

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
- Khoa Kinh tế và quản lý -



BÀI GIẢNG MÔN HỌC

KINH TẾ LƯỢNG

Người biên soạn: TS. Phạm Cảnh Huy

KINH TẾ LƯỢNG

1. Tên học phần: Kinh Tế Lượng (Econometrics)

2. Mô tả vắn tắt nội dung học phần:

Mục đích của Kinh Tế Lượng là giúp học viên nắm rõ và vận dụng được các mô hình hồi qui để ước lượng, dự đoán giá trị trung bình của tổng thể của các biến phụ thuộc theo giá trị của biến giải thích nhằm xác định mức độ quan hệ giữa các biến, từ đó thấy được bản chất của các hiện tượng và tìm được các biện pháp khắc phục. Môn học còn nhằm trang bị cho các học viên cách thức vận dụng các công cụ phân tích định lượng vào việc xử lý phân tích các vấn đề kinh tế cụ thể.

3. Nhiệm vụ của học viên:

Dự giờ giảng trên lớp và đọc giáo trình, làm bài tập theo nhóm về xử lý dữ liệu bằng ít nhất một phần mềm được môn học trang bị, tham dự thảo luận dưới sự hướng dẫn của giảng viên. Tham dự kiểm tra hết môn học theo lịch nhà trường qui định.

4. Tài liệu tham khảo thêm:

- Basic Econometrics, tác giả Damodar N. Gujarati, 1995.
- Introductory Econometrics, tác giả Craig A. Depken, 2006
- Econometric Analysis, tác giả William H. Greene, 2000.

5. Tiêu chuẩn đánh giá:

- Dự giờ đủ trên lớp theo yêu cầu của môn học
- Hoàn thành và đạt yêu cầu các bài tập về môn học trước khi thi hết môn
- Thi hết môn

6. Mục tiêu của học phần:

Nắm vững các mô hình kinh tế lượng để có thể lượng hoá các quan hệ kinh tế vĩ mô và vi mô được trang bị trước đây. Liên kết được các mô hình kinh tế lượng với các lý thuyết kinh tế vĩ mô và vi mô bằng các dữ liệu thực tế.

Đề xuất chính sách và dự báo dựa trên việc phân tích, kiểm định các mối quan hệ kinh tế vi mô và vĩ mô qua kết quả của mô hình khi ứng dụng số liệu thực tế.

7. Nội dung học phần:

- Chương I: Cơ bản về Kinh tế lượng và phân tích hồi qui
- Chương II: Mô hình hồi qui hai biến, ước lượng và kiểm định
- Chương III: Mô hình hồi qui đa biến
- Chương IV: Đa cộng tuyến
- Chương V: Hồi qui với biến giả
- Chương VI: Phương sai sai số thay đổi
- Chương VII: Tương quan chuỗi

CHƯƠNG I

CƠ BẢN VỀ KINH TẾ LƯỢNG VÀ PHÂN TÍCH HỒI QUI

1.1. Vài nét cơ bản về kinh tế lượng:

1.1.1. Giới thiệu về kinh tế lượng

Thuật ngữ tiếng Anh là Econometrics, nó được ghép từ 2 từ “Econo” có nghĩa là kinh tế và “Metrics” có nghĩa là đo lường- Vậy đó là “đo lường kinh tế”.

Theo nghĩa đơn giản, **kinh tế lượng**, liên quan đến việc áp dụng các phương pháp thống kê trong kinh tế. Không như *thống kê kinh tế*, trong đó các dữ liệu thống kê là chính yếu, kinh tế lượng được phân biệt bằng sự *hợp nhất* của lý thuyết kinh tế, công cụ toán học và các phương pháp luận thống kê. Mở rộng hơn, kinh tế lượng quan tâm đến (1) ước lượng các mối quan hệ kinh tế, (2) đối chiếu lý thuyết kinh tế với thực tế và kiểm định các giả thuyết liên quan đến hành vi kinh tế, và (3) dự báo các hành vi của các biến số kinh tế.

Người ta có thể định nghĩa như sau:

- + 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 để tìm ra lời giải bằng số.
- + Kinh tế lượng có thể được định nghĩa như là sự phân tích về lượng các vấn đề kinh tế hiện thời, dựa trên việc vận dụng đồng thời lý thuyết và thực tế được thực hiện bằng các phương pháp suy đoán thích hợp.

Ví dụ về ứng dụng của kinh tế lượng trong:

Ước lượng các mối quan hệ kinh tế

Kinh tế học thực nghiệm cung cấp rất nhiều ví dụ nhằm ước lượng các mối quan hệ kinh tế như:

1. Ước lượng cầu/cung của các sản phẩm, dịch vụ.
2. Ước lượng ảnh hưởng của chi phí bán hàng/quảng cáo đến doanh thu và lợi nhuận.
3. Giá của cổ phiếu với các đặc trưng của công ty phát hành cổ phiếu đó, cũng như với tình hình chung của nền kinh tế.
4. Đánh giá tác động của các chính sách tiền tệ và tài chính đến các biến như việc làm hoặc thất nghiệp, thu nhập, xuất khẩu và nhập khẩu, lãi suất, tỷ lệ lạm phát, và thâm hụt ngân sách.

Kiểm định giả thuyết

Cũng như bất kỳ ngành khoa học nào, một ưu điểm của kinh tế lượng là quan tâm đến **việc kiểm định giả thuyết** về các hành vi kinh tế. Ví dụ như:

1. Một doanh nghiệp có thể muốn xác định xem chiến dịch quảng cáo của mình có tác động làm tăng doanh thu hay không.
2. Các nhà phân tích có thể quan tâm xem nhu cầu có giãn hay không co giãn theo giá và thu nhập.
3. Công ty muốn biết lợi nhuận có tăng hay giảm theo qui mô hoạt động không.
4. Các nhà kinh tế học vĩ mô có thể muốn đánh giá hiệu quả của các chính sách nhà nước.

Dự báo

Khi các biến số được xác định và chúng ta đánh giá được tác động cụ thể của chúng đến chủ thể nghiên cứu, chúng ta có thể muốn sử dụng các mối quan hệ ước lượng để dự đoán các giá trị trong tương lai.

Ví dụ:

1. Các công ty dự báo doanh thu, lợi nhuận, chi phí sản xuất, và lượng tồn kho cần thiết.
2. Dự đoán có nhu cầu về năng lượng nhằm phục vụ việc hoạch định các chính sách có liên quan.
3. Dự báo các chỉ số thị trường chứng khoán và giá của một số cổ phiếu.
4. Dự đoán thu nhập, chi tiêu, lạm phát, thất nghiệp, và thâm hụt ngân sách và thương mại.
5. Các thành phố dự báo định kỳ mức tăng trưởng của địa phương qua các mặt như: dân số; việc làm; số nhà ở, nhu cầu về trường học, và dịch vụ công cộng; ...v.v

1.1.2. Mục đích của kinh tế lượng

Mục đích của kinh tế lượng là giải thích sự biến thiên của biến và các mối quan hệ của các biến, ví dụ:

Có 1 biến (chỉ tiêu) thay đổi (do lệch khỏi trung bình) mà chúng ta cần phải giải thích, ví dụ khi chúng ta nghiên cứu lượng bán một loại sản phẩm nào đó (Q) biến động, vậy cái gì tác động đến nó và các chỉ tiêu tác động lẫn nhau như thế nào.

1.1.3. Phương pháp luận của kinh tế lượng

Nêu ra các giả thuyết hay 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 mô hình toán học để mô tả mối quan hệ giữa các biến số này. Các phương trình này mô tả mối quan hệ giữa các biến số kinh tế với nhau. Một phương trình sẽ bao gồm một biến phụ thuộc và một hoặc nhiều biến giải thích. Sự tác động của một biến giải thích lên biến phụ thuộc được đo lường bằng hệ số của nó và hình thức hàm của phương trình. Một phương trình tiêu biểu như sau:

$$Y(t) = f\{x_1(t), x_2(t), \dots, x_n(t), u(t)\}$$

$Y(t)$ là biến phụ thuộc tại thời điểm t , biểu trưng cho chỉ tiêu cần nghiên cứu hay dự báo (ví dụ như GDP, việc làm, lạm phát, ...).

$x_1(t), x_2(t), \dots, x_n(t)$ là các biến giải thích tại thời điểm t , biểu trưng cho các nhân tố tác động lên biến phụ thuộc. Sự thay đổi của một hay nhiều biến này sẽ dẫn tới sự thay đổi của biến phụ thuộc.

$u(t)$ là sai số ngẫu nhiên, biểu trưng cho các nhân tố không xác định được tác động lên biến phụ thuộc tại thời điểm t .

Số hạng sai số $u(t)$, chúng ta cũng có thể ký hiệu là u_i (hay còn gọi là số hạng nhiễu ngẫu nhiên) là thành phần ngẫu nhiên không quan sát được và là sai biệt giữa Y_i và phần xác định $\beta_1 + \beta_2 X_i$.

Sau đây một tổ hợp của bốn nguyên nhân ảnh hưởng khác nhau:

1. *Biến bỏ sót.* Giả sử mô hình thực sự là $Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + v_i$ trong đó, Z_i là một biến giải thích khác và v_i là số hạng sai số thực sự, nhưng nếu ta sử dụng mô hình là $Y_i = \beta_1 + \beta_2 X_i + u_i$ thì $u_i = \beta_3 Z_i + v_i$. Vì thế, u_i bao hàm cả ảnh hưởng của biến Z bị bỏ sót.
2. *Phi tuyến tính.* u_i có thể bao gồm ảnh hưởng phi tuyến tính trong mối quan hệ giữa Y và X . Vì thế, nếu mô hình thực sự là $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + v_i$, nhưng lại được giả định bằng phương trình $Y_i = \beta_1 + \beta_2 X_i + u_i$, thì ảnh hưởng của X_i^2 sẽ được bao hàm trong u_i .
3. *Sai số đo lường.* Sai số trong việc đo lường X và Y có thể được thể hiện thông qua u .
4. *Những ảnh hưởng không thể dự báo.* Dù là một mô hình kinh tế lượng tốt cũng có thể chịu những ảnh hưởng ngẫu nhiên không thể dự báo được. Những ảnh hưởng này sẽ luôn được thể hiện qua số hạng sai số u_i .

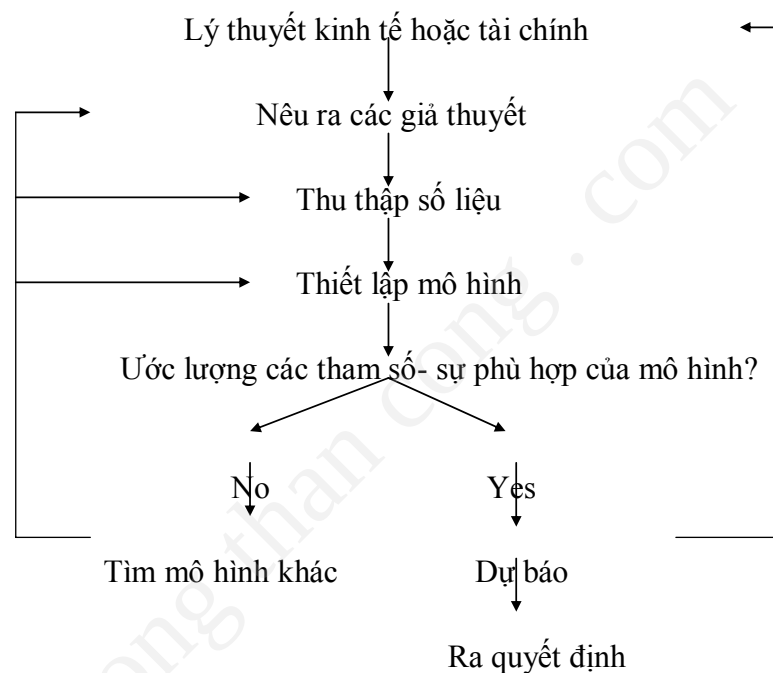
Việc xây dựng hệ thống các phương trình, với các biến giải thích lựa chọn thường được dựa trên nền tảng của lý thuyết kinh tế. Ví dụ như hàm tiêu dùng phải dựa trên lý thuyết về tiêu dùng, hàm đầu tư phải dựa trên lý thuyết về đầu tư, ... Điều này dẫn đến hệ quả là các nhà mô hình khác nhau có thể sẽ xây dựng các phương trình với các biến giải thích khác nhau, tùy thuộc vào việc áp dụng lý thuyết kinh tế nào. Điều đó cũng lý giải về sự đa dạng của các mô hình kinh tế lượng hiện nay.

Ví dụ, Giả sử chúng ta điều tra tất cả các hộ trong thành phố và tính thu nhập hàng tháng của họ (X) và tổng chi tiêu vào hàng hóa và dịch vụ (Y). Vì các hộ gia đình có cùng thu nhập sẽ có những mức chi tiêu khác nhau (có lẽ do khác biệt về các đặc điểm khác như số thành viên trong gia đình), một quan sát cụ thể (Y, X) sẽ không hoàn toàn chính xác nằm trên đường thẳng. Do vậy, mô hình hồi qui tuyến tính tương ứng với ví dụ này sẽ có dạng $Y = \beta_1 + \beta_2 X + u$

Trong thực tế, chúng ta sẽ không điều tra tất cả các hộ gia đình mà chỉ chọn một mẫu ngẫu nhiên từ tổng thể và sử dụng các quan sát này để ước lượng những tham số β_1 và β_2 cũng như thực hiện các kiểm định và kiểm tra tính phù hợp của giả định về mối liên hệ trung bình giữa chi tiêu và thu nhập là tuyến tính.

Sau khi xây dựng xong hệ thống các phương trình, chúng ta phải tập hợp đủ các số liệu cho các biến và tiến hành ước lượng các hệ số của các phương trình. Kỹ thuật hồi quy (regression) được áp dụng để ước lượng các hệ số của các phương trình. Sau khi ước lượng xong toàn bộ các phương trình của mô hình, chúng ta sẽ tiến hành mô phỏng (simulation) tác động của các thay đổi chính sách trong tương lai lên các biến kinh tế mà mình quan tâm (ví dụ như tăng trưởng, việc làm, lạm phát,...). Trên cơ sở đó, chúng ta có thể đánh giá tác động của chúng hoặc/và đề xuất ra các kịch bản dự báo.

Các bước thực hiện



1.2. Phân tích hồi qui

1.2.1. Các ví dụ trong lĩnh vực kinh tế về mối quan hệ nhân quả

Trong phân tích hồi qui, chúng ta cần ước lượng quan hệ toán học giữa các biến. Những mối quan hệ này còn được gọi là mối quan hệ hàm số. Chúng cố gắng mô tả các biến giải thích tác động lên biến phụ thuộc như thế nào.

- Biến giải thích là biến xảy ra.
- Biến phụ thuộc là biến kết quả.

Ví dụ: Khi chúng ta cố gắng giải thích chi tiêu dùng của mọi người, chúng ta có thể sử dụng các biến giải thích là thu nhập và độ tuổi. Khi giải thích giá của một chiếc ô tô, các biến giải thích có thể là kích cỡ, động cơ máy, độ tin cậy của hãng sản xuất cũng như độ an toàn của chiếc ô tô. Để giải thích giá của một ngôi nhà các biến giải thích có thể là kích cỡ, số phòng, tỷ lệ tội phạm của khu dân cư cũng như độ tuổi của ngôi nhà. Để dự đoán khả năng một học sinh cuối cấp trung học phổ thông vào đại học, chúng ta có thể

xem xét đến điểm các bài kiểm tra, trình độ giáo dục của cha mẹ cũng như thu nhập của gia đình anh ta.

Vậy phân tích hồi qui chính là 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 biến độc lập hay giải thích).

1.2.2. Mục đích của phân tích hồi qui:

Tưởng tượng rằng chúng ta có thông tin về thu nhập và chi tiêu tiêu dùng, chúng ta tin tưởng rằng chi tiêu tiêu dùng phụ thuộc vào thu nhập và chúng ta biểu diễn cả 2 biến này lên đồ thị. Biểu diễn biến phụ thuộc lên trục tung, còn biến giải thích (biến độc lập) lên trục hoành.

Mục đích của phân tích hồi qui là qua những điểm dữ liệu, chúng ta có thể kẻ ra một đường phù hợp nhất, sát nhất với các quan sát để sao cho có thể biểu diễn mối quan hệ giữa hai biến thu nhập và chi tiêu tiêu dùng một cách đáng tin cậy nhất.

1.2.3. Giới thiệu mô hình hồi qui tuyến tính đơn giản:

Để mô hình hóa quan hệ tuyến tính trong đó diễn tả sự thay đổi của biến Y theo biến X cho trước người ta sử dụng mô hình hồi qui tuyến tính đơn giản. Mô hình hồi qui tuyến tính đơn giản có dạng sau:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

- + Y_i : Giá trị của biến phụ thuộc Y trong lần quan sát thứ i.
- + X_i : Giá trị của biến độc lập X trong lần quan sát thứ i .
- + u_i : Giá trị đối với sự dao động ngẫu nhiên (nhiều ngẫu nhiên) hay sai số trong lần quan sát thứ i.
- + β_1 : là thông số diễn tả tung độ gốc (hệ số chặn) của đường hồi qui tổng thể, hay β_1 là giá trị trung bình của biến phụ thuộc Y khi biến độc lập X thay đổi 1 đơn vị.
- + β_2 : là thông số diễn tả độ dốc (hệ số góc) của đường hồi qui tổng thể, hay β_2 diễn tả sự thay đổi của giá trị trung bình của biến phụ thuộc Y khi biến độc lập X thay đổi 1 đơn vị.

Chúng ta có thể ước lượng các tham số (β_1, β_2) của phương trình hồi qui tổng thể bằng cách sử dụng số liệu của mẫu ngẫu nhiên thu thập được. Dựa vào số liệu của mẫu ta có phương trình hồi qui tuyến tính của mẫu.

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i}$$

Trong đó:

$\hat{Y}$ là ước lượng của giá trị trung bình của Y đối với biến X đã biết

$\hat{\beta}_1$ là ước lượng của β_1

$\hat{\beta}_2$ là ước lượng của β_2

1.3. Hồi qui và tương quan

Khi định mô hình ở dạng $Y_i = \beta_1 + \beta_2 X_i + u_i$, chúng ta ngầm giả định rằng X gây ra sự thay đổi của Y . Việc X và Y tương quan chặt với nhau không có nghĩa rằng sự thay đổi X dẫn đến sự thay đổi Y hay ngược lại. Ví dụ, hệ số tương quan giữa số lượng kanguru của Úc và tổng dân số nước này có thể là rất cao. Phải chăng điều này có nghĩa rằng sự thay đổi một biến sẽ làm cho biến kia thay đổi? Rõ ràng là không, vì ở đây chúng ta có một trường hợp **tương quan giả tạo**. Nếu chúng ta hồi quy một trong các biến với biến còn lại, chúng ta sẽ có sự **hồi qui giả tạo**. Lấy một ví dụ khác thực tế hơn, giả sử chúng ta hồi quy số lượng vụ trộm trong một thành phố với số hạng hằng số và số nhân viên cảnh sát (X) và sau đó quan sát thấy hệ số góc ước lượng có giá trị dương, có nghĩa rằng có tương quan thuận giữa X và Y . Phải chăng điều này có nghĩa rằng việc tăng số lượng cảnh sát sẽ làm tăng số vụ trộm, do đó ngầm kéo theo phải có chính sách giảm lực lượng cảnh sát? Rõ ràng kết luận này là không thể chấp nhận được. Điều xảy ra có thể là mối quan hệ nhân quả là ngược lại, có nghĩa là thành phố nên thuê thêm cảnh sát vì số vụ trộm tăng lên, và như vậy việc hồi quy X theo Y là hợp lý hơn.

Từ những ví dụ trên ta thấy rằng Hồi qui và tương quan khác nhau về mục đích và kỹ thuật. Phân tích tương quan xem xét mức độ kết hợp tuyến tính giữa hai biến, nhưng phân tích hồi qui lại ước lượng hoặc dự báo một biến trên cơ sở giá trị đã cho của các biến khác.

Về mặt kỹ thuật, trong phân tích hồi qui các biến không có tính chất đối xứng, biến phụ thuộc là biến 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.

1.4. Các dạng hàm trong kinh tế lượng

Giả sử ta có một mô hình kinh tế tiên đoán mối quan hệ giữa một biến phụ thuộc Y và một biến độc lập X . Trong nhiều trường hợp, mô hình này sẽ không cho chúng ta biết dạng hàm mà mối quan hệ này có trong dữ liệu, mặc dù mô hình này sẽ thường cho thấy một số ý niệm về dạng có thể có của mối quan hệ. Giải pháp thông thường là quyết định xem dạng hàm nào có khả năng mô tả tốt dữ liệu nhất, điều này hoặc phụ thuộc vào suy luận kinh tế hoặc phụ thuộc vào việc khảo sát dữ liệu. Sau đó, chúng ta thử xây dựng một số dạng hàm khác nhau và xem chúng có cho ra các kết quả tương tự hay không, và nếu không, thì phải xem dạng hàm nào cho ra các kết quả hợp lý nhất.

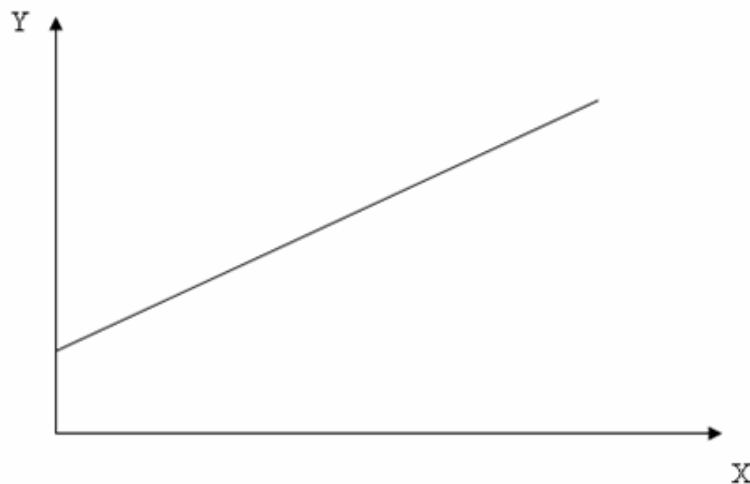
Phần này liệt kê một số dạng hàm được sử dụng phổ biến nhất, cho biết chúng biểu hiện như thế nào, mô tả các tính chất của chúng, và cho chúng ta một số ý tưởng về cách chọn lựa giữa các dạng hàm này.

1.4.1. Dạng Hàm Tuyến tính.

Dạng hàm này có phương trình:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Dạng hàm tuyến tính có thể được mô tả ở dạng như sau:



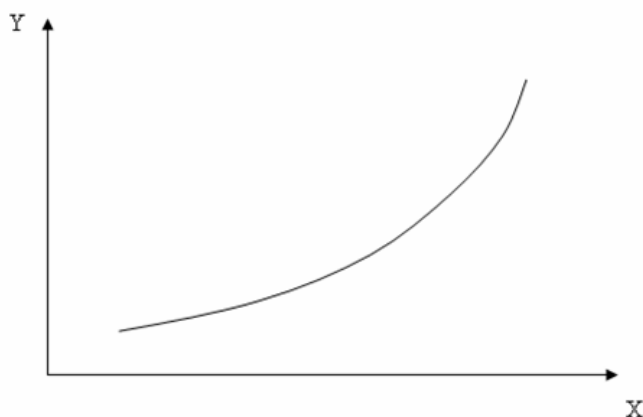
Ưu điểm của dạng hàm tuyến tính là tính đơn giản của nó. Mỗi lần X tăng thêm một đơn vị thì Y tăng thêm β_1 đơn vị, và điều này đúng bất kể các giá trị của X và Y là bao nhiêu. Nhược điểm của dạng hàm tuyến tính cũng chính là tính đơn giản của nó, bất cứ lúc nào tác động của X phụ thuộc vào các giá trị của X hoặc Y , thì dạng hàm tuyến tính không thể là dạng hàm phù hợp. Thí dụ, nếu ta có đường biểu diễn chi phí có dạng $C = \beta_0 + \beta_1 Q + \varepsilon$, thì dạng hàm tuyến tính ám chỉ là khi Q tăng thêm một đơn vị thì chi phí C tăng thêm β_1 đơn vị. Điều này chỉ có thể đúng trong trường hợp chi phí biên không đổi; nó không thể đúng trong trường hợp chi phí biên tăng dần (hay giảm dần). Nếu chúng ta nghĩ rằng chi phí biên tăng dần, chúng ta sẽ không muốn sử dụng dạng hàm tuyến tính.

1.4.2. Dạng Hàm Bậc hai.

Dạng hàm này cho phép giải thích tác động của X lên Y phụ thuộc vào giá trị hiện hành của X . Nó có phương trình:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon.$$

Dạng hàm bậc hai có thể được mô tả ở dạng như sau:



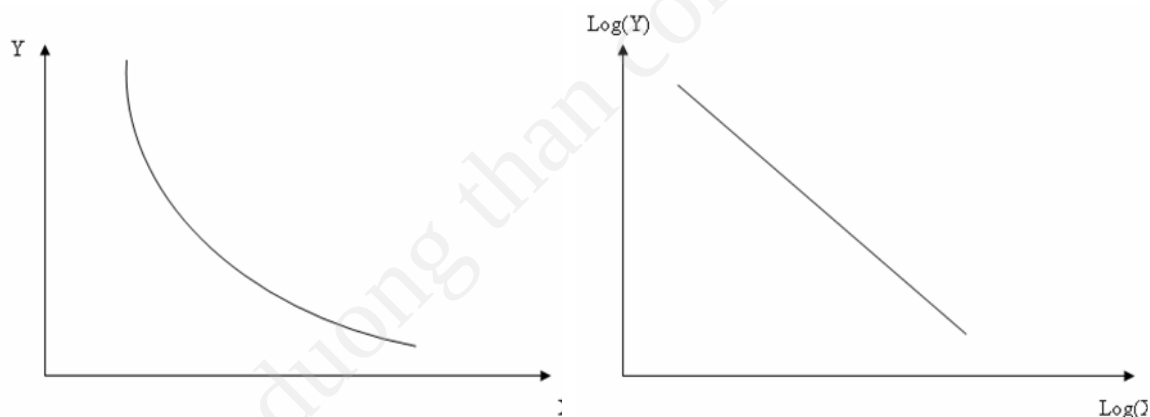
Ưu điểm của dạng hàm bậc hai là khi X tăng thêm một đơn vị thì Y tăng thêm $\beta_1 + 2\beta_2 X$ đơn vị. (Dễ dàng thấy được điều này bằng cách tính dY/dX từ phương trình nói trên.) Nếu $\beta_2 > 0$, thì khi X tăng lên, tác động bổ sung của X đến Y cũng tăng lên; nếu $\beta_2 < 0$, thì khi X tăng lên, tác động bổ sung của X đến Y giảm xuống. Nếu ta có đường biểu diễn chi phí $C = \beta_0 + \beta_1 Q + \beta_2 Q^2 + \varepsilon$, thì chi phí biên (tức là số đơn vị mà C tăng lên khi Q tăng lên thêm một đơn vị) sẽ là $MC = \beta_1 + 2\beta_2 Q$. Nếu $\beta_2 > 0$, thì ta có chi phí biên tăng dần; nếu $\beta_2 < 0$, thì ta có chi phí biên giảm dần. Lý thuyết kinh tế gợi ý rằng chúng ta thường có chi phí biên tăng dần (hay có thể không đổi); chúng ta có thể kiểm định liệu có phải $\beta_2 > 0$ (hay $\beta_2 = 0$) để kiểm định lý thuyết này.

1.4.3. Dạng Hàm Logarit.

Dạng hàm này có phương trình:

$$\log Y = \beta_0 + \beta_1 \log X + \varepsilon.$$

Đồ thị của dạng hàm này có thể được mô tả ở dạng như sau:



Có hai cách để nghĩ về dạng hàm này. Một cách để giải thích dạng hàm này là nếu X thay đổi 1% thì Y sẽ thay đổi $\beta_1\%$; đây là tính chất đặc biệt của quan hệ lôgarít. Cách giải thích thứ hai về dạng hàm này là β_1 là độ co giãn của Y theo X ; điều này suy ra từ định nghĩa của độ co giãn (chúng ta dễ dàng chứng minh điều này bằng một số biến đổi, bắt đầu từ $d(\log Y)/d(\log X)$ bằng với $(dY/dX)(X/Y)$ và sắp xếp các số hạng lại). Dạng hàm này thường được sử dụng khi chúng ta quan tâm đến việc ước lượng một loại độ co giãn nào đó. Người ta cũng sử dụng dạng hàm này phổ biến khi chúng ta sử dụng hàm Cobb Douglas; hàm Cobb-Douglas có dạng $Y = AX^{\beta_1}e^{\varepsilon}$.

và nếu lấy log ở cả hai vế, chúng ta được: $\ln Y = \ln(A) + \beta_1 \ln X + \varepsilon$

(Trong Eviews hàm này được viết dưới dạng $\log Y = \log(A) + \beta_1 \log X + \varepsilon$)

và ta có thể đặt $\beta_0 = \log(A)$. Vì thế cho nên dạng hàm lôgarít thường được sử dụng cho các hàm chi phí, các hàm sản xuất, các hàm hữu dụng, và các hàm khác mà chúng thường được mô tả dưới dạng hàm Cobb-Douglas.

Ví dụ về Hàm sản xuất Cobb-Douglas

$$Y = \beta_1 K^{\beta_2} L^{\beta_3} e^{\varepsilon}$$

Y = sản lượng

K = nhập lượng vốn

L = nhập lượng lao động

Đây là mối quan hệ phi tuyến, nhưng chúng ta có thể biến đổi quan hệ này:

$$\text{Nhu thế: } \ln Y = \ln \beta_1 + \beta_2 \ln K + \beta_3 \ln L + \varepsilon$$

Đây là mô hình tuyến tính trong các tham số nhưng không tuyến tính trong các biến số. Mô hình này tuyến tính theo lôgarít của các biến số. Mô hình này được gọi là mô hình lôgarít-lôgarít, lôgarít kép hay tuyến tính lôgarít

Hệ số độ dốc của một mô hình tuyến tính lôgarít đo lường độ co giãn của Y theo X. Như thế, hệ số nói trên là độ co giãn. Độ co giãn này không đổi đối với X&Y

- + $\beta_2 + \beta_3$ đo lường hiệu quả theo qui mô. Đáp ứng của sản lượng đối với thay đổi tương xứng trong các nhập lượng.
- + Nếu $\beta_2 + \beta_3 = 1$: hiệu quả không đổi. Tăng gấp đôi nhập lượng thì sản lượng sẽ tăng gấp đôi.
- + Nếu $\beta_2 + \beta_3 < 1$: hiệu quả giảm dần
- + Nếu $\beta_2 + \beta_3 > 1$: hiệu quả tăng dần

Ví dụ khi hồi qui theo dạng hàm này với dữ liệu về nông nghiệp của Đài Loan từ 16 quan sát, người ta thu được kết quả như sau:

$$\ln Y = -3.34 + 0.49 \ln K + 1.50 \ln L$$

$$t \quad (-1.36) \quad (4.80) \quad (0.54)$$

$$R^2 = 0.89$$

Y GNP tính bằng triệu đô la

K là vốn thực tính bằng triệu đô la

L tính bằng triệu ngày công lao động

Độ co giãn của sản lượng theo vốn là 0,49. Giữ nhập lượng lao động không đổi, gia tăng 1% nhập lượng vốn dẫn đến gia tăng 0,49% sản lượng.

Độ co giãn của sản lượng theo lao động là 1,50. Giữ nhập lượng vốn không đổi, gia tăng 1% nhập lượng lao động dẫn đến gia tăng 1,5% sản lượng.

Hiệu quả tăng theo qui mô bởi vì $\beta_2 + \beta_3 = 1,99$.

R^2 có nghĩa là 89% biến thiên trong lôgarít của sản lượng được giải thích bởi lôgarít của lao động và vốn.

Hay ví dụ chúng ta có thể lập mô hình cầu như một mô hình tuyến tính lôgarít và từ đó ước lượng độ co giãn của cầu tiêu dùng cà phê mỗi ngày:

Giả sử kết quả thu được như sau:

$$\ln Q = 0.78 - 0.25 \ln P_{\text{coffee}} + 0.38 \ln P_{\text{tea}}$$

- + Q là mức tiêu dùng cà phê mỗi ngày
- + P_{coffee} là giá cà phê mỗi cân Anh
- + P_{tea} là giá trà mỗi cân Anh

Độ co giãn theo giá riêng là -0,25.

- + Giữ các yếu tố khác không đổi, nếu giá gia tăng 1% thì lượng cầu sẽ giảm 0,25%.
- + Đây là không co giãn - giá trị tuyệt đối nhỏ hơn 1.

Độ co giãn theo giá-chéo là 0,38.

- + Giữ các yếu tố khác không đổi, nếu giá trà gia tăng 1%, thì lượng cầu cà phê sẽ gia tăng 0,38%
- + Nếu độ co giãn theo giá-chéo dương, thì cà phê và trà là các sản phẩm thay thế.
- + Nếu độ co giãn theo giá-chéo âm, thì đó là các sản phẩm bổ trợ.

1.4.4. Dạng Hàm Translog.

Phương trình của dạng hàm này là:

$$\log Y = \beta_0 + \beta_1 \log X + \beta_2 (\log X)^2 + \varepsilon.$$

Dạng hàm này có mối quan hệ với dạng hàm lôgarít giống như mối quan hệ giữa hàm bậc hai và hàm tuyến tính; nó đưa thêm một số hạng bình phương vào phương trình. Trong dạng hàm này, độ co giãn của Y theo X là $\beta_1 + 2\beta_2 \log X$.

1.4.5. Dạng Hàm bán-lôgarít (Semilog).

Dạng hàm này có phương trình sau:

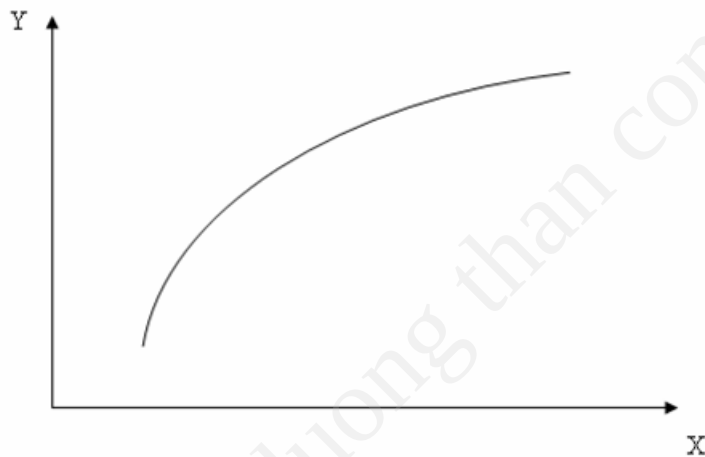
$$\log Y = \beta_0 + \beta_1 X + \varepsilon.$$

Sử dụng mô hình này khi chúng ta quan tâm đến tốc độ tăng trưởng của biến nào đó như GNP hay mức việc làm.

Dạng hàm bán-lôgarít có tính chất là nếu X tăng thêm 1 đơn vị thì Y tăng thêm $[\beta_1 \cdot 100] \%$. Đây không phải là một tính chất được mong muốn một cách phổ biến, nhưng có một số ứng dụng hữu ích cho dạng hàm này.

Ví dụ, quan hệ giữa tiền lương và trình độ giáo dục hầu như luôn luôn được biểu hiện dưới dạng hàm này như là $\log(SAL) = \beta_0 + \beta_1 ED + \varepsilon$. Điều này có nghĩa là nếu trình độ giáo dục của một người tăng 1 năm thì tiền lương của người đó tăng $[\beta_1 \cdot 100] \%$. Thí dụ, nếu $\beta_1 = 0,08$, nó có nghĩa là một năm tăng thêm trong trình độ giáo dục làm tăng tiền lương thêm 8%. Khi X tăng lên thì độ dốc của đường biểu diễn sẽ trở nên rất lớn, bởi vì khi X tăng lên thì tỷ lệ phần trăm gia tăng của X cũng lớn hơn.

Chúng ta cũng có thể đặt lôgarít cho X , nghĩa là dạng hàm trên trở thành, $Y = \beta_0 + \beta_1 \cdot \log X + \varepsilon$ và điều này có nghĩa là khi X tăng 1%, thì Y tăng $[\beta_1/100]$ đơn vị. Trong trường hợp này, khi X càng lớn thì độ dốc càng nhỏ, bởi vì X cần gia tăng nhiều hơn mới tạo ra được 1% gia tăng. Dạng hàm này không được sử dụng rộng rãi; dưới đây là đồ thị của nó.

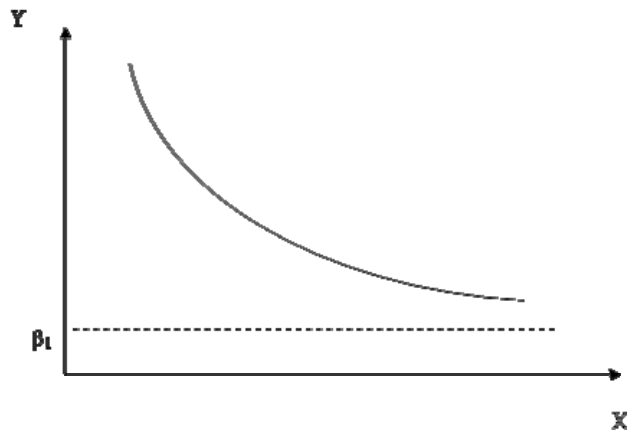


1.4.6. Dạng Hàm Nghịch đảo.

Dạng hàm này có phương trình như sau:

$$Y = \beta_0 + \beta_1 \left(\frac{1}{X} \right) + \varepsilon.$$

Đồ thị của dạng hàm này có thể được mô tả ở dạng như sau:



Dạng hàm nghịch đảo thường được sử dụng khi Y và X đều dương và khi đường biểu diễn quan hệ giữa chúng có lẽ dốc xuống (nghĩa là, $\beta_0 > 0$ và $\beta_1 > 0$). Trong trường hợp này, dạng hàm tuyến tính không được tốt bởi vì đường biểu diễn sẽ cắt trục tọa độ và Y sẽ trở nên âm đối với các giá trị X đủ lớn. Dạng hàm này thường được sử dụng cho các đường (cong) như đường cầu hay đường chi phí cố định (chi phí cố định trung bình trong sản xuất giảm xuống liên tục khi sản lượng tăng lên) cần có tính chất này.

Nếu chúng ta có nhiều biến độc lập, thì dạng quan hệ giữa Y và mỗi biến có thể giống nhau hoặc có thể khác nhau. Chúng ta có thể sử dụng cùng một dạng hàm như nhau cho mỗi biến; ví dụ, nếu chúng ta nghĩ rằng sản lượng phụ thuộc vào ba nhập lượng khác nhau thì có thể sử dụng dạng hàm tuyến tính cho mỗi nhập lượng:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

Hay dạng hàm translog cho mỗi nhập lượng:

$$\log Y = \beta_0 + \beta_1 \log X_1 + \beta_2 (\log X_1)^2 + \beta_3 \log X_2 + \beta_4 (\log X_2)^2 + \varepsilon.$$

Hay tổng quát hơn:

$$\log Y = \beta_0 + \sum_{i=1}^k \beta_i X_i + \sum_{i=1}^k \sum_{j=1}^k \gamma_{ij} (\log X_i)(\log X_j) + \varepsilon.$$

Chúng ta cũng có thể kết hợp vài dạng hàm khác nhau trong một hồi qui, thí dụ:

$$Y = \beta_0 + \beta_1 \log X + \beta_2 X_2 + \beta_3 X_2^2 + \beta_4 \left(\frac{1}{X_3}\right) + \varepsilon.$$

mặc dù, nếu làm thế, chúng ta thường phải có các lý do thỏa đáng để nghĩ rằng hình dạng của quan hệ giữa Y và X_1 là khác với các hình dạng của quan hệ giữa Y và X_2 , và Y và X_3 .

1.5. Các loại dữ liệu cho phân tích kinh tế lượng

Để ước lượng mô hình kinh tế đã đưa ra, cần có mẫu dữ liệu về các biến phụ thuộc và biến độc lập.

Có 3 loại số liệu được sử dụng để phân tích:

1. Các số liệu theo thời gian (chuỗi thời gian)
2. Các số liệu chéo
3. Các số liệu hỗn hợp của hai loại trên.

Các số liệu có thể dạng số lượng (như. GDP, tỷ giá hối đoái, Giá chứng khoán), hay dạng chất lượng (như. Nam/ nữ; có gia đình/ chưa có gia đình; Quá trình sản xuất A/quá trình sản xuất B).

1.5.1. Số liệu theo thời gian:

Là số liệu được thu thập trong một thời kỳ, như:

- Quan sát mức lạm phát và thất nghiệp của Mỹ từ 1962-1995
- Quan sát GDP của Mỹ từ 1960-1992
- Quan sát khả năng sinh lời của một công ty trong hơn 20 năm
- Quan sát giá vàng hàng ngày lúc đóng cửa trong hơn 30 năm.

Ví dụ, giả sử một thành phố muốn dự báo nhu cầu nhà ở cho năm hoặc mười năm trong tương lai. Việc này đòi hỏi phải xác định các biến có ảnh hưởng đến nhu cầu nhà ở của thành phố đó trong quá khứ, có được chuỗi dữ liệu theo thời gian trong nhiều năm ở quá khứ, và sử dụng chúng vào một mô hình thích hợp để tạo các giá trị dự báo của nhu cầu tương lai. Khoảng thời gian hoặc **thời đoạn của chuỗi thời gian** sẽ là hàng năm, hàng quý hoặc hàng tháng, tùy theo thành phố đó muốn xem xét thay đổi trong nhu cầu nhà ở hàng năm, hàng quý hay hàng tháng. Loại dữ liệu sẵn có thường sẽ quyết định thời đoạn của dữ liệu thu thập.

1.5.2. Số liệu chéo:

Là số liệu về một hay nhiều biến được thu thập tại một thời điểm ở nhiều địa phương, đơn vị khác nhau.

- Quan sát chiều cao và cân nặng của 1000 người
- Quan sát thu nhập, trình độ học vấn, và cân nặng của 1000 người
- Quan sát khả năng sinh lời của 20 công ty
- Quan sát GDP trên đầu người, dân số, và chi phí quốc phòng thực tế của 80 quốc gia.

Ví dụ, khi chúng ta muốn xem xét thu nhập ảnh hưởng như thế nào đến tiêu dùng của một người. Việc này đòi hỏi phải quan sát thu nhập và tiêu dùng của nhiều người trong một khoảng thời gian xác định.

1.5.3. Số liệu hỗn hợp:

Số liệu hỗn hợp theo không gian và thời gian, ví dụ như:

- Quan sát tỷ lệ lạm phát và mức tăng trưởng của 15 quốc gia trong khoảng thời gian từ 1970-1995
- Quan sát mức sản lượng và mức giá của 100 ngành trong hơn 12 quý
- Quan sát khả năng sinh lời của 20 công ty trong hơn 20 năm

Số liệu chuỗi thời gian thường người ta ký hiệu là t và tổng số quan sát là T , còn đối với số liệu chéo ta ký hiệu quan sát là i và tổng số quan sát là N .

Dữ liệu có thể thu thập được tại những nguồn sẵn có. Tuy nhiên, trong nhiều trường hợp những nguồn này không đủ để giải quyết vấn đề đặt ra hoặc những dữ liệu này không có sẵn. Trong trường hợp như vậy, cần tiến hành những khảo sát để thu thập các thông tin cần thiết. Ví dụ, chúng ta muốn quan tâm đến việc nghiên cứu xem người tiêu dùng sẽ phản ứng như thế nào đối với chính sách giá điện. Chính sách giá điện trong ngày là giá điện sẽ thay đổi theo những giờ khác nhau trong ngày, với giá cao trong những giờ cao điểm và giá thấp trong những giờ thấp điểm. Để có được dữ liệu phù hợp người ta chọn một số khách hàng và lắp đặt đồng hồ để ghi lại lượng điện sử dụng từng giờ trong ngày. Lượng điện tiêu thụ được thu thập trong vòng một năm, như vậy có được dữ liệu theo chuỗi thời gian cho một nhóm các hộ tiêu thụ nào đó.

TÓM TẮT

Kinh tế lượng liên quan đến ước lượng các mối liên hệ kinh tế, kiểm định giả thuyết các lý thuyết kinh tế, và dự báo các biến kinh tế hoặc các biến số khác. Khi nghiên cứu, chúng ta thường phải bắt đầu với một tập hợp các lý thuyết kinh tế, sau đó kết hợp chúng với những nhận định trực giác (hoặc kinh nghiệm, nghiên cứu trong quá khứ) để xây dựng một mô hình kinh tế lượng. Quá trình này liên quan đến quyết định chọn một hay nhiều biến phụ thuộc và xác định các biến độc lập có ảnh hưởng đến các biến phụ thuộc.

Bước tiếp theo là thu thập dữ liệu tương ứng. Khi có được các dữ liệu này, chúng ta sẽ ước lượng các thông số của một hoặc nhiều mô hình sơ bộ. Các mô hình này sẽ được kiểm định nhiều lần, dựa vào những kiểm định này, các mô hình được thiết lập lại và ước lượng lại cho đến khi thỏa mãn. Mô hình cuối cùng có thể được dùng để xây dựng các chính sách hoặc để dự báo các giá trị của các biến phụ thuộc trong nhiều tình huống khác nhau.

CHƯƠNG II

MÔ HÌNH HỒI QUY HAI BIẾN, ƯỚC LƯỢNG VÀ KIỂM ĐỊNH

Ở chương I phát biểu rằng bước đầu tiên trong phân tích kinh tế lượng là việc thiết lập mô hình mô tả được hành vi của các đại lượng kinh tế. Tiếp theo đó nhà phân tích kinh tế/ kinh doanh sẽ thu thập những dữ liệu thích hợp và ước lượng mô hình nhằm hỗ trợ cho việc ra quyết định. Trong chương này sẽ giới thiệu mô hình đơn giản nhất và phát triển các phương pháp ước lượng, phương pháp kiểm định giả thuyết và phương pháp dự báo. Mô hình này đề cập đến biến độc lập (Y) và một biến phụ thuộc (X). Đó chính là mô hình hồi quy tuyến tính hai biến (thường gọi là mô hình hồi quy đơn). Mặc dù đây là một mô hình đơn giản, và đôi khi có thể là phi thực tế, nhưng việc hiểu biết những vấn đề cơ bản trong mô hình này là nền tảng cho việc tìm hiểu những mô hình phức tạp hơn. Thực tế, mô hình hồi quy đơn tuyến tính có thể giải thích cho nhiều phương pháp kinh tế lượng. Trong chương này chỉ đưa ra những kết luận căn bản về mô hình hồi quy tuyến tính đơn biến.

Mục tiêu đầu tiên của một nhà kinh tế lượng là làm sao sử dụng dữ liệu thu thập được để ước lượng hàm hồi quy của tổng thể, đó là, ước lượng tham số của tổng thể β_1 và β_2 . Ký hiệu $\hat{\beta}_1$ là ước lượng mẫu của β_1 và $\hat{\beta}_2$ là ước lượng mẫu của β_2 . Khi đó mối quan hệ trung bình ước lượng là $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$. Đây được gọi là hàm hồi quy của mẫu.

Thuật ngữ *đơn* trong mô hình hồi quy tuyến tính đơn được sử dụng để chỉ rằng chỉ có duy nhất một biến giải thích (X) được sử dụng trong mô hình.

Trong chương tiếp theo khi nói về *mô hình hồi quy đa biến* sẽ bổ sung thêm nhiều biến giải thích khác. Thuật ngữ *hồi quy* xuất phát từ Francis Galton (1886), người đặt ra mối liên hệ giữa chiều cao của người con trai với chiều cao của người cha và quan sát thực nghiệm cho thấy có một xu hướng giữa chiều cao trung bình của người con trai với chiều cao của những người cha của họ để “hồi quy” cho chiều cao trung bình của toàn bộ tổng thể. $\beta_1 + \beta_2 X_i$ gọi là *phần xác định* của mô hình và là trung bình có điều kiện của Y theo X , đó là $E(Y_i) = \beta_1 + \beta_2 X_i$. Thuật ngữ *tuyến tính* dùng để chỉ rằng bản chất của các *thông số của tổng thể* β_1 và β_2 là tuyến tính (bậc nhất).

2.1. Khái niệm hàm hồi quy tổng thể

Tổng thể là toàn bộ các quan sát về các đối tượng hay con người cho mục đích nghiên cứu. Mục tiêu đầu tiên của một nhà kinh tế lượng là làm sao sử dụng dữ liệu thu thập được để ước lượng hàm hồi quy của tổng thể, đó là, ước lượng tham số của tổng thể β_1 và β_2 .

Cho Y là biến được giải thích, chọn $X_2, X_3, \dots, X_k$ là biến giải thích. Y là ngẫu nhiên và có 1 phân phối xác suất nào đó.

$\Rightarrow$ tồn tại $E(Y|X_2, X_3, \dots, X_k) =$ giá trị xác định

Do vậy $F(X_2, X_3, \dots, X_k) = E(Y|X_2, X_3, \dots, X_k)$ là hàm hồi quy tổng thể của Y theo $X_2, X_3, \dots, X_k$ (PRF-population regression function), hàm phụ thuộc ở mức độ trung bình của Y theo X .

Với một cá thể i , tồn tại $(X_{2i}, X_{3i}, \dots, X_{ki}, Y_i)$

Ta có $Y_i \neq F(X_2, X_3, \dots, X_k) \Rightarrow u_i = Y_i - F$

Do vậy: $Y_i = E(Y|X_2, X_3, \dots, X_k) + u_i$

Hồi qui tổng thể PRF:

$$Y = E(Y|X) + U$$

$$E(Y|X) = F(X)$$

2.2. Hàm hồi qui mẫu

Do không biết tổng thể, nên chúng ta không biết giá trị trung bình tổng thể của biến phụ thuộc là đúng ở mức độ nào. Do vậy chúng ta phải dựa vào dữ liệu mẫu để ước lượng.

Trên thực tế khi tổng thể lớn, tồn tại F nhưng không tìm được chính xác do:

- ✓ Không quan sát được (do thời gian hay tài chính không cho phép..)
- ✓ Tổng thể biến động
- ✓ Đặc điểm thông tin: không cần quan sát

Do vậy người ta phải tiến hành chọn mẫu, mẫu là một nhóm hay một bộ phận của tổng thể.

Hồi qui mẫu:

$$\text{Cho PRF: } Y = F(x_2, x_3, \dots, x_k) + u$$

Trên một bộ phận (mẫu) có n cá thể gọi $\hat{Y} = \hat{F}(X_2, X_3, \dots, X_k)$ là hồi qui mẫu (SRF - Sample regression function)

Với một cá thể mẫu $Y_i \neq \hat{F}(X_{2i}, X_{3i}, \dots, X_{ki})$

Sinh ra $e_i = Y_i - \hat{F}(X_{2i}, X_{3i}, \dots, X_{ki}) = Y_i - \hat{Y}_i$; e_i gọi là *Phần dư SRF*

Ký hiệu $\hat{\beta}_1$ là ước lượng mẫu của β_1 và $\hat{\beta}_2$ là ước lượng mẫu của β_2 . Khi đó mối quan hệ trung bình ước lượng là $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$. Đây được gọi là hàm hồi quy của mẫu.

Ước lượng SRF: Chọn 1 phương pháp nào đó để ước lượng các tham số của F qua việc tìm các tham số $\hat{F}$ và lấy giá trị quan sát của các tham số này làm giá trị xấp xỉ cho tham số của F .

2.3. Phương pháp bình phương nhỏ nhất

2.3.1. Tư tưởng của phương pháp bình phương nhỏ nhất

Trong kinh tế lượng, thủ tục ước lượng được dùng phổ biến nhất là phương pháp bình phương nhỏ nhất. Tiêu chuẩn tối ưu được sử dụng bởi phương pháp bình phương nhỏ nhất là cực tiểu hóa hàm mục tiêu.

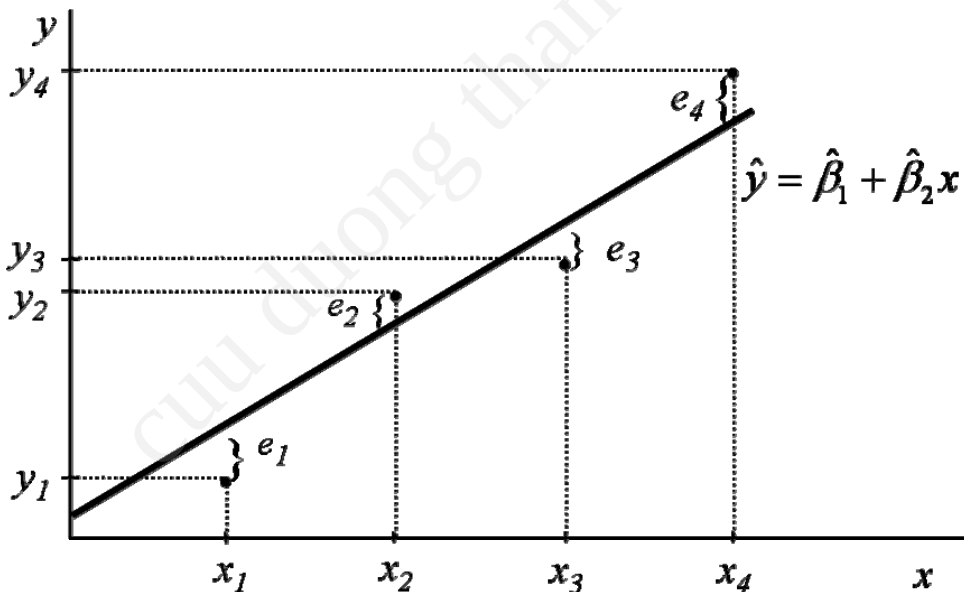
Phương pháp bình phương nhỏ nhất là một phương pháp được đưa ra bởi nhà toán học Đức Carl Friedrich Gauss, đây là một phương pháp mạnh và được rất nhiều người sử dụng, nó thường được ký hiệu là OLS (ordinary least squares). Tư tưởng của phương pháp này là cực tiểu tổng bình phương các phần dư. Do đó có thể nói để có được đường hồi qui “thích hợp” nhất, chúng ta chọn các ước lượng của tung độ gốc và độ dốc sao cho phần dư là nhỏ.

Chúng ta đặt:

y_i ký hiệu giá trị thực của biến y tại quan sát i

$\hat{y}_i$ ký hiệu giá trị của hàm hồi qui mẫu

e_i ký hiệu phần dư, $y_i - \hat{y}_i$



Do đó cực tiểu hoá $\sum (y_i - \hat{y}_i)^2$ sẽ tương đương với cực tiểu $\sum e_i^2$ từ đó tìm ra β_1 và β_2 .

Chúng ta có thể mô tả tổng quát như sau:

Xét hàm hồi qui tổng thể (PRF): $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k + u$

$$\beta_1 = E(Y | x_2, x_3, \dots, x_k = 0)$$

$$\beta_j = \frac{\partial Y}{\partial x_j}$$

Với mẫu:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \dots + \hat{\beta}_k X_k$$

Chúng ta sẽ sử dụng phương pháp bình phương nhỏ nhất để tìm các tham số $\hat{\beta}_j$ của hàm hồi quy mẫu.

2.3.2. Các giả thiết của phương pháp bình phương nhỏ nhất

Phương pháp bình phương nhỏ nhất (OLS) là phương pháp rất đáng tin cậy trong việc ước lượng các tham số của mô hình, tuy nhiên mô hình ước lượng phải thỏa mãn 6 giả thiết. Khi thỏa mãn các giả thiết, ước lượng bình phương nhỏ nhất (OLS) là ước lượng tuyến tính không chệch có hiệu quả nhất trong các ước lượng. Vì thế phương pháp OLS đưa ra **Ước Lượng Không chệch Tuyến Tính Tốt Nhất (BLUE)**.

Kết quả này được gọi là **Định lý Gauss–Markov**, theo lý thuyết này ước lượng OLS là BLUE; nghĩa là trong tất cả các tổ hợp tuyến tính không chệch của Y , ước lượng OLS có phương sai bé nhất. Các giả thiết như sau:

+ **Mô hình hồi quy là tuyến tính theo các hệ số:**

Điều này có nghĩa là quá trình thực hành hồi quy trên thực tế được miêu tả bởi mối quan hệ dưới dạng:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u$$

hoặc mối quan hệ thực tế đó có thể được viết lại ví dụ như dưới dạng lấy loga cả hai vế.

+ **$E(u_i) = 0$, kỳ vọng của các yếu tố ngẫu nhiên u_i bằng 0.**

Trung bình tổng thể sai số là bằng 0. Điều này có nghĩa là có một giá trị sai số mang dấu dương và một số sai số mang dấu âm. Do $\beta_1 + \beta_2 X_i$ là đường trung bình, nên có thể giả định rằng các sai số ngẫu nhiên trên sẽ bị loại trừ nhau, ở mức trung bình, *trong* tổng thể.

+ **$Cov(u_i, u_j) = 0$, Không có sự tương quan giữa các u_i**

Không có sự tương quan giữa các quan sát của yếu tố sai số (không có tương quan chuỗi). Nếu chúng ta xem xét các chuỗi số liệu thời gian (dữ liệu được thu thập từ một nguồn trong nhiều khoảng thời gian khác nhau). Yếu tố sai số u_i trong khoảng thời gian này không có bất kỳ một tương quan nào với yếu tố sai số trong khoảng thời gian trước đó.

+ $\text{Cov}(u_i, x_i) = 0$, U và X không tương quan với nhau

Điều này có nghĩa là khi bất kỳ biến giải thích nào mà lớn hơn hay nhỏ đi thì yếu tố sai số sẽ không thay đổi theo nó.

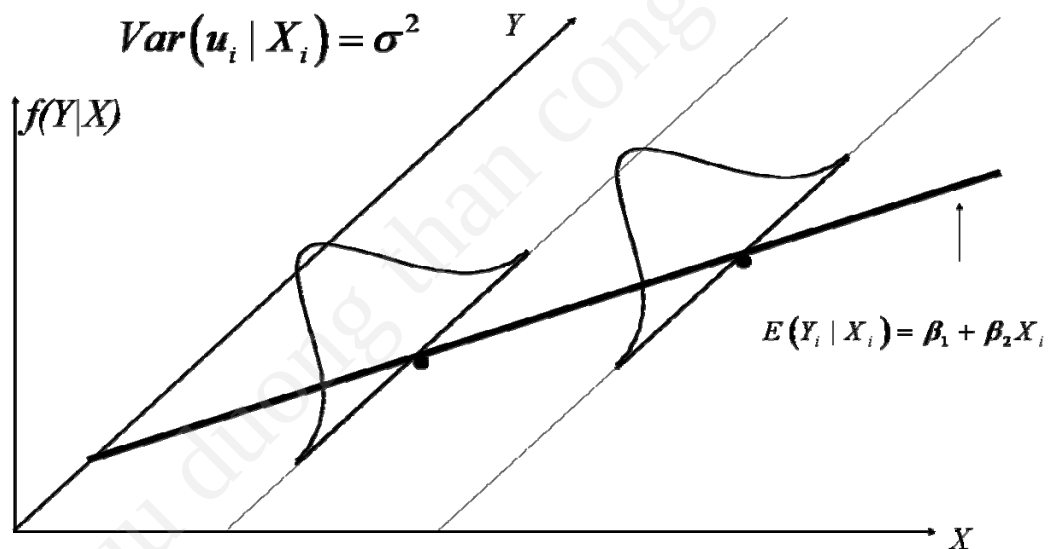
+ $\text{Var}(u_i) = \sigma^2$, Phương sai của sai số không đổi với mọi u_i

Tất cả giá trị u được phân phối giống nhau với cùng phương sai σ^2 , sao cho $\text{Var}(u_i) = E(u_i^2) = \sigma^2$. Điều này được gọi là phương sai của sai số không đổi.

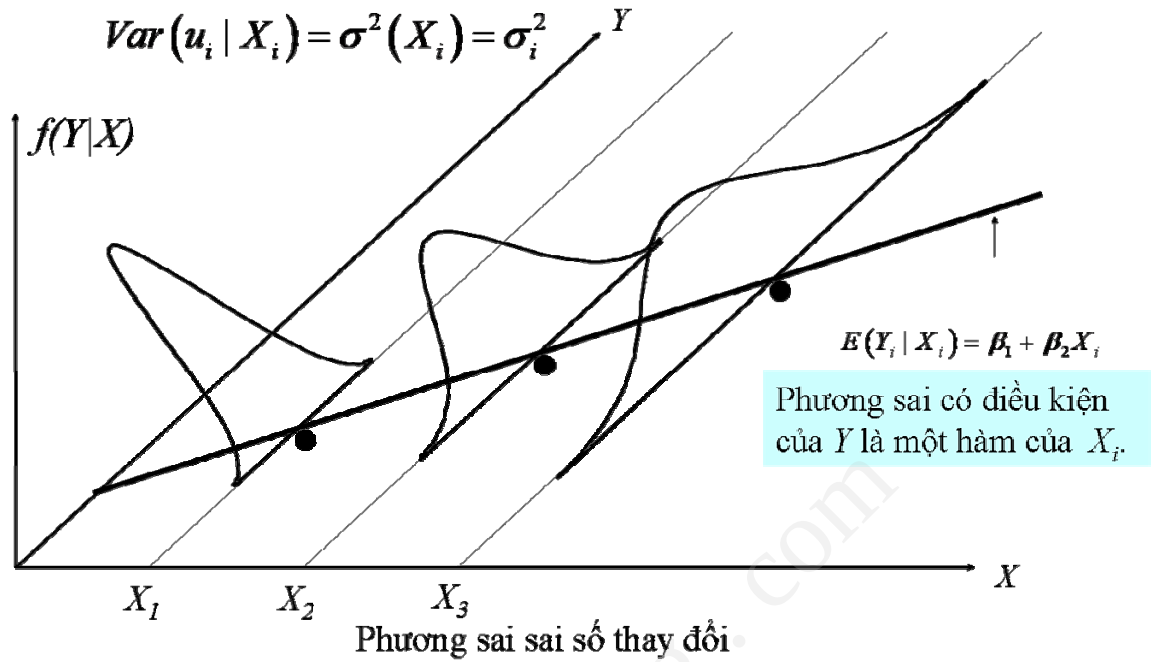
+ u_i Phân phối chuẩn

Điều này rất quan trọng khi phát sinh khoảng tin cậy và thực hiện kiểm định giả thuyết trong những phạm vi mẫu là nhỏ. Nhưng với phạm vi mẫu lớn hơn, điều này sẽ trở nên không mấy quan trọng.

Ví dụ về “Phương sai sai số không đổi” và “Phương sai sai số thay đổi”



Phương sai sai số không đổi



2.3.3. Ứng dụng phương pháp bình phương nhỏ nhất tìm tham số hồi qui:

Cho hàm hồi qui mẫu $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Ta đặt
$$L = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$$

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

$$\frac{\partial L}{\partial \hat{\beta}_1} = -2 \sum_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \quad (2.1)$$

$$\frac{\partial L}{\partial \hat{\beta}_2} = -2 \sum_i x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \quad (2.2)$$

$$\text{Từ (2.1), } \sum_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \Leftrightarrow \sum_i y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_i x_i = 0$$

Nhưng $\sum y_i = n\bar{y}$ và $\sum x_i = n\bar{x}$

$$\text{Do vậy ta có thể viết } n\bar{y} - n\hat{\beta}_1 - n\hat{\beta}_2 \bar{x} = 0 \text{ hay } \bar{y} - \hat{\beta}_1 - \hat{\beta}_2 \bar{x} = 0 \quad (2.3)$$

$$\text{Từ (2.2), } \sum_i x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \quad (2.4)$$

$$\text{Từ (2.3), } \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \quad (2.5)$$

Thay vào (2.4) với $\hat{\beta}_1$ từ (2.5), ta có:

$$\begin{aligned} \sum_i x_i (y_i - \bar{y} + \hat{\beta}_2 \bar{x} - \hat{\beta}_2 x_i) &= 0 \\ \sum_i x_i y_i - \bar{y} \sum_i x_i + \hat{\beta}_2 \bar{x} \sum_i x_i - \hat{\beta}_2 \sum_i x_i^2 &= 0 \\ \sum_i x_i y_i - n\bar{y}\bar{x} + \hat{\beta}_2 n\bar{x}^2 - \hat{\beta}_2 \sum_i x_i^2 &= 0 \end{aligned}$$

Tương đương với,

$$\hat{\beta}_2 (n\bar{x}^2 - \sum_i x_i^2) = n\bar{y}\bar{x} - \sum_i x_i y_i$$

Do vậy $\hat{\beta}_1, \hat{\beta}_2$ được xác định như sau:

$$\hat{\beta}_2 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \& \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$\hat{\beta}_1, \hat{\beta}_2$ 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.

2.3.4. Các tính chất của các tham số ước lượng: Định lý Gauss Markov

Với các giả thiết đã cho, thì các ước lượng $\hat{\beta}_1, \hat{\beta}_2$ được xác định bằng phương pháp bình phương nhỏ nhất là các ước lượng tuyến tính, không chệch tốt nhất (có phương sai nhỏ nhất).

- + “Ước lượng” - $\hat{\beta}$ là ước lượng điểm của β .
- + “Tuyến tính” - $\hat{\beta}$ là ước lượng tuyến tính (theo Y)
- + “Không chệch”- Giá trị kỳ vọng của $\hat{\beta}_1, \hat{\beta}_2$ đúng bằng giá trị của β_1, β_2
- + “Tốt nhất” - điều đó có nghĩa là ước lượng $\hat{\beta}$ có phương sai nhỏ nhất trong tất cả các lớp ước lượng tuyến tính không chệch. Chúng ta có thể chứng minh định lý Gauss-Markov.

2.4. Độ chính xác của ước lượng

Từ lý thuyết xác suất ta biết rằng phương sai của một biến ngẫu nhiên đo lường sự phân tán xung quanh giá trị trung bình. Phương sai càng bé, từng giá trị riêng biệt càng gần với giá trị trung bình. Tương tự, khi đề cập đến khoảng tin cậy, ta biết rằng phương sai của biến ngẫu nhiên càng nhỏ, khoảng tin cậy của các tham số càng bé. Như vậy, phương sai của một ước lượng là thông số để chỉ độ chính xác của một ước lượng. Do đó việc tính toán phương sai của $\hat{\beta}_1$ và $\hat{\beta}_2$ là rất cần thiết.

Do $\hat{\beta}_1$ và $\hat{\beta}_2$ thuộc vào các giá trị Y , mà Y lại phụ thuộc vào các biến ngẫu nhiên $u_1, u_2, \dots, u_n$, nên chúng cũng là biến ngẫu nhiên với phân phối tương ứng.

Với các giả thiết đã cho, phương sai và độ lệch chuẩn được tính như sau

$$\begin{aligned} Var(\hat{\beta}_2) &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}; & se(\hat{\beta}_2) &= \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} \\ Var(\hat{\beta}_1) &= \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \sigma^2; & se(\hat{\beta}_1) &= \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}} \sigma \end{aligned}$$

$$\sigma^2 = Var(u_i)$$

Trong các công thức trên σ^2 chưa biết, σ^2 được ước lượng bằng ước lượng không chệch của nó là:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - 2}$$

2.5. Độ phù hợp của mô hình

Để có thể biết mô hình giải thích được như thế nào hay bao nhiêu % biến động của biến phụ thuộc, người ta sử dụng R^2 .

Ta có:

$$\begin{aligned} \sum (y_i - \bar{y})^2 &= \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum [e_i + (\hat{y}_i - \bar{y})]^2 \\ &= \sum e_i^2 + 2 \sum e_i (\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2 \end{aligned}$$

Đặt:

$\sum (y_i - \bar{y})^2$: TSS - total sum of squares

$\sum (\hat{y}_i - \bar{y})^2$: ESS - explained sum of squares

$\sum e_i^2$: RSS - residual sum of squares

Do $\sum e_i (\hat{y}_i - \bar{y}) = 0 \Leftrightarrow [\sum e_i \hat{y}_i = 0; \sum e_i \bar{y} = 0]$

Do vậy ta có thể viết: $TSS = ESS + RSS$

Ý nghĩa của các thành phần:

- + TSS là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát Y_i và giá trị trung bình.
- + 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 qui mẫu và giá trị trung bình của chúng. Phần này đo độ chính xác của hàm hồi qui
- + 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 qui.
- + TSS được chia thành hai phần: một phần do ESS và một phần do RSS gây ra.

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

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} + \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

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

$$R^2 = \frac{(\sum (y_i - \bar{y})(\hat{y}_i - \bar{y}))^2}{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{y})^2} \Rightarrow R = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{y})^2}}$$

Tỉ số giữa tổng biến thiên được giải thích bởi mô hình cho tổng bình phương cần được giải thích được gọi là hệ số xác định, hay là trị thống kê “good of fit”. Từ định nghĩa R^2 chúng ta thấy R^2 đo tỷ lệ hay số % của toàn bộ sai lệch Y với giá trị trung bình được giải thích bằng mô hình. Khi đó người ta sử dụng R^2 để đo sự phù hợp của hàm hồi qui; $0 \leq R^2 \leq 1$

- + R^2 cao nghĩa là mô hình ước lượng được giải thích được một mức độ cao biến động của biến phụ thuộc.
- + Nếu R^2 bằng 0. Nghĩa là mô hình không đưa ra thông tin nào về biến phụ thuộc và dự đoán tốt nhất về giá trị của biến phụ thuộc là giá trị trung bình của nó. Các biến "giải thích" thực sự không đưa ra được một giải thích nào.

2.6. Ước lượng khoảng tin cậy và Kiểm định các hệ số hồi qui

2.6.1. Ước lượng khoảng tin cậy cho các hệ số hồi qui

a. Khái niệm chung:

Xét một tổng thể gồm N biến ngẫu nhiên X tuân theo luật phân phối có hàm mật độ xác suất là $f(x, \theta)$; trong đó θ là các tham số thống kê của tổng thể.

Ví dụ, trong phân phối chuẩn:

$$f(x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Rightarrow \theta = (\mu, \sigma^2), \quad -\infty < \mu < +\infty$$
$$0 < \sigma^2 < +\infty$$

Gọi $\{x_1, x_2, \dots, x_n\}$ là mẫu ngẫu nhiên, cỡ mẫu n được dùng lấy ra từ tổng thể tuân theo hàm mật độ xác suất $f(x, \theta)$. Ở đây dạng của hàm f xem như đã biết còn các tham số thống kê θ của tổng thể xem như chưa biết.

Vấn đề đặt ra ở phần này là dựa vào các mẫu quan sát $\{x_1, x_2, \dots, x_n\}$ ta ước lượng xem giá trị cụ thể của θ bằng bao nhiêu (bài toán đó gọi là ước lượng điểm) hoặc ước lượng xem θ nằm trong khoảng nào (bài toán ước lượng khoảng).

+ Ước lượng khoảng và giá trị ước lượng khoảng.

* Ước lượng khoảng: Ước lượng khoảng đối với tham số thống kê của tổng thể θ là một quy tắc dựa trên thông tin của mẫu để xác định miền (range) hay khoảng (interval) mà tham số θ hầu như nằm trong đó.

* Giá trị ước lượng khoảng: là giá trị cụ thể của miền hay khoảng mà tham số θ nằm trong đó.

+ Khoảng tin cậy và độ tin cậy:

Gọi θ là tham số thống kê chưa biết. Giả sử dựa trên thông tin của mẫu ta có thể xác định được 2 biến ngẫu nhiên A và B sao cho

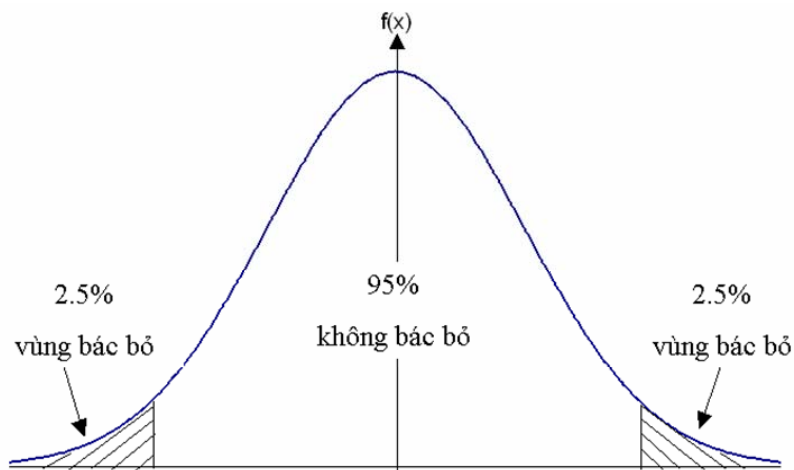
$$P(A < \theta < B) = 1 - \alpha \text{ với } 0 < \alpha < 1$$

Nếu giá trị cụ thể của biến ngẫu nhiên A và B là a và b thì khoảng (a, b) từ a đến b được gọi là khoảng tin cậy của θ với xác suất là $(1 - \alpha)$

Xác suất $(1 - \alpha)$ được gọi là độ tin cậy của khoảng.

Trong thực tế, độ tin cậy $(1 - \alpha)$ do nhà thống kê chọn theo yêu cầu của mình, thông thường độ tin cậy được chọn là 0,90; 0,95; 0,99...

α là mức ý nghĩa hay là xác suất sai lầm khi chọn khoảng tin cậy (a, b)



ta tìm được $t_{\alpha/2}(n-2)$ thỏa mãn

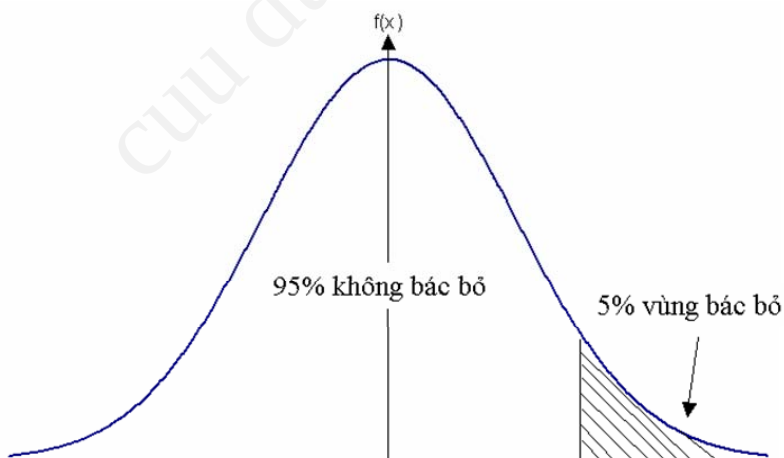
$$P(-t_{\alpha/2}(n-2) \leq \frac{\hat{\beta}_j - \beta_j}{Se(\hat{\beta}_j)} \leq t_{\alpha/2}(n-2)) = 1 - \alpha$$

Khoảng tin cậy $1-\alpha$ của β_j là:

$$\left[\hat{\beta}_j - t_{\alpha/2}(n-2)Se(\hat{\beta}_j); \hat{\beta}_j + t_{\alpha/2}(n-2)Se(\hat{\beta}_j) \right]$$

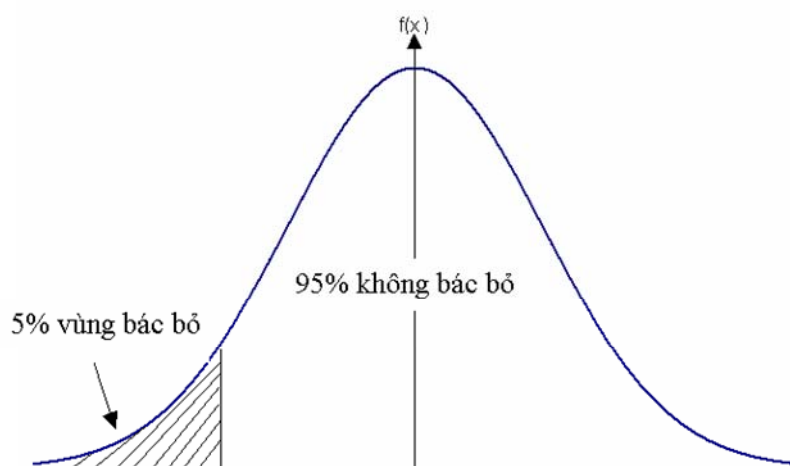
+ Ước lượng phía phải:

$$\beta_j \leq \hat{\beta}_j + t_{\alpha}(n-2)Se(\hat{\beta}_j)$$



+ Ước lượng phía trái:

$$\beta_j \geq \hat{\beta}_j - t_{\alpha}(n-2)Se(\hat{\beta}_j)$$



2.6.2. Kiểm định giả thuyết ý nghĩa của từng hệ số hồi qui

a. Khái niệm:

Giả thuyết thống kê là một giả sử hay một phát biểu có thể đúng, có thể sai liên quan đến tham số của tổng thể. Khi thực hiện kiểm định, người ta thiết lập cặp giả thiết thống kê, **Giả thuyết không** và **giả thuyết ngược lại** (giả thiết đối).

- + Giả thuyết không: là sự giả sử mà chúng ta muốn kiểm định thường được ký hiệu là H_0 .
- + Giả thuyết ngược lại: Việc bác bỏ giả thuyết không sẽ dẫn đến việc chấp nhận giả thuyết ngược lại. Giả thuyết ngược lại thường được ký hiệu là H_1 .

Ví dụ:

$$H_0 : \beta = 0.5$$

$$H_1 : \beta \neq 0.5$$

Miền bác bỏ và miền chấp nhận (rejection region & acceptance region)

Tất cả các giá trị có thể có của các đại lượng thống kê trong kiểm định có thể chia làm 2 miền: miền bác bỏ và miền chấp nhận.

- + Miền bác bỏ là miền chứa các giá trị làm cho giả thuyết H_0 bị bác bỏ.
- + Miền chấp nhận là miền chứa các giá trị giúp cho giả thuyết H_0 không bị bác bỏ.

Trong thực tế khi H_0 không bị bác bỏ cũng có nghĩa là nó được chấp nhận. Giá trị chia đôi hai miền được gọi là giá trị giới hạn (Critical value)

Giả thiết không và giả thiết đối có thể là giả thiết đơn hay giả thiết kép. Một giả thiết được gọi là đơn nếu nó đưa ra 1 giá trị cụ thể cho tham số (ví dụ $H_0: \beta=0.5$). Một giả thiết được gọi là kép nếu nó đưa ra một khoảng giá trị của phân bố xác suất (ví dụ $H_0: \beta > 0.5$). Liên quan đến vấn đề này người ta có kiểm định hai phía và kiểm định một phía.

Các bước kiểm định giả thuyết thống kê

Bước 1: Thành lập giả thuyết H_0

Ví dụ: $H_0: \beta = 0.5$

Bước 2: Thành lập giả thuyết H_1

Ví dụ: $H_1: \beta \neq 0.5$

Bước 3: Xác định mức ý nghĩa α

Bước 4: Chọn các tham số thống kê thích hợp cho việc kiểm định và xác định các miền bác bỏ, miền chấp nhận và giá trị giới hạn.

Bước 5: Tính toán biến ngẫu nhiên của kiểm định như biến Z (trong phân phối chuẩn), t (trong phân phối Student t) hay χ^2 (trong phân phối Chi bình phương).

Bước 6: Ra quyết định: Nếu các giá trị tính toán rơi vào miền bác bỏ H_0 thì ra quyết định bác bỏ H_0 . Ngược lại sẽ chấp nhận H_0 .

b. Kiểm định giả thiết đối với β_j

Có thể đưa ra giả thiết nào đó đối với β_j , chẳng hạn $\beta_j = \beta_j^*$. Nếu giả thiết này đúng thì:

$$T = \frac{\hat{\beta}_j - \beta_j}{\text{Se}(\hat{\beta}_j)} \sim T(n-2)$$

Ta có bảng sau đây:

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

Giả sử:

Kiểm định β_j ; $H_0: \beta_j = 0 \Leftrightarrow x_j$ không tác động

$H_1: \beta_j \neq 0 \Leftrightarrow x_j$ có tác động

$\beta_j < 0 \Leftrightarrow x_j$ có tác động ngược

$\beta_j > 0 \Leftrightarrow x_j$ có tác động thuận

2.6.3. Ước lượng Khoảng tin cậy và Kiểm định đối với σ^2

Ước lượng khoảng tin cậy đối với σ^2 :

Ta có:

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

Do đó khoảng tin cậy $1-\alpha$ của σ^2 được xác định từ:

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

Hay

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

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	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\frac{(n-2)\hat{\sigma}^2}{\sigma_0^2} > \chi^2_{\alpha/2}(n-2)$ hay $\frac{(n-2)\hat{\sigma}^2}{\sigma_0^2} < \chi^2_{1-\alpha/2}(n-2)$
Phía phải	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\frac{(n-2)\hat{\sigma}^2}{\sigma_0^2} > \chi^2_{\alpha}(n-2)$
Phía trái	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\frac{(n-2)\hat{\sigma}^2}{\sigma_0^2} < \chi^2_{1-\alpha}(n-2)$

2.7. Ứng dụng phân tích hồi qui

2.7.1. Dự báo giá trị trung bình của biến phụ thuộc

Giả sử ta biết rằng biến độc lập x và một giá trị x_0 nào đó mà ta cần đưa ra các kết luận về giá trị trung bình của biến phụ thuộc y , thì ta có:

$$E(y|x_0) = E(\beta_1 + \beta_2 x_0 + u_0) = \beta_1 + \beta_2 x_0$$

Khi đó đường hồi qui mẫu cho ước lượng điểm $E(y|x_0)$:

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0$$

$\hat{y}_0$ là ước lượng không chệch có phương sai nhỏ nhất của $E(y|x_0)$, tuy nhiên $\hat{y}_0$ vẫn khác giá trị thực của nó.

$\hat{y}_0$ có phân bố chuẩn với kỳ vọng $\beta_1 + \beta_2 x_0$ nên

$$\begin{aligned} \text{Var}(\hat{y}_0) &= E(\hat{\beta}_1 + \hat{\beta}_2 x_0 - \beta_1 - \beta_2 x_0)^2 \\ &= E[(\hat{\beta}_1 - \beta_1) + x_0(\hat{\beta}_2 - \beta_2)]^2 \\ &= E[(\hat{\beta}_1 - \beta_1)^2 + 2x_0(\hat{\beta}_2 - \beta_2)(\hat{\beta}_1 - \beta_1) + x_0^2(\hat{\beta}_2 - \beta_2)^2] \\ \text{Var}(\hat{y}_0) &= E[(\hat{\beta}_1 - \beta_1)^2] + E[x_0^2(\hat{\beta}_2 - \beta_2)^2] \\ &\quad + 2x_0 E[(\hat{\beta}_2 - \beta_2)(\hat{\beta}_1 - \beta_1)] \\ &= \text{var}(\hat{\beta}_1) + x_0^2 \text{var}(\hat{\beta}_2) + 2x_0 \text{cov}(\hat{\beta}_1, \hat{\beta}_2) \end{aligned}$$

$$\text{Var}(\hat{y}_0) = \text{var}(\bar{y} - \hat{\beta}_2 \bar{x}) + x_0^2 \text{var}(\hat{\beta}_2) + 2x_0 \text{cov}(\hat{\beta}_1, \hat{\beta}_2)$$

$$\text{Var}(\hat{y}_0) = \frac{\sigma^2}{n} + \bar{x}^2 \text{var}(\hat{\beta}_2) + x_0^2 \text{var}(\hat{\beta}_2) + 2x_0 \text{cov}(\hat{\beta}_1, \hat{\beta}_2)$$

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)$$

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

$$\begin{aligned} \text{Var}(\hat{y}_0) &= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2} + x_0^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2} - 2x_0 \bar{x} \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] \end{aligned}$$

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

$$T = \frac{\hat{y}_0 - (\beta_1 + \beta_2 x_0)}{Se(\hat{y}_0)} \sim T(n-2)$$

Khoảng tin cậy $1-\alpha$ của $E(y|x_0)$:

$$\begin{aligned} P(\hat{\beta}_1 + \hat{\beta}_2 x_0 - t_{\alpha/2}(n-2)Se(\hat{y}_0) \leq \beta_1 + \beta_2 x_0 \\ \leq \hat{\beta}_1 + \hat{\beta}_2 x_0 + t_{\alpha/2}(n-2)Se(\hat{y}_0)) = 1 - \alpha \\ \hat{y}_0 - t_{\alpha/2}(n-2)Se(\hat{y}_0) \leq E(y|x_0) \leq \hat{y}_0 + t_{\alpha/2}(n-2)Se(\hat{y}_0) \end{aligned}$$

2.7.2. Dự báo giá trị cá biệt của biến phụ thuộc

Giả sử chúng ta muốn dự báo giá trị cá biệt $y=y_0$ với $x=x_0$, khi đó ước lượng của y_0 là:

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0$$

$$E(e_0) = E(y_0 - \hat{y}_0) = E(y_0) - E(\hat{y}_0) = 0$$

$$\text{var}(y_0 - \hat{y}_0) = \text{var}(y_0) + \text{var}(\hat{y}_0)$$

$$\begin{aligned} &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] \end{aligned}$$

Ta có: $e_0 \sim N(0, \text{var}(e_0))$

Người ta chứng minh được:

$$\frac{e_0 - 0}{Se(e_0)} \sim T(n-2) \Rightarrow \frac{y_0 - \hat{y}_0}{Se(e_0)} \sim T(n-2)$$

Với giá trị $0 < \alpha < 1$ ta có thể tìm được giá trị $t_{\alpha/2}$ sao cho:

$$P\left(\left| \frac{y_0 - \hat{y}_0}{Se(e_0)} \right| \leq t_{\alpha/2}\right) = (1 - \alpha)$$

$$\Rightarrow P(\hat{y}_0 - t_{\alpha/2}(n-2)Se(e_0) \leq y_0 \leq \hat{y}_0 + t_{\alpha/2}(n-2)Se(e_0)) = 1 - \alpha$$

BÀI THỰC HÀNH

Cho bảng sau đây về lãi suất (Y) và tỷ lệ lạm phát (X) trong năm 1988 ở 9 nước. Giả sử rằng sự phụ thuộc $E(Y/X)$ có dạng tuyến tính. Hãy ước lượng hàm hồi qui và tính các đặc trưng của nó.

Y	11.9	9.40	7.50	4.00	11.30	66.30	2.20	10.30	7.60
X	7.20	4.00	3.10	1.60	4.80	51.00	2.00	6.60	4.40

Trước tiên với cách làm thủ công (không sử dụng các phần mềm chuyên dụng) để thực hiện hồi qui và tìm các đặc trưng của mô hình, ta lập bảng tính và tính như sau:

n	y_i	x_i	x_i^2	$x_i - \bar{X}$	$(x_i - \bar{X})^2$	$y_i - \bar{Y}$	$(y_i - \bar{Y})^2$	$(y_i - \bar{Y})(x_i - \bar{X})$	$\hat{y}_i$	e_i	e_i^2
1	11.90	7.20	51.84	-2.21	4.89	-2.60	6.76	5.75	11.74	0.16	0.03
2	9.40	4.00	16.00	-5.41	29.28	-5.10	26.01	27.60	7.74	1.66	2.76
3	7.50	3.10	9.61	-6.31	39.83	-7.00	49.00	44.18	6.61	0.89	0.78
4	4.00	1.60	2.56	-7.81	61.01	-10.50	110.25	82.02	4.74	-0.74	0.55
5	11.30	4.80	23.04	-4.61	21.26	-3.20	10.24	14.76	8.74	2.56	6.56
6	66.30	51.00	2601.00	41.59	1729.64	51.80	2683.24	2154.30	66.46	-0.16	0.03
7	2.20	2.00	4.00	-7.41	54.92	-12.30	151.29	91.16	5.24	-3.04	9.25
8	10.30	6.60	43.56	-2.81	7.90	-4.20	17.64	11.81	10.99	-0.69	0.47
9	7.60	4.40	19.36	-5.01	25.11	-6.90	47.61	34.58	8.24	-0.64	0.41
Tổng	130.50	84.70	2770.97	0.00	1973.85	0.00	3102.04	2466.14	130.50	0.00	20.83
TB	14.50	9.41									

Từ bảng tính trên, chúng ta dễ dàng tính được:

Các tham số hồi qui:

$$\hat{\beta}_2 = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sum (x_i - \bar{X})^2} = \frac{2466.14}{1973.85} = 1.249;$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 14.5 - 1.249 * 9.41 = 2.742$$

Hàm hồi qui mẫu:

Từ các tham số hồi qui ở trên, hàm hồi qui mẫu được ước lượng là:

$$\hat{Y}_i = 2.742 + 1.249X_i \quad \text{SRF}$$

Độ chính xác của các ước lượng:

Để tính độ chính xác của các ước lượng, do σ^2 chưa biết, nên ta phải tính ước lượng không chệch của nó, ước lượng không chệch tính được như sau:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{20.83}{9-2} = 2.975$$

Từ đó suy ra:

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} = \frac{2.975}{1973.85} = 0.0015; \quad Se(\hat{\beta}_2) = 0.0388$$

$$Var(\hat{\beta}_1) = \frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2} \sigma^2 = \frac{2770.97}{9 \cdot 1973.85} 2.975 = 0.464; \quad Se(\hat{\beta}_1) = 0.6811$$

Độ phù hợp của mô hình:

Từ bảng kết quả, chúng ta cũng tính được độ phù hợp của mô hình như sau:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{20.83}{3102.04} = 0.993$$

Từ đó cho thấy rằng 99.3% sự biến đổi của lãi suất tiết kiệm trong mẫu được giải thích bằng tỷ lệ lạm phát.

Khoảng tin cậy các hệ số β_j :

Với $\alpha=0.1 \Rightarrow$ tra bảng ta có $t_{\alpha/2(n-2)} = 1.895$.

Vậy khoảng tin cậy với β_1 và β_2 với độ tin cậy 90% là:

$$\hat{\beta}_1 - 1.895 Se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 1.895 Se(\hat{\beta}_1)$$

$$1.131 \leq \beta_1 \leq 4.353$$

$$\hat{\beta}_2 - 1.895 Se(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + 1.895 Se(\hat{\beta}_2)$$

$$1.158 \leq \beta_2 \leq 1.341$$

Kiểm định giả thiết:

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

Giả thiết H_0 về mặt kinh tế tức là chúng ta đưa ra giả thiết biến X không ảnh hưởng đến Y, trong thí dụ này có nghĩa là lạm phát không ảnh hưởng đến lãi suất ngân hàng.

$$t = \frac{\hat{\beta}_2 - \beta_2^*}{se(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - 0}{0.0388} = 32.2$$

Ta thấy $t > t_{\alpha/2(n-2)}$, do đó bác bỏ giả thuyết H_0 , hay có thể nói là ở mức ý nghĩa 10% ta bác bỏ giả thiết cho rằng lạm phát không ảnh hưởng đến lãi suất ngân hàng.

Dự báo:

Giả sử chúng ta muốn dự báo giá trị trung bình hay giá trị cá biệt cho Lãi suất tiết kiệm khi chúng ta biết một giá trị cụ thể của Tỷ lệ lạm phát, áp dụng những công thức như đã trình bày trong phần 2.7 chúng ta dễ dàng tính được các giá trị dự báo mong muốn.

Chúng ta vừa thực hiện ước lượng hàm hồi qui và tính các đặc trưng của nó bằng việc sử dụng những công thức và lập bảng tính, hiện nay có rất nhiều phần mềm ứng dụng hay các phần mềm phân tích dữ liệu khác do đó chúng ta dễ dàng tính được các tham số hồi qui cũng như những đặc trưng của nó mà không cần mất quá nhiều thời gian. Đối với những yêu cầu đơn giản, chúng ta cũng có thể thực hiện ngay trên EXCEL, ví dụ với bài thực hành trên chúng ta có thể thực hiện một số thao tác đơn giản như sau:

Tool → Data Analysis → Regression

Sau khi thực hiện khai báo các biến, chúng ta sẽ thu được kết quả hồi qui như sau:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.997
R Square	0.993
Adjusted R Square	0.992
Standard Error	1.725
Observations	9

ANOVA

	df	SS	MS	F	Significance F
Regression	1	3081.212	3081.212	1035.543	0.000
Residual	7	20.828	2.975		
Total	8	3102.040			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.742	0.681	4.024	0.005	1.131	4.353
X	1.249	0.039	32.180	0.000	1.158	1.341

Dựa vào kết quả hồi qui trên, chúng ta dễ dàng thực hiện những phân tích, đánh giá mô hình hồi qui nhận được.

CHƯƠNG III

MÔ HÌNH HỒI QUI ĐA BIẾN

Trong Chương II chúng ta giới hạn trong trường hợp đơn giản của mô hình là mô hình hồi qui hai biến hay hồi qui đơn. Trong lý thuyết cũng như trong thực tế, có nhiều trường hợp mà biến kinh tế cho không thể giải thích bằng các mô hình hồi qui đơn như vậy, ví dụ Lượng cầu phụ thuộc vào giá, thu nhập, giá các hàng hoá khác; Giá nhà ở phụ thuộc vào diện tích nhà, số phòng ngủ và số phòng tắm; Chi tiêu của hộ gia đình về thực phẩm phụ thuộc vào qui mô hộ gia đình, thu nhập, vị trí địa lý; Tỷ lệ tử vong trẻ em của quốc gia phụ thuộc vào thu nhập bình quân đầu người, trình độ giáo dục hay Lương của một người phụ thuộc vào trình độ giáo dục, kinh nghiệm, giới tính, độ tuổi ...

Bây giờ, chúng ta sẽ xem xét **hồi qui đa biến (hồi qui bội)**, nghĩa là liên hệ biến phụ thuộc Y cho trước với nhiều biến độc lập $X_2, X_3, \dots, X_k$.

3.1. Giới thiệu mô hình hồi qui đa biến

Mô hình hồi qui tuyến tính đa biến có dạng tổng quát như sau:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u \text{ PRF} \quad (3.1)$$

Trong đó:

β_1 : là hệ số tự do (hệ số chặn)

β_j : là hệ số hồi qui riêng

u : sai số ngẫu nhiên

3.2. Các giả thiết của mô hình hồi qui đa biến

Các giả thiết OLS cho mô hình hồi qui tuyến tính đơn được giải thích trong mô hình hồi qui đa biến:

1. Hàm hồi qui là tuyến tính theo các tham số
2. $E(u_i) = 0$ Kỳ vọng của các yếu tố ngẫu nhiên u_i bằng 0
3. $\text{Var}(u_i) = \sigma^2$ Phương sai bằng nhau và thuần nhất với mọi u_i
4. $\text{Cov}(u_i, u_j) = 0$ Không có sự tương quan giữa các u_i
5. $\text{Cov}(u_i, x_i) = 0$ U và X không tương quan với nhau
6. u_i Phân phối chuẩn

Giả thiết bổ sung cho mô hình hồi qui đa biến:

7. Giữa các $x_2, x_3, \dots, x_k$ không có quan hệ tuyến tính. Nếu $x_2, x_3, \dots, x_k$ có quan hệ tuyến tính thì người ta nói rằng có hiện tượng đa cộng tuyến.

Hay không tồn tại $\lambda_i \equiv 0$: $\lambda_1 x_{1i} + \lambda_2 x_{2i} + \lambda_3 x_{3i} + \dots + \lambda_k x_{ki} + v_i = 0$

Về mặt hình thức, nếu có hiện tượng đa cộng tuyến giữa các biến giải thích và trong mô hình có tất cả các biến này thì không thể tách được ảnh hưởng của từng biến lên biến phụ thuộc y .

Giả sử: $x_{2i} = 3x_{3i}$, Khi đó:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

$$= \beta_1 + (3\beta_2 + \beta_3)x_{3i} + u_i; \text{ đặt: } \beta_3' = 3\beta_2 + \beta_3 \Rightarrow y_i = \beta_1 + \beta_3' x_{3i} + u_i$$

$\Rightarrow$ Không thể tách được ảnh hưởng của β_2 và β_3

3.3. Ước lượng các tham số mô hình hồi qui đa biến

Ta có hàm hồi qui mẫu tổng quát được viết dưới dạng như sau:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \dots + \hat{\beta}_k x_{ki}$$

Các phần dư được định nghĩa giống như trong mô hình hồi qui đơn:

$$e_i = y_i - \hat{y}_i$$

Để ước lượng các tham số của mô hình, chúng ta sử dụng phương pháp bình phương nhỏ nhất-OLS (như đã giới thiệu ở Chương II) từ đó tìm ra $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \dots, \hat{\beta}_k$

Ta có:

$$\sum e_i^2 = \sum (y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \dots + \hat{\beta}_k x_{ki}))^2$$

Chúng ta có thiết lập các điều kiện bậc nhất cho phép tính tối thiểu này như sau :

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_1} = -2 \sum (y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \dots + \hat{\beta}_k x_{ki})) = 0$$

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_2} = -2 \sum (y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \dots + \hat{\beta}_k x_{ki})) x_{2i} = 0$$

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_k} = -2 \sum (y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \dots + \hat{\beta}_k x_{ki})) x_{ki} = 0$$

Hệ phương trình mà chúng ta có được gọi là hệ phương trình chuẩn. Chúng ta có thể giải K phương trình chuẩn này để tìm K hệ số beta chưa biết.

3.4. Trường hợp hàm hồi qui có 2 biến giải thích

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

$$\hat{\beta}_1 + \hat{\beta}_2 \bar{x}_2 + \hat{\beta}_3 \bar{x}_3 = \bar{y}$$

$$\hat{\beta}_1 \sum x_{2i} + \hat{\beta}_2 \sum x_{2i}^2 + \hat{\beta}_3 \sum x_{2i} x_{3i} = \sum y_i x_{2i}$$

$$\hat{\beta}_1 \sum x_{3i} + \hat{\beta}_2 \sum x_{2i} x_{3i} + \hat{\beta}_3 \sum x_{3i}^2 = \sum y_i x_{3i}$$

Trong đó:

$$\bar{x}_2 = \sum x_{2i} / n; \bar{x}_3 = \sum x_{3i} / n; \bar{y} = \sum y_i / n$$

Giải hệ phương trình ta tìm được:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}_2 - \hat{\beta}_3 \bar{x}_3$$

$$\hat{\beta}_2 = \frac{(\sum (y_i - \bar{y})(x_{2i} - \bar{x}_2))(\sum (x_{3i} - \bar{x}_3)^2) - (\sum (y_i - \bar{y})(x_{3i} - \bar{x}_3))(\sum (x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3))}{(\sum (x_{2i} - \bar{x}_2)^2)(\sum (x_{3i} - \bar{x}_3)^2) - (\sum (x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3))^2}$$

$$\hat{\beta}_3 = \frac{(\sum (y_i - \bar{y})(x_{3i} - \bar{x}_3))(\sum (x_{2i} - \bar{x}_2)^2) - (\sum (y_i - \bar{y})(x_{2i} - \bar{x}_2))(\sum (x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3))}{(\sum (x_{2i} - \bar{x}_2)^2)(\sum (x_{3i} - \bar{x}_3)^2) - (\sum (x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3))^2}$$

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

Phương sai và độ lệch chuẩn được tính bởi công thức sau:

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum (x_{2i} - \bar{x}_2)^2 (1 - r_{23}^2)}; \quad \text{Se}(\hat{\beta}_2) = \sqrt{\text{Var}(\hat{\beta}_2)}$$

$$\text{Var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum (x_{3i} - \bar{x}_3)^2 (1 - r_{23}^2)}; \quad \text{Se}(\hat{\beta}_3) = \sqrt{\text{Var}(\hat{\beta}_3)}$$

$$\text{Cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}^2 \sigma^2}{(1 - r_{23}^2) \sqrt{\sum (x_{2i} - \bar{x}_2)^2} \sqrt{\sum (x_{3i} - \bar{x}_3)^2}}$$

$$r_{23}^2 = \frac{(\sum (x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3))^2}{\sum (x_{2i} - \bar{x}_2)^2 \sum (x_{3i} - \bar{x}_3)^2}$$

Trong công thức trên σ^2 là phương sai của u_i , nhưng chưa biết. Ước lượng không chệch của nó bằng:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - 3}$$

3 là số tham số trong mô hình, trường hợp mô hình có k tham số $\beta_1, \beta_2, \beta_3, \dots, \beta_k$ thì :

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - k}$$

3.5. Hệ số xác định bội R^2 và hệ số xác định bội đã hiệu chỉnh $\bar{R}^2$

Khi đánh giá mức độ thích hợp, tổng bình phương toàn phần, tổng bình phương hồi qui, và tổng bình phương của sai số có cùng dạng như đã trình bày trước, và ở đây cũng có $TSS = RSS + ESS$

Trong mô hình hồi qui hai biến R^2 đo độ thích hợp của hàm hồi qui. 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 qui đa biến tỷ lệ của toàn bộ sự khác biệt của biến y do tất cả các biến x_2 và x_3 gây ra được gọi là hệ số xác định bội, ký hiệu là R^2 :

$$R^2 = \frac{\hat{\beta}_2 \sum (y_i - \bar{y})(x_{2i} - \bar{x}_2) + \hat{\beta}_3 \sum (y_i - \bar{y})(x_{3i} - \bar{x}_3)}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

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

Tuy nhiên, định nghĩa R^2 theo cách này sẽ phát sinh một vấn đề. Có thể thấy là việc thêm vào bất kỳ một biến nào (dù biến này có ý nghĩa hay không) thì R^2 cũng sẽ không bao giờ giảm. Chúng ta dễ thấy điều này bằng lý luận theo trực giác. Khi một biến mới được thêm vào và RSS được cực tiểu, chúng ta đang cực tiểu theo một tập rất nhiều biến số và vì vậy RSS mới có vẻ sẽ nhỏ hơn (ít nhất thì cũng không lớn hơn). Cụ thể hơn, giả sử số hạng $\beta_{k+1}X_{k+1,i}$ được thêm vào phương trình (3.1) và ta có được một mô hình mới. Nếu giá trị *cực tiểu* của tổng bình phương của mô hình mới này *lớn hơn* giá trị của mô hình cũ, thì ta đặt β_{k+1} bằng không và sử dụng các ước lượng cũ cho các giá trị β khác sẽ tốt hơn, và vì vậy các ước lượng mới không thể có RSS cực tiểu. Điều này kéo theo khi một biến mới được thêm vào, giá trị R^2 tương ứng không thể giảm đi mà còn có thể tăng thêm. Do vậy, người ta thường cố gắng thêm một biến mới vào chỉ để tăng R^2 không kể đến mức độ quan trọng của biến đó đối với vấn đề đang giải quyết.

Để ngăn chặn tình trạng “có đưa thêm biến vào mô hình” như đã nêu trên, một phép đo khác về mức độ thích hợp được sử dụng thường xuyên hơn. Phép đo này gọi là **R^2 hiệu chỉnh** hoặc **R^2 hiệu chỉnh theo bậc tự do** (kết quả này luôn được in ra khi thực hiện hồi qui bằng những phần mềm chuyên dụng). Để phát triển phép đo này, trước hết phải nhớ là R^2 đo lường tỷ số giữa phương sai của Y “được giải thích” bằng mô hình; một cách tương đương, nó bằng một trừ đi tỷ số “không được giải thích” do phương sai của sai số $Var(u)$. Ta có thể biểu diễn công thức tính $\bar{R}^2$ như sau:

$$\bar{R}^2 = 1 - \frac{Var(u)}{Var(Y)}$$

Chúng ta biết rằng một ước lượng không chệch của $Var(u)$ được tính bằng $RSS/(n-k)$, và một ước lượng không chệch của $Var(Y)$ được tính bằng $TSS/(n-1)$. Thay vào phương trình trên ta có:

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - \frac{RSS(n-1)}{TSS(n-k)} \\ &= 1 - \frac{(n-1)}{(n-k)}(1 - R^2) = 1 - \frac{\hat{\sigma}^2(n-1)}{TSS}\end{aligned}$$

Việc thêm vào một biến dẫn đến tăng R^2 nhưng cũng làm giảm đi một bậc tự do, bởi vì chúng ta đang ước lượng thêm một tham số nữa. R^2 hiệu chỉnh là một phép đo độ thích hợp tốt hơn bởi vì nó cho phép đánh đổi giữa việc tăng R^2 và giảm bậc tự do. Cũng cần lưu ý là vì $(n-1)/(n-k)$ không bao giờ nhỏ hơn 1. R^2 hiệu chỉnh sẽ không bao giờ lớn hơn R^2 . Tuy nhiên, mặc dù R^2 không thể âm, R^2 hiệu chỉnh có thể nhỏ hơn không. Ví dụ, khi $n = 26$, $k = 6$, và $R^2 = 0,1$, chúng ta có R^2 hiệu chỉnh $= -0,125$.

Từ đó chúng ta thấy rằng không thể dùng R^2 làm tiêu chuẩn để xem xét việc không đưa thêm hay đưa thêm một biến giải thích mới vào trong mô hình. Do vậy $\bar{R}^2$ thường được sử dụng để cân nhắc khi xem xét việc đưa thêm một biến giải thích mới vào mô hình. Chúng ta chỉ đưa thêm biến mới vào mô hình khi $\bar{R}^2$ tăng.

3.6. Khoảng tin cậy và kiểm định các hệ số hồi qui

3.6.1. Ước lượng khoảng tin cậy đối với β_j

Hồi qui tuyến tính bội cũng có các tính chất gần giống như hồi qui tuyến tính đơn nhưng bậc tự do đã thay đổi.

Các hệ số ước lượng tuân theo phân phối chuẩn

$$\hat{\beta}_j \sim N(\beta_j, Se(\hat{\beta}_j))$$

$$\frac{\hat{\beta}_j - \beta_j}{Se(\hat{\beta}_j)} \sim T(n-k)$$

Như trong trường hợp hồi qui đơn, ước lượng phương sai sai số dựa vào các phần dư bình phương tối thiểu. Trong đó k là số hệ số có trong phương trình hồi qui đa biến

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k}$$

+ **Ước lượng 2 phía**, ta tìm được $t_{\alpha/2}(n-3)$ thỏa mãn

$$P(-t_{\alpha/2}(n-3) \leq \frac{\hat{\beta}_j - \beta_j}{Se(\hat{\beta}_j)} \leq t_{\alpha/2}(n-3)) = 1 - \alpha$$

Khoảng tin cậy $1-\alpha$ của β_j là:

$$\left[\hat{\beta}_j - t_{\alpha/2}(n-3)Se(\hat{\beta}_j); \hat{\beta}_j + t_{\alpha/2}(n-3)Se(\hat{\beta}_j) \right]$$

3.6.2. Kiểm định giả thiết đối với β_j

Kiểm định ý nghĩa thống kê của các hệ số hồi qui có ý nghĩa hay không: kiểm định rằng biến giải thích có thực sự ảnh hưởng đến biến phụ thuộc hay không. Nói cách khác là hệ số hồi qui có ý nghĩa thống kê hay không. Cách làm tương tự như đối với hồi qui đơn, nhưng thay đổi độ tự do.

Ví dụ kiểm định hai phía, Giả thiết:

$$H_0 : \beta_j = \beta_0$$

$$H_1 : \beta_j \neq \beta_0$$

Luật quyết định: $|t_c| > t_{(\alpha/2, n-k)}^*$ bác bỏ giả thiết không

Chúng ta có thể sử dụng giá trị p-value trong EViews. Nếu p-value tính được nhỏ hơn mức ý nghĩa thì chúng ta bác bỏ giả thiết không.

3.7. Kiểm định mức độ ý nghĩa chung của mô hình (trường hợp đặc biệt của KIỂM ĐỊNH WALD)

3.7.1. Khái quát về kiểm định WALD:

Giả sử chúng ta có 2 mô hình dưới đây:

$$(U) : Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

$$(R) : Y = \beta_1 + \beta_2 X_2 + v$$

Mô hình U được gọi là **mô hình không giới hạn** (unrestrict), và Mô hình R được gọi là **mô hình giới hạn** (restrict). Đó là do β_3 và β_4 buộc phải bằng không trong Mô hình R. Ta có thể kiểm định giả thuyết liên kết $\beta_3 = \beta_4 = 0$ với giả thuyết đối là ít nhất một trong những hệ số này không bằng không. Kiểm định giả thuyết liên kết này được gọi là **kiểm định Wald**. Thủ tục như sau.

Đặt các mô hình giới hạn và không giới hạn là:

$$(U) : Y = \beta_1 + \beta_2 X_2 + \dots + \beta_m X_m + \beta_{m+1} X_{m+1} + \dots + \beta_k X_k + u$$

$$(R) : Y = \beta_1 + \beta_2 X_2 + \dots + \beta_m X_m + v$$

Mô hình R có được bằng cách bỏ bớt một số biến ở Mô hình U, đó là $X_{m+1}, X_{m+2}, \dots, X_k$. Vì vậy, *giả thuyết* là:

$$H_0 : \beta_{m+1} = \dots = \beta_k = 0$$

$$H_1 : \text{không phải đồng thời các tham số bằng 0}$$

Lưu ý rằng (U) chứa k hệ số hồi qui chưa biết và (R) chứa m hệ số hồi qui chưa biết. Do đó, Mô hình R có ít hơn $k - m$ thông số so với U. Câu hỏi chúng ta sẽ nêu ra là $k - m$ biến bị loại ra có ảnh hưởng *liên kết* có ý nghĩa đối với Y hay không.

Giả sử những biến bị loại này không có ảnh hưởng có ý nghĩa đối với Y. Chúng ta sẽ không kỳ vọng tổng bình phương sai số của Mô hình R (RSSR) quá khác biệt với tổng bình phương sai số của Mô hình U (RSSU). Nói cách khác, sai biệt $RSSR - RSSU$ có vẻ rất nhỏ. Nhưng giá trị này nhỏ như thế nào? Chúng ta biết là RSS rất nhạy với đơn vị đo lường, và vì vậy có thể làm giá trị này lớn hơn hay nhỏ hơn chỉ đơn giản bằng cách thay đổi thang đo. “Nhỏ” hoặc “lớn” được xác định bằng cách so sánh sai biệt trên với RSSU, tổng bình phương sai số của mô hình hoàn toàn không giới hạn. Vì vậy, $RSSR - RSSU$ được so sánh với RSSU. Nếu giá trị đầu “nhỏ” tương đối so với giá trị sau, chúng ta kết luận là việc loại bỏ các biến $X_{m+1}, X_{m+2}, \dots, X_k$ không thay đổi RSS đủ để có thể tin là các hệ số của chúng có ý nghĩa.

Do $RSSU/\sigma^2$ có phân phối chi bình phương với $n - k$ bậc tự do (n quan sát trừ k thông số trong Mô hình U). Có thể thấy $(RSSR - RSSU)/\sigma^2$ cũng là phân phối chi bình phương (vì tính chất cộng của chi bình phương) với bậc tự do bằng số biến số loại bỏ trong (R). Do tỷ số của hai phân bố chi bình phương độc lập có phân phối F có hai thông số: bậc tự do cho tử số của tỷ số, bậc tự do cho mẫu số. Trị thống kê sẽ căn cứ trên tỷ số F .

Thủ tục kiểm định như sau:

$$H_0 : \beta_{m+1} = \dots = \beta_k = 0$$

H_1 : không phải đồng thời các tham số bằng 0

Trị thống kê kiểm định đối với giả thiết này là :

$$F_c = \frac{[RSS_R - RSS_U]/(k-m)}{RSS_U/(n-k)} \sim F(\alpha, k-m, n-k)$$

$$= \frac{(R_U^2 - R_R^2)/(k-m)}{(1 - R_U^2)/(n-k)}$$

với R^2 là số đo độ thích hợp không hiệu chỉnh. Với *giả thuyết không*, F_c có phân phối F với $k-m$ bậc tự do đối với tử số và $n-k$ bậc tự do đối với mẫu số.

Nguyên tắc ra quyết định: Bác bỏ giả thiết không khi

$$F_c > F(\alpha, k-m, n-k)$$

Hoặc giá trị p-value của thống kê F nhỏ hơn mức ý nghĩa cho trước.

3.7.2. Kiểm định ý nghĩa của mô hình:

Trong mô hình hồi qui đa biến, giả thiết "không" cho rằng mô hình không có ý nghĩa được hiểu là tất cả các hệ số hồi qui riêng (các tham số độ dốc) đều bằng không.

Ứng dụng kiểm định Wald (thường được gọi là kiểm định F) được tiến hành cụ thể như sau:

Bước 1 *Giả thuyết không* là $H_0: H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$. Giả thuyết ngược lại là H_1 : có ít nhất một trong những giá trị β không bằng không.

Bước 2 Trước tiên hồi qui Y theo một số hạng không đổi và $X_2, X_3, \dots, X_k$, sau đó tính tổng bình phương sai số RSS_U . Kế đến tính RSS_R . Chúng ta đã định nghĩa phân phối F là tỷ số của hai biến ngẫu nhiên phân phối chi bình phương độc lập. Điều này cho ta trị thống kê:

$$F_c = \frac{[RSS_R - RSS_U]/(k-m)}{RSS_U/(n-k)} \sim F(\alpha, k-m, n-k)$$

Vì $H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$, dễ dàng thấy rằng trị thống kê kiểm định đối với giả thiết này sẽ là:

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

Bước 3 Từ số liệu trong bảng F tương ứng với bậc tự do $k-1$ cho tử số và $n-k$ cho mẫu số, và với mức ý nghĩa cho trước α , ta có $F^*(\alpha, k-1, n-k)$ sao cho diện tích bên phải của F^* là α .

Bước 4 Bác bỏ *giả thuyết không* ở mức ý nghĩa α nếu $F_c > F^*$. Đối với phương pháp giá trị p , tính giá trị $p = P(F > F_c | H_0)$ và bác bỏ *giả thuyết không* nếu giá trị p nhỏ hơn mức ý nghĩa.

3.8. Phương pháp ma trận giải bài toán hồi qui đa biến

Chúng ta thấy rằng trong hồi qui đa biến, nếu chúng ta sử dụng cách thức thay thế dần để tìm tham số hồi qui sẽ rất phức tạp và mất rất nhiều thời gian, đặc biệt đối với những mô hình có nhiều biến giải thích. Một cách đơn giản hơn để giải quyết đó là sử dụng phương pháp ma trận.

Giả sử chúng ta có hàm hồi qui tổng thể

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u \quad \text{PRF}$$

Chúng ta viết cho từng quan sát như sau:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \dots + \beta_k x_{k1} + u_1 \\ y_2 &= \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \dots + \beta_k x_{k2} + u_2 \\ &\vdots \\ y_n &= \beta_1 + \beta_2 x_{2n} + \beta_3 x_{3n} + \dots + \beta_k x_{kn} + u_n \end{aligned}$$

Viết dưới dạng ma trận ta có:

$$Y = X\beta + U$$

trong đó Y ma trận cấp $n \times 1$

X ma trận cấp $n \times k$

β ma trận cấp $k \times 1$

U ma trận cấp $n \times 1$

Với x_1 được biểu diễn là:

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Ví dụ nếu k bằng 2:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{21} \\ 1 & x_{22} \\ \vdots & \vdots \\ 1 & x_{2n} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Hàm hồi qui mẫu được viết như sau:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \dots + \hat{\beta}_k x_{ki}$$

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \dots + \hat{\beta}_k x_{ki} + e_i$$

Hay

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

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

Khi đó các ước lượng OLS tìm được bằng cách:

$$\sum e_i^2 = \sum (y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \dots + \hat{\beta}_k x_{ki}))^2 \Rightarrow \min$$

$$\begin{aligned} e'e &= \sum e_i^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - 2\hat{\beta}'X'Y + X'X\hat{\beta}^2 \end{aligned}$$

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

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

Từ đó chúng ta dễ dàng tính được các tham số của hàm hồi qui mẫu, cũng như việc ứng dụng các qui tắc tính phương sai và độ lệch chuẩn chúng ta cũng dễ dàng tìm được phương sai và độ lệch chuẩn của các tham số hồi qui.

BÀI THỰC HÀNH

Khi nghiên cứu số người sẽ di chuyển bằng xe buýt với nhiều yếu tố ảnh hưởng khác nhau. Người ta thu thập dữ liệu chéo cho 40 thành phố khắp nước Mỹ.

BUSTRAVL	FARE	GASPRICE	INCOME	POP	DENSITY	LANDAREA
2073.0	0.85	0.88	17293	537.1	4099	131.0
2136.1	0.75	1.03	17768	787.0	9798	80.3
1878.8	0.60	0.91	17823	587.1	12438	47.2
937.5	1.00	0.91	15163	338.0	8070	41.8
7343.3	0.50	0.97	17480	3090.0	13547	228.1
837.9	0.85	0.88	15329	399.0	5110	78.1
1648.0	1.00	0.91	16141	561.8	7110	79.0
739.1	0.75	0.89	15326	585.1	3234	180.9
1070.7	1.50	0.89	17115	1142.4	3431	333.0
274.6	1.50	0.89	17117	486.5	2027	240.2
312.9	0.75	0.87	16127	198.7	4113	48.4
1879.1	1.00	0.94	17242	549.8	4975	110.6
1941.0	0.60	0.99	17340	1253.0	8913	135.6
2317.6	1.50	0.87	15108	1603.0	2885	556.4
471.4	1.05	0.93	15809	741.2	2105	352.0
594.3	0.70	0.79	16321	490.4	1551	316.3
7632.9	0.60	0.93	18027	3478.9	7486	464.7
510.1	0.60	0.93	18023	423.3	8508	49.8
630.6	0.60	0.93	12349	304.0	4997	60.0
1650.9	1.00	1.03	17886	377.2	10994	34.3
1618.3	0.50	0.86	16537	664.0	6702	95.8
2009.8	1.15	0.96	13019	368.0	6714	55.1
1562.4	1.15	0.96	13019	265.0	5144	52.4
1139.4	0.60	0.88	13130	572.0	2832	199.4
13103.0	1.00	1.00	20513	7323.3	24288	301.5
3739.6	1.35	0.92	17409	1760.2	12944	136.0
525.7	0.75	0.91	15944	991.6	3059	324.0
2385.8	1.00	0.89	15207	396.6	8147	55.4
1698.5	1.15	0.93	15409	387.0	3751	103.3
544.0	1.00	0.87	17743	167.0	8309	18.9
1769.1	0.85	0.81	16309	495.9	8077	61.4
1065.0	0.50	0.85	15092	794.0	2318	262.7
803.1	1.25	0.98	18014	1027.2	3208	320.0
1616.7	0.75	0.90	21886	753.6	16240	46.4
146.5	0.75	0.90	20744	376.0	6988	53.9
18.1	0.75	0.90	21313	698.1	4422	158.0
2056.1	1.00	0.88	17539	548.3	3790	144.6
470.1	0.75	0.92	17633	295.7	3497	84.4
242.5	0.75	0.92	17643	259.8	4675	55.5
3933.5	0.60	0.96	15522	693.6	11068	62.7

Các biến được định nghĩa như sau:

BUSTRAVL = Mức độ giao thông bằng xe buýt ở đô thị tính theo ngàn hành khách mỗi giờ

FARE = Giá vé xe buýt tính bằng \$

GASPRICE = Giá một ga lông nhiên liệu tính bằng \$

INCOME = Thu nhập bình quân đầu người tính bằng \$

POP = Dân số thành phố tính bằng ngàn người

DENSITY = Mật độ dân số tính (người/dặm vuông)

LANDAREA = Diện tích thành phố (dặm vuông)

Đặc trưng tổng quát của mô hình, được xem có dạng như sau:

$$\text{BUSTRAV} = \beta_1 + \beta_2 \text{FARE} + \beta_3 \text{GASPRICE} + \beta_4 \text{INCOME} + \beta_5 \text{POP} + \beta_6 \text{DENSITY} + \beta_7 \text{LANDAREA} + u$$

Trước khi ước lượng mô hình, chúng ta sẽ xác định dấu của các biến, mức độ ưu tiên, cho các hệ số hồi qui. Ở đây chúng ta giả định những tiềm ẩn về phía cung không được xem là quan trọng. Bởi vì một sự gia tăng giá vé xe buýt có thể làm giảm nhu cầu đi xe buýt, nên chúng ta kỳ vọng β_2 sẽ âm. Trong lĩnh vực di chuyển, xe hơi sẽ là một thay thế đối với xe buýt, và vì vậy một sự gia tăng giá nhiên liệu có thể khiến một số người tiêu thụ chuyển sang đi xe buýt, vì vậy chúng ta kỳ vọng một hiệu ứng tích cực ở đây; nghĩa là β_3 sẽ dương. Khi thu nhập tăng, chúng ta kỳ vọng nhu cầu đối với hàng tiêu dùng cũng tăng lên, và vì vậy như thường lệ chúng ta kỳ vọng β_4 sẽ dương. Tuy nhiên, nếu hàng tiêu dùng thuộc loại hàng hóa “thấp cấp”, thì hiệu ứng thu nhập (nghĩa là, β_4) sẽ âm. Một sự gia tăng dân số hay mật độ dân số thường làm gia tăng nhu cầu di chuyển bằng xe buýt, vì vậy, chúng ta kỳ vọng β_5 và β_6 sẽ dương. Nếu diện tích đất tăng lên, thì thành phố sẽ trải rộng ra hơn và người tiêu thụ có thể thích dùng xe hơi như là phương tiện giao thông chính hơn, nếu đây là một tình huống thì β_7 được kỳ vọng sẽ âm.

Kết quả hồi qui được tính toán bởi phần mềm Eviews.

Mô hình 1:

Dependent Variable: BUSTRAVL

Method: Least Squares

Date: 06/07/08 Time: 00:05

Sample: 1 40

Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2744.680	2641.672	1.038994	0.3064
FARE	-238.6544	451.7281	-0.528314	0.6008
GASPRICE	522.1132	2658.228	0.196414	0.8455
INCOME	-0.194744	0.064887	-3.001294	0.0051
POP	1.711442	0.231364	7.397176	0.0000
DENSITY	0.116415	0.059570	1.954253	0.0592
LANDAREA	-1.155230	1.802638	-0.640855	0.5260
R-squared	0.921026	Mean dependent var		1933.175
Adjusted R-squared	0.906667	S.D. dependent var		2431.757
S.E. of regression	742.9113	Akaike info criterion		16.21666
Sum squared resid	18213267	Schwarz criterion		16.51221
Log likelihood	-317.3332	F-statistic		64.14338
Durbin-Watson stat	2.082671	Prob(F-statistic)		0.000000

Qua bảng kết quả chúng ta thấy rằng R^2 hiệu chỉnh là 0,907, có nghĩa rằng 90,7% sự biến đổi của BUSTRAVL được giải thích chung bởi các biến trong mô hình (đã được điều chỉnh do giảm bậc tự do của mô hình). Đối với một nghiên cứu chéo, R^2 hiệu chỉnh như vậy là khá cao.

Nhìn vào giá trị p -value (được cho ở cột cuối cùng), khi kiểm định hai phía cho thấy rằng chỉ INCOME, POP, và DENSITY có các hệ số có nghĩa ở mức 10%. Hằng số và các hệ số của FARE, GASPRICE, và LANDAREA không có ý nghĩa về mặt thống kê ngay cả ở mức lớn hơn.

Kiểm định ý nghĩa của mô hình bằng kiểm định F, ta thấy rằng $F = 64,14$ và mức ý nghĩa của F, $P_F < 0.1$ do đó ta có thể kết luận mô hình trên là có ý nghĩa.

Nhưng chúng ta sẽ phải làm gì với những hệ số không có ý nghĩa. Qui tắc chung là bỏ qua ý nghĩa của hằng số hoặc là không cần nó. Tuy nhiên, FARE, GASPRICE, và LANDAREA cần phải xem xét loại bỏ khỏi mô hình bởi vì không có bằng chứng chứng tỏ chúng có những ảnh hưởng có nghĩa lên BUSTRAVL.

Chúng ta có thể thực hiện một bỏ tất cả chúng, ước lượng một mô hình được giới hạn, và thực hiện kiểm định Wald F-test như đã trình bày. Để tạo thực hiện việc này, chúng ta lấy ra tổng bình phương sai số và số bậc tự do cho mô hình không giới hạn vừa mới được ước lượng. Tuy nhiên, chúng ta cũng cần cân trọng, vì việc cùng lúc loại bỏ một vài biến cũng có thể bỏ mất những biến có ý nghĩa hoặc là những biến quan trọng về mặt lý thuyết. Do đó, cách làm thận trọng và nhạy bén hơn là loại bỏ dần từng biến. Có một vài lý do đối với việc loại bỏ các biến với các hệ số không có nghĩa. Thứ nhất, một mô hình đơn giản hơn dễ diễn giải hơn một mô hình phức tạp. Thứ hai, việc bỏ bớt một biến làm tăng bậc tự do và vì vậy cải thiện sự chính xác của các hệ số còn lại. Cuối cùng, như chúng ta sẽ thấy trong chương tiếp theo, nếu các biến giải thích có tương quan chặt với nhau nó sẽ gây khó khăn cho sự diễn giải riêng từng hệ số. Việc loại trừ các biến làm giảm cơ hội nảy sinh những tương quan này và vì vậy nó làm cho việc diễn giải có ý nghĩa hơn.

Điểm bắt đầu cho quá trình loại bỏ là nhận diện biến có hệ số hồi qui ít có nghĩa nhất. Điều này được thực hiện bằng cách nhìn vào giá trị p -value cao nhất. Từ kết quả mô hình A, chúng ta để ý rằng hệ số cho GASPRICE có giá trị p -value cao nhất và vì vậy ít có ý nghĩa nhất. Do đó, biến này bị loại bỏ khỏi đặc trưng mô hình và chúng ta sẽ thực hiện hồi qui với những biến còn lại. Kết quả hồi qui thu được sau khi loại bỏ biến GASPRICE như sau:

Mô hình 2:

Dependent Variable: BUSTRAVL

Method: Least Squares

Date: 06/07/08 Time: 00:07

Sample: 1 40

Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3215.856	1090.469	2.949058	0.0057
FARE	-225.6595	440.4936	-0.512288	0.6118
INCOME	-0.195716	0.063777	-3.068778	0.0042
POP	1.716808	0.226474	7.580596	0.0000
DENSITY	0.118216	0.058023	2.037401	0.0495
LANDAREA	-1.195297	1.765554	-0.677010	0.5030
R-squared	0.920934	Mean dependent var	1933.175	
Adjusted R-squared	0.909307	S.D. dependent var	2431.757	
S.E. of regression	732.3323	Akaike info criterion	16.16783	
Sum squared resid	18234559	Schwarz criterion	16.42116	
Log likelihood	-317.3565	F-statistic	79.20400	
Durbin-Watson stat	2.079321	Prob(F-statistic)	0.000000	

Ta thấy rằng việc loại bỏ biến đã cải thiện các ràng buộc lựa chọn mô hình, đã cải thiện độ chính xác của các hệ số còn lại bằng cách làm cho chúng có ý nghĩa nhiều hơn, dễ dàng thấy được điều này qua giá trị p -value. Biến có hệ số ít ý nghĩa nhất (giá trị p -value cao nhất), bây giờ là FARE. Nhưng vé xe buýt là một thước đo giá cả mà theo cách nói lý thuyết kinh tế là một yếu tố quan trọng của nhu cầu.

Do đó, chúng ta không nên loại bỏ nó ngay cả khi giá trị p -value cho rằng chúng ta có thể bỏ. Do vậy bước kế tiếp ta loại bỏ LANDAREA, biến có giá trị p -value cao nhất kế tiếp. Tiếp tục thực hiện hồi qui, ta thu được:

Mô hình 3:

Dependent Variable: BUSTRAVL

Method: Least Squares

Date: 06/07/08 Time: 00:08

Sample: 1 40

Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3111.181	1071.067	2.904749	0.0063
FARE	-295.7306	424.8354	-0.696106	0.4910
INCOME	-0.202197	0.062564	-3.231821	0.0027
POP	1.588337	0.122654	12.94973	0.0000
DENSITY	0.149027	0.035713	4.172925	0.0002
R-squared	0.919868	Mean dependent var	1933.175	
Adjusted R-squared	0.910710	S.D. dependent var	2431.757	
S.E. of regression	726.6434	Akaike info criterion	16.13122	
Sum squared resid	18480373	Schwarz criterion	16.34233	
Log likelihood	-317.6243	F-statistic	100.4449	
Durbin-Watson stat	1.995180	Prob(F-statistic)	0.000000	

Ta thấy rằng biến DENSITY đã gia tăng ý nghĩa đáng kể. Tuy nhiên, biến FARE có giá trị p -value là 49%, quá cao không thể chấp nhận được. Điều này gợi ý rằng, với sự có mặt của các biến khác, giá cả có thể không ảnh hưởng lên nhu cầu đi xe buýt. Nói cách khác, khi có nhu cầu đi xe buýt, người tiêu thụ có thể không nhạy cảm lắm với giá cả.

Do vậy, loại bỏ FARE là cần thiết và lại tiếp tục hồi qui với những biến còn lại, ta thu được kết quả:

Mô hình 4:

Dependent Variable: BUSTRAVL

Method: Least Squares

Date: 06/07/08 Time: 00:09

Sample: 1 40

Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2815.703	976.3007	2.884053	0.0066
INCOME	-0.201273	0.062101	-3.241076	0.0026
POP	1.576575	0.120612	13.07148	0.0000
DENSITY	0.153421	0.034898	4.396311	0.0001
R-squared	0.918759	Mean dependent var		1933.175
Adjusted R-squared	0.911989	S.D. dependent var		2431.757
S.E. of regression	721.4228	Akaike info criterion		16.09497
Sum squared resid	18736228	Schwarz criterion		16.26386
Log likelihood	-317.8993	F-statistic		135.7080
Durbin-Watson stat	1.878671	Prob(F-statistic)		0.000000

Qua kết quả ta thấy rằng mô hình này có các trị thống kê lựa chọn mô hình thấp nhất và tất cả các hệ số đều có nghĩa rất lớn. Hơn nữa, các hệ số đối với INCOME, POP, và DENSITY không khác với các hệ số giữa mô hình 3 và mô hình 4. Vì vậy có thể kết luận việc loại bỏ FARE không quá nghiêm trọng.

Dựa trên tất cả các ràng buộc, mô hình 4 dường như là “tốt nhất” và được chọn như là mô hình cuối cùng cho việc diễn dịch.

Các hệ số của thu nhập, kích thước dân số, và mật độ dân số có ý nghĩa rất lớn. Lý thuyết kinh tế chuẩn cho rằng ảnh hưởng thu nhập lên nhu cầu đối với bất cứ hàng hóa nào đều dương, nhưng hệ số ước lượng của INCOME thì lại âm. Điều này, không gây ngạc nhiên, gợi ý rằng đi xe buýt là một loại hàng hóa “thấp cấp”. Khi thu nhập tăng lên, người ta có khuynh hướng sử dụng xe hơi để di chuyển, và vì vậy lượng đi xe buýt sẽ giảm xuống. Nếu thu nhập đầu người tăng lên khoảng 100 đô la, thì về trung bình, đi xe buýt được kỳ vọng giảm khoảng 20,13 ngàn người mỗi giờ. Hệ số của POP và DENSITY dương hay có thể nói khi dân số hay mật độ dân số tăng lên, thì có nhiều người di chuyển bằng xe buýt hơn.

Mặc dù vậy, để thực sự có những kết luận thật thỏa đáng, chúng ta còn phải thực hiện những kiểm định và phân tích sâu hơn nữa.

CHƯƠNG IV

ĐA CỘNG TUYẾN

Các biến giải thích được xác định trong một mô hình kinh tế lượng thường xuất phát từ lý thuyết hoặc những hiểu biết của chúng ta cũng như từ kinh nghiệm quá khứ. Dữ liệu về các biến này đặc biệt xuất phát từ những thực nghiệm không kiểm soát và thường tương quan với nhau. Ví dụ, dân số và tổng sản phẩm quốc nội là hai chuỗi dữ liệu tương quan chặt chẽ. Trong chương trước, chúng ta phát biểu là hệ số hồi qui đối với một biến cụ thể là số đo tác động riêng phần của biến này, nghĩa là tác động của nó khi tất cả các biến khác trong mô hình được giữ ở những mức cố định và chỉ có giá trị của biến này thay đổi. Tuy nhiên, khi hai biến giải thích cùng tương quan chặt; chúng ta không thể chỉ đơn giản giữ một biến không đổi và thay đổi biến còn lại vì khi biến sau thay đổi thì biến đầu thay đổi. Cũng vậy, thay đổi mô hình bằng cách loại bỏ hoặc thêm vào một biến có thể làm thay đổi kết quả một cách nghiêm trọng, khiến cho việc diễn dịch các ước lượng sẽ khó khăn hơn. Đây chính là vấn đề **đa cộng tuyến**, vấn đề xuất hiện khi các biến giải thích có các quan hệ gần như tuyến tính.

4.1. Bản chất của đa cộng tuyến

Trường hợp lý tưởng là các biến không có tương quan với nhau; mỗi một biến X_j chứa một thông tin riêng về Y , thông tin không chứa trong bất kỳ biến X_j khác. Trong thực tế, khi điều này xảy ra ta không gặp hiện tượng đa cộng tuyến. Ở trường hợp ngược lại, ta gặp hiện tượng đa cộng tuyến.

Giả sử ta phải ước lượng hàm hồi qui Y gồm k biến giải thích $X_1, X_2, \dots, X_k$:

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

Đa cộng tuyến xảy ra khi một biến giải thích được biểu diễn dưới dạng tổ hợp tuyến tính của các biến giải thích còn lại đối với mọi điểm của tập số liệu. Hay có thể nói, nếu tồn tại các λ_i không đồng nhất bằng 0 làm cho:

$$\lambda_2 x_{2i} + \lambda_3 x_{3i} + \dots + \lambda_k x_{ki} + v_i = 0; \text{ Trong đó } v_i \text{ là nhiễu; } E(v_i) = 0; \text{Var}(v_i) = \sigma_{v_i}^2 \geq 0$$

Trường hợp này chúng ta có thể nói là có đa cộng tuyến

Nói chung hồi qui đa biến là có đa cộng tuyến, vấn đề là ở mức nào. Trường hợp $\text{Var}(v_i) = 0, \Rightarrow v_i = 0$ do $E(v_i) = 0$, khi đó ta có $\lambda_2 x_{2i} + \lambda_3 x_{3i} + \dots + \lambda_k x_{ki} = 0$, trường hợp này được gọi là đa cộng tuyến hoàn hảo. Nhưng thực tế $\text{Var}(v_i) = 0$ rất khó xảy ra, chỉ có khi số liệu quá ít hoặc đưa vào x_i sai. Khi $\text{Var}(v_i) > 0$, ta có đa cộng tuyến không hoàn hảo, $\text{Var}(v_i)$ lớn thì đa cộng tuyến thấp.

Ví dụ: Giả sử chúng ta ước lượng hàm tiêu dùng. Y = tiêu dùng, X_2 = thu nhập và X_3 = của cải.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u; \quad X_3 = 5X_2$$

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 5X_2 + u$$

$$Y = \beta_1 + (\beta_2 + 5\beta_3)X_2 + u$$

Chúng ta có thể ước lượng $(\beta_2 + 5\beta_3)$ nhưng không ước lượng riêng từng hệ số hồi qui. Hay có thể nói không thể có nghiệm duy nhất cho từng hệ số hồi qui (xem lại cách tính các hệ số hồi qui). Như vậy các hệ số hồi qui sẽ không xác định được.

4.2. Nguồn gốc của đa cộng tuyến

➤ Do phương pháp thu thập dữ liệu:

- Các giá trị của các biến độc lập phụ thuộc lẫn nhau trong mẫu, nhưng không phụ thuộc lẫn nhau trong tổng thể.
- Ví dụ: người có thu nhập cao hơn khuynh hướng sẽ có nhiều của cải hơn. Điều này có thể đúng với mẫu mà không đúng với tổng thể. Trong tổng thể sẽ có các quan sát về các cá nhân có thu nhập cao nhưng không có nhiều của cải và ngược lại.

➤ Các biến độc lập vĩ mô được quan sát theo dữ liệu chuỗi thời gian

Ví dụ: Nhập khẩu quốc gia phụ thuộc vào GDP và CPI (các chỉ số này được thu thập từ dữ liệu chuỗi thời gian).

4.3. Ước lượng khi có đa cộng tuyến hoàn hảo

Khi có đa cộng tuyến hoàn hảo thì các hệ số hồi qui là không xác định còn các sai số tiêu chuẩn là vô hạn. Chúng ta dễ dàng thấy được từ ví dụ trong phần 4.1. Ta cũng có thể sử dụng công thức tính $\hat{\beta}_2$ & $\hat{\beta}_3$ như đã trình bày ở phần hồi qui đa biến, ta cũng thấy rằng $\hat{\beta}_2$ & $\hat{\beta}_3$ là không xác định do tử số và mẫu số = 0

4.4. Ước lượng khi có đa cộng tuyến không hoàn hảo

Trường hợp chúng ta giả thiết giữa x_2 và x_3 có đa cộng tuyến không hoàn hảo theo nghĩa:

$$x_{2i} = \alpha x_{3i} + v_i$$

Trong đó $\alpha \neq 0$, v_i là nhiễu ngẫu nhiên

Trong trường hợp này, theo phương pháp bình phương bé nhất ta có thể thu được $\hat{\beta}_2$ & $\hat{\beta}_3$ nhưng việc giải thích nó sẽ gặp rất nhiều khó khăn.

4.5. Hậu quả của đa cộng tuyến

Khi có hiện tượng đa cộng tuyến trong mô hình, chúng ta có thể sẽ gặp phải những hậu quả sau:

- Sai số chuẩn của các hệ số hồi qui lớn, làm cho khoảng tin cậy lớn và thống kê t ít ý nghĩa và khi đó các ước lượng không thật chính xác. Chúng ta dễ đi đến không có cơ sở bác bỏ giả thiết “không” nhưng trong thực tế không đúng như vậy. Hay có thể nói nó sẽ làm cho chúng ta không xác định đúng tác động của mỗi biến độc lập đến biến phụ thuộc, thậm chí các hệ số hồi qui bị ước lượng sai dấu.
- Các mẫu ngẫu nhiên khác nhau ít nhưng có thể cho kết quả ước lượng khác nhau nhiều, do $Var(\hat{\beta}_j)$ quá lớn.

4.6. Phát hiện đa cộng tuyến

Tính đa cộng tuyến thường xuất hiện dưới một số đặc điểm như sau:

Giá trị R^2 cao với các giá trị của trị thống kê t thấp: Mọi hệ số hồi qui đều không có ý nghĩa (nghĩa là có giá trị t thấp) nhưng trị thống kê F của kiểm định lại rất có ý nghĩa.

Những giá trị cao cho các hệ số tương quan: Các tương quan từng mỗi cặp giữa các biến giải thích có thể cao, Xin lưu ý rằng một hệ số tương quan cao giữa biến phụ thuộc và một biến độc lập không phải là một dấu hiệu của tính đa cộng tuyến. Thực ra một tương quan như vậy rất được mong muốn.

Các hệ số hồi qui nhạy với các đặc trưng: Mặc dù một sự tương quan cao giữa các cặp biến độc lập là một điều kiện đủ cho tính đa cộng tuyến, điều kiện đảo lại không cần thiết phải đúng. Nói cách khác, tính đa cộng tuyến có thể hiện diện mặc dù sự tương quan giữa hai biến giải thích thể hiện không cao. Điều này là do ba hay nhiều hơn các biến có thể gần tuyến tính. Trong trường hợp như vậy, bằng chứng thật sự của tính đa cộng tuyến là sự quan sát cho thấy rằng các hệ số hồi qui bị thay đổi đáng kể khi các biến được thêm vào hoặc bỏ ra.

Dưới đây là một số cách thông dụng để phát hiện đa cộng tuyến:

4.6.1. Xét hệ số tương quan và tương quan riêng:

Xây dựng ma trận hệ số tương quan cặp và quan sát để nhận diện độ mạnh của các tương quan từng cặp biến số độc lập

$$\text{Giả sử } Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

Nếu như tương quan $R(X_2, X_3)$; $R(X_2, X_4)$; $R(X_3, X_4)$ đủ lớn thì có đa cộng tuyến.

Hoặc nếu như tương quan $R(Y, X_2)$ lớn mà sau khi kiểm định tác động của X_2 đến y không bác bỏ được giả thiết H_0 thì có thể nói là có đa cộng tuyến.

4.6.2. Phát hiện qua hồi qui phụ:

Hồi qui một biến giải thích nào đó mà phụ thuộc tuyến tính vào một biến giải thích khác (có hệ số chặn) gọi là hồi qui phụ.

Nếu hồi qui $X_j = \alpha_1 + \sum \alpha_i X_i$ ($i \neq j$)

kiểm định F: $F = [R^2/(k-1)] / [(1-R^2)/(n-k)]$.

k số biến độc lập trong hồi qui phụ

Nếu $F > F^*$ thì chúng ta có thể kết luận rằng R^2 khác không có ý nghĩa thống kê và điều này có nghĩa là có đa cộng tuyến.

4.6.3. Thừa số tăng phương sai (Variance inflation factor-VIF)

Với mỗi X_j , làm hồi qui phụ: $X_j = \alpha_1 + \sum \alpha_i X_i$ ($i \neq j$)

Tính: $VIF = 1/(1-R_j^2)$

Ta thấy rằng khi R_j^2 tăng làm VIF tăng và làm tăng mức độ đa cộng tuyến

Người ta quan niệm rằng khi Khi $VIF \geq 10$ thì có đa cộng tuyến.

4.6.4. Độ đo Theil:

Ý tưởng của phương pháp này là khi không có đa cộng tuyến thì đóng góp của các cá thể là đóng góp chung còn khi có đa cộng tuyến thì đóng góp của cá thể nhỏ hơn nhiều đóng góp chung. Thực hiện như sau:

Ước lượng k-1 hồi qui, bỏ X_j ra khỏi mô hình.

Khi đó hàm hồi qui là:

$$Y_i = \beta_1 + \sum \beta_i X_i + \varepsilon_i \quad (i \neq j);$$

Sau khi hồi qui người ta tìm được $R^2_{(-j)}$, đó là độ phù hợp của hàm hồi qui (không có X_j), sau đó tính $R^2 - R^2_{(-j)}$ là phần đóng góp của X_j cho Y

Tính $m = R^2 - \sum (R^2 - R^2_{(-j)})$

Người ta quan niệm $m/R^2 > 0.5$ thì có đa cộng tuyến.

4.7. Các giải pháp khắc phục

4.7.1. Bỏ qua đa cộng tuyến:

Nếu chúng ta ít quan tâm đến việc đánh giá tác động của các biến độc lập đến biến phụ thuộc mà chú trọng chủ yếu vào việc dự báo, thì tính đa cộng tuyến có thể không phải là một vấn đề nghiêm trọng. Khi đó ta có thể bỏ qua nó mà không phải chịu một hậu quả xấu nào đáng kể. Hay có thể nói nếu chỉ phục vụ dự báo thì có thể bỏ qua đa cộng tuyến.

4.7.2. Loại bỏ biến:

Tính đa cộng tuyến là do những mối quan hệ chặt chẽ giữa các biến giải thích, cách chắc chắn nhất để loại bỏ hoặc giảm bớt các tác động của tính đa cộng tuyến là bỏ một hoặc nhiều biến ra khỏi mô hình. Việc loại bỏ các biến có trị thống kê t thấp nói chung sẽ làm cải thiện mức ý nghĩa của các biến còn lại. Điều quan trọng xảy ra trong tình huống này là các biến còn lại có khả năng nắm giữ những tác động của các biến bị loại bỏ có liên quan chặt chẽ với chúng.

4.7.3. Tái thiết lập mô hình:

Trong nhiều tình huống, tái xác định mô hình có thể làm giảm tính đa cộng tuyến.

4.7.4. Bổ sung dữ liệu hoặc tìm dữ liệu mới:

Vì đa cộng tuyến đặc trưng cho mẫu, nên có thể lấy mẫu khác liên quan đến cùng các biến trong mẫu ban đầu mà đa cộng tuyến có thể không còn nghiêm trọng nữa. Do đó việc bổ sung dữ liệu hoặc tìm dữ liệu mới đôi khi được đề xuất dựa trên cơ sở cho rằng việc gia tăng kích thước mẫu sẽ làm cải thiện độ chính xác của một ước lượng và do đó giảm thiểu được những yếu tố phản tác dụng của tính đa cộng tuyến.

4.7.5. Sử dụng thông tin tiên nghiệm:

Ví dụ chúng ta có thể biết tác động biên của của cải lên tiêu dùng chỉ bằng 1/10 so với tác động biên của thu nhập lên tiêu dùng.

$$\beta_3 = 0.10 \beta_2$$

Chạy mô hình với điều kiện tiên nghiệm.

$$Y = \beta_1 + \beta_2 X_2 + 0.10 \beta_2 X_3 + u$$

$$Y = \beta_1 + \beta_2 X \text{ trong đó } X = X_2 + 0.1 X_3$$

Khi ước lượng được β_2 thì suy ra β_3 từ mối quan hệ tiên nghiệm trên.

BÀI TẬP THỰC HÀNH

Sau khi ước lượng hàm tiêu dùng, với X_2 là thu nhập và X_3 là của cải. Chúng ta thu được kết quả như sau:

$$\hat{Y} = 1.60 - 0.10 X_2 + 0.18 X_3$$

$$p\text{-value} \quad 0.25 \quad 0.86 \quad 0.24$$

$$R^2 = 0.99$$

$$P_F = 0.03$$

Từ kết quả, cho thấy rằng:

- R^2 rất cao giải thích 99% biến đổi của hàm tiêu dùng
- Không có biến độc lập nào có ý nghĩa (thống kê t quá thấp).

- Có một biến sai dấu.
- Giá trị thống kê F rất cao dẫn đến bác bỏ giả thuyết “không” và cho rằng mô hình ước lượng có ý nghĩa.

Chúng ta có thể thấy được rằng, điều này do nguyên nhân biến thu nhập và của cải tương quan rất mạnh với nhau do đó không thể nào ước lượng được tác động biên chính xác cho thu nhập hoặc của cải lên tiêu dùng.

Xét tương quan giữa X_2 và X_3 ta có:

$R(X_2, X_3) = 0,99$, đây thực sự là một tương quan mạnh, hay có thể nói chúng ta hầu như chúng ta có đa cộng tuyến hoàn hảo giữa X_2 và X_3

Hồi qui tiêu dùng theo thu nhập:

$$\hat{Y} = 1.57 + 0.82X_2$$

$$p - value \quad 0.03 \quad 0.001$$

$$R^2 = 0.99$$

$$P_F = 0.001$$

Ta thấy biến thu nhập trở nên có ý nghĩa thống kê, nhưng trước lúc đó trong mô hình đầu thì không có ý nghĩa.

Hay tương tự hồi qui thu nhập Y theo của cải:

$$\hat{Y} = 1.57 + 0.34X_2$$

$$p - value \quad 0.03 \quad 0.012$$

$$R^2 = 0.99$$

$$P_F = 0.012$$

Biến của cải trở nên có ý nghĩa thống kê, nhưng trước lúc đó trong mô hình đầu thì không có ý nghĩa.

Qua kết quả đó, chúng ta có thể kết luận rằng có hiện tượng đa cộng tuyến trong mô hình, hay có quan hệ tuyến tính chặt giữa hai biến thu nhập và của cải. Khi đó chúng ta sẽ phải sử dụng biện pháp nào đó để khắc phục tình trạng trên.

CHƯƠNG V

HỒI QUI VỚI BIẾN GIẢ

5.1. Bản chất của biến giả

Tất cả các biến chúng ta đã giới thiệu ở những chương trước đều là biến định lượng; nghĩa là các biến này có các đặc tính có thể đo lường bằng số. Tuy nhiên, hành vi của các biến kinh tế cũng có thể phụ thuộc vào các nhân tố **định tính** như giới tính, trình độ học vấn, mùa, v.v... ví dụ chúng ta xem xét mô hình hồi qui tuyến tính đơn sau:

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

Gọi Y là mức tiêu thụ năng lượng trong một ngày và X là nhiệt độ trung bình. Khi nhiệt độ tăng trong mùa hè, chúng ta sẽ kỳ vọng mức tiêu thụ năng lượng sẽ tăng. Vì vậy, hệ số độ dốc β có khả năng là số dương. Tuy nhiên, trong mùa đông, khi nhiệt độ tăng ví dụ từ 20 đến 40 độ, năng lượng được dùng để sưởi ấm sẽ ít hơn, và mức tiêu thụ sẽ có vẻ giảm khi nhiệt độ tăng. Điều này cho thấy β có thể âm trong mùa đông. Vì vậy, bản chất của quan hệ giữa mức tiêu thụ năng lượng và nhiệt độ có thể được kỳ vọng là phụ thuộc vào biến định tính “mùa”.

Vậy chúng ta phải làm gì để có thể mô tả được tác động của những biến định tính, hay làm thế nào để có thể đưa các biến định tính vào mô hình. Công cụ xử lý đó chính là biến giả (dummy). Chúng ta sẽ giải thích trong nhiều trường hợp khác nhau từ đơn giản đến phức tạp. Biến giả thể hiện các biến định tính.

- Mô hình chỉ có biến giải thích là biến giả
- Mô hình có biến giải thích định lượng và biến giả

Biến giả thường được ký hiệu là D, và chỉ nhận 1 trong 2 giá trị là 0 và 1.

Ví dụ:

Chúng ta bắt đầu với việc xem xét trường hợp đơn giản nhất trong đó một biến định tính chỉ có hai lựa chọn. Ví dụ, giữa hai ngôi nhà có cùng các đặc trưng, một có thể có hồ bơi trong khi ngôi nhà còn lại không có. Tương tự, giữa hai nhân viên của một công ty có cùng tuổi, học vấn, kinh nghiệm v.v..., một người là nam và người kia là nữ. Vậy làm thế nào để đo lường tác động của giới tính đến lương và tác động của sự hiện diện của hồ bơi đến giá nhà. Cụ thể chúng ta sẽ xem xét ví dụ về lương và đặt Y_i là tiền lương hàng tháng của nhân viên thứ i trong một công ty. Để đơn giản, ở đây chúng ta bỏ qua các biến khác có ảnh hưởng đến lương và chỉ tập trung vào giới tính. Vì biến giới tính không phải là một biến định lượng nên chúng ta định nghĩa một **biến giả** (gọi là D), biến giả này là chỉ nhận giá trị 1 trong 2 giá trị, bằng 1 đối với nhân viên nam và bằng 0 đối với nhân viên nữ. Chúng ta sẽ thiết lập và ước lượng một mô hình sử dụng biến giả như một biến giải thích. Dạng đơn giản nhất của mô hình như sau:

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

Chúng ta giả sử là số hạng sai số ngẫu nhiên thỏa mãn tất cả các giả thiết của phương pháp bình phương nhỏ nhất. Chúng ta có thể lấy kỳ vọng có điều kiện của Y với D cho trước và được các phương trình sau:

Nam: $E(Y|D = 1) = \beta_1 + \beta_2$

Nữ: $E(Y|D = 0) = \beta_1$

Trong đó:

- Hệ số chặn β_1 của hồi qui tuyến tính là tiền lương trung bình của nhân viên nữ, trong khi độ dốc β_2 của đường hồi qui đo sự khác nhau về tiền lương giữa nhân viên nam và nữ.
- Kiểm định giả thiết $H_0: \beta_2 = 0$ cung cấp kiểm định về giả thiết là không có sự khác nhau tiền lương giữa nhân viên nam và nhân viên nữ.
- Thủ tục biến giả có thể dễ dàng mở rộng cho trường hợp có nhiều hơn 2 lựa chọn (phạm trù)..

5.2. Hồi qui với biến giả

Phần này ta xét mô hình hồi qui 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 được làm tương tự.

5.2.1. Trường hợp khi biến chất chỉ có 2 lựa chọn (2 phạm trù):

Giả sử rằng chúng ta phải ước lượng mối quan hệ giữa tiền lương của một người lao động và số năm công tác của họ

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

Trong đó: y_i = tiền lương hàng năm của một người lao động i .

x_i = số năm công tác của một người lao động i .

Ở đây chúng ta hoàn toàn chưa nói là người lao động là nam hay nữ.

Bây giờ giả sử rằng chúng ta muốn tìm hiểu xem liệu có phải người lao động nữ bị phân biệt đối xử về việc trả lương của họ không? Ta bắt đầu bằng việc ước lượng chung cho cả người lao động nam và nữ bằng cách xác định một biến mới ghi nhận sự có mặt hay vắng mặt của thuộc tính "nữ"

$D_i = 0$ nếu quan sát i thuộc về một người lao động nữ

$D_i = 1$ nếu quan sát i thuộc về một người lao động nam

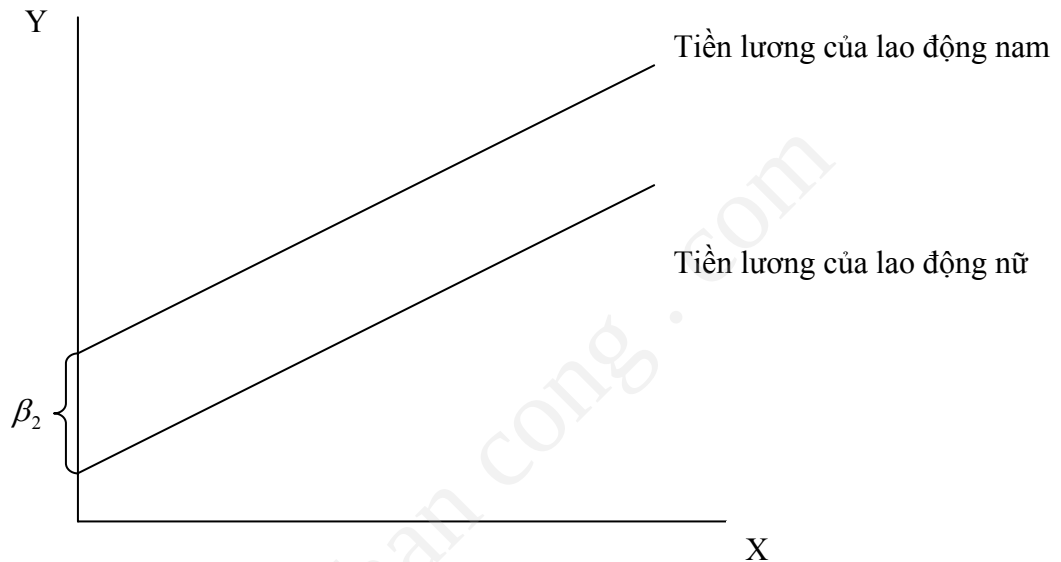
Khi đó mô hình hồi qui có dạng như sau:

$$y_i = \beta_1 + \beta_2 D_i + \beta_3 x_i + u_i$$

Giá trị kỳ vọng có điều kiện như sau:

$$E(y|D_i=0) = \beta_1 + \beta_3 x_i \quad \text{nữ}$$

$$E(y|D_i=1) = (\beta_1 + \beta_2) + \beta_3 x_i \quad \text{nam}$$



Hệ số β_2 là *chênh lệch tung độ gốc*. Dễ dàng kiểm định xem liệu chênh lệch này có ý nghĩa thống kê hay không: đơn giản là tính giá trị thống kê t cho β_2 rồi so với giá trị tới hạn tra bảng t^* như chúng ta đã làm khi kiểm định ý nghĩa thống kê của các hệ số hồi qui.

5.2.2. Trường hợp với một biến lượng và hai biến chất:

Giả sử rằng chúng ta phải ước lượng mối quan hệ giữa tiền lương của các đối tượng và số năm công tác của họ, bây giờ chúng ta giả thiết thêm rằng, ngoài giới tính thì vùng mà lao động làm việc cũng ảnh hưởng tới thu nhập. Như vậy ta sẽ có hai biến chất:

- Vùng có 3 phạm trù
- Giới tính có 2 phạm trù

Mô hình sẽ có dạng:

$$y_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 D_{3i} + \beta_5 x_i + u_i$$

Trong đó: y_i = thu nhập hàng năm của người lao động i.

x_i = số năm công tác của người lao động i.

$D_1 = 1$ nếu lao động của các doanh nghiệp miền Bắc

$D_1 = 0$ nếu lao động không thuộc các doanh nghiệp miền Bắc

$D_2 = 1$ nếu lao động của các doanh nghiệp miền Nam

$D_2 = 0$ nếu lao động không thuộc các doanh nghiệp miền Nam

$D_3 = 1$ nếu lao động là nam

$D_3 = 0$ nếu lao động là nữ

Phạm trù cơ sở bây giờ là người lao động thuộc các doanh nghiệp miền Trung. Giả sử $E(u_i) = 0$, khi đó thu nhập trung bình của một lao động nữ ở 1 doanh nghiệp miền Trung là:

$$E(y|D_1 = 0, D_2 = 0, D_3 = 0, x_i) = \beta_1 + \beta_5 x_i$$

5.3. Mô tả tác động của các biến chất (biến định tính):

5.3.1. Tác động chỉ làm thay đổi tung độ gốc (hệ số chặn):

Trường hợp chúng ta chỉ xem xét tác động của biến định tính trong việc thay đổi tung độ gốc, ta sẽ thiết lập dạng hàm tổng quát như sau:

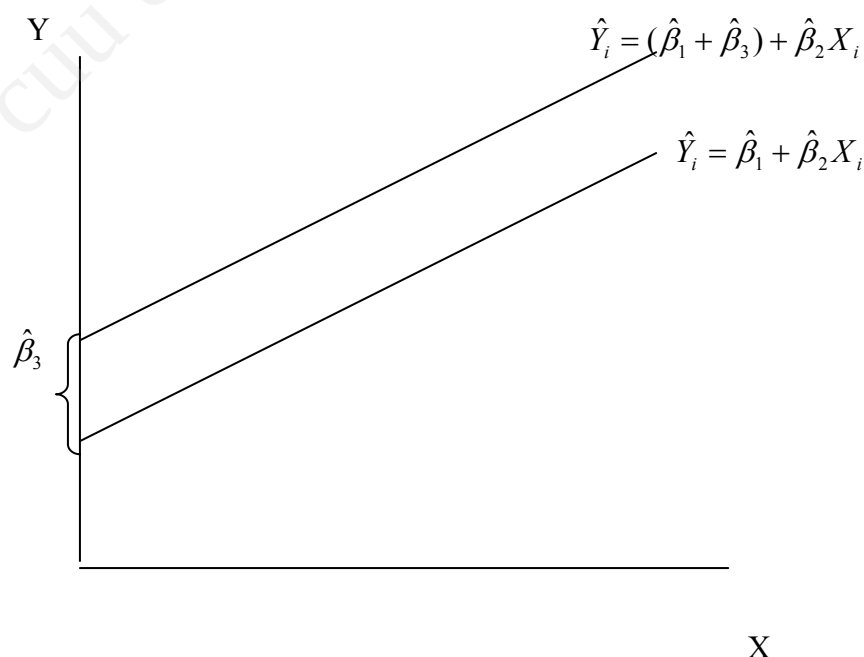
$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i$$

Lấy kỳ vọng tương ứng với $D = 0$ và $D = 1$ sẽ cho ta thấy sự thay đổi về tung độ gốc do tác động của biến định tính.

$$E(Y|D = 0) = \beta_1 + \beta_2 X_i$$

$$E(Y|D = 1) = (\beta_1 + \beta_3) + \beta_2 X_i$$

Hình 4.1: Ví dụ về việc dịch chuyển tung độ gốc



5.3.2. Tác động chỉ làm thay đổi số hạng độ dốc (hệ số góc):

Khi đó dạng hàm sẽ được thiết lập có dạng như sau:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i X_i + u_i$$

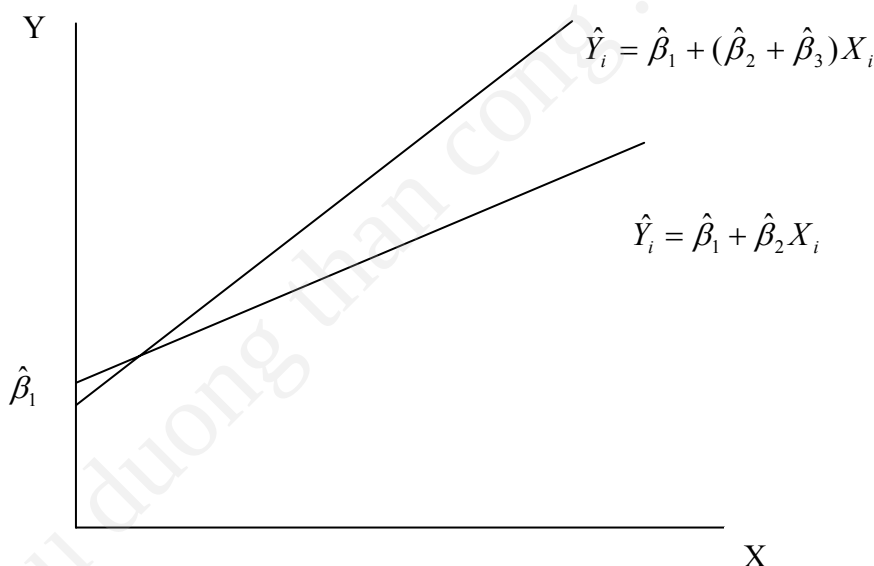
Có nghĩa là để ước lượng mô hình này, chúng ta nhân biến giả với X và tạo một biến mới, DX . Rồi chúng ta hồi qui Y theo một số hạng không đổi, X , và DX .

Lấy kỳ vọng tương ứng với $D = 0$ và $D = 1$ sẽ cho ta thấy sự thay đổi về số hạng độ dốc do tác động của biến định tính.

$$E(Y|D = 0) = \beta_1 + \beta_2 X_i$$

$$E(Y|D = 1) = \beta_1 + (\beta_2 + \beta_3) X_i$$

Hình 4.2: Ví dụ về việc dịch chuyển số hạng độ dốc



5.3.3. Tác động làm dịch chuyển cả tung độ gốc và số hạng độ dốc:

Để mô tả tác động của biến định tính đối với việc làm dịch chuyển cả tung độ gốc và số hạng độ dốc (hay có thể gọi là trường hợp tổng quát), chúng ta thiết lập mô hình như sau:

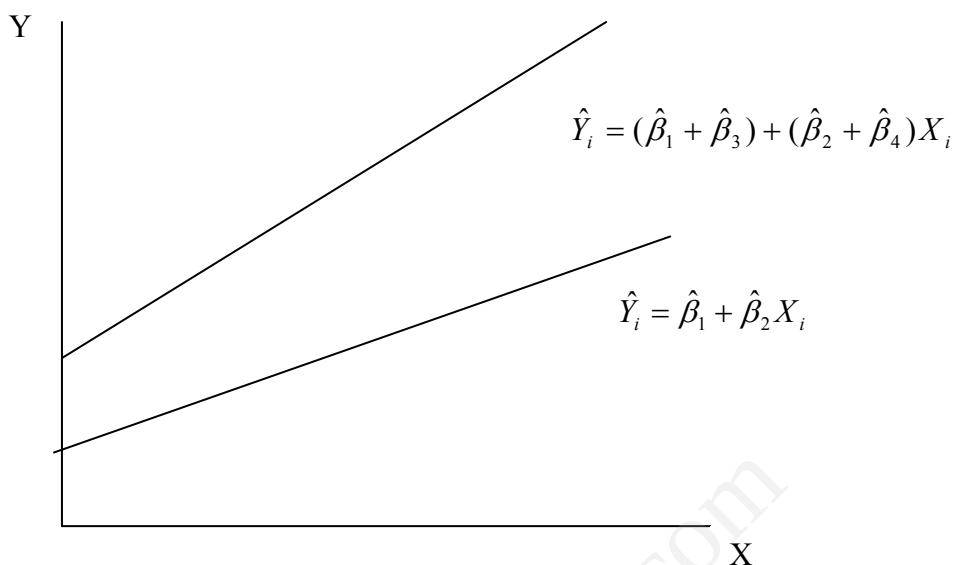
$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta_4 D_i X_i + u_i$$

Lấy kỳ vọng tương ứng với $D = 0$ và $D = 1$ sẽ cho ta thấy sự thay đổi về tung độ gốc và số hạng độ dốc do tác động của biến định tính.

$$E(Y|D = 0) = \beta_1 + \beta_2 X_i$$

$$E(Y|D = 1) = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) X_i$$

Hình 4.3: Ví dụ về việc dịch chuyển cả tung độ gốc và số hạng độ dốc



5.4. Ước Lượng Những Tác động Mùa

Một ví dụ về việc sử dụng biến giả trong ước lượng tác động mùa của các biến độc lập. Xem xét quan hệ $E = \beta_1 + \beta_2 T + u$, đã được giới thiệu trước đây, giữa việc tiêu thụ điện năng và nhiệt độ. Trong mùa hè, khi nhiệt độ tăng, nhu cầu máy lạnh sẽ đẩy việc tiêu thụ điện năng lên. Do vậy chúng ta kỳ vọng β có dấu dương, E và T có tương quan dương. Tuy nhiên, vào mùa đông, khi nhiệt độ tăng (từ 20 độ lên 40 độ), nhu cầu cho việc sưởi ấm nhà trở nên thấp hơn và từ đó chúng ta mong đợi β có dấu âm về mùa đông, hay E và T có tương quan âm. Bằng cách nào chúng ta có thể ghi nhận được tác động lên E của biến định tính “mùa” có bốn loại: xuân, hạ, thu, đông? Việc này thực hiện được bằng cách xác định ba biến giả; được gọi là: **biến giả theo mùa**. Mùa xuân được sử dụng làm mùa điều khiển:

$D_2 = 1$ nếu là mùa hè

$D_2 = 0$ nếu là mùa khác

$D_3 = 1$ nếu là mùa thu

$D_3 = 0$ nếu là mùa khác

$D_4 = 1$ nếu là mùa đông

$D_4 = 0$ nếu là mùa khác

Vậy ta có thể biểu diễn dạng hàm tổng quát khi đưa vào mô hình biến định tính mùa như sau:

$$E = \alpha_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_4 D_4 + \beta_1 T + \beta_2 D_2 T + \beta_3 D_3 T + \beta_4 D_4 T + u$$

Những mô hình ước lượng cho từng mùa khi đó được biểu diễn như sau:

Mùa xuân: $\hat{E} = \hat{\alpha}_1 + \hat{\beta}_1 T$

$$\text{Mùa hè:} \quad \hat{E} = (\hat{\alpha}_1 + \hat{\alpha}_2) + (\hat{\beta}_1 + \hat{\beta}_2)T$$

$$\text{Mùa thu:} \quad \hat{E} = (\hat{\alpha}_1 + \hat{\alpha}_3) + (\hat{\beta}_1 + \hat{\beta}_3)T$$

$$\text{Mùa đông:} \quad \hat{E} = (\hat{\alpha}_1 + \hat{\alpha}_4) + (\hat{\beta}_1 + \hat{\beta}_4)T$$

α_2 là độ lệch của hệ số tung độ gốc mùa hè so với hệ số tung độ gốc của mùa xuân, và β_2 là độ lệch của hệ số độ dốc mùa hè so với hệ số độ dốc của mùa xuân. Có thể thực hiện nhiều kiểm định đối với những mô hình này. Ví dụ, giả thuyết hợp lý là không có sự khác biệt trong quan hệ giữa mùa thu và mùa xuân.

So sánh các phương trình của mùa thu và mùa xuân, giả thuyết hàm ý rằng $\alpha_3 = \beta_3 = 0$. Điều này được kiểm định bằng kiểm định Wald.

5.5. Sử dụng biến giả để Kiểm định sự thay đổi về cấu trúc

Mối quan hệ giữa các biến phụ thuộc và độc lập có thể có một **sự thay đổi về cấu trúc** (còn được gọi là **sự bất ổn định về cấu trúc** hay **những gián đoạn về cấu trúc**); có nghĩa là, mối quan hệ có thể thay đổi từ thời đoạn này sang thời đoạn khác.

Kiểm định thống kê đối với thay đổi về cấu trúc được gọi là **Kiểm định Chow** (sau khi Gregory Chow [1960] lần đầu tiên công bố kỹ thuật này). Phần này trình bày hai phương pháp kiểm định đối với thay đổi về cấu trúc. Phương pháp thứ nhất bao gồm việc chia mẫu thành hai hay nhiều nhóm, ước lượng mô hình một cách riêng biệt đối với từng thời đoạn và với cả mẫu chung lại, và sau đó xây dựng một trị thống kê F sử dụng để tiến hành kiểm định. Ở phương pháp thứ hai, chúng ta sử dụng các biến giả.

5.5.1. Kiểm định dựa trên việc phân cắt mẫu (Kiểm định Chow)

Giả sử chúng ta muốn kiểm định xem có một sự thay đổi về cấu trúc hay không

vào thời điểm t . Thủ tục sẽ là phải chia mẫu gồm n quan sát thành hai nhóm, nhóm 1 gồm n_1 quan sát đầu tiên và nhóm 2 gồm những quan sát còn lại $n_2 = n - n_1$. Ước lượng mô hình một cách riêng biệt (với k hệ số hồi qui) đối với từng nhóm một và tính toán tổng bình phương các phần dư RSS_1 và RSS_2 . Do đó, tổng các bình phương không giới hạn được tính bằng $RSS_U = RSS_1 + RSS_2$. Khi lấy số này chia cho σ^2 , kết quả sẽ có phân phối chi-square với bậc tự do d.f. là $n_1 - k + n_2 - k = n - 2k$. Ước lượng mô hình lần nữa nhưng với chung cả mẫu, và thu được giá trị RSS_R . Trị thống kê kiểm định sẽ là:

$$F_c = \frac{(RSS_R - RSS_U) / k}{RSS_U / (n_1 + n_2 - 2k)} = \frac{(RSS_R - RSS_1 - RSS_2) / k}{(RSS_1 + RSS_2) / (n_1 + n_2 - 2k)} \sim F(k, n_1 + n_2 - 2k)$$

Thủ tục kiểm định là để bác bỏ *giả thuyết không* rằng không có thay đổi về cấu

trúc nào nếu F_c vượt quá giá trị $F^*(k, n-2k)$, điểm nằm trên phân phối F với bậc tự do d.f. là k và $n - 2k$ mà vùng từ đó tính sang bên phải bằng với mức ý nghĩa.

5.5.2. Kiểm định dựa trên việc sử dụng biến giả

Kiểm định cũng có thể được tiến hành bằng cách sử dụng kỹ thuật dùng biến giả. Các bước tiến hành như sau:

Ước lượng hồi qui cả bộ số liệu với dạng hàm như sau:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta_4 D_i X_i + u_i$$

Trong đó:

$D = 1$ nếu số liệu được lấy từ thời kỳ 1 (bộ số liệu 1, n_1 quan sát)

$D = 0$ nếu số liệu được lấy từ thời kỳ 2 (bộ số liệu 2, n_2 quan sát)

Sau khi ước lượng ta thu được tổng bình phương các phần dư, ký hiệu là $RSSU$.

Ước lượng mô hình lần nữa cũng với cả bộ số liệu không sử dụng biến giả, mô hình là $Y_i = \beta_1 + \beta_2 X_i + u_i$ và thu được giá trị $RSSR$. Trị thống kê kiểm định sẽ là:

$$F_c = \frac{(RSSR - RSSU) / k}{RSSU / (n_1 + n_2 - 2k)} \sim F(k, n_1 + n_2 - 2k); \text{ (ở đây } k = 2 \text{)}$$

Thủ tục kiểm định là để bác bỏ *giả thuyết không* rằng không có thay đổi về cấu

trúc nào cũng là F_c vượt quá giá trị $F^*(k, n-2k)$.

BÀI THỰC HÀNH:

Khi nghiên cứu SAVINGS là lượng tiết kiệm của người dân Mỹ trong một thời đoạn cho trước và biến độc lập thu nhập INCOME. Người ta thấy rằng có hai thời đoạn trong suốt khoảng thời gian 1970-1995, có thể gây ra những thay đổi trong mô hình hành vi tiết kiệm. Bởi vì sự liên tục tăng mạnh về giá dầu lửa trên thế giới hồi năm 1979 cho đến hết năm 1981, từ 13,5 đôla/thùng lên 36đôla/thùng đã gây nên cuộc khủng hoảng năng lượng trầm trọng trên thế giới. Tác động của những biến động mạnh mẽ về giá dầu lửa thế giới cùng sự tăng vọt về lãi suất cho vay thực tế ở Mỹ và trên khắp thế giới đã là một nhân tố quan trọng gây ra cuộc khủng kinh tế thế giới nói chung và cuộc khủng hoảng nợ 1982 nói riêng. Vậy hành vi tiết kiệm có thể có sự khác biệt giữa hai thời kỳ là 1970-1981 và thời kỳ 1982-1995.

YEAR	SAVINGS (Y)	INCOME (X)	DUM (D)
1970	61.00	727.10	0
1971	68.60	790.20	0
1972	63.60	855.30	0

1973	89.60	965.00	0
1974	97.60	1054.20	0
1975	104.40	1159.20	0
1976	96.40	1273.00	0
1977	92.50	1401.40	0
1978	112.60	1580.10	0
1979	130.10	1769.50	0
1980	161.80	1973.30	0
1981	199.10	2200.20	0
1982	205.50	2347.30	1
1983	167.00	2522.40	1
1984	235.70	2810.00	1
1985	206.20	3002.00	1
1986	196.50	3187.60	1
1987	168.40	3363.10	1
1988	189.10	3640.80	1
1989	187.80	3894.50	1
1990	208.70	4166.80	1
1991	246.40	4343.70	1
1992	272.60	4613.70	1
1993	214.40	4790.20	1
1994	189.40	5021.70	1
1995	249.30	5320.80	1

Bây giờ chúng ta sẽ xem xét việc ứng dụng kỹ thuật biến giả trong nghiên cứu có sự thay đổi về cấu trúc của 2 thời kỳ này không.

Đầu tiên ta tiến hành hồi qui có sử dụng biến giả, mô hình hồi qui có dạng như sau: $Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 D_t X_t + u_t$

Dt = 0 ứng với thời kỳ 1: $Y_t = \alpha_1 + \beta_1 X_t + u_t$

Dt = 1 ứng với thời kỳ 2: $Y_t = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2)X_t + u_t$

Kết quả hồi qui thu được như sau:

Dependent Variable: SAVINGS

Method: Least Squares

Sample: 1970 1995

Included observations: 26

SAVINGS=C(1)+C(2)*DUM+C(3)*INCOME+C(4)*DUM*INCOME

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	1.016117	20.16483	0.050391	0.9603
C(2)	152.4786	33.08237	4.609058	0.0001
C(3)	0.080332	0.014497	5.541347	0.0000
C(4)	-0.065469	0.015982	-4.096340	0.0005
R-squared	0.881944	Mean dependent var	162.0885	
Adjusted R-squared	0.865846	S.D. dependent var	63.20446	
S.E. of regression	23.14996	Akaike info criterion	9.262501	
Sum squared resid	11790.25	Schwarz criterion	9.456055	
Log likelihood	-116.4125	Durbin-Watson stat	1.648454	

Tiếp theo ta tiến hành hồi qui toàn bộ số liệu với mô hình hồi qui có dạng:
 $Y_t = \alpha_1 + \beta_1 X_t + u_t$, và thu được kết quả:

Dependent Variable: SAVINGS

Method: Least Squares

Sample: 1970 1995

Included observations: 26

SAVINGS=C(1)+C(2)*INCOME

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	62.42267	12.76075	4.891772	0.0001
C(2)	0.037679	0.004237	8.893776	0.0000
R-squared	0.767215	Mean dependent var	162.0885	
Adjusted R-squared	0.757515	S.D. dependent var	63.20446	
S.E. of regression	31.12361	Akaike info criterion	9.787614	
Sum squared resid	23248.30	Schwarz criterion	9.884391	
Log likelihood	-125.2390	Durbin-Watson stat	0.859717	

Lập thống kê kiểm định, ta có:

$$F_c = \frac{(RSSR - RSSU)/k}{RSSU/(n_1 + n_2 - 2k)} = \frac{(23,248.3 - 11,790.25)/2}{11,790.25/22} = 10.69$$

Trong đó $F_{2,22} = 3.44$, do đó bác bỏ H_0 cho rằng không có sự thay đổi về cấu trúc, hay có thể nói rằng hành vi tiết kiệm của người dân Mỹ ở 2 thời kỳ này là khác nhau với mức ý nghĩa bằng 5%.

CHƯƠNG VI

PHƯƠNG SAI SAI SỐ THAY ĐỔI

6.1. Khái niệm

Trong việc tính toán các giá trị ước lượng bình phương tối thiểu thông thường (OLS), chúng ta đã thiết lập giả thuyết cho rằng các số hạng sai số u_i có phân phối giống nhau với trị trung bình bằng không và phương sai bằng nhau.

$$\text{Var}(u_i | X_i) = E(u_i^2 | X_i) = \sigma^2$$

Giả thuyết phương sai bằng nhau được hiểu là **phương sai của sai số không đổi** (có nghĩa là phân tán như nhau).

Phương sai σ^2 là một đại lượng đo lường mức độ phân tán của các số hạng sai số u_i xung quanh giá trị trung bình bằng 0. Một cách tương đương, đó là một đại lượng đo lường mức độ phân tán của giá trị biến phụ thuộc quan sát được (Y) xung quanh đường hồi qui $\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k$. Phương sai của sai số không đổi có nghĩa là mức độ phân tán như nhau cho tất cả các quan sát.

Tuy nhiên, trong nhiều trường hợp đối với dữ liệu chéo, giả định này bị vi phạm. Khi giả định bị vi phạm, khi đó ta gọi là phương sai sai số thay đổi.

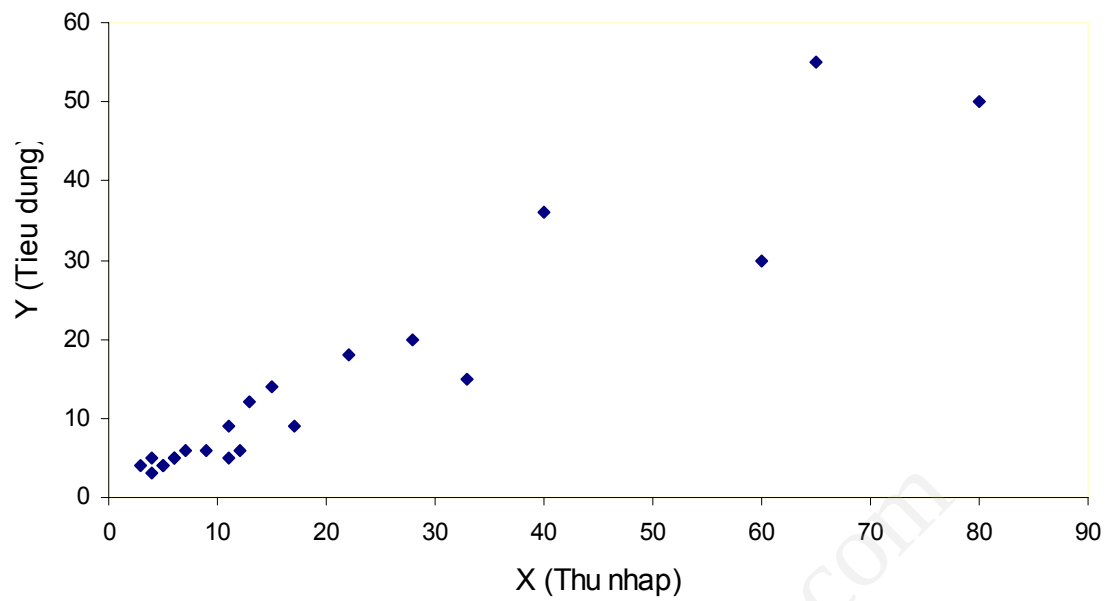
$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Và

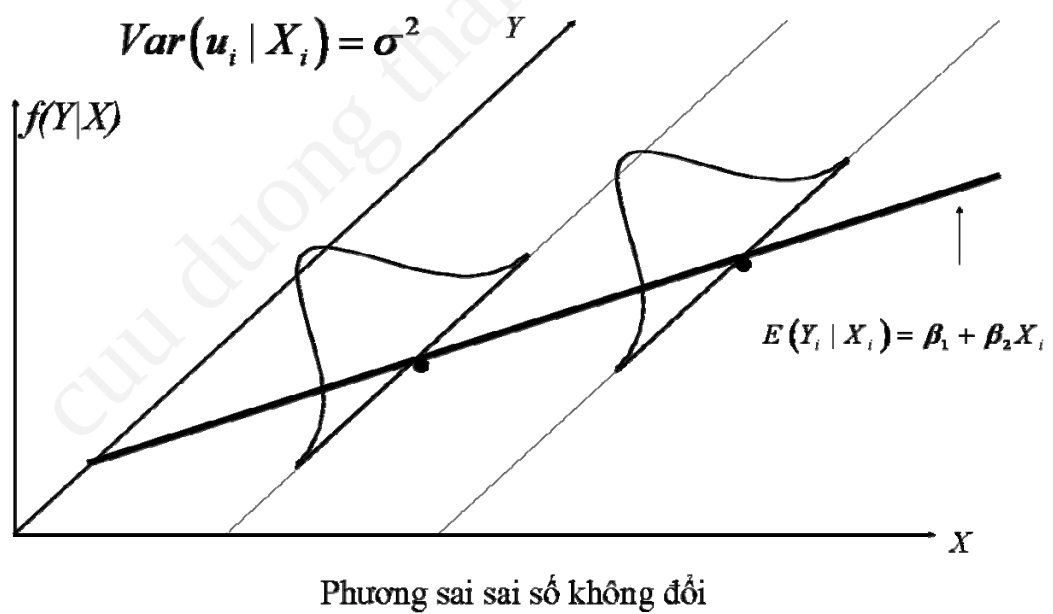
$$\text{Var}(u_i | X_i) = \sigma_i^2 = \sigma^2(X_i)$$

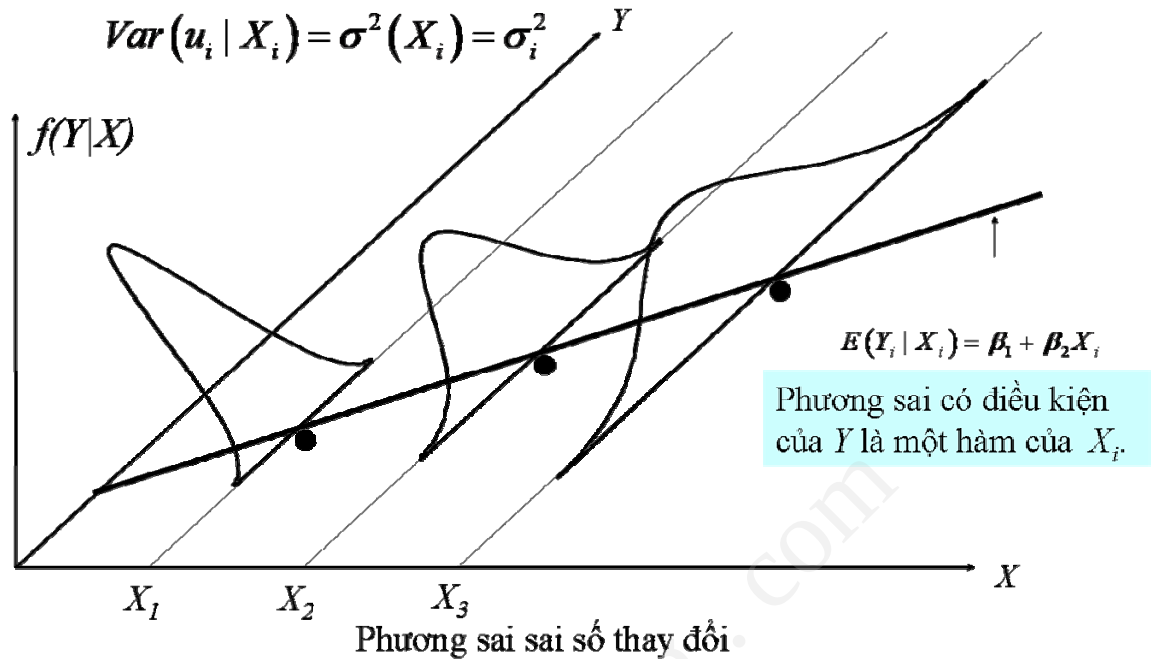
Phương sai sai số thay đổi, có nghĩa là phương sai sai số là khác nhau đối với các quan sát. Trường hợp đặc biệt là nó là một hàm của biến giải thích, X_i .

Ví dụ, giả sử như chúng ta tiến hành điều tra một mẫu ngẫu nhiên các hộ gia đình và thu được thông tin về tổng chi phí tiêu dùng của từng hộ gia đình và thu nhập của họ trong một năm cho trước. Những hộ gia đình với mức thu nhập thấp không có nhiều linh động trong chi tiêu. Phần lớn thu nhập sẽ tập trung vào các nhu cầu cần bản chẳng hạn như thức ăn, chỗ ở, quần áo, và đi lại. Do vậy, chi tiêu giữa những hộ gia đình có thu nhập thấp như thế sẽ không khác nhau nhiều. Mặt khác, những gia đình giàu có có sự linh động rất lớn trong chi tiêu. Một vài gia đình là những người tiêu dùng lớn; những người khác có thể là những người tiết kiệm nhiều và đầu tư nhiều vào bất động sản, thị trường chứng khoán, Điều này hàm ý rằng tiêu dùng thực có thể khác nhiều so với mức thu nhập trung bình. Hay nói cách khác, rất có khả năng những hộ gia đình có thu nhập cao có mức độ phân tán xung quanh giá trị tiêu dùng trung bình lớn hơn những hộ gia đình có thu nhập thấp. Trong trường hợp như thế, biểu đồ phân tán giữa tiêu dùng và thu nhập sẽ chỉ ra những điểm của mẫu gần với đường hồi qui hơn cho những hộ gia đình thu nhập thấp nhưng những điểm phân tán rộng hơn cho những hộ gia đình thu nhập cao. (xem Hình dưới đây). Hiện tượng như vậy được gọi là **phương sai của sai số thay đổi** (có nghĩa là phân tán không như nhau).



Chúng ta có thể biểu diễn phương sai không đổi và phương sai thay đổi như dưới đây:





6.2. Nguyên nhân

- Do bản chất của mối quan hệ, ví dụ khi X lớn hơn X_0 thì quan hệ Y vào X thay đổi, ví dụ:
 - Tiêu dùng và thu nhập
 - Sản lượng sản xuất ở các DN có qui mô khác nhau
- Chọn mô hình sai (sai dạng, thiếu biến), ví dụ: hàm chi phí lại sử dụng dạng tuyến tính.

6.3. Hậu quả

Những chứng minh của tính chất tuyến tính và không chệch không phụ thuộc vào phương sai sai số không đổi. Do vậy, các tính chất tuyến tính và không chệch không bị vi phạm do việc bỏ qua hiện tượng phương sai của sai số thay đổi và sử dụng OLS để ước lượng các tham số của mô hình. Tuy nhiên, trong khi chứng minh định lý Gauss-Markov, chúng ta phải sử dụng giả thuyết cho rằng $Var(u_i) = \sigma^2$ nhằm làm cực tiểu phương sai. Bởi vì giả thuyết đó không còn đúng nữa, nên không thể khẳng định rằng ước lượng OLS hiệu quả hơn. Điều này có nghĩa là ước lượng OLS khi này sẽ không còn hiệu quả. Có thể tìm một ước lượng tuyến tính không chệch khác mà có giá trị phương sai thấp hơn ước lượng OLS.

Dễ dàng thấy được điều này:

Ước lượng vẫn là không chệch. Để đơn giản giả sử chúng ta xét mô hình hồi qui đơn, khi đó ta có:

$$\hat{\beta}_2 = \beta_2 + \frac{\sum (x_i - \bar{x})u_i}{\sum (x_i - \bar{x})^2}$$

$$Var(\hat{\beta}_2) = \frac{\sum (x_i - \bar{x})^2 \sigma_i^2}{(\sum (x_i - \bar{x})^2)^2}$$

$$\& E(\hat{\beta}_2) = \beta_2$$

Bây giờ hãy xét phương sai của hàm ước lượng:

$$Var(\hat{\beta}_2) = \frac{\sum (x_i - \bar{x})^2 \sigma_i^2}{(\sum (x_i - \bar{x})^2)^2}$$

Ta dễ dàng thấy rằng nếu phương sai sai số không đổi thì công thức trên sẽ trở thành:

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Vậy có thể nói rằng khi chúng ta gặp phương sai sai số thay đổi, nếu ta vẫn coi là không đổi và ước lượng bằng OLS, khi đó:

- Chúng ta vẫn nhận được các ước lượng là không chệch.
- Nhưng các ước lượng không còn hiệu quả nữa.
- Chúng ta sẽ không thoả mãn được BLUE (Best *Linear* Unbiased).

6.4. Biện pháp khắc phục

- Nếu chúng ta gặp phương sai sai số thay đổi, ước lượng OLS sẽ không còn thoả mãn BLUE (phương sai nhỏ nhất, ước lượng chính xác nhất).
- Chúng ta cần xem xét phương pháp ước lượng mới. có thể sử dụng phương pháp bình phương nhỏ nhất có trọng số-WLS.
- Khi chúng ta gặp phương sai sai số thay đổi ta có thể sử dụng phương pháp bình phương nhỏ nhất có trọng số-WLS để nhận được các ước lượng BLUE.

Thủ tục WLS

- Xét hàm hồi qui khi phương sai thay đổi:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Và

$$Var(u_i | X_i) = \sigma_i^2 = \sigma^2 (X_i)$$

$$\text{Giả sử: } Var(u_i | X_i) = \sigma_i^2 = \sigma^2 X_i^2$$

ta đặt:

$$u_i^* = \frac{u_i}{X_i}$$

Khi đó

$$Var(u_i^* | X_i) = Var\left(\frac{u_i}{X_i} | X_i\right) = \frac{1}{X_i^2} Var(u_i | X_i) = \frac{1}{X_i^2} \sigma^2 \cdot X_i^2 = \sigma^2$$

Vậy ta đã có được phương sai không đổi.

- Sau khi biến đổi u_i thành u_i^* chúng ta sẽ thu được phương sai sai số không đổi (thuần nhất), và chúng ta có thể sử dụng OLS để tìm mô hình hồi qui đã được biến đổi.

$$\frac{Y_i}{X_i} = \beta_1 \cdot \frac{1}{X_i} + \beta_2 + \frac{u_i}{X_i}$$

$$\text{Suy ra: } Y_i^* = \beta_1 X_i^* + \beta_2 + u_i^*$$

Hồi qui Y_i^* với X_i^* . Khi đó phương sai của u_i^* là không đổi, ước lượng OLS sẽ có BLUE.

Đối với dạng hàm khác của phương sai sai số thay đổi.

$$\text{Nếu: } Var(u_i | X_i) = \sigma_i^2 = \sigma^2 X_i$$

$$\text{Khi đó ta sẽ sử dụng trọng số là: } \frac{1}{\sqrt{X_i}}$$

Ta nhân $\frac{1}{\sqrt{X_i}}$ cả hai vế của mô hình.

$$\frac{Y_i}{\sqrt{X_i}} = \beta_1 \cdot \frac{1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + \frac{u_i}{\sqrt{X_i}}$$

Suy ra:

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + u_i^*$$

- Hồi qui Y_i^* với X_{1i}^* và X_{2i}^* bằng OLS.
 - Ta cũng nhận được các ước lượng OLS là BLUE

6.5. Phát hiện phương sai sai số thay đổi

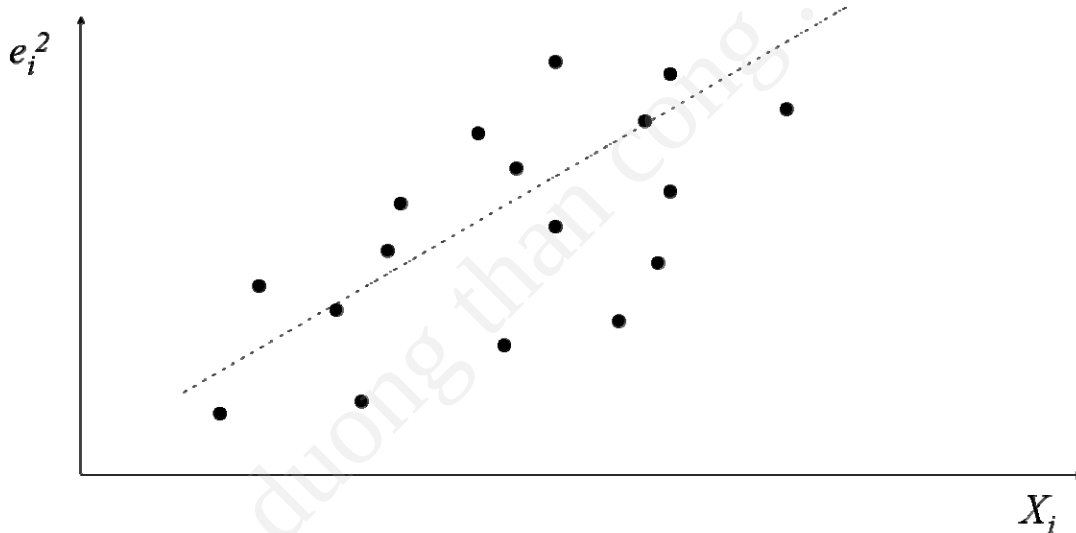
- Do sai số ngẫu nhiên, u_i , là không biết được, do đó chúng ta không xác định được phương sai của u_i .
- Chúng ta sẽ sử dụng phần dư OLS là e_i , để phát hiện vấn đề này.

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

Có rất nhiều đề xuất khác nhau cho việc phát hiện (kiểm định) phương sai sai số thay đổi, sau đây là một số kiểm định thường dùng:

6.5.1. Phương pháp đồ thị

Chúng ta biểu diễn bình phương của phần dư theo X , nếu đồ thị cho thấy rằng khi X thay đổi sẽ làm cho bình phương phần dư thay đổi (tăng hoặc giảm) thì ta có thể nói rằng có hiện tượng phương sai sai số thay đổi.



6.5.2. Kiểm định Park

- Khảo sát mối quan hệ của phần dư e_i với X_i .
- Thực hiện hồi qui phụ.

$$\ln(e_i^2) = \alpha_1 + \alpha_2 \cdot \ln(X_i) + \varepsilon_i$$

- Chạy hồi qui OLS dạng log của phần dư theo X_i , và kiểm định ý nghĩa của α_2 bởi thống kê T (t -test).
 - α_2 có ý nghĩa $\Rightarrow$ có phương sai sai số thay đổi
 - α_2 không có ý nghĩa $\Rightarrow$ không có phương sai thay đổi

6.5.3. Kiểm định Glejser

- Thủ tục kiểm định tương tự như kiểm định Park.
- Thực hiện hồi qui phụ: sử dụng trị tuyệt đối của phần dư.

$$|e_i| = \alpha_1 + \alpha_2 \cdot X_i + \varepsilon_i$$

- Thực hiện hồi qui bằng OLS, và kiểm định ý nghĩa của α_2 bằng thống kê T (t -test).
 - α_2 có ý nghĩa $\Rightarrow$ có phương sai sai số thay đổi
 - α_2 không có ý nghĩa $\Rightarrow$ không có phương sai thay đổi

6.5.4. Kiểm định Goldfeld-Quandt (G-Q)

- Kiểm định này so sánh tổng bình phương các phần dư của hai nhóm số liệu/ quan sát trong mẫu.
- Kiểm định này phù hợp khi chúng ta biết phương sai của sai số có quan hệ với *chỉ một* biến giải thích.
- Nếu có nhiều hơn một biến giải thích trong hồi qui (hồi qui đa biến), chúng ta không biết biến giải thích nào có liên quan với sai số ngẫu nhiên, do vậy kiểm định này không phù hợp.
- Sắp xếp quan sát biến giải thích X theo thứ tự tăng dần.
 - Chúng ta giả định là biết phương sai của sai số là có quan hệ với biến X .
- Bỏ qua c quan sát ở giữa và chia $(n-c)$ quan sát còn lại thành hai nhóm.
- Sử dụng OLS hồi qui hai nhóm quan sát và tính tổng bình phương các phần dư (RSS) cho mỗi nhóm. Ta gọi RSS_1 (cho nhóm có giá trị X bé) và RSS_2 (cho nhóm có giá trị X lớn)

Tính toán kiểm định G-Q.

$$F^*_{df_2, df_1} = \frac{RSS_2 / df_2}{RSS_1 / df_1} \sim F$$

$$df_1 = n_1 - k_1, df_2 = n_2 - k_2. \text{ nếu, } n_1 = n_2 \text{ và } k_1 = k_2. \text{ Thì ta có, } df_1 = df_2.$$

Giả thiết H_0 là G-Q không có phương sai sai số thay đổi.

Bác bỏ H_0 nếu giá trị thống kê tính được lớn hơn giá trị tra bảng với α cho trước ($F^* > F^c$).

6.5.5. Kiểm định White

- Xem xét hồi qui: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$

- Ta thu được phần dư: $e_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \hat{\beta}_4 X_{4i}$
- Thực hiện hồi qui phụ: Hồi qui e_i^2 với tất cả các X và bình phương cũng như tích của chúng:

$$e_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{4i} + \alpha_5 X_{2i}^2 + \alpha_6 X_{3i}^2 + \alpha_7 X_{4i}^2 + \alpha_8 (X_{2i} \cdot X_{3i}) + \alpha_9 (X_{2i} \cdot X_{4i}) + \alpha_{10} (X_{3i} \cdot X_{4i}) + \varepsilon_i$$
- Giả thiết $H_0: \alpha_2 = \alpha_3 = \dots = \alpha_{10} = 0$ (phương sai không đổi)
 - không là bất kỳ hàm nào của các X .
- Thống kê kiểm định: $n.R^2 \sim \chi_{df=p-1}^2$
 - n : Số quan sát trong hồi qui phụ.
 - R^2 : R^2 của hồi qui phụ.
 - p : Số tham số trong hồi qui phụ.
- Bác bỏ H_0 nếu $\chi_{df}^{2*} \sim \chi_{df}^{2c}$

BÀI THỰC HÀNH

Khi sử dụng OLS để hồi qui với biến phụ thuộc là chi phí cho nghiên cứu và phát triển (R&D) và biến độc lập là sản lượng bán (SALES). Chúng ta nghi ngại rằng có thể xảy ra có phương sai sai số thay đổi. Chi phí R&D thay đổi lớn hơn đối với những công ty có sản lượng bán hàng lớn hơn so với những công ty có sản lượng bán hàng nhỏ hơn.

Chạy kiểm định phương sai sai số thay đổi, lưu lại phần dư của OLS. Phần dư OLS có thể lưu lại sau khi ước lượng phần dư.

Thực hiện kiểm định Park:

Dependent Variable: LOG(E2)

Method: Least Squares

Sample: 1 18

Included observations: 18

LOG(E2)=C(1)+C(2)*LOG(SALES)

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	5.687749	6.635175	0.857212	0.4040
C(2)	0.701435	0.603332	1.162602	0.2620
R-squared	0.077897	Mean dependent var		13.36642
Adjusted R-squared	0.020266	S.D. dependent var		2.721342
S.E. of regression	2.693626	Akaike info criterion		4.924093
Sum squared resid	116.0899	Schwarz criterion		5.023023
Log likelihood	-42.31683	Durbin-Watson stat		1.567339

Từ kết quả Kiểm định Park cho thấy α_2 không có ý nghĩa thống kê ở mức 10%

Thực hiện kiểm định Glejser:

Dependent Variable: ABS(E)

Method: Least Squares

Sample: 1 18

Included observations: 18

ABS(E)=C(1)+C(2)*SALES

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	578.5710	678.6950	0.852476	0.4065
C(2)	0.011939	0.005704	2.093059	0.0526
R-squared	0.214951	Mean dependent var		1650.432
Adjusted R-squared	0.165886	S.D. dependent var		2069.046
S.E. of regression	1889.657	Akaike info criterion		18.03062
Sum squared resid	57132868	Schwarz criterion		18.12955
Log likelihood	-160.2756	Durbin-Watson stat		1.743294

Kiểm định Glejser: α_2 không có ý nghĩa thống kê ở mức 5%, nhưng có ý nghĩa ở mức 10%.

Kiểm định White

Dependent Variable: E2

Method: Least Squares

Sample: 1 18

Included observations: 18

E2=C(1)+C(2)*SALES+C(3)*SALES^2

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-6219665.	6459809.	-0.962825	0.3509
C(2)	229.3508	126.2197	1.817077	0.0892
C(3)	-0.000537	0.000449	-1.194952	0.2507
R-squared	0.289583	Mean dependent var		6767046.
Adjusted R-squared	0.194861	S.D. dependent var		14706011
S.E. of regression	13195639	Akaike info criterion		35.77968
Sum squared resid	2.61E+15	Schwarz criterion		35.92808
Log likelihood	-319.0171	Durbin-Watson stat		1.694567

Thống kê kiểm định: $n \cdot R^2 = 18 \cdot 0.2896 = 5.2128$ ($p\text{-value} = 0.0738$)

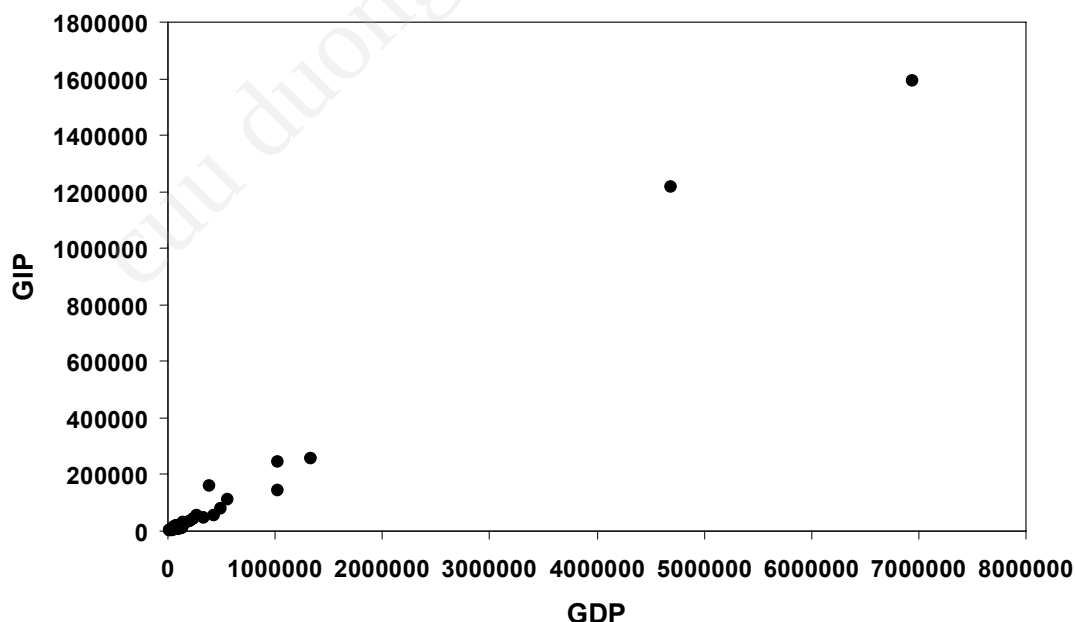
Giá trị tới hạn của phân phối Chi-square là 5.99 và 4.61 cho 5% và 10%.

Bác bỏ H_0 (phương sai sai số không đổi) tại 10%, nhưng không có cơ sở bác bỏ tại mức 5%.

Bài thực hành Kiểm định GOLDFELD-QUANDT

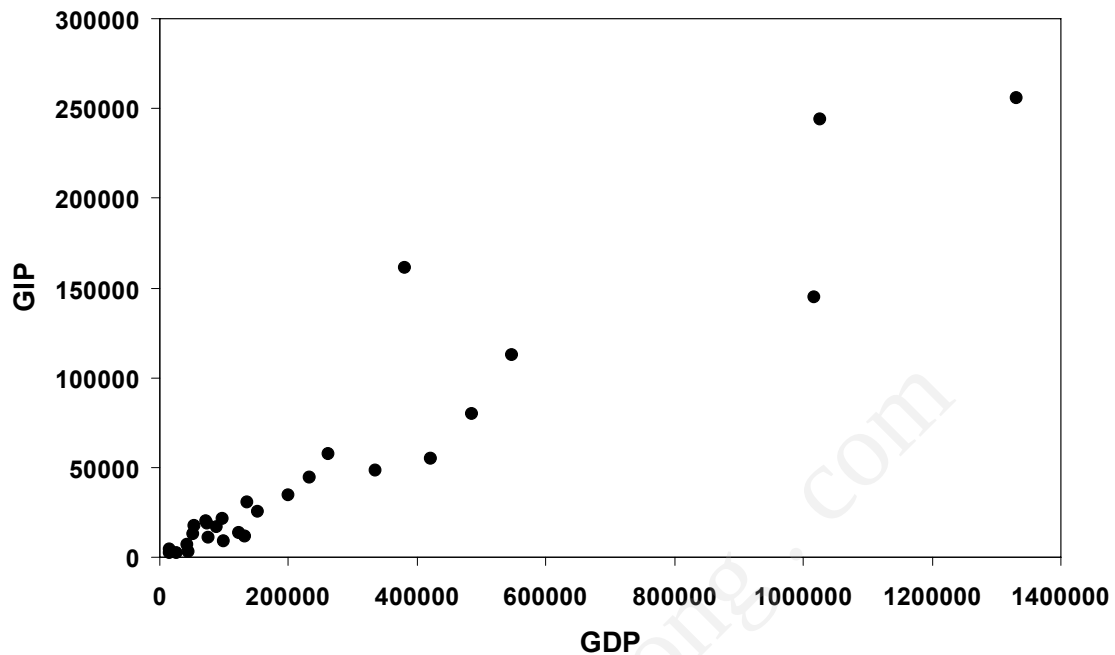
Nghiên cứu về tổng sản phẩm công nghiệp GIP với GDP được tính bằng tỷ \$ của 30 quốc gia năm 1997. (Dữ liệu từ UNIDO). Mẫu được lấy từ những quốc gia có GDP lớn hơn 10 tỷ \$ và GDP/đầu người tối thiểu là 2000\$

Country	Pop	GDP	GIP
Japan	124793	4688974	1218217
Switzerland	7104	261388	57503
Denmark	5207	151266	25927
Norway	4314	122926	13484
USA	264646	6935700	1592989
France	57856	1330998	256316
Belgium	10093	232006	44517
Sweden	8751	198432	34806
Singapore	3268	71039	20648
Netherlands	15382	334286	48595
Hong Kong	6044	130823	11758
Finland	5085	97624	21581
Canada	29109	547203	112617
Italy	57177	1016286	145013
UK	58005	1024609	244397
Ireland	3536	52662	17572
Kuwait	1754	24848	2797
Israel	5362	74121	11349
Spain	39577	483652	80104
Greece	10413	98861	9392
Portugal	9824	87352	17025
Korea, S.	44501	380820	161318
Slovenia	1925	14386	4520
Mexico	89564	420788	55073
Hungary	10162	41506	7227
Malaysia	19695	72505	18874
Chile	13994	50919	13096
Syria	13840	44753	3317
Slovakia	5325	13746	2720
Turkey	59903	135961	31115

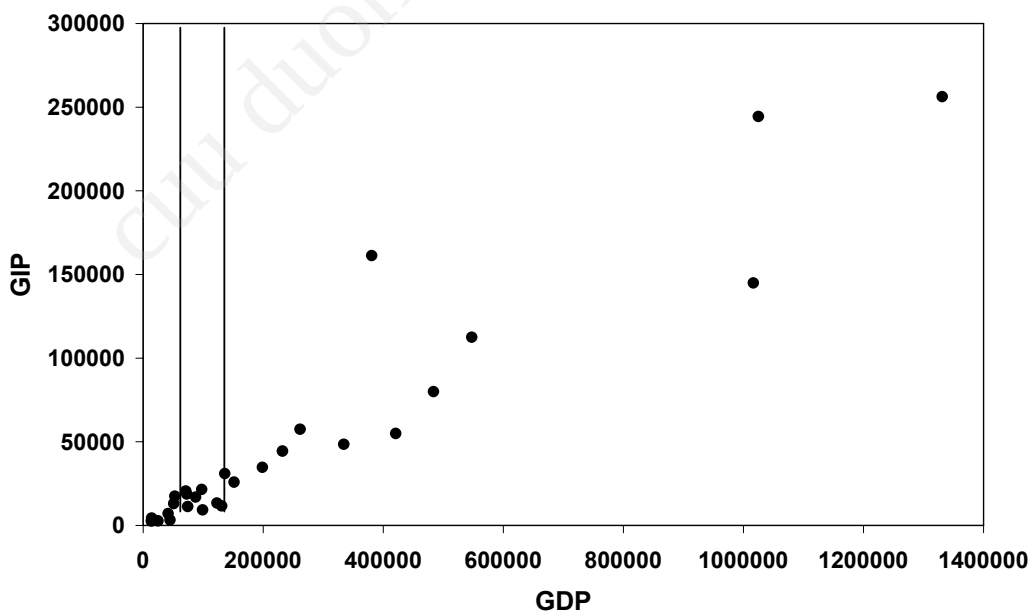


Nhìn vào biểu đồ phân bố của GIP với GDP ở trên, ta thấy rằng Nhật bản và Mỹ có sự khác biệt rất lớn, chúng ta bỏ đi 2 quan sát này.

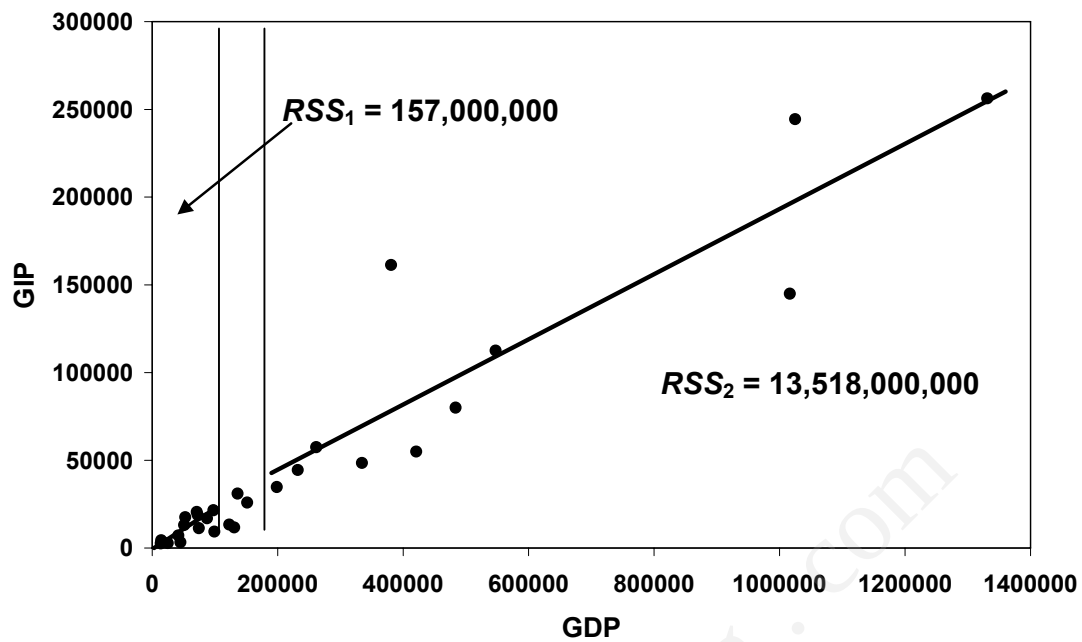
Vẽ lại biểu đồ Nhìn vào biểu đồ phân bố này, ta thấy có sự thay đổi khá lớn về GIP so với GDP. Ta nghĩ rằng có hiện tượng phương sai sai số thay đổi



Mẫu được chia thành 3 phần, phần 1 chứa 3/ 8 quan sát với những giá trị nhỏ của biến X, phần 2 gồm 3/ 8 quan sát với những giá trị lớn, và phần còn lại là 1/ 4 quan sát ở giữa. Bây giờ chúng ta còn 28 quan sát, do đó từng phần sẽ là 11 quan sát dưới, 11 quan sát trên và 6 quan sát ở giữa.



Thực hiện hồi qui đối với các quan sát dưới và trên ta tìm được tổng bình phương các phần dư, ký hiệu là RSS1 và RSS2.

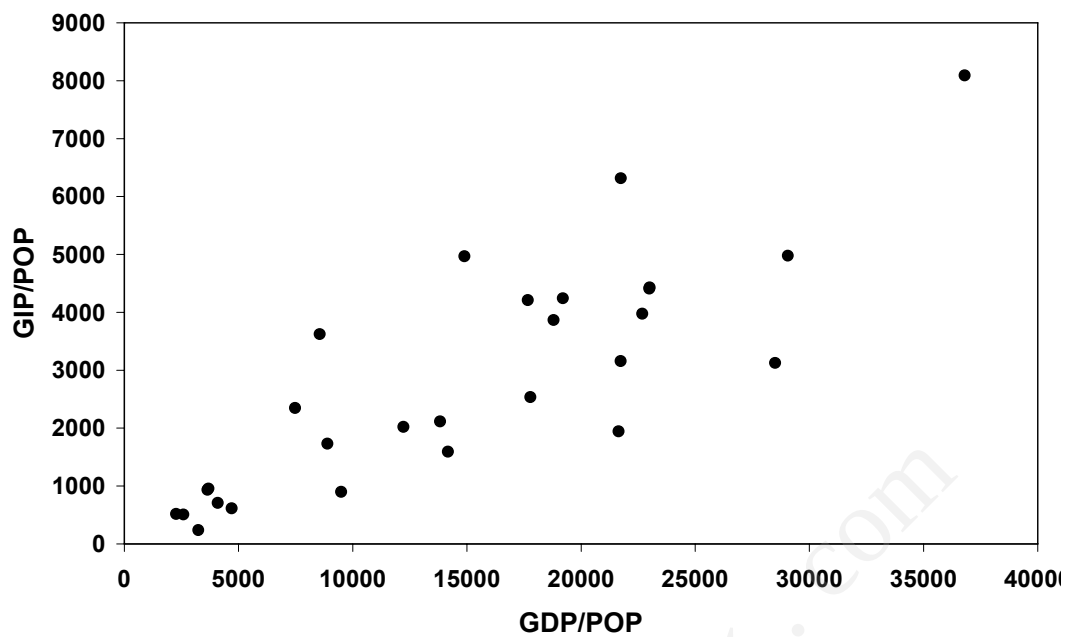


Ta sử dụng thống kê F để kiểm định. Nhìn vào kết quả ta thấy bác bỏ giả thiết H_0 cho rằng không có phương sai sai số thay đổi, do vậy chúng ta phải tìm cách để biến đổi hồi qui

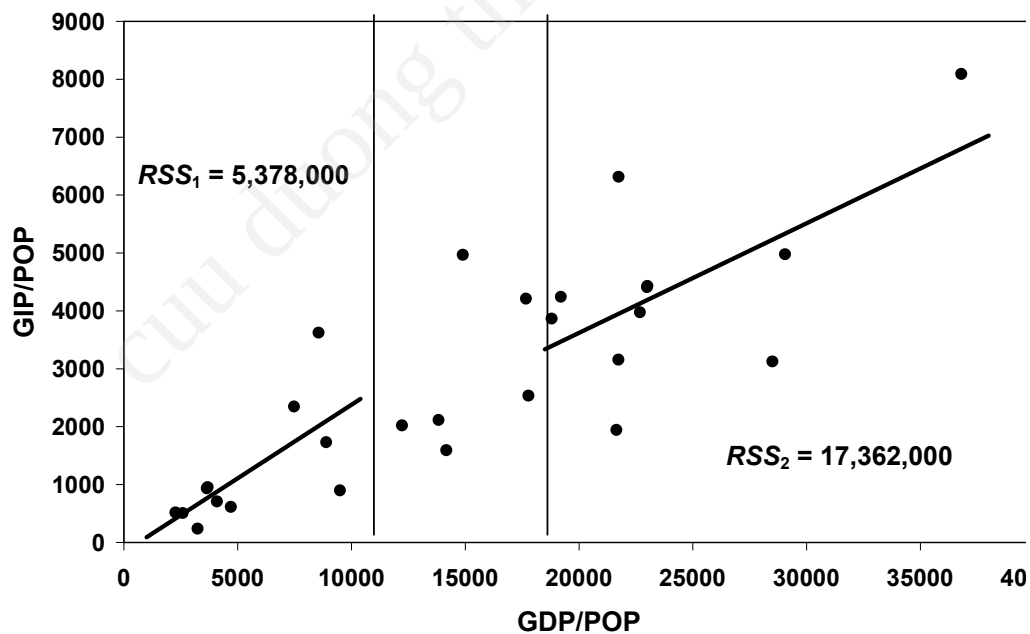
$$F(n_2, n_1) = \frac{RSS_2 / n_2}{RSS_1 / n_1} = \frac{13,518,000,000 / 9}{157,000,000 / 9} = 86.1$$

$$F(9, 9)_{5\%} = 3.18$$

Chúng ta xem xét việc sử dụng trọng số là dân số (POP), khi đó biến phụ thuộc là “GIP/POP” và biến độc lập là “GDP/POP”. Khi đó ta thu được sơ đồ tán xạ như trên, ta sẽ kiểm tra xem có hiện tượng phương sai sai số thay đổi hay không.



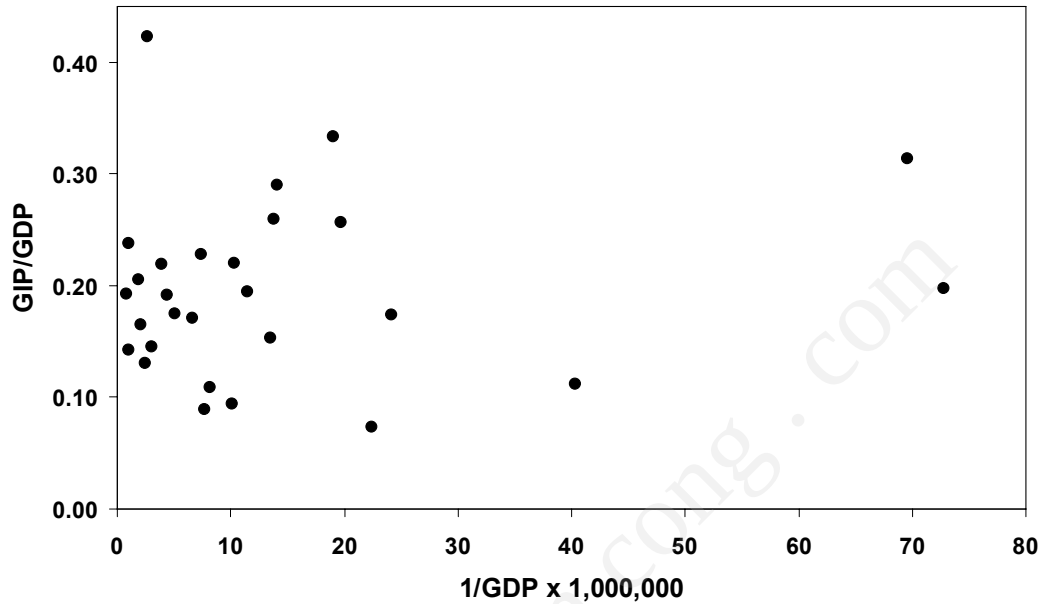
Sử dụng kiểm định GOLDFELD-QUANDT, thực hiện hồi qui 2 bộ số liệu ta thu được RSS_1 và RSS_2



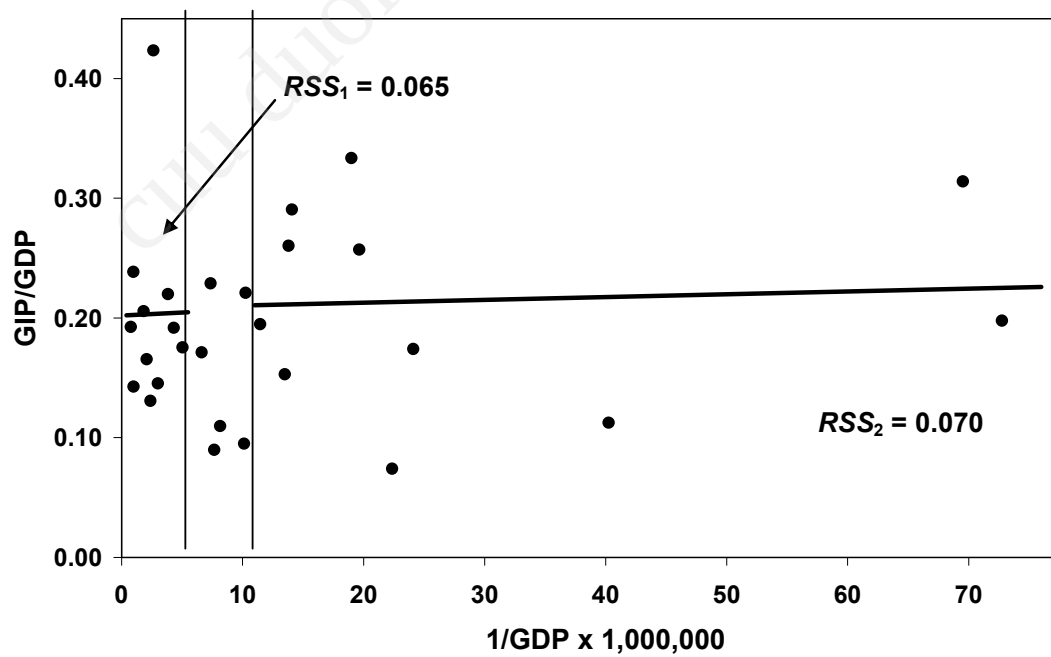
Ta sử dụng thống kê F để kiểm định. Nhìn vào kết quả ta thấy bác bỏ giả thiết H_0 cho rằng không có phương sai sai số thay đổi. Ta sẽ lại phải thực hiện một biến đổi khác.

$$F(n_2, n_1) = \frac{RSS_2 / n_2}{RSS_1 / n_1} = \frac{17,362,000 / 9}{5,378,000 / 9} = 3.23; F(9, 9)_{5\%} = 3.18$$

Giả sử $\text{Var}(u_i) = \sigma_i^2$ và $\sigma_i = \lambda \cdot X_i$ khi đó ta sẽ sử dụng trọng số là $1/\text{GDP}$. Khi đó ta thu được sơ đồ tán xạ như đồ thị.



Tiếp tục sử dụng kiểm định GOLDFELD-QUANDT, thực hiện hồi qui 2 bộ số liệu ta thu được RSS1 và RSS2

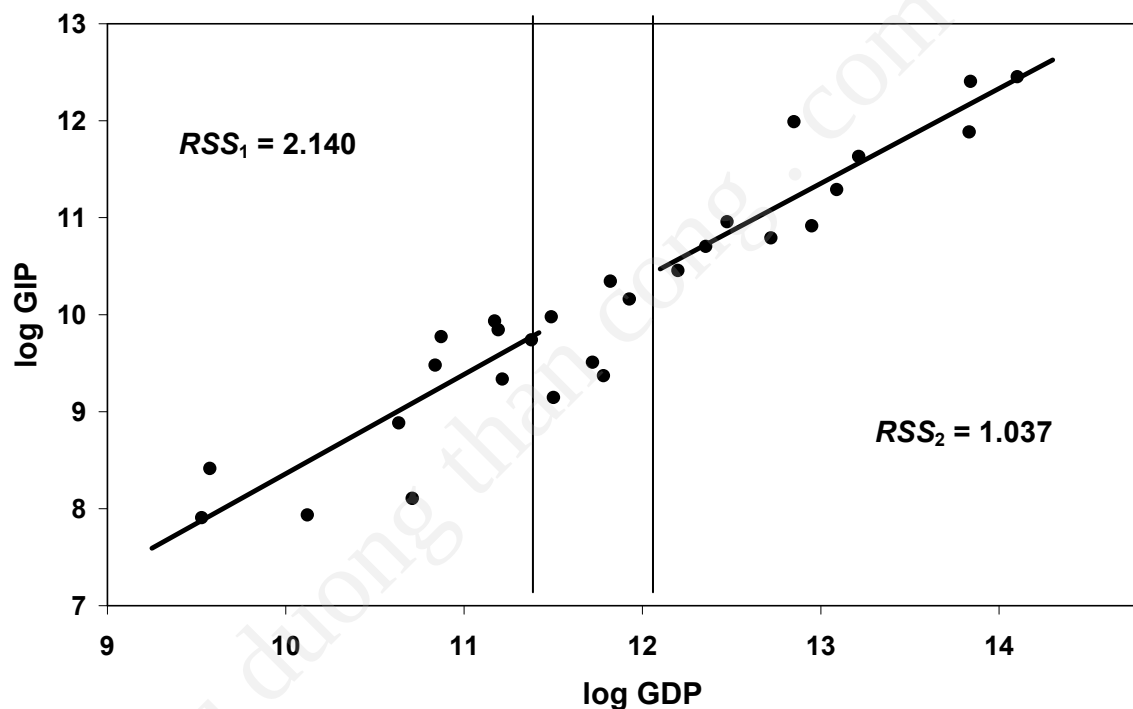


Ta sử dụng thống kê F để kiểm định. Nhìn vào kết quả ta thấy không có cơ sở bác bỏ giả thiết H_0 cho rằng không có phương sai sai số thay đổi, do vậy chúng ta chấp nhận là phương sai sai số không đổi với độ tin cậy 95%

$$F(n_2, n_1) = \frac{RSS_2 / n_2}{RSS_1 / n_1} = \frac{0.070 / 9}{0.065 / 9} = 1.08$$

$$F(9, 9)_{5\%} = 3.18$$

Bây giờ chúng ta sẽ sử dụng thêm phép biến đổi loga. Sử dụng điểm định G-Q chúng ta thực hiện hồi qui và thu được kết quả.



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

$$Y = e^{\beta_1} X^{\beta_2} e^u$$

$$F(n_2, n_1) = \frac{RSS_1 / n_1}{RSS_2 / n_2} = \frac{2.140 / 9}{1.037 / 9} = 2.06$$

$$F(9, 9)_{5\%} = 3.18$$

Nhìn vào kết quả ta thấy không có cơ sở bác bỏ giả thiết H_0 cho rằng không có phương sai sai số thay đổi, do vậy chúng ta chấp nhận là phương sai sai số không đổi với độ tin cậy 95%

Qua các kết quả thu được bởi từng hàm hồi qui, thực hiện so sánh:

$$\hat{GIP} = 604 + 0.194GDP \quad R^2 = 0.89$$

(5700) (0.013)

$$\frac{\hat{GIP}}{POP} = 612 \frac{I}{POP} + 0.182 \frac{GDP}{POP} \quad R^2 = 0.70$$

(1371) (0.016)

$$\frac{\hat{GIP}}{GDP} = 0.189 + 533 \frac{I}{GDP} \quad R^2 = 0.02$$

(0.019) (841)

$$\log \hat{GIP} = -1.694 + 0.999 \log GDP \quad R^2 = 0.90$$

(0.785) (0.066)

Sau khi so sánh các kết quả thu được, ta thấy rằng lựa chọn mô hình (hàm) hồi qui dạng loga là thuyết phục nhất. Đảm bảo giả thiết phương sai sai số không đổi và có R^2 cao.

$$\log \hat{GIP} = -1.694 + 0.999 \log GDP \quad R^2 = 0.90$$

(0.785) (0.066)

$$GIP = e^{-1.694} GDP^{0.999} = 0.184 GDP^{0.999}$$

CHƯƠNG VII

TƯƠNG QUAN CHUỖI

Giả thiết của OLS chúng ta đã nêu ở chương II và chương III phát biểu rằng các số hạng sai số u_t là độc lập, không có tương quan với nhau. Tính chất này gọi là độc lập chuỗi. Mặc dù vậy khi phân tích dữ liệu chuỗi thời gian, giả thiết này thường sẽ bị vi phạm. Các số hạng sai số cho các thời đoạn không quá cách xa có thể có tương quan. Tính chất này được gọi là tương quan chuỗi hay tự tương quan (các thuật ngữ này sẽ được sử dụng thay thế nhau). Chúng ta đã biết rằng một số nhân tố giải thích cho sự có mặt của số hạng sai số u_t . Đó có thể là các biến bị loại bỏ, bỏ qua sự phi tuyến, các sai số đo lường, hay các tác động không dự đoán được. Ba nhân tố đầu tiên trong các nhân tố này có thể dẫn đến các sai số tương quan chuỗi.

7.1. Nguyên nhân của tương quan chuỗi

Tương quan chuỗi xảy ra có thể do nhiều nguyên nhân khác nhau, dưới đây là một số những nguyên nhân thường gặp gây nên hiện tượng tương quan chuỗi:

7.1.1. Nguyên nhân khách quan

a. Quán tính:

Nét nổi bật của hầu hết các chuỗi thời gian trong kinh tế là quán tính. Chúng ta có thể thấy các chuỗi thời gian như tổng sản phẩm, chỉ số giá mang tính chu kỳ. Chẳng hạn nếu chúng ta ở đầu thời kỳ tăng trưởng kinh tế, tổng sản phẩm có xu hướng đi lên, giá trị thời điểm sau thường cao hơn thời điểm trước. Vì vậy hồi qui chuỗi thời gian, các quan sát kế tiếp có nhiều khả năng phụ thuộc lẫn nhau.

Một ví dụ khác của tương quan chuỗi cũng có thể được tìm thấy trong dữ liệu thị trường chứng khoán. Giá của một chứng khoán đặc biệt nào đó hoặc một chỉ số thị trường chứng khoán tại thời điểm đóng cửa của những ngày liên tiếp hoặc trong những giờ liên tiếp có thể tương quan theo chuỗi.

b. Hiện tượng mang nhện:

Chúng ta có thể thấy hiện tượng này xảy ra đối với việc cung một số hàng hoá nào đó, cung hàng hoá phản ứng lại với giá có trễ một khoảng thời gian.

c. Hiện tượng trễ:

Trong hồi qui chuỗi thời gian, nhiều trường hợp ta thấy rằng biến phụ thuộc ở thời kỳ t không những phụ thuộc vào các biến độc lập mà còn phụ thuộc vào chính biến đó ở thời kỳ $t-1$. Chúng ta dễ dàng thấy được điều này khi hồi qui tiêu dùng và thu nhập, tiêu dùng ở thời điểm hiện tại không những phụ thuộc vào thu nhập mà còn phụ thuộc vào tiêu dùng (thói quen tiêu dùng) ở thời kỳ trước đó. Như vậy nếu bỏ qua số hạng trễ ta sẽ gặp hiện tượng tương quan chuỗi giữa các sai số.

7.1.2. Nguyên nhân chủ quan

a. Bỏ sót biến:

Ví dụ, giả sử một biến phụ thuộc Y_t tương quan với các biến độc lập X_{1t} và X_{2t} , nhưng chúng ta không tính đến biến X_{2t} trong mô hình. Tác động của biến này sẽ được bao gộp qua số hạng sai số u_t . Bởi vì nhiều biểu hiện chuỗi thời gian kinh tế có chiều hướng theo thời gian, X_{2t} có thể phụ thuộc vào $X_{2,t-1}$, $X_{2,t-2}$,... Điều này sẽ biến thành sự tương quan rõ ràng giữa u_t và u_{t-1} , u_{t-2} ,... do đó vi phạm giả thiết độc lập chuỗi. Vậy, các biến bị loại bỏ có thể tạo sự tự tương quan trong các sai số.

b. Định dạng hàm sai:

Tương quan chuỗi cũng có thể được gây nên bởi đặc trưng sai về dạng hàm. Ví dụ, giả sử mối quan hệ giữa Y và X là bậc hai nhưng ta giả thiết là tuyến tính. Vậy số hạng sai số u_t sẽ phụ thuộc vào X^2 . Nếu X tăng hoặc giảm theo thời gian, u_t cũng sẽ biểu hiện chiều hướng như vậy, cho thấy sự tự tương quan.

c. Do đo lường, xử lý số liệu:

Sai số có hệ thống trong đo lường cũng gây nên sự tự tương quan. Ví dụ, giả sử một công ty đang cập nhật số liệu hàng hóa tồn kho trong một thời đoạn cho trước. Nếu có một sai sót có tính hệ thống xảy ra trong cách đo lường, dự trữ tồn kho tích lũy sẽ phản ánh các sai số đo lường tích lũy. Các sai số này sẽ cho thấy như là sự tương quan chuỗi.

7.2. Các hệ quả khi bỏ qua tương quan chuỗi

Như ta đã đề cập trong chương 2, với giả thiết kỳ vọng của u_t bằng 0 và không có sự tương quan giữa u và X , các ước lượng OLS là không chệch. Vì sự chứng minh các tính chất này không phụ thuộc vào giả thiết $\text{Cov}(u_t, u_s) \neq 0$, giả thiết bị vi phạm bởi sự có mặt của tự tương quan, do đó các ước lượng OLS (và các dự báo dựa trên chúng) là không chệch ngay cả khi các số hạng sai số tương quan theo chuỗi. Nhưng khi đó các ước lượng sẽ không còn hiệu quả nữa.

Ta có thể cho thấy rằng nếu tương quan chuỗi trong u_t là dương và biến độc lập X_t tăng lên theo thời gian (trường hợp thường thấy), thì phương sai phần dư ước lượng ($\hat{\sigma}^2$) sẽ là một ước lượng quá thấp và giá trị của R^2 sẽ là một ước lượng quá cao. Nói cách khác, độ thích hợp sẽ bị phóng đại và các sai số chuẩn ước lượng sẽ nhỏ hơn các sai số chuẩn thực sự.

Chúng ta vừa biện luận rằng trong trường hợp thông thường khi mà tương quan chuỗi là dương và biến độc lập tăng lên theo thời gian, các sai số chuẩn ước lượng nhỏ hơn các sai số thực, và do đó sẽ là ước lượng quá thấp. Điều này có nghĩa là các trị thống kê t sẽ là các ước lượng quá cao, và do vậy một hệ số hồi quy có vẻ có ý nghĩa nhưng thực tế có thể là không. Các phương sai ước lượng của các thông số sẽ bị chệch. Vì vậy, các kiểm định t và F không còn hợp lệ.

Từ đó ta có thể thấy rằng nếu tương quan chuỗi giữa các số hạng nhiễu ngẫu nhiên trong mô hình hồi quy bị bỏ qua và thủ tục OLS được dùng để ước lượng các thông số, thì sẽ có các tính chất sau:

- ❖ Các ước lượng và các dự báo dựa trên chúng sẽ vẫn không chệch.

- ❖ Các ước lượng OLS không còn BLUE nữa và sẽ không hiệu quả. Các dự báo cũng sẽ không hiệu quả.
- ❖ Các phương sai ước lượng của các hệ số hồi qui sẽ chệch, và do đó các kiểm định các giả thuyết sẽ không còn hợp lệ. Nếu tương quan chuỗi là dương và biến độc lập X_t tăng lên theo thời gian, thì các sai số chuẩn sẽ là các ước lượng quá thấp của các giá trị thực. Điều này có nghĩa rằng R^2 tính toán sẽ là một ước lượng quá cao, cho thấy một sự thích hợp tốt hơn so với tồn tại trên thực tế. Cũng vậy, các trị thống kê t trong trường hợp như vậy sẽ có khuynh hướng có ý nghĩa hơn giá trị thực sự của chúng.

7.3. Kiểm định tương quan chuỗi bậc nhất

7.3.1. Tương quan chuỗi bậc nhất

Đầu tiên, ta xét trường hợp đặc biệt nhất của tương quan chuỗi gọi là **tương quan chuỗi bậc nhất**. Mặc dù ta dùng mô hình hồi qui tuyến tính đơn để khảo sát các vấn đề, tất cả kết quả cũng khái quát hóa cho trường hợp hồi qui đa biến. Nếu tương quan chuỗi tồn tại, thì $\text{Cov}(u_t, u_s) \neq 0$, nghĩa là, sai số cho thời đoạn t là tương quan với sai số cho thời đoạn s . Giả thiết của tự tương quan bậc nhất được phát biểu chính thức như sau:

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t \quad -1 < \rho < 1$$

Vậy sai số u_t quan hệ với sai số của thời đoạn trước (u_{t-1}), một số hạng sai số mới (ε_t), và một thông số mới ρ . Bởi vì ρ là hệ số của số hạng sai số trễ một thời đoạn, được gọi là **hệ số tự tương quan bậc nhất**. Quá trình này được gọi là **quá trình tự hồi qui bậc nhất**, ký hiệu là **AR(1)**.

Các sai số mới ε_t giả thiết thỏa mãn các điều kiện: Các sai số ε_t tuân theo phân phối một cách độc lập kỳ vọng bằng 0 và phương sai không đổi sao cho $E(\varepsilon_t) = 0$, $\text{Var}(\varepsilon_t) = \sigma_\varepsilon^2$, và $\text{Cov}(\varepsilon_t, \varepsilon_{t+s}) = 0$ với $s \neq 0$. Vậy các số hạng sai số mới được giả thiết để có cùng tính chất với các tính chất mà thủ tục OLS giả thiết u_t phải có.

Bởi vì u_t phụ thuộc vào u_{t-1} , ta có thể kỳ vọng là chúng tương quan. Lưu ý rằng u_t không phụ thuộc trực tiếp vào u_{t-2} ; tuy nhiên, lại phụ thuộc gián tiếp qua u_{t-1} bởi vì u_{t-1} phụ thuộc trực tiếp vào u_{t-2} . Vậy, u_t tương quan với tất cả sai số quá khứ. Nếu đồng phương sai là dương, thì có *tự tương quan dương*, và khi đồng phương sai âm, ta có *tự tương quan âm*.

7.3.2. Kiểm định Durbin-Watson

Mặc dù biểu đồ phần dư là một phương pháp đồ thị hữu ích để nhận dạng sự có mặt của tương quan chuỗi, các kiểm định chính thức cho tự tương quan vẫn là cần thiết. Trong phần này, ta trình bày kiểm định phổ biến nhất cho tương quan chuỗi bậc nhất, gọi là **kiểm định Durbin-Watson (DW)**. Kiểm định DW có một lợi thế rất lớn là dựa trên các phần dư đã ước lượng được, vì vậy ngày nay nó rất thông dụng.

Các bước thực hiện kiểm định Durbin-Watson cho AR(1) được mô tả cho mô hình hồi qui đa biến sau đây:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t \quad -1 < \rho < 1$$

Bước 1 Ước lượng mô hình bởi bình phương tối thiểu thông thường và tính toán các phần dư e_t

Bước 2 Tính toán thông kê Durbin-Watson:

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

Vì các $\sum e_t^2$ và $\sum e_{t-1}^2$ chỉ khác nhau 1 quan sát, chúng xấp xỉ bằng nhau, do đó nếu đặt $\sum e_t^2 = \sum e_{t-1}^2$ thì lúc đó:

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

Bây giờ chúng ta định nghĩa: $\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}$

Khi đó: $d \approx 2(1 - \hat{\rho})$

Trong đó $\hat{\rho}$ là hệ số tự tương quan bậc nhất của mẫu, đó là ước lượng của ρ . Vì $-1 \leq \rho \leq 1$, nên suy ra $0 \leq d \leq 4$.

Từ đó chúng ta thấy rằng $\hat{\rho} = 0$ suy ra $d \approx 2$ tức là nếu không tồn tại tương quan chuỗi thì d được kỳ vọng bằng 2. Nếu d càng gần 0 thì càng chứng tỏ có tương quan thuận chiều và d càng gần 4 thì có tương quan ngược chiều.

Phân phối chính xác của d phụ thuộc vào các quan sát trên các biến X . Phân phối này được dùng để xây dựng các vùng giới hạn cho kiểm định DW.

Bước 3

Để kiểm định $H_0: \rho = 0$ đối lại $\rho > 0$ (kiểm định một phía), tra bảng, các giá trị tới hạn cho thông kê Durbin-Watson, và viết các số dL và dU . Lưu ý rằng bảng này cho giá trị k' , là số các hệ số hồi qui được ước lượng, ngoại trừ số hạng hằng số. Bác bỏ H_0 nếu $d \leq dL$. Nếu $d \geq dU$, ta không thể bác bỏ H_0 . Nếu $dL < d < dU$, kiểm định chưa thể kết luận.

Bước 4 Để kiểm định tương quan chuỗi âm (nghĩa là, cho $H_1: \rho < 0$), dùng $4 - d$. Điều này được thực hiện khi d lớn hơn 2. Nếu $4 - d \leq dL$, ta kết luận rằng có tự tương quan âm ý nghĩa. Nếu $4 - d \geq dU$, ta kết luận rằng không có tự tương quan âm. Kiểm định chưa thể kết luận nếu $dL < 4 - d < dU$.

Durbin và Watson xếp thành bảng các giá trị tới hạn cho các phân phối của các giới hạn đối với d , với các giá trị khác nhau của kích thước mẫu n và số các hệ số k' . Khi kiểm định là chưa thể kết luận, ta có thể thử dùng kiểm định nhân tử lagrange được mô tả tiếp sau.

Qui tắc của kiểm định DW được trình bày như sau:

Tự tương quan dương ($\rho > 0$)			Tự tương quan âm ($\rho < 0$)		
Bác bỏ $\rho = 0$	Miền chưa thể kết luận	Chấp nhận $\rho = 0$	Miền chưa thể kết luận	Bác bỏ $\rho = 0$	
0	dL	dU	2	$4 - dU$	$4 - dL$

Kiểm định DW là không hợp lệ nếu vài biến X là hiệu ứng trễ của biến phụ thuộc nghĩa là, nếu chúng có dạng $Y_{t-1}, Y_{t-2}, \dots$

7.3.3. Kiểm định nhân tử Lagrange

Kiểm định nhân tử Lagrange (LM) rất hữu dụng trong việc nhận dạng tương quan chuỗi không chỉ với bậc nhất mà cũng cho cả các bậc cao hơn, nhưng ở đây ta tự hạn chế cho trường hợp bậc nhất. Trường hợp tổng quát sẽ được xét đến trong phần sau.

Để bắt đầu kiểm định này, chúng ta xuất phát từ:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + \rho u_{t-1} + \varepsilon_t$$

Do đó kiểm định đối với $\rho = 0$ có thể được xử lý như kiểm định LM đối với việc thêm biến u_{t-1} (là chưa biết, và do đó ta có thể dùng e_{t-1} hay có thể ký hiệu $\hat{u}_{t-1}$ để thay vào). Các bước thực hiện kiểm định LM như sau:

Bước 1 Bước này đúng như Bước 1 của kiểm định DW; nghĩa là ước lượng Phương trình theo OLS và tính toán các phần dư của nó.

Bước 2 Hồi qui e_t theo một hằng số, $X_{2t}, \dots, X_{kt}$ và e_{t-1} , dùng $n - 1$ quan sát từ 2 đến n . Kế đến tính toán $LM = (n - 1)R^2$ từ hồi qui phụ này. $n - 1$ được dùng bởi vì số quan sát hiệu quả là $n - 1$.

Bước 3 Bác bỏ giả thuyết không của tự tương quan có giá trị bằng không và củng cố giả thuyết $\rho \neq 0$ nếu $(n - 1)R^2 > \chi^2(1, \alpha)$, giá trị $\chi^2(1, \alpha)$ trong phân phối chi-square với 1 bậc tự do sao cho diện tích vùng bên phải của nó là α . Một cách khác, tính toán giá trị $p = \text{Prob}(\chi^2 > LM)$, vùng bên phải của LM trong phân phối χ^2 . Nếu giá trị $p < \alpha$, chắc chắn bác bỏ $H_0: \rho = 0$ và kết luận rằng tự tương quan là có ý nghĩa.

Nếu đã có tương quan chuỗi trong các phần dư, ta có thể kỳ vọng e_t có quan hệ với e_{t-1} .

Lưu ý rằng kiểm định LM không có tình trạng không thể kết luận như của kiểm định DW. Tuy nhiên, đó là kiểm định mẫu lớn và cần ít nhất 30 bậc tự do để có ý nghĩa.

7.4. Tương quan chuỗi bậc cao

Trong phần này, chúng ta sẽ thảo luận về các thủ tục để kiểm định tính tự tương quan bậc cao hơn và để ước lượng các thông số mô hình khi các *số hạng nhiễu* tuân theo tính tương quan bậc tổng quát. Các đặc trưng chung của mô hình với số hạng sai số tự hồi quy được cho như sau:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + u_t \quad (7.1)$$

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_p u_{t-p} + \varepsilon_t \quad (7.2)$$

Phương trình (7.2) còn được gọi là quá trình tự hồi quy bậc p của các phần dư hay AR(p). Nếu chúng ta có dữ liệu theo từng quý, chúng ta có thể kỳ vọng rằng mô hình tự hồi quy *bậc bốn* là hợp lý. Tương tự, dữ liệu theo tháng có khả năng có tính tự tương quan bậc 12 và dữ liệu theo giờ có thể có tính tương quan chuỗi bậc 24. Vì thế, chúng ta cần những đại lượng để xác định cấu trúc sai số tự hồi quy bậc tổng quát cũng như các thủ tục ước lượng có thể áp dụng cho trường hợp bậc cao hơn.

7.4.1. Kiểm định tính tự tương quan bậc cao LM Breusch – Godfrey

Kiểm định LM, mô tả trong phần trước được dùng để kiểm định tính tương quan chuỗi bậc nhất, dễ dàng được mở rộng áp dụng cho bậc cao hơn với cỡ mẫu không được nhỏ. Kiểm định này được gọi là kiểm định Breusch (1978) – Godfrey (1978). Việc thực hiện thủ tục này sẽ rõ ràng hơn nếu chúng ta kết hợp hai phương trình 7.1 và 7.2 như sau:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_p u_{t-p} + \varepsilon_t \quad (7.3)$$

Giả thuyết không là mỗi giá trị trong số các giá trị ρ đều bằng zero (nghĩa là $\rho_1 = \rho_2 = \dots = \rho_p = 0$) đối với giả thuyết ngược lại cho rằng ít nhất một trong các giá trị ρ này khác zero. *Giả thuyết không* này rất giống giả thuyết mà chúng ta đã trình bày trong chương 6 để kiểm định sự thêm vào các biến mới. Trong trường hợp này, những biến mới là $u_{t-1}, u_{t-2}, \dots, u_{t-p}$ mà có thể ước lượng bằng phần dư $e_{t-1}, e_{t-2}, \dots, e_{t-p}$. Các bước tiến hành kiểm định như sau:

Bước 1 Ước lượng phương trình (7.1) theo phương pháp OLS và tính ra các giá trị phần dư e_t .

Bước 2 Hồi quy phần dư e_t theo tất cả các biến độc lập trong phương trình (7.1) cũng như các phần dư $u_{t-1}, u_{t-2}, \dots, u_{t-p}$. Số lần quan sát hiệu quả cho bước hồi quy phụ này phải là $n-p$ vì $t-p$ được xác định cho thời đoạn từ $p+1$ đến n .

Bước 3 Tính toán giá trị $(n-p)R^2$ trong phần hồi quy phụ ở bước 2. Nếu giá trị này lớn hơn $\chi_p^2(\alpha)$ thì giá trị của phân phối Chi-square với p bậc tự do là phần diện tích phía bên phải điểm α , và từ đây chúng ta sẽ bác bỏ giả thuyết $H_0: \rho_1 = \rho_2 = \dots = \rho_p = 0$ và chấp nhận giả thuyết H_1 : có ít nhất một trong số các giá trị ρ khác không.

Mặc dù bản thân thủ tục kiểm định thì không phức tạp nhưng chúng ta cần phải quyết định số bậc p đối với mô hình tự hồi quy cho trước bởi phương trình (7.2). Tính định kỳ của dữ liệu (theo quý, tháng, tuần, hay theo bất cứ khoảng thời gian nào) sẽ góp phần xác định bậc p . Trong bước 2, người ta đã chỉ ra rằng cỡ mẫu thích hợp là $n-p$. Hơn nữa, bước hồi quy phụ có hệ số tự hồi quy p cộng thêm các hệ số k cho $k-1$ biến giải thích và số hạng hằng số.

Vì vậy, $n-p$ tối thiểu phải bằng $p+k$ (nếu không chúng ta sẽ có bậc tự do âm). Điều này có nghĩa là n tối thiểu phải bằng $k+2p$ trước khi có thể ước lượng được phần hồi quy phụ. Nếu cỡ mẫu không tương xứng, chúng ta có thể bỏ bớt một vài số hạng tự hồi quy. Ví dụ, với dữ liệu theo tháng, chúng ta có thể thiết lập các biến trễ ứng với $t=1, 2, 3$, và 12 và đặt các hệ số tự hồi quy khác bằng zero.

7.4.2. Kiểm định Engel's Arch

Các loại tương quan chuỗi được trình bày cho đến nay chỉ xem xét đến số hạng sai số u_t . Do đó trong $AR(p)$ chúng ta xem như u_t phụ thuộc tuyến tính vào p sai số quá khứ $u_{t-1}, u_{t-2}, \dots, u_{t-p}$.

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_p u_{t-p} + \varepsilon_t$$

Đó một loại tương quan chuỗi thường gặp trong dữ liệu theo thời gian, đặc biệt khi tạo ra các dự báo. Một số các chuyên gia dự báo nhận thấy rằng phương sai của sai số dự báo không phải là một hằng số mà thay đổi theo từng thời đoạn. Ví dụ, khi Cục Dự Trữ Liên Bang chuyển từ kiểm soát lãi suất sang kiểm soát tăng trưởng tiền tệ, như đã được thực hiện trước kia, lãi suất trở nên biến động rất nhiều (đó là, chúng bắt đầu thay đổi nhiều xung quanh giá trị trung bình). Vì vậy, các sai số dự báo liên quan đến dự báo lãi suất có tính chất gọi là phương sai của sai số thay đổi. Mặc dù có thể kỳ vọng chỉ là một sự “thay đổi về cấu trúc” trong phương sai, người ta lại nhận thấy rằng phương sai đã thay đổi đều đặn. Sự thay đổi phương sai của sai số tương tự cũng được nhận thấy khi chính sách tỷ giá hối đoái chuyển từ cố định sang linh hoạt. Trong trường hợp linh hoạt, tỷ giá hối đoái dao động mạnh, làm cho phương sai dự báo lớn hơn. Trong lý thuyết tiền tệ và lý thuyết tài chính, tập danh mục tài sản tài chính là hàm của trung bình và phương sai kỳ vọng của suất thu lợi. Sự biến động mạnh của giá chứng khoán hay suất thu lợi thường là biểu hiện của phương sai không cố định theo thời gian. Engel (1982) đã đưa ra một phương pháp tiếp cận mới để mô hình hóa phương sai của sai số thay đổi đối với dữ liệu theo thời gian. Tác giả đặt tên cho mô hình là **ARCH (autoregressive conditional heteroscedascity – phương sai của sai số thay đổi có điều kiện tự hồi qui)**. Quá trình các phương sai được tính được giả định như sau:

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \dots + \alpha_p u_{t-p}^2$$

Phương trình (7.3) được biết đến như là quá trình ARCH bậc p . Thuật ngữ *tự hồi qui* được dùng vì phương sai sai số tại thời điểm t được giả định là phụ thuộc vào bình phương các số hạng sai số trước đó. Tương tự, phương sai tại thời điểm t là có điều kiện theo phương sai ở các giai đoạn trước và do đó hình thành thuật ngữ *phương sai của sai số thay đổi có điều kiện*.

Kiểm định ARCH có giả thiết ban đầu $H_0 : \sigma_1 = \sigma_2 = \dots = \sigma_p = 0$. Các bước kiểm định như sau:

Bước 1 Ước lượng phương trình (7.1) theo OLS

Bước 1 Tính phần dư $\hat{u}_t = Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{2t} - \hat{\beta}_3 X_{3t} - \dots - \hat{\beta}_k X_{kt}$, bình phương các phần dư và tính $\hat{u}_{t-1}^2, \hat{u}_{t-2}^2, \hat{u}_{t-3}^2, \dots, \hat{u}_{t-k}^2$.

Bước 3 Hồi qui $\hat{u}_t^2$, theo một hằng số, $\hat{u}_{t-1}^2, \hat{u}_{t-2}^2, \hat{u}_{t-3}^2, \dots, \hat{u}_{t-k}^2$. Đây là hồi qui phụ với $n-p$ quan sát.

Bước 4 Từ R^2 của hồi qui phụ, tính $(n-p)R^2$. Theo điều kiện giả thuyết H_0 , $(n-p)R^2$ có phân bố chi bình phương với p bậc tự do. Bác bỏ H_0 nếu $(n-p)R^2 > \chi_p^2(\alpha)$, điểm nằm trên χ_p^2 với diện tích α nằm về phía bên phải.

7.5. Xử lý tương quan chuỗi

7.5.1. Thay đổi dạng hàm số

Tương quan chuỗi có thể do mô hình bị định dạng sai, ví dụ, giả sử rằng quan hệ là bậc hai và đáng ra ta hồi qui Y theo X và X^2 . Nếu X tăng hoặc giảm có hệ thống theo thời gian, hồi qui của Y chỉ theo X sẽ hiển nhiên thể hiện sự tương quan chuỗi. Giải pháp ở đây là thiết lập lại mô hình có tính đến số hạng bậc hai sao cho không có tương quan chuỗi xuất hiện.

Khi một hàm đã được hiệu chỉnh mà không thể loại bỏ được tính chất tự tương quan, người ta có thể sử dụng một số thủ tục khác để đưa ra các ước lượng những giá trị nhận được từ phương pháp OLS. Lưu ý rằng các phương pháp này chỉ được áp dụng cho loại dữ liệu chuỗi thời gian. Đối với loại dữ liệu chéo, người ta có thể tái sắp xếp các giá trị quan sát theo bất cứ hình thái nào và nhận được một trị thống kê DW chấp nhận được. Tuy nhiên, điều này cũng nói lên rằng thực hiện kiểm định DW cho loại dữ liệu chéo là không có ý nghĩa. Vì dữ liệu chuỗi thời gian không thể tái sắp xếp được nên chúng ta cần quan tâm đến mối tương quan chuỗi có thể có giữa chúng.

7.5.2. Thiết lập mô hình trong các sai phân bậc nhất

Trong việc nghiên cứu kinh tế lượng thực nghiệm, cách thức mà người ta thường sử dụng để vượt qua vấn đề này là thiết lập một số mô hình theo sai phân bậc nhất, nghĩa là hiệu số giữa giá trị tại mốc thời gian t và $t-1$. Trong trường hợp này, chúng ta sẽ ước lượng $\Delta Y_t = \beta_0 + \beta_1 \Delta X_t + \varepsilon_t$, trong đó $\Delta Y_t = Y_t - Y_{t-1}$ và $\Delta X_t = X_t - X_{t-1}$.

7.5.3. Thủ tục tính lặp Cochrane – Orcutt

Thủ tục Tính lặp Cochrane – Orcutt (CORC) (tác giả Cochrane và Orcutt, 1949) yêu cầu có sự biến đổi mô hình hồi quy:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + u_t \quad (7.4)$$

thành dạng mô hình có thể áp dụng bằng thủ tục OLS. Phương trình trên được viết lại theo thời đoạn $t-1$, chúng ta có:

$$Y_{t-1} = \beta_1 + \beta_2 X_{2(t-1)} + \beta_3 X_{3(t-1)} + \dots + \beta_k X_{k(t-1)} + u_{t-1} \quad (7.5)$$

Lấy (7.4) trừ đi (7.5) sau khi đã nhân số hạng của (7.5) với ρ , chúng ta có

$$Y_t - \rho Y_{t-1} = \beta_1(1 - \rho) + \beta_2 [X_{2t} - \rho X_{2(t-1)}] + \beta_3 [X_{3t} - \rho X_{3(t-1)}] + \dots + \beta_k [X_{kt} - \rho X_{k(t-1)}] + \varepsilon_t$$

Với biểu thức:

$u_t = \rho u_{t-1} + \varepsilon_t$, phương trình trên có thể viết lại như sau:

$$Y_t^* = \beta_1^* + \beta_2^* X_{2t}^* + \beta_3^* X_{3t}^* + \dots + \beta_k^* X_{kt}^* + \varepsilon_t \quad (7.6)$$

Trong đó:

$$Y_t^* = Y_t - \rho Y_{t-1}; \quad \beta_1^* = \beta_1(1 - \rho); \quad X_{it}^* = X_{it} - \rho X_{i(t-1)}; \quad t = 2, 3, \dots, n \text{ \& } i = 2, \dots, k$$

Quá trình biến đổi tạo ra biến Y^* và các biến X^* còn được gọi là phép lấy sai phân gần đúng, hay phép lấy sai phân tổng quát. β_1^* chỉ là số hạng hằng số mới. Lưu ý rằng số hạng sai số trong phương trình (7.6) thỏa mãn mọi tính chất cần thiết để có thể áp dụng được thủ tục bình phương tối thiểu. Nếu biết được giá trị của ρ , chúng ta có thể áp dụng phương pháp OLS cho phương trình (7.6) và giá trị ước lượng nhận được là BLUE. Tuy nhiên, giá trị ρ chưa biết nên chúng ta cần phải ước lượng chúng từ mẫu quan sát. Các bước tiến hành thủ tục Cochrane – Orcutt được trình bày như sau:

Bước 1 Ước lượng phương trình (7.4) bằng phương pháp OLS và tính toán phần dư của nó e_t .

Bước 2 Ước lượng hệ số tương quan chuỗi bậc nhất (còn gọi là $\hat{\rho}$) từ phương trình.

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

Bước 3 Biến đổi các biến như sau:

$$Y_t^* = Y_t - \rho Y_{t-1}; \quad \beta_1^* = \beta_1(1 - \rho); \quad X_{it}^* = X_{it} - \rho X_{i(t-1)}; \quad t = 2, 3, \dots, n \text{ \& } i = 2, \dots, k$$

Lưu ý rằng các biến có dấu hình sao (*) được xác định chỉ với t nhận giá trị từ 2 đến n vì có $t-1$ số hạng xuất hiện.

Bước 4 Hồi quy Y_t^* theo hằng số, theo các X_{ti}^* và tính giá trị ước lượng của phương trình (7.6) được biến đổi bằng phương pháp OLS.

Bước 5 Sử dụng những giá trị ước lượng này cho các giá trị β trong phương trình (7.4) và tính được một tập mới các giá trị ước lượng u_t . Sau đó, quay tính lặp bước 2 với những giá trị mới này cho đến khi có thể áp dụng được *quy tắc dừng* tiếp theo đây.

Bước 6 Thủ tục tính lặp trên đây có thể dừng lại khi hiệu số giá trị ước lượng của ρ từ hai kết quả liên tiếp tính được không lớn hơn giá trị chọn trước nào đó, như 0,001 chẳng hạn.

Giá trị $\hat{\rho}$ cuối cùng này sẽ được dùng để tính giá trị ước lượng CORC từ phương trình (7.6).

Vì số hạng hằng số cũng được nhân với $1 - \hat{\rho}$ nên giá trị $\hat{\beta}_1$ nhận được sẽ bằng $\beta_1^*/(1 - \hat{\rho})$, với β_1^* là số hạng hằng số ước lượng trong *phương trình biến đổi* (7.6).

7.5.4. Thủ tục tìm kiếm Hildreth – Lu

Một giải pháp thường được dùng để thay thế thủ tục Cochrane – Orcutt là **thủ tục tìm kiếm Hildreth – Lu (HILU)** (của tác giả Hildreth – Lu, 1960). Thủ tục này bao gồm các bước sau:

Bước 1 Chọn một giá trị ρ (gọi là ρ_1). Sử dụng giá trị này, biến đổi các biến và ước lượng phương trình (7.6) bằng thủ tục OLS.

Bước 2 Từ các giá trị ước lượng này, tính $\hat{\varepsilon}_t$ từ phương trình (7.6) và tính ra giá trị tổng bình phương sai số tương ứng. Gọi giá trị này là $RSS(\rho_1)$. Tiếp tục chọn một giá trị khác nữa cho ρ (gọi là ρ_2) và lặp lại bước 1 và 2.

Bước 3 Bằng cách thay đổi giá trị của ρ từ -1 đến $+1$ theo với bước nhảy có tính hệ thống nào đó (như với bước nhảy là 0,05 hoặc 0,01), chúng ta sẽ nhận được một chuỗi các giá trị $RSS(\rho)$. Hãy chọn ρ nào có giá trị RSS nhỏ nhất. Đây là giá trị ρ cuối cùng có thể cực tiểu hoá một cách bao quát tổng bình phương sai số của mô hình biến đổi.

Phương trình (7.6) ước lượng với giá trị ρ cuối cùng này là kết quả tối ưu.

So sánh hai thủ tục

Một cách căn bản, thủ tục HILU sẽ tìm kiếm các giá trị ρ nằm trong khoảng -1 và $+1$ mà khiến cho giá trị tổng bình phương phần dư của phương trình (7.6) đạt cực tiểu. Nếu khoảng cách giữa các bước nhảy là nhỏ thì thủ tục phải thực hiện rất nhiều số lần hồi quy; vì thế khi so sánh với thủ tục CORC, phương pháp HILU đòi hỏi sự hỗ trợ tính toán của máy tính rất lớn. Ngược lại, thủ tục CORC lặp lại nhiều lần để giá trị $RSS(\rho)$ đạt *cực tiểu cục bộ (local minimum)* và như vậy thủ tục có thể bỏ qua giá trị *cực tiểu toàn cục (global minimum)* nếu mô hình có nhiều hơn một điểm cực tiểu cục bộ.

Hầu hết các chương trình máy tính tốt đều cho phép áp dụng cả hai thủ tục lặp lại và tìm kiếm; đây là một giải pháp tối ưu khi dùng cả hai thủ tục nhằm bảo đảm rằng phương pháp CORC sẽ không bỏ qua giá trị cực tiểu toàn cục. Thủ tục Hildreth – Lu và thủ tục lai kết hợp chỉ thích hợp cho quá trình AR(1), và đây chính là một điểm giới hạn rất lớn. Vì lý do trên mà thủ tục tìm kiếm này không được ứng dụng phổ biến.

BÀI THỰC HÀNH

Xét dữ liệu về doanh số và tồn kho của công nghiệp Mỹ giai đoạn 1950 – 1991. Các biến được định nghĩa như sau:

SALES = Doanh số của ngành Công nghiệp Mỹ, (triệu \$)

INVENTORIES = Giá trị tồn kho của ngành Công nghiệp Mỹ, (triệu \$)

Chúng ta sẽ xem xét thực hiện hồi qui bằng một mô hình đơn giản, SALES là một hàm tuyến tính theo INVENTORIES. Thực hiện một số cách kiểm định tương quan chuỗi bậc nhất (bậc 1) và giải pháp khắc phục hiện tượng tương quan chuỗi.

YEAR	SALES	INVENTORIES	YEAR	SALES	INVENTORIES
1950	38596	59822	1971	117023	188991
1951	43356	70242	1972	131227	203227
1952	44840	72377	1973	153881	234406
1953	47987	76122	1974	178201	287144
1954	46443	73175	1975	182412	288992
1955	51694	79516	1976	204386	318345
1956	54063	87304	1977	229786	350706
1957	55879	89052	1978	260755	400929
1958	54021	87055	1979	298328	452636
1959	59729	92097	1980	328112	510124
1960	60827	94719	1981	356909	547169
1961	61159	95580	1982	348771	575486
1962	65662	101049	1983	370501	591858
1963	68995	105463	1984	411427	651527
1964	73682	111504	1985	423940	665837
1965	80283	120929	1986	431786	664654
1966	87187	136824	1987	459107	711745
1967	90918	145681	1988	496334	767387
1968	98794	156611	1989	522344	813018
1969	105812	170400	1990	540788	835985
1970	108352	178594	1991	533838	828184

Thực hiện hồi qui với dạng hàm tuyến tính của SALES theo INVENTORIES:

$$Sales = \beta_1 + \beta_2 Inventories + u$$

Ta thu được kết quả như sau:

Dependent Variable: SALES

Method: Least Squares

Date: 06/08/08 Time: 11:43

Sample: 1950 1991

Included observations: 42

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-911.6914	1165.337	-0.782342	0.4386
INVENTORIES	0.642845	0.002889	222.5282	0.0000
R-squared	0.999193	Mean dependent var		199479.4
Adjusted R-squared	0.999173	S.D. dependent var		166654.0
S.E. of regression	4793.439	Akaike info criterion		19.83433
Sum squared resid	9.19E+08	Schwarz criterion		19.91708
Log likelihood	-414.5210	F-statistic		49518.79
Durbin-Watson stat	1.374931	Prob(F-statistic)		0.000000

Vậy ta có:

$$Sales = -911.69 + 0.64 Inventories + \hat{u}$$

Thực hiện kiểm định Durbin-Watson tương quan chuỗi bậc nhất:

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Kiểm định giả thiết thống kê:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Trị thống kê Durbin-Watson: $d = DW = 1.3749$

Với $\alpha = 5\%$, $k=1$ và $n=42$, tra bảng ta có $d_U = 1.55$, $d_L = 1.46$.

Suy ra $d < d_L$, bác bỏ H_0 , có hiện tượng tương quan chuỗi bậc nhất.

Dựa vào giản đồ tương quan phần dư của phép hồi qui nói trên để xác nhận có tương quan chuỗi bậc nhất:

Kết quả tính toán được như sau:

Date: 06/08/08 Time: 11:46
Sample: 1950 1991
Included observations: 42

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
. **	. **	1 0.309	0.309	4.3012	0.038
. *	. *	2 0.165	0.077	5.5606	0.062
** .	*** .	3 -0.227	-0.331	7.9954	0.046
** .	** .	4 -0.301	-0.195	12.393	0.015
** .	* .	5 -0.306	-0.105	17.075	0.004
** .	* .	6 -0.269	-0.188	20.790	0.002
* .	. .	7 -0.082	-0.046	21.142	0.004
. .	. .	8 0.011	-0.057	21.149	0.007
. .	** .	9 -0.046	-0.280	21.266	0.012
. .	** .	10 -0.054	-0.222	21.437	0.018
. .	. .	11 0.035	-0.011	21.508	0.028
. *	. .	12 0.112	-0.036	22.283	0.034
. **	. .	13 0.199	-0.004	24.811	0.024
. *	. .	14 0.159	-0.033	26.471	0.023
. *	* .	15 0.110	-0.068	27.293	0.026
. .	. .	16 0.016	-0.025	27.312	0.038
. .	. .	17 -0.054	0.048	27.530	0.051
* .	. *	18 -0.061	0.072	27.813	0.065
* .	* .	19 -0.122	-0.072	29.017	0.066
* .	* .	20 -0.134	-0.108	30.525	0.062

Để kiểm tra tương quan chuỗi bậc nhất, ta kiểm định giả thiết:

$H_0: AC_1 = 0$

$H_1: AC_1 \neq 0$

Ta có Q-Stat = 4.3012; p-value = 0.038 < 0.05,

Suy ra bác bỏ H_0 . Có tương quan chuỗi bậc nhất ở mức 5%.

Kiểm định LM để xác định có tương quan chuỗi bậc nhất:

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	4.142795	Probability	0.048648
Obs*R-squared	4.033058	Probability	0.044617

Test Equation:

Dependent Variable: RESID

Method: Least Squares

Date: 06/08/08 Time: 11:49

Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.608282	1122.093	0.005889	0.9953
INVENTORIES	3.47E-05	0.002782	0.012486	0.9901
RESID(-1)	0.310847	0.152722	2.035386	0.0486
R-squared	0.096025	Mean dependent var	-4.95E-12	
Adjusted R-squared	0.049668	S.D. dependent var	4734.622	
S.E. of regression	4615.546	Akaike info criterion	19.78100	
Sum squared resid	8.31E+08	Schwarz criterion	19.90512	
Log likelihood	-412.4009	F-statistic	2.071398	
Durbin-Watson stat	2.047526	Prob(F-statistic)	0.139652	

Từ kết quả kiểm định LM, ta có $nR^2 = 4.033$; $p\text{-value} = 0.044617 < 0.05$. Suy ra có hiện tượng tương quan chuỗi bậc nhất.

Khắc phục:

Để khắc phục hiện tượng tương quan chuỗi bậc nhất, ta thực hiện hồi qui sau:

$$SALES_t = \beta_1 + \beta_2 INVENTORIES_t + \rho_1 u_{t-1} + \varepsilon_t$$

Và thu được kết quả:

Dependent Variable: SALES

Method: Least Squares

Date: 06/08/08 Time: 11:50

Sample(adjusted): 1951 1991

Included observations: 41 after adjusting endpoints

Convergence achieved after 5 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-975.7059	1695.130	-0.575593	0.5683
INVENTORIES	0.643007	0.004057	158.4934	0.0000
AR(1)	0.310721	0.154573	2.010185	0.0515
R-squared	0.999254	Mean dependent var		203403.4
Adjusted R-squared	0.999215	S.D. dependent var		166748.4
S.E. of regression	4672.711	Akaike info criterion		19.80722
Sum squared resid	8.30E+08	Schwarz criterion		19.93261
Log likelihood	-403.0481	F-statistic		25450.19
Durbin-Watson stat	2.044187	Prob(F-statistic)		0.000000
Inverted AR Roots	.31			

Nhìn vào kết quả hồi qui, dễ dàng thấy được hiện tượng tương quan chuỗi đã được khắc phục.

PHỤ LỤC

Bảng tra giá trị phân phối T

Bảng tra giá trị phân phối F

Bài đọc tham khảo

CuuDuongThanCong.com

BẢNG TRA GIÁ TRỊ PHÂN PHỐI T (T-Stat)

Ví dụ: Giá trị phân phối T với $\alpha = 0.05$ ($\alpha/2 = 0.025$) và n (bậc tự do) bằng 10 là:

$$T(0.025, 10) = 2.228$$

n	α	0.1	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
	$\alpha/2$	0.05	0.045	0.04	0.035	0.03	0.025	0.02	0.015	0.01	0.005
1		6.314	7.026	7.916	9.058	10.579	12.706	15.895	21.205	31.821	63.657
2		2.920	3.104	3.320	3.578	3.896	4.303	4.849	5.643	6.965	9.925
3		2.353	2.471	2.605	2.763	2.951	3.182	3.482	3.896	4.541	5.841
4		2.132	2.226	2.333	2.456	2.601	2.776	2.999	3.298	3.747	4.604
5		2.015	2.098	2.191	2.297	2.422	2.571	2.757	3.003	3.365	4.032
6		1.943	2.019	2.104	2.201	2.313	2.447	2.612	2.829	3.143	3.707
7		1.895	1.966	2.046	2.136	2.241	2.365	2.517	2.715	2.998	3.499
8		1.860	1.928	2.004	2.090	2.189	2.306	2.449	2.634	2.896	3.355
9		1.833	1.899	1.973	2.055	2.150	2.262	2.398	2.574	2.821	3.250
10		1.812	1.877	1.948	2.028	2.120	2.228	2.359	2.527	2.764	3.169
11		1.796	1.859	1.928	2.007	2.096	2.201	2.328	2.491	2.718	3.106
12		1.782	1.844	1.912	1.989	2.076	2.179	2.303	2.461	2.681	3.055
13		1.771	1.832	1.899	1.974	2.060	2.160	2.282	2.436	2.650	3.012
14		1.761	1.821	1.887	1.962	2.046	2.145	2.264	2.415	2.624	2.977
15		1.753	1.812	1.878	1.951	2.034	2.131	2.249	2.397	2.602	2.947

BẢNG TRA GIÁ TRỊ PHÂN PHỐI T (T-Stat)											
n	α	0.1	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
	$\alpha/2$	0.05	0.045	0.04	0.035	0.03	0.025	0.02	0.015	0.01	0.005
16		1.746	1.805	1.869	1.942	2.024	2.120	2.235	2.382	2.583	2.921
17		1.740	1.798	1.862	1.934	2.015	2.110	2.224	2.368	2.567	2.898
18		1.734	1.792	1.855	1.926	2.007	2.101	2.214	2.356	2.552	2.878
19		1.729	1.786	1.850	1.920	2.000	2.093	2.205	2.346	2.539	2.861
20		1.725	1.782	1.844	1.914	1.994	2.086	2.197	2.336	2.528	2.845
21		1.721	1.777	1.840	1.909	1.988	2.080	2.189	2.328	2.518	2.831
22		1.717	1.773	1.835	1.905	1.983	2.074	2.183	2.320	2.508	2.819
23		1.714	1.770	1.832	1.900	1.978	2.069	2.177	2.313	2.500	2.807
24		1.711	1.767	1.828	1.896	1.974	2.064	2.172	2.307	2.492	2.797
25		1.708	1.764	1.825	1.893	1.970	2.060	2.167	2.301	2.485	2.787
26		1.706	1.761	1.822	1.890	1.967	2.056	2.162	2.296	2.479	2.779
27		1.703	1.758	1.819	1.887	1.963	2.052	2.158	2.291	2.473	2.771
28		1.701	1.756	1.817	1.884	1.960	2.048	2.154	2.286	2.467	2.763
29		1.699	1.754	1.814	1.881	1.957	2.045	2.150	2.282	2.462	2.756
30		1.697	1.752	1.812	1.879	1.955	2.042	2.147	2.278	2.457	2.750
31		1.696	1.750	1.810	1.877	1.952	2.040	2.144	2.275	2.453	2.744
32		1.694	1.748	1.808	1.875	1.950	2.037	2.141	2.271	2.449	2.738
33		1.692	1.747	1.806	1.873	1.948	2.035	2.138	2.268	2.445	2.733
34		1.691	1.745	1.805	1.871	1.946	2.032	2.136	2.265	2.441	2.728

BẢNG TRA GIÁ TRỊ PHÂN PHỐI T (T-Stat)											
n	α	0.1	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
	$\alpha/2$	0.05	0.045	0.04	0.035	0.03	0.025	0.02	0.015	0.01	0.005
35		1.690	1.744	1.803	1.869	1.944	2.030	2.133	2.262	2.438	2.724
36		1.688	1.742	1.802	1.867	1.942	2.028	2.131	2.260	2.434	2.719
37		1.687	1.741	1.800	1.866	1.940	2.026	2.129	2.257	2.431	2.715
38		1.686	1.740	1.799	1.864	1.939	2.024	2.127	2.255	2.429	2.712
39		1.685	1.739	1.798	1.863	1.937	2.023	2.125	2.252	2.426	2.708
40		1.684	1.737	1.796	1.862	1.936	2.021	2.123	2.250	2.423	2.704
50		1.676	1.729	1.787	1.852	1.924	2.009	2.109	2.234	2.403	2.678
60		1.671	1.723	1.781	1.845	1.917	2.000	2.099	2.223	2.390	2.660
70		1.667	1.719	1.776	1.840	1.912	1.994	2.093	2.215	2.381	2.648
80		1.664	1.716	1.773	1.836	1.908	1.990	2.088	2.209	2.374	2.639
90		1.662	1.714	1.771	1.834	1.905	1.987	2.084	2.205	2.368	2.632
100		1.660	1.712	1.769	1.832	1.902	1.984	2.081	2.201	2.364	2.626

BẢNG TRA GIÁ TRỊ PHÂN PHỐI F-VỚI $\alpha = 0.05$

Ví dụ: Giá trị phân phối F với $\alpha = 0.05$ và n_1 (bậc tự do của tử số) bằng 3; n_2 (bậc tự do của mẫu số) bằng 10 là:
 $F(0.05, 3, 10) = 3.708$

n_1/n_2	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494

BẢNG TRA GIÁ TRỊ PHẦN PHỐI F-VỚI $\alpha = 0.05$

n1/n2	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323	2.255	2.199	2.153
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189	2.142
33	4.139	3.285	2.892	2.659	2.503	2.389	2.303	2.235	2.179	2.133
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170	2.123
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161	2.114

BẢNG TRA GIÁ TRỊ PHÂN PHỐI F-VỚI $\alpha = 0.05$

n1/n2	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153	2.106
37	4.105	3.252	2.859	2.626	2.470	2.356	2.270	2.201	2.145	2.098
38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138	2.091
39	4.091	3.238	2.845	2.612	2.456	2.342	2.255	2.187	2.131	2.084
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927

BÀI ĐỌC THAM KHẢO 1

PHÂN TÍCH HỒI QUI TRONG THƯƠNG MẠI

QUẢNG TÂY – ASEAN VÀ TĂNG TRƯỞNG GDP QUẢNG TÂY

I. Lựa chọn số liệu mẫu và tạo mô hình

Số liệu năm về mức giao dịch xuất nhập khẩu Quảng Tây – Asean (X) và GDP Quảng Tây (Y) từ năm 1998 đến năm 2004, vẽ biểu đồ rời, sau đó tính toán các con số liên quan và tiến hành phân tích hồi quy.

Chú thích: Năm 2004, thể chế ngoại hối có nhiều cải cách lớn, tỷ giá hối đoái đồng NDT đã được hợp nhất và dần ổn định, liên tục giữ ở mức tỷ lệ từ 1/8,3124 đến 1/8,2770 (USD/NDT). Số liệu GDP tính theo USD trong bảng sau được căn cứ vào tỷ giá hối đoái trung bình năm từ 1998 – 2004 để quy từ GDP Quảng Tây tính bằng NDT sang GDP Quảng Tây tính bằng USD.

Bảng 1 Số liệu về thương mại Quảng Tây – Asean và GDP Quảng Tây

Đơn vị: tỷ USD

Năm	Tổng mức XNK	XK	NK	GDP(tỷ NDT)	GDP(USD)	Tỷ giá bình quân năm
1998	0,39659	0,31861	0,07798	190,304	22,98607	8,2791
1999	0,36883	0,28872	0,08011	195,327	23,59506	8,2783
2000	0,43950	0,31011	0,12939	205,015	24,76505	8,2784
2001	0,41906	0,25914	0,15991	223,119	26,95651	8,2770
2002	0,62726	0,44238	0,18488	245,536	29,84514	8,2770
2003	0,82619	0,55235	0,21384	273,513	33,04494	8,2770
2004	1,00100	0,63600	0,36500	332,010	40,11188	8,2771

Bảng 2 Hệ số tương quan

	Y	X
Y	1,000000	0,976391
X	0,976391	1,000000

Từ Bảng 2 có thể thấy, hệ số tương quan giữa mức thương mại Quảng Tây và GDP Quảng Tây năm 1998 – 2004 là 0,976391, là một mức tương quan rất cao, thương mại XNK Quảng Tây – Asean có vai trò quyết định trong thúc đẩy tăng trưởng kinh tế Quảng Tây, nói một cách khác tăng trưởng GDP Quảng Tây phụ thuộc rất nhiều vào thương mại XNK Quảng Tây – Asean, giữa chúng có thể lập ra một mô hình tuyến tính.

II. Lập mô hình và tiến hành phân tích hồi quy

Vì chỉ phân tích tính tương quan giữa thương mại Quảng Tây – Asean và GDP Quảng Tây, có thể giả định là các nhân tố khác ảnh hưởng đến tăng trưởng GDP đều bình ổn, do đó có thể sử dụng mô hình hồi quy tuyến tính đơn để phân tích. Dùng Y biểu thị GDP Quảng Tây, dùng X biểu thị mức thương mại Quảng Tây – Asean, lập một mô hình hồi quy tuyến tính đơn.

$$Y = \alpha + \beta X + u$$

Căn cứ vào các số liệu của Bảng 1 từ năm 1998-2004, vận dụng phần mềm kinh tế số lượng Eviews, sử dụng phương pháp nhị thừa nhỏ nhất (OLS), tiến hành phân tích hồi quy đối với X và Y, ta được kết quả như sau:

Từ những kết quả đạt được từ phần mềm Eviews, ta có phương trình hồi quy tương ứng như sau:

$$\begin{aligned} \hat{Y} = & 14,53003 + 24,41980X \\ & (9,612073) \quad (10,10718) \\ & (0,0002) \quad (0,0002) \end{aligned}$$

$$R^2 = 0,953339 \quad F = 102,1550 \quad \text{Prob}(F - \text{statistic}) = 0,000162$$

Kiểm nghiệm: kiểm nghiệm ý nghĩa kinh tế, thông qua quan sát ký hiệu và trị số của hệ số lượng đánh giá tham số, có thể thấy ý nghĩa kinh tế phù hợp của mô hình này. Hệ số của X là 24,41980 cho thấy, trong năm 1998 – 2004, thương mại Quảng Tây – Asean hàng năm tăng thêm 0,1 tỷ USD, và kết quả là GDP khu vực tăng thêm khoảng 2,441890 tỷ USD.

Thống kê kiểm nghiệm: $R^2 = 0,953339$ cho thấy, xác suất đi kèm của kiểm nghiệm T đối với hệ số hạng thường số và hệ số hồi quy lần lượt là 0,0002 và 0,0002, cho thấy chúng đều có thể thông qua được kiểm nghiệm với xác suất dưới 5%, từ đó phủ nhận giả thuyết ban đầu là hệ số = 0.

Từ $F = 102,1550$ Prob(F – statistic) = 0,000162 có thể thấy, không cần tra Bảng cũng có thể biết chắc phương trình thông qua, có nghĩa là tính rõ ràng của phương trình rất tốt, nói lên tính tương quan cao giữa 2 biến lượng tăng trưởng GDP Quảng Tây và mức thương mại Quảng Tây – Asean.

Từ kiểm nghiệm trên có thể rút ra, mặc dù có rất nhiều nhân tố ảnh hưởng đến tăng trưởng GDP Quảng Tây, nhưng trong đó mức thương mại XNK Quảng Tây – Asean có vai trò quyết định thúc đẩy tăng trưởng kinh tế Quảng Tây.

Mô hình kinh tế lượng ứng dụng trong nghiên cứu thị trường du lịch Việt Nam

Ngành Du lịch Việt Nam đã và đang được Đảng và Nhà nước ta lựa chọn phát triển và xây dựng là ngành kinh tế mũi nhọn. Du lịch Việt Nam nói chung và du lịch quốc tế đến Việt Nam nói riêng, đóng vai trò quan trọng trong việc tạo ra ngân sách cho ngành du lịch Việt Nam. Nếu từ năm 1990 du lịch quốc tế đã đóng góp cho ngành khoảng 109 triệu đô la, thì năm 2002 du lịch quốc tế đã đóng góp cho ngành khoảng 1,017 tỷ đô la (tương đương 5,3 triệu tấn gạo xuất khẩu). Một công việc do ngành du lịch trực tiếp tạo ra, có thể tạo ra được từ (1-3) công việc gián tiếp khác cho các ngành sản xuất trong nước: ngành công nghiệp, nông nghiệp, giao thông vận tải và các ngành dịch vụ khác. Như vậy nếu đẩy mạnh phát triển được thị trường du lịch Việt Nam, sẽ tạo được nhiều cơ hội cho phát triển kinh tế của các vùng và địa phương có tài nguyên du lịch, mặc dù nơi đó có thể rất xa xôi hẻo lánh. Trên thế giới cũng như thực tiễn ở Việt Nam cho thấy nhiều nơi đã phát triển đi lên từ du lịch.

Để phát triển ngành du lịch Việt Nam thực sự trở thành ngành kinh tế mũi nhọn, một trong những yêu cầu trước tiên là phải hiểu biết và nắm vững được thị trường du lịch trong nước cũng như thị trường du lịch quốc tế. Với tư cách là những nhà cung ứng, ngành du lịch Việt Nam quảng bá các yếu tố hấp dẫn tới khách du lịch trong và ngoài nước, và cung ứng các hàng hoá dịch vụ cho khách du lịch. Cầu về hàng hoá dịch vụ du lịch, có thể là khách trong nước hay ngoài nước. Với điều kiện và vị thế của Việt Nam, thu hút nguồn khách quốc tế đến Việt Nam, có vai trò quan trọng trong chiến lược phát triển du lịch, và tạo cơ hội cho thị trường du lịch Việt Nam phát triển.

Vấn đề đặt ra cần phải có cơ sở khoa học để hiểu biết và nắm được thị trường khách quốc tế đến Việt Nam, có như vậy mới có những quyết định đúng đắn. Trong quản lý kinh tế hiện đại, ngoài sự hiểu biết về mặt định tính các yếu tố và các mối quan hệ của thị trường, người ta còn cần định lượng được các yếu tố và các mối quan hệ giữa các yếu tố.

Để hiểu biết được các yếu tố ảnh hưởng và định lượng được những yếu tố ảnh hưởng tới lượng khách du lịch quốc tế, hiện nay người ta thường sử dụng mô hình kinh tế lượng. Một khi đã xây dựng được mô hình kinh tế lượng, việc tiến hành dự báo thị trường như lượng cầu, xác định độ co giãn cầu v.v..., hoặc cần ra quyết định trong những tình huống với mức tin cậy nhất định, thì mô hình kinh tế lượng tỏ ra có ưu thế.

Thông thường, người ta hay tiến hành dự báo qui mô thị trường. Tổng cầu thị trường được biểu thị dưới dạng số lượt khách du lịch quốc tế sẽ đến thăm từ một nước xuất phát tới một điểm đến nước ngoài, hoặc bằng khoản tiêu dùng du lịch của khách tham quan từ nước xuất phát đến đất nước du lịch.

Khi tiến hành dự báo qui mô thị trường bằng mô hình kinh tế lượng, người ta sử dụng phân tích hồi qui để ước lượng mối quan hệ giữa biến dự báo gọi là biến phụ thuộc (hay biến được giải thích), với một hay nhiều biến độc lập (hoặc biến giải thích) khác. Ước lượng tiến hành dựa vào các dữ liệu đã có trước, giá trị dự báo tương lai là giá trị ngẫu nhiên xác định trên cơ sở sử dụng mô hình hồi qui đã xây dựng.

Bước đầu tiên trong việc xây dựng mô hình dự báo qui mô thị trường là cần xác định các biến độc lập có thể gây ảnh hưởng tới cầu du lịch quốc tế.

Giả sử hàm cầu thị trường có dạng:

$$Y = f(X_1, X_2, \dots, X_k) \quad (1)$$

Trong đó:

Y : Cầu đối với du lịch quốc tế từ nơi xuất phát cụ thể tới điểm đến khảo sát.

$X_1, X_2, \dots, X_k$: Là các biến độc lập hay các biến giải thích.

Biến dự báo:

Cầu du lịch được đo bằng lượng khách du lịch tới thăm hoặc giá trị tiêu dùng du lịch, mà được xem xét dưới dạng tiền.

Thu nhập:

Thu nhập thực tế tính bằng tiền, tại đất nước du khách sinh sống, thường được sử dụng làm biến giải thích ở mô hình (1). Nếu chuyển đi với mục đích nghỉ ngơi, thăm hỏi gia đình bạn bè thì sử dụng là dạng thu nhập có thể sử dụng được; còn nếu chuyển đi với mục đích kinh doanh, thì sử dụng dạng thu nhập chung (như thu nhập quốc gia chẳng hạn).

Giá cả hàng hoá dịch vụ:

Yếu tố giá cả cũng được thường xuyên đưa vào làm biến giải thích trong hàm cầu. Với du lịch quốc tế, cần lưu ý có hai yếu tố giá cả - liên quan đến các chi phí vận chuyển để tới được điểm đến và các chi phí khi ở tại điểm đến. Các chi phí vận chuyển có thể được tính qua giá vé máy bay, hoặc cước phí vận chuyển đường bộ từ điểm xuất phát của du khách tới điểm đến du lịch.

Có thể tính chi phí du lịch tại điểm đến theo mức chi phí riêng cho khách du lịch. Ngoài ra, chỉ số giá tiêu dùng ở điểm đến có khả năng sử dụng thay thế biến chi phí du lịch. Tỷ giá hối đoái cũng có thể được đưa vào như một biến giải thích trong mô hình (1).

Giá hàng hoá dịch vụ thay thế:

Kinh tế học đã chứng tỏ rằng hàng hóa và dịch vụ thay thế có thể đóng vai trò quan trọng để xác định cầu. Khách du lịch tiềm năng có thể so sánh giá cả của những kỳ nghỉ ở nước ngoài với các giá kỳ nghỉ ở trong nước, trước khi họ quyết định mua. Tất nhiên, họ cũng so sánh các chi phí kỳ nghỉ ở các nước khác nhau. Như thế, chi phí lựa hành thay thế và chi phí tiêu dùng du lịch có thể đóng vai trò quan trọng để xác định cầu du lịch quốc tế tới điểm đến đã cho từ một nơi xuất phát cụ thể. Giá thay thế có thể đưa vào mô hình (1) bao gồm: trung bình trọng số chi phí phương tiện vận tải thay thế và trung bình trọng số chi phí tiêu dùng du lịch. Các trọng số phải phản ánh được sự liên quan của nó với tính hấp dẫn của điểm đến, tới các khách du lịch một cách rõ ràng và thường dựa trên cơ sở thị phần có trước đó.

Hoạt động xúc tiến:

Các tổ chức du lịch quốc gia thường tiêu dùng khá nhiều tiền cho việc xúc tiến du lịch ở nước ngoài, nhằm thu hút khách tới đất nước mình. Do vậy, việc chi phí cho xúc tiến du lịch mong muốn đóng vai trò quan trọng trong việc xác định mức cầu du lịch quốc tế, vì thế chi phí xúc tiến du lịch cần được đưa vào như một biến giải thích trong hàm cầu (1).

Các biến giả:

Vào những năm 1973 và 1979, thế giới lâm vào cuộc khủng hoảng dầu lửa có thể làm cầu du lịch quốc tế suy giảm tạm thời. Hoặc đe dọa do chiến tranh hay dịch bệnh xảy ra, cũng làm suy giảm tạm thời cầu du lịch quốc tế. Sử dụng các biến giả đưa vào trong mô hình kinh tế lượng, nhằm xem xét một sự kiện nào đó tác động một cách tạm thời cầu du lịch quốc tế.

Biến trễ phụ thuộc:

Nhằm phản ánh những thói quen sự trung thành của cầu và sự chậm trễ cứng nhắc của cung, người ta có thể đưa vào các biến trễ phụ thuộc. Chẳng hạn đã có lần khách du lịch đến thăm một điểm đến nào đó mà họ cảm thấy thích thú và hài lòng, họ sẽ có xu hướng thích quay trở lại đó. Ngoài ra, sự hài lòng và hiểu biết về điểm đến sẽ tiếp tục được quảng bá tới những du khách tiềm năng hác thông qua các câu chuyện hấp dẫn và những bức ảnh đã được ghi lại, từ đó sẽ làm giảm bớt sự rủi ro khi họ lựa chọn điểm đến. Những điều đó dường như làm kéo dài sự trung thành với điểm đến.

Ngoài ra, sự cung cấp các dịch vụ cũng khó có thể gia tăng nhanh chóng trong thời gian ngắn. Tất cả các vấn đề trên, sẽ được thể hiện thông qua sự có mặt của biến trễ phụ thuộc trong mô hình (1).

Dạng mô hình toán thông dụng nhất thường được lựa chọn để biểu diễn hàm cầu du lịch quốc tế là dạng hàm loga tuyến tính:

$$Y = aX_1^{b_1} X_2^{b_2} \dots X_k^{b_k} e^u$$

Trong đó :Y là cầu đối với du lịch quốc tế, là biến phụ thuộc.

$X_1, \dots, X_k$ là các biến giải thích, hay biến độc lập. k

u là sai số ngẫu nhiên.

a, $b_1, \dots, b_k$ là các tham số.

e = 2,781 là số lôgarit tự nhiên.

Để có thể ước lượng được biểu thức (2), tiến hành loga hoá biểu thức (2):

$$\ln Y = \ln a + b_1 \ln X_1 + b_2 \ln X_2 + \dots + b_k \ln X_k + u \quad (3)$$

Với biểu thức (3), ta có thể ước lượng theo mô hình toán này thông qua các phần mềm Mfit, hay Eviews.

Qua (3), cũng dễ dàng xác định được độ co giãn của cầu theo các biến. Thật vậy, lấy đạo hàm hai vế biểu thức (2) :

Nhưng độ co giãn của Y với X được xác định là :

Thay thế biểu thức (4) vào biểu thức (5) ta được :

Từ đó ta thấy b_1 chính là hệ số co giãn của Y theo X_1 .

Một cách tổng quát b_j là hệ số co giãn của Y theo X_j , $j = 1, 2, \dots, k$.

Xét cụ thể việc xây dựng một mô hình kinh tế lượng nhằm đánh giá cầu du lịch quốc tế ở thị trường trọng điểm tới điểm đến du lịch Việt Nam.

Căn cứ vào các nguồn dữ liệu từ asean centre, nguồn của Tổng cục du lịch, tham khảo ý kiến của các chuyên gia du lịch. Chúng tôi đã tiến hành xây dựng mô hình hàm cầu

khách du lịch Nhật Bản mà ngành du lịch Việt Nam đã và đang xúc tiến thu hút cho điểm đến du lịch Việt Nam.

Qua quá trình nghiên cứu, chúng tôi thấy nguồn du khách Nhật Bản là nguồn khách khá hấp dẫn đối với các nước Asean, trong đó Thái Lan và Singapo là hai nước dẫn đầu về thu hút khách Nhật. Năm 2001, khách Nhật Bản tới Asean là trên 16,2 triệu. Thái Lan đón 1,18 triệu khách chiếm 7,2%, còn Singapo đón 0,755 triệu khách chiếm 4,6%, Việt Nam đón 0,204 triệu khách chiếm 1,2%. Phải chăng do giá tour đến Việt Nam cao hơn các nước trong khu vực, và đầu tư cho chính sách xúc tiến của Việt Nam còn quá yếu, nên lượng khách Nhật vào Việt Nam còn rất nhỏ bé so với tiềm năng. Để có thể kiểm định những phán đoán trên, chúng tôi đã lựa chọn dạng mô hình toán cho hàm cầu khách Nhật Bản tới Việt Nam có dạng dưới đây :

$$Y = a \cdot PVN^{b_1} \cdot XT^{b_2} \cdot e^u \quad (7)$$

Trong đó :

Y : Số lượng khách du lịch trong năm từ thị trường Nhật Bản tới điểm đến Việt Nam.

PVN: Mức giá tour trọn gói trung bình của kỳ nghỉ trong năm từ Nhật Bản tới Việt Nam.

XT: Ngân sách xúc tiến trong năm của ngành du lịch Việt Nam tới thị trường Nhật Bản.

$e = 2,781$

u : Sai số ngẫu nhiên.

Tiến hành ln hai vế biểu thức (7) ta có :

$$\ln Y = \ln a + b_1 \ln PVN + b_2 \ln XT + u \quad (8)$$

Với các dữ liệu trong bảng 1, bằng sự hỗ trợ của phần mềm tính toán Eviews 3.1, kết quả ước lượng có được các tham số của mô hình (8) như sau, xem bảng 2 :

$$\ln Y = 14,67446 - 1,785864 \cdot \ln PVN + 0,873611 \cdot \ln XT + u \quad (9)$$

Mô hình (9) là mô hình đã được lựa chọn "tốt nhất", các hệ số ước lượng là tin cậy và phù hợp, giá trị R^2 là khá cao ($R^2 = 0,999$) chứng tỏ quan hệ giữa các biến là liên quan chặt, hai biến độc lập PVN và XT đã giải thích được tới 99,9% giá trị của biến phụ thuộc Y. Mô hình cũng không có những khuyết tật.

Sử dụng mô hình (9) ta có thể xác định được lượng cầu Y phụ thuộc vào những yếu tố nào. Khi tiến hành dự báo lượng khách từ thị trường Nhật Bản đến Việt Nam, chỉ cần có được những thông tin trong năm dự báo: Mức giá trung bình của các tour trọn gói từ thị trường Nhật Bản tới Việt Nam, ngân sách xúc tiến du lịch của ngành du lịch Việt Nam đầu tư cho thị trường Nhật Bản.

Ngoài ra mô hình (9), còn cho ta biết các thông tin về các loại hệ số: co giãn cầu theo giá ($b_1=1,785864$), co giãn cầu theo xúc tiến du lịch ($b_2=0,873611$). ý nghĩa của hệ số co giãn cầu b_1 là: khi giá tour trọn gói tăng 1%, thì lượng cầu sẽ giảm 1,785864%. Cũng tương tự, với hệ số co giãn cầu theo xúc tiến b_2 , khi tăng ngân sách xúc tiến lên 1%, thì lượng cầu sẽ tăng 0,873611%.

Để xác định tính chính xác của mô hình (9), mô hình cần được tiếp tục kiểm nghiệm qua thực tiễn, thông qua liên tiếp các bước lặp, ta sẽ có được một mô hình đúng đắn và hoàn toàn tin cậy.

Trên đây là những nghiên cứu bước đầu về ứng dụng mô hình kinh tế lượng trong nghiên cứu thị trường du lịch của Việt Nam. Mô hình đã xác định được các yếu tố ảnh hưởng chính tới lượng cầu du lịch tới Việt Nam của thị trường du khách Nhật Bản. Mô hình cũng định lượng được mối quan hệ của các yếu tố ảnh hưởng đó. các nhà quản trị có thể dựa vào mô hình để ra các quyết định cần thiết. Hy vọng, việc ứng dụng mô hình kinh tế lượng sẽ được phổ biến rộng rãi với công tác nghiên cứu thị trường của Việt Nam nói chung và của ngành du lịch nói riêng.

Bảng 1: Số lượt khách Nhật đến Việt Nam, giá tour trọn gói, Ngân sách xúc tiến (Nguồn Asean Centre, Tổng cục du lịch Việt Nam)

Năm	Y (Lượt khách)	PVN (Đô la)	XT (Đô la)
1998	95258.00	1200.000	50000.00
1999	113514.0	1190.000	60000.00
2000	152725.0	1160.000	80000.00
2001	204860.0	1149.000	110000.0
2002	275000.0	1135.000	150000.0

Bảng 2 : Bản báo cáo của chương trình Eview

Dependent Variable: LY				
Method: Least Squares				
Sample: 1998 2002				
Included observations: 5				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
LXT	0.873611	0.006248	139.8119	0.0001
LPVN	-1.785864	0.118404	-15.08278	0.0044
C	14.67446	0.905715	16.20207	0.0038
R - squared	0.999997	Mean dependent var		11.95901
Adjusted R- squared	0.999994	S.D. dependent var		0.430251
S.E. of regression	0.001067	Akaike info criterion		-10.56371
Sum squared resid	2.28E-06	Schwarz criterion		-10.79804
Log likelihood	29.40927	F-statistic		325028.0
Durbin - Watson stat	2.713355	Prob(F-statistic)		0.000003