3 không có quan hệ tuyến tính.
- U_i có phân bố $N(0, \sigma^2)$.

Trong mô hình hồi quy bội có thêm một giả thiết mới - giả thiết thứ 4 -giữa các biến X_2, X_3 không có quan hệ tuyến tính. Nếu như X_2, X_3 có quan hệ tuyến tính với nhau thì người ta nói rằng có hiện tượng đa cộng tuyến.

Ý nghĩa của các hệ số β_2 và β_3

$E(Y / X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$

$\frac{\partial E}{\partial X_2} = \beta_2$. Điều này có nghĩa là khi chúng ta giữ nguyên yếu tố X_3 thì giá trị trung bình của biến phụ thuộc Y sẽ thay đổi (tăng hoặc giảm tuỳ thuộc vào dấu của β_2) β_2 đơn vị cho mỗi đơn vị tăng của yếu tố X_2.

$\frac{\partial E}{\partial X_{3i}} = \beta_3$ điều này có nghĩa là giá trị trung bình của biến Y tăng (hoặc giảm) β_3 đơn vị cho mỗi đơn vị tăng của X_3. Như vậy các hệ số hồi qui riêng (hệ số góc) phản ánh ảnh hưởng của một biến giải thích đối với giá trị trung bình của biến phụ thuộc khi giá trị của biến giải thích khác chứa trong mô hình không đổi.

3.1.2. Ước lượng các tham số của mô hình

Để ước lượng các tham số của mô hình:

$E(Y / X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$

Chúng ta sử dụng phương pháp bình phương tối thiểu OLS, tư tưởng chính của phương pháp này đã được trình bày ở chương 2.

Giả sử chúng ta có n quan sát, quan sát thứ i có 3 giá trị ứng với Y, X_2 và X_3, kí hiệu (Y_i, X_{2i}, X_{3i}).

Hàm hồi quy mẫu SRF được xây dựng từ n quan sát này có dạng:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$$

Trong đó $\hat{\beta}_i$: i=1, 2, 3 là ước lượng tương ứng của β_i : i = 1, 2, 3.

Khi đó $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i$; e_i là phần dư ứng với quan sát thứ i.

$$e_i = Y_i - \hat{Y}_i = Y_i - \beta_1 - \beta_2 X_{2i} - \beta_3 X_{3i}$$

Phương pháp OLS tính giá trị của các tham số $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ sao cho:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 X_{2i} - \beta_3 X_{3i})^2 \Rightarrow \min$$

Các tham số $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ được tính từ hệ phương trình chuẩn sau đây:

$$\hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3 = \bar{Y}$$

$$\hat{\beta}_1 \sum_{i=1}^{n} X_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 + \hat{\beta}_3 \sum_{i=1}^{n} X_{2i} \sum_{i=1}^{n} X_{3i} = \sum_{i=1}^{n} Y_i X_{2i}$$

$$\hat{\beta}_1 \sum_{i=1}^{n} X_{3i} + \hat{\beta}_3 \sum_{i=1}^{n} X_{3i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} \sum_{i=1}^{n} X_{3i} = \sum_{i=1}^{n} Y_i X_{3i}$$

Trong đó: $\bar{X}_2 = \sum_{i=1}^{n} X_{2i} / n; \bar{X}_3 = \sum_{i=1}^{n} X_{3i} / n; \bar{Y} = \sum_{i=1}^{n} Y_i / n$

Đặt: $y_i = Y_i - \bar{Y}; x_{2i} = X_{2i} - \bar{X}_2; x_{3i} = X_{3i} - \bar{X}_3$

Giải hệ phương trình ta được:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} y_i x_{2i}\right)\left(\sum_{i=1}^{n} x_{3i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{3i}\right)\left(\sum_{i=1}^{n} x_{2i} x_{3i}\right)}{\left(\sum_{i=1}^{n} x_{2i}^2\right)\left(\sum_{i=1}^{n} x_{3i}^2\right) - \left(\sum_{i=1}^{n} x_{2i} x_{3i}\right)^2}$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} y_i x_{3i}\right)\left(\sum_{i=1}^{n} x_{2i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{2i}\right)\left(\sum_{i=1}^{n} x_{2i} x_{3i}\right)}{\left(\sum_{i=1}^{n} x_{2i}^2\right)\left(\sum_{i=1}^{n} x_{3i}^2\right) - \left(\sum_{i=1}^{n} x_{2i} x_{3i}\right)^2}$$

$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ được gọi là các ước lượng bình phương nhỏ nhất.

3.1.3. Phương sai và độ lệch chuẩn của các ước lượng bình phương nhỏ nhất

Phương sai và độ lệch chuẩn của các ước lượng bình phương nhỏ nhất được cho bởi các công thức sau đây:

$$\text{var}(\hat{\beta}_2) = \frac{\sum_{i=1}^{n} x_{3i}^2}{\left(\sum_{i=1}^{n} x_{2i}^2\right)\left(\sum_{i=1}^{n} x_{3i}^2\right) - \left(\sum_{i=1}^{n} x_{2i}x_{3i}\right)^2}\sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_{2i}^2(1 - r_{23}^2)}$$

Trong đó r_{23} là hệ số tương quan mẫu giữa biến X_2 và X_3.

$$se(\hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_2)}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sum_{i=1}^{n} x_{2i}^2}{\left(\sum_{i=1}^{n} x_{2i}^2\right)\sum_{i=1}^{n} x_{3i}^2 - \left(\sum_{i=1}^{n} x_{2i}x_{3i}\right)^2}\sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_{3i}^2(1 - r_{23}^2)}$$

$$se(\hat{\beta}_3) = \sqrt{\text{var}\left(\hat{\beta}_3\right)}$$

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}^2\sigma^2}{(1 - r_{23}^2)\sqrt{\sum_{i=1}^{n} x_{2i}^2}\sqrt{\sum_{i=1}^{n} x_{3i}^2}}$$

Trong đó r_{23} là hệ số tương quan giữa biến X_2 và X_3.

$$r_{23}^2 = \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}$$

Trong các công thức trên σ^2 là phương sai của U_i nhưng chưa biết. Ước lượng không chệch của σ^2 là: $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-3} = \frac{RSS}{n-3}$

3 là số tham số của mô hình, trong trường hợp tổng quát nếu mô hình có k tham số $\beta_1, \beta_2, ..., \beta_k$ thì $\hat{\sigma}^2 = \sum e_i^2 \big/ (n-k)$

3.2. MÔ HÌNH HỒI QUY K BIẾN

3.2.1. Mô hình

Phần này giới thiệu mô hình hồi quy bội k biến bằng ngôn ngữ ma trận. Với ngôn ngữ ma trận kết hợp với kỹ thuật tính toán cho phép chúng ta giải quyết các vấn đề của phân tích hồi quy một cách nhanh chóng, chính xác.

Hàm hồi quy tổng thể có dạng:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + U_i$$

Trong đó β_1 là hệ số tự do (hệ số chặn)

β_j : j = 2..k là các hệ số hồi quy riêng.

Giả sử chúng ta có n quan sát, mỗi quan sát gồm k giá trị (Y_i, X_{2i},..., X_{ki})

$Y_1 = \beta_1 + \beta_2 X_{21} + ... + \beta_k X_{k1} + U_1$

$Y_2 = \beta_1 + \beta_2 X_{22} + ... + \beta_k X_{k2} + U_2$

$Y_n = \beta_1 + \beta_2 X_{2n} + ... + \beta_k X_{kn} + U_n$

Ký hiệu: $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}; U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix};\quad X = \begin{bmatrix} 1 & X_{21} & X_3 & \dots & X_4 \\ 1 & X_{21} & X_3 & \cdots & X_4 \\ \vdots & X_{21} & X_3 & \cdots & X_4 \\ 1 & X_{21} & X_3 & \cdots & X_4 \end{bmatrix}$

Khi đó ta có: $Y = X\beta + U$

Giả thiết 4 nói rằng giữa các biến độc lập không có quan hệ tuyến tính với nhau, khi đó các cột của ma trận X là độc lập tuyến tính. Do đó hạng của ma trận X bằng số cột của ma trận này tức là R(X) = k, ma trận X không suy biến.

3.2.2. Ước lượng các tham số

Hàm hồi qui mẫu SRF có dạng:

$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}$

$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki} + e_i$

hay $Y = X\hat{\beta} + e$

Trong đó $e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = Y - X\hat{\beta}$

Các ước lượng OLS được tìm bằng cách:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_k X_{ki})^2 \Rightarrow \min$$

$\sum_{i=1}^{n} e_i^2$ là tổng bình phương của các phần dư (RSS)

$$e'e = \sum_{i=1}^{n} e_i^2 = \left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right) = Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$\frac{\partial\left(e'e\right)}{\partial\hat{\beta}} = -2X^{'}Y + 2X^{'}X\hat{\beta} \Rightarrow X^{'}Y = X^{'}X\hat{\beta}$$

$$\begin{bmatrix} n & \sum X_{2i} & \sum X_{3i} & \cdots & \sum X_{ki} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i}X_{3i} & & \sum X_{2i}X_{ki} \\ \cdots & & & & \\ \sum X_{ki} & \sum X_{ki}X_{2i} & \sum X_{ki}X_{3i} & & \sum X_{ki}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \cdots & & & \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Với giả thiết 4, X không suy biến, nên X'X cũng không suy biến, do đó tồn tại $(X'X)^{-1}$. Từ đó: $\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$

3.2.3. Ma trận hiệp phương sai

Để kiểm định giả thiết, tìm khoảng tin cậy, cũng như thực hiện các suy luận thống kê khác cần phải tìm $\text{var}(\hat{\beta}_i)$ i = 1..k và $Cov(\hat{\beta}_i, \hat{\beta}_j)$. Phương pháp ma trận cho phép chúng ta tìm chúng một cách dễ dàng.

Ma trận hiệp phương sai của:

$$Cov(\hat{\beta}) = \begin{bmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) & \cdots & Cov(\hat{\beta}_1, \hat{\beta}_k) \\ Cov(\hat{\beta}_1, \hat{\beta}_2) & Var(\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & & & \\ Cov(\hat{\beta}_k, \hat{\beta}_1) & Cov(\hat{\beta}_k, \hat{\beta}_2) & \cdots & Var(\hat{\beta}_k) \end{bmatrix}$$

$Cov(\hat{\beta})$ được xác định như thế nào?

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$$

$$Y = X\beta + U$$

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}\left(X\beta + U\right) = \beta + (X^{'}X)^{-1}X^{'}U$$

$$\hat{\beta} - \beta = (X^{'}X)^{-1}X^{'}U$$

$$Cov(\hat{\beta}) = E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)^{'}\right] = E\left\{\left[(X^{'}X)^{-1}X^{'}U\right]\left[(X^{'}X)^{-1}X^{'}U\right]^{'}\right\}$$

$$= E\left[(X^{'}X)^{-1}X^{'}UU^{'}X(X^{'}X)^{-1}\right] = (X^{'}X)^{-1}XE(UU^{'})X(X^{'}X)^{-1}$$

$$= (X^{'}X)^{-1}X\sigma^2X(X^{'}X)^{-1}$$

$$Cov(\hat{\beta}) = \sigma^2(X^{'}X)^{-1}$$

Trong công thức trên $(X'X)^{-1}$ là ma trận nghịch đảo của ma trận (X'X), σ^2 là $Var(U_i)$, nhưng chưa biết nên chúng ta phải dùng ước lượng không chệch của σ^2 là:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-k)$$

$$e'e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i^{'}\right)^2 = Y^{'}Y - 2\hat{Y}^{'}Y + \hat{Y}^{'}Y$$

$$= Y^{'}Y - 2\hat{\beta}^{'}X^{'}Y + \hat{\beta}^{'}X^{'}X\hat{\beta} = Y^{'}Y - \hat{\beta}^{'}X^{'}Y$$

3.2.4. Hệ số xác định bội

Trong mô hình hồi quy hai biến, r^2 đo độc thích hợp của hàm hồi quy. Nó chính là tỷ lệ của toàn bộ sự biến đổi của biến phụ thuộc Y do biến giải thích X gây ra. Trong mô hình hồi quy bội tỷ lệ của toàn bộ sự khác biệt của biến Y do tất cả các biến giải thích $X_2, X_3, \ldots, X_k$ gây ra được gọi là hệ số xác định bội, ký hiệu R^2.

Hệ số xác định bội R^2 có thể tính bằng một trong hai công thức sau:

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Phần trước đã chứng minh:

$$e'e = \sum_{i=1}^{n} e_i^2 = Y^{'}Y - \hat{\beta}^{'}X^{'}Y$$

$$= \sum_{i=1}^{n} e_i \sum e_i\left(Y_i - \hat{Y}_i\right) = \sum_{i=1}^{n} e_i(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \ldots - \hat{\beta}_k X_{ki})$$

$$= \sum_{i=1}^{n} e_i(Y_i - \overline{Y} - \hat{\beta}_2(X_{2i} - \overline{X}) - \ldots - \hat{\beta}_k(X_{ki} - \overline{X}_{ki})$$

$$= \sum_{i=1}^{n} e_i(y_i - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_k x_{ki}) = \sum_{i=1}^{n} e_i y_i - 0$$

$$= \sum_{i=1}^{n} y_i(y_i - \hat{\beta}_2 X_{2i} - \ldots - \hat{\beta}_k X_{ki})$$

$$= \sum_{i=1}^{n} y_i y_i - \hat{\beta}_2 \sum_{i=1}^{n} y_i x_{2i} - \ldots - \hat{\beta}_k \sum_{i=1}^{n} y_i x_{ki}$$

$$TSS = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} Y_i^2 - 2\sum_{i=1}^{n} Y_i\overline{Y} + n\overline{Y}^2 = Y^{'}Y - n\overline{Y}^2$$

$$ESS = TSS - e'e = \hat{\beta}^{'}X^{'}Y - n\overline{Y}$$

$$R^2 = \frac{\hat{\beta}^{'}X^{'}Y - n\overline{Y}^2}{Y^{'}Y - n\overline{Y}^2}$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\hat{\beta}_2 \sum_{i=1}^{n} y_i x_{2i} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} y_i x_{ki}}{\sum_{i=1}^{n} y_i^2}$$

$0 \le R^2 \le 1$. Nếu $R^2 = 1$, có nghĩa là đường hồi quy giải thích 100% sự thay đổi của Y. Nếu $R^2 = 0$, có nghĩa là mô hình không giải thích sự thay đổi nào của Y.

Một tính chất quan trọng của R^2 là nó là hàm không giảm của số biến giải thích có trong mô hình. Dễ thấy rằng $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$ không phụ thuộc vào số biến giải thích trong mô hình, nhưng $\sum_{i=1}^{n} e_i^2$ là hàm giảm của số này. Do đó, nếu tăng số biến giải thích trong mô hình thì R^2 cũng tăng. Vấn đề đặt ra là khi nào thì đưa thêm biến giải thích mới vào mô hình?

Không thể dùng R^2 làm tiêu chuẩn để xem xét việc đưa thêm hay không đưa thêm một biến giải thích mới vào mô hình. Bởi vì R^2 còn phụ thuộc vào số bậc tự do của $\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ và $\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$ tương ứng là (n – k) và (n – 1). Trong đó k là số các tham số (kể cả hệ số chặn) của mô hình.

Người ta dùng hệ số xác định bội đã hiệu chỉnh, ký hiệu $\bar{R}^2$ để cân nhắc khi xem xét việc thêm biến giải thích mới vào mô hình.

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2 / (n-k)}{\sum_{i=1}^{n} y_i^2 / (n-1)} = 1 - \frac{\hat{\sigma}^2}{S_y^2} = 1 - (1 - R^2)\frac{n-1}{n-k}$$

$\bar{R}^2$ có các tính chất sau:

- Nếu k > 1, $\bar{R}^2 \le R^2 \le 1$, điều này có nghĩa là nếu số biến giải thích tăng lên thì $\bar{R}^2$ tăng chậm hơn so với R^2.

- $R^2 \ge 0$, nhưng $\bar{R}^2$ có thể âm. Như vậy khi $\bar{R}^2$ còn tăng thì ta còn phải đưa thêm biến mới. $\bar{R}^2$ còn có thể tăng khi mà hệ số của biến mới trong hàm hồi quy khác không. Khi nào biết được hệ số của biến mới trong hàm hồi quy khác không? Khi mà giả thiết:

$H_0 : \beta_k = 0$; $H_1 : \beta_k \ne 0$.

bị bác bỏ, trong đó X_k là biến chúng ta định đưa thêm vào mô hình.

Giả sử chúng ta có mô hình hồi quy bội:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + U_i$$

Kí hiệu R_{tj} là hệ số tương quan giữa biến thứ t và thứ j. Nếu t = 1 thì r_{tj} là hệ số tương quan giữa các biến Y và biến X_j.

$$r_{1j}^2 = \frac{\left(\sum_{i=1}^{n} y_i x_{ij}\right)^2}{\sum_{i=1}^{n} y_i^2 \sum_{i=1}^{n} x_{ij}^2}; r_{tj}^2 = \frac{\left(\sum_{i=1}^{n} x_{ti} x_{ji}\right)^2}{\sum_{i=1}^{n} x_{ti}^2 \sum_{i=1}^{n} x_{ji}^2}$$

Trong đó: $x_{ji} = X_{ji} - \bar{X}_j$

Dễ dàng thấy rằng: $r_{tj} = r_{jt}$; $r_{jj} = 1$.

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{1k} \\ r_{21} & r_{22} & r_{23} & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & r_{k3} & r_{kk} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{1k} \\ r_{21} & 1 & r_{23} & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & r_{k3} & 1 \end{bmatrix}$$

3.2.5. Hệ số tương quan riêng

Chúng ta đã biết hệ số tương quan r đo mức độ phụ thuộc tuyến tính giữa hai biến. Đối với mô hình hồi quy 3 biến:

$$Y_i = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + U_i$$

Chúng ta định nghĩa $r_{12,3}$ là hệ số tương quan giữa biến Y và X_2 trong khi X_3 không đổi.

$r_{13,2}$ là hệ số tương quan riêng giữa biến Y và X_3 trong khi X_2 không đổi.

$r_{23,1}$ là hệ số tương quan riêng giữa biến X_2 và X_3 trong khi Y không đổi.

Ta có thể dễ dàng chỉ ra rằng:

$$r_{12,3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$$

$$r_{13,2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$$

$$r_{23,1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{13}^2\right)}}$$

Hệ số tương quan riêng đã được định nghĩa như trên được gọi là hệ số tương quan bậc nhất. Từ "bậc" ở đây ngụ ý chỉ số hạng sau dấu phẩy vì thế $r_{12,34}$ là hệ số tương quan riêng bậc 2; còn r_{12}, r_{13} là các hệ số tương quan bậc không.

Giữa hệ số xác định bội và các hệ số tương quan bậc không và hệ số tương quan bậc nhất có các mối liên hệ sau:

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$R^2 = r_{12}^2 + (1 - r_{12}^2)r_{13,2}^2$$

$$R^2 = r_{13}^2 + (1 - r_{13}^2)r_{12,3}^2$$

Ma trận R nói ở trên được gọi là ma trận hệ số tương quan riêng cấp 0.

3.2.6. Kiểm định giả thiết

Với giả thiết $U \sim N(0, \sigma^2)$ ta có thể kiểm định giả thiết, tìm khoảng tin cậy cho các hệ số hồi quy riêng.

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

Thành phần $\hat{\beta}_i$ có phân bổ chuẩn với kỳ vọng β_i và phương sai bằng σ^2 nhân với phần tử nằm trên dòng thứ i và cột i của ma trận $(X'X)^{-1}$ hay chính là phần tử thứ i trên đường chéo chính của ma trận $Cov(\hat{\beta})$. Tuy nhiên do σ^2 chưa biết, nên ta phải dùng ước lượng không chệch của σ^2 là: $\hat{\sigma}^2 = \sum_{i=1}^{n} e_i^2 / (n-k)$

Khi đó $t = \dfrac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)}$ có phân bố t(n-k). Với tiêu chuẩn này có thể tìm khoảng tin cậy, kiểm định giả thiết về các hệ số hồi quy riêng.

Khoảng tin cậy với hệ số tin cậy 1 - α của β_i được xác định:

$$P(-t_{\alpha/2}(n-k) < \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} < t_{\alpha/2}(n-k)) = 1 - \alpha \text{; do đó:}$$

$$(\hat{\beta}_i - t_{\alpha/2}(n-k)se(\hat{\beta}_i) < \beta_i < \hat{\beta}_i + t_{\alpha/2}(n-k)se(\hat{\beta}_i))$$

Chúng ta có thể kiểm định giả thiết $\beta_i = \beta_i^*$

Tiêu chuẩn dùng để kiểm định: $t = \dfrac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \sim t(n-k)$

Tuỳ theo các giả thiết H_1, chúng ta có các miền bác bỏ sau đây:

Loại giả thiết	**H_0**	**H_1**	**Miền bác bỏ**
Hai phía	$\beta_i = \beta_i^*$	$\beta_i \neq \beta_i^*$	$\lvert t \rvert > t_{\alpha/2}(n-k)$
Bên trái	$\beta_i = (\geq)\beta_i^*$	$\beta_i < \beta_i^*$	$t < -t_{\alpha}(n-k)$
Bên phải	$\beta_i = (\leq)\beta_i^*$	$\beta_i > \beta_i^*$	$t > t_{\alpha}(n-k)$

Nếu $\beta_i^* = 0$, chúng ta muốn kiểm định biến độc lập X_i không ảnh hưởng đến biến phụ thuộc.

Chúng ta kiểm định giả thiết:

$\beta_2 = \beta_3 = ... = \beta_k = 0$ hay $R^2 = 0$.

Chúng ta đã trình bày kỹ thuật phân tích phương sai và mối quan hệ giữa R^2 và F. Bằng ngôn ngữ ma trận có thể trình bày tổng quát các vấn đề đó.

Bảng 3.2. Phân tích phương sai cho mô hình hồi quy bội k biến.

Nguồn biến thiên	Tổng bình phương	Bậc tự do	Phương sai
Từ hàm hồi quy (ESS)	$\hat{\beta}' X' Y - n\overline{Y}^2$	k - 1	$\dfrac{\hat{\beta}' X' Y - n\overline{Y}^2}{k-1}$
Phần dư (RSS)	$Y' Y - \hat{\beta}' X' Y$	n – k	$\dfrac{Y' Y - \hat{\beta}' X' Y}{n-k}$
Tổng	$Y' Y - n\overline{Y}^2$	n – 1	

Do $R^2 = \dfrac{\hat{\beta}' X' Y - n\overline{Y}^2}{Y' Y - n\overline{Y}^2}$

Nên $\hat{\beta}' X' Y - n\overline{Y}^2 = R^2\left(Y' Y - n\overline{Y}^2\right)$

$$Y' Y - \hat{\beta}' X' Y = (1-R^2)\left(Y' Y - n\overline{Y}^2\right)$$

Ta có bảng sau đây:

Bảng 3.3. Phân tích phương sai đối với R^2

Nguồn biến thiên	Tổng bình phương	Bậc tự do	Phương sai
Từ hàm hồi quy (ESS)	$R^2\left(Y' Y - n\overline{Y}^2\right)$	k – 1	$R^2\left(Y' Y - n\overline{Y}^2\right)/(k-1)$
Phần dư (RSS)	$(1-R^2)\left(Y' Y - n\overline{Y}^2\right)$	n – k	$(1-R^2)\left(Y' Y - n\overline{Y}^2\right)/(n-k)$
Tổng	$Y' Y - n\overline{Y}^2$	n - 1	

Với giả thiết $\hat{\beta} \sim N(\beta, \sigma^2)$ thì giả thiết:

H_0: $\beta_2 = \beta_3 = ... = \beta_k = 0$ (hay $R^2 = 0$)

H_1: có ít nhất một $\beta_i \neq 0$ (hay $R^2 > 0$)

được kiểm định bằng tiêu chuẩn:

$$F = \frac{ESS/(k-10}{RSS/(n-k)} \sim F((k-1),(n-k))$$

$$F = \frac{R^2(n-k)}{(1-R^2)(k-1)}$$

Hồi quy có điều kiện ràng buộc – kiểm định F:

Giả sử rằng chúng ta có hàm hồi quy:

$$Y = \beta_1 + \beta_2 X_{2i} + ... + \beta_{ki} X_{ki} + U_i$$

Bây giờ chúng ta kiểm định giả thiết:

H_0: $\beta_{k-m+1} = \beta_{k-m+2} = ... = \beta_k = 0$

Với giả thiết này thì hàm hồi quy có dạng:

$$Y = \beta_1 + \beta_2 X_{2i} + ... + \beta_{k-m} X_{k-m} + U_i$$

(được gọi là hàm hồi quy thu hẹp hay hàm hồi quy có điều kiện ràng buộc).

Ký hiệu e_R; Véc tơ phần dư từ hàm hồi quy có điều kiện ràng buộc.

e_{UR}: Véc tơ phần dư từ hàm hồi quy ban đầu (không điều kiện ràng buộc).

m: Số biến bị loại khỏi mô hình ban đầu (số điều kiện ràng buộc)

n: số quan sát.

Khi đó tiêu chuẩn kiểm định giả thiết H_0.

$$\frac{(e'_R e_R - e'_{UR} e_{UR})/m}{e'_{UR} e_{UR}/(n-k)} \sim F(m,(n-k))$$

Nếu $F > F_\alpha(m,n-k)$ thì giả thiết H_0 bị bác bỏ.

Đôi khi ta dùng công thức sau đây:

$$\frac{(e'_R e_R - e'_{UR} e_{UR})/m}{e'_{UR} e_{UR}/(n-k)} = \frac{(TSS - ESS_R) - (TSS - ESS_{UR})/m}{(TSS - ESS_{UR})/(n-k)}$$

$$= \frac{(ESS_{UR} - ESS_R)/m}{(TSS - ESS_{UR})/(n-k)}$$

Chia cả tử số và mẫu cho TSS, ta được công thức rút gọn:

$$F = \frac{(R^2_{UR} - R^2_R)/m}{(1-R^2_{UR})/(n-k)} \sim F(m,(n-k))$$

Cách trình bày như trên chỉ là trường hợp riêng của kiểm định "tổ hợp tuyến tính của các hệ số hồi quy". Với giả thiết về tổ hợp tuyến tính của các hệ số hồi quy, có thể làm cho biến phụ thuộc không còn là biến phụ thuộc ban đầu. Khi biến phụ thuộc thay đổi thì công thức tính F qua các R^2 sẽ không dùng được. Trong trường hợp này ta dùng công thức ban đầu, tính F qua RSS.

3.2.7. Dự báo

Chúng ta có thể sử dụng mô hình hồi quy vào dự báo: dự báo giá trị trung bình và giá trị cá biệt.

Cho $X^0 = \begin{bmatrix} 1 \\ X_2^0 \\ X_3^0 \\ \vdots \\ X_k^0 \end{bmatrix}$

Dự báo giá trị trung bình $E(Y/X^0)$

$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k = X^{'}\hat{\beta}$

Với $X = X^0$ ta có $(\hat{Y}_0 / X^0) = X^{0'}\hat{\beta} \Rightarrow \text{var}(\hat{Y}_0 / X_0) = X^{0'}\text{var}(\hat{\beta})X^0$

$\text{var}(\hat{Y}_0 / X_i^0) = \sigma^2 X^{0'}(X^{'}X)^{-1}X^0$ vì $\text{var}(\hat{\beta}) = \sigma^2 (X^{'}X)^{-1}$.

Nhưng σ^2 chưa biết nên phải dùng ước lượng không chệch $\hat{\sigma}^2$ của nó.

$Var(\hat{Y} / X^0) = \hat{\sigma}^2 X^{0'}(X^{'}X)^{-1}X^0$

$se(\hat{Y} / X^0) = \sqrt{\hat{\sigma}^2 X^{0'}(X^{'}X)^{-1}X^0}$

$\hat{Y}_0 - t_{\alpha/2}(n-k)se(\hat{Y}_0 / X^0) \le E(Y / X_0) \le \hat{Y}_0 + t_{\alpha/2}(n-k)se(\hat{Y}_0 / X^0)$

Dự báo giá trị cá biệt:

$Y_i = X\hat{\beta} + e_i \Rightarrow \text{var}(Y_0 / X^0) = \text{var}(X^{0'}\hat{\beta}) + \sigma^2$

$Var(Y_0 / X^0) = \hat{\sigma}^2 \left[1 + X^{0'}\left(X^{'}X\right)^{-1} X^0 \right]$

$se(Y_0 / X^0) = \sqrt{Var(Y_0 / X^0)}$

$\hat{Y}_0 - t_{\alpha/2}(n-k)se(\hat{Y}_0 / X^0) \le (Y / X_0) \le \hat{Y}_0 + t_{\alpha/2}(n-k)se(\hat{Y}_0 / X^0)$

Với việc trình bày mô hình hồi quy bằng ngôn ngữ ma trận đã cung cấp cho chúng ta một công nghệ mà nhờ đó có thể sử dụng kỹ thuật tính toán, tự động hoá toàn bộ quá trình tính toán, phân tích và dự báo.

Câu hỏi chương 3.

1. Thế nào là mô hình hồi quy ba biến? Mô tả ý nghĩa hình học của mô hình hồi quy như thế nào?
2. Dùng công cụ ma trận mô tả các biến, các phương trình trong mô hình hồi quy tuyến tính k biến như thế nào?
3. Phân tích về hệ số hồi quy với mô hình tổng quát có những điểm già khác so với mô hình hồi quy hai biến.

4. Ta có các biến Y (Lượng cam bán được-tấn/tháng), X1 (Giá cam-nghàn đồng/kg), X2 (giá quýt-nghàn đồng/kg) với số liệu trong bảng sau:

Y	14	13	12	10	8	9	8	7	6	6
X1	2	2	3	4	5	5	6	7	8	9
X2	7	6	7	6	5	6	4	5	4	5

Dùng phương pháp ma trận để trả lời các câu hỏi sau:

a) Xác định hàm SRF của mô hình?

b) Nêu ý nghĩa của mô hình hồi quy riêng?

c) Xác định khoảng tin cậy với độ tin cậy 90%.

d) Biến giá quýt có ảnh hưởng đến lượng cam bán ra không?

e) Kiểm định giả thiết nếu giá cam tăng lên 1000đ/kg, lượng cam bán ra giảm đi 01 tấn/tháng?

5. Cho kết quả ước lượng của Y, X2, X3 như sau:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,931479147
R Square	0,867653401
Adjusted R Square	0,841184081
Standard Error	3,460980395
Observations	13

ANOVA

	Df	*SS*	*MS*	*F*	*Significance F*
Regression	2	785,2930701	392,6465351	32,77959	4,06E-05
Residual	10	119,783853	11,9783853		
Total	12	905,0769231			

	Coefficients	*Standard Error*	*T Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7,140388394	9,735596246	-0,73343103	0,480136	-28,8326	14,55187176
X Variable 1	0,145255037	0,024077944	6,032700957	0,000126	0,091606	0,19890404
X Variable 2	0,100329287	0,102597224	0,977894754	0,35119	-0,12827	0,328930146

Yêu cầu:

a) Xác định hàm hồi quy mẫu.

b) Ý nghĩa kinh tế của các hệ số nhận được.

c) Kiểm định sự phù hợp của hàm hồi quy.

d) Kiểm định ảnh hưởng của X2, X3 đến Y

e) Có nhận định rằng: khi X2 lên 1 đơn vị, thì Y tăng lên 0,15 đơn vị. Bạn có tin điều đó không, với mức ý nghĩa 5%.

CHƯƠNG 4. HỒI QUY VỚI BIẾN GIẢ

Chương này nhằm tới việc vận dụng biến giả để lượng hóa các biến định tính. Ngoài ra cũng trình bày các khía cạnh ứng dụng khác của biến giả như phân tích tác động thời vụ, kiểm định tính ổn định cấu trúc của mô hình hồi quy.

Trong các mô hình hồi quy tuyến tính mà chúng ta đã xem xét từ các chương trước cho đến nay thì các biến giải thích đều là các biến số lượng. Các biến đó có thể nhận giá trị bằng số. Chẳng hạn, tiền lương của cán bộ, doanh số bán ra của một cửa hàng, chi tiêu cho quảng cáo, cung tiền, ... là những biến số lượng. Nhưng trong thực tế có nhiều trường hợp các biến giải thích (hoặc thậm chí cả biến phụ thuộc) là biến chất lượng. Trong chương này ta sẽ nghiên cứu hồi quy khi biến giải thích là biến chất lượng.

Nội dung cơ bản của chương này bao gồm:

O Bản chất của biến giả

O Mô hình hồi quy có một biến lượng và một biến chất

O Mô hình hồi quy có một biến lượng và hai biến chất

O Kết hợp hai hồi quy

O Ảnh hưởng tương tác giữa các biến giả

O Sử dụng biến giả trong phân tích mùa

4.1. BẢN CHẤT CỦA BIẾN GIẢ - MÔ HÌNH CÓ BIẾN GIẢ

Biến chất lượng như đã nói ở trên thường chỉ ra có hoặc không có một thuộc tính nào đó, chẳng hạn nam hay nữ; khu vực tư nhân hay nhà nước... vấn đề đặt ra là làm thế nào để lượng hoá được những thuộc tính này. Trong phân tích hồi quy người ta sử dụng kỹ thuật gọi là kỹ thuật biến giả. Kỹ thuật này cho phép ta lượng hoá được những thuộc tính như vậy. Chẳng hạn để giải thích cho việc một số thanh niên vào trường đại học, một số khác thì không, chúng ta tạo ra biến giả mà nhận giá trị là 1 nếu thanh niên vào đại học và nhận giá trị là 0 nếu thanh niên đó không vào đại học. Chúng ta cũng sẽ chỉ ra biến giả có thể được sử dụng như thế nào trong phạm vi hồi quy để giải thích cho sự kiện là có những quan sát trong phạm trù (thuộc tính) đã cho gắn với một tập các tham số hồi quy còn các quan sát khác trong phạm trù thứ 2 (thứ 3) lại gắn với những tham số hồi quy khác. Biến giả được sử dụng trong mô hình hồi quy giống như biến số lượng thông thường.

Giả sử một công ty sử dụng 2 quá trình sản xuất (kí hiệu quá trình sản xuất A và quá trình sản xuất B) để sản xuất ra một loại sản phẩm. Giả sử sản phẩm thu được từ mỗi một quá trình sản xuất là đại lượng ngẫu nhiên có phân phối chuẩn và có kỳ vọng khác nhau nhưng phương sai như nhau. Chúng ta có thể biểu thị quá trình sản xuất đó như một phương trình hồi quy.

$$Y_i = \beta_1 + \beta_2 D_i + U_i \qquad (4.1)$$

trong đó Y_i là sản lượng sản phẩm gắn với quá trình thứ i.

D_i Là biến giả nhận 1 trong 2 giá trị:

$$D_i \begin{cases} = 1 \text{ Nếu sản lượng sản phẩm thu được từ quá trình sản xuất A} \\ = 0 \text{ Nếu sản lượng thu được từ quá trình sản xuất B.} \end{cases}$$

Mô hình hồi quy trên đây giống như mô hình hồi quy 2 biến mà chúng ta gặp trước đây chỉ khác là biến số lượng X được thay bằng biến giả D. Căn cứ vào mô hình này chúng ta có thể biết được sản lượng trung bình do quá trình sản xuất A có khác với sản lượng trung bình do quá trình sản xuất B tạo ra hay không?

Hệ số chặn β_1 của hồi quy tuyến tính do sản lượng trung bình gắn với quá trình sản xuất B, trong khi đó độ dốc β_2 của đường hồi quy do sự khác nhau về sản lượng sinh ra do việc thay đổi từ quá trình sản xuất B đến quá trình sản xuất A.

Điều này có thể thấy bằng 2 cách lấy giá trị kỳ vọng cả 2 vế của phương trình (4.1) ứng với $D_i = 0$ và $D_i = 1$:

$$E(Y_i / D_i = 0) = \beta_1$$

$$E(Y_i / D_i = 1) = \beta_1 + \beta_2$$

Kiểm định giả thiết H_0: $\beta_2 = 0$ cung cấp kiểm định về giả thiết là không có sự khác nhau về sản lượng do quá trình sản xuất A và B tạo ra.

Thủ tục biến giả có thể dễ dàng mở rộng cho trường hợp có nhiều hơn 2 phạm trù. Chẳng hạn trong thí dụ ở trên ta giả thiết có 3 quá trình sản xuất khác nhau có thể sử dụng để sản xuất ra sản phẩm và người ta hy vọng giải thích cho vấn đề là sản lượng được sản xuất ra cho mỗi quá trình có thể không như nhau. Trong trường hợp này ta sẽ đưa vào 2 biến giả là D_1 và D_2. Chúng ta sẽ xét mô hình:

$$Y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} U_i \qquad (4.2)$$

Trong đó:

$$D_1 = \begin{cases} 1 \text{ Nếu sản lượng sản phẩm thu được từ quá trình sản xuất A} \\ 0 \text{ Nếu sản lượng sản phẩm thu được từ quá trình khác.} \end{cases}$$

$$D_2 = \begin{cases} 1 \text{ Nếu sản lượng sản phẩm thu được từ quá trình sản xuất B.} \\ 0 \text{ Nếu sản lượng sản phẩm thu được từ quá trình khác.} \end{cases}$$

Như vậy 3 quá trình sản xuất này được biểu thị dưới dạng các kết hợp sau của các giá trị của các biến giả:

Quá trình sản xuất	D_1	D_2
A	1	0
B	0	1
C	0	0

Bằng việc lấy kỳ vọng cho mỗi một trong 3 trường hợp này chúng ta có thể giải thích kết quả hồi quy:

$$E(Y_i / D_i = 1; D_2 = 0) = \beta_1 + \beta_2$$

$$E(Y_i / D_1 = 0; D_2 = 1) = \beta_1 + \beta_3$$

$$E(Y_i / D_1 = 0; D_2 = 0) = \beta_1$$

Hệ số chặn của hồi quy biểu thị giá trị kỳ vọng của sản lượng do quá trình sản xuất C tạo ra. Hệ số góc thứ nhất do sự thay đổi trung bình về sản lượng do việc chuyển từ quá trình sản xuất C sang quá trình sản xuất A và hệ số góc thứ 2 tức là β_3 đo thay đổi trung bình về sản lượng khi thay đổi từ quá trình sản xuất C sang quá trình sản xuất B.

Kiểm định giả thiết H_0: $\beta_2 = 0$ có nghĩa là không có sự khác nhau giữa quá trình sản xuất A và quá trình sản xuất C. Giả thiết H_0: $\beta_3 = 0$ cũng có ý nghĩa tương tự nhưng lại so sánh 2 quá trình sản xuất B và C.

Chú ý:

1. Để phân biệt 2 phạm trù nam hoặc nữ hay quá trình sản xuất A hoặc B người ta dùng một biến giả. Để phân biệt 3 phạm trù người ta dùng 2 biến giả. Một cách tổng quát để phân biệt N phạm trù người ta dùng N – 1 biến giả. Số biến giả thấp hơn số phạm trù là 1 để tránh tính đa cộng tuyến hoàn hảo. Để phân biệt 3 quá trình sản xuất A, B và C ta chỉ sử dụng 2 biến giả D_1 và D_2.

2. Phạm trù được gán giá trị là phạm trù cơ sở. Phạm trù được gọi là cơ sở theo nghĩa việc so sánh được tiến hành với phạm trù này. Như vậy trong mô hình trên quá trình sản xuất C là phạm trù cơ sở, nghĩa là nếu ta ước lượng hồi quy (4.2) với $D_1 = 0$; $D_2 = 0$, thì chỉ có quá trình sản xuất C, hệ số chặn sẽ là $\hat{\beta}_1$.

3. Hệ số $\hat{\beta}_2$ gắn với biến giả D_1 được gọi là hệ số chặn chênh lệch, vì nó cho biến giá trị của số hạng chặn của phạm trù nhận giá trị bằng 1 sẽ khác bao nhiêu với hệ số chặn của phạm trù cơ sở.

4.2. MÔ HÌNH HỒI QUY MỘT BIẾN LƯỢNG VÀ MỘT BIẾN CHẤT

Trong mục này ta sẽ xét mô hình hồi quy chỉ có một biến lượng và một biến chất với số phạm trù nhiều hơn hoặc bằng 2. Trường hợp có nhiều biến lượng và một biến chất thì thủ tục cũng được xét tương tự như ta sẽ làm dưới đây chỉ khác là số biến lượng tăng lên. Để dễ theo dõi trong mục này ta chia ra làm 2 trường hợp: trường hợp 1 khi biến chất chỉ có hai phạm trù, trường hợp 2 khia biến chất có nhiều hơn hai phạm trù.

4.2.1. Trường hợp biến chất chỉ có 2 phạm trù

Trong trường hợp này, mô hình hồi quy sẽ đơn giản vì theo chú ý ở trên khi biến chất có 2 phạm trù thì chỉ cần đặt 1 biến giả là đủ. Thí dụ ta xét mô hình sau:

$$Y_i = \beta_1 + \beta_2 D_i + U_i \qquad (4.3)$$

Trong đó: Y_i: là tiền lương hàng tháng của một công nhân cơ khí i.

X_i: Bậc thợ của công nhân i.

$$D_i = \begin{cases} 1 \text{ Nếu công nhân i làm việc trong khu vực tư nhân.} \\ 0 \text{ Nếu công nhân i làm việc trong khu vực quốc doanh.} \end{cases}$$

Mô hình có một biến lượng đó là bậc thợ của người công nhân và một biến chất chỉ rõ công nhân đó làm việc thuộc khu vực nào. Nếu ta giả thiết $E(U_i) = 0$ thì (4.3) có thể cho ta thấy liệu tiền lương của người công nhân làm việc ở khu vực tư nhân có khác tiền lương của người công nhân làm việc ở khu vực nhà nước không nếu các điều kiện khác không thay đổi. Bằng cách lấy kỳ vọng cả 2 vế (4.3) ta được:

Tiền lương trung bình của người công nhân cơ khí làm việc trong khu vực nhà nước:

$$E(Y_i / X_i, D_i = 0) = \beta_1 + \beta_3 X_i \qquad (4.3.1)$$

Tiền lương trung bình của người công nhân cơ khí làm việc trong khu vực tư nhân:

$$E(Y_i / X_i, D_i = 1) = (\beta_1 + \beta_2) + \beta_3 X_i \qquad (4.3.2)$$

Mô hình này giả thiết rằng mức tiền lương trung bình của người công nhân ngành cơ khí làm việc ở khu vực tư nhân khác với mức tiền lương trung bình của công nhân cơ khí làm việc ở khu vực nhà nước nhưng tốc độ tăng lương trung bình theo bậc thì như nhau.

Nếu giả thiết về tốc độ đã nêu trên là có giá trị thì kiểm định giả thiết rằng 2 hồi quy (4.3.1) và (4.3.2) có cùng hệ số chặn có thể tiến hành dễ dàng bằng cách ước lượng hồi quy (4.3) và chú ý rằng ý nghĩa về mặt thống kê của $\hat{\beta}_2$ đã được ước lượng trên cơ sở của kiểm định t. Nếu t chỉ ra rằng $\hat{\beta}_2$ là có ý nghĩa về mặt thống kê thì chúng ta từ bỏ giả thiết H_0 là tiền lương của công nhân cơ khí ở 2 khu vực kinh tế là như nhau.

4.2.2. Trường hợp biến chất có nhiều hơn 2 phạm trù

Khi biến chất có nhiều hơn 2 phạm trù thì vấn đề cũng không phức tạp hơn nhiều bởi vì theo chú ý ở trên nếu số phạm trù là N thì ta đưa vào mô hình hồi quy N – 1 biến giả làm biến giải thích.

Thí dụ căn cứ vào số liệu chéo người ta muốn hồi quy thu nhập hàng năm của một cán bộ giảng dạy đại học đối với tuổi nghề giảng dạy và vùng mà anh ta giảng dạy. Vì biến vùng là biến chất, trên thực tế chúng ta có thể căn cứ vào 3 vùng khác nhau trong cả nước là Bắc, Trung, Nam. Như vậy trong trường hợp này, biến chất của ta có 3 phạm trù, theo chú ý ở trên ta sẽ đưa vào mô hình hồi quy 2 biến.

Giả sử rằng cả 3 hồi quy có cùng độ dốc nhưng khác nhau hệ số chặn, chúng ta có mô hình sau:

$$Y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 X_i + U_i \qquad (4.4)$$

Trong đó: Y_i là thu nhập hàng năm của một giảng viên đại học

X_i: Tuổi nghề của giảng viên.

$D_1 = \begin{cases} 1 & \text{Nếu giảng viên i thuộc một trường đại học ở miền Bắc} \\ 0 & \text{Nếu giảng viên thuộc một trường không phải ở miền Bắc.} \end{cases}$

$D_2 = \begin{cases} 1 & \text{Nếu giảng viên i thuộc một trường đại học miền Nam} \\ 0 & \text{Nếu giảng viên thuộc một trường không phải ở miền Nam.} \end{cases}$

Như vậy, ta coi giảng viên thuộc một trường đại học ở miền Trung là phạm trù cơ sở, hệ số chặn chênh lệch β_2 và β_3 cho chúng ta biết chặn của các phạm trù khác với chặn của phạm trù cơ sở bao nhiêu. Chúng ta có thể tính được nếu giả thiết $E(U_i) = 0$ thì từ (4.4) ta có:

Thu nhập trung bình của một cán bộ giảng dạy ở một trường đại học ở miền Trung:

$$E(Y_i / D_1 = 0, D_2 = 0, X_i) = \beta_1 + \beta_4 X_i \qquad (4.4.1)$$

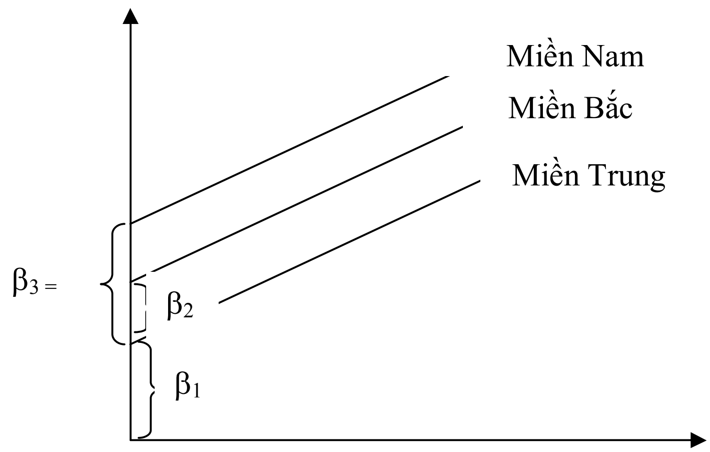
Thu nhập trung bình của một cán bộ giảng dạy ở một trường đại học miền Bắc:

$$E(Y_i / D_1 = 1, D_2 = 0, X_i) = (\beta_1 + \beta_2) + \beta_4 X_i \qquad (4.4.2)$$

Thu nhập trung bình của một cán bộ giảng dạy ở trường đại học miền Nam:

$$E(Y_i / D_1 = 0, D_2 = 1, X_i) = (\beta_1 + \beta_3) + \beta_4 X_i \qquad (4.4.3)$$

Giả sử $\beta_1 > 0$ ta có minh hoạ sau:

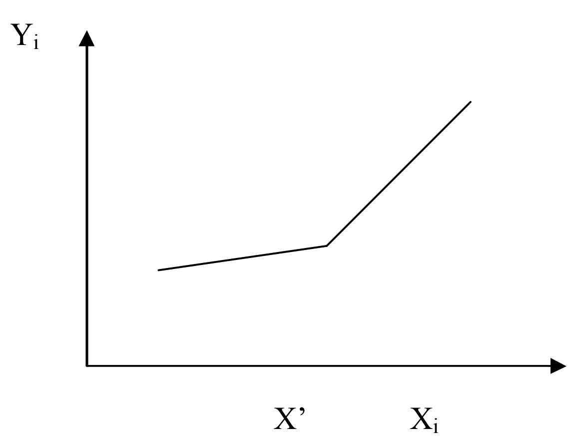


Hình 4.3. Thu nhập của một cán bộ giảng dạy đại học

Sau khi ước lượng hồi quy (4.4) chúng ta dễ thấy rằng liệu có sự khác nhau về thu nhập của cán bộ giảng dạy ở các miền khác nhau của đất nước không.

4.3. HỒI QUY TUYẾN TÍNH TỪNG KHÚC

Hầu hết các mô hình kinh tế lượng mà chúng ta nghiên cứu cho đến nay đều là các mô hình liên tục theo nghĩa là cả biến độc lập và biến phụ thuộc lấy một số lớn giá trị và sự thay đổi nhỏ trong một biến này có ảnh hưởng đo được đến biến khác. Điều này đã được cải biên khi chúng ta sử dụng thủ tục biến giả để giải thích cho sự khác nhau về hệ số chặn hay độ dốc hoặc cả hệ số chặn và độ dốc. Bây giờ chúng ta mở rộng sự phân tích cho phép thay đổi độ dốc, nhưng hạn chế rằng đoạn thẳng được ước lượng vẫn là liên tục. Thí dụ chỉ ra ở hình 4.5, mô hình đúng là một mô hình liên tục hay có sự thay đổi về kết cấu.



Hình 4.5. Mô hình hồi quy tuyến tính từng khúc.

Nếu chúng ta xem xét tiêu dùng của nước ta trước và sau khi chuyển đổi thì chúng ta thấy mô hình có dạng như hình 4.5.

Ở đây cần nhấn mạnh rằng mô hình đang xem xét khác với các mô hình biến giả đã được trình bày trong mục trước bởi vì chúng ta giả thiết rằng không có sự mất liên tục hoặc sự dịch chuyển trong mức tiêu dùng từ năm này qua năm khác. Mô hình như vậy chúng ta gọi là mô hình tuyến tính từng khúc, ở hình 4.5 gồm 2 đoạn. Chúng ta sẽ thấy mô hình có thể ước lượng được bằng phương pháp bình phương nhỏ nhất với việc sử dụng biến giả thích hợp.

Để ước lượng mô hình đã cho trong hình 4.5, chúng ta giả thiết rằng tiêu dùng nước ta trong 2 thời kỳ trước và sau chuyển đổi khác nhau. Gọi năm chuyển đổi kinh tế (từ cơ chế kế hoạch hoá sang cơ chế thị trường) là t_0. Ta xét mô hình sau:

$$Y_i = \beta_1 + \beta_2 X_t + \beta_3 (X_t - X_{t_0}) D_t + U_t \qquad (4.13)$$

Y_t: Tiêu dùng

X_t: Thu nhập

X_{t_0} : Thu nhập trong năm bắt đầu chuyển giai đoạn từ cơ chế có kế hoạch sang cơ chế thị trường.

$$D_t = \begin{cases} 1 \text{ Nếu } t > t_0 \\ 0 \text{ Nếu là giá trị khác của } t \end{cases}$$

Với giả thiết $E(U_t) = 0$ chúng ta thấy ngay rằng: trung bình của tiêu dùng trong những năm trước khi chuyển đổi kinh tế là:

$$E(Y_i / D_t = 0, X_t) = \beta_1 + \beta_2 X_t \qquad (4.14)$$

và với $D_t = 1$ thì ta có:

$$E(Y_i / D_t = 1, X_t) = \beta_1 - \beta_3 X_{t_0} + (\beta_2 + \beta_3) X_t \quad (4.15)$$

Vậy β_2 cho độ dốc của đường hồi quy trước khi chuyển đổi. $(\beta_2 + \beta_3)$ cho độ dốc của đường hồi quy sau khi chuyển đổi.

Chú ý rằng không có sự gián đoạn vì:

$$E(Y_{t_0}) = \beta_1 + \beta_2 X_{t_0} = \beta_1 - \beta_3 X_{t_0} + (\beta_2 + \beta_3) X_{t_0}$$
$$= \beta_1 + \beta_2 X_{t_0}$$

Ta cũng chú ý rằng khi $\beta_3 = 0$ thì phương trình (4.13) sẽ trở thành phương trình của đường thẳng, vậy kiểm định $\beta_3 = 0$ sẽ cung cấp cho ta kiểm định đơn giản về sự thay đổi cấu trúc.

Nhưng vấn đề sẽ như thế nào nếu mô hình có nhiều thay đổi về cấu trúc ứng với t_0 và t_1 thì mô hình thích hợp sẽ là:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 (X_t - X_{t_0}) D_1 + \beta_4 (X_t - X_{t_1}) D_2 + U_t$$

Trong đó:

$$D_2 = \begin{cases} 1 \text{ Nếu } t > t_1 \\ 0 \text{ Nếu t nhận giá trị khác} \end{cases}$$

$$D_4 = \begin{cases} 1 \text{ Nếu } t > t_0 \\ 0 \text{ Nếu t nhận giá trị khác.} \end{cases}$$

Vậy phương trình cho mỗi một trong 3 giai đoạn là như sau:

$$E(Y_t) = \begin{cases} \beta_1 + \beta_2 X_t & \text{với } 0 < t \le t_0 \\ \beta_1 - \beta_3 X_{t_0} + (\beta_2 + \beta_3) X_t & \text{với } t_0 < t \le t_1 \\ \beta_1 - \beta_3 X_{t_0} - \beta_4 X_{t_1} + (\beta_2 + \beta_3 + \beta_4) X_t & \text{với } t > t_1 \end{cases}$$

Câu hỏi chương 4

1. Trình bày mô hình hồi quy với biến giả trong trường hợp mô hình chỉ có một biến chất và biến chất có 2 phạm trù, biến chất có nhiều hơn 2 phạm trù.
2. Trình bày mô hình hồi quy với biến giả trong trường hợp có một biến lượng và một biến chất. Trong đó biến chất có 2 phạm trù, biến chất có nhiều hơn 2 phạm trù.
3. Trình bày mô hình hồi quy với biến giả có một biến lượng và 2 biến chất, mô hình biến giả có kết hợp hai hồi quy, mô hình biến giả trong phân tích mùa.
4. Trình bày mô hình hồi quy tuyến tính từng khúc trường hợp chỉ có một thay đổi về cấu trúc (tức là chỉ có một năm chuyển đổi).

5. Giả thiết rằng *Thu nhập* hàng năm của một công nhân dệt may phụ thuộc vào *Tay nghề*, *Khu vực làm việc* (nhà nước và tư nhân) và *Giới tính* (nam và nữ). Khu vực nhà nước và giới tính nữ là những phạm trù cơ sở. Giữa 2 biến giả khu vực làm việc và giới tính có sự tương tác lẫn nhau.

   a) Thiết lập mô hình.
   b) Nếu giả sử thêm rằng thu nhập hàng năm của công nhân dệt may còn phụ thuộc vào đặc tính *Vùng* (Bắc, Trung, Nam) thì mô hình mới sẽ như thế nào? (miền Trung là phạm trù cơ sở).

CHƯƠNG 5
ĐA CỘNG TUYẾN VÀ TỰ TƯƠNG QUAN

Trọng tâm của chương 5 là bàn về một số giả thiết của mô hình hồi quy tuyến tính cổ điển bao gồm bản chất của hiện tượng, hậu quả, nguyên nhân, cách phát hiện và biện pháp khắc phục.

Nội dung cơ bản của chương này bao gồm:

O Đa cộng tuyến

- Bản chất của đa cộng tuyến

- Ước lượng khi có đa cộng tuyến hoàn hảo

- Ước lượng khi có đa cộng tuyến không hoàn hảo

- Hậu quả của đa cộng tuyến

- Phát hiện sự tồn tại của đa cộng tuyến

- Biện pháp khắc phục

O Hiện tượng tự tương quan

- Nguyên nhân của hiện tượng tự tương quan

- Ước lượng bình phương nhỏ nhất khi có hiện tượng tự tương quan

- Hậu quả của hiện tượng tự tương quan

- Phát hiện có tự tương quan

5.1. HIỆN TƯỢNG ĐA CỘNG TUYẾN

Trong mô hình hồi qui bội, giả thiết 5 nói rằng giữa các biến X_i không có quan hệ tuyến tính . Vậy, nếu các biến X_i có quan hệ tuyến tính thì chuyện gì sẽ xảy ra.

5.1.1. Bản chất của đa cộng tuyến

Thuật ngữ **đa cộng tuyến** (multicollinearity) do Ragnar Frisch đưa ra năm 1934. Ý tưởng ban đầu là để chỉ hiện tượng các biến độc lập trong mô hình hồi qui có quan hệ tuyến tính hoàn hảo với nhau.

Giả sử trong mô hình có k biến: $Y_i = \beta_1 + \beta_2X_{2i} + \beta_3X_{3i} + ... + \beta_kX_{ki} + u_i$ (5.1)

Quan hệ tuyến tính hoàn hảo (đa cộng tuyến hoàn hảo) tồn tại giữa các biến X_i nếu:

$\lambda_2X_{2i} + \lambda_3X_{3i} + ... + \lambda_k X_{ki} = 0$ với các λ_i (i = 2,3,...,k) không đồng thời bằng 0 (5.2)

Ngày nay, khái niệm đa cộng tuyến được sử dụng theo nghĩa rộng hơn. Nó bao gồm trường hợp đa cộng tuyến hoàn hảo và đa cộng tuyến không hoàn hảo.

Đa cộng tuyến không hoàn hảo xảy ra khi:

$$\lambda_2X_{2i} + \lambda_3X_{3i} + ... + \lambda_kX_{ki} + v_i = 0 \quad (5.3)$$

với λ_i (i = 2,3,...,k) không đồng thời bằng 0 và v_i là yếu tố ngẫu nhiên

Nguyên nhân của đa cộng tuyến

☞ Do bản chất các biến trong mô hình: Chẳng hạn, trong kinh tế các biến số, các chỉ tiêu kinh tế đều có quan hệ với nhau ở mức độ nhất định.

☞ Do thu thập số liệu: Phương pháp thu thập số liệu có thể sinh ra đa cộng tuyến nếu ta thu thập số liệu có giá trị liên hệ trên một biến số.

5.1.2. Hậu quả của đa cộng tuyến

5.1.2.1. Trường hợp đa cộng tuyến hoàn hảo

Xét mô hình hồi qui 3 biến Y, X_2, X_3 có dạng độ lệch dưới đây:

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + e_i \tag{5.4}$$

Trong đó: $y_i = Y_i - \overline{Y}$ $\quad x_i = X_i - \overline{X}$ (5.5)

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{5.6}$$

Các ước lượng bình phương nhỏ nhất sẽ là:

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} y_i x_{2i}\right)\left(\sum_{i=1}^{n} x_{3i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{3i}\right)\left(\sum_{i=1}^{n} x_{2i}x_{3i}\right)}{\left(\sum_{i=1}^{n} x_{2i}^2\right)\left(\sum_{i=1}^{n} x_{3i}^2\right) - \left(\sum_{i=1}^{n} x_{2i}x_{3i}\right)^2} \tag{5.7}$$

$$\hat{\beta}_3 = \frac{\left(\sum_{i=1}^{n} y_i x_{3i}\right)\left(\sum_{i=1}^{n} x_{2i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{2i}\right)\left(\sum_{i=1}^{n} x_{2i}x_{3i}\right)}{\left(\sum_{i=1}^{n} x_{3i}^2\right)\left(\sum_{i=1}^{n} x_{2i}^2\right) - \left(\sum_{i=1}^{n} x_{2i}x_{3i}\right)^2} \tag{5.8}$$

Giả sử $X_{3i} = \lambda X_{2i} \Rightarrow x_{3i} = \lambda x_{2i}$ trong đó λ là hằng số khác 0, thay điều kiện này vào (5.7) ta có:

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} y_i x_{2i}\right)\left(\lambda^2\sum_{i=1}^{n} x_{2i}^2\right) - \left(\lambda\sum_{i=1}^{n} y_i x_{2i}\right)\left(\lambda\sum_{i=1}^{n} x_{2i}^2\right)}{\left(\sum_{i=1}^{n} x_{2i}^2\right)\left(\lambda^2\sum_{i=1}^{n} x_{2i}^2\right) - \lambda^2\left(\sum_{i=1}^{n} x_{2i}^2\right)^2} = \frac{0}{0} \tag{5.9}$$

$\Rightarrow \hat{\beta}_2$ không xác định. Tương tự, ta cũng có thể chỉ ra $\hat{\beta}_3$ cũng không xác định

Cách khác, ta có thể thay $x_{3i} = \lambda x_{2i}$ vào (5.4) ta được

$$\begin{aligned} y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3(\lambda x_{2i}) + e_i \\ &= (\hat{\beta}_2 + \lambda\hat{\beta}_3)x_{2i} + e_i \end{aligned} \tag{5.10}$$

$= \hat{\alpha} x_{2i} + e_i$

Áp dụng phương pháp tính ước lượng bình phương nhỏ nhất ta có

$$\hat{\alpha} = \left(\hat{\beta}_2 + \hat{\beta}_3 \lambda\right) = \frac{\sum_{i=1}^{n} x_{2i} y_i}{\sum_{i=1}^{n} x_{2i}^2} \tag{5.11}$$

Như vậy, dù α được xác định một cách duy nhất thì cũng không thể xác định được $\hat{\beta}_2$, $\hat{\beta}_3$ từ một phương trình 2 ẩn.

Như vậy, trong trường hợp đa cộng tuyến hoàn hảo ta không có lời giải duy nhất cho các hệ số hồi qui mà chỉ có lời giải duy nhất cho tổ hợp của các hệ số hồi qui.

5.1.2.2. Trường hợp đa cộng tuyến không hoàn hảo

Trong thực tế, đa cộng tuyến hoàn hảo ít khi xảy ra, thường xảy ra hiện tượng đa cộng tuyến không hoàn hảo ở các số liệu chuỗi thời gian.

Với mô hình trên, giả sử ta có đa cộng tuyến không hoàn hảo: $x_{3i} = \lambda x_{2i} + v_i$ (5.12)

Trong đó, $\lambda \neq 0$ và v là biến ngẫu nhiên sao cho $\sum x_{2i} v_i = 0$ (không có quan hệ tương quan giữa x_{2i} và v_i).

Bằng phương pháp OLS, ta thu được ước lượng của hệ số hồi qui:

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} y_i x_{2i}\right)\left(\lambda^2 \sum_{i=1}^{n} x_{2i}^2 + \sum_{i=1}^{n} v_i^2\right) - \left(\lambda \sum_{i=1}^{n} y_i x_{2i} + \sum_{i=1}^{n} y_i v_i\right)\left(\lambda \sum_{i=1}^{n} x_{2i}^2\right)}{\left(\sum_{i=1}^{n} x_{2i}^2\right)\left(\lambda^2 \sum_{i=1}^{n} x_{2i}^2 + \sum_{i=1}^{n} v_i^2\right) - \lambda^2 \left(\sum_{i=1}^{n} x_{2i}^2\right)^2} \tag{5.13}$$

Như vậy, không có lý do gì để nói rằng (5.13) là không ước lượng được

✯ Về mặt thực hành, sự xuất hiện của đa cộng tuyến dẫn đến những hậu quả sau:

☞ Phương sai và hiệp phương sai bị phóng đại

$$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{2i}^2 (1 - r_{23}^2)} \tag{5.14}$$

$$\operatorname{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{3i}^2 (1 - r_{23}^2)}$$

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum_{i=1}^{n} x_{2i}^2 \sum_{i=1}^{n} x_{3i}^2}} \tag{5.15}$$

Trong đó, r_{23} là hệ số tương quan giữa X_2 và X_3. Nếu đa cộng tuyến xảy, r_{23} sẽ dần tới $r_{23} \rightarrow \pm 1$ thì phương sai và hiệp phương sai sẽ rất lơn.

Sự tăng lên của phương sai và hiệp phương sai có thể thấy được qua hệ số phóng đại phương sai (VIF):

$$VIF = \frac{1}{(1 - r_{23}^2)} \quad (r_{23} \rightarrow \pm 1 \Rightarrow \text{VIF dần đến vô cùng}) \tag{5.16}$$

$$\Rightarrow \text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{2i}^2} VIF \qquad \text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{3i}^2} VIF$$

Do đó, hệ số VIF cho biết phương sai của các ước lượng bị phóng đại như thế nào khi có đa cộng tuyến.

☞ Khoảng tin cậy rộng hơn

Ta có khoảng tin cậy 95% của β_i như sau: $\left[\hat{\beta}_i - 1.96se(\hat{\beta}_i); \hat{\beta}_i + 1.96se(\hat{\beta}_i)\right]$

Có đa cộng tuyến ⇒ var ($\hat{\beta}_i$) lớn ⇒ se($\hat{\beta}_i$) lớn ⇒ khoảng tin cậy lớn

☞ Thống kê t thấp

Khi kiểm định giả thiết H_0: $\beta_i = 0$, ta sử dụng thống kê t = $\hat{\beta}_i$/se($\hat{\beta}_i$) và so sánh nó với giá trị tới hạn t_α. Có đa cộng tuyến ⇒ var ($\hat{\beta}_i$) lớn ⇒ se($\hat{\beta}_i$) lớn ⇒ t nhỏ ⇒ tăng khả năng chấp nhận H_0.

☞ R^2 cao nhưng thống kê t thấp

Xét mô hình hồi qui k biến

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i \tag{5.17}$$

Khi có đa cộng tuyến, như đã nói ở phần trên, các hệ số góc ước lượng được có thể không có ý nghĩa về mặt thống kê do thống kê t nhỏ. Nhưng trong khi đó, R^2 lại có thể rất cao nên khả năng bác bỏ giả thiết H_o: $\beta_2 = \beta_3 = ... = \beta_k = 0$ bằng kiểm định F là rất cao.

☞ Dấu của các ước lượng của hệ số hồi qui có thể sai

☞ Các ước lượng OLS và các sai số tiêu chuẩn của chúng trở nên rất nhạy đối với những thay đổi trong số liệu. Thêm vào hay bớt đi các biến cộng tuyến với các biến khác, độ lớn và dấu của các ước lượng sẽ có thể thay đổi rất nhiều.

5.1.3. Cách pháp hiện đa cộng tuyến

5.1.3.1. R^2 cao nhưng thống kê t có ý nghĩa thấp

Trong trường hợp hệ số tương quan cao, thống kê t thấp thì có khả năng tồn tại đa cộng tuyến

5.1.3.2. Hệ số tương quan cặp giữa 2 biến giải thích cao

Trong trường hợp hệ số tương quan cặp cao, ví dụ vượt qua 0.8 thì có khả năng tồn tại đa cộng tuyến. Tuy nhiên, hệ số tương quan cặp cao chỉ là điều kiện đủ chứ không phải là điều kiện cần của đa cộng tuyến. Đa cộng tuyến có thể xảy ra ngay cả khi tương quan cặp thấp.

5.1.3.3.Xem xét các hệ số tương quan riêng

Trong hồi qui Y theo X_2, X_3, X_4. Nếu ta nhận thấy $R^1_{1.234}$ (R^2 thu được từ hồi qui) cao trong khi $r^2_{12,34}$, $r^2_{13,24}$, $r^2_{14,23}$ tương đối thấp thì có thể gợi ý rằng X_2, X_3, X_4 có tương quan cao và ít nhất một trong các biến này không cần thiết cho mô hình.

($r^2_{12,34}$ là hệ số tương quan riêng giữa x_1 và x_2 trong điều kiện kiểm soát x_3 và x_4)

5.1.3.4. Hồi qui phụ

Vì đa cộng tuyến phát sinh khi một hay nhiều biến giải thích có quan hệ tuyến tính (hoàn hảo hoặc không hoàn hảo) với các biến giải thích khác nên ta có thể pháp hiện đa cộng tuyến bằng cách hồi qui hồi qui mỗi biến giải thích theo các biến giải thích còn lại, gọi là hồi qui phụ, từ đó thu được R^2 tương ứng, kí hiệu là R_i^2. (Hồi qui chính là hồi qui Y theo các X_i)

Giả sử xét mô hình (hồi qui chính): $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + U$

(5.18)

Hồi qui phụ có dạng:

$$X_i = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + ... + \alpha_{i-1} X_{i-1} + \alpha_{i+1} X_{i+1} + ... + \alpha_k X_k + v \qquad (5.19)$$

Ước lượng (5.19) ta thu được R_i^2. Ta kiểm định giả thiết

H_0: Không có đa cộng tuyến $\Leftrightarrow \alpha_2 = \alpha_3 = ... = \alpha_{i-1} = \alpha_{i+1} = ... = \alpha_k = 0$

H_1: Có đa cộng tuyến $\Leftrightarrow \exists \alpha_i = 0$

Sử dụng tiêu chuẩn kiểm định F:

$$F_i = \frac{R_i^2}{1 - R_i^2} . \frac{n - (k-1)}{k - 1 - 1} = \frac{R_i^2}{1 - R_i^2} . \frac{n - k + 1}{k - 2} \qquad (5.20)$$

$F_i \sim F_\alpha(k\text{-}2,n\text{-}k+1)$

$F_i > F_\alpha(k\text{-}2,n\text{-}k+1) \Rightarrow$ bác bỏ giả thiết $H_0 \Rightarrow$ có đa cộng tuyến

$F_i < F_\alpha(k\text{-}2,n\text{-}k+1) \Rightarrow$ không có cơ sở để bác bỏ giả thiết H_0

5.1.3.5. Hệ số phóng đại phương sai (VIF)

Một cách tổng quát, VIF được tính bằng công thức: $VIF(X_i) = \dfrac{1}{1-R_i^2}$

R_i^2 là R^2 thu được từ mô hình hồi qui (5.19). Khi R_i tăng thì VIF tăng, và VIF đặc biệt tăng mạnh khi R_i tăng từ 0.9 đến 1. Khi VIF > 10 là dấu hiệu của đa cộng tuyến nhưng điều này cũng không nhất thiết đúng.

5.1.3.6. Tiêu chuẩn Theil

$$m = R^2 - \sum_{i=2}^{k}(R^2 - R_{-i}^2) \qquad (5.22)$$

Trong đó, R^2 là hệ số xác định bội trong mô hình hồi qui chính

$$Y = \beta_1 + \beta_2X_2 + \beta_3X_3 + ... + \beta_kX_k + U \qquad (5.23)$$

R_{-i}^2 là hệ số xác định bội thu được từ việc hồi qui mô hinh (5.23) sau khi đã bỏ đi biến X_i

$(R^2 - R_{-i}^2)$ được gọi là mức độ đóng góp của X_i đối với R^2.

- Nếu có đa cộng tuyến :m = 0
- Nếu không có đa cộng tuyến : $m \neq 0$

Tuy nhiên chỉ số này không xác định được mức độ nghiêm trọng của đa cộng tuyến.

5.1.4. Các phương pháp khắc phục

5.1.4.1. Dùng thông tin tiên nghiệm

Hồi qui mô hình:

$Y_i = \beta_1 + \beta_2X_{2i} + \beta_3X_{3i} + U_i$

với Y: tiêu dùng, X_2: thu nhập, X_3: tài sản

Rõ ràng, thu nhập và tài sản là những biến có quan hệ cộng tuyến cao. Giả sử, từ những nghiên cứ khác, ta biết rằng $\beta_3 = 0.1\beta_2$, nghĩa là tỉ lệ thay đổi của tiêu dùng theo tài sản bằng 1/10 tỉ lệ thay đổi của tiêu dùng theo thu nhập.

Ta có thể viết lại hồi qui như sau:

$$Y_i = \beta_1 + \beta_2X_{2i} + 0.1\beta_2X_{3i} + U_i$$

$$\beta_1 + \beta_2X_i + U_i \qquad (\text{với } X_i = X_{2i} + 0.1X_{3i}) \qquad (5.24)$$

Từ việc ước lượng phương trình (5.24) ta thu được $\hat{\beta}_2$. Sau đó, ta dựa vào thông tin tiên nghiệm để tìm ra ước lượng $\hat{\beta}_3$

5.1.4.2. Thu thập thêm số liệu hoặc lấy thêm mẫu mới

Đa cộng tuyến là một đặc trưng của mẫu nên có thể tồn tại mẫu khác liên quan đến cùng các biến trong mẫu ban đầu và sử dụng số liệu từ mẫu mới này làm cho đa cộng tuyến bớt nghiêm trọng hơn. Điều này chỉ làm được khi chi phí cho việc lấy mẫu khác không quá cao.

5.1.4.3. Bỏ bớt biến

Khi mô hình có đa cộng tuyến thì cách *"đơn giản nhất"* là loại bỏ các biến cộng tuyến ra khỏi mô hình. Biện pháp này được tiến hành như sau:

Trong ví dụ về mô hình hồi qui tiêu dùng Y theo thu nhập X_2 và tài sản X_3. Dựa vào giá trị của R^2 và $\overline{R}^2$ trong các hồi qui Y theo X_2 và Y theo X_3 để quyết định nên bỏ biến nào trong 2 biến X_2 và X_3 khỏi mô hình.

Giả sử, R^2 trong hồi qui của Y đối với tất cả các biến X_2, X_3 là 0,94; R^2 khi loại biến X_2 là 0,87 và R^2 khi loại X_3 là 0,92; Như vậy, trong trường hợp này, ta loại X_3 khỏi mô hình và khắc phục được vấn đề đa cộng tuyến.

Tuy nhiên khi loại bỏ biến tài sản khỏi mô hình có thể làm mô hình mắc lỗi kỹ thuật. Do đó, nếu lý thuyết kinh tế khẳng định tiêu dùng phụ thuộc vào cả thu nhập và tài sản thì trong mô hình nên có biến tài sản.

Đa cộng tuyến có thể làm giảm tính chính xác của các ước lượng thu được nhưng việc bỏ biến có thể dẫn đến một sai lầm nghiêm trọng hơn: đó là sự ngộ nhận về các giá trị đúng của các tham số ước lượng (giá trị ước lượng sai nhưng ta tin là nó đúng).

5.1.4.4. Thực hiện phép biến đổi với các biến số

Sử dụng sai phân cấp 1

Khi chúng ta hồi qui với các biến số là các chuỗi thời gian, đa cộng tuyến có thể xảy ra do xu hướng vận động (tăng, giảm) của các biến theo thời gian. Nếu có tồn tại đa cộng tuyến, có thể dùng sai phân cấp 1 để khắc phục.

Thí dụ: Chúng ta có số liệu chuỗi thời gian biểu thị liên hệ giữa biến tiêu dùng Y và các biến phụ thuộc thu nhập X_2 và tài sản X_3 bằng mô hình sau:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + U_t \qquad (5.25)$$

Trong đó t là thời gian. Phương trình trên đúng với t thì cũng đúng với t-1 nghĩa là:

$$Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + \beta_3 X_{3t-1} + u_{t-1} \qquad (5.26)$$

Từ (5.30) và (5.31) ta được:

$$Y_t - Y_{t-1} = \beta_2 (X_{2t} - X_{2t-1}) + \beta_3 (X_{3t} - X_{3t-1}) + u_t - u_{t-1} \quad (5.27)$$

Đặt : $y_t = Y_t - Y_{t-1}$

$x_{2t} = X_{2t} - X_{2t-1}$

$x_{3t} = X_{3t} - X_{3t-1}$

$V_t = U_t - U_{t-1}$

Ta được: $y_t = \beta_2 x_{2t} + \beta_3 x_{3t} + v_t$ (5.28)

Phương trình (5.28) được gọi là phương trình có dạng sai phân cấp 1. Phương trình sai phân cấp 1 có các biến thể hiện sự khác biệt giữa các thời kì.

Mô hình hồi qui dạng (5.28) thường làm giảm tính nghiêm trọng của đa cộng tuyến vì X_2 và X_3 có thể tương quan cao nhưng không có lý do tiên nghiệm nào chắc chắn rằng sai phân của chúng cũng tương quan cao.

Tuy nhiên biến đổi sai phân bậc nhất sinh ra một số vấn đề (1) chẳng hạn nhưsố hạng sai số v_t trong (5.28) có thể không thoả mãn giả thiết của mô hình hồi qui tuyến tính cổ điển là các nhiễu không tương quan. Khi đó, biện pháp sửa chữa này lại có thể dẫn đến những khuyết tật nghiêm trọng hơn cho mô hình so với trước khi sửa chữa. (2) Mất đi một quan sát do tiến hành sai phân sẽ làm giảm môt bậc tự do của mô hình vì thế cần cân nhắc khi áp dụng biện pháp này đối với các mẫu nhỏ. (3) Phương pháp này không thể áp dụng cho số liệu chéo (không có yếu tố thời gian).

5.1.4.5. Hồi qui đa thức

Khi mô hình có đa cộng tuyến, ta có thể khắc phục bằng cách đưa vào mô hình các dạng khác nhau của biến độc lập. Biến độc lập có thể ở dạng bậc 2, bậc 3… Phương pháp này cũng làm giảm đa cộng tuyến trong mô hình.

5.1.4.6. Một số biện pháp khác

Ngoài những biện pháp đã đưa ra người ta còn sử dụng một số biện pháp khác để khác phục hiện tượng đa cộng tuyến sau:

- Hồi qui thành phần chính (Phân tích nhân tố)
- Hồi qui ngọn sóng
- Sử dụng các ước lượng từ bên ngoài

5.2. HIỆN TƯỢNG TỰ TƯƠNG QUAN

Một trong các giả thiết của mô hình hồi quy tuyến tính cổ điển là không có tự tương quan hay tương quan chuỗi các nhiễu U_i trong hàm hồi quy tổng thể. Nhưng trong thực tế liệu hiện tượng đó có xảy ra hay không? Nguyên nhân của hiện tượng đó là gì? Nếu có hiện tượng tự tương quan thì liệu nó còn áp dụng được phương pháp bình phương bé nhất nữa không? Làm thế nào để biết rằng hiện tượng tự tương quan xảy ra? Cách khắc phục? ... Đó là một loạt các câu hỏi mà chúng ta cần phải giải quyết trong phần này.

5.2.1. Bản chất của hiện tượng tự tương quan

5.2.1.1. Khái niệm

Thuật ngữ tự tương quan (autocorrelation) có thể hiểu là sự tương quan giữa các thành phần của chuỗi các quan sát được sắp xếp theo thứ tự thời gian (trong các số liệu chuỗi thời gian) hoặc không gian (trong số liệu chéo).

Mô hình hồi quy tuyến tính cổ điển giả thiết không có tương quan giữa các yếu tố ngẫu nhiên U_i, nghĩa là: $cov(u_i, u_j) = 0$

Giả thiết này có nghĩa là yếu tố ngẫu nhiên của bất kỳ quan sát nào cũng không bị ảnh hưởng bởi yếu tố ngẫu nhiên của các quan sát khác.

Xét mô hình hồi qui 2 biến: $Y_t = \beta_1 + \beta_2 X_{2t} + u_t$ (7.1)

Giả sử: $cov(u_t, u_{t+s}) \neq 0$ ($\forall \ s \neq 0$) và nhiễu có quan hệ với nhau theo dạng sau

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad (7.2)$$

trong đó, ρ được gọi là hệ số tự hiệp phương sai (autocovariance), ε_t là nhiễu ngẫu nhiên thoả mãn các giả thiết của OLS.

$$E(\varepsilon_t) = 0$$

$$Var(\varepsilon_t) = \sigma_\varepsilon^2$$

$Cov(\varepsilon_t, \varepsilon_{t+s}) = \gamma_s$ (covarian giữa hai thời kì không phụ thuộc vào thời kì, t, mà chỉ phụ thuộc vào khoảng cách giữa các thời kì)

u_{t-1}: trễ một thời kỳ của $u_t \Rightarrow$ (7.2) được ký hiệu là AR(1)

Khi đó, (7.2) được gọi là tự hồi qui bậc nhất và ta nói có tự tương quan bậc nhất trong mô hình.

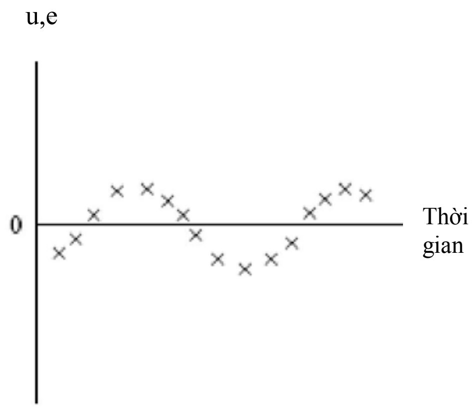
+ Tương tự ta có tự tương quan bậc 2:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t \quad AR(2)$$

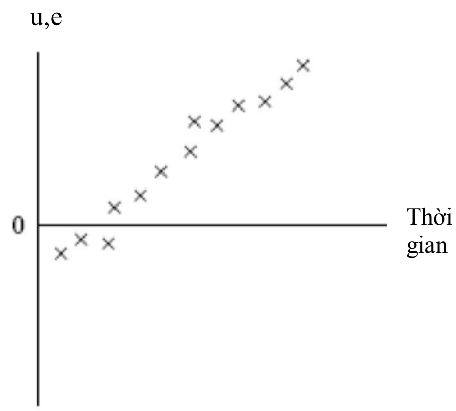
+ Tổng quát: Tự tương quan bậc p:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_p u_{t-p} + \varepsilon_t \qquad AR(p)$$

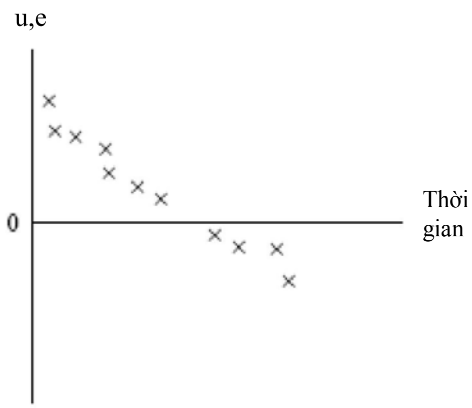
Khi giữa các u_i có tự tương quan, đồ thị của U_i theo thời gian có thể ở 1 trong các dạng ở đồ thị (a) – (d). Đồ thị (e) không tạo lên một hình dạng nhất định chứng tỏ không có tự tương quan



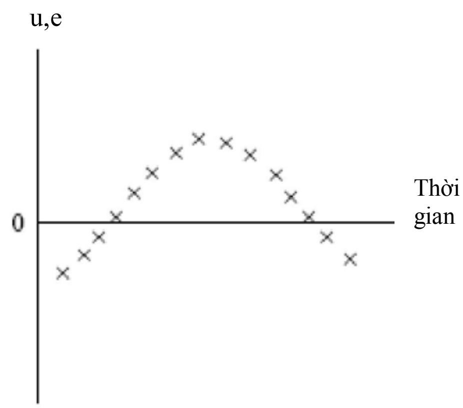
(a)



(b)



(c)



(d)

5.2.1.2. Nguyên nhân của tự tương quan

a. Nguyên nhân khách quan

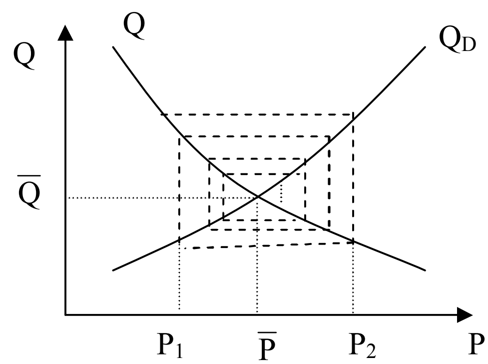
- Do tính quán tính của các số liệu trong kinh tế. Ví dụ, tổng sản phẩm, chỉ số giá… thời kỳ sau phụ thuộc thời kỳ trước.

- Hiện tượng mạng nhện kiến yếu tố nhiễu không còn ngẫu nhiên mà mang tính hệ thống.

Ví dụ, Lượng cung, lượng cầu của thời kỳ này chịu ảnh hưởng của giá thời kỳ trước.

$Q_D = f(p_{t-1})$

$Q_S = f(p_{t-1})$



- Trong phân tích hồi qui chuỗi thời gian, mô hình có thể chứa các biến phụ thuộc ở thời kỳ trễ là các biến độc lập. Nếu chúng ta bỏ qua các yếu tố trễ này sẽ là sai số ngẫu nhiên mang tính hệ thống và dẫn tới hiện tượng tự tương quan.

Ví dụ: Hàm tiêu dùng: $C_t = \beta_1 + \beta_2 Y_t + \beta_3 C_{t-1} + u_t$

(do người tiêu dùng thường thay đổi thói quen tiêu dùng).

b. Do yếu tố chủ quan

- Do quá trình xử lý số liệu có thể dẫn đến hiện tượng các yếu tố ngẫu nhiên tự tương quan với nhau. Ví dụ quá trình lấy trung bình trượt để làm trơn số liệu, hay quá trình ngoại suy… đều có thể dẫn đến hiện tượng tự tương quan.

- Do mô hình đã bỏ sót 1 hay 1 số biến thích hợp. Điều này không chỉ dẫn đến ước lượng chệch mà còn dẫn đến hiện tượng tự tương quan.

Giả sử, mô hình đúng có dạng: $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$

Trong đó, Y là lượng cầu về thịt bò, X_2 là giá thịt bò, X_3 là thu nhập của người tiêu dùng, X_4 là giá thịt lợn và t là thời gian.

Nhưng vì một lý do nào đó, ta hồi qui mô hình: $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + v_t$

$\Rightarrow v_t = \beta_4 X_{4t} + u_t$. Vậy, rõ ràng các v_t sẽ tương quan với nhau vì trong v_t có yếu tố giá (thụ thuộc lẫn nhau theo thời gian)

5.2.2. Hậu quả của việc sử dụng phương pháp bình phương nhỏ nhất khi có tự tương quan

$\hat{\beta}$ vẫn là ước lượng không chệch, nhưng không phải là các ước lượng có phương sai nhỏ nhất.

- σ^2 là ước lượng chệch của σ_i^2 (thông thường nó ước lượng nhỏ hơn giá trị thực của $\hat{\sigma}^2$).
- Kiểm định T và kiểm định F mất ý nghĩa.
- R^2 tính toán được có thể là độ đo không đáng tin cậy cho R^2 thực

5.2.3. Cách phát hiện

5.2.3.1. Dùng đồ thị phần dư

Ta sử dụng phần dư để đại diện cho nhiễu u_t.

☞ Vẽ đồ thị phần dư theo thời gian

☞ Vẽ đồ thị phần dư chuẩn hoá theo thời gian. Việc chuẩn hoá giúp triệt tiêu đơn vị, do đó, ta có thể so sánh các phần dư chuẩn hoá trong các hồi qui khác nhau với nhau.

Có $\frac{U_t}{\sigma} \sim N(0,1) \Rightarrow$ với các mẫu lớn, $\frac{e_t}{\hat{\sigma}_t}$ phân phối xấp xỉ N(0,1)

☞ Vẽ đồ thị phần dư theo các giá trị trễ

Dùng phương pháp OLS ước lượng mô hình xuất phát $\Rightarrow e_t$

Để phát hiện AR(1): vẽ đồ thị e_t phụ thuộc e_{t-1}

Để phát hiện AR(p): vẽ đồ thị e_t phụ thuộc e_{t-p}

5.2.3.2. Kiểm định Durbin-Watson (DW)

a. Giả thiết của Dubin-Watson

Sử dụng thống kê Durbin-Watson để pháp hiện tương quan chuỗi là một phương pháp phổ biên. Thống kê Durbin-Watson được kí hiệu là d:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad (7.3)$$

Thống kê d được tính dựa trên phần dư nên chắc chắn tính được. Vì lợi thế này, thống kê d thường được trình bày trong kết quả hồi qui từ các phần mềm kinh tế lượng. Tuy nhiên, cũng cần chú ý tới những giả thiết sau đối với thống kê d.

☞ Mô hình hồi qui phải có hệ số chặn.

☞ Các biến giải thích là biến phi ngẫu nhiên (xác định trong các mẫu nhắc lại). Giả thiết này rất khó đáp ứng trong mô hình kinh tế liên quan đến chuỗi thời gian.

☞ Nhiễu u_t được hình thành từ quá trình tự hồi qui bậc 1 ($u_t = \rho u_{t-1} + \varepsilon_t$) nên không thể sử dụng thống kê d pháp hiện tự tương quan bậc cao.

☞ Nhiễu U_t được giả thiết phân bố chuẩn. Nếu không, thống kê d sẽ không đáng tin cậy.

☞ Mô hình hồi qui gốc không chứa độc lập là biến trễ của biến phụ thuộc (mô hình tự hồi qui $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \ldots + \beta_k X_{kt} + \gamma Y_{t-1} + u_t$) . Trường hợp ngược lại, giá trị thống kê d thường xấp xỉ 2, chứng tỏ không có tự tương quan ngay cả khi hiện tượng này xuất hiện trong mô hình.

☞ Không có quan sát bị thiếu trong số liệu.

b. Xây dựng thống kê d:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

Vì các $\sum e_t^2$ và $\sum e_{t-1}^2$ chỉ khác nhau 1 quan sát nên coi chúng xấp xỉ nhau khi đó ta có:

$$d = 2\left(1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}\right) \qquad (7.4)$$

Đặt $$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

Ki đó $$d \approx 2(1 - \hat{\rho}) \qquad (7.5)$$

Gọi $\hat{\rho}$ là hệ số tương quan bậc nhất của mẫu, đó là ước lượng của ρ.

Vì $0 \le \rho \le 1$ nên ta suy ra rằng $0 \le d \le 0$

Nếu $\hat{\rho} = 0 \Rightarrow d = 2$: không có tự tương quan

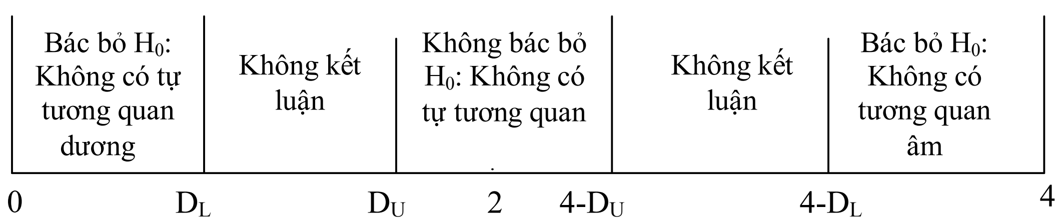
Nếu $\hat{\rho} = 1 \Rightarrow d = 0$: tồn tại tự tương quan dương hoàn hảo

Nếu $\hat{\rho} = -1 \Rightarrow d = 4$: tồn tại tự tương quan âm hoàn hảo

c. Quy tắc

Durbin và Watson đưa ra các giá trị tới hạn $D_{L\ (Lower)}$ và $D_{U\ (Upper)}$ để thực hiện kiểm định giả thiết về sự tồn tại tương quan chuỗi. Việc ra quyết định dựa trên thống kê d và các giá trị tới hạn của nó được thực hiện như sau

Giả thiết H_0	Kết luận	Nếu
Không có tự tương quan dương	Bác bỏ	$0 < 0 < d_L$
Không có tự tương quan dương	Không kết luận	$d_L \le d \le d_U$
Không có tương quan âm	Bác bỏ	$4 - d_L < d < 4$
Không có tương quan âm	Không kết luận	$4 - d_U \le d \le 4 - d_L$
Không có tự tương quan	Không bác bỏ	$d_U < d < 4 - d_U$



5.2.3.3. Kiểm định Breusch-Godfrey (BG)

Xét mô hình: $Y_t = \beta_1 + \beta_2 X_t + u_t$

Trong đó, $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_p u_{t-p} + \varepsilon_t$ AR(p)

Với ε_t thoả mãn các giả thiết của OLS.

Giả thiết H_0: $\rho_1 = \rho_2 = \ldots = \rho_p = 0$ (không có tự tương quan) được kiểm định như sau:

**** Các bước kiểm định***

Bước 1: Dùng OLS ước lượng mô hình xuất phát, thu được e_t

Bước 2: Ước lượng mô hình: $e_t = \beta_1 + \beta_1 X_t + \rho_1 e_{t-1} + \rho_2 e_{t-2} + ... + \rho_p e_{t-p} + v_t$

Thu được R^2 (kích thước mẫu chỉ còn n-p).

Bước 3: Với n đủ lớn, ta có (n-p)R^2 có phân phối xấp xỉ $\chi^2(p)$

Kiểm định giả thiết:

H_0: $\rho_1 = \rho_2 = ... = \rho_p = 0$ (không có tự tương quan)

H_1: $\exists \rho_i \neq 0$ (có tự tương quan)

Nếu $(n-p)R^2 > \chi^2_\alpha(p)$: Bác bỏ giả thiết H_0

Nếu $(n-p)R^2 < \chi^2_\alpha(p)$: Không đủ cơ sở bác bỏ giả thiết H_0

* Ngoài ra có thể sử dụng tiêu chuẩn F để kiểm định giả thiết H_0 bằng cách thực hiện hồi quy có điều kiện ràng buộc.

Chú ý trong thực hành kiểm định BG:

☞ BG có thể áp dụng đối với:

+ mô hình gốc có chứa biến giải thích là biến trễ của biến phụ thuộc (Y_{t-1}, Y_{t-2}…)

+ nhiễu có dạng trung bình trượt của phần dư thoả mãn các điều kiện OLS:

$u_t = \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + \quad + \lambda_p \varepsilon_{t-p}$

☞ Khi độ dài của trễ p = 1, kiểm định BG chính là kiểm định Durbin-Washson

☞ Hạn chế của kiểm định BG là việc xác định độ dài của trễ (p)

5.2.4. Khắc phục hiện tượng tự tương quan

Trong trường hợp tự tương quan thuần tuý, tuỳ vào sự hiểu biết của chúng ta về bản chất của sự phụ thuộc qua lại giữa các nhiễu, ta có thể thực hiện khắc phục bằng 1 trong số các biện pháp sau.

5.2.4.1.Trường hợp ρ đã biết

Xét mô hình hồi qui 2 biến giản đơn sau:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad (7.1)$$

Giả sử có tương quan bậc 1: $u_t = \rho_1 u_{t-1} + \varepsilon_t$ (7.6)

với ε_t thoả mãn các giả thiết của OLS.

Ta có: $Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + u_{t-1}$ (7.7)

Nhân 2 vế phương trình (7.7) với ρ: $\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1}$ (7.8)

Trừ (7.1) cho (7.8) ta được:

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + u_t - \rho u_{t-1}$$

$$\Delta Y = \beta_1^* + \beta_2 X_i^* + \varepsilon_t \quad (7.9)$$

(7.9): phương trình sai phân tổng quát

Hay $Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t$

(7.9) có sai số ngẫu nhiên ε_t thoả mãn các giả thiết của phương pháp OLS $\Rightarrow$ không có tự tương quan.

5.2.4.2. Trường hợp ρ chưa biết

a. Dùng sai phân cấp 1

Mô hình xuất phát: $Y_t = \beta_1 + \beta_2 X_t + u_t$ (7.1)

$u_t = \rho_1 u_{t-1} + \varepsilon_t$, với ε_t thoả mãn các giả thiết của OLS.

$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t$: Phương trình sai phân tổng quát

Với $-1 \le \rho \le 1$, $\rho \ne 0$

+ Nếu $\rho=1 \Rightarrow Y_t - Y_{t-1} = 0 + \beta_2(X_t - X_{t-1}) + \varepsilon_t$

$$\Rightarrow \Delta Y_t = \beta_2 \Delta X_t + \varepsilon_t \quad (7.10)$$

(7.10): phương trình sai phân không có hệ số chặn

+ Nếu $\rho=-1 \Rightarrow Y_t + Y_{t-1} = 2\beta_1 + \beta_2(X_t + X_{t-1}) + \varepsilon_t$

$$\Rightarrow \frac{Y_t + Y_{t-1}}{2} = \beta_1 + \beta_2 \frac{X_t + X_{t-1}}{2} + v_t \quad (7.11)$$

(7.11) hồi quy trung bình 2 thời kỳ liên tiếp của biến phụ thuộc.

Từ các giả thiết về ρ, ta ước lượng được các hệ số $\hat{\beta}$ mà không cần biết giá trị thực của ρ (phương trình (7.10) và (7.11))

b. Dùng thống kê d

Sau khi thực hiện hồi qui, các phần mềm thường cho ta báo cáo bao gồm giá trị của thống kê Durbin-Watson d. $d=2(1-\hat{\rho})$

$$\Rightarrow \hat{\rho} = 1 - \frac{d}{2}$$

Đẳng thức này gợi cho ta cách thức đơn giản để thu được ước lượng của ρ từ thống kê d.

Sau đó, sử dụng $\hat{\rho}$ như ước lượng của ρ và thay vào phương trình sai phân tổng quát.

$Y_t - \hat{\rho}Y_{t-1} = \beta_1(1-\hat{\rho}) + \beta_2(X_t - \hat{\rho}X_{t-1}) + u_t - \hat{\rho}u_{t-1}$ hay $Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t$

Ước lượng phương trình sai phân tổng quát, ta được các ước lượng của các hệ số ban đầu $\hat{\beta}_1^*, \hat{\beta}_1^* \Rightarrow \hat{\beta}_1 = \dfrac{\hat{\beta}_1^*}{1-\hat{\rho}}$; $\hat{\beta}_2 = \hat{\beta}_2^*$

c. Ước lượng ρ dựa trên phần dư

Nếu phần dư có tự tương quan dưới dạng tự hồi qui bậc 1 [AR(1)], $u_t = \rho_1 u_{t-1} + \varepsilon_t$.

Để thu được ước lượng của ρ, ta hồi qui e_t theo e_{t-1}, vì e_t là ước lượng đúng của u_t.

Vậy, để ước lượng ρ từ phần dư, ta thực hiện theo các bước sau:

Bước 1: ước lượng mô hình gốc bằng OLS $\Rightarrow e_t$

Bước 2: hồi qui $e_t = \rho.e_{t-1} + \varepsilon_t \Rightarrow \hat{\rho}$

Chú ý: $\hat{\rho}$ thu được bằng phương pháp này không khác $\hat{\rho}$ thu được khi dựa trên thống kê d (do thống kê d cũng được tính dựa trên giả thiết có tự tương quan bậc 1)

d. Dùng thủ tục Cochrane - Ocutt

Các phương pháp trên chỉ đưa ra 1 giá trị ước lượng $\hat{\rho}$. Phương pháp Cochrane – Ocutt ước lượng ρ lặp lại nên còn được gọi là phương pháp lặp Cochrane – Ocutt.

Phương pháp này sử dụng các phần dư e_t đã được ước lượng để thu thông tin về ρ chưa biết. Các bước:

Bước 1: Ước lượng mô hình gốc $Y_t = \beta_1 + \beta_2 X_t + U_t$ (6.10) $\Rightarrow e_t$

Bước 2: Sử dụng các phần dư đã ước lượng để ước lượng hồi quy:

$e_t = \rho e_{t-1} + v_t$ (7.12) $\Rightarrow \hat{\rho}$

Bước 3: Sử dụng $\hat{\rho}$ thu được từ bước 2 để ước lượng phương trình sai phân tổng quát

$$Y_t - \hat{\rho}Y_{t-1} = \beta_1(1-\hat{\rho}) + \beta_2(X_t - \hat{\rho}X_{t-1}) + u_t - \hat{\rho}u_{t-1}$$

hay $Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t$

$\Rightarrow \hat{\beta}_1^*, \hat{\beta}_1^* \Rightarrow \hat{\beta}_1 = \dfrac{\hat{\beta}_1^*}{1-\hat{\rho}}; \ \hat{\beta}_2 = \hat{\beta}_2^*$

Bước 4. Vì chúng ta chưa biết $\hat{\rho}$ thu được từ (7.12) có phải là ước lượng tốt nhất của ρ hay không nên ta thay $\hat{\beta}_1, \hat{\beta}_2$ vào phương trình hồi qui gốc (7.12) $\Rightarrow e_t^*$

$$e_t^* = Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_t$$

Quay trở lại bước 2 và cứ tiếp tục cho đến khi hai ước lượng kế tiếp nhau của ρ khác nhau không đáng kể, chẳng hạn bé hơn 0,01 hoặc 0,05.

Trong thực tế, dùng 3 đến 4 bước lặp là đủ.

Chú ý: ở bước hai, mô hình hồi qui có thể là AR(1) hoặc tự hồi qui ở các bậc cao hơn AR(2), AR(3) …

Câu hỏi chương 5.

1. Trình bày khái niệm, nguyên nhân, hậu quả, cách phát hiện và biện pháp khắc phục của hiện tượng đa cộng tuyến.

1. Trình bày khái niệm, nguyên nhân, hậu quả, cách phát hiện và biện pháp khắc phục của hiện tượng tự tương quan.

3. Cho C là tiêu dùng, I là thu nhập và W là phúc lợi có quan hệ với nhau dưới dạng mô hình: $\hat{C}_i = \hat{\beta}_1 + \hat{\beta}_2 I_i + \hat{\beta}_3 W_i$. Dựa trên số liệu thu thập được ta có: $\sum i_i w_i = 5242; \sum i_i^2 = 1125,6; \sum w_i^2 = 25008$

Dùng tương quan cặp để kiểm tra xem mô hình có hiện tượng đa cộng tuyến không?

4. Cho mô hình: $\hat{Y}_i = 1,553 + 1,415 X_{2i} - 0,155 X_{3i}$. Dựa trên số liệu thu thập được ta có: $\sum y_i x_{2i} = 1135,5; \sum y_i x_{3i} = 5124,7; \sum y_i^2 = 899,78; se(\hat{\beta}_2) = 11,56$

Dùng hệ số xác định bội R^2 và tỉ số t để kiểm tra xem mô hình có hiện tượng đa cộng tuyến không?

5. Cho mô hình: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$. Dựa trên số liệu thu thập được ta có: $n = 20; R_{23}^2 = 0,875; F(1,18) = 4,41;$. Trong đó là R_{23}^2 hệ số xác định trong hồi quy của biến X_2 theo biến X_3.Dùng hồi quy phụ để kiểm tra xem mô hình có hiện tượng đa cộng tuyến không?

6. Cho mô hình: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$. Dựa trên số liệu thu thập được ta có:

$R^2 = 0,987; r_{12}^2 = 0,691; r_{13}^2 = 0,296;$ .Dùng độ đo Theil để kiểm tra xem mô hình có hiện tượng đa cộng tuyến không?

7. Cho mô hình: $\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_t$. Dựa trên số liệu thu thập được ta có:

$$\sum_{t=2}^{20} \left(\mathrm{e}_{\mathrm{t}} - e_{t-1} \right)^2 = 3994; \sum_{\mathrm{t}=1}^{20} \mathrm{e}_{\mathrm{t}}^2 = 2661; \ \mathrm{n} = 20; \mathrm{d}_{\mathrm{L}} = 1,201; d_U = 1,411;$$

Dùng kiểm định Durbin-Watson để xem mô hình có hiện tượng tự tương quan không?

CHƯƠNG 6. PHƯƠNG SAI SAI SỐ THAY ĐỔI

Trọng tâm của chương 6 là bàn về một số giả thiết của mô hình hồi quy tuyến tính cổ điển bao gồm bản chất của hiện tượng, hậu quả, nguyên nhân, cách phát hiện và biện pháp khắc phục.

O Nội dung cơ bản của chương này bao gồm:

O Nguyên nhân của phương sai của sai số thay đổi

O Phương pháp bình phương nhỏ nhất có trọng số

O Hậu quả của phương sai của sai số thay đổi

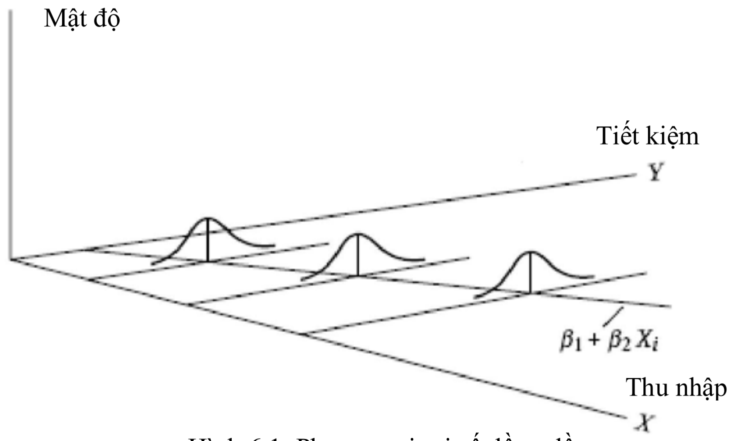
O Cách phát hiện phương sai của sai số thay đổi

O Biện pháp khắc phục

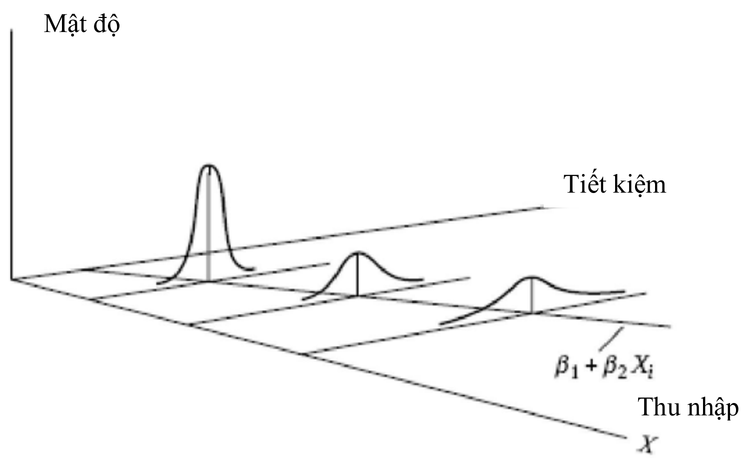
6.1. BẢN CHẤT CỦA HIỆN TƯỢNG

Một giả thiết quan trọng khác của OLS là các nhiễu ngẫu nhiên U_i trong hàm hồi quy tổng thể (PRF) có phương sai đồng đều (homoscedasticity): $Var(U_i) = Var(U_j) = \sigma^2$, $\forall$ $(i \neq j)$. Đồ thị 6.1 cho thấy phương sai có điều kiện của Y_i (cũng chính là phương sai của u_i) bằng nhau khi các giá trị X_i thay đổi.

Đồ thị 6.2 chỉ ra trường hợp ngược lại, phương sai của u_i thay đổi khi giá trị của X_i thay đổi.



Hình 6.1: Phương sai sai số đồng đều



Hình 6.2: Phương sai sai số thay đổi

✭ Phương sai sai số thay đổi có thể do những nguyên nhân sau:

☞ Mô hình sửa sai: con người thường học trong quá trình sản xuất. Vì thế, những lỗi mắc phải sẽ ngày càng ít đi theo thời gian. Trong trường hợp này, phương sai của các quan sát được kì vọng là sẽ nhỏ dần theo thời gian ⇒ phương sai sai số thay đổi.

☞ Do bản chất của các mối liên hệ kinh tế: chẳng hạn khi thu nhập tăng người ta có nhiều lựa chọn hơn trong phạm vi tiêu dùng thu nhập đó. Trong hồi qui tiết kiệm theo thu nhập ta thấy khi thu nhập tăng thì σ^2_i cũng tăng theo do người ta có nhiều lựa chọn hơn cho hành vi tiết kiệm.

☞ Do kỹ thuật thu thập số liệu ngày càng được nâng cao nên σ^2_i cũng giảm dần

☞ Trong mẫu có thể chứa những giá trị ngoại lai của các biến số.

☞ Do việc định dạng mô hình (dạng hàm số, số biến số trong mô hình) sai nên phương sai của sai số cũng thay đổi. Ví dụ, mô hình bỏ sót một số biến hoặc biến đổi các biến số trong mô hình không phù hợp…

6.2. HẬU QUẢ CỦA PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI

Xét mô hình $Y_i = \beta_1 + \beta_2 X_i + u_i$

Có: ☞ $Var(\hat{\beta}_2) = E\left[\hat{\beta}_2 - E(\hat{\beta}_2)\right]^2$

$$= E\left[\hat{\beta}_2 - \beta_2\right]^2$$

từ (2.10) có : ☞ $\hat{\beta}_2 = \beta_1 \sum_{i=1}^{n} k_i + \beta_2 \sum_{i=1}^{n} k_i X_i + \sum_{i=1}^{n} k_i u_i$ với $k_i = \dfrac{x_i}{\sum_{i=1}^{n} x_i^2}$

$$= \beta_2 + \sum_{i=1}^{n} k_i u_i$$

$$\Rightarrow Var(\hat{\beta}_2) = E\left[\sum_{i=1}^{n} k_i u_i\right]^2 = E\left[k_1^2 u_1^2 + k_2^2 u_2^2 + ... + k_n^2 u_n^2 + 2k_1 k_2 u_1 u_2 + ... + 2k_{n-1} k_n u_{n-1} u_n\right]$$

Do giả thiết không có tự tương quan nên $E(u_1 u_2) = 0$

$$Var(\hat{\beta}_2) = E\left[k_1^2 u_1^2 + k_2^2 u_2^2 + ... + k_n^2 u_n^2\right]$$

$$= k_1^2 E(u_1^2) + k_2^2 E(u_2^2) + ... + k_n^2 E(u)_n^2$$

$$= k_1^2 \sigma_1^2 + k_2^2 \sigma_2^2 + ... + k_n^2 \sigma_n^2$$

$$= \sum_{i=1}^{n} k_i^2 \sigma_i^2$$

$$\Rightarrow Var(\hat{\beta}_2) = \frac{\sum_{i=1}^{n} x_i^2 \sigma_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \qquad (6.1)$$

Công thức tính phương sai của $\hat{\beta}_2$ trong trường hợp phương sai sai số thay đổi cho bởi (6.1) rõ ràng khác với công thức tính phương sai của $\hat{\beta}_2$ trong trường hợp phương sai đồng nhất: $\text{var}(\hat{\beta}_2) = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$

Người ta chứng minh được rằng:

- Các ước lượng β nhận được vẫn không chệch, tuyến tính nhưng mất tính hiệu quả khi có hiện tượng phương sai số thay đổi (không còn là ước lượng có phươn sai nhỏ nhất nữa).

- Ước lượng của các phương sai sẽ bị chệch (do các phần mềm thống kê đều áp dụng công thức $\text{var}(\hat{\beta}_2) = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$ để tính phương sai cho ước lượng), như vậy khi kiểm định F và T mất hiệu lực.

6.3. CÁC BIỆN PHÁP PHÁT HIỆN PHƯƠNG SAI CỦA SAI SỐ THAY ĐỔI

6.3.1. Bản chất của vấn đề nghiên cứu

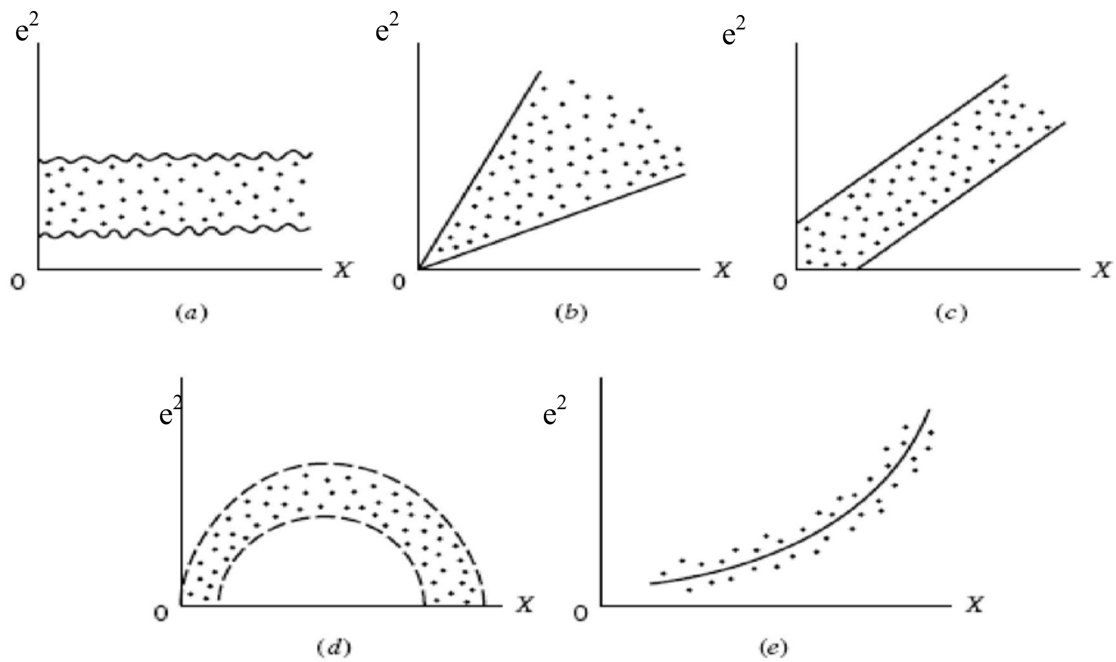
Bản chất của vấn đề nghiên cứu cũng gợi ý cho ta rằng có thể xảy ra hiện tượng phương sai sai số thay đổi hay không. Trên thực tế, ở các số liệu chéo liên quan đến các đơn vị không thuần nhất hay xảy ra hiện tượng phương sai sai số thay đổi.

6.3.2. Đồ thị phần dư

Do trên thực tế, chúng ta không có giá trị thực của sai số ngẫu nhiên u_i nên ta có thể sử dụng phần dư e_i để đại diện cho chúng nếu kích kỡ mẫu đủ lớn.

Ta có thể vẽ đồ thị e_i^2 theo X_i hoặc theo Y_i.

Chú ý: Trong trường hợp mô hình có nhiều biến độc lập có thể vẽ lần lượt phần dư phụ thuộc vào từng biến độc lập.



Hình 6.3

Hình 6.3a cho thấy phương sai sai số đồng đều. Các hình 6.3b-6.3e thể hiện phương sai sai số thay đổi.

6.3.3. Kiểm định Park

Park đã công thức hoá phương pháp đồ thị bằng giả thiết σ_i^2 là một hàm của biến độc lập. Hàm này có dạng

Trong đó: v_i là sai số ngẫu nhiên.

Ước lượng mô hình: $\ln\sigma_i^2 = \ln\sigma^2 + \beta\ln X_i + v_i$ (X>0)

Kiểm định giả thiết

H_o: phương sai sai số đồng đều $\Leftrightarrow \beta = 0$

H_1: Phương sai sai số thay đổi $\Leftrightarrow \beta \neq 0$

Vì σ_i^2 chưa biết nên Park gợi ý sử dụng e_i^2 là đại diện cho σ_i^2. Như vậy, quá trình kiểm định sẽ được thực hiện như sau:

Bước 1: Dùng OLS để ước lượng mô hình ban đầu $\Rightarrow$ phần dư $e_i \Rightarrow e_i^2 \Rightarrow \ln e_i^2$

Bước 2: ước lượng mô hình :

$$\ln e_i^2 = \ln\sigma^2 + \beta\ln X_i + v_i$$

$$\ln e_i^2 = \beta_0 + \beta\ln X_i + v_i \text{ với } \beta_0 = \ln\sigma^2 \qquad (6.2)$$

Bước 3: Kiểm định giả thiết

Tính thống kê T, $t = \dfrac{\hat\beta}{Se(\hat\beta)}$

Nếu $|t| > t_{\alpha/2}$:Bác bỏ giả thiết $H_o \Rightarrow$ phương sai sai số thay đổi

Nếu $|t| < t_{\alpha/2}$:Không đủ cơ sở bác bỏ giả thiết $H_o \Rightarrow$ phương sai sai số đồng đều

6.3.4. Kiểm định Glejer

* Kiểm định Glejer cũng có cùng tư tưởng với kiểm định Park: σ_i^2 là một hàm của biến độc lập. Glejer hồi qui $|e_i|$ phụ thuộc vào các biến số độc lập và theo kinh nghiệm của ông, hồi qui có thể thực hiện với một trong số các dạng hàm sau:

$$|e_i| = \beta_1 + \beta_2 X_i + v_i$$

$$|e_i| = \beta_1 + \beta_2\frac{1}{X_i} + v_i$$

$$|e_i| = \beta_1 + \beta_2\frac{1}{\sqrt{X_i}} + v_i$$

$$|e_i| = \beta_1 + \beta_2\sqrt{X_i} + v_i$$

$$|e_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i$$

$$|e_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i$$

* Kiểm định giả thiết

H_o: phương sai sai số đồng đều $\Leftrightarrow \beta_2 = 0$

H_1: Phương sai sai số thay đổi $\Leftrightarrow \beta_2 \neq 0$

* Các bước:

Bước 1: Dùng OLS để ước lượng mô hình ban đầu $\Rightarrow$ phần dư $e_i \Rightarrow |e_i|$

Bước 2: Ước lượng một trong các dạng mô hình ở trên

Bước 3: Kiểm định (tương tự như Kiểm định Park)

6.3.5. Kiểm định WHITE

Xét mô hình: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$

Các bước thực hiện kiểm định White

Bước 1: Ước lượng mô hình xuất pháp bằng OLS $\Rightarrow$ phần dư $e_i \Rightarrow e_i^2$

Bước 2: Dùng OLS để ước lượng mô hình:

$$e_i^2 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_2^2 + \alpha_5 X_3^2 + \alpha_6 X_2 X_3 + v_i \quad (6.3)$$

Trong đó : v là sai số ngẫu nhiên

Chú ý:

Ước lượng mô hình $\Rightarrow R^2$

Bước 3: Kiểm định giả thiết

H_o: phương sai sai số đồng đều $\Leftrightarrow R^2 = 0$ ($\alpha_2 = \ldots = \alpha_6 = 0$)

H_1: Phương sai sai số thay đổi $\Leftrightarrow R^2 > 0$ ($\exists ! \alpha_i \neq 0$, i = 2,… 6)

Có thể chỉ ra rằng nR^2 phân phối xấp xỉ $\chi^2(df)$, df bằng số hệ số trong mô hình (6.3) không kể hệ số chặn.

Nếu $nR^2 > \chi^2_\alpha(df)$: giả thiết Ho bị bác bỏ

Nếu $nR^2 < \chi^2_\alpha(df)$: không đủ cơ sở bác bỏ giả thiết Ho

6.3.6. Kiểm định dựa trên biến phụ thuộc

Kiểm định này dựa trên ý tưởng sau:

$$\sigma_i^2 = \alpha_1 + \alpha_2 E(Y_i) + U_i$$

Do σ_i^2 và $E(Y_i)$ đều chưa biết nên ta sử dụng các ước lượng của chúng tương ứng là: e_i^2 và $\hat{Y}_i$.

Bước 1: Ước lượng mô hình ban đầu bằng phương pháp OLS $\Rightarrow e_i$ và $\hat{Y}_i \Rightarrow e_i^2$ và $\hat{Y}_i^2$

Bước 2: Ước lượng mô hình sau bằng OLS:

$e_i^2 = \alpha_1 + \alpha_2 \hat{Y}_i^2 + v_i$

Từ kết quả ước lượng thu được R^2 tương ứng.

Bước 3: Có thể sử dụng hai tiêu chuẩn sau để kiểm định giả thiết.

H_o: phương sai sai số đồng đều $\Leftrightarrow R^2 = 0$

H_1: Phương sai sai số thay đổi $\Leftrightarrow R^2 > 0$

a.Tiêu chuẩn χ^2

nR^2 có phân bố xấp xỉ $\chi^2(1)$. Nếu nR^2 lớn hơn $\chi^2_\alpha(1)$ thì Ho bị bác bỏ. Trường hợp ngược lại thì kết luận không có cơ sở bác bỏ giả thiết Ho.

b. Tiêu chuẩn F

$$F = \frac{R^2}{1-R^2}.\frac{n-2}{1} = \left(\frac{\hat{\alpha}_2^2}{se(\hat{\alpha}_2^2)}\right)^2 \text{ có phân bố } F_\alpha(1,n\text{-}2)$$

Nếu F> $F_\alpha(1,n\text{-}2)$ thì hệ số $\alpha_2 \neq 0$, có nghĩa là Ho bị bác bỏ, hay phương sai sai số thay đổi. Ngược lại, phương sai của sai số đồng đều.

6.4. CÁC BIỆN PHÁP KHẮC PHỤC

Khi mô hình có hiện tượng phương sai của sai số thay đổi các ước lượng OLS vẫn là các ước lượng không chệch và tính vững nhưng ước lượng đó không còn là ước lượng hiệu quả nữa. Vì thế phải tìm cách để khắc phục là điều cần thiết. Việc khắc phục chia làm 2 trường hợp: biết σ_i^2 và chưa biết σ_i^2.

Xét mô hình: $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ (6.4)

$Var(u_i) = \sigma_i^2$

6.4.1. Trường hợp đã biết σ_i^2

Khi đã biết σ_i^2 ta có thể khắc phục bằng cách sử dụng phương pháp bình phương nhỏ nhất có trọng số (WLS: Weighted Least Squares).

a. Phương pháp bình phương nhỏ nhất tổng quát (GLS: General Least Squares)

Khi tiến hành ước lượng, về mặt ý tưởng, ta mong muốn các quan sát thuộc nhóm có giá trị của biến phụ thuộc biến động lớn so với trung bình của tổng thể nhận được trọng số bé hơn so với các quan sát thuộc nhóm có giá trị của biến phụ thuộc biến động ít so với trung bình của tổng thể. Việc này sẽ giúp ta ước lượng hàm hồi qui tổng thể (PRF) chính xác hơn.

Xét mô hình: $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ (6.4)

Có $var(u_i) = \sigma_i^2$

(6.4) có thể viết dưới dạng $Y_i = \beta_1 X_{0i} + \beta_2 X_{2i} + U_i$ với $X_{0i} = 1$ (6.5)

Chia các số hạng của mô hình (6.5) cho σ_i ta được mô hình hiệu chỉnh sau

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{X_{0i}}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{U_i}{\sigma_i} \tag{6.6}$$

(6.6) $\Leftrightarrow Y_i^* = \beta_1 X_{0i}^* + \beta_2 X_i^* + v_i$ trong đó $v_i = \frac{U_i}{\sigma_i}$; $Y_i^* = \frac{Y_i}{\sigma_i}$; $X_{0i}^* = \frac{X_{0i}}{\sigma_i}$; $X_i^* = \frac{X_i}{\sigma_i}$

Ta có: var(v_i)= $\text{var}\left(\frac{U_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2}\text{var}(U_i) = 1 \ \forall i$ vì var(u_i)= σ_i^2

Kết luận: Mô hình với các biến : Y^*, X_0, X^* là mô hình có phương sai đồng đều.

$\Rightarrow$ Phương pháp bình phương nhỏ nhất tổng quát (GLS) là phương pháp OLS dựa trên các biến số đã được biến đổi để thoả mãn giả thiết của OLS

* Phương pháp OLS thu được $\hat{\beta}$: minRSS = min $\sum_{i=1}^{n} e_i^2$

* Phương pháp WLS thu được $\hat{\beta}^*$: min RSS = min $\sum_{i=1}^{n} w_i e_i^2$

$$= \min \sum_{i=1}^{n} w_i \left(Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* \mathbf{X}_i\right)^2 = f(\beta_1^*, \beta_2^*)$$

Trong đó: $w_i = \frac{1}{\sigma_i^2}$

$$\frac{\partial f}{\partial \hat{\beta}_1^*} = \sum_{i=1}^{n} 2 w_i \left(Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* \mathbf{X}_i\right)(-1) = 0$$

$$\hat{\beta}_1^* \sum_{i=1}^{n} w_i + \hat{\beta}_2^* \sum_{i=1}^{n} w_i \mathbf{X}_i = \sum_{i=1}^{n} w_i Y_i$$

$$\frac{\partial f}{\partial \hat{\beta}_2^*} = \sum_{i=1}^{n} 2 w_i \left(Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* \mathbf{X}_i\right)(-X_i) = 0$$

$$\hat{\beta}_1^* \sum_{i=1}^{n} w_i X_i + \hat{\beta}_2^* \sum_{i=1}^{n} w_i \mathbf{X}_i^2 = \sum_{i=1}^{n} w_i X_i Y_i$$

$$\Rightarrow \quad \hat{\beta}_1^* = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i} - \hat{\beta}_2^* \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

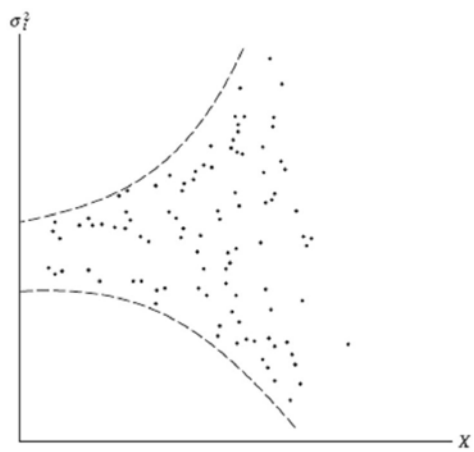
$$\hat{\beta}_2^* = \frac{\sum_{i=1}^{n} w_i \left(\sum_{i=1}^{n} w_i X_i Y_i\right) - \left(\sum_{i=1}^{n} w_i X_i\right)\left(\sum_{i=1}^{n} w_i Y_i\right)}{\left(\sum_{i=1}^{n} w_i\right)\left(\sum_{i=1}^{n} w_i X_i^2\right) - \left(\sum_{i=1}^{n} w_i X_i\right)^2}$$

6.4.2. Trường hợp σ_i^2 chưa biết

Trong nghiên cứu kinh tế rất ít khi ta biết được σ_i^2. Do đó, người ta thường đưa ra các giải thiết về σ_i^2 và thực hiện biến đổi mô hình gốc để phương sai sai số không đổi. Áp dụng phương pháp OLS cho các mô hình đã được biến đổi này đồng nghĩa với việc áp dụng phương pháp WLS cho mô hình gốc với trọng số tương ứng rút ra từ các phép biến đổi.

Bằng phương pháp đồ thị hoặc cách tiếp cận Park hoặc Glejser, ta có thể đưa ra các giả thiết sau về σ_i^2

a. Giả thiết 1: $\sigma_i^2 = \sigma^2 X_i^2$



Biến đổi mô hình gốc bằng cách chia 2 vế của mô hình gốc (6.4) cho X_i ($X_i \neq 0$) ta được:

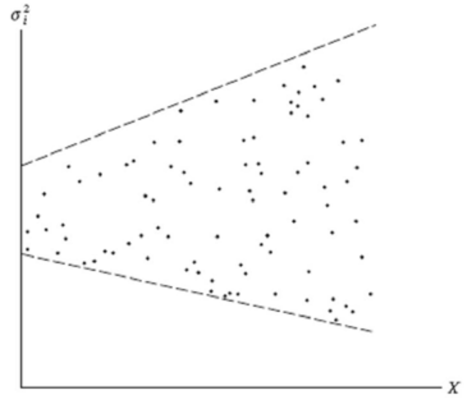
$$\frac{Y_i}{X_i} = \beta_1 \frac{1}{X_i} + \beta_2 + \frac{U_i}{X_i}$$

$$\Rightarrow \quad Y_i^* = \beta_1 X_i^* + \beta_2 + v_i \text{ trong đó, } v_i = \frac{U_i}{X_i};\ Y_i^* = \frac{Y_i}{X_i};\ X_i^* = \frac{1}{X_i}$$

Ta có: $\mathrm{var}(v_i) = \mathrm{var}\left(\frac{U_i}{X_i}\right) = \frac{1}{X_i^2}\mathrm{var}(U_i) = \sigma^2 \quad \forall i$

Chú ý: Vai trò của hệ số chặn và hệ số góc trong mô hình ban đầu và mô hình sau khi biến đổi đã đổi chỗ cho nhau.

b. Giả thiết 2: $\sigma_i^2 = \sigma^2 X_i$ ***(với $X_i > 0$)***



Biến đổi mô hình gốc cách chia 2 vế của mô hình gốc (6.4) cho $\sqrt{X_i}$

$$\frac{Y_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + \frac{U_i}{\sqrt{X_i}}$$

$$\Rightarrow \quad Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + v_i$$

trong đó $v_i = \frac{U_i}{\sqrt{X_i}}$; $Y_i^* = \frac{Y_i}{\sqrt{X_i}}$; $X_{1i}^* = \frac{1}{\sqrt{X_i}}$; $X_{2i}^* = \sqrt{X_i}$

Ta thấy, $\mathrm{var}(v_i) = \sigma^2 \ \forall i$

Chú ý: Trong trường hợp này, mô hình biến đổi không có hệ số chặn.

c. Giả thiết 3: $\sigma_i^2 = \sigma^2 \left(E(Y_i)\right)^2$

Thực hiện phép biến đổi mô hình gốc bằng cách chia 2 vế của mô hình gốc (6.4) cho $E(Y_i)$ ta được:

$$\frac{Y_i}{E(Y_i)} = \beta_1 \frac{1}{E(Y_i)} + \beta_2 \frac{X_i}{E(Y_i)} + \frac{U_i}{E(Y_i)}$$

$$= \beta_1 \frac{1}{E(Y_i)} + \beta_2 \frac{X_i}{E(Y_i)} + v_i \qquad \text{với } v_i = \frac{U_i}{E(Y_i)}$$

$\Rightarrow \mathrm{var}(v_i) = \sigma^2 \ \forall i$

Do $E(Y_i)$ chưa biết nên chúng ta dùng ước lượng của nó là $\hat{Y}_i$ bằng cách thực hiện các bước sau:

Bước 1: Dùng OLS ước lượng mô hình gốc (6.4), từ đó nhận được $\hat{Y}_i$ tương ứng.

Bước 2: Sử dụng $\hat{Y}_i$ để biến đổi mô hình gốc như sau:

$$\frac{Y_i}{\hat{Y}_i} = \beta_1 \frac{1}{\hat{Y}_i} + \beta_2 \frac{X_i}{\hat{Y}_i} + \frac{U_i}{\hat{Y}_i}$$

Phép biến đổi này được sử dụng trong thực hành khi cỡ mẫu tương đối lớn. Vì khi đó ước lượng $\hat{Y}_i$ gần với $E(Y_i)$ hơn.

d. Biến đổi loga dạng hàm

Đôi khi thay cho việc dự đoán về σ_i^2 người ta xác định hoặc biến đổi dạng hàm

Giả sử thay cho việc ước lượng mô hình gốc chúng ta sẽ ước lượng mô hình:

$$\ln Y_i = \beta_1 + \beta_2 \ln X_{2i} + u_i$$

Phép biến đổi loga sẽ làm giảm sự khác biệt về độ lớn của các biến số trong mô hình. Do đó, nó làm giảm khả năng xảy ra phương sai của sai số thay đổi.

Ví dụ, hai quan sát có giá trị là 8 và 80. Tức là, giá trị của quan sát này gấp 10 lần giá trị của quan sát kia. Nhưng ln(80) = 1.328 và ln(8) = 2.079. Như vậy, sau phép biến đổi loga, giá trị của quan sát này chỉ gấp đôi giá trị quan sát kia.

Một trong những ưu thế khác của phép biến đổi loga là hệ số β_2 là hệ số co dãn của Y đối với X.

Câu hỏi chương 6.

1. Trình bày khái niệm, nguyên nhân và hậu quả, cách phát hiện của hiện tượng phương sai sai số thay đổi.

2. Trình bày cách khắc phục hiện tượng phương sai sai số thay đổi theo giả thiết 1 và giả thiết 2.

3. Từ một mẫu người ta thiết lập được mô hình $\hat{Y}_i = 4,5 + 0,84X_i$. Tính ra phần dư e_i và người ta tính được hạng d_i = hạng $|e_i|$ – hạng X_i. Cho các thông số sau: n = 20, $t_{0,025}(18) = 2,101$; $\sum d_i^2 = 549$. Dùng kiểm định tương quan hạng Spearman để phát hiện xem có hiện tượng phương sai sai số thay đổi trong mô hình không?

4. Từ một mẫu n = 30 quan sát theo mô hình $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ người ta sắp xếp các quan sát theo thứ tự tăng dần của biến X. Bỏ 6 quan sát ở giữa, chia số mẫu còn lại thành 2 nhóm gọi là nhóm 1 và nhóm 2. Ước lượng OLS cho mỗi nhóm trên ta có các thông số sau:

Nhóm 1: $TSS_1 = 5564$; $ESS_1 = 2892$.

Nhóm 2: $TSS_2 = 19913$; $ESS_2 = 6331$.

Dùng kiểm định Goldfeld để phát hiện xem có hiện tượng phương sai sai số thay đổi trong mô hình chung hay không biết rằng $F_{0,05}(10,10) = 2,98$,

CHƯƠNG 7. CHỌN MÔ HÌNH VÀ KIỂM ĐỊNH VIỆC CHỌN MÔ HÌNH

Trọng tâm của chương 7 là việc đánh giá và lựa chọn mô hình hồi quy theo một số tiêu chí được áp dụng rộng rãi trong thực nghiệm, chương này cũng trình bày một số kiểm định liên quan đến sai số đặc trưng trọng việc chỉ định mô hình như việc kiểm định thừa biến, thiếu biến, kiểm định việc lựa chọn dạng hàm của mô hình.

Nội dung cơ bản của chương này bao gồm:

O Các thuộc tính tốt của một mô hình

- Tính kiệm

- Tính thống nhất

- Tính thích hợp

- Tính vững về mặt lý thuyết

- Khả năng về dự đoán

O Các loại sai lầm khi chỉ định

- Bỏ sót biến thích hợp

- Đưa vào biến không thích hợp

- Chọn dạng hàm không đúng

O Phát hiện những sai lầm chỉ định-kiểm định

7.1. CÁC SAI LẦM ĐỊNH DẠNG

7.1.1. Các thuộc tính của một mô hình tốt

☞ Tính kiệm: Mô hình là sự biểu diễn đơn giản nhưng hoàn chỉnh của hiện tượng. Mô hình càng đơn giản càng tốt.

☞ Tính thống nhất: nghĩa là với cùng 1 tập số liệu, giá trị các tham số phải thống nhất

☞ Tính vững về mặt lý thuyết: Mô hình phải có tính phù hợp về mặt lý thuyết, không được sai phạm các vấn đề kinh tế cơ bản. Ví dụ, hệ số tiêu dùng biên phải thuộc [0,1].

☞ Tính thích hợp: Vì mục đích của mô hình là giải thích sự thay đổi của biến phụ thuộc do sự thay đổi của các biến độc lập nên R^2, $\overline{R}^2$ cao là một tiêu chuẩn cần thiết

☞ Khả năng dùng cho dự báo của mô hình: dự báo dựa trên mô hình phải phù hợp với thực tế.

7.1.2. Các sai lầm định dạng

Định dạng mô hình nghĩa là xác định mô hình có chứa những biến số nào và dạng của quan hệ giữa biến phụ thuộc với các biến độc lập trong mô hình. Khi định dạng mô hình có thể gặp một số sai lầm sau.

a. Mô hình bỏ sót biến thích hợp

Giả sử mô hình đúng có dạng: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ (8.1)

Nhưng vì một lý do nào đó, chúng ta ước lượng mô hình: $Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i$ (8.2)

$\Rightarrow v_i = \beta_2 X_{3i} + u_i$

Tức là, trong quá trình xây dựng mô hình ta đã bỏ sót biến X_3 mà đáng lẽ phải có mặt trong mô hình.

Việc bỏ sót như vậy sẽ gây ra một số hậu quả:

✯ Trường hợp X_2 và X_3 có tương quan với nhau ($r_{23} \neq 0$) $\Rightarrow$ bỏ sót X_3 sẽ khiến X_2 và v_i có quan hệ tương quan $\Rightarrow$ vi phạm giả thiết (1) của OLS (các biến giải thích và sai số ngẫu nhiên không có quan hệ tương quan) $\Rightarrow$ Các tham số ước lượng được bị chệch ngay cả khi kỡ mẫu lớn

☞Ước lượng của α_2 được tính theo công thức sau:

$$\hat{\alpha}_2 = \frac{\sum_{i=1}^{n} x_{2i} y_i}{\sum_{i=1}^{n} x_{2i}^2} = \frac{\sum_{i=1}^{n} x_{2i}(\beta_2 x_{2i} + \beta_3 x_{3i} + u_i)}{\sum_{i=1}^{n} x_{2i}^2} = \frac{\beta_2 \sum_{i=1}^{n} x_{2i}^2 + \beta_3 \sum_{i=1}^{n} x_{2i} x_{3i} + \sum_{i=1}^{n} x_{2i} u_i}{\sum_{i=1}^{n} x_{2i}^2}$$

$$\hat{\alpha}_2 = \beta_2 + \beta_3 \frac{\sum_{i=1}^{n} x_{2i} x_{3i}}{\sum_{i=1}^{n} x_{2i}^2} + \frac{\sum_{i=1}^{n} x_{2i} u_i}{\sum_{i=1}^{n} x_{2i}^2}$$

$$\Rightarrow E(\hat{\alpha}_2) = \beta_2 + \beta_3 \frac{\sum_{i=1}^{n} x_{2i} x_{3i}}{\sum_{i=1}^{n} x_{2i}^2}$$

$\Rightarrow \beta_3 \dfrac{\sum_{i=1}^{n} x_{2i} x_{3i}}{\sum_{i=1}^{n} x_{2i}^2}$ là phần chệch trong ước lượng của α.

$$\beta_3 \frac{\sum_{i=1}^{n} x_{2i} x_{3i}}{\sum_{i=1}^{n} x_{2i}^2} = \beta_3 \frac{\sum_{i=1}^{n} x_{2i} x_{3i} / (n-1)}{\sum_{i=1}^{n} x_{2i}^2 / (n-1)} = \beta_3 \frac{\overline{\text{cov}}(X_2, X_3)}{\overline{\text{var}}(X_2)}$$

Từ công thức phần chệch trên cho ta thấy, chệch sẽ không đáng kể khi:

+ $\beta_3 \to 0 \Rightarrow X_3$ không nên có mặt trong mô hình

+ $\overline{\text{cov}}(X_2, X_3) \to 0 \Rightarrow$ rất ít trong thực tế vì các biến số kinh tế thường có tương quan chặt chẽ.

☞ $E(\hat{\alpha}_1) = E(\bar{Y} - \hat{\alpha}_2\bar{X}_2) = E(\hat{\beta}_1 + \hat{\beta}_2\bar{X}_2 + \hat{\beta}_3\bar{X}_3 - \hat{\alpha}_2\bar{X}_2)$

$= \beta_1 + (\beta_2 - \alpha_2)\bar{X}_2 + \beta_3\bar{X}_3 \neq \beta_1$

✯ Trường hợp X_2 và X_3 không có tương quan với nhau ($r_{23} = 0$) $\Rightarrow$ ước lượng của α_2 không bị chệch (như phần trình bầy trên) nhưng ước lượng α_1 vẫn bị chệch.

$E(\hat{\alpha}_1) = \beta_1 + \beta_3\bar{X}_3 \neq \beta_1$

✯ Phương sai của yếu tố ngẫu nhiên, σ^2 , bị ước lượng chệch ($\hat{\sigma}^2 = \dfrac{RSS}{df}$, trong hai mô hình RSS và df đều khác nhau. $RSS_1 < RSS_2$ và $df_1 = n - 3 < df_2 = n - 2$ (do mô hình (8.1) có nhiều biến giải thích hơn)

✯ $\text{var}(\hat{\alpha}_2) = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_{2i}^2}$ là ước lượng chệch của phương sai của ước lượng đúng $\hat{\beta}_2$,

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{2i}^2(1 - r_{23}^2)} = \frac{\sigma^2}{\sum_{i=1}^{n} x_{2i}^2} VIF$$

$0 < r^2_{23} < 1 \Rightarrow \text{var}(\hat{\alpha}_2) < \text{var}(\hat{\beta}_2) \Rightarrow$ ở đây có sự đánh đổi giữa tính chệch và tính hiệu quả

✯ Từ các hậu quả trên, khoảng tin cậy và kiểm định giả thiết sẽ dẫn tới những kết luận sai lầm về ý nghĩa thống kê của tham số ước lượng được

✯ Giá trị dự báo cũng như khoảng tin cậy được dự báo không đáng tin cậy do mô hình sai.

b. Mô hình chứa biến số không thích hợp

Giả sử mô hình đúng là $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ (8.3)

Nhưng trên thực tế, chúng ta ước lượng mô hình: $Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + v_i$ (8.4)

Hậu quả của việc trong mô hình chứa biến không thích hợp như sau:

✯ Các tham số ước lượng từ mô hình vẫn là ước lượng không chệch và vững. Nghĩa là $E(\hat{\alpha}_1) = \beta_1$; $E(\hat{\alpha}_2) = \beta_2$ (Vì X_3 là biến không thích hợp trong mô hình nên $E(\hat{\alpha}_3) = 0$)

✯ Phương sai của yếu tố nhẫu nhiên, σ^2, vẫn được ước lượng đúng

✯ Khoảng tin cậy và kiểm định giả thiết vẫn hợp lệ và đáng tin cậy

✯ Tuy nhiên, mất tính hiệu quả do ước lượng thu được của phương sai không phải là nhỏ nhất var($\hat{\alpha}_i$) > var($\hat{\beta}_i$).

$$\text{Ví dụ, } \operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{2i}^2} > \operatorname{var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{2i}^2(1-r_{23})}$$

7.1.3. Dạng h□m khụng đỳng

Một loại sai lầm chỉ định lớn nhất của mô hình đó là dạng hàm sai. Nếu mắc phải sai lầm này các hệ số thu được từ mô hình hồi quy sai sẽ không chính xác vì bị đánh giá gián tiếp thông qua thông tin khác của mô hình, do đó, sẽ cho kết luận không chính xác về ảnh hưởng của biến độc lập đến biến phụ thuộc.

Giả sử mô hình đúng là: $Y_t = \beta_1 + \beta_2X_{2t} + \beta_3X_{3t} + u_t$ (8.5)

Nhưng mô hình sai được ước lượng có dạng:

$\ln(Y_t) = \beta_1 + \beta_2 \ln(X_{2t}) + \beta_2X_{3t} + u_t$

(8.6)

7.2. CÁCH PHÁT HIỆN CÁC SAI LẦM ĐỊNH DẠNG

Ta cũng biết những hậu quả do sai lầm định dạng gây ra, vấn đề đặt ra là cách phát hiện và tìm biện pháp khắc phục chúng.

7.2.1. Phát hiện mô hình chứa biến không thích hợp

Giả sử, ta có mô hình hồi qui:

$Y_t = \beta_1 + \beta_2X_{2t} + \beta_3X_{3t} + \beta_4X_{4t} + \beta_5X_{5t} + u_t$ (8.7)

☞ Nếu ta không biết biến X_5 có thực sự cần thiết trong mô hình không thì dùng kiểm định t để kiểm định về sự bằng 0 của hệ số biến tương ứng.

H_0: $\beta_5 = 0$

H_1: $\beta_5 \neq 0$

Nếu giả thiết H_0 được chấp nhận, chứng tỏ X_5 không cần thiết cho mô hình

☞ Nếu có hơn một biến, ví dụ X_4 và X_5, bị nghi ngờ không thích hợp với mô hình thì dùng kiểm định F (Hồi qui có điều kiện ràng buộc)

H_0: $\beta_4 = \beta_5 = 0$

H_1: β_4, β_5 không đồng thời bằng 0

Nếu H_0 được chấp nhận, chứng tỏ X_4, X_5 không cần thiết cho mô hình.

☞ Trong các trường hợp giả thiết về sự bằng 0 của các tham số ước lượng thì việc bỏ hay giữ lại các biến này cần được cân nhắc kỹ. Vì

+ Khi bỏ đi một số biến số có thể dẫn đến một số giả thiết khác của mô hình không được đảm bảo.

+ Nếu trong các lý thuyết khẳng định sự có mặt của các yếu tố này trong mô hình thì dù giả thiết hệ số ước lượng của các yếu tố này có bằng 0 được chấp nhận cũng không nên loại bỏ các biến số này khỏi mô hình.

7.2.2. Phát hiện mô hình bỏ sót biến thích hợp

Xét mô hình: $Y_i = \beta_1 + \beta_2 X_i + u_i$

Giả sử ta có nghi nghờ mô hình bị bỏ sót biến z, khi đó tuỳ thuộc vào việc có hay không có các quan sát tương ứng đối với biến Z để kiểm định việc bỏ sót hay không bỏ sót biến này.

+ *Trường hợp 1*: có quan sát đối với biến Z

Ta ước lượng mô hình: $Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + u_i$

Sau đó kiểm định giả thiết:

H_0: $\beta_3 = 0$

H_1 : $\beta_3 \neq 0$

Nếu kết quả cho thấy không đủ cơ sở bác bỏ giả thiết H_0, nghĩa là không bỏ sót biến Z. Còn ngược lại thì mô hình bỏ sót biến Z.

+ *Trường hợp 2*: chưa có quan sát đối với biến Z: khi đó người ta sẽ tìm một biến đại diện cho Z, ví dụ Z', và ước lượng lại mô hình có biến Z', từ đó có kết luận về việc bỏ sót hay không bỏ sót biến Z.

a. Kiểm định Ramsey

Ý tưởng của kiểm định Ramsay như sau. Sau khi hồi qui mô hình gốc ta thu được $\hat{Y}$ và phần dư e. Nếu đồ thị của phần dư e theo $\hat{Y}$ có một dáng điệu đặc biệt nào đó thể hiện mối quan hệ giữa e và $\hat{Y}$ thì chứng tỏ rằng nếu ta đưa các dạng khác nhau của $\hat{Y}$ vào phương trình hồi qui với tư cách biến giải thích sẽ làm tăng R^2. Trường hợp R^2 tăng có ý nghĩa về mặt thống kê, gợi ý rằng dạng hàm ban đầu là sai.

Kiểm định Ramsey được tiến hành theo các bước sau:

Giả sử mô hình được ước lượng là $Y = \beta_1 + \beta_2 X + u$ (R)

Bước 1: ước lượng mô hình xuất phát, ta thu được $\hat{Y}$ (giá trị biến phụ thuộc nhận được từ hàm hồi quy mẫu) và R^2_R

Tính $\hat{Y}^2, \hat{Y}^3 \ldots$

Bước 2: Ước lượng mô hình: $Y = \beta_1 + \beta_2 X + \beta_3 \hat{Y}^2 + \beta_4 \hat{Y}^3 + \ldots + v$ (UR)

thu được R^2_{UR}

Bước 3: Kiểm định giả thiết :

H_0: $\beta_3 = \beta_4 = \ldots = 0$

$H_1 : \beta_i \neq 0$ (i=3,4,...)

Tính thống kê F:

$$F = \frac{\left(R^2_{UR} - R^2_R\right)/m}{\left(1 - R^2_{UR}\right)/(n-k)}$$

Trong đó: m là các số hạng mới đưa vào mô hình xuất phát

k là số các hệ số của mô hình mới

Với n khá lớn F có phân bố F(m, n-k).

Nếu F < F(m, n-k) $\Rightarrow$ không đủ cơ sở bác bỏ $H_0 \Rightarrow$ mô hình không bỏ sót biến số.

Nếu F > F(m, n-k) $\Rightarrow$ bác bỏ H_0, mô hình đã bỏ sót biến số nào đó.

b. Kiểm định bằng nhân tử Lagrange (LM)

- Bước 1: Ước lượng mô hình xuất phát $Y_i = \beta_1 + \beta_2 X_i + u_i$

ta thu được $\hat{Y}_i$ và phần dư e_i

- Bước 2: Ước lượng mô hình sau:

$$e_i = \beta_1 + \beta_2 X_i + \beta_3 \hat{Y}^2 + \beta_4 \hat{Y}^3 + ... + v_i$$

Kết quả ước lượng thu được R^2. Với n khá lớn nR^2 có phân bố xấp xỉ $\chi^2(m)$, trong đó m là số các biến số $\hat{Y}^2$, $\hat{Y}^3$,..., $\hat{Y}^p$.

Dựa vào giá trị nR^2 kết luận về kiểm định:

H_0: $\beta_3 = \beta_4 = ... = 0 \Leftrightarrow (R^2 = 0)$, mô hình đúng

$H_1 : \beta_i \neq 0$ (i=3,4,...) $\Leftrightarrow (R^2 \neq 0)$, mô hình sai

c. Kiểm định Durbin-Watson d

Trong trường hợp dạng hàm sai do bỏ sót biến, phần dư u sẽ chứa những thông tin cần thiết để giải thích sự thay đổi của biến phụ thuộc. Và khi đó, đồ thị của u sẽ có những dáng điệu nhất định (không còn ngẫu nhiên). Vì vậy, ta có thể sử dụng thống kê Durbin-Watson, được tính toán từ các u_i để phát hiện sai lầm định dạng này.

Giả sử mô hình ban đầu $Y = \beta_1 + \beta_2 X + u$

Bước 1: Ước lượng mô hình gốc nhận được phần dư tương ứng.

Bước 2: Nếu ta nghi ngờ đã bỏ sót biến Z thì sắp xếp lại e_i theo giá trị tăng dần của biến Z, trong trường hợp biến Z không có thì sắp xếp e_i theo 1 trong các biến độc lập nào đó.

Bước 3: Tính thống kê d:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2}$$

Bước 4: Kiểm định giả thiết

H_0: Dạng hàm đúng

H_1 : Dạng hàm sai

Dựa vào bảng Durbin-Watson và mức ý nghĩa, ta tìm được các giá trị d_L và d_U tương ứng. Kết luận về dạng hàm sai được khẳng định nếu giá trị d cho thấy tồn tại dấu hiệu của tự tương quan.

* Chú ý: Các tiêu chuẩn Ramsey và LM cũng được sử dụng để kiểm định giả thiết

H_0: Dạng hàm đúng

H_1 : Dạng hàm sai

7.2.3. Kiểm định về tính phân bố chuẩn của U

Giả thiết:

H_0: u có phân bố chuẩn

H_1 : u không có phân bố chuẩn

Để sử dụng các kiểm định T, kiểm định F, χ^2, trong hầu hết các trường hợp chúng ta giả thiết rằng các sai số ngẫu nhiên u_i có phân bố chuẩn. Do tổng thể chưa biết nên ta cũng không biết được u_i và do đó phải đoán nhận thông qua e_i.

Để kiểm định giả thiết e_i có phân bố chuẩn hay không người ta thường dùng tiêu chuẩn χ^2. Còn trong hầu hết các phần mềm kinh tế lượng thường sử dụng kiểm định Jarque-Bera (JB). Thống kê JB đo sự khác nhau của hệ số bất đối xứng và độ nhọn của chuỗi. Thống kế JB được tính như sau:

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right]$$

Trong đó, S là hệ số bất đối xứng, K là hệ số nhọn, n là kích thước mẫu.

Trong trường hợp tổng quát S và K được tính như sau:

$$S = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3 / n}{S_X^3} \qquad K = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4 / n}{S_X^4}$$

Với n khá lớn JB có phân bố xấp xỉ $\chi^2(2)$.

Nếu JB > $\chi^2_\alpha(2)$: Bác bỏ H_0

Nếu JB < $\chi^2_\alpha(2)$: Không đủ cơ sơ bác bỏ H_0.

Câu hỏi chương 7

1. Trình bày các thuộc tính của một mô hình tốt, cho ví dụ

2. Các sai lầm cơ bản khi chọn và kiểm định việc chọn mô hình, cho ví dụ.

PHỤ LỤC

Bảng 1. Phân bố T - student

k/α	0.001	0.002	0.005	0.01	0.02	0.025	0.05	0.1	0.2
1	318.309	159.153	63.657	31.821	15.895	12.706	6.341	3.078	1.376
2	22.327	15.764	9.925	6.695	4.849	4.303	2.920	1.886	1.061
3	10.215	8.053	5.841	4.541	3.482	3.182	2.353	1.638	0.978
4	7.173	5.951	4.604	3.747	2.999	2.776	2.132	1.533	0.941
5	5.893	5.030	4.032	3.365	2.757	2.571	2.015	1.476	0.920
6	5.208	4.524	3.707	3.143	2.612	2.447	1.943	1.440	0.906
7	4.785	4.207	3.499	2.998	2.517	2.365	1.895	1.415	0.896
8	4.501	3.991	3.355	2.896	2.449	2.306	1.860	1.379	0.889
9	4.297	3.835	3.250	2.821	2.398	2.262	1.833	1.383	0.883
10	4.144	3.716	3.169	2.764	2.359	2.228	1.812	1.372	0.879
11	4.025	3.624	3.106	2.718	2.328	2.201	1.796	1.363	0.876
12	3.930	3.550	3.055	2.681	2.303	2.179	1.782	1.356	0.873
13	3.852	3.489	3.012	2.650	2.282	2.160	1.771	1.350	0.870
14	3.787	3.438	2.977	2.624	2.264	2.145	1.761	1.345	0.868
15	3.733	3.395	2.947	2.602	2.249	2.131	1.753	1.341	0.866
16	3.686	3.358	2.921	2.583	2.235	2.120	1.746	1.337	0.865
17	3.646	3.326	2.898	2.567	2.224	2.110	1.740	1.333	0.863
18	3.610	3.298	2.878	2.552	2.214	2.101	1.734	1.330	0.862
19	3.579	3.273	2.861	2.539	2.205	2.093	1.729	1.328	0.861
20	3.552	3.251	2.845	2.528	2.179	2.086	1.725	1.325	0.860
21	3.527	3.231	2.831	2.518	2.189	2.080	1.721	1.323	0.859
22	3.505	3.214	2.819	2.508	2.183	2.074	1.717	1.321	0.858
23	3.485	3.198	2.807	2.500	2.177	2.069	1.714	1.319	0.858
24	3.467	3.183	2.797	2.492	2.172	2.064	1.711	1.318	0.857
25	3.450	3.170	2.787	2.485	2.167	2.060	1.708	1.316	0.856
26	3.435	3.158	2.779	2.479	2.162	2.056	1.706	1.315	0.856
27	3.421	3.147	2.771	2.473	2.158	2.052	1.703	1.314	0.855
28	3.408	3.136	2.763	2.476	2.154	2.048	1.701	1.313	0.855
29	3.396	3.127	2.256	2.426	2.150	2.045	1.699	1.311	0.854

Bảng 1. Phân bố T - student (tiếp theo)

k/α	0.001	0.002	0.005	0.01	0.02	0.025	0.05	0.1	0.2
30	3.385	3.118	2.750	2.457	2.147	2.042	1.697	1.310	0.854
31	3.375	3.109	2.744	2.453	2.144	2.040	1.696	1.309	0.853
32	3.365	3.102	2.738	2.449	2.141	2.037	1.694	1.309	0.853
33	3.356	3.094	2.733	2.445	2.138	2.035	1.692	1.308	0.853
34	3.348	3.088	2.728	2.441	2.136	2.032	1.691	1.307	0.852
35	3.340	3.081	2.724	2.438	2.133	2.030	1.690	1.306	0.852
36	3.333	3.075	2.719	2.434	2.131	2.028	1.688	1.306	0.852
37	3.326	3.070	2.715	2.431	2.129	2.026	1.687	1.305	0.851
38	3.319	3.064	2.712	2.429	2.127	2.024	1.686	1.304	0.851
39	3.313	3.059	2.708	2.426	2.125	2.023	1.685	1.304	0.851
40	3.307	3.055	2.704	2.423	2.123	2.021	1.684	1.303	0.851
41	3.301	3.050	2.701	2.421	2.121	2.020	1.683	1.303	0.850
42	3.296	3.046	2.698	2.418	2.120	2.018	1.682	1.302	0.850
43	3.291	3.042	2.695	2.416	2.118	2.017	1.681	1.302	0.850
44	3.286	3.038	2.692	2.414	2.116	2.015	1.680	1.301	0.850
45	3.281	3.034	2.690	2.412	2.115	2.014	1.679	1.301	0.850
46	3.277	3.030	2.687	2.410	2.114	2.013	1.679	1.300	0.850
47	3.273	3.027	2.685	2.408	2.112	2.012	1.678	1.300	0.849
48	3.269	3.024	2.682	2.407	2.111	2.011	1.677	1.299	0.849
49	3.265	3.021	2.680	2.405	2.110	2.010	1.677	1.299	0.849
50	3.261	3.018	2.678	2.403	2.109	2.009	1.676	1.299	0.849
51	3.258	3.015	2.676	2.402	2.108	2.008	1.675	1.298	0.849
52	3.255	3.012	2.674	2.400	2.107	2.007	1.675	1.298	0.849
53	3.251	3.009	2.672	2.399	2.106	2.006	1.674	1.298	0.848
54	3.248	3.007	2.670	2.397	2.105	2.005	1.674	1.297	0.848
55	3.245	3.005	2.668	2.396	2.104	2.004	1.673	1.297	0.848
56	3.242	3.002	2.667	2.395	2.103	2.003	1.673	1.297	0.848
57	3.239	3.000	2.665	2.394	2.102	2.002	1.672	1.297	0.848
58	3.237	2.998	2.663	2.392	2.101	2.002	1.672	1.296	0.848
59	3.234	2.996	2.662	2.391	2.100	2.001	1.671	1.296	0.848
60	3.232	2.994	2.660	2.390	2.099	2.000	1.671	1.296	0.848
61	3.229	2.992	2.659	2.389	2.099	2.000	1.670	1.296	0.848

Bảng 2. Bảng phân bố F; α = 0.01

k_2/k_1	1	2	3	4	5	6	7	8
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029
9	19.561	8.022	6.992	6.422	6.075	5.802	5.613	5.467
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140
15	8.863	6.359	5.417	4.893	4.556	4.318	4.142	4.004
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791
18	8.258	6.013	5.092	4.579	4.248	4.015	3.841	3.705
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324
26	7.721	5.526	4.673	4.140	3.818	3.591	3.421	3.288
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173

Bảng 2. Bảng phân bố F (tiếp theo); α = 0,01

k_2/k_1	9	10	11	12	13	14	15	16
2	99.388	99.399	99.408	99.416	99.422	99.428	99.433	99.437
3	27.345	27.229	27.133	27.052	26.983	26.924	26.872	26.827
4	14.659	14.546	14.452	14.374	14.307	14.249	14.198	14.154
5	10.158	10.051	9.963	9.888	9.825	9.770	9.722	9.680
6	7.976	7.874	7.790	7.718	7.657	7.605	7.559	7.519
7	6.719	6.620	6.538	6.469	6.410	6.359	6.314	6.275
8	5.911	5.814	5.734	5.667	5.609	5.559	5.515	5.477
9	5.351	5.257	5.178	5.111	5.055	5.005	4.962	4.924
10	4.942	4.849	4.772	4.706	4.650	4.601	4.558	4.520
11	4.632	4.539	4.462	4.397	4.342	4.293	4.251	4.213
12	4.388	4.286	4.220	4.155	4.100	4.052	4.010	3.972
13	4.191	4.100	4.025	3.960	3.905	3.857	3.815	3.778
14	4.030	3.939	3.864	3.800	3.745	3.698	3.656	3.619
15	3.895	3.805	3.730	3.666	3.612	3.564	3.522	3.485
16	3.780	3.691	3.616	3.553	3.498	3.451	3.409	3.372
17	3.682	3.593	3.519	3.455	3.401	3.353	3.312	3.275
18	3.597	3.508	3.434	3.371	3.316	3.269	3.227	3.190
19	3.523	3.434	3.360	3.297	3.242	3.195	3.153	3.116
20	3.457	3.368	3.294	3.231	3.177	3.130	3.088	3.051
21	3.398	3.310	3.236	3.173	3.119	3.072	3.030	2.993
22	3.346	3.258	3.184	3.121	3.067	3.019	2.978	2.941
23	3.299	3.211	3.137	3.074	3.020	2.973	2.931	2.894
24	3.256	3.168	3.094	3.032	2.977	2.930	2.889	2.852
25	3.217	3.129	3.056	2.993	2.939	2.892	2.850	2.813
26	3.182	3.094	3.021	2.958	2.904	2.857	2.815	2.778
27	3.149	3.062	2.988	2.926	2.871	2.824	2.783	2.746
28	3.120	3.032	2.959	2.896	2.842	2.795	2.753	2.716
29	3.092	3.005	2.931	2.868	2.814	2.767	2.726	2.689
30	3.067	2.979	2.906	2.843	2.789	2.742	2.700	2.663
31	3.043	2.955	2.882	2.820	2.765	2.718	2.677	2.640
32	3.021	2.934	2.860	2.798	2.744	2.696	2.655	2.618
33	3.000	2.913	2.840	2.777	2.723	2.676	2.634	2.597

Bảng 2. Bảng phân bố F với α = 0,05

k2/k1	1	2	3	4	5	6	7	8
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266

Bảng 2. Bảng phân bố F với α = 0,05 (tiếp theo)

k2/k1	9	10	11	12	13	14	15	16
2	19.385	19.396	19.405	19.413	19.419	19.424	19.429	19.433
3	8.812	8.786	8.763	8.745	8.729	8.715	8.692	8.692
4	5.999	5.964	5.936	5.912	5.891	5.873	5.858	5.844
5	4.772	4.735	4.704	4.678	4.655	4.636	4.619	4.604
6	4.099	4.060	4.027	4.000	3.976	3.956	3.938	3.922
7	3.677	3.637	3.603	3.575	3.550	3.529	3.511	3.494
8	3.388	3.347	3.313	3.284	3.259	3.237	3.218	3.202
9	3.179	3.137	3.102	3.073	3.048	3.025	3.006	2.989
10	3.020	2.978	2.943	2.913	2.887	2.865	2.845	2.828
11	2.896	2.854	2.818	2.788	2.761	2.739	2.719	2.701
12	2.796	2.753	2.717	2.687	2.660	2.637	2.617	2.599
13	2.714	2.671	2.635	2.604	2.577	2.554	2.533	2.515
14	2.646	2.602	2.565	2.534	2.507	2.484	2.463	2.445
15	2.588	2.544	2.507	2.475	2.448	2.424	2.403	2.385
16	2.538	2.494	2.456	2.425	2.397	2.373	2.352	2.333
17	2.494	2.450	2.413	2.381	2.353	2.329	2.308	2.289
18	2.456	2.412	2.374	2.342	2.314	2.290	2.269	2.250
19	2.423	2.378	2.340	2.308	2.280	2.256	2.234	2.215
20	2.393	2.348	2.310	2.278	2.250	2.225	2.203	2.184
21	2.366	2.321	2.283	2.250	2.222	2.197	2.176	2.156
22	2.342	2.297	2.259	2.226	2.198	2.173	2.151	2.131
23	2.320	2.275	2.236	2.204	2.175	2.150	2.128	2.109
24	2.300	2.255	2.216	2.183	2.155	2.130	2.108	2.088
25	2.282	2.236	2.198	2.165	2.136	2.111	2.089	2.069
26	2.265	2.220	2.181	2.148	2.119	2.094	2.072	2.052
27	2.250	2.204	2.166	2.132	2.103	2.078	2.056	2.036
28	2.236	2.190	2.151	2.118	2.089	2.064	2.041	2.021
29	2.223	2.177	2.138	2.104	2.075	2.050	2.027	2.007
30	2.211	2.165	2.126	2.092	2.063	2.037	2.015	1.995
31	2.199	2.153	2.114	2.080	2.051	2.026	2.003	1.983
32	2.189	2.142	2.103	2.070	2.040	2.015	1.992	1.972
33	2.179	2.133	2.093	2.060	2.030	2.004	1.982	1.961

Bảng 3. Bảng phân phối Chi - square

k/α	**0.01**	**0.99**	**0.025**	**0.975**	**0.05**	**0.95**
1	6.6349	0.0002	5.0239	0.0010	3.8415	0.0039
2	9.2103	0.0201	7.3778	0.0506	5.9915	0.1026
3	11.3449	0.1148	9.3483	0.2158	7.8147	0.3518
4	13.2767	0.2971	11.1433	0.4844	9.4877	0.7107
5	15.0863	0.5543	12.8325	0.8312	11.0705	1.1455
6	16.8119	0.8721	14.4494	1.2373	12.5916	1.6354
7	18.4753	1.2390	16.0128	1.6899	14.0671	2.1673
8	20.0902	1.6465	17.5345	2.1797	15.5073	2.7326
9	21.6660	2.0879	19.0228	2.7004	16.9190	3.3251
10	23.2093	2.5582	20.4832	3.2470	18.3070	3.9403
11	24.7250	3.0535	21.9200	3.8157	19.6751	4.5748
12	26.2170	3.5706	23.3367	4.4038	21.0261	5.2260
13	27.6882	4.1069	24.7356	5.0088	22.3620	5.8919
14	29.1414	4.6604	26.1189	5.6287	23.6848	6.5706
15	30.5779	5.2293	27.4884	6.2621	24.9958	7.2609
16	31.9999	5.8122	28.8454	6.9077	26.2962	7.9616
17	33.4087	6.4078	30.1910	7.5642	27.5871	8.6718
18	34.8053	7.0149	31.5264	8.2307	28.8693	9.3905
19	36.1909	7.6327	32.8523	8.9065	30.1435	10.1170
20	37.5662	8.2604	34.1696	9.5908	31.4104	10.8508
21	38. 9322	8.8972	35.4789	10.2829	32.6706	11.5913
22	40.2894	9.54253	36.7807	10.9823	33.9244	12.3380
23	41.6384	10.1957	38.0756	11.6886	35.1725	13.0905
24	42.9798	10.8564	39.3641	12.4012	36.4150	13.8484
25	44.3141	11.5240	40.6465	13.1197	37.6525	16.6114
26	45.6471	12.1981	41.9232	13.8439	38.8851	15.3792
27	46.9629	12.8785	43.1945	14.5734	40.1133	16.1514
28	48.2782	13.5647	44.4608	15.3079	41.3371	16.9279

Bảng 3. Bảng phân phối Chi - square (tiếp)

k/α	0.01	0.99	0.025	0.975	0.05	0.95
29	49.5879	14.2565	45.7223	16.0471	42.5570	17.7084
30	50.8922	14.9535	46.9792	16.7908	43.7730	18.4927
31	52.1914	15.6555	48.2319	17.5387	44.9853	19.2806
32	53.4858	16.3622	49.4804	18.2908	46.1943	20.0719
33	54.7755	17.0735	50.7251	19.0467	47.3999	20.8665
34	56.0609	17.7891	51.9660	19.8063	48.6024	21.6643
35	57.3421	18.5089	53.2033	20.5694	49.8018	22.4650
36	58.6192	19.2327	54.4373	21.3359	50.9985	23.2686
37	59.8925	19.9602	55.6680	22.1056	52.1923	24.0749
38	61.1621	20.6914	56.8955	22.8785	53.3835	24.8839
39	62.4281	21.4262	58.1201	23.6543	54.5722	25.6954
40	63.6907	22.1643	59.3417	24.4330	55.7585	26.5093
41	64.9501	22.9056	60.5606	25.2145	56.9424	27.3256
42	66.2062	23.6501	61.7768	25.9987	58.1240	28.1440
43	67.4593	24.3976	62.9904	26.7854	59.3035	28.9647
44	68.7095	25.1480	64.2015	27.5746	60.4809	29.7875
45	69.9568	25.9013	65.4102	28.3662	61.6562	30.6123
46	71.2014	26.6572	66.6165	29.1601	62.8296	31.4390
47	72.4433	27.4158	67.8206	29.9562	64.0011	32.2676
48	73.6826	28.1770	69.0226	30.7545	65.1708	33.0981
49	74.9195	28.9406	70.2224	31.5549	66.3386	33.9303
50	76.1539	29.7067	71.4202	32.3574	67.5048	34.7643
51	77.3860	30.4750	72.6160	33.1618	68.6693	35.5999
52	78.6158	31.2457	73.8099	33.9681	69.8322	36.4371
53	79.8433	32.0185	75.0019	34.7763	70.9935	37.2759
54	81.0688	32.7934	76.1920	35.5863	72.1532	38.1162
55	82.2921	33.5707	77.3805	36.3981	73.3115	38.9580
56	83.5134	34.3495	78.5672	37.2116	74.4683	39.8013
57	84.7328	35.1305	79.7522	38.0267	75.6237	40.6459
58	85.9502	35.9135	80.9356	38.8435	76.7778	41.4920
59	87.1657	36.6982	82.1174	39.6619	77.9305	42.3393
60	88.3794	37.48349	83.2977	40.4817	79.0819	43.1880

Bảng 4. Bảng thống kê d (Durbin - Watson) với α = 0,05

	k =1		k =2		k = 3		k =4		k = 5	
n	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**
6	0.610	1.400								
7	0.700	1.356	0.467	1.896						
8	0.763	1.332	0.559	1.777	0.368	2.287				
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588		
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822
11	0.927	1.324	0.658	1.604	0.595	1.928	0.444	2.283	0.316	2.645
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.445	2.390
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991
21	1.221	1.420	1.125	1.535	1.026	1.669	0.927	1.812	0.829	1.964
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833

Bảng 4. Bảng thống kê d (Durbin - Watson) với α = 0,05 (tiếp)

	k =1		k =2		k = 3		k =4		k = 5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820

Bảng 4. Bảng thống kê d (Durbin - Watson) với $\alpha = 0,05$ (tiếp)

	k =6		k =7		k = 8		k =9		k = 10	
n	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**
6										
7										
8										
9										
10										
11	0.203	3.005								
12	0.268	2.832	0.171	3.149						
13	0.328	2.692	0.230	2.985	0.147	3.266				
14	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360		
15	0.447	2.472	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	0.603	2.257	0.502	2.461	0.407	2.667	0.321	2.873	0.244	3.073
19	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	0.692	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	0.732	2.124	0.637	2.290	0.547	2.460	0.461	2.633	0.380	2.806
22	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.734
23	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	0.837	2.035	0.751	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	0.868	2.012	o.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	0.951	1.958	0.874	2.071	0.798	2.188	0.723	2.309	0.650	2.431
29	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.682	2.396
30	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363

Bảng 4. Bảng thống kê d (Durbin - Watson) với α = 0,05 (tiếp)

	k =6		k =7		k = 8		k =9		k = 10	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
31	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.795	2.281
34	1.080	1.891	1.015	1.979	0.950	2.069	0.885	2.162	0.821	2.257
35	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.114	1.877	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.198
38	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.433	1.802	1.401	1.837	1.369	1.873	1.337	1.910	1.305	1.948
75	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.707	1.831	1.697	1.841	1.686	1.852	1,675	1.863	1.665	1.874

Bảng 4. Bảng thống kê d (Durbin - Watson) với α = 0,05 (tiếp)

	k =11		k =12		k = 13		k =14		k = 15	
n	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**
16	0.098	3.503								
17	0.138	3.378	0.087	3.557						
18	0.177	3.265	0.123	3.441	0.078	3.603				
19	0.220	3.159	0.160	3.335	0.111	3.496	0.070	3.642		
20	0.263	3.063	0.200	3.234	0.145	3.395	0.100	3.542	0.063	3.676
21	0.307	2.976	0.240	3.141	0.182	3.300	0.132	3.448	0.091	3.583
22	0.349	2.897	0.281	3.057	0.220	3.211	0.166	3.358	0.120	3.495
23	0.391	2.826	0.322	2.979	0.259	3.128	0.202	3.272	0.153	3.409
24	0.431	2.761	0.362	2.908	0.297	3.053	0.239	3.193	0.186	3.327
25	0.470	2.702	0.400	2.844	0.335	2.983	0.275	3.119	0.221	3.251
26	0.508	2.649	0.438	2.784	0.373	2.919	0.312	3.051	0.256	3.179
27	0.544	2.600	0.475	2.730	0.409	2.859	0.348	2.987	0.291	3.112
28	0.578	2.555	0.510	2.680	0.445	2.805	0.383	2.928	0.325	3.050
29	0.612	2.515	0.544	2.634	0.479	2.755	0.418	2.874	0.359	2.992
30	0.643	2.477	0.577	2.592	0.512	2.708	0.451	2.823	0.392	2.937
31	0.674	2.443	0.608	2.553	0.545	2.665	0.484	2.776	0.425	2.887
32	0.703	2.411	0.638	2.517	0.576	2.625	0.515	2.733	0.457	2.840
33	0.731	2.382	0.668	2.484	0.606	2.588	0.546	2.692	0.488	2.796
34	0.758	2.355	0.695	2.454	0.634	2.554	0.575	2.654	0.518	2.754
35	0.783	2.330	0.722	2.425	0.662	2.521	0.604	2.619	0.547	2.716
36	0.808	2.306	0.748	2.398	0.689	2.492	0.631	2.586	0.575	2.680
37	0.831	2.285	0.772	2.374	0.714	2.464	0.657	2.555	0.602	2.646
38	0.854	2.265	0.796	2.351	0.739	2.438	0.683	2.526	0.628	2.614
39	0.875	2.246	0.819	2.329	0.763	2.413	0.707	2.499	0.653	2.585
40	0.896	2.228	0.840	2.309	0.785	2.391	0.731	2.473	0.678	2.557
45	0.988	2.156	0.938	2.225	0.887	2.296	0.838	2.367	0.788	2.439
50	1.064	2.103	1.019	2.163	0.973	2.225	0.927	2.287	0.882	2.350
55	1.129	2.062	1.087	2.116	1.045	2.170	1.003	2.225	0.961	2.281
60	1.184	2.031	1.145	2.079	1.106	2.127	1.068	2.177	1.029	2.227
65	1.231	2.006	1.195	2.049	1.160	2.093	1.124	2.138	1.088	2.183
70	1.272	1.986	1.239	2.026	1.206	2.066	1.172	2.106	1.139	2.148
75	1.308	1.970	1.277	2.006	1.247	2.043	1.215	2.080	1.184	2.118
80	1.340	1.957	1.311	1.991	1.283	2.024	1.253	2.059	1.224	2.093
85	1.369	1.946	1.342	1.977	1.315	2.009	1.287	2.040	1.260	2.073
90	1.395	1.937	1.369	1.966	1.344	1.995	1.318	2.025	1.292	2.055
95	1.418	1.929	1.394	1.956	1.370	1.984	1.345	2.012	1.321	2.040
100	1.439	1.923	1.416	1.948	1.393	1.974	1.371	2.000	1.347	2.026
150	1.579	1.892	1.564	1.908	1.550	1.924	1.535	1.940	1.519	1.956
200	1.654	1.885	1.643	1.896	1.632	1.908	1.621	1.919	1.610	1.931

Bảng 4. Bảng thống kê d (Durbin - Watson) (tiếp)

	k=16		k =17		k = 18		k =19		k = 20	
n	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**	**d_L**	**d_U**
16										
17										
18										
19										
20										
21	0.058	3.705								
22	0.083	3.619	0.052	3.731						
23	0.110	3.535	0.076	3.650	0.048	3.753				
24	0.141	3.454	0.101	3.572	0.070	3.678	0.044	3.773		
25	0.172	3.376	0.130	3.494	0.094	3.604	0.065	3.702	0.041	3.790
26	0.205	3.303	0.160	3.420	0.120	3.531	0.087	3.632	0.060	3.724
27	0.238	3.233	0.191	3.349	0.149	3.460	0.112	3.563	0.081	3.658
28	0.271	3.168	0.222	3.283	0.178	3.392	0.138	3.495	0.104	3.592
29	0.305	3.107	0.254	3.219	0.208	3.327	0.166	3.431	0.129	3.528
30	0.337	3.050	0.286	3.160	0.238	3.266	0.195	3.368	0.156	3.465
31	0.370	2.996	0.317	3.103	0.269	3.208	0.224	3.309	0.183	3.406
32	0.401	2.946	0.349	3.050	0.299	3.153	0.253	3.252	0.211	3.348
33	0.432	2.899	0.379	3.000	0.329	3.100	0.283	3.198	0.239	3.293
34	0.462	2.854	0.409	2.954	0.359	3.051	0.312	3.147	0.267	3.240
35	0.492	2.813	0.439	2.910	0.388	3.005	0.340	3.099	0.295	3.190
36	0.520	2.774	0.467	2.868	0.417	2.961	0.369	3.053	0.323	3.142
37	0.548	2.738	0.495	2.829	0.445	2.920	0.397	3.009	0.351	3.097
38	0.575	2.703	0.522	2.792	0.472	2.880	0.424	2.968	0.378	3.054
39	0.600	2.671	0.549	2.757	0.499	2.843	0.451	2.929	0.404	3.013
40	0.626	2.641	0.575	2.724	0.525	2.808	0.477	2.892	0.430	2.974
45	0.740	2.512	0.692	2.586	0.644	2.659	0.598	2.733	0.553	2.807
50	0.836	2.414	0.792	2.479	0.747	2.544	0.703	2.610	0.660	2.675
55	0.919	2.338	0.877	2.396	0.836	2.454	0.795	2.512	0.754	2.571
60	0.990	2.778	0.951	2.330	0.913	2.382	0.874	2.434	0.836	2.487
65	1.052	2.229	1.016	2.276	0.980	2.323	0.944	2.371	0.908	2.419
70	1.105	2.189	1.072	2.232	1.038	2.275	1.005	2.318	0.971	2.362
75	1.153	2.156	1.121	2.195	1.090	2.235	1.058	2.275	1.027	2.315
80	1.195	2.129	1.165	2.165	1.136	2.201	1.106	2.238	1.076	2.275
85	1.232	2.105	1.205	2.139	1.177	2.172	1.149	2.206	1.121	2.241
90	1.266	2.085	1.240	2.116	1.213	2.148	1.187	2.179	1.160	2.211
95	1.296	2.068	1.271	2.097	1.247	2.106	1.222	2.156	1.197	2.186
100	1.324	2.053	1.301	2.080	1.277	2.108	1.253	2.135	1.229	2.164
150	1.504	1.972	1.489	1.989	1.474	2.006	1.458	2.023	1.443	2.040
200	1.599	1.943	1.588	1.955	1.576	1.967	1.565	1.979	1.554	1.991