TRƯỜNG ĐẠI HỌC TÀI CHÍNH MARKETING
KHOA CƠ BẢN
BỘ MÔN TOÁN – THỐNG KÊ
TRẦN KIM THANH (CHỦ BIÊN)
NGUYỄN VĂN PHONG
NGUYỄN TRUNG ĐÔNG

BÀI GIẢNG

KINH TẾ LƯỢNG

MÃ SỐ: CS – K21 – 13

(LƯU HÀNH NỘI BỘ)

TP. HỒ CHÍ MINH – 2015

LỜI NÓI ĐẦU

Kinh tế lượng là môn học được đưa vào giảng dạy cho các lớp sinh viên thuộc hầu hết các chuyên ngành của trường Đại học Tài chính – Marketing. Vì vậy, một tài liệu được biên soạn thống nhất theo đề cương môn học là nhu cầu cần thiết cho các giảng viên và sinh viên. Để đáp ứng nhu cầu đó, được sự đồng ý của Bộ môn Toán – Thống kê, của Khoa Cơ Bản và của Ban Giám hiệu trường Đại học Tài chính – Marketing, chúng tôi biên soạn cuốn Bài giảng này.

Tài liệu này trình bày những nội dung cơ bản, dựa trên đề cương học phần Kinh tế lượng của Bộ môn Toán – Thống kê, sử dụng Eviews 8 làm phần mêm hỗ trợ và được chia làm 7 chương và 4 bảng phụ lục thống kê:

- *Chương 1*: Trình bày tổng quan về kinh tế lượng, những khái niệm liên quan đến số liệu, hàm hồi quy tổng thể, hàm hồi quy mẫu, mô hình kinh tế lượng.
- *Chương 2*: Trình bày mô hình hồi quy hai biến, mô hình hồi quy đơn giản nhất, tuy ít hiện hữu, nhưng rất quan trọng về mặt phương pháp luận. Trong đó trình bày chi tiết phương pháp bình phương tối thiểu để ước lượng các hệ số hồi quy, cùng những bài toán thống kê cơ bản trên mô hình hồi quy hai biến: Ước lượng khoảng tin cậy cho các tham số mô hình; Kiểm định giả thuyết về mô hình. Phần cuối chương trình bày một số ứng dụng của mô hình hai biến và một số mô hình tuyến tính hóa được thường gặp trong thực tế.
- *Chương 3*: Khảo sát mô hình hồi quy nhiều biến, trong đó trình bày phương pháp bình phương tối thiểu để tìm ước lượng cho các hệ số hồi quy, khảo sát hệ số xác định hiệu chỉnh, ma trận trương quan mẫu, các bài toán thống kê trên mô hình hồi quy nhiều biến: Ước lượng các tham số, kiểm định giả thuyết về mô hình, một kiểm định thường được ứng dụng nhiều trong mô hình hồi quy nhiều biến là kiểm định Wald.
- *Chương 4*: Biến giả trong phân tích hồi quy. Chương này đề cập đến việc lượng hóa biến định tính (biến giả) để đưa vào mô hình hồi quy và sự cần thiết phải sử dụng biến giả, đồng thời giới thiệu kỹ thuật sử dụng biến giả để xử lý các vấn đề trong mô hình hồi quy.
- *Chương 5*: Đề cập đến những vấn đề thực tế có thể xảy ra trong một mô hình hồi quy, mà chúng vi phạm giả thiết của phương pháp bình phương tối thiểu thông dụng, một phương pháp được sử dụng trong kinh tế lượng để ước lượng mô hình hồi quy tổng thể. Đó là các vấn đề: Đa cộng tuyến giữa các biến giải thích; Phương sai nhiễu thay đổi; Tự tương quan của nhiễu. Mỗi vấn đề này đều được đề cập với ba nội dung: Phân tích nguyên nhân; Cách phát hiện (thông qua các biểu hiện của mô hình, của đồ thị và qua các kiểm định); Biện pháp khắc phục.
- *Chương 6*: Phân tích đặc trưng và lựa chọn mô hình, Chương này trình bày những vấn đề chính sau đây: Phân tích đặc trưng mô hình (Các thuộc tính của một mô hình tốt, các loại sai lầm chỉ định, cách tiếp cận để lựa chọn mô hình); Các kiểm định về sai lầm chỉ định; Ứng dụng hồi quy trong phân tích, dự báo.

- *Chương phụ lục*: Trình bày có tính chất hệ thống lại những vấn đề của Lý thuyết Xác suất – Thống kê toán, cần thiết cho việc phân tích và giải quyết các bài toán trên mô hình hồi quy của Kinh tế lượng, tạo cơ sơ nền tảng cho người học để nắm bắt tốt hơn nội dung bài giảng.

Cuốn tài liệu này do TS. Trần Kim Thanh làm chủ biên và được biên soạn dựa trên cơ sở đề cương chi tiết của Bộ môn Toán - Thống kê, tổng hợp các bài giảng Kinh tế lượng qua nhiều năm giảng dạy, các tài liệu tham khảo và các ý kiến đóng góp quý báu của các giảng viên Bộ môn Toán - Thống kê và các đồng nghiệp. Nội dung của tài liệu được biên soạn phù hợp với thời lượng 3 tín chỉ, bao gồm 7 chương. Trong đó có những nội dung được trình bày trên lớp, có những nội dung yêu cầu sinh viên tự đọc với sự hướng dẫn của giáo viên. Sau mỗi chương đều có phần bài tập để người học thực hành, kèm theo những hướng dẫn cần thiết.

Nhóm biên soạn tài liệu này gồm :

- TS. Trần Kim Thanh (Chủ biên), biên soạn phần lý thuyết các chương.
- ThS. Nguyễn Văn Phong, sưu tầm và biên soạn phần bài tập cho các chương 1,2, 3, 4.
- ThS. Nguyễn Trung Đông, sưu tầm và biên soạn phần bài tập cho các chương 5, 6, đọc và chỉnh sửa bản in.

Chúng tôi xin chân thành cảm ơn Bộ môn Toán – Thống kê và các đồng nghiệp đã đóng góp những ý kiến quý báu cho cuốn Bài giảng này.

Hy vọng đây sẽ là một tài liệu đáp ứng được yêu cầu về giảng dạy và học tập đối với học phần Kinh tế lượng trong nhà trường.

Nhóm biên soạn đã hết sức cố gắng để hoàn thành cuốn tài liệu này, tuy nhiên không tránh khỏi những thiếu sót. Chúng tôi mong được sự đóng góp ý kiến của các đồng nghiệp và bạn đọc tài liệu này ngày càng hoàn thiện hơn.

Nhóm tác giả

Chương 1. TỔNG QUAN VỀ KINH TẾ LƯỢNG

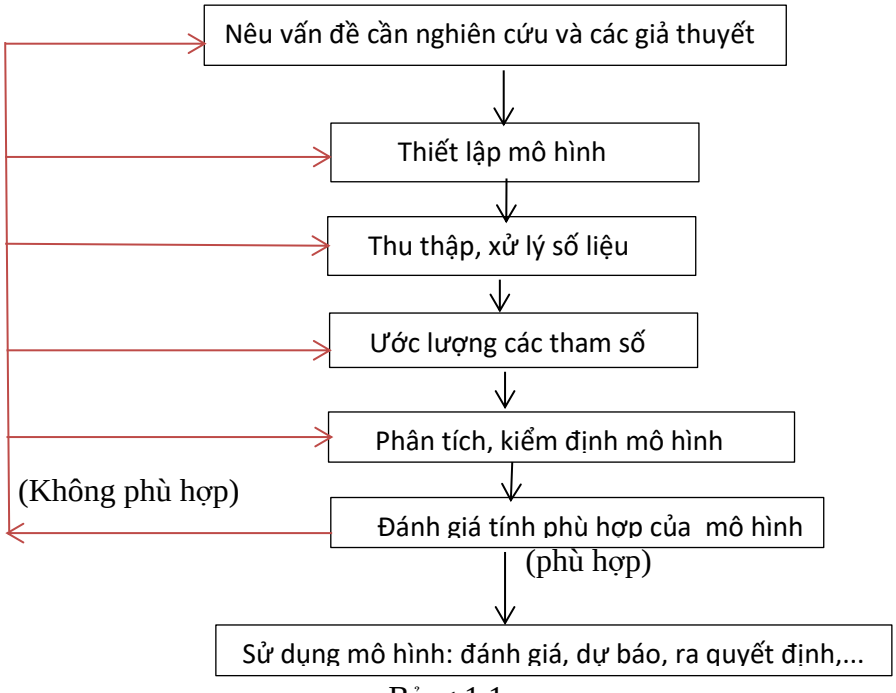
Chương này trình bày tổng quan về kinh tế lượng: Khái niệm về kinh tế lượng; mô hình kinh tế lượng, trong đó đặc biệt là các khái niệm về hàm hồi quy tổng thể, hàm hồi quy mẫu; các vấn đề cơ bản của kinh tế lượng.

1.1. Các khái niệm mở đầu

1.1.1. Khái niệm về kinh tế lượng

Kinh tế lượng, theo thuật ngữ tiếng Anh là *Econometrics*. Theo đó hiểu một cách đơn giản thì kinh tế lượng là đo lường kinh tế. Một cách đầy đủ và chi tiết hơn thì Kinh tế lượng là khoa học nghiên cứu những vấn đề thực nghiệm của các quy luật kinh tế, là sự kết hợp chặt chẽ giữa các số liệu thực tế, lý thuyết kinh tế và công cụ toán học không thể thiếu được là lý thuyết Xác suất Thống kê kết hợp với các phần mềm vi tính hỗ trợ, nhằm lượng hóa các quy luật kinh tế nói riêng và thực tiễn nói chung thông qua những mô hình toán học phù hợp với thực tế, ước lượng các tham số, phân tích, đánh giá và dự báo các chỉ tiêu kinh tế, xã hội. Kinh tế lượng vì thế còn được áp dụng trong các lĩnh vực khoa học kỹ thuật, môi trường, dân số, giáo dục, v.v....

1.1.2. Sơ đồ tổng quan về kinh tế lượng: Có thể hình dung một cách tổng quan về quá trình xây dựng và ứng dụng của kinh tế lượng qua sơ đồ sau:



Bảng 1.1

a. Vấn đề nghiên cứu và các giả thuyết

Vấn đề nghiên cứu có thể dựa trên cơ sở lý thuyết kinh tế, kinh nghiệm thực tế, kết quả của những nghiên cứu trước đó. Từ đó cần phải xác định được các biến kinh tế và mối quan hệ giữa chúng. Mối quan hệ này là sự phụ thuộc giữa một biến, gọi là biến phụ thuộc hay biến cần được giải thích, đối với các biến còn lại, gọi là các biến giải thích (có nhiều tài liệu còn gọi là c+ác biến độc lập, nhưng một cách chính xác, ta nên gọi là các biến giải thích)

Chẳng hạn, lý thuyết kinh tế chỉ ra rằng: *Chỉ tiêu tiêu dùng tăng khi thu nhập tăng nhưng sự gia tăng trong tiêu dùng không nhiều như sự gia tăng trong thu nhập*. Trên cơ sở này, ta xác định được hai biến kinh tế cần khảo sát là *Thu nhập* và *Tiêu dùng*, trong đó Tiêu dùng sẽ phụ thuộc vào Thu nhập và vấn đề cần nghiên cứu ở đây là: Khi thu nhập thay đổi 1 đơn vị thì tiêu dùng sẽ thay đổi một lượng là bao nhiêu?

b. Thiết lập mô hình kinh tế lượng

Lý thuyết kinh tế cho biết quy luật về mối quan hệ giữa các biến kinh tế một cách định tính, nhưng không lượng hóa được mối quan hệ này, tức là không nêu cụ thể dạng hàm biểu diễn mối quan hệ đó. Trên cơ sở các học thuyết kinh tế, sử dụng công cụ toán học, kinh tế lượng sẽ định dạng các mô hình cho các trường hợp cụ thể, tức là thiết lập mô hình kinh tế lượng.

Trong lý thuyết xác suất, ta biết hàm hồi quy:

$$E(Y|X) = f(X) \qquad (1)$$

là mô hình toán học mô tả sự phụ thuộc của giá trị trung bình (có điều kiện) của biến quan sát Y vào biến quan sát X.

Tuy nhiên do tác động ngẫu nhiên mà các giá trị của biến Y thường lệch khỏi giá trị trung bình (quan hệ giữa Y và X là quan hệ phụ thuộc thống kê), nên độ lệch:

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

là một biến ngẫu nhiên. Vì thế: $Y = E(Y|X) + U$

Và mô hình sau đây được gọi là mô hình kinh tế lượng:

$$\begin{cases} E(Y|X) = f(X) \\ Y = f(X) + U \end{cases} \qquad (2)$$

Trong đó số hạng U, gọi là số hạng nhiễu, là một biến ngẫu nhiên (nên còn gọi là sai số ngẫu nhiên), đại diện cho các tác động ngẫu nhiên của các yếu tố khác ngoài X. Chẳng hạn nếu X là thu nhập, Y là tiêu dùng thì U đại diện cho tác động của các yếu tố ngẫu nhiên khác ngoài thu nhập, như: hoàn cảnh gia đình, sở thích, tập quán tiêu dùng,…ảnh hưởng đến việc tiêu dùng.

c. Thu thập, xử lý số liệu

Trong mô hình kinh tế lượng được xác lập, tức là đã xác lập được dạng của hàm hồi quy f(X), có các tham số chưa biết mà ta cần ước lượng. Chẳng hạn dạng hồi quy là tuyến tính, tức là f(X) = a + b.X Để ước lượng mô hình kinh tế lượng, ta cần tới việc thu thập và xử lý các số liệu về các biến trong mô hình.

d. Ước lượng các tham số: Các tham số trong mô hình kinh tế lượng là các hằng số chưa biết của tổng thể. Ở đây chúng ta sẽ dùng phương pháp thông dụng nhất, đó là phương pháp bình phương bé nhất thông thường (*Ordinary Least Squares*) hay cò gọi là phương pháp bình phương tối thiểu thông thường, viết tắt là: OLS.

e. Kiểm định giả thuyết về tính phù hợp của mô hình

Mục đích kiểm định giả thuyết là:

- Xác định mức độ phù hợp về mặt lý thuyết của mô hình
- Xác định mức độ phù hợp của dạng mô hình với số liệu điều tra và phát hiện dấu hiệu có thể bị vi phạm các giả thiết cổ điển của mô hình kinh tế lượng.

Chẳng hạn về quan hệ thu nhập X – tiêu dùng Y, nếu ta định dạng mô hình kinh tế lượng là:

$$\begin{cases} E(Y|X) = a + b.X \\ Y = a + b.X + U \end{cases}$$

thì do quan hệ giữa Y và X thực tế là đồng biến, tức là phải có $b > 0$. Mặt khác do sự gia tăng trong tiêu dùng không nhanh nhiều như trong thu nhập, có nghĩa là $b < 1$. Vậy phải kiểm định $b \in (0, 1)$, đó là sự kiểm định về tính phù hợp với lý thuyết kinh tế của mô hình. Ngoài ra người ta còn quan tâm tới mức độ thích hợp cũng như các tính chất của một mô hình tốt. Nếu mô hình ước lượng chưa đạt được các tiêu chuẩn của một mô hình tốt thì cần kiểm tra lại bước b/và c/. Nếu mô hình được đánh giá là tốt thì sử dụng mô hình để đánh giá, dự báo, ra quyết định,...

h. Đánh giá, dự báo

Khi mô hình được đánh giá là phù hợp, là tốt, ta sử dụng nó để đánh giá, phân tích, dự báo về mối liên hệ giữa biến phụ thuộc với các biến giải thích, qua đó đánh giá, dự báo và ra quyết định đối với những vấn đề có liên quan.

1.2. Khái niệm về hồi quy và phân tích hồi quy

1.2.1. Số liệu cho phân tích hồi quy

a. Phân loại số liệu: Số liệu được chia làm 3 loại: *Các số liệu theo thời gian* (hay là chuỗi thời gian), *các số liệu chéo* và *các số liệu hỗn hợp*.

- *Các số liệu theo thời gian* là các số liệu về một biến hay một véc tơ quan sát trên cùng một đối tượng quan sát (cùng một địa phương, một đơn vị, ...) ở những thời kỳ (ngày, tuần, tháng, năm,...) khác nhau.
- *Các số liệu chéo* là các số liệu về một biến hay một véc tơ quan sát được thu thập trong cùng một thời kỳ ở nhiều đối tượng quan sát (nhiều địa phương, đơn vị,...) khác nhau. - *Các số liệu hỗn hợp* hay các số liệu chéo và theo chuỗi thời gian là sự kết hợp của hai loại nói trên, đó là các số liệu về một biến hay một véc tơ quan sát trên các đối tượng quan sát (các địa phương, các đơn vị, ...) khác nhau ở những thời kỳ (ngày, tuần, tháng, năm,...) khác nhau.

Ví dụ 1:

K. sát giá vàng: $\begin{cases} \text{Trong 10 ngày tại tp. HCM} & \rightarrow Số\ liệu\ theo\ thời\ gian. \\ \text{Trong ngày hôm qua tại : Hà Nội, tp. HCM, Đ. nẵng} & \rightarrow Số\ liệu\ chéo. \\ \text{Trong 10 ngày tại: Hà Nội, tp. HCM, Đà nẵng} & \rightarrow Số\ liệu\ hỗn\ hợp. \end{cases}$

Việc phân loại số liệu là cần thiết đối với người sử dụng, vì mỗi loại số liệu đều có những đặc tính ưu việt hay hạn chế riêng đối với mô hình.

b. Nguồn số liệu: Số liệu được sử dụng trong phân tích hồi quy được thu thập từ hai nguồn: Số liệu điều tra thực tế và số liệu thử nghiệm.

Số liệu thử nghiệm nhận được từ việc tiến hành thử nghiệm trong những điều kiện nhất định nào đó (có thể do người thử nghiệm, quan sát đặt ra) để quan sát, đo đạc. Nguồn số liệu này thường xuất hiện trong các lĩnh vực khoa học, kỹ thuật. Chẳng hạn người ta áp dụng các chế độ canh tác khác nhau trên các thửa ruộng để quan sát tác động của chúng trên năng suất của một giống lúa.

Số liệu thực tế không chịu tác động ảnh hưởng của người điều tra, quan sát. Chẳng hạn các số liệu về giá vàng, giá bất động sản, tỷ lệ hộ nghèo, mức thu nhập,...không nằm trong sự kiểm soát của người điều tra, quan sát, là những số liệu thực tế. Đối với các số liệu thực tế, việc phân tích ảnh hưởng của một yếu tố nào đó đối với biến phụ thuộc sẽ khó khăn hơn do người ta không kiểm soát được những tác động của chúng.

Chất lượng của số liệu là ở chỗ nó có tính khách quan, có phản ánh đúng thực chất của hiện tượng, đối tượng quan sát, nghiên cứu hay không. Có thể chỉ ra các nguyên nhân sau khiến cho chất lượng số liệu thường không hoàn hảo:

- Vấn đề sai số trong các phép đo, quan sát.
- Vấn đề sai số, sai lầm, bỏ sót trong quá trình thu thập số liệu.
- Vấn đề lựa chọn phương pháp điều tra, chọn mẫu.
- Mức độ tổng hợp và tính chất bảo mật của số liệu.

Vậy chúng ta chỉ có thể tìm hàm hồi quy phù hợp nhất với số liệu đã có.

1.2.2. Hàm hồi quy tổng thể PRF (*Population regression function*)

Trung bình có điều kiện (hay kỳ vọng có điều kiện) của biến Y theo tập biến (hay véc tơ) X là $E(Y|X)$ được gọi là hàm hồi quy tổng thể của Y theo X, tức là hàm hồi quy được xây dựng dựa trên kết quả nghiên cứu khảo sát tổng thể, viết tắt là PRF.

Ví dụ 2: Tổng thể là 60 hộ gia đình ở một khu vực nhỏ với 2 tiêu chí quan sát: X (USD) là mức thu nhập hàng tuần của một hộ, Y (USD) là mức chi tiêu 1 tuần của một hộ. Điều tra toàn bộ tổng thể ta có kết quả sau, trong đó Y_x là các giá trị của biến Y ứng với $X = x$, ở đây có nghĩa là mức chi tiêu 1 tuần của các hộ có cùng mức thu nhập hang tuần là $X = x$, còn n_x là tổng số hộ có cùng mức thu nhập hang tuần là $X = x$ và hàm PRF là $E(Y|X)$ có các giá trị tương ứng với giá trị của X được chỉ ra ở dòng cuối cùng của bảng tính sau:

X / Y	80	100	120	140	160	180	200	220	240	260
	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
		88		113	125	140		160	189	185
				115				162		191
Y_x	Y_{x1}	Y_{x2}	Y_{x3}	Y_{x4}	Y_{x5}	Y_{x6}	Y_{x7}	Y_{x8}	Y_{x9}	Y_{x10}
n_x	325	462	445	707	678	750	685	1043	966	1211
$E(Y\|X)$	65	77	89	101	113	125	137	149	161	173

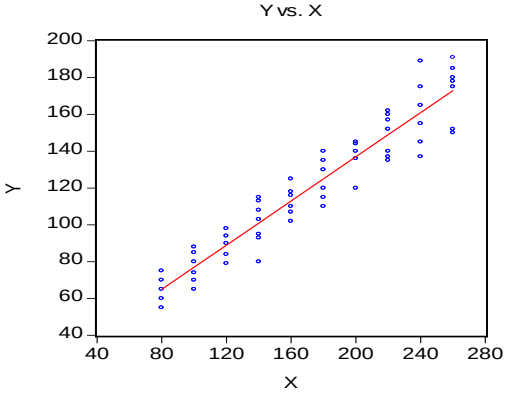
Bảng 2: *Khảo sát về thu nhập và chi tiêu của 60 hộ gia đình*

Trong bảng ta có:

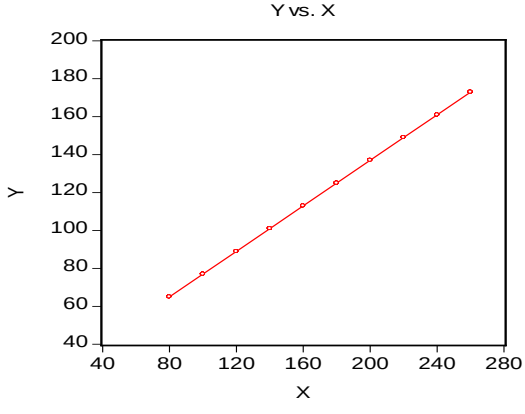
- Mức chi tiêu bình quân trong tuần của những hộ có cùng mức thu nhập 80USD là:

$$E(Y|X = x1 = 80) = EY_{x1} = \frac{55+60+65+70+75}{5} = \frac{325}{5} = 65(USD), \dots; E(Y|x10) = 173.$$

Nhờ phần mềm Eviews, hình 1.1.a cho thấy biểu đồ phân tán của chi tiêu Y theo thu nhập X, hình 1.1.b là đồ thị mô tả sự phụ thuộc của giá trị trung bình có điều kiện của tiêu dùng Y theo giá trị của thu nhập X.



Hình 1.1.a



Hình 1.1.b

Đặt: $U = Y - E(Y|X)$ thì U là một biến ngẫu nhiên. Đó là sai số giữa biến quan sát Y với trung bình có điều kiện của Y theo X. Ta gọi U là sai số ngẫu nhiên.

Trong ví dụ trên, ta có:

$y_1 - E(Y|X = x1) = 55 - 65 = -10, y_2 - E(Y|X = x1) = 60 - 65 = -5,$
$y_3 - E(Y|X = x1) = 65 - 65 = 0, y_4 - E(Y|X = x1) = 70 - 65 = 5,$
$y_5 - E(Y|X = x1) = 75 - 65 = 10, \dots$
$y_{54} - E(Y|X = x10) = 150 - 173 = -23, y_{55} - E(Y|X = x10) = -21,$

$y_{56} - E(Y|X = x10) = 2, y_{57} - E(Y|X = x1) = 5, y_{58} - E(Y|X = x10) = 7, \ldots,$
$y_{60} - E(Y|X = x10) = 18$
Như vậy sai số ngẫu nhiên U tập trung khá đối xứng xung quanh số 0. Mặt khác ta có: $= E\big(Y - E(Y|X)\big) = EY - EY = 0$. Điều này cho thấy sai số ngẫu nhiên U là đại lượng ngẫu nhiên có phân phối xấp xỉ phân phối chuẩn với giá trị trung bình EU = 0.
Cần lưu ý rằng: Hàm hồi quy tổng thể PRF của Y theo X là một hàm của X, vì vậy nếu X là biến ngẫu nhiên thì $E(Y|X)$ là một biến ngẫu nhiên, nếu X là biến tất định (không ngẫu nhiên) thì $E(Y|X)$ là một hàm số tất định. Trong ví dụ trên, với tổng thể là 60 hộ gia đình thì $E(Y|X)$ là biến ngẫu nhiên có 10 giá trị: $E(Y|X) = 65, nếu\ X = 80;\ E(Y|X) = 77, nếu\ X = 100, \ldots, E(Y|X) = 173, nếu\ X = 260.$

Với biến ngẫu nhiên U thỏa mãn một số tính chất nào đó (như là tính chất của sai số ngẫu nhiên), ta gọi: $E(Y|X) + U$ là hàm hồi quy tổng thể ngẫu nhiên, hay PRF ngẫu nhiên của Y theo X. Cần nhớ rằng PRF ngẫu nhiên luôn là biến ngẫu nhiên. Mô hình:

$$\begin{cases} E(Y|X) = f(X) & (3a) \\ Y = E(Y|X) + U & (3b) \end{cases}$$

cho phép ta xấp xỉ biến cần giải thích Y bởi hàm hồi quy tổng thể ngẫu nhiên, gọi là *mô hình kinh tế lượng*.
Trong mô hình (3a, 3b), ta vẫn gọi U là sai số ngẫu nhiên. Thành phần U xuất hiện trong mô hình với vai trò là tác động ngẫu nhiên của những yếu tố khác mà chúng không được đưa vào mô hình. Sự có mặt của U thể được giải thích bởi những nguyên nhân sau:
* Ta không biết hết được các yếu tố ảnh hưởng đến biến phụ thuộc, tác động của chúng đối với biến phụ thuộc nằm ngoài khả năng nhận biết của chúng ta.
* Ta không thể có được số liệu cho mọi yếu tố ảnh hưởng, kể cả khi biết chúng có ảnh hưởng đến biến phụ thuộc.
* Mô hình sẽ trở nên quá phức tạp nếu ta đưa hết các yếu tố ảnh hưởng vào mô hình. Vì thế thông thường người ta chỉ giữ lại những yếu tố có ảnh hưởng quan trọng trong mô hình, các yếu tố khác có ảnh hưởng không được đưa vào sẽ nhập vào thành phần nhiễu.
* Sai số ngẫu nhiên trong thu thập số liệu.
Chú ý:
a/ Nếu hàm PRF chỉ có 1 biến giải thích thì được gọi là hàm hồi quy đơn hay hồi quy hai biến. Nếu PRF có nhiều hơn 2 biến giải thích thì được gọi là hàm hồi quy nhiều chiều hay hồi quy bội, hồi quy nhiều biến.
b/ Nếu số liệu điều tra là số liệu theo thời gian thì mô hình kinh tế lượng (3a), (3b) được quy ước viết:

$$\begin{cases} E(Y|X_t) = f(X_t) & (3a') \\ Y_t = E(Y|X_t) + U_t & (3b') \end{cases}$$

Trong đó chỉ số t biểu thị thời điểm hay thời kỳ của số liệu.
Nếu số liệu điều tra là số liệu chéo thì mô hình kinh tế lượng (3a), (3b) được quy ước viết:

$$\begin{cases} E(Y|X_i) = f(X_i) & (3a'') \\ Y_i = E(Y|X_i) + U_i & (3b'') \end{cases}$$

Trong đó i là chỉ số thứ tự được sắp của quan sát.

c/ Việc định dạng hàm hồi quy tổng thể là vấn đề rất quan trọng, ảnh hưởng rất lớn đến tính phù hợp, tính chính xác của các ước lượng, đánh giá, dự báo hay ra các quyết định dựa trên mô hình. Đối với vấn đề này, ta cần dựa vào nhiều yếu tố, trước hết là bản chất của mối liên hệ giữa biến phụ thuộc với các biến giải thích trên cơ sở lý thuyết kinh tế. Về phương diện trực quan, ta dựa vào biểu đồ phân tán mô tả sự biến thiên của dãy các số liệu quan sát.

Chẳng hạn trong ví dụ trên, dựa vào bản chất của mối liên hệ giữa tiêu dùng đối với thu nhập và biểu đồ phân tán của dãy các số liệu (tập trung khá gần với một đường thẳng), ta định dạng hàm PRF xác định và PRF ngẫu nhiên như sau:

$$\begin{cases} E(Y|X) = f(X) = a + bX & (4a) \\ Y = E(Y|X) + U = a + bX + U & (4b) \end{cases}$$

Trong mô hình (4a, 4b): a, b là các tham số chưa biết được gọi là các hệ số hồi quy, trong đó a gọi là tung độ độ gốc hay hệ số tự do hoặc hệ số bị chặn, b gọi là độ dốc hay hệ số góc của đường thẳng hồi quy.

d/ *Mô hình hồi quy được gọi là tuyến tính nếu hàm hồi quy tuyến tính đối với các tham số của mô hình* (lưu ý rằng nó có thể không tuyến tính theo biến giải thích). Từ nay về sau, trong giáo trình này, ta *chỉ khảo sát mô hình hồi quy tuyến tính hoặc đưa được về dạng tuyến tính*.

Chẳng hạn các mô hình hồi quy sau đây là tuyến tính:

$Y = a + b_1X + b_2X^2 + U : mô\ hình\ Parabol$

$Y = b_1 + b_2 lnX + U: mô\ hình\ lin - log$

$Y = b_1 + b_2.\frac{1}{X} + U: mô\ hình\ nghịch\ đảo$

Các mô hình sau đây không phải là mô hình tuyến tính:

$$Y = \frac{1}{a} + bX + U \qquad (a)$$

$$Y = \frac{1}{1 + e^{a+bX}} + U \qquad (b)$$

Tuy nhiên (a) có thể đưa về mô hình tuyên tính:

$$Y = a' + bX + U \quad (a' = \frac{1}{a}) \qquad (a')$$

(b) có thể đưa về mô hình tuyến tính:

$$\ln\left(\frac{1}{Y} - 1\right) = a + bX + U' \qquad (b')$$

1.2.3. Hàm hồi quy mẫu SRF (*Sample Regression Function)*

Trong thực tế, người ta thường không thể điều tra toàn bộ tổng thể. Khi đó thay vì điều tra tổng thể, ta chỉ có thể dựa vào mẫu và hàm hồi quy xây dựng trên mẫu được gọi là

hàm hồi quy mẫu, viết tắt là SRF (*Sample Regression Function*). Hàm hồi quy mẫu SRF là hình ảnh của hàm hồi quy tổng thể PRF thông qua mẫu điều tra. Tuy nhiên khi thay đổi mẫu thì nói chung hàm hồi quy mẫu thay đổi. Vậy với số liệu mẫu, làm sao xây dựng một hàm hồi quy mẫu SRF gần nhất hay xấp xỉ tốt nhất cho hàm hồi quy tổng thể PRF? Ký hiệu $\hat{Y}$ là hàm hồi quy mẫu SRF thì $\hat{Y}$ thực chất là một ước lượng của hàm hồi quy tổng thể PRF. Khi đã định dạng hàm hồi quy tổng thể PRF (có chứa các tham số chưa biết gọi là các tham số của mô hình) thì hàm hồi quy mẫu SRF được định dạng tương ứng. Khi đó việc tìm ước lượng $\hat{Y}$ cho PRF được quy về tìm các ước lượng cho các tham số chưa biết của mô hình. Chẳng hạn nếu PRF xác định và ngẫu nhiên được định dạng là tuyến tính:

$$\begin{cases} E(Y|X) = a + bX, \\ Y = E(Y|X) + U = a + bX + U \end{cases}$$

thì hàm hồi quy mẫu được định dạng tương ứng là:

$$\begin{cases} \hat{Y} = \hat{a} + \hat{b}X \\ Y = \hat{Y} + \hat{U} = \hat{a} + \hat{b}X + \hat{U} \end{cases}$$

với $\hat{a},\ \hat{b},\ \hat{U}$ tương ứng là các ước lượng của $a,\ b,\ U$. Ta gọi $\hat{U}$ là phần dư hay thặng dư (*residuals*). Để tìm hàm hồi quy ước lượng $\hat{Y}$, người ta sử dụng phương pháp bình phương bé nhất sẽ được đưa vào trong chương sau. Chẳng hạn từ tổng thể 60 hộ gia đình trong ví dụ trên, ta lấy mẫu 10 hộ:

X	80	100	120	140	160	180	200	220	240	260
Y	60	74	90	108	116	130	136	140	145	175

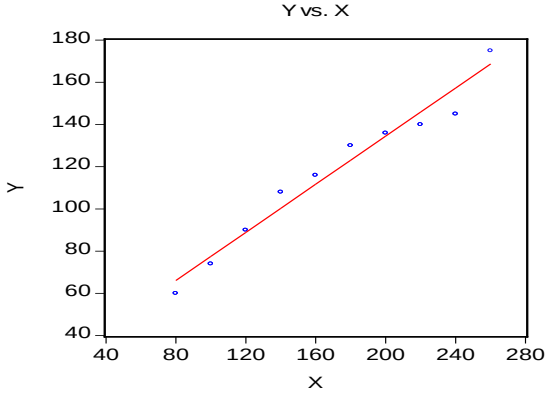
Từ mẫu này, biểu đồ phân tán của Y theo X (được cho bởi Eviews) cho thấy các điểm phân tán sắp xếp rất gần với một đường thẳng, kết hợp với bản chất mối quan hệ giữa thu nhập và tiêu dùng, ta nhận dạng

PRF: $\begin{cases} E(Y|X) = a + bX \\ Y = a + bX + U \end{cases}$

Do đó:

SRF: $\begin{cases} \hat{Y} = \hat{a} + \hat{b}X \\ Y = \hat{a} + \hat{b}X + \hat{U} \end{cases}$

Trong đó việc tìm các ước lượng $\hat{a}, \hat{b}$ sẽ được đề cập trong chương 2.



Hình 1.2
Biểu đồ phân tán Y theo X từ mẫu

Chương 2.

MÔ HÌNH HỒI QUY HAI BIẾN

Mô hình hồi quy hai biến hay mô hình hồi quy đơn là dạng đơn giản nhất, tuy ít có ý nghĩa về mặt thực tế, nhưng lại là cơ sở cho việc khảo sát mô hình hồi quy bội. Trong chương này ta tập trung vào mô hình hồi quy tuyến tính hai biến với các vấn đề về ước lượng và kiểm định giả thuyết có liên quan.

Nhắc lại rằng: Hồi quy hồi tuyến tính hai biến với biến giải thích X và biến phụ thuộc Y có:

- Mô hình PRF (mô hình lý thuyết hay mô hình tổng thể):

$$\begin{cases} E(Y|X) = a + bX, & (1a) \\ Y = E(Y|X) + U = a + bX + U & (1b) \end{cases}$$

Trong đó a, b là các hệ số hồi quy: a được gọi là hệ số bị chặn hay hệ số tự do, nó là tung độ gốc của đường thẳng hồi quy (1a); b được gọi là hệ số hồi quy của biến X, nó là độ dốc hay hệ số góc của đường thẳng hồi quy (1a).

Để thấy được ý nghĩa của hệ số hồi quy b, từ (1a) và (1b), cho biến X lần lượt lấy giá trị x, x + 1, ta có:

$$\begin{cases} b = E(Y|X = x + 1) - E(Y|X = x) & (2a) \\ b = Y_{x+1} - Y_x - (U_{x+1} - U_x) & (2b) \end{cases}$$

Theo (2a), *b chính là lượng tăng hay giảm bình quân (theo dự báo qua mô hình) của biến phụ thuộc Y khi biến giải thích X tăng lên 1 đơn vị.*

Theo (2b), *b chính là lượng tăng hay giảm (theo dự báo qua mô hình) của biến phụ thuộc Y khi biến giải thích X tăng thêm 1 đơn vị trong điều kiện các yếu tố khác không thay đổi* (vì khi đó: $U_{x+1} = U_x$)

- Mô hình SRF (hay mô hình ước lượng):

$$\begin{cases} \hat{Y} = \hat{a} + \hat{b}X & (3a) \\ Y = \hat{Y} + \hat{U} = \hat{a} + \hat{b}X + \hat{U} & (3b) \end{cases}$$

Như vậy: $\hat{b}$ chính là lượng tăng hay giảm bình quân (theo dự báo qua mô hình ước lượng) của biến phụ thuộc Y khi biến giải thích X tăng lên 1 đơn vị. Nói theo một cách khác: $\hat{b}$ chính là lượng tăng hay giảm (theo dự báo qua mô hình ước lương) của biến phụ thuộc Y khi biến giải thích X tăng thêm 1 đơn vị trong điều kiện các yếu tố khác không thay đổi .

2.1. Ước lượng các tham số hồi quy

Xét hồi quy tuyến tính hai biến với biến giải thích X và biến phụ thuộc Y có

- Mô hình PRF (mô hình lý thuyết hay mô hình tổng thể):

$$\begin{cases} E(Y|X) = a + bX, \\ Y = E(Y|X) + U = a + bX + U \end{cases} \quad (2.1)$$

- Mô hình SRF (hay mô hình ước lượng):

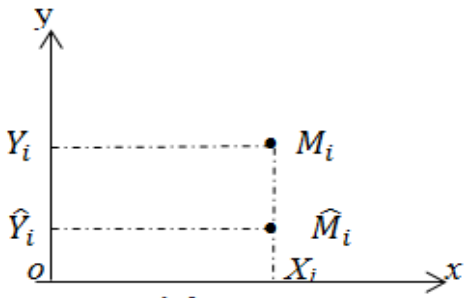
$$\begin{cases} \hat{Y} = \hat{a} + \hat{b}X \\ Y = \hat{Y} + \hat{U} = \hat{a} + \hat{b}X + \hat{U} \end{cases} \quad (2.2)$$

Trong phần này ta tìm các ước lượng $\hat{a},\ \hat{b}$ cho các hệ số hồi quy a, b của mô hình tổng thể (2.1) tốt nhất theo nghĩa dưới đây.

2.1.1. Phương pháp bình phương bé nhất thông thường OLS

(OLS = *Ordinary Least Squares*)

Trước hết ta mô tả ý tưởng của phương pháp OLS như sau: Trong mặt phẳng Oxy, ta gọi các điểm $M_i(X_i, Y_i)$ là các điểm thực nghiệm (điểm quan sát) $\widehat{M}_i(X_i, \hat{Y}_i)$ là các điểm hồi quy ước lượng, $i = \overline{1,n}$. Khi đó: $\hat{U}_i^2 = \left(Y_i - \hat{Y}_i\right)^2$ là bình phương khoảng cách từ điểm quan sát M_i đến điểm hồi quy ước lượng $\widehat{M}_i$.



Hình 2.1

Ta muốn tìm các ước lượng $\hat{a}$, $\hat{b}$ cho các hệ số hồi quy a, b sao cho tổng bình phương các khoảng cách từ các điểm quan sát đến các điểm ước lượng là bé nhất. Điều này có nghĩa là:

$$F(\hat{a},\hat{b}) = \sum_{i=1}^{n} \hat{U}_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{a} - \hat{b}X_i\right)^2 \to min$$

Vậy bài toán bây giờ là: tìm điểm $(\hat{a},\hat{b})$ mà tại đó hàm hai biến $F(\hat{a},\hat{b})$ đạt trị nhỏ nhất.

Ta có: $\frac{\partial F}{\partial \hat{a}} = -2\sum_{i=1}^{n}\left(Y_i - \hat{a} - \hat{b}X_i\right)$, $\frac{\partial F}{\partial \hat{b}} = -2\sum_{i=1}^{n} X_i\left(Y_i - \hat{a} - \hat{b}X_i\right)$,

Hệ phương trình:

$$\begin{cases} \dfrac{\partial F}{\partial \hat{a}} = 0 \\ \dfrac{\partial F}{\partial \hat{b}} = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n}\left(Y_i - \hat{a} - \hat{b}X_i\right) = 0 \\ \displaystyle\sum_{i=1}^{n} X_i\left(Y_i - \hat{a} - \hat{b}X_i\right) = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{a} = \bar{Y} - \hat{b}.\bar{X} \\ \hat{b} = \dfrac{\overline{X.Y} - \bar{X}.\bar{Y}}{S^2(X)} \end{cases} \qquad (*)$$

Tính:

$$A = \frac{\partial^2 F}{\partial \hat{a}^2} = 2n, B = \frac{\partial^2 F}{\partial \hat{a}\partial \hat{b}} = 2\sum_{i=1}^{n} X_i = 2n\bar{X}, C = \frac{\partial^2 F}{\partial \hat{\beta}^2} = 2\sum_{i=1}^{n} X_i^2 = 2n\overline{X^2}$$

$$\Longrightarrow A > 0, \qquad \Delta = AC - B^2 = 4n^2\left(\overline{X^2} - \bar{X}^2\right) = 4n^2 S^2(X) > 0$$

Suy ra hàm $F(\hat{a},\hat{b})$ đạt cực trị duy nhất tại điểm $(\hat{a},\hat{b})$ *xác định bởi* $(*)$ là điểm cực tiểu. Vì thế $F(\hat{a},\hat{b})$ đạt trị nhỏ nhất tại điểm này.

Vậy: $\hat{a} = \bar{Y} - \hat{b}.\bar{X}$, $\hat{b} = \frac{\overline{X.Y} - \bar{X}.\bar{Y}}{S^2(X)}$ là các ước lượng cần tìm.

Tóm lại: Bằng phương pháp bình phương bé nhất thông thường OLS, đối với mô hình hồi quy tuyến tính PRF của Y theo X là:

$$\begin{cases} E(Y|X) = a + bX, \\ Y = E(Y|X) + U = a + bX + U \end{cases}$$

ta tìm được mô hình SRF (hay mô hình ước lượng):

$$\begin{cases} \hat{Y} = \hat{a} + \hat{b}X \\ Y = \hat{Y} + \hat{U} = \hat{a} + \hat{b}X + \hat{U} \end{cases} \quad (2.3)$$

Trong đó: $\hat{a} = \bar{Y} - \hat{b}.\bar{X}, \ \hat{b} = \frac{\overline{X.Y} - \bar{X}.\bar{Y}}{S^2(X)}$ (*)

(Nhắc lại các đại lượng thống kê: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i : Trung\ bình\ mẫu\ của\ X$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i : Trung\ bình\ mẫu\ của\ Y$$

$$\overline{X.Y} = \frac{1}{n}\sum_{i=1}^{n} X_i.Y_i : Trung\ bình\ mẫu\ của\ X.Y$$

$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 : Trung\ bình\ mẫu\ của\ X^2$$

$$S^2(X) = \overline{X^2} - \bar{X}^2 : Phương\ sai\ mẫu\ của\ X \ )$$

2.1.2. Chú ý:

a. Để thiết lập mô hình SRF (hay mô hình ước lượng), ta có thể lập bảng tính, sau này cùng với những tính toán phức tạp hơn, ta có thể sử dụng phần mềm hỗ trợ.

b. Mô hình SRF (2) có thể viết lại dưới dạng:

$$\begin{cases} \hat{Y} - \bar{Y} = \hat{b}(X - \bar{X}) \\ Y - \bar{Y} = \hat{b}(X - \bar{X}) + \hat{U} \end{cases} \quad (2.4)$$

Trong đó: $\hat{b} = \frac{\overline{X.Y} - \bar{X}.\bar{Y}}{S^2(X)}$

Từ đây suy ra:

* Hàm SRF tìm theo phương pháp OLS là $y = \hat{Y}$ có đồ thị luôn đi qua điểm trung bình mẫu $(\bar{X}, \bar{Y})$

* Với một mẫu xác định thì hệ số hồi quy $\hat{b}$ là số xác định, nó biểu thị lượng tăng hay giảm của trung bình biến phụ thuộc Y khi biến độc lập X tăng một đơn vị. Dấu của $\hat{b}$ biểu thị xu thế thuận, nghịch của tương quan giữa Y và X.

Khi đó trên mẫu này ta có:

$\bar{\hat{Y}} = \bar{Y}, \ \bar{\hat{U}} = 0, \ E\hat{Y} = E\bar{\hat{Y}} = E\bar{Y} = EY, \ E\hat{U} = E\bar{\hat{U}} = 0 = EU.$ (2.5)

* $\hat{U}$ không tương quan với X, $\hat{U}$ không tương quan với $\hat{Y}, tức\ là$:

$$cov(\hat{U}, X) = cov(\hat{U}, \hat{Y}) = 0 \quad (2.6)$$

c. Với một mẫu cụ thể thì $\hat{a}, \hat{b}$ là các hằng số xác định, nhưng khi mẫu thay đổi thì hệ thức (*) cho thấy rằng $\hat{a}, \hat{b}$ là các đại lượng ngẫu nhiên.
d. Ký hiệu: $x = X - \bar{X},\ y = Y - \bar{Y},\ \hat{y} = \hat{Y} - \bar{\hat{Y}},\ \hat{u} = \hat{U} - \bar{\hat{U}}$ tương ứng là các độ lệch của các biến $X,\ Y,\ \hat{Y},\ \hat{U}$ so với trung bình mẫu của chúng. Từ mô hình hồi quy SRF nói trên, ta có:

$$\hat{y} = \hat{b}.x,\ \hat{u} = \hat{U},\ \overline{\hat{y}.\hat{u}} = 0\ v\text{à}:\ y = \hat{y} + \hat{u} \qquad (2.7)$$

Ví dụ 1: Với một mẫu điều tra về mức thu nhập X và mức tiêu dùng Y gồm 10 hộ gia đình từ tổng thể 60 hộ trong ví dụ trước đây ở chương 1, ta có các số liệu sau:

X	80	100	120	140	160	180	200	220	240	260
Y	60	74	90	108	116	130	136	140	145	175

Để thấy được các bước tính toán, vào Excel, lập bảng tính:

X	Y	X^2	Y^2	X.Y
80	60	6400	3600	4800
100	74	10000	5476	7400
120	90	14400	8100	10800
140	108	19600	11664	15120
160	116	25600	13456	18560
180	130	32400	16900	23400
200	136	40000	18496	27200
220	140	48400	19600	30800
240	145	57600	21025	34800
260	175	67600	30625	45500
Tổng:	**1700**	**1174**	**322000**	**148942**	**218380**

Suy ra:
$\bar{X} = \frac{1700}{10} = 170,\ \bar{Y} = \frac{1174}{10} = 117{,}4$

$\overline{X^2} = \frac{322000}{10} = 32200,$

$\overline{Y^2} = \frac{148942}{10} = 14894{,}2,$

$\overline{X.Y} = \frac{218380}{10} = 21838$

$$\hat{b} = \frac{\overline{X.Y} - \bar{X}.\bar{Y}}{S^2(X)} = \frac{21838 - 170.117{,}4}{32200 - 170^2} = 0{,}569657,$$
$$\hat{a} = \bar{Y} - \hat{b}.\bar{X} = 117{,}4 - 0{,}569657.170 = 20{,}55831$$

Vậy ta có mô hình SRF: $\begin{cases} \hat{Y} = 20{,}55831 + 0{,}569657.X \\ Y = \hat{Y} + \hat{U} = 20{,}55831 + 0{,}569657.X + \hat{U} \end{cases}$

Hệ số $\hat{b} = 0{,}569657$ cho thấy khi thu nhập của các hộ tăng thêm 1 USD thì bình quân mức tiêu dùng tăng lên 0,569657 USD.

2.2. Hệ số xác định

a. Các tổng bình phương độ lệch: Xét mô hình SRF (3) nhận được bằng phương pháp OLS. Ký hiệu:

$$TSS = n\overline{y^2} = \sum y_i^2 = nS^2(Y) = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n.(\bar{Y})^2 \qquad (2.9)$$

là *tổng bình phương các độ lệch giữa các giá trị quan sát thực tế Y_i và giá trị trung bình $\bar{Y}$ của các quan sát*, còn gọi là *tổng bình phương các độ lệch của Y* (trên mẫu).(*TSS = Total Sum of Squares*), nó cho thấy toàn bộ sự biến thiên của biến phụ thuộc Y.

$$ESS = n.\overline{\hat{y}^2} = \sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2 = (\hat{b})^2 . \sum x_i^2 = nS^2(X).(\hat{b})^2 \qquad (2.10)$$

là *tổng bình phương các độ lệch giữa giá trị của biến hồi quy mẫu ước lượng của Y với giá trị trung bình của chúng*, còn gọi là *tổng bình phương độ lệch của Y được giải thích bởi SRF* (*ESS: Explained Sum of Squares*).

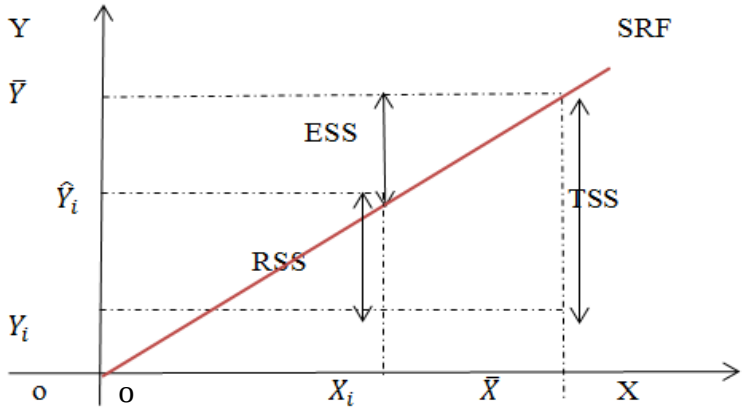
$$RSS = n.\overline{\hat{u}^2} = \sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2 \qquad (2.11)$$

là tổng bình phương các độ lệch giữa các giá trị quan sát Y_i và giá trị ước lượng (tính toán) $\hat{Y}_i$, còn được gọi là tổng bình phương các độ lệch của Y không được giải thích bởi SRF, hay tổng bình phương các phần dư (thặng dư) RSS do các yếu tố ngẫu nhiên gây ra (*RSS: Residual Sum of Squares*).

Nhận xét: Từ các tính chất của hàm SRF được chỉ ra ở trên, ta có:

$$\sum y_i^2 = \hat{b}^2 . \sum x_i^2 + \sum \hat{u}_i^2 \qquad (2.12)$$

Hay $$TSS = ESS + RSS \qquad (2.13)$$



Hình 2.2

b. Hệ số xác định

Từ (2.13) ta có: $$\frac{ESS}{TSS} + \frac{RSS}{TSS} = 1 \qquad (2.14)$$

Với một mẫu cụ thể, khi sử dụng phương pháp OLS, ta nhận được TSS là hằng số xác định, còn giá trị ESS và RSS còn thay đổi tùy theo dạng hàm hồi quy.

Mức độ phù hợp của hàm hồi quy mẫu SRF (hay của mô hình kinh tế lượng) với các số liệu quan sát được đánh giá qua mức độ gần nhau giữa các giá trị ước lượng $\hat{Y}_i$ *và các giá trị thực tế* Y_i. Vì thế tổng RSS càng bé (tức là càng gần về 0) thì SRF càng phù hợp.

Tuy nhiên ta lại không biết được RSS tăng đến hằng số nào thì mô hình kém phù hợp nhất. Vì vậy người ta đưa ra một đại lượng để đo mức độ phù hợp của hàm hồi quy mẫu SRF với các số liệu quan sát, gọi là hệ số xác định R^2 như sau:

$$\boldsymbol{R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}} \quad (2.15)$$

Từ định nghĩa, dễ thấy hệ số xác định R^2 có các tính chất sau:

Tính chất 1: $0 \le R^2 \le 1$

Tính chất 2: Khi $R^2 = 1$ thì hàm hồi quy mẫu SRF thích hợp một cách hoàn hảo với các số liệu quan sát, khi đó $\hat{Y}_i = Y_i,\ \forall i = 1,2,\dots,n,\ hay\ RSS = 0$, ta nói tất cả các sai lệch của Y_i (so với trị trung bình) đều được giải thích bởi SRF

Tính chất 3: Khi $R^2 = 0$ thì hàm hồi quy mẫu SRF không thích hợp, tất cả các sai lệch của Y_i (so với giá trị trung bình) đều không được giải thích bởi SRF (vì khi đó RSS = TSS, hay $\hat{Y}_i = \bar{Y},\ \forall i = 1,2,\dots,n.$).

Trong thực hành, đối với mô hình hồi quy hai biến, ta có thể sử dụng một trong các cách tính hệ số xác định như sau:

$$R^2 = \hat{b}^2.\frac{S^2(X)}{S^2(Y)} \quad (2.16)$$

$$R^2 = \frac{(\sum x_i y_i)^2}{(\sum x_i^2)(\sum y_i^2)} \quad (2.17)$$

$$R^2 = r_{XY}^2 = r_{X\hat{Y}}^2 \quad (2.18)$$

(r_{XY}: *hệ số tương quan mẫu giữa* X *và* Y; $r_{X\hat{Y}}$ *là hệ số tương quan mẫu giữa* X *và* $\hat{Y}$; lưu ý là phép biến đổi bậc nhất không làm thay đổi hệ số tương quan)

Ví dụ 2: Với mẫu trong ví dụ 1 và các kết quả tính toán đã có thì:

$$TSS = \sum y_i^2 = \sum Y_i^2 - n.\bar{Y}^2 = 148942 - 10.(117{,}4)^2 = 11114{,}4$$

$$ESS = \hat{b}^2\left(\sum X_i^2 - n.\bar{X}^2\right) = (0{,}569657)^2.(322000 - 10.170^2) = 10708{,}80022$$

$$RSS = TSS - ESS = 11114{,}4 - 10708{,}80022 = 405{,}59978$$

$$R^2 = \frac{ESS}{TSS} = \frac{10708{,}80022}{11114{,}4} = 0{,}9635$$

Vì $\hat{b} = 0{,}569657 > 0\ nên\ r_{XY} = \sqrt{R^2} = \sqrt{0{,}9635} = 0{,}98158$

Như vậy trong hàm hồi quy SRF, biến X giải thích được 96,35% sự thay đổi của biến phụ thuộc Y, 3,65% sự thay đổi còn lại của Y do các yếu tố ngẫu nhiên khác gây ra. Xu thế tương quan ở đây là thuận. Hàm SRF phù hợp khá cao với mẫu quan sát.

Chú ý:

c1. Thực tế người ta không có một tiêu chuẩn chung để đánh giá mức độ cao thấp của R^2 và không nên chỉ dựa vào R^2 để đánh giá mức độ phù hợp của mô hình mà còn phải dựa vào các yếu tố khác như kinh nghiệm thực tế, khả năng dự báo chính xác,....

c2. Theo kinh nghiệm thực tế, đối với số liệu chuỗi thời gian thì $R^2 > 0{,}9$ được xem phù hợp tốt, đối với số liệu chéo thì $R^2 > 0{,}7$ được xem phù hợp tốt.

c3. Theo công thức định nghĩa thì R^2 chính là tỷ lệ hay phần trăm sự biến thiên của biến phụ thuộc Y được giải thích bởi mô hình.

2.3. Các giả thiết của phương pháp OLS

Mục đích của việc xây dựng mô hình kinh tế lượng là dựa vào đó người ta giải quyết các bài toán thống kê: phân tích, đánh giá, lựa chọn, ước lượng, dự báo,…. Muốn có một mô hình ước lượng tốt thì trước hết các hệ số hồi quy ước lượng phải có những tính chất tốt. Để có được các ước lượng $\hat{a}\ cho\ a, \hat{b}\ cho\ b$ tìm theo phương pháp OLS có các tính chất tốt, mô hình cần đáp ứng các điều kiện sau đây mà người ta thường gọi là các giả thiết của mô hình hồi quy tuyến tính cổ điển:

Giả thiết 1: Mặc dù biến độc lập X là biến ngẫu nhiên, nhưng các giá trị của X thường được xác định trước, tức là phép lấy mẫu về biến X là không ngẫu nhiên.

Chẳng hạn trong việc khảo sát quan hệ giữa tiêu dùng Y và thu nhập X thì các số liệu về mức thu nhập X đã được định trước.

Giả thiết 2: Nhiễu U là đại lượng ngẫu nhiên có $E(U|X) = 0$, tức là nhiễu có giá trị trung bình bằng 0 và không phụ thuộc vào giá trị của X.

Giả thiết 3: Nhiễu U có phương sai có điều kiện $Var(U|X) = \sigma^2 = const$ (không phụ thuộc vào các giá trị của X).

Nhiễu U là mức độ dao động của các giá trị của biến Y xung quanh trung bình có điều kiện $E(Y|X)$. Giả thiết 3 có nghĩa là dao động này có biên độ không đổi khi giá trị của X thay đổi. Tuy nhiên trong thực tế, không phải giả thiết này lúc nào cũng được thỏa mãn. Chẳng hạn như chi tiêu của những người có mức thu nhập thấp và thu nhập cao thường có xu hướng khác nhau: Chi tiêu của nhóm thu nhập thấp thường chỉ tập trung vào những mặt hàng thiết yếu, nhưng ngoài những mặt hàng thiết yếu thì đối với nhóm thu nhập cao còn có các khoản chi tiêu cho những nhu cầu giải trí, mặt hàng xa xỉ, tức là không có sự đồng đều về chi tiêu giữa các nhóm này. Khi đó nếu ta quan sát thu nhập và chi tiêu của cả hai nhóm này thì dễ có hiện tượng phương sai nghiễu thay đổi.

Giả thiết 4: Không có sự tương quan giữa các sai số ngẫu nhiên

Giả thiết này được giải thích như sau: Sai số ngẫu nhiên $U = Y - E(Y|X)$ là một biến quan sát mà ứng với mẫu ngẫu nhiên $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ là n biến ngẫu nhiên tương ứng:

$$U_1 = Y_1 - E(Y|X_1), U_2 = Y_2 - E(Y|X_2), \dots, U_n = Y_n - E(Y|X_n).$$

Giả thiết không có sự tương quan giữa các sai số ngẫu nhiên có nghĩa là:

$$cov(U_i, U_j) = E\{(U_i - EU_i).(U_j - EU_j)\} = 0, \forall i \neq j$$

Giả thiết này có thể bị vi phạm khi đối tượng điều tra có sự ràng buộc, phụ thuộc nhau về tiêu chuẩn điều tra Y. Chẳng hạn: Khi khảo sát về thu nhập X và tiêu dùng Y mà đối tượng khảo sát là các thành viên trong một gia đình thì mặc dù các thành viên có mức thu nhập X khác nhau, nhưng những yếu tố ngoài thu nhập tác động lên chi tiêu như hoàn cảnh gia đình, tập quán, sở thích, thói quen tiêu dùng đều có thể tác động gần như tương đồng đến chi tiêu của các thành viên trong gia đình, điều này dẫn đến các tác động ngẫu nhiên có sự tương quan.

Giả thiết 5: X và U không tương quan, tức là:

$$cov(U, X) = E\{(U - EU).(X - EX)\} = 0$$

Điều này cũng có nghĩa là các thành phần X_i của mẫu ngẫu nhiên về X không tương quan với sai số ngẫu nhiên U_i tương ứng, tức là:

$$cov(U_i, X_i) = E\{(U_i - EU_i).(X_i - EX_i)\} = 0, \forall i.$$

Nếu biến giải thích X có tương quan với nhiễu U thì ta không thể tách rời ảnh hưởng của

biến giải thích X và của nhiễu lên biến phụ thuộc Y. Để minh họa cho giả thiết 5, ta quan sát thu nhập X và chi tiêu Y, với yếu tố hoàn cảnh gia đình là nhiễu có thể tác động lên hành vi tiêu dùng của thành viên trong gia đình thì *giả thiết 5* ở đây là xem yếu tố hoàn cảnh gia đình không tác động đến thu nhập của thành viên đó.
Giả thiết 6: Sai số ngẫu nhiên U là đại lượng ngẫu nhiên có phân phối chuẩn: $U \sim N(0, \sigma^2)$.

Chú ý:
- *Giả thiết 1* có thể được bỏ đi trong lý thuyết kinh tế lượng hiện đại
- Một giả thiết khá hiển nhiên là cỡ mẫu n lớn hơn số tham số của mô hình .
- Giả thiết về quy luật chuẩn của nhiễu được thỏa mãn khá rộng rãi trong thực tế và được ứng dụng để ước lượng, kiểm định và dự báo về các tham số trong mô hình, tuy nhiên giả thiết về phương sai không thay đổi có thể bị vi phạm.

2.4. Các tính chất của các hệ số hồi quy

Xét mô hình hồi quy PRF: $\begin{cases} E(Y|X) = a + bX, \\ Y = E(Y|X) + U = a + bX + U \end{cases}$

có mô hình hồi quy ước lượng SRF: $\begin{cases} \hat{Y} = \hat{a} + \hat{b}X \\ Y = \hat{Y} + \hat{U} = \hat{a} + \hat{b}X + \hat{U} \end{cases}$

các ước lượng $\hat{a}, \hat{b}$ nhận được nhờ dựa vào mẫu $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ nên chúng là các đại lượng ngẫu nhiên mà trên một mẫu cụ thể, chúng là các giá trị xác định và là các ước lượng điểm của các tham số a, b.
Các tính chất tốt của $\hat{a}, \hat{b}$ tìm theo phương pháp OLS được chỉ ra trong định lý sau:
Định lý Gauss – Markov: *Với mô hình hồi quy tuyến tính cổ điển:*

$$\begin{cases} E(Y|X) = a + bX, \\ Y = E(Y|X) + U = a + bX + U \end{cases}$$

thì các ước lượng $\hat{a}, \hat{b}$ tìm theo phương pháp OLS là các ước lượng tuyến tính không chệch, có phương sai bé nhất trong lớp các ước lượng tuyến tính không chệch của các tham số a, b.
Lưu ý:
- Các ước lượng $\hat{a}, \hat{b}$ là các ước lượng tuyến tính do biểu thức của chúng là hàm tuyến tính theo các thành phần mẫu $(Y_1, Y_2, \dots, Y_n)$, cụ thể ta có:

$\hat{b} = \sum_{j=1}^{n} c_j Y_j$ $\left(c_j = \frac{X_j - \bar{X}}{nS^2(X)} \text{ là hằng số}, j = 1,2, \dots, n\right)$

$\hat{a} = \bar{Y} - \hat{b}.\bar{X} = \sum_{j=1}^{n} \left(\frac{1}{n} - c_j \bar{X}\right).Y_j$

- Các ước lượng $\hat{a}, \hat{b}$ là các ước lượng không chệch của các tham số a, b có nghĩa là: $E\hat{a} = a$, $E\hat{b} = b$.
- Định lý Gauss – Markov cho thấy $\hat{a}, \hat{b}$ là các ước lượng hiệu quả nhất cho các tham số a, b. Tính tuyến tính, không chệch và hiệu quả nhất được gọi tắt là tính chất **BLUE (BLUE:** *Bets Linear Unbiased Estimators*).

- Với X là biến quan sát có phân phối chuẩn (hoặc xấp xỉ chuẩn) $N(a_0, b_0^2)$ và với các giả thiết của phương pháp OLS thì biến phụ thuộc Y cũng có phân phối chuẩn (hoặc xấp xỉ chuẩn) $N(a + ba_0, b^2 b_0^2 + \sigma^2)$.
- Do các ước lượng $\hat{a}, \hat{b}$ là các hàm tuyến tính theo các thành phần mẫu $(Y_1, Y_2, \dots, Y_n)$ nên chúng có phân phối chuẩn (hoặc xấp xỉ chuẩn):

$$\begin{cases} \hat{a} \sim N(a, \sigma_{\hat{a}}^2) \\ \hat{b} \sim N(b, \sigma_{\hat{b}}^2) \end{cases}$$

- Với các ước lượng $\hat{a}, \hat{b}$ tìm được bằng phương pháp OLS và với giả thiết mẫu về biến X là không ngẫu nhiên, ta có các công thức sau đây xác định phương sai, ký hiệu var(.) (var: variance) và độ lệch chuẩn (hay sai số chuẩn) của chúng , ký hiệu se(.) (se: standard error):

$$\sigma_{\hat{a}}^2 = var(\hat{a}) = \frac{\sum X_i^2}{n^2 S^2(X)} . \sigma^2 \quad ; \quad se(\hat{a}) = \sqrt{var(\hat{a})} \; ; \tag{2.19}$$

$$\sigma_{\hat{b}}^2 = var(\hat{b}) = \frac{\sigma^2}{nS^2(X)}; \quad se(\hat{b}) = \sqrt{var(\hat{b})} \; (\text{Ở đây } \sigma^2 = varU). \tag{2.20}$$

- Trong (2.19), (2.20) thì $\sigma^2 = var(U)$ là phương sai nhiễu của tổng thể, nói chung chưa biết, người ta dùng một ước lượng điểm của σ^2 là:

$$\hat{\sigma}^2 = \frac{RSS}{n-2} \; (\text{RSS} = \text{TSS} - \text{ESS} = n\{S^2(Y) - S^2(X).\hat{b}^2\}; \tag{2.21}$$

$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ được gọi là sai số chuẩn của hồi quy, ký hiệu là **SER** (**SER**: Standard Error of the Regression). Thay (2.21) vào (2.19), (2.20) ta nhận được các ước lượng điểm $\widehat{var}(\hat{a})$, $\widehat{var}(\hat{b})$ của $var(\hat{a}), var(\hat{b})$ là:

$$\widehat{var}(\hat{a}) = \frac{\sum X_i^2}{n^2 S^2(X)} . \hat{\sigma}^2 = \frac{\sum X_i^2}{n^2(n-2)S^2(X)} . \text{RSS}; \tag{2.22}$$

$$\widehat{var}(\hat{b}) = \frac{\hat{\sigma}^2}{nS^2(X)} = \frac{\text{RSS}}{n(n-2)S^2(X)}.$$

2.5. Khoảng tin cậy cho các tham số trong mô hình

Xét mô hình hồi quy PRF: $\begin{cases} E(Y|X) = a + bX, \\ Y = E(Y|X) + U = a + bX + U \end{cases}$ (2.23)

có mô hình hồi quy ước lượng SRF: $\begin{cases} \hat{Y} = \hat{a} + \hat{b}X \\ Y = \hat{Y} + \hat{U} = \hat{a} + \hat{b}X + \hat{U} \end{cases}$ (2.24)

trong đó các ước lượng $\hat{a}, \hat{b}$ tìm theo phương pháp OLS.
Chúng ta sẽ tìm khoảng tin cậy cho các hệ số hồi quy và phương sai nhiễu

2.5.1. Khoảng tin cậy cho các hệ số hồi quy

Trong mục này ta dùng $\hat{a}, \hat{b}$ để ước lượng khoảng tin cậy cho các hệ số hồi quy $a, \; b$.
* Với độ tin cậy: $\gamma = 1 - \alpha$, ta có khoảng tin cậy cho a là:

$$\left(\hat{a} - t_{\frac{\alpha}{2}}^{(n-2)} . \widehat{se}(\hat{a}); \quad \hat{a} + t_{\frac{\alpha}{2}}^{(n-2)} . \widehat{se}(\hat{a}) \right) \tag{2.25}$$

* Với độ tin cậy: $\gamma = 1 - \alpha$, ta có khoảng tin cậy cho b là:

$$\left(\hat{b} - t_{\frac{\alpha}{2}}^{(n-2)}.\widehat{se}(\hat{b});\ \hat{b} + t_{\frac{\alpha}{2}}^{(n-2)}.\widehat{se}(\hat{b})\right) \quad (2.26)$$

trong đó $t_{\frac{\alpha}{2}}^{(n-2)}$ là giá trị tới hạn (critical value) mức $\frac{\alpha}{2}$ của phân phối Student với n – 2 bậc tự do, tra từ bảng giá trị tới hạn của phân phối Student (bảng phụ lục I).

Ví dụ 3: Với độ tin cậy 95%, dựa vào mẫu 1 về thu nhập X và tiêu dùng Y trong ví dụ 2, hãy ước lượng khoảng tin cậy cho các tham số a, b trong mô hình hồi quy tuyến tính (2.23).

Giải: Ta có

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{405,59978}{8} = 50,6999725;\quad \widehat{se}(\hat{a}) = \sqrt{\frac{\sum X_i^2.\hat{\sigma}^2}{n.\sum x_i^2}} = 7,033554027;$$

$$\widehat{se}(\hat{b}) = \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}} = \sqrt{\frac{50,6999725}{322000-10.170^2}} = 0,039196464$$

Với độ tin cậy $\gamma = 1-\alpha = 0,95 \Rightarrow \frac{\alpha}{2} = 0,025, tra\ bảng\ có\ t_{0,025}^{(8)} = 2,306$

$$\begin{cases} \hat{b} - t_{\frac{\alpha}{2}}^{(n-2)}.\widehat{se}(\hat{b}) = 0,569657 - 2,306.0,039196464 = 0,479269953 \\ \hat{b} + t_{\frac{\alpha}{2}}^{(n-2)}.\widehat{se}(\hat{b}) = 0,569657 + 2,306.0,039196464 = 0,660044046 \end{cases}$$

Vậy với độ tin cậy 95%, dựa vào mẫu 1, ta có KTC cho b là:

$$(0,479269953\ ;\ 0,660044046\ )$$

Chú thích: Việc tìm khoảng tin cậy cho các hệ số hồi quy với ba mức độ tin cậy 90%, 95%, 99% có thể được thực hiện bởi Eviews 7, sau khi chạy hồi quy.

2.5.2. Khoảng tin cậy cho phương sai của nhiễu

Với độ tin cậy $\gamma = 1-\alpha$, khoảng tin cậy cho phương sai nhiễu σ^2 là:

$$\left(\frac{(n-2).\hat{\sigma}^2}{\chi_{\alpha/2}^2}\ ;\ \frac{(n-2).\hat{\sigma}^2}{\chi_{1-\alpha/2}^2}\right) \quad (2.27)$$

trong đó: $\chi_{\frac{\alpha}{2}}^2$, $\chi_{1-\frac{\alpha}{2}}^2$ *là các giá trị tới hạn của phân phối Khi − bình phương*,$(n-2)$ *bậc tự do* (tra từ bảng phụ lục III)

Ví dụ 4: Trong ví dụ trước với mẫu 1 về tiêu dùng Y và thu nhập X, ta ước lượng KTC cho σ^2 với độ tin cậy 95%.

Với độ tin cậy $\gamma = 1-\alpha = 0,95 \Rightarrow \frac{\alpha}{2} = 0,025$, tra bảng giá trị tới hạn của phân phối Chi-Square với bậc tự do n – 2 = 8, ta có: $\chi_{\alpha/2}^2 = 17,5345$; $\chi_{1-\alpha/2}^2 = 2,1797$.

$$\frac{(n-2).\hat{\sigma}^2}{\chi_{\alpha/2}^2} = \frac{8.50,6999725}{17,5345} = 23,1315\ ;\ \frac{(n-2).\hat{\sigma}^2}{\chi_{1-\alpha/2}^2} = \frac{8.50,6999725}{2,1797} = 186,0806$$

Vậy với độ tin cậy 95%, KTC cần tìm cho phương sai nhiễu σ^2 là:

$$(23,1315;\ 186,0806\ )$$

2.6. Kiểm định giả thuyết về mô hình

2.6.1. Kiểm định giả thuyết về hệ số hồi quy

Giả sử θ là một hằng số mà ta chưa biết và không thể biết chính xác. Dựa vào những thông tin nhất định, người ta có các nhận định sau: $\theta = \theta_0, \theta < \theta_0, \theta > \theta_0, \theta \neq \theta_0$. Để xác minh nhận định nào là phù hợp với thực tế, là chấp nhận được – như đã biết trong lý thuyết kiểm định giả thuyết thống kê, tùy thuộc vào bản chất của từng vấn đề liên quan mà ta xác định đối thuyết là một trong ba nhận định: $\theta < \theta_0, \theta > \theta_0, \theta \neq \theta_0$ để có một trong ba bài toán:

- Kiểm định hai phía: $Giả\ thuyết\ H_0: \theta = \theta_0, đối\ thuyết\ H_1: \theta \neq \theta_0$
- Kiểm định phía phải: $Giả\ thuyết\ H_0: \theta = \theta_0, đối\ thuyết\ H_1: \theta > \theta_0$
- Kiểm định phía trái: $Giả\ thuyết\ H_0: \theta = \theta_0, đối\ thuyết\ H_1: \theta < \theta_0$

Có ba phương pháp để kiểm định: Phương pháp khoảng tin cậy, phương pháp giá trị tới hạn, phương pháp giá trị p-value. Dưới đây, để đơn giản và tránh lặp lại trong cách trình bày, ta dùng ký hiệu θ thay thế cho hệ số hồi quy a hoặc b.

2.6.1.1. Phương pháp khoảng tin cậy:

a. Kiểm định hai phía: Cho trước mức ý nghĩa α, với độ tin cậy $\gamma = 1 - \alpha$ ta có khoảng tin cậy đối xứng cho θ là:

$$\left(\hat{\theta} - t_{\frac{\alpha}{2}}^{(n-2)}.\ \widehat{se}\left(\hat{\theta}\right);\ \hat{\theta} + t_{\frac{\alpha}{2}}^{(n-2)}.\ \widehat{se}\left(\hat{\theta}\right)\right).$$

- Từ số liệu điều tra, tính giá trị các đầu mút khoảng tin cậy.
- Nếu: $$\boldsymbol{\theta_0 \notin \left(\hat{\theta} - t_{\frac{\alpha}{2}}^{(n-2)}.\ \widehat{se}\left(\hat{\theta}\right);\ \hat{\theta} + t_{\frac{\alpha}{2}}^{(n-2)}.\ \widehat{se}\left(\hat{\theta}\right)\right)} \quad (2.28\ a)$$

thì ta bác bỏ giả thuyết H_0, tạm thời chấp nhận đối thuyết H_1. Nếu ngược lại thì tạm thời chấp nhận giả thuyết H_0. bác bỏ đối thuyết H_1.

b. Kiểm định phía phải:

Cho trước mức ý nghĩa α, với độ tin cậy $\gamma = 1 - \alpha$ ta có khoảng tin cậy bên phải cho θ là:

$$\left(\hat{\theta} - t_{\alpha}^{(n-2)}.\ \widehat{se}\left(\hat{\theta}\right);\ +\infty\right).$$

- Từ số liệu điều tra, tính giá trị các đầu mút khoảng tin cậy.
- Nếu: $$\boldsymbol{\theta_0 \notin \left(\hat{\theta} - t_{\alpha}^{(n-2)}.\ \widehat{se}\left(\hat{\theta}\right);\ +\infty\right)} \quad (2.28b)$$

thì ta bác bỏ giả thuyết H_0, tạm thời chấp nhận đối thuyết H_1. Nếu ngược lại thì tạm thời chấp nhận giả thuyết H_0. bác bỏ đối thuyết H_1.

c. Kiểm định phía trái:

Cho trước mức ý nghĩa α, với độ tin cậy $\gamma = 1 - \alpha$ ta có khoảng tin cậy bên trái cho θ là:

$$\left(-\infty\ ;\ \ \hat{\theta} + t_{\alpha}^{(n-2)}.\ \widehat{se}\left(\hat{\theta}\right)\right).$$

- Từ số liệu điều tra, tính giá trị các đầu mút khoảng tin cậy.
- Nếu: $$\boldsymbol{\theta_0 \notin \left(-\infty\ ;\ \ \hat{\theta} + t_{\alpha}^{(n-2)}.\ \widehat{se}\left(\hat{\theta}\right)\right)},\quad (2.28\ c)$$

thì ta bác bỏ giả thuyết H_0, tạm thời chấp nhận đối thuyết H_1. Nếu ngược lại thì tạm thời chấp nhận giả thuyết H_0. bác bỏ đối thuyết H_1.

2.6.1.2. Phương pháp giá trị tới hạn

Đây chính là phương pháp kiểm định thông thường trong Thống kê, kiểm định giả thuyết về giá trị trung bình $\theta_0 = E\hat{\theta}$

($lưu\ ý$: $E(\hat{a}) = a,\ \mathrm{E}(\hat{b}) = b$ và θ đóng vai trò là a hoặc b):

$Gt\ H_0: \theta = \theta_0$; đt $H_1: \theta < \theta_0 (trái)/\theta > \theta_0 (phải)/\theta \neq \theta_0 (hai\ phía)$.

Tiêu chuẩn kiểm định với mức ý nghĩa :

$$W = \left\{ t = \frac{\hat{\theta} - \theta_0}{\widehat{se}(\hat{\theta})} < -t_\alpha^{(n-2)} \right\}, nếu\ đối\ thuyết\ H_1: \theta < \theta_0, \quad (2.29)$$

$$W = \left\{ t = \frac{\hat{\theta} - \theta_0}{\widehat{se}(\hat{\theta})} > t_\alpha^{(n-2)} \right\}, nếu\ đối\ thuyết\ H_1: \theta > \theta_0, \quad (2.30)$$

$$W = \left\{ |t| = \left| \frac{\hat{\theta} - \theta_0}{\widehat{se}(\hat{\theta})} \right| > t_{\alpha/2}^{(n-2)} \right\}, nếu\ đối\ thuyết\ H_1: \theta \neq \theta_0, \quad (2.31)$$

Bước 1: $Tra\ bảng\ phân\ vị\ Student, tìm\ giá\ trị\ tới\ hạn\ t_\alpha^{(n-2)} hoặc\ t_{\frac{\alpha}{2}}^{(n-2)}$.

Bước 2: $Dựa\ vào\ số\ liệu, tính\ t = \dfrac{\hat{\theta} - \theta_0}{\widehat{se}(\hat{\theta})}$ $và\ so\ sánh\ với\ giá\ trị\ tới\ hạn$

* Nếu W xảy ra thì bác bỏ H_0, chấp nhận H_1

* Nếu W không xảy ra thì tạm thời chấp nhận H_0, bác bỏ H_1.

2.6.1.2. Phương pháp giá trị p-value

Bước 1: Từ mẫu điều tra, tính giá trị: $t_0 = \frac{\hat{\theta} - \theta_0}{\widehat{se}(\hat{\theta})}$.

Bước 2: Tính: $p - value = P(|t| > |t_0|)$, t là biến ngẫu nhiên có phân phối Student với $(n-2)$ bậc tự do. ($p - value$ được cung cấp bởi phần mềm ứng dụng)

Bước 3: Với mức ý nghĩa α cho trước, quy tắc kiểm định là:

$-$ Đối với kiểm định 2 phía: nếu $p - value < \alpha$ thì bác bỏ H_0;

$-$Đối với kiểm định một phía: nếu $p - value < 2\alpha$ thì bác bỏ H_0.

Ví dụ 5: Từ mẫu điều tra giữa biến giải thích X và biến phụ thuộc Y sau đây:

X	1	2	3	4	5	6	7
Y	8	6	6	5	4	4	3

trong mô hình hồi quy tuyến tính cổ điển, với mức ý nghĩa 5%, hãy kiểm định giả thuyết $H_0: b = 0$ ($Y\ không\ phụ\ thuộc\ X$), đối thuyết $H_1: b \neq 0$ ($Y\ phụ\ thuộc\ X$).

Giải: Từ mẫu ta có: $\hat{b} = \frac{\overline{X.Y} - \bar{X}.\bar{Y}}{S^2(X)} = -0{,}75;\ \widehat{se}(\hat{b}) = 0.088928 \Rightarrow t_0 = \frac{-0{,}75}{0{,}115728} = -8.433803;$

$$p - value = P(|t| > 8.433803) = 0.0004 < \alpha = 0{,}05.$$

$Vậy\ ta\ bác\ bỏ\ H_0, có\ nghĩa\ là\ Y\ có\ phụ\ thuộc\ tương\ quan\ với\ X.$

2.6.2. Kiểm định giả thuyết về phương sai của nhiễu

Với mức ý nghĩa α cho trước, hãy kiểm định giả thuyết $H_0: \sigma^2 = \sigma_0^2$ với đối thuyết $H_1: \sigma^2 \neq \sigma_0^2 / \sigma^2 > \sigma_0^2 / \sigma^2 < \sigma_0^2$. Các phương pháp giải quyết bài toán này được tóm tắt trong bảng sau:

Bài toán kiểm định	P.pháp kiểm định	Tiêu chuẩn bác bỏ giả thuyết
$\begin{cases} Gt\ H_0: \sigma^2 = \sigma_0^2 \\ Đt\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases}$	Khoảng tin cậy	$\sigma_0^2 \notin \left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2}};\ \frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}} \right)$
	Giá trị tới hạn	$\chi^2 \notin (\chi^2_{1-\alpha/2};\ \chi^2_{\alpha/2})$
	Giá trị p – value	$p - value \notin \left[\frac{\alpha}{2};\ 1 - \frac{\alpha}{2} \right]$
$\begin{cases} Gt\ H_0: \sigma^2 = \sigma_0^2 \\ Đt\ H_1: \sigma^2 > \sigma_0^2 \end{cases}$	Khoảng tin cậy	$\sigma_0^2 \leq \frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha}}$
	Giá trị tới hạn	$\chi^2 > \chi^2_{\alpha}$
	Giá trị p – value	$p - value < \alpha$
$\begin{cases} Gt\ H_0: \sigma^2 = \sigma_0^2 \\ Đt\ H_1: \sigma^2 < \sigma_0^2 \end{cases}$	Khoảng tin cậy	$\sigma_0^2 \geq \frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha}}$
	Giá trị tới hạn	$\chi^2 < \chi^2_{1-\alpha}$
	Giá trị p – value	$p - value > 1 - \alpha$
Ghi chú: *Từ số liệu tính*: $\chi^2 = \frac{(n-2)\hat{\sigma}^2}{\sigma_0^2}$; χ^2_{α} *tra từ bảng giá trị tới hạn của phân phối khi – bình phương* $(n-2)$ *bậc tự do*		

Bảng 2.1

Ví dụ 6: Xét mẫu 1 về tiêu dùng Y và thu nhập X, với mức ý nghĩa 5%, ta muốn xác minh xem phương sai σ^2 của nhiễu có vượt quá 1000 hay không.
Giải: Đây là bài toán kiểm định giả thuyết về phương sai nhiễu:

$$\begin{cases} Gt\ H_0: \sigma^2 = \sigma_0^2 = 50 \\ Đt\ H_1: \sigma^2 > 50 \end{cases}$$

Với phương pháp khoảng tin cậy, tiêu chuẩn bác bỏ H_0 là: $W = \{\sigma_0^2 \leq \frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha}}\}$
Với độ mức ý nghĩa $\alpha = 0{,}05$ tra bảng giá trị tới hạn của phân phối Chi-Square với bậc tự do n – 2 = 8, ta có:
$\chi^2_{\alpha} = 15{,}5073;\ \ \hat{\sigma}^2 = 50{,}6999725 \ \Rightarrow (n-2)\hat{\sigma}^2: \chi^2_{\alpha} = 26{,}0586 \ \Rightarrow \sigma_0^2 = 50 \in$ KTC
Do đó nhận H_0: p.sai nhiễu chưa vượt quá 50

2.6.3. Kiểm định giả thuyết về sự phù hợp của mô hình

Như đã biết hệ số xác định R^2 càng gần 1 thì mô hình càng có ý nghĩa, hệ số xác định R^2 càng gần 0 thì mô hình càng ít có ý nghĩa. Để đánh giá mức độ thích hợp của mô hình hồi quy, nghĩa là xem mô hình hồi quy giải thích được bao nhiêu % sự thay đổi của biến phụ thuộc Y ta dùng hệ số xác định R^2. Vì vậy, với mẫu cụ thể, khi nhận được kết quả với một mô hình hồi quy ta quan tâm đến việc đánh giá xem hệ số xác định R^2 có khác không có ý nghĩa thống kê hay không. Điều này có nghĩa là ta cần kiểm định giả thuyết H_0: $R^2 = 0$, với đối thuyết H_1: $R^2 > 0$.

Trong trường hợp mô hình hồi quy hai biến, giả thuyết H_0: $R^2 = 0$ có nghĩa là biến giải thích X không ảnh hưởng đến biến phụ thuộc Y, tức là tương đương với điều kiện: $b = 0$. Để kiểm định giả thuyết H_0: $R^2 = 0$, người ta thường dùng hai phương pháp: Phương pháp giá trị tới hạn và phương pháp p – value như sau:

2.6.3.1. Phương pháp giá trị tới hạn

Tiêu chuẩn bác bỏ giả thuyết H_0: $W = \{F > F_\alpha(1, n-2)\}$ (2.32)

($F_\alpha(1, n-2)$ *là giá trị tới hạn của phân phối* F, *bậc tự do* $(1, n-2)$);

$$F = \frac{ESS/1}{RSS/(n-2)} = \frac{R^2.(n-2)}{1-R^2} \quad (2.33)$$

Bước 1: Tra bảng giá trị tới hạn của phân phối F để tìm $F_\alpha(1, n-2)$

Bước 2: Dựa vào số liệu, tính F và so sánh với giá trị tra bảng $F_\alpha(1, n-2)$:

- Nếu W xảy ra thì bác bỏ H_0, chấp nhận H_1.
- Nếu W không xảy ra thì chấp nhận H_0, bác bỏ H_1.

Lưu ý: Giá trị của thống kê F có thể được cho bởi các phần mềm ứng dụng.

2.6.3.2. Phương pháp giá trị p – value

Tiêu chuẩn bác bỏ giả thuyết H_0: $p - value < \alpha$ (2.34)

Với: $p - value = P(F > F_0)$, F *là phân phối Fisher với 2 bậc tự do* $(1, n-2)$;

$$F_0 = \frac{R^2.(n-2)}{1-R^2}.$$

Bước 1: Từ mẫu điều tra, tính F_0.

Bước 2: Tính $p - value = P(F > F_0)$ và so sánh với α

Lưu ý: Giá trị của thống kê F và p – value của thống kê F có thể được cung cấp bởi các phần mềm ứng dụng.

Ví dụ 7: Xét mẫu 1 về tiêu dùng Y và thu nhập X, với mức ý nghĩa 5%, ta muốn xác minh xem mô hình SRF: $\begin{cases} \hat{Y} = 20{,}55831 + 0{,}569657.X \\ Y = \hat{Y} + \hat{U} = 20{,}55831 + 0{,}569657.X + \hat{U} \end{cases}$ có phù hợp (với mẫu điều tra) hay không.

Giải: Ta dùng phương pháp giá trị tới hạn: Tra bảng F ta có $F_\alpha(1, n-2) = F_{0,05}(1; 8) = 5{,}318$

Theo kết quả tính toán trước đây, ta có: $R^2 = \frac{ESS}{TSS} = \frac{10708{,}80022}{11114{,}4} = 0{,}9635 \Rightarrow F_0 = \frac{R^2.(n-2)}{1-R^2} = \frac{8\,.0{,}9635}{1-0{,}9635} = 211{,}178 > F_{0,05}(1; 8) = 5{,}318$. Vậy ta bác bỏ giả thuyết H_0 và cho rằng mô hình SRF phù hợp với mẫu điều tra.

2.6.4. Một số chú ý trong kiểm định giả thuyết về mô hình

a. Khi giải quyết bài toán kiểm định về mô hình, nếu không nói gì về mức ý nghĩa α thì nhầm định α = 5%. Chẳng hạn khi kiểm định về hệ số hồi quy:

– $Đối\ với\ kiểm\ định\ 2\ phía: nếu\ p - value < 0{,}05\ thì\ bác\ bỏ\ H_0$;

– $Đối\ với\ kiểm\ định\ một\ phía: nếu\ p - value < 0{,}1\ thì\ bác\ bỏ\ H_0$.

b. Như đã biết trong lý thuyết Xác suất- Thống kê: Khi ta chấp nhận H_0 thì không có nghĩa là H_0 đúng hoàn toàn, khi ta bác bỏ H_0 thì không có nghĩa là H_0 sai hoàn toàn. Sai lầm khi ta bác bỏ giả thuyết H_0 mà thực tế nó đúng gọi là sai lầm loại 1, sai lầm khi ta chấp nhận giả thuyết H_0 mà thực tế nó sai gọi là sai lầm loại 2. Xác suất sai lầm loại 1 chính là p – value.

c. Việc xác lập giả thuyết H_0 và đối thuyết H_1 không được đưa ra tùy tiện mà phải dựa vào bản chất của các mối quan hệ giữa các biến và yêu cầu của bài toán. Tránh khuynh hướng gò ép giả thuyết, đối thuyết để biện minh cho kết quả thực nghiệm đang tiến hành.

d. Phân biệt ý nghĩa thống kê và ý nghĩa thực tế của các đánh giá, kết luận: Chẳng hạn trong thống kê, một sự sai khác giữa giá trị ước lượng với giá trị thực có thể xem là bé, nhưng trong thực tế sự sai khác đó lại không nhỏ và rất đáng kể. Ví dụ giả sử sai số giữa ước lượng của hệ số hồi quy b so với giá trị thực là 0,05, về mặt thống kê có thể xem là bé, nhưng trong kinh tế lại không nhỏ, thậm chí là sai số đáng kể khi đó là mức tăng trưởng GDP (*Gross Domestic Product*: Tổng sản phẩm quốc nội) của một quốc gia.

e. Khi thực hiện các kiểm định giả thuyết về mô hình, các giá trị cần thiết như: sai số chuẩn của hồi quy, sai số chuẩn của các hệ số hồi quy ước lượng, hệ số xác định, giá trị của thống kê **t**, giá trị của thống kê **F** và các giá trị p- value tương ứng,... được chỉ ra trong bảng kết quả hồi quy(bảng *Equation*) của phần mềm Eviews.

2.6.5. Mô hình hồi quy với việc thay đổi đơn vị đo của biến

Vấn đề đặt ra là: khi thay đổi đơn vị đo của các biến, ta có cần thiết lập lại từ đầu mô hình hồi quy hay không?

Giả sử mô hình hồi quy SRF của Y theo X là: $Y = \hat{a} + \hat{b}.X + \hat{U}$ (2.35)

Đặt Y' = k.Y, X' = h. X, khi đó mô hình hồi quy SRF của Y' theo X' là:

$$Y' = \widehat{a'} + \widehat{b'}.X' + \widehat{U'}. \quad (2.36)$$

trong đó $\hat{a}, \hat{b}, \widehat{a'}, \widehat{b'}$ tìm được theo phương pháp OLS. Từ các công thức của $\hat{a}, \hat{b}, \widehat{a'}, \widehat{b'}$ ta có:

$$\widehat{a'} = k.\hat{a}; \quad \widehat{b'} = \frac{k}{h}.\hat{b} \quad (2.37)$$

Vì vậy mô hình hồi quy SRF của Y' theo X' là:

$$Y' = k.\hat{a} + \frac{k}{h}.\hat{b}.X' + \widehat{U'} \quad (2.38)$$

Điều này có nghĩa là sau khi dùng phép đổi biến Y' = k.Y, X' = h. X nói chung và đổi đơn vị đo cho các biến nói riêng, ta không cần thiết lập lại từ đầu mô hình hồi quy: Từ mô hình (2.35) của Y theo X, ta suy ra mô hình hồi quy SRF của Y' theo X' là (2.38). Ngoài ra ta có hệ thức:

$$\widehat{\sigma'}^2 = k.\hat{\sigma}^2;\ R^2_{X'Y'} = R^2_{XY};\ var(\widehat{a'}) = k^2.var(\hat{a});\ var(\widehat{b'}) = \frac{k^2}{h^2}.var(\hat{b}) \quad (2.39)$$

Việc thay đổi đơn vị đo của các biến không ảnh hưởng đến những tính chất của các ước lượng nhận được theo phương pháp OLS.

Ví dụ 8: Với một mẫu điều tra về mức thu nhập X (USD) và mức tiêu dùng Y (USD) gồm 10 hộ gia đình từ tổng thể 60 hộ trong ví dụ trước đây ở chương 1, ta có các số liệu sau:

X	80	100	120	140	160	180	200	220	240	260
Y	60	78	90	108	114	132	138	144	150	174

Hãy thiết lập SRF tuyến tính mô tả sự phụ thuộc của Tiêu dùng tính theo EUR và thu nhập tính theo ngàn VNĐ, biết 1 USD = 20 ngàn VNĐ, 1 EUR = 1,2 USD.

Giải: Từ số liệu ta tính được: $\hat{b} = 0{,}578182;\ \hat{a} = 20{,}50909$

SRF tuyến tính của Y theo X là:

$\hat{Y} = 20{,}50909 + 0{,}578182.X$ (*)

Gọi X' là mức thu nhập hàng tuần của một hộ tính theo ngàn VNĐ, Y' là mức tiêu dùng hàng tuần của một hộ tính theo EUR.

- Nếu chuyển số liệu trên sang cho X', Y' ta có bảng số liệu:

X'	1600	2000	2400	2800	3200	3600	4000	4400	4800	5200
Y'	50	65	75	90	95	110	115	120	125	145

Tính trực tiếp ta có: $\widehat{b'} = 0{,}024091;\ \widehat{a'} = 17{,}09091$

SRF tuyến tính của Y' theo X' là: $\widehat{Y'} = 17{,}09091 + 0{,}024091.X'$ (**)

- Nếu dùng công thức đổi đơn vị đo (2.37), từ giả thiết:

$$Y' = \frac{1}{1{,}2}.Y; X' = 20.X, \text{tức là } k = \frac{1}{1{,}2}, h = 20$$

ta có: $$\widehat{a'} = k.\hat{a} = 17{,}09091;\ \widehat{b'} = \frac{k}{h}\hat{b} = 0{,}024091$$

Tức là ta nhận lại đúng như kết quả tính trực tiếp (**)

Nhận xét: Để nhận được kết quả (2.38), ta chỉ cần thay trong (2.37): $Y = \frac{1}{k}Y'; X = \frac{1}{h}X'$.

2.7. Trình bày kết quả hồi quy

Chúng ta chỉ có thể thực hiện bằng cách tính tay trong một số trường hợp đơn giản.Nói chung chúng ta phải thực hiện các bước tính toán nhờ vào các phần mềm hỗ trợ như: Eviews (Econometrics Views), Rats (Regression Analysis Tempral Series).

a/ Trường hợp đơn giản, khi không dùng phần mềm ứng dụng để chạy hồi quy, cần trình bày các kết quả tính:

- Các hệ số hồi quy và hồi quy ước lượng SRF:

$$\hat{b} = \frac{\overline{XY} - \bar{X}\bar{Y}}{S^2(X)};\ \hat{a} = \bar{Y} - \hat{b}.\bar{X};\ \hat{Y} = \hat{a} + \hat{b}.X$$

- Các tổng bình phương các độ lệch: TSS, ESS, RSS

$$TSS = n.S^2(Y); ESS = n\hat{b}.S^2(X); RSS = TSS - ESS$$

- Hệ số $R^2, \hat{\sigma}$: $R^2 = \frac{ESS}{TSS} = \frac{S^2(X)}{S^2(X)}.\hat{b}$; $\hat{\sigma} = \sqrt{\frac{RSS}{n-2}}$

- Sai số chuẩn của các hệ số hồi quy (nếu cần): $\widehat{se}(\hat{a})$, $\widehat{se}(\hat{b})$, ...

$$\widehat{se}(\hat{a}) = \frac{\hat{\sigma}}{S(X)}.\sqrt{\frac{\overline{X^2}}{n}};\ \widehat{se}(\hat{b}) = \frac{\hat{\sigma}}{S(X)\sqrt{n}}$$

- Các g.trị của thống kê t và F (nếu cần): $t = \frac{\hat{b}}{\widehat{se}(\hat{b})}; F = \frac{R^2.(n-2)}{1-R^2}$

Chẳng hạn trong ví dụ trên, ta có kết quả hồi được tính toán trực tiếp và trình bày như sau:

* $\hat{b} = 0{,}578182;\ \hat{a} = 20{,}50909;\ \hat{Y} = 20{,}50909 + 0{,}578182.X;$
* $RSS = 257{,}8909; TSS = 11289{,}60013; ESS = 11031{,}70923;$
* $R^2 = \frac{ESS}{TSS} = 0{,}977157;\ \hat{\sigma} = \sqrt{\frac{RSS}{n-2}} = 5{,}677708;$
* $\widehat{se}(\hat{a}) = 5{,}608465;\ \widehat{se}(\hat{b}) = 0{,}031255;$
* $t = \frac{\hat{b}}{\widehat{se}(\hat{b})} = 3{,}656810; F = \frac{R^2.(n-2)}{1-R^2} = 342{,}2132$

b/ Với sự trợ giúp của Eviews, các kết quả của việc phân tích hồi quy được chỉ ra các thông tin trong bảng dưới đây:

Dependent Variable: ...
Method: Least Squares
Date: ..... Time: .....
Sample: .....
Included observations: .....

Variable	Coefficient	Std. Error	t-Statistic	Prob.
...	.....	.....	.....	.....
...	.....	.....	.....	.....

R-squared	.....	Mean dependent var	.....
Adjusted R-squared	.....	S.D. dependent var	.....
S.E. of regression	.....	Akaike info criterion	.....
Sum squared resid	.....	Schwarz criterion	.....
Log likelihood	.....	Hannan-Quinn criter.	.....
F-statistic	.....	Durbin-Watson stat	.....
Prob(F-statistic)	.....		

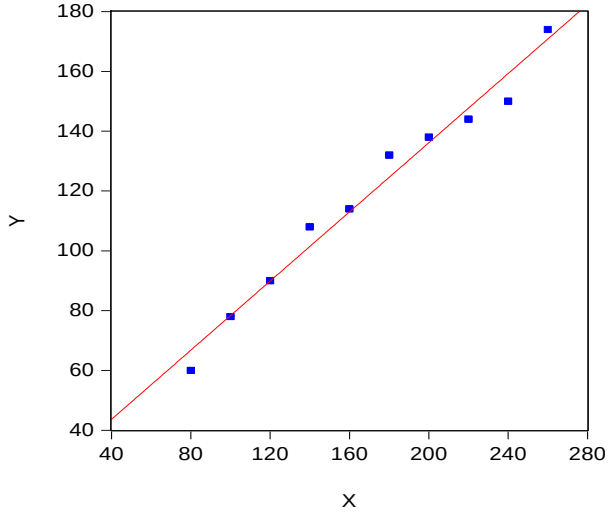
Bảng 2.2

Chú giải:
* *Dependent Variable*: Biến phụ thuộc
* *Method: Least Squares*: Phương pháp (ước lượng): Phương pháp bình phương bé nhất
* *Date*:..... *Time*:...: ngày...giờ (thực hiện)
* *Sample*:...: Mẫu sử dụng hay phạm vi quan sát được sử dụng
* *Included observations*: ...: Tổng số quan sát trong mẫu thực hiện
* Cột *Variable* : cho biết danh sách các biến giải thích trong mô hình. Lưu ý là C dùng để chỉ hằng số trong hàm hồi quy tương ứng với hằng số a, cũng được coi là một biến (biến hằng).
* Cột *Coefficient*: Cho biết giá trị của các hệ số hồi quy ước lượng $\hat{a}$ và $\hat{b}$ tương ứng với C và X
* Cột *Std. Error*: cho biết giá trị của các sai số chuẩn: $\widehat{se}(\hat{a})$, $\widehat{se}(\hat{b})$
* Cột *t-Statistic*: cho biết giá trị của thống kê t ứng với giả thuyết tham số hồi quy = 0 (lấy cột Coefficient chia cho cột Std. Error)
* Cột *Prob.* : cho biết giá trị p – value của thống kê **t** tương ứng.
* *R-squared*: hệ số xác định R^2; Mean dependent var.: trung bình mẫu của biến phụ thuộc ($\bar{Y}$)
* *Adjusted R-squared* : Hệ số xác định điều chỉnh $\bar{R}^2$
* *Sum Squared resid* : RSS (tổng bình phương các phần dư)
* *Log likelihood* : Ln hàm hợp lý
* *Durbin – Watson stat*: Thống kê Durbin – Watson
* *S.D. dependent var.*:Độ lệch mẫu điều chỉnh của biến phụ thuộc Y (S'(Y))
* *S.E. of regression*: sai số chuẩn của hàm hồi quy: $\hat{\sigma}$,
* *Akaike info criterion*: tiêu chuẩn Akaike
* *Schwarz criterion*: Tiêu chuẩn Schwarz
* *F- statistic*: thống kê F
* *Hannan-Quinn criterion*: Tiêu chuẩn Hannan-Quinn
* *Prob(F- statistic)*: Xác suất P(F > F- statistic) (p-value của thống kê F)

Chẳng hạn, với số liệu trong ví dụ trên về Tiêu dùng Y (USD) và thu nhập X (USD) của 10 hộ, Eviews cung cấp biểu đồ phân tán (hình bên) của Y theo X. Các điểm phân tán rất gần xung quanh một đường thẳng. Đây là cơ sở trực quan để ta nhận dạng hồi quy PRF của Y theo X là dạng bậc nhất

$$\hat{Y} = a + b.X$$

và kết quả hồi quy được cho bởi Eviews như sau và chúng ta có thể thấy được sự trùng hợp kết quả giữa hai cách: tính toán trực tiếp và sử dụng phần mềm ứng dụng



Dependent Variable: Y
Method: Least Squares
Date: 06/24/15 Time: 12:00
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	20.50909	5.608465	3.656810	0.0064
X	0.578182	0.031255	18.49901	0.0000

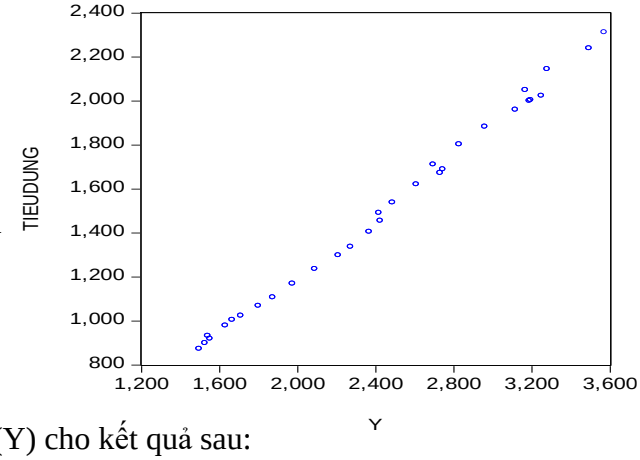
R-squared	0.977157	Mean dependent var	118.8000
Adjusted R-squared	0.974301	S.D. dependent var	35.41751
S.E. of regression	5.677708	Akaike info criterion	6.487829
Sum squared resid	257.8909	Schwarz criterion	6.548346
Log likelihood	-30.43914	Hannan-Quinn criter.	6.421442
F-statistic	342.2132	Durbin-Watson stat	1.562483
Prob(F-statistic)	0.000000		

Ví dụ 9: Các số liệu về thu nhập (Y) và tiêu dùng (C) trong khoảng thời gian từ năm 1958 đến năm 1988 được cho ở bảng dưới đây. Sử dụng phần mềm Eviews để chạy hồi quy của thu nhập Y theo tiêu dùng C (lưu ý là khi khai biến tiêu dùng, hoặc ta để nguyên tên tiêu dùng, hoặc ta dùng một ký tự khác C (vì ký tự C mặc định là hệ số bị chặn trong mô hình):

Năm	C	Y	Năm	C	Y	Năm	C	Y
1958	873.8	1494.9	1969	1298.9	2208.4	1980	1883.7	2958.7
1959	899.8	1525.7	1970	1337.7	2271.3	1981	1960.9	3115.2
1960	919.7	1551.1	1971	1405.8	2365.6	1982	2004.4	3192.3
1961	932.9	1539.3	1972	1456.6	2423.3	1983	2000.4	3187.2
1962	979.3	1629.1	1973	1492	2416.2	1984	2024.2	3248.7
1963	1005.1	1665.2	1974	1538.7	2484.8	1985	2050.7	3166
1964	1025.1	1708.7	1975	1621.8	2608.5	1986	2145.9	3277.6
1965	1069	1799.4	1976	1689.6	2744	1987	2239.9	3492
1966	1108.3	1873.3	1977	1674	2729.3	1988	2313	3570
1967	1170.6	1973.3	1978	1711.9	2695			
1968	1236.3	2087.6	1979	1803.9	2826.7			

Bảng 2.3

a/ Với mẫu điều tra này, Eviews cho ta biểu đồ phân tán của TIEUDUNG theo Y(thu nhập) sau: Biểu đồ này cho thấy các điểm quan sát thực nghiệm rất gần một đường thẳng, đây là hình ảnh trực quan cho phép ta nhận dạng hồi quy của TIEUDUNG theo thu nhập Y là tuyến tính, nên mô hình kinh tế lượng ở đây được nhận dạng là:



$$\begin{cases} E(TIEUDUNG|Y) = a + b.Y \\ TIEUDUNG = a + bY + U \end{cases}$$

b/ Hồi quy tiêu dùng (C) theo thu nhập (Y) cho kết quả sau:

Dependent Variable: TIEUDUNG
Method: Least Squares
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-161.5118	22.37920	-7.217049	0.0000
Y	0.684186	0.008848	77.32386	0.0000

R-squared	0.995173	Mean dependent var	1512.061
Adjusted R-squared	0.995007	S.D. dependent var	448.3518
S.E. of regression	31.68220	Akaike info criterion	9.811728
Sum squared resid	29109.09	Schwarz criterion	9.904243
F-statistic	5978.979	Durbin-Watson stat	0.683880
Prob(F-statistic)	0.000000		

Bảng 2.4

Từ bảng 2.4, ta có: hàm hồi quy SRF ước lượng của *TIEUDUNG* theo Y là:

$$TIEUDUNG = -161{,}5118 + 0{,}684186.Y + \widehat{U}$$

$R^2 = 0{,}995173$,

$\hat{b} = 0.684186 > 0$ *nên* $r_{TIEUDUNG,\widehat{TIEUDUNG}} = \sqrt{R^2} = \sqrt{0.995173} \approx 0.9976$, cho thấy mô hình phù hợp rất tốt với số liệu điều tra, biến X giải thích được 99.52% sự thay đổi của biến Y; 0.48% còn lại là do tác động của nhiễu ngẫu nhiên không đưa vào mô hình.

Hệ số $\hat{b} = 0{,}684186$ cho thấy khi thu nhập Y tăng 1(đơn vị tiền tệ) thì bình quân *TIEUDUNG* tăng 0,684186 (đơn vị tiền tệ).

c/ Tiến hành ước lượng khoảng tin cậy cho các hệ số hồi quy a, b, Eviews cho kết quả sau:

Coefficient Confidence Intervals
Included observations: 31

		90% CI		95% CI		99% CI	
Variable	Coefficient	Low	High	Low	High	Low	High
C	-161.5118	-199.5368	-123.4867	-207.2823	-115.7412	-223.1975	-99.82605
Y	0.684186	0.669152	0.699221	0.666090	0.702283	0.659797	0.708576

Bảng 2.5

Theo đó, các khoảng tin cậy tương ứng với các độ tin cậy 90%, 95%, 99%:

- cho a là: (-199.5368 , -123.4867); (-207.2823, -115.7412); (-223.1975, -99.82605)
- cho b là: (0.669152, 0.699221); (0.666090, 0.702283); (0.659797, 0.708576)

d/ Tiến hành kiểm định các giả thuyết: $H_0: a = 0$; $H_0: b = 0$ bằng phương pháp p – value, căn cứ vào cột **Prob** trong bảng hồi quy, đối với cả a và b, ta đều thấy: p – value < 0,0001 << 0,01 nên ta bác bỏ cả hai giả thuyết trên, tức là thừa nhận cả a và b đều khác không một cách có ý nghĩa.

e/ Khoảng tin cậy cho phương sai nhiễu có dạng: $\left(\frac{(n-2).\hat{\sigma}^2}{\chi^2_{\alpha/2}} \; ; \; \frac{(n-2).\hat{\sigma}^2}{\chi^2_{1-\alpha/2}} \right)$

Từ bảng kết quả hồi quy, có $\hat{\sigma}^2 = 31.68220^2 \approx 1003{,}76180$; $n = 31$;

tra bảng với $\alpha = 0{,}05$, có: $\boldsymbol{\chi^2_{\frac{\alpha}{2}}} = 45{,}7223$; $\boldsymbol{\chi^2_{1-\frac{\alpha}{2}}} = 16{,}0471$. Từ đó:

$\frac{(n-2).\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}} = \frac{29.1003{,}7618}{45{,}7223} = 636{,}64978$; $\frac{(n-2).\hat{\sigma}^2}{\chi^2_{1-\alpha/2}} = \frac{29.1003{,}7618}{16{,}0471} = 1813{,}97836$

Khoảng tin cậy cần tìm cho phương sai nhiễu là: $(636{,}64978,\ 1813{,}97836)$

2.8. Một số ứng dụng của mô hình hồi quy tuyến tính

Trong mục này, ta giới thiệu một số mô hình phi tuyến có thể tuyến tính hóa được và những mô hình thực tế có liên quan.

2.8.1. Một số khái niệm cần thiết

2.8.1.1. Biên tế và hệ số co giãn

Giả sử đại lượng Y là hàm của đại lượng X: Y = f(X), khi đó các số gia $\boldsymbol{\Delta X}, \boldsymbol{\Delta Y}$ còn được gọi là các lượng thay đổi tuyệt đối của X và của Y và $\frac{\boldsymbol{\Delta X}}{\boldsymbol{X}}, \frac{\boldsymbol{\Delta Y}}{\boldsymbol{Y}}$ được gọi là lượng thay đổi tương đối của X và của Y.

* Ta gọi đại lượng sau đây là biên tế của Y theo X:

$$\boldsymbol{M_{YX}} = \boldsymbol{\Delta Y/\Delta X} \qquad (2.40)$$

Ta có: $\Delta Y = M_{YX}.\Delta X$, như vậy biên tế của Y theo X cho biết lượng thay đổi tuyệt đối của biến phụ thuộc Y khi biến độc lập thay đổi 1 đơn vị. Với giả thiết f(X) có đạo hàm, khi ΔX khá nhỏ ta có: $\boldsymbol{M_{YX} \approx f'(X)}$. (2.41)

* Hệ số co giãn của Y theo X là: $\boldsymbol{E_{YX}} = \frac{\boldsymbol{\Delta Y/Y}}{\boldsymbol{\Delta X/X}}$ (2.42)

Từ (2.42) suy ra: $\frac{\Delta Y}{Y} = E_{YX}.\frac{\Delta X}{X}$. Như vậy hệ số co giãn $\boldsymbol{E_{YX}}$ là lượng thay đổi (%) của biến phụ thuộc Y khi X thay đổi 1%.

Khi ΔX khá nhỏ ta có:

$$\boldsymbol{E_{YX} = \frac{\Delta Y/Y}{\Delta X/X} = \frac{\Delta Y}{\Delta X}.\frac{Y}{X} \approx f'(X).\frac{Y}{X}} \qquad (2.43)$$

Chú ý:

- Biên tế phụ thuộc vào các đơn vị đo của X và Y, nhưng hệ số co giãn thì không phụ thuộc vào đơn vị đo của các biến.

2.8.1.2. Mô hình hồi quy qua gốc tọa độ

Mô hình hồi quy qua gốc tọa độ là một trường hợp riêng của mô hình hồi quy tuyến tính với tung độ gốc a = 0. Hàm hồi quy qua gốc tọa độ có thể viết dưới dạng:

$$PRF: \begin{cases} E(Y|X) = b.X \\ Y = b.X + U \end{cases} ; \qquad SRF: \begin{cases} \hat{Y} = \hat{b}.X \\ Y = \hat{b}.X + \hat{U} \end{cases} \qquad (2.44)$$

trong đó, ước lượng $\hat{b}$ của b được tìm bằng phương pháp OLS,

hơn nữa ta có: $var(\hat{b}) = \frac{\sigma^2}{\sum X_i^2}$; σ^2 *có ước lượng*: $\hat{\sigma}^2 = \frac{\sum \hat{U}_i^2}{n-1} = \frac{RSS}{n-1}$

2.8.2. Một số mô hình tuyến tính hóa được:

Mục này giới thiệu một số mô hình hồi quy phi tuyến thường gặp mà bằng phép đổi biến thích hợp có thể đưa được về mô hình tuyến tính.

2.8.1.1. Mô hình tuyến tính Log

Xét mô hình: $Y = \gamma . X^b . e^U \ (\gamma > 0)$ (2.46)

Đây là một mô hình phi tuyến. tuy nhiên mô hình này có dạng tương đương:

$$lnY = a + b.lnX + U \ (với\ a = ln\gamma) \quad (2.47)$$

gọi là mô hình tuyến tính log. Đặt Y* = lnY, X* = lnX thì (2.47) có dạng:

Y* = a + b. X* + U

là mô hình hồi quy tuyến tính đối với các biến X*, Y* và nếu các giả thiết của mô hình hồi quy tuyến tính được thỏa mãn thì ta có thể tìm các ước lượng $\hat{a}$, $\widehat{b}$ cho a, b (và do đó có ước lượng $\hat{\gamma} = e^{\hat{a}}$ cho γ) bằng phương pháp OLS.

Với mô hình (2.47), ta có: $\frac{dY}{Y} = b.\frac{dX}{X}$, hệ số co giãn của mô hình là:

$$E_{Y/X} = \frac{dY/Y}{dX/X} = \frac{dY}{dX}.\frac{X}{Y} = b$$

Vì thế mô hình tuyến tính log, hay mô hình tuyến tính kép còn có các tên gọi: mô hình hệ số co giãn không đổi, mô hình log – log.

Ta biết rằng các ước lượng $\hat{a}$, $\widehat{b}$ tìm theo phương pháp OLS là các ước lượng không chệch cho a, b. Trong khi ước lượng $\hat{\gamma} = e^{\hat{a}}$ tương ứng lại là ước lượng chệch cho γ. Tuy nhiên trong thực tế, người ta chú ý nhiều đến vai trò của hệ số b, nên $\hat{\gamma} = e^{\hat{a}}$ là ước lượng chệch cho γ không phải là vấn đề đáng quan ngại.

2.8.1.2. Mô hình bán logarit (semi log)

Đó là mô hình chỉ có một biến xuất hiện dưới dạng logarit. Mô hình này được chia làm 2 dạng:

a. Mô hình log – lin: $\boldsymbol{lnY = a + b.X + U}$ (2.48)

Ta có: $M_{YX} = \frac{\Delta Y}{\Delta X} = \frac{e^{b.\Delta X}-1}{\Delta X}.Y \approx b.Y$; $E_{YX} = \frac{\frac{\Delta Y}{Y}}{\frac{\Delta X}{X}} \approx b.X$

Trong nghiên cứu thực nghiệm, mô hình dạng log-lin thích hợp cho các trường hợp như khảo sát tốc độ tăng trưởng hay suy thoái của các biến kinh tế tầm vĩ mô: lượng cung tiền, thâm hụt thương mại, năng suất, dân số, lao động, GDP, GNP,.... Cần lưu ý sự khác biệt giữa mô hình log-lin và mô hình có xu hướng tuyến tính là mô hình có dạng:

$$Y = a + b.t + U \quad (2.49)$$

Tùy thuộc vào việc ta quan tâm tới ước lượng thay đổi tương đối hay tuyệt đối của biến phụ thuộc theo thời gian mà lựa chọn mô hình nào. Nếu quan tâm tới lượng thay đổi tuyệt đối của biến phụ thuộc thì mô hình có xu hướng tuyến tính tỏ ra thích hợp hơn. Vì biến phụ thuộc xuất hiện dưới hai dạng khác nhau nên không thể so sánh hệ số xác định R^2 của hai mô hình này. Ngoài ra cần lưu ý là cả hai mô hình này chỉ thích hợp với các biến có số liệu chuỗi thời gian có tính chất dừng, tức là trung bình và phương sai của các biến này không phụ thuộc vào thời điểm quan sát đầu và thời điểm quan sát cuối mà chỉ phụ thuộc vào khoảng cách giữa hai thời điểm này.

Ví dụ 10: Xét công thức lãi suất gộp**:** $\boldsymbol{Y_t = Y_0.(1+r)^t}$

trong đó r là tốc độ tăng trưởng gộp theo thời gian của Y; Y_0 là giá trị của Y tại t = 0, Y_t là lãi suất gộp tại thời điểm t. Ta có: $\boldsymbol{lnY_t = a + b.t}$ $(a = lnY_0,\ b = ln(1+r))$

Đưa thêm vào sai số ngẫu nhiên, ta nhận được mô hình log-lin:

$$\boldsymbol{lnY_t = a + b.t + U_t}$$

Như vậy: $E_{T_t/t} = \frac{\frac{\Delta Y_t}{Y_t}}{\frac{\Delta t}{t}} = \frac{\Delta Y_t}{\Delta t}.\frac{t}{Y_t} \approx b.t = ln(1+r).t$ $(khi\ \Delta t\ khá\ bé)$: hệ số b > 0 (< 0) biểu thị tốc độ tăng trưởng (suy thoái) của lãi suất gộp Y_t.

b. Mô hình lin – log: $\boldsymbol{Y = a + b.lnX + U}$ (2.48)

Đây là mô hình tuyến tính (theo tham số).Từ Y = a + b. lnX, ta có:

$\frac{dY}{dX} = \frac{b}{X} \Rightarrow \frac{\Delta Y}{\Delta X} \approx \frac{b}{X} \Rightarrow \Delta Y/\frac{\Delta X}{X} \approx b$ $(khi\ \Delta X\ khá\ bé)\ hay$:

$\Delta Y(lượng\ thay\ đổi\ tuyệt\ đối\ của\ Y) \approx b.\frac{\Delta X}{X}$ $(khi\ \Delta X\ khá\ bé)$ (2.49)

Nếu thay đổi của X thể hiện bằng % $(100.\frac{\Delta X}{X})$ thì lượng thay đổi tuyệt đối của Y sẽ là:

$$\Delta Y \approx \frac{b}{100}.\left(100.\frac{\Delta X}{X}\right) = 0{,}01.b.\left(100.\frac{\Delta X}{X}\right)$$

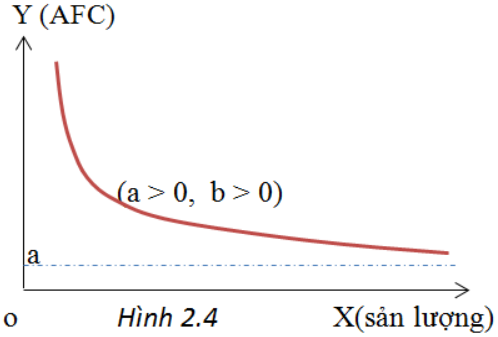
Do đó **nếu X thay đổi 1% thì Y thay đổi (0,01).b đơn vị**.

Người ta thường sử dụng mô hình lin – log trong trường hợp quan tâm đến lượng thay đổi tuyệt đối của biến phụ thuộc khi biến độc lập thay đổi 1%. Nó có thể được dùng để khảo sát một số quan hệ như: diện tích sử dụng của căn nhà tác động tới giá nhà, diện tích trồng trọt tác động tới sản lượng của cây trồng, lượng cung tiền ảnh hưởng tới GNP,...

c. Mô hình nghịch đảo: $Y = a + b.\frac{1}{X} + U$ (2.50)

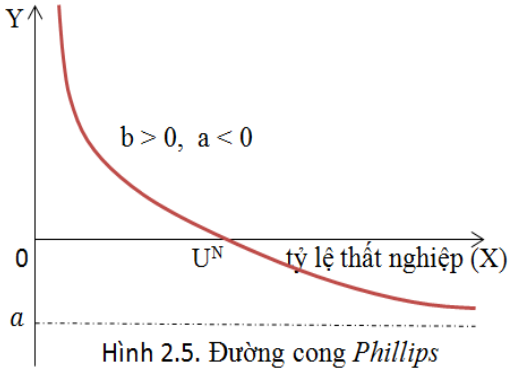
Trong thực tế, một số trường hợp có thể áp dụng mô hình nghịch đảo là:

c1. Mối quan hệ giữa Y là chi phí sản xuất cố định bình quân (AFC: *Average fixed cost*) và sản lượng X. Khi sản lượng tăng thì chi phí sản xuất cố định bình quân trên một sản phẩm có khuynh hướng giảm dần, nhưng không vượt quá một mức tối thiểu a. Quan hệ này được mô tả bằng đồ thị sau (hình 2.4):



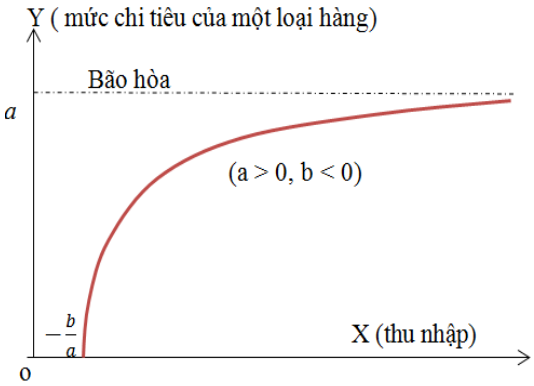
Hình 2.4

c2. Mối quan hệ giữa tỷ lệ thay đổi tiền lương Y và tỷ lệ thất nghiệp X, trong lý thuyết kinh tế được biểu diễn bằng đường cong Phillips (hình 2.5). Khi tỷ lệ thất nghiệp tăng, nhưng vẫn ở dưới mức thất nghiệp tự nhiên U^N thì tiền lương tăng (nhưng vẫn ở mức Y > 0), nhưng mức tăng có khuynh hướng giảm dần (biểu thị đường cong có hướng dốc xuống tiến về giá trị 0). Khi tỷ lệ thất nghiệp vượt quá mức tỷ lệ thất nghiệp tự nhiên U^N, tiền lương sẽ giảm (tương ứng Y < 0), nhưng mức giảm của tiền lương có khuynh hướng tăng dần (biểu thị đường cong càng xa dần giá trị 0) và tỷ lệ giảm sút tiền lương không vượt quá $|a|$.



Hình 2.5. Đường cong *Phillips*

c3. Quan hệ chi tiêu Y của người tiêu dùng đối với một loại hàng và tổng thu nhập X biểu diễn bằng đường cong Engel (hình 2.6) Theo Lý thuyết kinh tế: chi tiêu hàng hóa tăng khi thu nhập tăng. Tuy nhiên đối với một loại hàng hóa thì thu nhập của người tiêu dùng phải đạt không dưới một mức tối thiểu – b/a, mà người ta gọi là ngưỡng thu nhập hay thu nhập tới hạn, thì người tiêu dùng mới sử dụng loại hàng này.



Hình 2.6. Đường cong *Engel*

Mặt khác nhu cầu về loại hàng này là có hạn, tức là dù thu nhập có còn tăng lên bao nhiêu đi nữa thì người tiêu dùng cũng không tiêu thụ thêm loại hàng này nữa, đó là mức tiêu dùng bão hòa a của loại hàng này. Cần lưu ý rằng mối quan hệ thu nhập – tiêu dùng ở đây được xét đối với một loại hàng hóa có đặc điểm nhất định nào đó, chẳng hạn là mặt hàng xa xỉ, mà không phải là nhu yếu phẩm thông thường. Mô hình nghịch đảo thích hợp cho những trường hợp này. Nếu quan hệ là tổng chi tiêu và thu nhập thì mô hình tuyến tính theo biến tỏ ra thích hợp hơn.

- Bảng sau đây chỉ ra một số đặc tính cần lưu ý của các mô hình hồi quy phi tuyến hai biến thông dụng nói trên, trong đó để đơn giản cho trình bày, ta để ở dạng mô hình toán học, còn mô hình kinh tế lượng tương ứng phải cộng thêm thành phần sai số ngẫu nhiên, đồng thời công thức dẫn xuất từ biên tế là công thức gần đúng và chỉ có ý nghĩa khi X thay đổi nhỏ.

Mô hình	Dạng hàm	Hệ số góc	Hệ số co giãn	Ý nghĩa của hệ số góc
*Tuyến tính	Y = a + b.X	b	$b.\frac{X}{Y}$	**Lượng thay đổi của Y khi X tăng 1 đơn vị.**
*Tuyến tính log (log kép)	lnY = a + b.lnX .	b.Y/X .	b	**Khi X tăng 1% thì Y thay đổi b%.**
*Log –lin	lnY = a + b.X	b.Y	b.X	**Khi X tăng 1 đơn vị thì Y thay đổi 100b%**
*Lin –log	Y = a + b.lnX	b.(1/X)	$b.\frac{1}{Y}$	**Khi X tăng 1% thì Y thay đổi b/100 đơn vị.**
*Nghịch đảo	$Y = a + \frac{b}{X}$	$-b.\frac{1}{X^2}$	$-b.\frac{1}{XY}$	

Bảng 2.6

Lưu ý: Trong thực hành, đối với mô hình hồi quy ước lượng, trong công thức hệ số co giãn, hệ số góc, người ta thay giá trị của một biến bởi trung bình mẫu của biến đó

2.8.3. So sánh hệ số xác định giữa các mô hình

Một tiêu chí quan trọng để đánh giá sự phù hợp của hàm hồi quy là hệ số xác định R^2. Tuy nhiên khi có nhiều hàm hồi quy khác nhau thì ta nên chọn hàm hồi quy nào một khi ta dựa vào tiêu chuẩn R^2 lớn nhất? Với những mô hình hồi quy khác nhau, để so sánh các hệ số xác định, cần phải đảm bảo các yêu cầu chung sau đây:

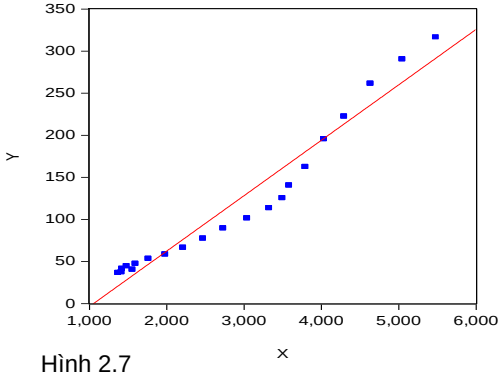
- Điều tra quan sát ở các mô hình có cùng cỡ mẫu.
- Các mô hình có cùng số biến độc lập. Nếu điều này không thỏa thì ta sẽ dùng hệ số xác định hiệu chỉnh $\bar{R}^2$ mà ta sẽ xác định sau.
- Các biến phụ thuộc xuất hiện trong các hàm hồi quy phải cùng dạng hoặc được đưa về cùng dang (điều này không yêu cầu đối với các biến giải thích)

Ví dụ 11: Có số liệu về số lượng máy điện thoại Y (ngàn cái) và X là GDP tính theo đầu người (tính theo đô la Singapore) từ 1960 đến 1981:

Năm	X	Y	Năm	X	Y	Năm	X	Y
1961	1365	37	1968	1974	59	1975	3575	141
1962	1409	38	1969	2204	67	1976	3784	163
1963	1549	41	1970	2462	78	1977	4025	196
1964	1416	42	1971	2723	90	1978	4286	223
1965	1473	45	1972	3033	102	1979	4628	262
1966	1589	48	1973	3317	114	1980	5038	291
1967	1757	54	1974	3487	126	1981	5472	317

Bảng 2.7 *Nguồn: D.N.Gujarati*

a/ Biểu đồ phân tán của Y theo X,
và kết quả hồi quy ước lượng cho mô hình
$Y = a + b.X + U$
được Eviews cung cấp như sau:



Hình 2.7

Dependent Variable: Y
Method: Least Squares
Sample: 1961 1981
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-69.28657	11.40038	-6.077568	0.0000
X	0.065862	0.003613	18.22724	0.0000

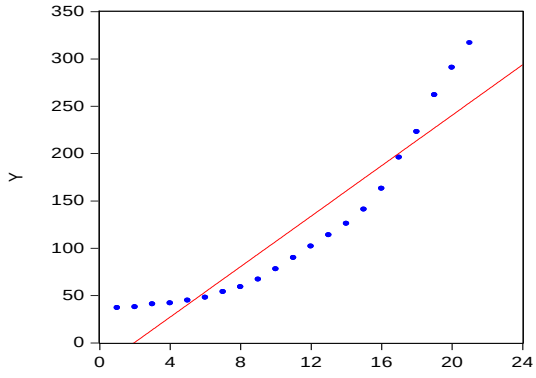
R-squared	0.945905	Mean dependent var	120.6667
Adjusted R-squared	0.943058	S.D. dependent var	88.76167
S.E. of regression	21.18084	Akaike info criterion	9.034464
Sum squared resid	8523.930	Schwarz criterion	9.133942
Log likelihood	-92.86187	Hannan-Quinn criter.	9.056053
F-statistic	332.2324	Durbin-Watson stat	0.155590
Prob(F-statistic)	0.000000		

Bảng 2.8

b/ Biểu đồ phân tán của Y theo T (biến xu thế: T = 1, ứng với năm 1961, T = 2 ứng với năm 1962,…) và kết quả hồi quy ước lượng cho mô hình:

$$Y = a + b.T + U$$

được Eviews cung cấp như sau:



Hình 2.8

Dependent Variable: Y
Method: Least Squares
Sample: 1961 1981
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-25.90476	14.99410	-1.727663	0.1003
T	13.32468	1.194127	11.15851	0.0000

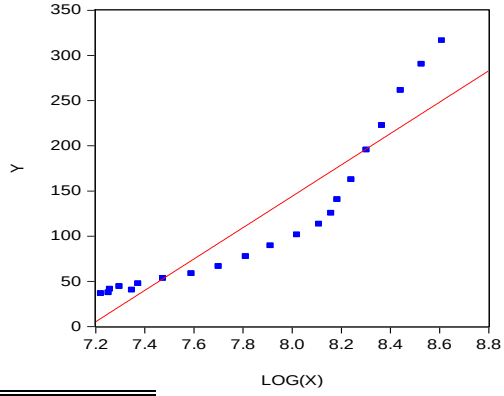
R-squared	0.867607	Mean dependent var	120.6667
Adjusted R-squared	0.860639	S.D. dependent var	88.76167
S.E. of regression	33.13568	Akaike info criterion	9.929491
Sum squared resid	20861.50	Schwarz criterion	10.02897
Log likelihood	-102.2597	Hannan-Quinn criter.	9.951081
F-statistic	124.5123	Durbin-Watson stat	0.120083
Prob(F-statistic)	0.000000		

Bảng 2.9

c/ Biểu đồ phân tán của Y theo ln(X) và kết quả hồi quy ước lượng cho mô hình:

$$Y = a + b.\ln(X) + U$$

được Eviews cung cấp như sau:



Hình 2.9

Dependent Variable: Y
Method: Least Squares
Sample: 1961 1981
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1244.184	133.2847	-9.334780	0.0000
LOG(X)	173.5521	16.91942	10.25756	0.0000

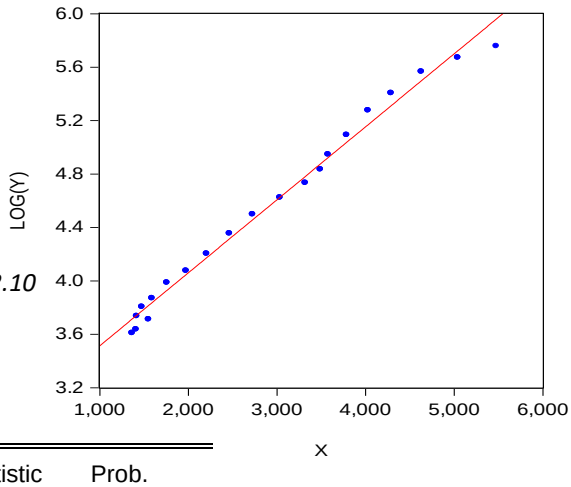
R-squared	0.847043	Mean dependent var	120.6667
Adjusted R-squared	0.838992	S.D. dependent var	88.76167
S.E. of regression	35.61630	Akaike info criterion	10.07388
Sum squared resid	24101.90	Schwarz criterion	10.17336
Log likelihood	-103.7757	Hannan-Quinn criter.	10.09547
F-statistic	105.2176	Durbin-Watson stat	0.135336
Prob(F-statistic)	0.000000		

Bảng 2.10

d/ Biểu đồ phân tán của lnY theo X
và kết quả hồi quy ước lượng cho mô hình:

$$\widehat{\ln(Y)} = a + b.X$$

được Eviews cung cấp như sau:



Hình 2.10

Dependent Variable: LOG(Y)
Method: Least Squares
Date: 06/27/15 Time: 13:10
Sample: 1961 1981
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.967819	0.042584	69.69393	0.0000
X	0.000547	1.35E-05	40.49965	0.0000

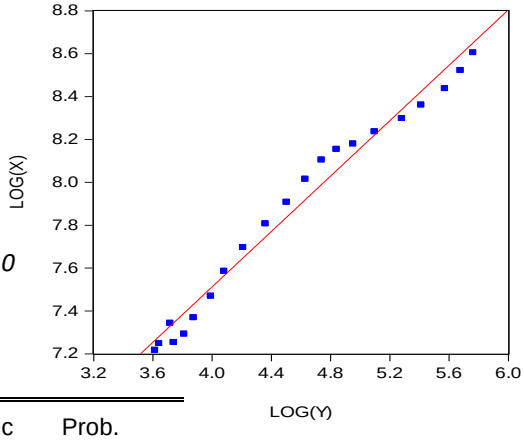
R-squared	0.988549	Mean dependent var	4.544341
Adjusted R-squared	0.987946	S.D. dependent var	0.720615
S.E. of regression	0.079116	Akaike info criterion	-2.145401
Sum squared resid	0.118929	Schwarz criterion	-2.045923
Log likelihood	24.52671	Hannan-Quinn criter.	-2.123812
F-statistic	1640.222	Durbin-Watson stat	0.524373
Prob(F-statistic)	0.000000		

Bảng 2.11

e/ Biểu đồ phân tán của ln(Y) theo ln(X)
và kết quả hồi quy ước lượng cho mô hình:

$$\widehat{\ln(Y)} = a + b.\ln(X)$$

được Eviews cung cấp như sau:



Hình 2.10

Dependent Variable: LOG(Y)
Method: Least Squares
Sample: 1961 1981
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7.366443	0.403580	-18.25275	0.0000
LOG(X)	1.514555	0.051231	29.56312	0.0000

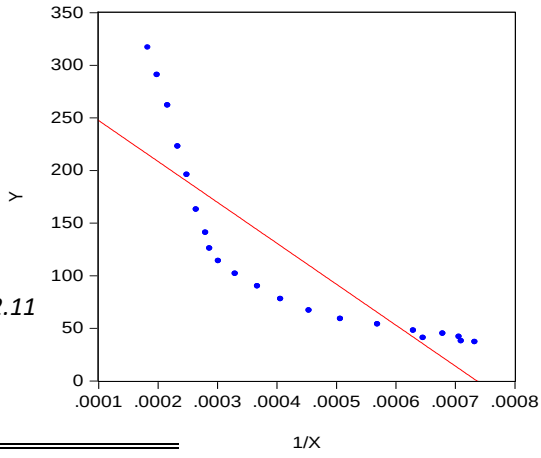
R-squared	0.978723	Mean dependent var	4.544341
Adjusted R-squared	0.977603	S.D. dependent var	0.720615
S.E. of regression	0.107845	Akaike info criterion	-1.525860
Sum squared resid	0.220978	Schwarz criterion	-1.426381
Log likelihood	18.02153	Hannan-Quinn criter.	-1.504270
F-statistic	873.9783	Durbin-Watson stat	0.313608
Prob(F-statistic)	0.000000		

Bảng 2.11

k/ Biểu đồ phân tán của Y theo 1/X
và kết quả hồi quy ước lượng cho mô hình:

$$\hat{Y} = a + \frac{b}{X}$$

được Eviews cung cấp như sau:



Hình 2.11

Dependent Variable: Y
Method: Least Squares
Sample: 1961 1981
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	286.3879	25.71385	11.13749	0.0000
1/X	-389006.6	55153.22	-7.053198	0.0000

R-squared	0.723627	Mean dependent var	120.6667
Adjusted R-squared	0.709081	S.D. dependent var	88.76167
S.E. of regression	47.87532	Akaike info criterion	10.66547
Sum squared resid	43548.87	Schwarz criterion	10.76495
Log likelihood	-109.9874	Hannan-Quinn criter.	10.68706
F-statistic	49.74760	Durbin-Watson stat	0.135977
Prob(F-statistic)	0.000001		

Bảng 2.12

h/ Từ các biểu đồ và các kết quả hồi quy cho các mô hình a/, b/, c/ và k/, nhận thấy mô hình a/ có $R^2 = 0{,}945905$ là lớn nhất. Trong hai mô hình d/ và e/ thì mô hình d/ có hệ số $R^2 = 0{,}988549$ cao hơn. Kết hợp so sánh hệ số xác định và biểu đồ phân tán, ta chọn mô hình d/:

$$\ln(Y) = 2{,}967819 + 0{,}000547.X + \hat{U}$$

m/ Hệ số co giãn của Y theo X hoặc T trong các mô hình hồi quy ước lượng:
- Tính các giá trị thống kê của các biến X, Y, T, X.Y nhờ Eviews:

	Y	X	T	X*Y
Mean (Trung bình mẫu)	120.6667	2884.095	11.00000	455777.7
Median (Trung vị)	90.00000	2723.000	11.00000	245070.0
Maximum	317.0000	5472.000	21.00000	1734624.
Minimum	37.00000	1365.000	1.000000	50505.00
Std. Dev.(Độ lệch mẫu)	88.76167	1310.726	6.204837	501633.9
Skewness (Hệ số bất đối xứng)	0.964006	0.418868	6.34E-17	1.301942
Kurtosis (Hệ số nhọn)	2.667312	1.946650	1.794545	3.574871
Sum	2534.000	60566.00	231.0000	9571332.
Sum Sq. Dev.(Tổng các bình phương độ lệch)	157572.7	34360076	770.0000	5.03E+12
Observations (Cỡ mẫu)	21	21	21	21

* Với mô hình a/ $\hat{Y} = -69{,}28657 + 0{,}065862.X$,

$E_{Y/X} = \hat{b}.\frac{\bar{X}}{\bar{Y}} = 0{,}065862.\frac{2884{,}095}{120{,}6667} = 1{,}574190$

* Với mô hình b/ $\hat{Y} = -25{,}90476 + 13{,}32468.T$,

$E_{Y/T} = \hat{b}.\frac{\bar{T}}{\bar{Y}} = 13{,}32468.\frac{11}{120{,}6667} = 15{,}821043$

* Với mô hình c/ $\hat{Y} = -1244{,}184 + 173{,}5521.\ln(X)$,

$E_{Y/X} = \hat{b}.\frac{1}{\bar{Y}} = 173{,}5521.\frac{1}{120{,}6667} = 1{,}438277$

* Với mô hình d/ $\widehat{\ln(Y)} = 2{,}967819 + 0{,}000547.X$,

$E_{Y/X} = \hat{b}.\bar{X} = 0{,}000547.2884{,}095 = 1{,}577900$

* Với mô hình e/ $\widehat{\ln(Y)} = -7{,}366443 + 1{,}514555.ln(X)$,

$E_{Y/X} = \hat{b} = 1{,}514555$

* Với mô hình h/ $\hat{Y} = 286{,}3879 - \frac{389006{,}6}{X}$,

$E_{Y/X} = -\hat{b}.\frac{1}{\overline{XY}} = -\frac{389006{,}6}{455777{,}7} = -0{,}853501$

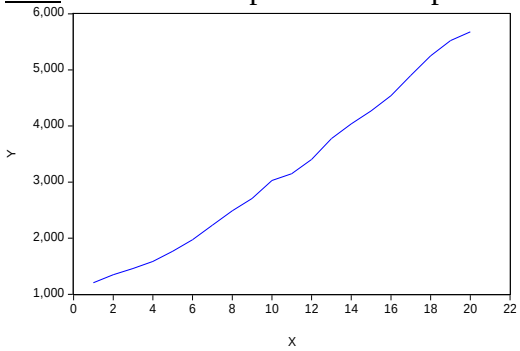
Ví dụ 12: Bảng sau cho biết tổng giá trị sản phẩm nội địa GDP(tỷ USD) của một quốc gia theo thời gian X(X=1 ứng với 1972, X = 2 ứng với 1973,..., X = 20 ứng với 1991) từ 1972 đến 1991:

X	GDP	X	GDP	X	GDP	X	GDP
1	1207	6	1974	11	3150	16	4540
2	1350	7	2233	12	3405	17	4900
3	1459	8	2489	13	3777	18	5251
4	1586	9	2708	14	4039	19	5522
5	1768	10	3030	15	4269	20	5678

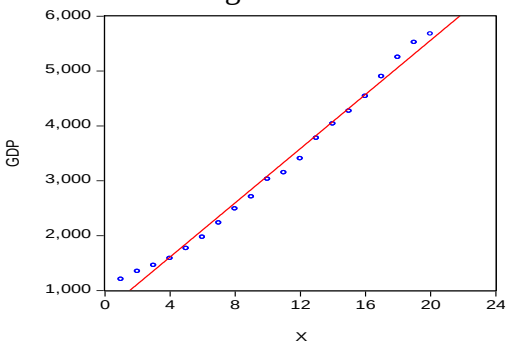
Bảng 2.13

a. Vẽ Line Graph của GDP theo thời gian X.
b. Vẽ biểu đồ phân tán của GDP theo X và cho nhận xét.
c. Hãy ước lượng mô hình: $GDP_t = a + b.X_t + U_t$. Nêu ý nghĩa của hệ số ước lượng $\hat{b}$.
d. Ước lượng mô hình trên với GDP tính theo đô la hiện hành trong giai đoạn 1972-1987. Sử dụng mô hình ước lượng này để dự báo GDP cho các năm 1988, 1989, 1990, 1991.
e. Vẽ Line Graph của GDP thực tế và GDP dự báo từ 1972 đến 1991.

Giải: a-b/ Line Graph và biểu đồ phân tán của GDP theo thời gian:



Hình 2.12a



Hình 2.12b

b/ Từ biểu đồ nhận thấy các điểm quan sát tập trung gần một đường thẳng, đó là cơ sở để nhận dạng hồi quy là: $GDP_t = a + b.X_t + U_t$.
c/ Chạy hồi quy của Y theo X, bảng kết quả nhận được dưới đây cho ta SRF ngẫu nhiên của GDP theo X: $GDP_t = 625.1053 + 246.8233.X_t + \widehat{U}_t$

Dependent Variable: GDP
Method: Least Squares
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	625.1053	71.44939	8.748924	0.0000
X	246.8233	5.964480	41.38220	0.0000

R-squared	0.989598	Mean dependent var	3216.750
Adjusted R-squared	0.989020	S.D. dependent var	1467.881
S.E. of regression	153.8096	Akaike info criterion	13.00395
Sum squared resid	425833.0	Schwarz criterion	13.10352
F-statistic	1712.487	Durbin-Watson stat	0.286667
Prob(F-statistic)	0.000000		

Bảng 2.14

$\hat{b} = 246.8233$ cho thấy đây là mức tăng bình quân hàng năm của GDP là 246.8233 tỷ USD.
d/ Chạy hồi quy của GDP theo X từ 1972 đến 1987, nhận được bảng kết quả hồi quy:

Dependent Variable: GDP
Method: Least Squares
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	736.6250	63.44092	11.61120	0.0000
X	229.3971	6.560904	34.96425	0.0000

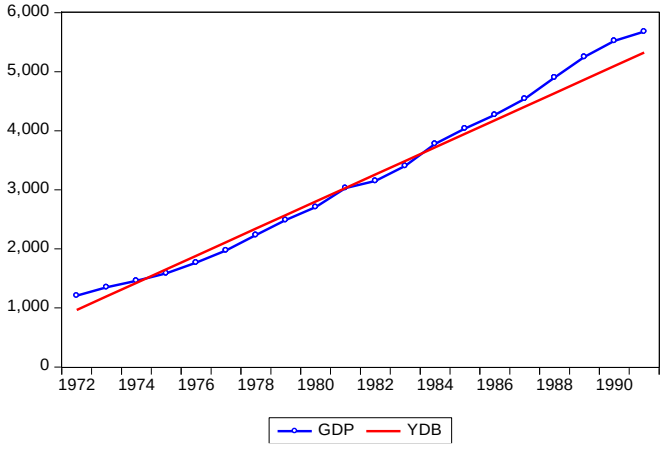
R-squared	0.988678	Mean dependent var	2686.500
Adjusted R-squared	0.987869	S.D. dependent var	1098.384
S.E. of regression	120.9771	Akaike info criterion	12.54555
Sum squared resid	204896.4	Schwarz criterion	12.64212
F-statistic	1222.498	Durbin-Watson stat	0.396588
Prob(F-statistic)	0.000000		

Bảng 2.15

SRF ngẫu nhiên của GDP theo X: $GDP_t = 736{,}6250 + 229{,}3971.X_t + \widehat{U}_t$
* Sử dụng mô hình ước lượng này để dự báo GDP cho các năm1988, 1989, 1990, 1991, có nghĩa là lần lượt thay X_t bởi các giá trị 17, 18, 19, 20 tương ứng với số thứ tự của các năm trên. Eviews cung cấp kết quả dự báo như sau:

Modified: 1972 1991 // fit(f=actual) ydb
4636.375 (dự báo GDP 1988)
4865.772 (dự báo GDP 1989)
5095.169 (dự báo GDP 1990)
5324.566 (dự báo GDP 1991)

e/ Line Graph của GDP thực tế và GDP dự báo từ 1972 đến 1991:



Hình 2.13

****

Bài tập.

1. Bảng số liệu sau là về Y là biến tổng sản phẩm được sản xuất của một ngành công nghiệp trong vòng 15 năm của một quốc gia và chi phí về vốn X1 và lao động X2 để sản xuất của ngành đó, với X3 là biến xu thế (biến thời gian)

Năm	Y	X1	X2	X3	Năm	Y	X1	X2	X3
1	8911.4	120.753	281.5	1	9	23052.6	153.714	616.7	9
2	10873.2	122.242	284.4	2	10	26128.2	164.783	695.7	10
3	11132.5	125.263	289.9	3	11	29563.7	176.864	790.8	11
4	12086.5	128.539	375.8	4	12	33367.6	188.146	816.0	12
5	12767.5	131.427	375.2	5	13	38354.8	205.841	848.4	13
6	16347.1	134.267	402.5	6	14	46868.3	221.748	873.4	14
7	19542.7	139.038	478.0	7	15	54308.9	239.715	999.2	15
8	21075.9	146.450	553.4	8					

a/ Vẽ biểu đồ phân tán của:
a1. Y theo X_1, Y theo X_2, Y theo X_3, Y theo X_3.
a2. lnY theo lnX_1, lnY theo lnX_2, lnY theo lnX_3.
b. Chạy hồi quy SRF cho các mô hình và cho biết hệ số xác định:
$Y = a + b.X_1 + U$, $Y = a + b.X_2 + U$, $Y = a + b.X_3 + U$.
$lnY = a + blnX_1 + U$, $lnY = a + blnX_2 + U$, $lnY = a + blnX_3 + U$.
c. Ước lượng khoảng tin cậy 95% cho các hệ số của X_1, X_2, X3 ở các mô hình trên.

2. Số liệu về tổng chi phí Y và sản lượng X được cho như sau:

X	1	2	3	4	5	6	7	8	9	10
Y	195	225	242	245	258	260	275	298	350	425

a. Vẽ biểu đồ phân tán và line Graph của Y theo X.
b. Từ số liệu, chạy hồi quy ước lượng SRF cho các mô hình:
*b*1. $Y = a + bX + U$, b2. $Y = a + blnX + U$,
b3. $lnY = a + bX + U$, b4. $lnY = a + blnX + U$.
c. Sử dụng các mô hình SRF ở trên để ước lượng tổng chi phí khi sản lượng X = 12.

3. Số liệu về lợi nhuận Y(tỷ VNĐ) và doanh thu X(tỷ VNĐ) của một số doanh nghiệp thuộc một ngành dịch vụ ở Tp. Hồ Chí Minh năm 2004 cho ở bảng sau:

Y	15	17	20	21	24	26	27	35
X	120	130	145	149	155	162	165	174

a/ Vẽ biểu đồ phân tán của Y theo X và cho nhận xét.
b/ Dựa vào bảng số liệu, sử dụng phương pháp OLS, hãy thiết lập mô hình hồi quy SRF ngẫu nhiên: $Y = \hat{a} + \hat{b}X + \hat{U}$ của lợi nhuận Y theo doanh thu X.

c/ Ước lượng khoảng tin cậy 95% cho các hệ số hồi quy trong mô hình: Y = a + bX + U.
d/ Ước lượng khoảng tin cậy 95% cho phương sai nhiễu trong mô hình: Y = a + bX + U.
e/ Dựa vào mẫu, hãy kiểm định giả thuyết về sự phù hợp của mô hình: Y = a + bX + U.
4. Giả sử có số liệu điều tra về lãi suất ngân hàng X(% /năm) và tổng vốn đầu tư Y(tỷ đồng) ở một địa phương A qua 10 năm liên tục như sau:

X	7,0	7,0	6,5	6,5	6,0	6,0	5,5	5,5	5,0	4,3
Y	29	32	32	34	34	36	45	47	50	54

a. Vẽ biểu đồ phân tán và đường hồi quy thực nghiệm của Y theo X.
b. Lập mô hình hồi quy tuyến tính SRF của Y theo X. Cho biết ý nghĩa thực tế của hệ số hồi quy ước lượng.
c. Với số liệu nói trên, xác minh Y có phụ thuộc thống kê vào X hay không.
d. Dựa vào số liệu trên, trong mô hình PRF, với độ tin cậy 90%, 95%, 99%, hãy chỉ ra các khoảng tin cậy tương ứng cho lượng thay đổi bình quân vốn đầu tư khi lãi suất tăng 1% (trong điều kiện các yếu tố ảnh hưởng khác không đổi)
e. Xác định khoảng tin cậy cho phương sai nhiễu với độ tin cậy 95%.
HD:
- SRF ngẫu nhiên của Y theo X: $Y = 96{,}76232 - 9{,}690104.X + \widehat{U}$
Hệ số hồi quy ước lượng: $\hat{b} = -9{,}690104$, cho thấy khi lãi suất ngân hàng tăng 1% (trong điều kiện các yếu tố khác không đổi) thì bình quân vốn đầu tư trên địa bàn địa phương A giảm 9,690104 tỷ đồng.
- Việc xác minh Y có phụ thuộc thống kê vào X hay không, chính là kiểm định giả thuyết về hệ số hồi quy b: $Giả\ thuyết\ H_0: b = 0, đối\ thuyết\ H_1: b \neq 0.$

5. Để ước lượng cho mô hình hồi quy bậc nhất PRF: Y = a + b.X + U, trong đó X là lãi suất ngân hàng (%/năm), Y là tổng vốn đầu tư (tỷ đồng) của một địa phương, từ mẫu điều tra, Eviews cho kết sau đây:

Dependent Variable: Y
Method: Least Squares
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	96.76232	6.255720	15.46781	0.0000
X	-9.690104	1.044814	-9.274481	0.0000

R-squared	0.914908	Mean dependent var	39.30000
Adjusted R-squared	0.904272	S.D. dependent var	8.832390
S.E. of regression	2.732742	Akaike info criterion	5.025344
Sum squared resid	59.74302	Schwarz criterion	5.085861
Log likelihood	-23.12672	Hannan-Quinn criter.	4.958957
F-statistic	86.01600	Durbin-Watson stat	1.612948
Prob(F-statistic)	0.000015		

Bảng 2.7

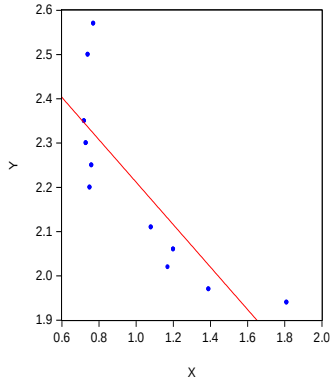
a/ Số liệu được sử dụng là loại số liệu gì? Viết SRF ước lượng cho PRF nói trên.
b/ Với mức ý nghĩa 5%, hãy cho biết mô hình SRF thu được có phù hợp với kết quả điều tra hay không.
c/ Hãy cho biết hệ số xác định và nêu ý nghĩa của nó.
d/ Tính các tổng bình phương các độ lệch: TSS, ESS, RSS.
e/ Sử dụng SRF nói trên để dự báo tổng vốn đầu tư của địa phương này khi lãi suất ngân hàng là 8%/năm.
g/ Hãy viết mô hình SRF tuyến tính của vốn đầu tư tính theo USD, với tỷ giá quy đổi là: 1 USD = 20000 VND, sử dụng công thức

HD: Sử dụng công thức đổi đơn vị đo, với $h = 1, k = \frac{1000000000}{20000} = 50000.$

6/ Từ bảng số liệu về các biến X, Y, ta có biểu đồ phân tán và bảng kết quả hồi quy sau:

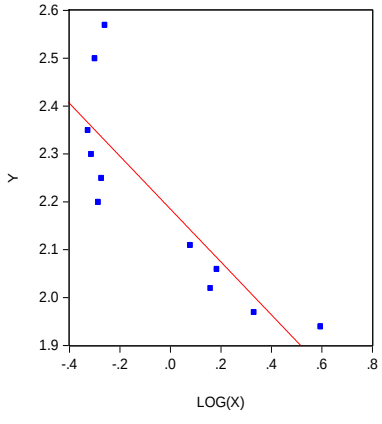
Dependent Variable: Y
Method: Least Squares
Sample: 1970 1980

Variable	Coefficient	Std. Error	Prob.
C	2.691124	0.121622	0.0000
X	-0.479529	0.114022	0.0023
R-squared	0.662757	2.206364	
Prob(F-statistic)	0.002288		



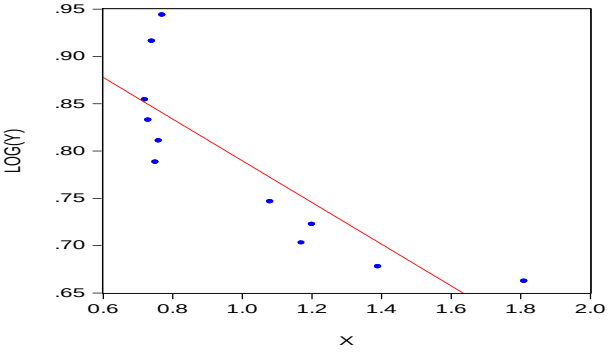
Dependent Variable: Y
Method: Least Squares
Sample: 1970 1980

Variable	Coefficient	Std. Error	Prob.
C	2.184839	0.036200	0.0000
LOG(X)	-0.552059	0.117262	0.0011
R-squared	0.711210		
Prob(F-statistic)	0.001108		



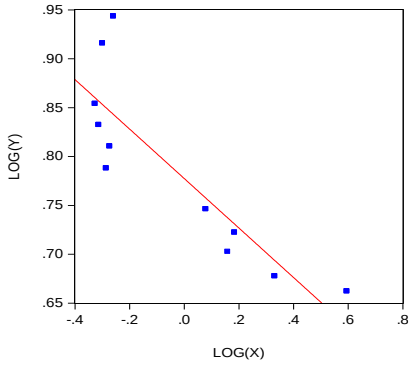
Dependent Variable: LOG(Y)
Method: Least Squares
Sample: 1970 1980

Variable	Coefficient	Std. Error	Prob.
C	1.009965	0.051630	0.0000
X	-0.220278	0.048403	0.0014
R-squared	0.697076		
Prob(F-statistic)	0.001384		



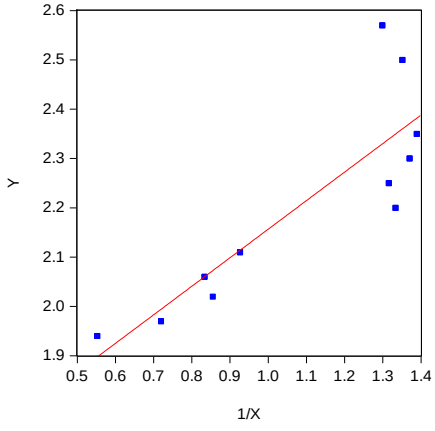
Dependent Variable: LOG(Y)
Method: Least Squares
Sample: 1970 1980

Variable	Coefficient	Std. Error	Prob.
C	0.777418	0.015242	0.0000
LOG(X)	-0.253046	0.049374	0.0006
R-squared	0.744800		
Prob(F-statistic)	0.000624		



Dependent Variable: Y
Method: Least Squares
Sample: 1970 1980

Variable	Coefficient	Std. Error	Prob.
C	1.577738	0.131788	0.0000
1/X	0.578952	0.117100	0.0008
R-squared	0.730893		
Prob(F-statistic)	0.000798		



a/ Từ kết quả trên, bạn chọn mô hình SRF nào? Tại sao?
b/ Với mô hình SRF đã được chọn, hãy dự báo giá trị của Y khi X = 2,0.

7. Bảng số liệu sau khảo sát về nhu cầu tiêu thụ cà phê ở Mỹ trong thời kỳ 1970 – 1980, trong đó X là giá một tách cà phê, Y là số tách mà một người dùng trong một ngày:

Năm	Y	X	Năm	Y	X	Năm	Y	X
1970	2.57	0.77	1974	2.25	0.76	1978	1.97	1.39
1971	2.5	0.74	1975	2.2	0.75	1979	2.06	1.2
1972	2.35	0.72	1976	2.11	1.08	1980	2.02	1.17
1973	2.3	0.73	1977	1.94	1.81			

a/ Hãy thiết lập SRF ước lượng cho mô hình PRF: $lnY = a + b.lnX + U$
b/ Từ SRF thiết lập được, giá trị của hệ số $\hat{b}$ ước lượng cho b nói lên điều gì?
c/ Hãy ước lượng khoảng tin cậy cho mức tăng hay giảm bình quân của lượng tiêu thụ (%) ở quốc gia này khi giá bán lẻ cà phê tăng 1%.

8. Theo số liệu của Tổng cục thống kê Việt nam về biến X là Tổng sản phẩm trong nước GDP và biến phụ thuộc Y(tỷ đồng) là chi tiêu tiêu dùng cá nhân trong thời kỳ 1995 – 2003, có bảng dưới đây:

Năm	X	Y	Năm	X	Y	Năm	X	Y
1995	195,567	142,916	1998	244,596	172,498	2001	292,535	190,577
1996	213,833	155,909	1999	256,272	176,976	2002	313,247	205,114
1997	231,264	165,125	2000	273,666	182,420	2003	336,243	221,545

a/ Hãy vẽ biểu đồ phân tán của Y theo X, cho nhận xét.
b/ Hãy thiết lập mô hình SRF ước lượng cho mô hình PRF: $Y = a + b.X + U$
9. Xét mô hình hồi quy Parabol: $Y = a + b.X^2 + U$
a/ Hãy xác định biên tế và hệ số co giãn của Y theo X.
b/ Sử dụng phương pháp bình phương tối thiểu thông dụng OLS để tìm các ước lượng $\hat{a}$, $\hat{b}$ cho a và b.
c/ Hãy thiết lập SRF: $Y = \hat{a} + \hat{b}.X^2 + \hat{U}$ (trong đó $\hat{a}$, $\hat{b}$ được tính bởi công thức tìm được ở b/) từ bảng số liệu sau:

X	1	2	3	4	5	6	7	8	9	10
Y	1,8	6	9,5	18	25	38	48	66	80	103

HD: Đặt X' = X^2
10/ Xét mô hình hồi quy $\frac{1}{lnY} = a + b.X + U$
a/ Hãy xác định biên tế và hệ số co giãn của Y theo X
b/ Sử dụng phương pháp OLS để tìm các ước lượng $\hat{a}$, $\hat{b}$ cho a và b.
c/ Thiết lập SRF: $\frac{1}{lnY} = \hat{a} + \hat{b}.X + \hat{U}$ (trong đó $\hat{a}$, $\hat{b}$ được tính bởi công thức tìm được ở b/) từ bảng số liệu sau:

X	1	2	3	4	5	6	7	8	9
Y	2	6	10	18	25	38	49	65	81

HD: Đặt Y' = 1/lnY.

****

Chương 3. HỒI QUY NHIỀU BIẾN

Trong thực tế, một đại lượng kinh tế không phải chỉ phụ thuộc vào một biến số kinh tế khác mà phụ thuộc vào nhiều biến số kinh tế khác nhau. Chẳng hạn nhu cầu về một loại hàng hóa không chỉ có phụ thuộc vào thu nhập của người tiêu dùng mà còn phụ thuộc vào nhiều yếu tố khác như: giá bán, thị hiếu của người tiêu dùng,.... Do đó mô hình hồi quy hai biến ở chương trước chưa đáp ứng được yêu cầu của thực tế. Chương này khảo sát mô hình hồi quy nhiều biến, tức là mô hình mà trong đó biến phụ thuộc được xét trong sự phụ thuộc vào hai hoặc nhiều hơn hai biến giải thích, cùng với các bài toán thống kê như ước lượng, kiểm định,…. Các ý tưởng, phương pháp và kết quả nghiên cứu hồi quy hai biến là cơ sở cho việc nghiên cứu mô hình hồi quy nhiều biến hay hồi quy bội.

3.1. Hàm hồi quy tổng thể và hàm hồi quy mẫu nhiều biến

3.1.1. Các khái niệm

Giả sử ta đang quan tâm đến véc tơ quan sát k chiều: (Y, X_1, X_2, ..., X_{k-1}), trong đó biến Y phụ thuộc vào k – 1 biến X_1, X_2, ..., X_{k-1} . Khi đó trung bình có điều kiện của Y với điều kiện véc tơ ngẫu nhiên X = (X_1, X_2, ..., X_{k-1}) là hàm của X = (X_1, X_2, ..., X_{k-1}):

$$E(Y|X) = E\big(Y\big|(X_1, X_2, \dots, X_{k-1})\big) = f(X) = f(X_1, X_2, \dots, X_{k-1})$$

Ta gọi hàm này là hàm hồi quy tổng thể PRF của Y theo X = (X_1, X_2, ..., X_{k-1}), hay PRF nhiều biến.

Như đã biết, hàm hồi quy xây dựng trên mẫu gọi là hàm hồi quy mẫu, viết tắt là SRF. Để hình dung được SRF, ta cần nhắc lại rằng: quan hệ giữa Y và X là phụ thuộc thống kê, ứng với mỗi giá trị $x = (x_1, x_2, \dots, x_{k-1})$ *của véc tơ* $X = (X_1, X_2, \dots, X_{k-1})$ không phải chỉ có một giá trị của Y, mà có cả một phân bố các giá trị của Y, nghĩa là có cả một biến quan sát mà ta ký hiệu là Y_x. Trung bình mẫu của biến Y_x là $\bar{Y}_x$ được gọi là trung bình mẫu có điều kiện của Y với điều kiện $X = (X_1, X_2, \dots, X_{k-1})$ *lấy giá trị* $x = (x_1, x_2, \dots, x_{k-1})$. Khi đó SRF của Y theo $X = (X_1, X_2, \dots, X_{k-1})$ là hàm của véc tơ ngẫu nhiên $= (X_1, X_2, \dots, X_{k-1})$, nhận giá trị là $\bar{Y}_x$ khi $X = (X_1, X_2, \dots, X_{k-1})$ *lấy giá trị* $x = (x_1, x_2, \dots, x_{k-1})$.

Ta vẫn dùng ký hiệu $\hat{Y}$ để chỉ hàm hồi quy mẫu, đó là một ước lượng của hàm hồi quy tổng thể PRF: $\hat{Y} = \hat{f}(X)$.

Ta đưa vào biến ngẫu nhiên U là tác động của những yếu tố ngẫu nhiên khác ngoài $X_1, X_2, \dots, X_{k-1}$ không được đưa vào, khiến cho giá trị của Y lệch khỏi $E(Y|X)$. Như vậy ta có mô hình sau đây gọi là mô hình PRF nhiều biến:

$$\begin{cases} E(Y|X) = f(X) = f(X_1, X_2, \dots, X_{k-1}) \\ Y = E(Y|X) + U \end{cases} \qquad (3.1)$$

Vẫn như trong hồi quy hai biến, ta gọi U là sai số ngẫu nhiên hay thặng dư. Ta có $\hat{U} = Y - \hat{Y}$ là một ước lượng của sai số ngẫu nhiên U.

Mô hình SRF nhiều biến là: $\begin{cases} \hat{Y} = \hat{f}(X) \\ Y = \hat{Y} + \hat{U} \end{cases}$ (3.2)

Trong phần tiếp theo của chương này, ta khảo sát mô hình hồi quy nhiều biến, tuyến tính theo biến và theo các tham số, tức là:

PRF: $\begin{cases} E(Y|X_1, X_2, \dots, X_{k-1}) = a_0 + a_1.X_1 + a_2.X_2 + \cdots + a_{k-1}.X_{k-1} \\ Y = a_0 + a_1.X_1 + a_2.X_2 + \cdots + a_{k-1}.X_{k-1} + U \end{cases}$ (3.3)

SRF: $\begin{cases} \hat{Y} = \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2 + \cdots + \hat{a}_{k-1} X_{k-1} \\ Y = \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2 + \cdots + \hat{a}_{k-1} X_{k-1} + \hat{U} \end{cases}$ (3.4)

với a_0: hệ số tự do; a_j: hệ số hồi quy riêng theo biến thứ $j, j = \overline{1, k-1}$

$\hat{a}_j$ là ước lượng của $a_j, j = \overline{1, k-1}$.

Hệ số a_j cho biết ảnh hưởng riêng của biến X_j lên trung bình có điều kiện của Y khi các biến còn lại không thay đổi. *Đó là lượng tăng (nếu* $a_j > 0$*) hay giảm (nếu* $a_j < 0$*) của biến phụ thuộc Y khi biến* X_j *tăng lên 1 đơn vị trong điều kiện các yếu tố khác không thay đổi*, hay: a_j *là lượng tăng hay giảm bình quân của biến phụ thuộc Y khi biến* X_j *tăng lên một đơn vị.*

$\hat{Y}$ là ước lượng của $E(Y|X_1, X_2, \dots, X_{k-1})$; $\hat{U}$ là ước lượng của U.

Giả sử ta có mẫu ngẫu nhiên kích thước n về véc tơ quan sát (Y, X_1, X_2, ..., X_{k-1}) là: $(Y_i, X_{1i}, X_{2i}, \dots, X_{k-1,i}), i = 1,2, \dots, n$

Đặt $U_i = Y_i - (a_0 + a_1.X_{1i} + a_2.X_{2i} + \cdots + a_{k-1}.X_{k-1,i})$;

$$\mathcal{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k-1,1} \\ 1 & X_{12} & \cdots & X_{k-1,2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{1n} & \cdots & X_{k-1,n} \end{pmatrix}; \mathcal{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}; \mathcal{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}; a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{k-1} \end{pmatrix};$$

$$\hat{\mathcal{Y}} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix}; \quad \hat{\mathcal{U}} = \begin{pmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_n \end{pmatrix}; \quad \hat{a} = \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_{k-1} \end{pmatrix}$$

Mô hình PRF (3.3) có dạng ma trận: $\mathcal{Y} = \mathcal{X}.a + \mathcal{U}$ (3.3a)

Mô hình SRF (3.4) có dạng ma trận: $\begin{cases} \hat{\mathcal{Y}} = \mathcal{X}.\hat{a} \\ \mathcal{Y} = \mathcal{X}.\hat{a} + \hat{\mathcal{U}} \end{cases}$ (3.4a)

3.1.2. Ước lượng các tham số hồi quy

Sử dụng phương pháp OLS, ta tìm ước lượng $\hat{a}_j$ *của* $a_j, j = 0,1, \dots, k-1$ sao cho:

$F(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_{k-1}) = \sum \hat{U}_i^2 = \sum \left[Y_i - \left(\hat{a}_0 + \hat{a}_1 X_{1i} + \cdots + \hat{a}_{k-1} X_{k-1,i}\right)\right]^2 \to \min.$

Với giả thiết ma trận $\boldsymbol{\mathcal{X}^T}.\ \boldsymbol{\mathcal{X}}$ khả nghịch (tức là $det\boldsymbol{\mathcal{X}} \neq \mathbf{0}$), người ta chứng minh được nghiệm duy nhất của hệ phương trình tuyến tính:

$$\frac{\partial F(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_{k-1})}{\partial \hat{a}_j} = 0, \forall j = 1,2, \dots k-1$$

là: $\hat{\boldsymbol{a}} = (\boldsymbol{\mathcal{X}^T}.\ \boldsymbol{\mathcal{X}})^{-1}.(\boldsymbol{\mathcal{X}^T}.\mathcal{Y})$ (3.5)

chính là ước lượng cần tìm.

3.2. Hệ số xác định và hệ số tương quan

Với mẫu ngẫu nhiên kích thước n về véc tơ quan sát (Y, X_1, X_2, ..., X_{k-1}) là:

$$(Y_i, X_{1i}, X_{2i}, \dots, X_{k-1,i}), i = 1,2, \dots, n$$

ta định nghĩa các tổng bình phương độ lệch như trước đây:

$TSS = \sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - n.(\bar{Y})^2 = \mathcal{Y}^T.\mathcal{Y} - n.(\bar{Y})^2;$ (3.6)

$ESS = \sum(\hat{Y}_i - \bar{Y})^2 = \hat{\boldsymbol{a}}^T.(\boldsymbol{X}^T.\mathcal{Y}) - n.(\bar{Y})^2;$ (3.7)

$RSS = \sum \hat{U}_i^2 = TSS - ESS$ (3.8)

Các tổng bình phương các độ lệch TSS, ESS, RSS trong mô hình hồi quy nhiều biến có ý nghĩa như các tổng bình phương các độ lệch tương ứng trong mô hình hồi quy hai biến.

Hệ số xác định là: $R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$ (3.9)

Ý nghĩa và tính chất của hệ số xác định giống như trước đây đã chỉ ra trong chương trước. Ngoài ra ta cần lưu ý các kết quả khảo sát sau đây:

* $TSS = \sum(Y_i - \bar{Y})^2$ có bậc tự do là (*n – 1*) và không phụ thuộc vào số biến độc lập trong mô hình.

* $RSS = \sum \hat{U}_i^2$ có bậc tự do là (*n – k*) và có giá trị giảm khi số biến giải thích trong mô hình tăng

* $R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$ có giá trị tăng theo số biến giải thích trong mô hình.

Vì vậy khi số biến giải thích càng nhiều thì R^2 càng lớn, tuy nhiên khi đó mô hình sẽ phức tạp hơn và khó phân tích hơn. Ngoài ra, khi có nhiều biến giải thích thì khả năng có tương quan cao giữa chúng dễ xảy ra, đồng thời bậc tự do của ESS và RSS sẽ giảm đi. Do đó cần thận trọng cân nhắc giữa việc đưa thêm biến giải thích vào để tăng trị số của R^2 với độ phức tạp phức tạp của mô hình cũng sẽ tăng lên.

3.2.1. Hệ số xác định hiệu chỉnh (*Adjusted R – squared*)

Trong mô hình hồi quy nhiều biến, khi đưa vào nhiều biến giải thích thì số bậc tự do bị giảm đi. Để hạn chế bất lợi này, người ta điều chỉnh hệ số xác định bằng cách đưa thêm bậc tự do của các tổng bình phương vào công thức sau để có hệ số xác định hiệu chỉnh:

$$\bar{R}^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} = R^2 + (1 - R^2).\frac{1-k}{n-k} \quad (3.10)$$

Hệ số xác định hiệu chỉnh $\bar{R}^2$ có các tính chất sau:

* $\bar{R}^2 \leq R^2 \leq 1, khi\ \ k > 0$

* Khi k càng lớn thì $\bar{R}^2$ *càng nhỏ hơn* R^2.

* $\bar{R}^2$ có thể ≤ 0 (*khi đó quy ước*: $\bar{R}^2 = 0$).

$\bar{R}^2$ được sử dụng để thay thế cho R^2 khi xem xét có nên đưa thêm biến giải thích mới vào mô hình hay không. Thường thì một biến giải thích nên được đưa thêm vào khi nó làm tăng giá trị của $\bar{R}^2$ và hệ số hồi quy của biến này phải khác không một cách có ý nghĩa thống kê.

3.2.2. Hệ số tương quan (*Coefficient of Correlation*)

Nhắc lại: với 2 biến ngẫu nhiên ξ và ζ: Hệ số tương quan giữa chúng là:

$$\rho_{\xi\zeta} = \frac{E(\xi - E\xi)(\zeta - E\zeta)}{\sqrt{var\xi}\sqrt{var\zeta}}$$

Hệ số tương quan đo mức độ phụ thuộc tương quan tuyến tính giữa hai biến. Với mẫu ngẫu nhiên kích thước n về véc tơ quan sát (Y, X_1, X_2, ..., X_{k-1}) là:

$$(Y_i, X_{1i}, X_{2i}, \dots, X_{k-1,i}), i = 1,2, \dots, n$$

* Hệ số tương quan mẫu giữa biến phụ thuộc Y và biến giải thích X_j là

$$r_{0j} = \frac{\overline{YX_j} - \bar{Y}.\bar{X}_j}{S(Y).S(X_j)} = \frac{\sum y_i.x_{ji}}{\sqrt{\sum y_i^2.\sum x_{ji}^2}} \quad (3.11)$$

* Hệ số tương quan mẫu giữa các biến X_s và X_j là:

$$r_{sj} = \frac{\overline{X_sX_j} - \bar{X}_s.\bar{X}_j}{S(X_s).S(X_j)} = \frac{\sum x_{si}.x_{ji}}{\sqrt{\sum x_{si}^2.\sum x_{ji}^2}} \quad (3.12)$$

$trong$ đó $y_i = Y_i - \bar{Y}$; $x_{ji} = X_{ji} - \bar{X}_j$, $i = 1,2, \dots, n; j = 1,2, \dots, k-1$

* Ma trận các hệ số tương quan mẫu là:

$$R = \begin{bmatrix} 1 & r_{01} & \cdots & r_{0,k-1} \\ r_{10} & 1 & \cdots & r_{1,k-1} \\ \vdots & \vdots & \cdots & \vdots \\ r_{k-1,0} & r_{k-1,1} & \cdots & 1 \end{bmatrix}$$

Lưu ý: Đối với mô hình hồi quy nhiều biến, việc tính toán trực tiếp các biểu thức có liên quan nói trên là rất khó khăn, phức tạp. Để thực hiện các tính toán này, cần dựa vào các phần mềm ứng dụng. Trong tài liệu này, chúng ta sử dụng phần mềm Eviews hỗ trợ.

3.2.3. Hệ số tương quan mẫu riêng phần (*Partial correlation coefficients*) (**Tham khảo**)

Hệ số tương quan được xét ở trên còn được gọi là hệ số tương quan bậc 0, xét mối tương quan giữa 2 biến mà không quan tâm đến sự thay đổi của các biến còn lại. Trong mô hình hồi quy k biến, để xét mối tương quan riêng phần giữa biến phụ thuộc Y và một biến giải thích X_j nào đó, ta phải cố định (k – 2) biến còn lại, khi đó ta có hệ số tương quan riêng phần bậc (k – 2).

* Với mô hình 3 biến: Y(biến phụ thuộc), X_1, X_2.

a. Để xác định hệ số tương quan riêng của Y và X_1 (loại bỏ tác động của X_2) tiến hành như sau:

- Chạy hồi quy của Y theo X_2 và xác định: $\hat{Y} = \hat{\alpha}_0 + \hat{\beta}_0.X_2$
- Chạy hồi quy của X_1 theo X_2 và xác định: $\hat{X}_1 = \hat{\alpha}_1 + \hat{\beta}_1.X_2$
- Loại bỏ tác động của X_2 lên Y và của X_2 lên X_1 bằng cách:

$$\text{Đặt } Y^* = Y - \hat{Y};\ X_1^* = X_1 - \hat{X}_1$$

- Hệ số tương quan riêng giữa Y và X_1 chính là hệ số tương quan giữa Y^* và X_1^*

b. Các công thức tính các hệ số tương quan riêng phần bậc 1:

$$r_{01.2} = \frac{r_{01} - r_{02}.r_{12}}{\sqrt{(1 - r_{02}^2)(1 - r_{12}^2)}}:$$

($Hệ\ số\ tương\ quan\ riêng\ giữa\ Y\ và\ X_1, với\ X_2\ không\ đổi.$)

$$r_{02.1} = \frac{r_{02} - r_{01}.r_{12}}{\sqrt{(1-r_{01}^2)(1-r_{12}^2)}}) \quad (3.15)$$

($Hệ\ số\ tương\ quan\ riêng\ giữa\ Y\ và\ X_2, với\ X_1\ không\ đổi.$)

$$r_{12.0} = \frac{r_{12} - r_{01}.r_{02}}{\sqrt{(1-r_{01}^2)(1-r_{02}^2)}}:$$

($Hệ\ số\ tương\ quan\ riêng\ giữa\ X_1\ và\ X_2, với\ Y\ không\ đổi$).

* Xét mô hình hai biến: $Y = a' + b'.X_1 + U, có\ hệ\ số\ xác\ định\ R_{2b}^2$ và mô hình ba biến: $Y = a_0 + a_1.X_1 + a_2.X_2 + V\ có\ hệ\ số\ xác\ định\ R_{3b}^2$

Khi đó người ta chứng minh được rằng:

$$R_{3b}^2 = R_{2b}^2 + (1 - R_{2b}^2).(r_{02.1})^2 \quad (3.16)$$

Kết quả này cũng chỉ ra khi số biến tăng thì hệ số xác định của mô hình cũng tăng

* Với mô hình 4 biến: Y(biến phụ thuộc), X_1, X_2, X_3 là các biến độc lập, để xác định hệ số tương quan riêng của Y và X_1 (loại bỏ tác động của X_2, X_3) ta tiến hành như sau:

- Chạy hồi quy của Y theo X_2 và X_3, xác định: $\hat{Y} = \hat{\alpha}_0 + \hat{\beta}_0.X_2 + \hat{\gamma}_0.X_3$
- Chạy hồi quy của X_1 theo X_2 và X_3, xác định: $\hat{X}_1 = \hat{\alpha}_1 + \hat{\beta}_1.X_2 + \hat{\gamma}_1.X_3$
- Loại bỏ tác động của X_2 và X_3 lên Y và của X_2 và X_3 lên X_1 bằng cách:

$$Đặt\ Y^* = Y - \hat{Y};\ X_1^* = X_1 - \hat{X}_1$$

- Hệ số tương quan riêng giữa Y và X_1 chính là hệ số tương quan giữa $Y^* và\ X_1^*$.

Tiến hành tương tự đối mô hình k biến độc lập tổng quát.

3.2.4. Các giả thiết của phương pháp OLS

Giả thiết 1: Trung bình của nhiễu không thay đổi và = 0. Như vậy:

$$E\mathcal{U} = \begin{pmatrix} EU_1 \\ EU_2 \\ \vdots \\ EU_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Giả thiết 2: Nhiễu có phương sai thuần nhất và không có tương quan chuỗi. Như vậy: $\begin{cases} VarU_i = \sigma^2 \\ cov(U_i, U_j) = E(U_i.U_j) = 0 \end{cases}, \forall i, j = \overline{1,n}, i \neq j$, hay một cách tương đương ta có:

$$E(\mathcal{U}.\mathcal{U}^T) = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2.I_{n\times n} \quad (3.17)$$

Giả thiết 3: Ma trận $\mathcal{X}$ đã được xác định theo nghĩa: Mẫu về biến X không chọn ngẫu nhiên.

Giả thiết 4: $r(\mathcal{X})(hạng\ của\ ma\ trận\ \mathcal{X}) = k$, hay không có cột nào của ma trận $\mathcal{X}$ là tổ hợp tuyến tính của các cột khác, tức là không có hiện tượng cộng tuyến xảy ra giữa các biến độc lập (giả thiết này cũng có nghĩa là $det\mathcal{X} \neq 0$, tức là ước lượng các hệ số hồi quy theo phương pháp OLS luôn tìm được và duy nhất) .

Giả thiết 5: $\mathcal{U} \sim N(0, \sigma^2.I)$: $\mathcal{U}$ *là véc tơ ngẫu nhiên có phân phối chuẩn n chiều*

3.2.5. Các tính chất của hệ số hồi quy

Từ các giả thiết của mô hình, các ước lượng được tìm theo phương pháp OLS nên có các tính chất tương tự như trong mô hình hồi quy hai biến, cụ thể ta có:

* Đồ thị hàm SRF đi qua điểm: $(\bar{Y}, \bar{X}_1, \dots, \bar{X}_{k-1})$

* $\bar{\hat{Y}} = \bar{Y}$

* $\bar{\hat{U}} = 0$

* $cov(\hat{U}, X) = 0$, *tức là* $\hat{U}$ *không tương quan với* X

* $cov(\hat{U}, \hat{Y}) = 0$, *tức là* $\hat{U}$ *không tương quan với* $\hat{Y}$

* $\hat{\boldsymbol{a}} = (\boldsymbol{X^T}.\boldsymbol{X})^{-1}.(\boldsymbol{X^T}.\mathcal{Y})$ được xác định duy nhất với một mẫu quan sát cụ thể , là một véc tơ ngẫu nhiên có phân phối chuẩn với giả thiết U có phân phối chuẩn.

* $\hat{\boldsymbol{a}} = (\boldsymbol{X^T}.\boldsymbol{X})^{-1}.(\boldsymbol{X^T}.\mathcal{Y})$ có ma trận hiệp phương sai:

$$\text{cov}(\hat{\boldsymbol{a}}) = \begin{bmatrix} var(\hat{a}_0) & cov(\hat{a}_0, \hat{a}_1) & \cdots & cov(\hat{a}_0, \hat{a}_{k-1}) \\ cov(\hat{a}_1, \hat{a}_0) & var(\hat{a}_1) & \cdots & cov(\hat{a}_1, \hat{a}_{k-1}) \\ \cdots & \cdots & \cdots & \cdots \\ cov(\hat{a}_{k-1}, \hat{a}_0) & cov(\hat{a}_{k-1}, \hat{a}_0) & \cdots & var(\hat{a}_{k-1}) \end{bmatrix} \qquad (3.18)$$

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

* Vì σ^2 chưa biết nên người ta dùng ước lượng $\hat{\sigma}^2 = \frac{RSS}{n-k}$ thay thế cho σ^2.

***** Với các giả thiết của mô hình hồi quy tuyến tính cổ điển thì

$$\hat{\boldsymbol{a}} = (\boldsymbol{X^T}.\boldsymbol{X})^{-1}.(\boldsymbol{X^T}.\mathcal{Y})$$

là ước lượng tuyến tính không chệch, có phương sai bé nhất trong trong lớp tất cả các ước lượng tuyến tính không chệch của a (tính chất **BLUE)**.

3.3. Các bài toán thống kê trên mô hình hồi quy nhiều biến

Mục này khảo sát bài toán ước lượng khoảng tin cậy cho các tham số, kiểm định giả thuyết thống kê liên quan đến mô hình

3.3.1. Khoảng tin cậy cho các tham số trong mô hình

Với mẫu kích thước n cho mô hình k tham số (1 biến phụ thuộc, (k – 1) biến giải thích) thỏa mãn giả thiết nhiễu U có phân phối chuẩn và ước lượng

$\hat{\boldsymbol{a}} = (\boldsymbol{X^T}.\boldsymbol{X})^{-1}.(\boldsymbol{X^T}.\mathcal{Y})$ (tìm theo phương pháp OLS), ta có:

$t = \dfrac{\hat{a}_j - a_j}{\widehat{se}(\hat{a}_j)} \sim t(n-k)$(*phân phối Student* $(n-k)$*bậc tự do*);

$\chi^2 = (n-k)\dfrac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k)$ (phân phối Chi-square, (n – k) bậc tự do). Vì thế:

- Với độ tin cậy $(1-\alpha)$, khoảng tin cậy cho các hệ số hồi quy a_j *l*à :

$$\left(\hat{a}_j - t_{\frac{\alpha}{2}}^{(n-k)}.\widehat{se}(\hat{a}_j);\quad \hat{a}_j + t_{\frac{\alpha}{2}}^{(n-k)}.\widehat{se}(\hat{a}_j)\right) \qquad (3.19)$$

- Với độ tin cậy $(1-\alpha)$, ta có khoảng tin cậy cho phương sai nhiễu σ^2:

$$\left(\frac{(n-k).\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}(n-k)};\quad \frac{(n-k).\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}(n-k)}\right) \qquad (3.20)$$

Trong đó: $t_{\frac{\alpha}{2}}^{(n-k)}$ là giá trị tới hạn mức $\frac{\alpha}{2}$ của phân phối Student (n – k) bậc tự do, tra từ bảng phụ lục I; $\chi^2_{\lambda}(n-k)$ là giá trị tới hạn mức λ của phân phối Chi – Square, (n – k) bậc tự do, tra từ bảng phụ lục III.

3.3.2. Kiểm định giả thuyết về mô hình

3.3.2.1. Kiểm định giả thuyết về hệ số hồi quy

$$H_0: a_j = a^*, H_1: a_j < a^*/a_j > a^*/a_j \neq a^* \ (a^* \text{ là hằng số cho trước})$$

Bảng sau tóm tắt kiểm định giả thuyết về hệ số hồi quy nhiều biến**:**

Bt k.định	P.p. k.định	Tiêu chuẩn bác bỏ giả thuyết H_0
$\begin{cases}H_0: a_j = a^*\\H_1: a_j < a^*\end{cases}$	Khoảng tin cậy	$a^* \geq \hat{a}_j + t_{\alpha}^{(n-k)}.\widehat{se}(\hat{a}_j)$
	Giá trị tới hạn	$t_0 < -t_{\alpha}^{(n-k)}$
	p – value	$p - value < 2\alpha$
$\begin{cases}H_0: a_j = a^*\\H_1: a_j > a^*\end{cases}$	Khoảng tin cậy	$a^* \leq \hat{a}_j - t_{\alpha}^{(n-k)}.\widehat{se}(\hat{a}_j)$
	Giá trị tới hạn	$t_0 > t_{\alpha}^{(n-k)}$
	p – value	$p - value < 2\alpha$
$\begin{cases}H_0: a_j = a^*\\H_1: a_j \neq a^*\end{cases}$	Khoảng tin cậy	$a^* \notin \left(\hat{a}_j - t_{\frac{\alpha}{2}}^{(n-k)}.\widehat{se}(\hat{a}_j); \hat{a}_j + t_{\frac{\alpha}{2}}^{(n-k)}.\widehat{se}(\hat{a}_j)\right)$
	Giá trị tới hạn	$t_0 \notin \left[-t_{\frac{\alpha}{2}}^{(n-k)};\quad t_{\frac{\alpha}{2}}^{(n-k)}\right]$
	p – value	$p - value < \alpha$
Ghi chú	$t_0 = \frac{\hat{a}_j - a^*}{\widehat{se}(\hat{a}_j)}$; $p - value = P(\lvert t\rvert > \lvert t_0\rvert)$	

Bảng 3.1

Ví dụ 3.1. Lượng hàng bán được Y (tấn / tháng), giá bán X_1 (ngàn đồng/kg) của một mặt hàng A và thu nhập X_2 (triệu đồng / tháng) của người tiêu dùng, qua điều tra, có số liệu sau:

Y	4	5	6	7	7	8	8	9
X_2	2	3	3	4	5	5	6	7
X_1	10	9	9	8	7	7	6	6

a/ Chạy hồi quy ước lượng cho mô hình: $Y = a_0 + a_1X_1 + a_2X_2 + U$ và cho biết Y có

thực sự phụ thuộc thống kê vào X_1, vào X_2 ? Cho biết ý nghĩa của các hệ số hồi quy ước lượng của các biến X_1, X_2.

b/ Hãy ước lượng khoảng tin cậy cho mức tăng hay giảm bình quân của lượng hàng bán được khi thu nhập của người tiêu dùng tăng thêm 1 triệu đồng/tháng với độ tin cậy 90%, 95%, 99%.

c/ Hãy ước lượng khoảng tin cậy cho mức tăng bình quân của lượng hàng bán được khi giá bán giảm 1 ngàn đồng/kg.

d/ Tính hệ số co giãn của Y theo X_2 từ hàm hồi quy ước lượng tìm được ở a/ và giải thích kết quả.

e/ Hãy ước lượng khoảng tin cậy 95% cho phương sai nhiễu.

Giải:

a/ Chạy hồi quy ta có:

Dependent Variable: Y
Method: Least Squares
Sample: 1 8
Included observations: 8

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.571429	9.322345	0.812181	0.4536
X2	0.571429	0.711423	0.803219	0.4583
X1	-0.428571	0.805593	-0.531995	0.6175

R-squared	0.912088	Mean dependent var	6.750000
Adjusted R-squared	0.876923	S.D. dependent var	1.669046
S.E. of regression	0.585540	Akaike info criterion	2.047432
Sum squared resid	1.714286	Schwarz criterion	2.077223
Log likelihood	-5.189728	Hannan-Quinn criter.	1.846507
F-statistic	25.93750	Durbin-Watson stat	2.440476
Prob(F-statistic)	0.002292		

Từ đó có mô hình hồi quy ước lượng:

$$Y = 7{,}571429 - 0{,}428571.X_1 + 0{,}571429.X_2 + \widehat{U}$$

- Việc xác minh xem Y có thực sự phụ thuộc thống kê vào X_1 chính là bài toán kiểm định: $H_0: a_1 = 0, H_1: a_1 \neq 0$. Dùng phương pháp p-value: từ kết quả hồi quy có: p-value = 0,6175 > $\alpha = 0{,}05$, nên ta chấp nhận H_0 và cho rằng Y không thực sự phụ thuộc thống kê vào X_1.

- Việc xác minh xem Y có thực sự phụ thuộc thống kê vào X_2 chính là bài toán kiểm định: $H_0: a_2 = 0, H_1: a_2 \neq 0$. Dùng phương pháp p-value: từ kết quả hồi quy có: p-value = 0,4583 > $\alpha = 0{,}05$, nên ta chấp nhận H_0 và cho rằng Y không thực sự phụ thuộc thống kê vào X_2.

Ta có thể lý giải kết quả kiểm định trên như sau: Mặt hàng mà chúng ta đang xét thuộc loại thiết yếu hay nhu yếu phẩm, do đó cho dù giá cả có tăng nhẹ hay lương có chút thay đổi thì người tiêu dùng vẫn phải mua với một lượng cần thiết để dùng.

- Ý nghĩa của các hệ số hồi quy ước lượng:

* $\hat{a}_2 = 0{,}571429$: cho thấy khi thu nhập của người tiêu dùng tăng thêm 1 triệu đồng/tháng thì bình quân lượng hàng bàn được tăng thêm 0,571429 tấn/tháng.

* $\hat{a}_1 = -0{,}428571$: cho thấy khi giá bán giảm bớt 1 ngàn đồng/kg thì bình quân lượng hàng bàn được tăng thêm 0,428571 tấn/tháng.

b, c/ Các khoảng tin cậy cần tìm được được Eviews cung cấp trong bảng sau đây:

Coefficient Confidence Intervals
Date: 06/29/15 Time: 17:02
Sample: 1 8
Included observations: 8

Variable	Coefficient	90% CI Low	90% CI High	95% CI Low	95% CI High	99% CI Low	99% CI High
C	7.571429	-11.21355	26.35641	-16.39242	31.53528	-30.01760	45.16046
X2	0.571429	-0.862123	2.004980	**-1.257342**	**2.400199**	-2.297130	3.439987
X1	-0.428571	-2.051880	1.194737	-2.499413	1.642271	-3.676836	2.819694

Chú ý: Nếu sử dụng thông tin từ bảng hồi quy, ta có thể tính toán trực tiếp để tìm khoảng tin cậy (điều này là bắt buộc đối với các bạn sinh viên khi làm bài mà không thể sử dụng phần mềm hỗ trợ), chẳng hạn đối với khoảng tin cậy 95% cho a_2:

$$(\hat{a}_2 - \varepsilon,\ \ \hat{a}_2 + \varepsilon)\ \ (\varepsilon = t_{\frac{\alpha}{2}}^{n-k}.\widehat{se}(\hat{a}_2))$$

Từ bảng hồi quy, có: $\hat{a}_2 = 0{,}571429;\ \widehat{se}(\hat{a}_2) = 0{,}711423$

Vơi độ tn cậy $\gamma = 1 - \alpha = 0{,}95, có: \alpha = 0{,}05;\ t_{\frac{\alpha}{2}}^{n-k} = t_{0{,}025}^{5} = 2{,}571;\ \varepsilon = 1{,}829069$

$$\begin{cases} \hat{a}_2 - \varepsilon = 0{,}571429 - 1{,}829069 = -1{,}25764 \\ \hat{a}_2 + \varepsilon = 0{,}571429 + 1{,}829069 = 2{,}400498 \end{cases}$$

Khoảng tin cậy cần tìm cho a_2 là: (-1,25764; 2,400494), phù hợp với kết quả mà Eviews xung cấp trong bảng ước lượng trên. Sai số không đáng kể là do quá trình tính toán, làm tròn.

d/ Từ PRF: $Y = a_0 + a_1X_1 + a_2X_2 + U$, ta có hệ số co giãn của Y theo X_2 là:

$$E_{Y/X2} = f'(X_2).\frac{Y}{X_2} = a_2.\frac{Y}{X_2}$$

Vì thế từ hàm hồi quy ước lượng, ta có hệ số co giãn của Y theo X_2 là:

$$E_{Y/X2} = \hat{a}_2.\frac{\bar{Y}}{\bar{X}_2} = 0{,}571429.\frac{6{,}75}{4{,}375} = 0{,}881633$$

Điều này cho thấy: khi mức lương bình quân của người tiêu dùng tăng 1% thì bình quân lượng hàng bán được tăng 0,881633%.

e/ Khoảng tin cậy cho phương sai nhiễu có dạng:

$$\left(\frac{(n-k).\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}(n-k)}\ ;\ \frac{(n-k).\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}(n-k)}\right)$$

Có n = 8, k = 3, độ tin cậy $\gamma = 1 - \alpha = 0{,}95, nên\ \alpha = 0{,}05;\ \hat{\sigma} = 0{,}585540$

$$\chi^2_{\frac{\alpha}{2}}(n-k) = \chi^2_{0{,}025}(5) = 12{,}8325\ ; \chi^2_{1-\frac{\alpha}{2}}(n-k) = \chi^2_{0{,}975}(5) = 0{,}8312$$

$$\frac{(n-k).\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}(n-k)} = 0{,}133589;\ \frac{(n-k).\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}(n-k)} = 2{,}062422$$

Vậy khoảng tin cậy cần tìm cho phương sai nhiễu là: (0,133589; 2,062422)

3.3.2.1. Kiểm định giả thuyết về phương sai của nhiễu

$$H_0: \sigma^2 = \sigma_0^2 \, , H_1: \sigma^2 < \sigma_0^2 / \sigma^2 > \sigma_0^2 / \sigma^2 \neq \sigma_0^2 \quad (\sigma_0^2: hằng\ số\ cho\ trước)$$

Bảng 3.2 (Tóm tắt các bài toán kiểm định giả thuyết về phương sai nhiễu)

Bài toán kiểm định	Phương pháp kiểm định	Tiêu chuẩn bác bỏ giả thuyết H_0
$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 < \sigma_0^2 \end{cases}$	Khoảng tin cậy	$\sigma_0^2 \geq \dfrac{(n-k).\hat{\sigma}^2}{\chi^2_{1-\alpha}(n-k)}$
	Giá trị tới hạn	$\chi_0^2 < \chi^2_{1-\alpha}(n-k)$
	Giá trị p – value	$p - value > 1 - \alpha$
$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 > \sigma_0^2 \end{cases}$	Khoảng tin cậy	$\sigma_0^2 \leq \dfrac{(n-k).\hat{\sigma}^2}{\chi^2_{\alpha}(n-k)}$
	Giá trị tới hạn	$\chi_0^2 > \chi^2_{\alpha}(n-k)$
	Giá trị p – value	$p - value < \alpha$
$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases}$	Khoảng tin cậy	$\sigma_0^2 \notin \left(\dfrac{(n-k).\hat{\sigma}^2}{\chi^2_{\alpha/2}(n-k)}; \dfrac{(n-k).\hat{\sigma}^2}{\chi^2_{1-\alpha/2}(n-k)} \right)$
	Giá trị tới hạn	$\chi_0^2 \notin [\chi^2_{1-\frac{\alpha}{2}}(n-k); \chi^2_{\frac{\alpha}{2}}(n-k)]$
	Giá trị p – value	$p - value \notin \left[\dfrac{\alpha}{2}; 1 - \alpha/2\right]$
Ghi chú	$\chi_0^2 = (n-k)\dfrac{\hat{\sigma}^2}{\sigma_0^2}$; $p - value = P(\chi^2 > \chi_0^2)$; $\chi^2_{\frac{\alpha}{2}}(n-k)\ tra\ bảng\ giá\ trị\ tới\ hạn\ phân\ bố\ \chi^2(n-k)$	

3.3.2.2. Kiểm định giả thuyết về sự phù hợp của mô hình

Kiểm định giả thuyết $H_0: R^2 = 0\ (hay: a_1 = a_2 = \cdots = a_{k-1} = 0)$,
đối thuyết $H_1: R^2 > 0\ (hay: có\ ít\ nhất\ một\ hệ\ số\ a_j \neq 0)$
Giả thuyết $H_0: a_1 = a_2 = \cdots = a_{k-1} = 0$ được gọi là giả thuyết đồng thời. Nếu ta bác bỏ H_0, thì có nghĩa là mô hình được xem là phù hợp. Nếu ngược lại thì mô hình được xem là không phù hợp.

Xét thống kê: $F = \dfrac{R^2.(n-k)}{(1-R^2).(k-1)}$. Nếu giả thuyết H_0 thì người chỉ ra được rằng F có phân phối F với các bậc tự do là k – 1, n – k . Vì vậy từ số liệu điều tra, tính: $F_0 = \dfrac{R^2.(n-k)}{(1-R^2).(k-1)}$

Ta có thể sử dụng phương pháp giá trị tới hạn hay giá trị p – value. Với mức ý nghĩa α:
Nếu: $F_0 > F_\alpha(k-1, n-k)$, hoặc: $p - value = P(F > F_0) < \alpha$,
thì bác bỏ giả thuyết H_0.

Chú ý: Giá trị thống kê F và p-value của thống kê F được Eviews cho biết trong bảng hồi quy.

Ví dụ 3.2: Xét tập số liệu trong ví dụ 1.

a/ Hãy xác minh xem phương sai nhiễu có vượt quá 0,4 hay không.

b/ Mô hình thu được trong kết quả hồi quy ở ví dụ 1 có phù hợp với thức tế điều tra hay không?

Giải:

a/ Ta có bài toán kiểm định giả thuyết về phương sai nhiễu: $H_0: \sigma^2 = 0{,}4$; $H_1: \sigma^2 > 0{,}4$

Dùng phương pháp khoảng tin cậy, tiêu chuẩn bác bỏ giả thuyết H_0 là:

$$\sigma_0^2 \leq \frac{(n-k).\hat{\sigma}^2}{\chi_\alpha^2(n-k)}$$

Ta có:

$\sigma_0^2 = 0{,}4; n = 8, k = 3, \hat{\sigma}^2 = (0{,}58554)^2 = 0{,}342857;\ \chi_\alpha^2(n-k) = \chi_{0,05}^2(5) = 11{,}070$

Từ đó: $0{,}4 > \frac{(n-k).\hat{\sigma}^2}{\chi_\alpha^2(n-k)} = 0{,}15485$ .... Vậy ta chấp nhận H_0 và cho rằng phương sai nhiễu chưa vượt quá 0,4.

3.3.2.3. Kiểm định Wald

Kiểm định Wald có nhiều ứng dụng trong các bài toán kiểm định và được chạy trên phần mềm Eviews.

Xét hai mô hình hồi quy sau:

$(U): Y = \ a_0 + a_1.X_1 + a_2.X_2 + \cdots a_{m-1}.X_{m-1} + \cdots + a_{k-1}.X_{k-1} \ + \ U;$

$(R): Y = \ a_0 + a_1.X_1 + a_2.X_2 + \cdots a_{m-1}.X_{m-1} + \ V$

(U) *gọi là mô hình không bị ràng buộc* $(Unrestricted\ model)$;

(R)*gọi là mô hình bị ràng buộc* $(Restricted\ model)$

Điều kiện ràng buộc trong mô hình (R) chính là hệ số hồi quy của các biến giải thích: X_m, X_{m+1},..., X_{k-1} đồng thời bằng 0.

Để xác minh có điều kiện ràng buộc hay không ràng buộc, ta kiểm định:

Giả thuyết $H_0: a_m = \cdots = a_{k-1} = 0$, *đối thuyết* $H_1: a_m^2 + \cdots + a_{k-1}^2 > 0$.

Kiểm định Wald giải bài toán này qua các bước sau:

Bước 1: Hồi quy mô hình (U) gồm *k* tham số, tính RSS_U có $(n-k)$ bậc tự do.

Bước 2: Hồi quy mô hình (R) gồm *m* tham số, tính RSS_R có $(n-m)$ bậc tự do.

Bước 3: Sử dụng thống kê F_W sau đây:

$$F_W = \frac{\frac{RSS_R - RSS_U}{k-m}}{\frac{RSS_U}{n-k}} = \frac{\frac{R_U^2 - R_R^2}{k-m}}{\frac{(1-R_U^2)}{(n-k)}};$$

Nếu giả thuyết H_0 đúng thì trong lý thuyết xác suất thống kê đã chỉ ra được khi đó F_W là biến ngẫu nhiên có phân phối Fisher với các bậc tự do là: $k-m, n-k$.

Vì thế, với mức ý nghĩa α, tiêu chuẩn bác bỏ giả thuyết H_0 là:

$$F_W > \ F_\alpha(k-m, n-k).$$

($F_\alpha(k-m, n-k)$ là giá trị tới hạn mức α của phân phối Fisher F(k – m, n – k) tra từ bảng phụ lục IV)

Lưu ý:

a/ Kiểm định Wald được sử dụng với nhiều mục đích khác nhau liên quan đến hệ số hồi quy, như: kiểm định tổ hợp tuyến tính, kiểm định thừa biến. Đối với giả thuyết H_0 là: $a_j = 0$, thì kết luận của kiểm định Wald tương đương với kết luận theo kiểm định t.
b/ Thường các phần mềm ứng dụng về kinh tế lượng đều lập trình thủ tục kiểm định Wald dựa trên cùng một nguyên tắc là so sánh RSS của mô hình bị ràng buộc và RSS của mô hình không bị ràng buộc. Các phần mềm này thường cung cấp cho người sử dụng các giá trị F_W và *p – value* của F_W là: *p – value* $= P(F > F_W)$, người sử dụng có thể dùng phương pháp giá trị tới hạn hay phương pháp *p – value* để kiểm định giả thuyết H_0.

3.3.2.4. Kiểm định tổ hợp tuyến tính về hệ số hồi quy (Tham khảo)

Ta xét bài toán kiểm định giả thuyết về mối liên hệ với ràng buộc tuyến tính giữa các hệ số hồi quy.

Xét mô hình hồi quy: *(U):* $Y = a_0 + a_1.X_1 + a_2.X_2 + U$

Ta muốn kiểm định: $Giả\ thuyết\ H_0: a_1 = a_2, với\ đối\ thuyết\ H_1: a_1 \neq a_2$

Để giải bài toán này ta có 3 phương pháp sau:

PP1: $K.định\ Wald$: Đặt $Z_1 = X_1 + X_2\ , Z_2 = X_1 - X_2, mô\ hình\ (U) trở\ thành$:

$$Y = a_0 + a_1'.Z_1 + a_2'.Z_2 + U. \quad \left(a_1' = \frac{(a_1 + a_2)}{2};\ a_2' = \frac{(a_1 - a_2)}{2}\right)$$

Khi đó $giả\ thuyết\ H_0: a_1 = a_2$ tương đương $với\ H_0: a_2' = 0$

$Bài\ toán\ trở\ thành\ bài\ toán\ kiểm$ định $Wald$:

$Mô\ hình\ không\ bị\ ràng\ buộc\ (U): Y = a_0 + a_1'.Z_1 + a_2'.Z_2 + U.$

$Mô\ hình\ ràng\ buộc\ (R): Y = a_0 + a_1'.Z_1 + U.$

$Với$ điều $kiện\ ràng\ buộc\ H_0: a_2' = 0$.

$$Áp\ dụng\ kiểm\ định\ Wald, với: F_W = \frac{\frac{(R_U^2 - R_R^2)}{(3-2)}}{\frac{(1-R_U^2)}{(n-3)}} \sim F(1, n-3);$$

$Nếu\ F_W > F_\alpha(1, n-3)\ \ thì\ bác\ bỏ\ giả\ thuyết\ H_0.$

PP2: Kiểm định t gián tiếp

Đặt: $V_1 = X_1 + X_2\ , V_2 = X_2\ \ \rho = a_2 - a_1$, ta có mô hình:

$$(U): Y = a_0 + a_1 V_1 + \rho V_2 + U\ ,$$

Việc kiểm định giả thuyết $H_0: a_1 = a_2$ trở thành kiểm định: $H_0: \rho = 0$

Dùng kiểm định t thông thường:

$$t = \frac{\hat{\rho} - 0}{\widehat{se}(\hat{\rho})}, trong\ đó\ \hat{\rho}\ là\ ước\ lượng\ của\ \rho. Nếu\ H_0\ đúng\ thì\ t \sim t(n-3)$$

Do đó nếu: $|t| > t_{\frac{\alpha}{2}}(n-3) (\alpha\ là\ mức\ ý\ nghĩa)$ thì bác bỏ H_0

PP3: Kiểm định t trực tiếp: $t = \frac{\hat{\rho} - 0}{\widehat{se}(\hat{\rho})} = \frac{\hat{a}_2 - \hat{a}_1}{\widehat{se}(\hat{a}_2 - \hat{a}_1)}$

$\left(\widehat{se}(\hat{a}_2 - \hat{a}_1) = \sqrt{\widehat{var}(\hat{a}_2 - \hat{a}_1)} = \sqrt{\widehat{var}(\hat{a}_1) + \widehat{var}(\hat{a}_2) - 2cov(\hat{a}_1, \hat{a}_2)}\ \right)$

Ví dụ 3.3: Hàm sản xuất Cobb – Douglas.

Hàm sản xuất là mô hình toán học, biểu diễn mối quan hệ giữa sản lượng (đầu ra) phụ thuộc vào các yếu tố đầu tư như lao động và vốn. Gọi Y là sản lượng, X_1 là lực lượng lao động, X_2 là vốn thì hàm sản xuất Cobb – Douglas có dạng:

$$Y = \alpha . X_1^{\beta_1} . X_2^{\beta_2} .$$

Mô hình toán học trên được đưa về dạng toán học tương đương:

$$lnY = \beta_0 + \beta_1 lnX_1 + \beta_2 lnX_2 \ \ (\beta_0 = ln\alpha)$$

Bằng cách bổ sung thêm sai số ngẫu nhiên U, ta nhận được mô hình kinh tế lượng hồi quy tuyến tính:

$$lnY = \beta_0 + \beta_1 lnX_1 + \beta_2 lnX_2 + U \ (\beta_0 = ln\alpha)$$

Đây là dạng tuyến tính log, với đặc điểm là các hệ số hồi quy riêng chính là hệ số co giãn: $\beta_1$là hệ số co giãn của sản lượng đối với lao động khi vốn không đổi, $\beta_2$là hệ số co giãn của sản lượng đối với vốn khi lao động không đổi.

Nhận thấy nếu tăng gấp đôi quy mô đầu tư về lao động và lượng vốn, ta có:

$$Y^* = Y(2X_1, 2X_2) = \alpha . (2X_1)^{\beta_1} . (2X_2)^{\beta_2} = 2^{\beta_1+\beta_2} . \alpha . X_1^{\beta_1} . X_2^{\beta_2} = 2^{\beta_1+\beta_2} . Y$$

Từ đó cho thấy:

$* \ khi \ \beta_1 + \beta_2 = 1 \ thì: Y^* = Y(2X_1, 2X_2) = 2Y$, tức là sản lượng không đổi theo quy mô và như thế việc mở rộng quy mô được đánh giá là không hiệu quả.

$* \ Khi \ \beta_1 + \beta_2 > 1 \ thì: Y^* = Y(2X_1, 2X_2) > 2Y$, tức là sản lượng tăng hơn quy mô, và như thế việc mở rộng quy mô được đánh giá là hiệu quả

$* \ Khi \ \beta_1 + \beta_2 < 1 \ thì: Y^* = Y(2X_1, 2X_2) < 2Y$, tức là sản lượng tăng không bằng quy mô, và như thế việc mở rộng quy mô được đánh giá là kém hiệu quả.

Vì vậy trong thực tế, người ta thường muốn xác minh xem có xảy ra tình huống sản lượng cùng cấp độ với quy mô hay không, tức là cần kiểm định:

Giả thuyết $H_0: \beta_1 + \beta_2 = 1, \ đối \ thuyết \ H_1: \beta_1 + \beta_2 \neq 1.$

Để minh họa, ta xét bảng số liệu dưới đây về sản lượng Y(đo bằng chỉ tiêu GDP thực: đơn vị tính: triệu pesos) và lực lượng lao động X1 được đo bằng tổng lao động (đơn vị: ngàn người), vốn cố định X2 (đơn vị: triệu pesos) ở Mexico từ 1955-1974.

Bảng 3.4. Số liệu về GDP, lượng LĐ và vốn cđ của Mexico từ 1955-1974

Năm	GDP	Lượng LĐ	Vốn c.định	Năm	GDP	Lượng LĐ	Vốn c.định
1955	114043	8310	182113	*1965*	212323	11746	315715
1956	120410	8529	193749	*1966*	226977	11521	337642
1957	129187	8738	205192	*1967*	241194	11540	363599
1958	134705	8952	215130	*1968*	260881	12066	391847
1959	139960	9171	225021	*1969*	277498	12297	422382
1960	150511	9569	237026	*1970*	296530	12955	455049
1961	157897	9527	248897	*1971*	306712	13338	484677
1962	165286	9662	260661	*1972*	329030	13738	520553
1963	178491	10334	275466	*1973*	354057	15924	561531
1964	199457	10981	295378	*1974*	374977	14154	609825

(Nguồn: Sources of Growth: A study of seven Latin American Economies, Victor J.Elias, (D.N. Gujarati))

Chạy hồi quy kinh tế lượng cho hàm sản xuất Cobb – Douglas, ta có bảng kết quả cho bởi bảng 3.5, trong đó: Mô hình hồi quy:

$$lnY = -1{,}652419 + 0{,}339732 . lnX_1 + 0{,}845997 . lnX_2 + \widehat{U}$$

Bảng 3.5 chỉ ra: trong trường hợp lượng lao động không thay đổi, nếu vốn cố định đầu tư tăng (giảm) 1% thì trung bình GDP thực của Mexico sẽ tăng (giảm) xấp xỉ 0,8459%; trong trường hợp vốn cố định đầu tư không thay đổi, nếu lượng lao động tăng (giảm) 1% thì trung bình. GDP thực của Mexico sẽ tăng (giảm) xấp xỉ 0,3397%. Hơn nữa hệ số xác định R^2 = 0,995080 là rất cao và các hệ số hồi quy đều có ý nghĩa thống kê, do đó mô hình phù hợp tốt với số liệu quan sát được, biến lnY được giải thích phần lớn bởi giá tri logarit của lao động và vốn.

Dependent Variable: LOG(Y)
Method: Least Squares
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.652419	0.606198	-2.725873	0.0144
LOG(X1)	0.339732	0.185692	1.829548	0.0849
LOG(X2)	0.845997	0.093352	9.062488	0.0000

R-squared	0.995080	Mean dependent var	12.22605
Adjusted R-squared	0.994501	S.D. dependent var	0.381497
S.E. of regression	0.028289	Akaike info criterion	-4.155221
Sum squared resid	0.013604	Schwarz criterion	-4.005861
Log likelihood	44.55221	Hannan-Quinn criter.	-4.126064
F-statistic	1719.231	Durbin-Watson stat	0.425667
Prob(F-statistic)	0.000000		

Bảng 3.5. Kết quả chạy hồi quy hàm sản xuất Cobb – Douglas của Mexico.

Theo kết quả hồi quy ta có: $\beta_1 + \beta_2 = 0{,}339732 + 0{,}845997 = 1{,}185729$,
giá trị này sai khác 1 không rõ ràng. Ta dùng kiểm định Wald để kiểm chứng điều đó, tức là kiểm định $H_0: \beta_1 + \beta_2 = 1,\ H_1: \beta_1 + \beta_2 \neq 1$.
Kết quả kiểm định nhờ phần mềm Eviews cho bởi bảng sau:

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
t-statistic	1.944078	17	0.0686
F-statistic	3.779440	(1, 17)	**0.0686**
Chi-square	3.779440	1	0.0519

Null Hypothesis: C(2)+C(3)=1
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
-1 + C(2) + C(3)	0.185730	0.095536

Restrictions are linear in coefficients.

Bảng 3.6.

Từ bảng 3.6, ta có p – value = 0,0686 > 0,05, nên với mức ý nghĩa 5% ta chấp nhận giả

thuyết H_0, nghĩa là có thể xem sản lượng không đổi theo quy mô sản xuất.

***Ví dụ 3.4*:** Tiến hành khảo sát giá bán X_1(ngàn đồng/kg), chi phí quảng cáo X_2 (triệu đồng/tháng) và lượng hàng bán được Y(tấn/tháng), có số liệu sau:

Y	X_1	X_2	Y	X_1	X_2	Y	X_1	X_2
20	2.5	10	16	4.7	7.1	12	7.7	7.5
19	3.1	9.2	15	5.3	6.9	15	5.9	6.9
18	3.5	8.8	15	5.8	6.5	16	4.8	6.7
18	4.2	8.4	14	5.9	6.8	12	7.2	6.5
17	4.6	8	14	6.4	6.6	10	8.3	7.2
17	3.8	7.6	13	6.8	7.0	11	8.5	8.3
16	4.2	7.2	12	7.2	7.8			

Bảng 3.7.

1**.** Ước lượng mô hình: $Y_i = a + b_1X_1 + b_2X_2 + U_i$

2. Cho biết ý nghĩa các hệ số hồi quy ước lượng của mô hình trên. Y có thực sự phụ thuộc thống kê vào X_1? Y có thực sự phụ thuộc thống kê vào X_2?

3. Đánh giá mức độ phù hợp của mô hình trên.

4. Dựa vào số liệu quan sát, hãy tìm ma trận tương quan của véc tơ ngẫu nhiên quan sát: (Y, X_1, X_2) và ma trận hiệp phương sai của các hệ số hồi quy ước lượng ở mô hình ở 1//

5. Hãy ước lượng khoảng tin cậy cho các hệ số hồi quy a, b_1, b_2, với độ tin cậy 95%.

6. Hãy ước lượng khoảng tin cậy cho phương sai nhiễu với độ tin cậy 95%.

Giải.

1. Chạy hồi quy, có bảng Equation sau:

Dependent Variable: Y
Method: Least Squares
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	**21.53448**	1.381611	15.58650	0.0000
X1	**-1.490279**	0.080691	-18.46905	0.0000
X2	**0.224088**	0.144743	1.548179	0.1400

R-squared	0.967652	Mean dependent var	15.00000
Adjusted R-squared	0.963846	S.D. dependent var	2.752989
S.E. of regression	0.523459	Akaike info criterion	1.680766
Sum squared resid	4.658165	Schwarz criterion	1.830126
Log likelihood	-13.80766	Hannan-Quinn criter.	1.709923
F-statistic	254.2644	Durbin-Watson stat	1.905397
Prob(F-statistic)	0.000000		

Bảng 3.7

2. Từ bảng kết quả hồi quy: đối với hệ số hồi quy của X_1 ta có p-value < 0,00005 << 0,05, nên Y phụ thuộc thống kê vào X_1; đối với hệ số hồi quy của X_2 ta có p-value = 0,1400 > 0,05, nên Y phụ thuộc thống kê không đáng kể vào X_2.

$\hat{b}_1$= -1.490279 cho thấy: khi giá bán tăng thêm 1 ngàn đồng/kg thì bình quân lượng hàng bán được giảm 1,490279 tấn/tháng.
$\hat{b}_2$= 0,224088 cho thấy: khi chi phí quảng cáo tăng thêm 1 triệu đồng/tháng thì bình quân lượng hàng bán được tăng thêm 0,224088 tấn/tháng
3. Từ bảng kết quả hồi quy, có: R^2 = 0,967652, cho thấy mô hình phù hợp tốt với số liệu quan sát.
4. Ma trận tương quan của véc tơ ngẫu nhiên quan sát: (Y, X_1, X_2) là:

	Y	X1	X2
Y	1.000000	-0.981372	0.564429
X1	-0.981372	1.000000	-0.516204
X2	0.564429	-0.516204	1.000000

Bảng 3.8

Ma trận hiệp phương sai của các hệ số hồi quy ước lượng ở mô hình ở a/ là:

	C	X1	X2
C	1.908850	-0.081459	-0.191456
X1	-0.081459	0.006511	0.006029
X2	-0.191456	0.006029	0.020951

Bảng 3.9

5. Từ kết quả ước lượng cho bởi Eviews: khoảng tin cậy cho các hệ số hồi quy a, b_1, b_2, với độ tin cậy 95% được chỉ ra như sau:

Coefficient Confidence Intervals
Sample: 1 20
Included observations: 20

Variable	Coefficient	95% CI Low	High
C	21.53448	18.61953	24.44942
X1	-1.490279	-1.660522	-1.320037
X2	0.224088	-0.081293	0.529469

Bảng 3.10

6. Khoảng tin cậy cho phương sai nhiễu σ^2 là:

$$\left(\frac{(n-k).\hat{\sigma}^2}{\chi^2_{\alpha/2}} \;;\; \frac{(n-k).\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right) \; (n=20, k=3, \alpha=0{,}05)$$

Từ bảng hồi quy ta có: $\hat{\sigma} = 0{,}523459$. Tra bảng có: $\chi^2_{\frac{\alpha}{2}}(n-k) = 30{,}1910$

$\chi^2_{1-\frac{\alpha}{2}}(n-k) = 7{,}5642 \Rightarrow \frac{(n-k).\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}} = \frac{17.(0{,}523459)^2}{30{,}1910} = 0{,}154289639;$

$\frac{(n-3).\hat{\sigma}^2}{\chi^2_{1-\alpha/2}} = 0{,}615816414.$

Khoảng tin cậy cần tìm cho phương sai nhiễu là: $(0{,}154289639,\ 0{,}615816414)$

****

Bài tập.

1. Biến Y phụ thuộc vào các biến X và T, qua số liệu điều ta, có bảng hồi quy sau

Dependent Variable: Y
Method: Least Squares
Sample: 1 14
Included observations: 14

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.158549	……	-0.782237	0.4506
T	……	1.575147	2.227274	0.0478
X	0.362917	0.808769	…….	0.6623

R-squared	0.982475	Mean dependent var	30.57143
Adjusted R-squared	0.979288	S.D. dependent var	17.77515
S.E. of regression	2.558114	Akaike info criterion	4.903827
Sum squared resid	…….	Schwarz criterion	5.040768
Log likelihood	-31.32679	Hannan-Quinn criter.	4.891151
F-statistic	308.3341	Durbin-Watson stat	2.546077
Prob(F-statistic)	0.000000		

a/ Hãy điền những thông tin còn thiếu vào chỗ trống (…). Viết SRF nhận được từ bảng kết quả hồi quy.
b/ Tính các tổng TSS, ESS
c/ Ước lượng khoảng tin cậy 95% cho phương sai nhiễu.
d/ Đánh giá mức độ phù hợp của mô hình với số liệu mẫu.

2. Từ mẫu điều tra về Y (tấn/tháng) là lượng cam bán được, X (ngàn đồng/kg) là giá cam và Z (ngàn đồng/kg) là giá quýt, để ước lượng cho mô hình:

$$Y = a_0 + a_1.X + a_2.Z + U$$

chạy hồi quy có kết quả sau:

Dependent Variable: Y
Method: Least Squares
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.04680	1.649159	……	0.0001
X	-0.127598	…….	-8.553684	0.0001
Z	……	0.028203	2.153557	0.0683

R-squared	0.981819	Mean dependent var	9.300000
Adjusted R-squared	0.976624	S.D. dependent var	2.869379
S.E. of regression	……	Akaike info criterion	1.433330
Sum squared resid	1.347213	Schwarz criterion	1.524105
Log likelihood	-4.166649	Hannan-Quinn criter.	1.333749
F-statistic	189.0086	Durbin-Watson stat	1.697425
Prob(F-statistic)	0.000001		

a/ Hãy điền các thông tin còn thiếu vào các ô trống (…) và thiết lập SRF ước lượng cho mô hình trên.

b/ Tính các tổng TSS, ESS.
c/ Đánh giá mức độ phù hợp của mô hình được thiết lập.

3. Từ bảng hồi quy ở bài tập 2, hãy thực hiện các yêu cầu sau:
a/ Dự báo lượng cam bán được, nếu giá cam là 55 ngàn đồng/kg và giá quýt là 45 ngàn đồng/kg
b/ Y phụ thuộc chủ yếu vào biến nào? Tại sao?
c/ Nêu ý nghĩa của các hệ số hồi quy của X và của Z trong SRF thu được
c/ Ước lượng khoảng tin cậy 95% cho mức tăng bình quân của lượng cam bán được khi giá cam giảm bớt 1 ngàn đồng/kg

4. Từ số liệu sử dụng trong bài tập 2, chạy hồi quy ước lượng cho mô hình:

$$Y = a_0 + a_1.lnX + a_2.lnZ + U$$

có kết quả sau:

Dependent Variable: Y
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	17.66274	4.004652	4.410555	0.0031
LOG(X)	-4.744952	0.438304	-10.82571	0.0000
LOG(Z)	2.659525	0.729265	3.646857	0.0082

R-squared	0.985784	Mean dependent var	9.300000
Adjusted R-squared	0.981722	S.D. dependent var	2.869379
S.E. of regression	0.387926	Akaike info criterion	1.187320
Sum squared resid	1.053405	Schwarz criterion	1.278095
Log likelihood	-2.936598	Hannan-Quinn criter.	1.087739
F-statistic	242.7017	Durbin-Watson stat	2.370928
Prob(F-statistic)	0.000000		

a/ Thiết lập SRF ước lượng cho mô hình trên. Nêu ý nghĩa của các hệ số hồi quy ước lượng của ln(X) và ln(Z).
b/ Từ SRF được thiết lập, tính hệ số co giãn của Y theo X và cho biết ý nghĩa của giá trị này, biết $\bar{X} = 46{,}5$.
c/ Ước lượng khoảng tin cậy cho mức tăng bình quân của lượng cam bán được khi giá cam giảm 1%.
d/ Với mức ý nghĩa 5%, xác minh xem phương sai nhiễu có vượt quá 0,15 hay không.
e/ Bạn chọn mô hình nào trong hai mô hình SRF được thiết lập ở bài tập 2 và bài tập 4, tại sao?
h/ Dùng mô hình SRF được thiết lập trong bài tập này để dự báo lượng cam bán được khi giá cam là 55 ngàn đồng/kg và giá quýt là 45 ngàn đồng/kg.

5. Với biến Y: lượng hàng A bán được (tấn /tháng) và các biến X_1:thu nhập của người tiêu dùng(triệu đồng /tháng), X_2 :giá bán của mặt hàng A(ngàn đồng /kg), có mẫu sau:

Y	4	4	6	7	6	8	7	9	8	9
X_1	2	2	4	4	5	6	5	6	5	7
X_2	6	7	7	6	7	5	4	8	7	7

a/ Tìm hàm hồi quy mẫu: $\hat{Y} = \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2 + \hat{U}$ và cho biết ý nghĩa của các hệ số hồi quy riêng.
b/ Tìm ma trận hiệp phương sai của các hệ số hồi quy
c/ Với mức ý nghĩa 5%, dựa vào mẫu quan sát trên, hãy kiểm định giả thuyết $H_0: a_2 = 0$; $H_1: a_2 \neq 0$. Cho biết ý nghĩa của kiểm định này.
d/ Với độ tin cậy 95%, dựa vào mẫu điều tra, hãy ước lượng khoảng tin cậy cho a_2.
e/ Hãy kiểm định sự phù hợp của hàm SRF với mức ý nghĩa 5%.
f/ Dựa vào số liệu có được, với độ tin cậy 95% hãy xác định khoảng tin cậy cho phương sai nhiễu.
g/ Hàm SRF thay đổi như thế nào nếu đơn vị đo của Y là tấn /năm, của X_2 là triệu đồng /tấn.
h/ Sử dụng mô hình a/ tìm khoảng dự báo cho giá trị trung bình và giá trị cá biệt về lượng hàng Y khi X1 = 8, X2 = 9 với độ tin cậy 95%

6. Bảng sau đây là số liệu của ngành công nghiệp Việt Nam từ 1976 – 1991:

Năm	Sản lượng	Chi phí lao động	Vốn
1976	65344	2033.4	23.88
1977	72399	2151.2	25.79
1978	78300	2092.4	28.32
1979	74594	2134.8	31.31
1980	66925	2250.3	33.74
1981	67594	2232.7	35.99
1982	73463	2273.2	38.14
1983	83034	2365.1	40.67
1984	93953	2460.2	43.23
1985	103258	2571.8	45.36
1986	109632	2587.0	46.80
1987	130551	2844.7	47.70
1988	137819	2945.0	49.20
1989	133311	2531.4	51.60
1990	139350	2251.0	52.99
1991	145621	2115.0	55.60

Bảng 3.12

a. Dùng hàm sản xuất Cobb – Douglas dạng: $Q = \alpha_0 L^{\alpha_1}.K^{\alpha_2}$ để ước lượng các tham số α_1, α_2 và cho biết ý nghĩa của chúng (trong đó: Q là sản lượng công nghiệp, L là lao động, K là vốn).
b. Chạy hồi quy ước lượng cho mô hình: $Ln(Q/L) = a_0 + a_1 \ln L + a_2 Ln(K/L) + U$; kiểm định $H_0 : a_2 = 0$, $H_1 : a_2 \neq 0$ ở mức ý nghĩa 3%
c. Ước lượng khoảng tin cậy đối xứng cho các hệ số hồi quy trong mô hình ở b/ với độ tin cậy 95%.
d. Ước lượng khoảng tin cậy đối xứng cho phương sai nhiễu trong mô hình ở b/ với độ tin cậy 97%.

7. Bảng số liệu sau là về Y là biến tổng sản phẩm được sản xuất của một ngành công nghiệp trong vòng 15 năm của một quốc gia và chi phí về vốn X1 và lao động X2 để sản xuất của ngành đó, với X3 là biến xu thế(biến thời gian)

a. Hồi quy SRF cho các mô hình sau và nêu ý nghĩa của các hệ số hồi quy:

$a1.\ \ lnY = a + b_1 lnX_1 + b_2 lnX_2 + U,$

$a2.\ \ lnY = a + b_1 lnX_1 + b_2 lnX_2 + X_3 + V$

b. Từ bảng số liệu, hãy tìm ma trận tương quan của véc tơ quan sát: (Y, X_1, X_2) và ma trận hiệp phương sai của các hệ số hồi quy ước lượng cho mô hình a1/ và a2/. Biến Y có thực sự phụ thuộc thống kê vào biến xu thế hay không?

Năm	Y	X1	X2	X3
1	8911.4	120.753	281.5	1
2	10873.2	122.242	284.4	2
3	11132.5	125.263	289.9	3
4	12086.5	128.539	375.8	4
5	12767.5	131.427	375.2	5
6	16347.1	134.267	402.5	6
7	19542.7	139.038	478.0	7
8	21075.9	146.450	553.4	8
9	23052.6	153.714	616.7	9
10	26128.2	164.783	695.7	10
11	29563.7	176.864	790.8	11
12	33367.6	188.146	816.0	12
13	38354.8	205.841	848.4	13
14	46868.3	221.748	873.4	14
15	54308.9	239.715	999.2	15

Bảng 3.13

8. Bảng số liệu sau đây là mẫu điều tra về doang thu Y, chi phí cho quảng cáo X_1 và tiền lương X_2 của 12 nhân viên tiếp thị (đơn vị: triệu đồng):

Y	102	140	127	128	139	138	144	159	161	163	180	106
X_1	15	25	18	16	17	15	23	22	25	24	26	19
X_2	7	11	10	12	12	15	12	14	14	16	17	6

a. Thiết lập hồi quy ước lượng cho mô hình $Y = a_0 + a_1X_1 + a_2X_2 + U$.

b. Ước lượng khoảng tin cậy đối xứng cho các hệ số hồi quy của mô hình trên với độ tin cậy 90%, 95%, 99%.

c. Tìm ma trận tương quan giữa các biến trong mô hình và ma trận hiệp phương sai cho các hệ số hồi quy ước lượng.

d. Tính các giá trị thống kê (các giá trị của các đặc trưng mẫu) của các biến trong mô hình.

e. Định mẫu từ quan sát thứ 1 đến quan sát thứ 9 và chạy hồi quy ước lượng cho mô hình $Y = a_0 + a_1X_1 + a_2X_2 + U$.

g. Dùng mô hình ước lượng ở e/ để dự báo cho doanh thu Y ứng với các quan sát thứ 10, thứ 11 và thứ 12 của các biến X_1, X_2. Vẽ đồ thị Line graph của Y_{db} và Y quan sát thực tế.

Chương 4. BIẾN GIẢ TRONG PHÂN TÍCH HỒI QUY

Biến giả có vai trò rất quan trọng trong phân tích hồi quy.
Chương này giúp người học:
- Nắm được khái niệm về biến giả và sự cần thiết phải đưa biến giả vào mô hình hồi quy.
- Kỹ thuật sử dụng biến giả trong các mô hình kinh tế lượng.

4.1. Các khái niệm về biến giả

4.1.1. Khái niệm về biến giả: Trong các chương trước, các biến giải thích là các biến định lượng hay các tiêu chuẩn số lượng. Tuy nhiên trong thực tế có những trường hợp biến giải thích là biến định tính hay tiêu chuẩn chất lượng như: màu sắc, âm thanh, giới tính, chủng tộc, tôn giáo, hình thức sở hữu, nghề nghiệp,... mà chúng ta cảm nhận được ảnh hưởng không bỏ qua được của chúng vào biến phụ thuộc đang xét. Để đưa những thuộc tính của biến định tính vào mô hình hồi quy định lượng, người ta lượng hóa các thuộc tính bằng cách sử dụng kỹ thuật biến giả (*dummy variables*). Biến định tính sau khi được lượng hóa để đưa vào mô hình được gọi là biến giả.

4.1.2. Các ví dụ.
Ví dụ 4.1: Khảo sát lượng hàng A bán được theo hai khu vực bán là thành thị và nông thôn. Ký hiệu Y là lượng hàng A bán được, Y là một biến định lượng phụ thuộc vào khu vực bán là một biến định tính gồm 2 thuộc tính: thành thị và nông thôn. Để biểu thị sự phụ thuộc của lượng hàng bán được Y vào biến định tính là khu vực bán trong một mô hình kinh tế lượng, ta lượng hóa biến định tính khu vực bằng cách đặt:

$$D(x) = \begin{cases} 0, nếu\ lượng\ hàng\ x\ được\ bán\ ở\ nông\ thôn, \\ 1, nếu\ lượng\ hàng\ x\ được\ bán\ ở\ thành\ thị. \end{cases}$$

Khi đó D là một biến định lượng, thay cho vai trò của biến định tính khu vực, được đưa vào mô hình kinh tế lượng sau đây:

$$\begin{cases} E(Y|D) = a + b.D \\ Y = a + b.D + U \end{cases} \quad (4.1)$$

Theo đó: $E(Y|D = 0) = a$: là lượng hàng bình quân bán được ở khu vực nông thôn.
$E(Y|D = 1) = a + b$: là lượng hàng bình quân bán được ở khu vực thành thị.
Vì thế: $b = E(Y|D = 1) - E(Y|D = 0)$: là mức chênh lệch bình quân về lượng hàng bán được ở khu vực thành thị so với khu vực nông thôn. Lúc này lượng hàng bình quân bán được ở khu vực nông thôn là tiêu chuẩn để so sánh. Ta gọi thuộc tính " nông thôn" là thuộc tính cơ sở (hay phạm trù cơ sở), tương ứng với giá trị D = 0.
Chú ý: Khi biến định tính có nhiều hơn hai thuộc tính, ta có thể sử dụng giải pháp biến giả có nhiều hơn hai giá trị hoặc sử dụng giải pháp nhiều biến giả có giá trị 0 và 1. Tuy nhiên giải pháp đầu ít được sử dụng bởi những nguyên nhân sau: Khi so sánh giá trị trung bình của biến phụ thuộc tương ứng với các thuộc tính khác nhau thì việc phân tích mô hình sẽ khó khăn hơn: Biến giả có nhiều giá trị trở thành biến định lượng thông thường nên dễ tương quan với các biến giải thích khác trong mô hình. *Vì thế người ta thường sử*

dụng nhiều biến giả với hai giá trị 0 và 1, trên nguyên tắc: đối với mỗi biến định tính thì số biến giả được sử dụng bằng số thuộc tính – 1.

Ví dụ 4.2: Khảo sát lương nhân viên theo trình độ, với ba mức: Đại học, thạc sĩ, tiến sĩ, tức là theo 3 thuộc tính khác nhau. Theo đó ta sử dụng 2 biến giả $D_1 và D_2$ như sau:

$$D_1(x) = \begin{cases} 1, nếu\ x\ có\ trình\ độ\ thạc\ sĩ \\ 0, nếu\ x\ có\ trình\ độ\ khác \end{cases};$$

$$D_2(x) = \begin{cases} 1, nếu\ x\ có\ trình\ độ\ tiến\ sĩ \\ 0, nếu\ x\ có\ trình\ độ\ khác \end{cases}.$$

Như vậy trình độ nhân viên công sở được xác định bởi cặp giá trị của hai biến giả:

$Đại\ học: D_1 = 0; D_2 = 0. Thạc\ sĩ: D_1 = 1; D_2 = 0. Tiến\ sĩ: D_1 = 0; D_2 = 1$

Khi đó mô hình hồi quy có dạng:

$$E(Y|D_1, D_2) = a + b_1.D_1 + b_2 D_2 \qquad (4.2)$$

trong đó các mức giá trị kỳ vọng có điều kiện mang ý nghĩa như sau:

$E(Y|D_1 = 0; D_2 = 0) = a$: là mức lương bình quân của nhân viên có trình độ đại học.

$E(Y|D_1 = 1; D_2 = 0) = a + b_1$: mức lương bình quân của nhân viên trình độ thạc sỹ

$E(Y|D_1 = 0; D_2 = 1) = a + b_2$: mức lương bình quân của nhân viên trình độ tiến sỹ

b_1: là chênh lệch mức lương bình quân của nhân viên có trình độ thạc sỹ với nhân viên có trình độ đại học.

b_2: là chênh lệch mức lương bình quân của nhân viên có trình độ tiến sỹ với nhân viên có trình độ đại học.

Như vậy mức lương bình quân của nhân viên có trình độ đại học là tiêu chuẩn để so sánh, ta gọi thuộc tính "*trình độ đại học*" là thuộc tính cơ sở (hay là phạm trù cơ sở) tương ứng với cặp giá trị $D_1 = 0, D_2 = 0$.

Ví dụ 4.3: Hãy thiết lập mô hình hồi quy trong đó: doanh số bán sản phẩm A phụ thuộc vào: giá bán, kiểu dáng 1, kiểu dáng 2, kiểu dáng 3, khu vực nông thôn, khu vực thành thị.

Mô hình này có 4 biến: doanh số Y là biến định lượng, nó là biến phụ thuộc; các biến giải thích là: giá bán X (biến định lượng), và hai biến định tính là: kiểu dáng (có 3 thuộc tính) và khu vực bán (có 2 thuộc tính). Vậy ta cần 2 biến giả (nhị phân) cho kiểu dáng và một biến giả (nhị phân) cho khu vực bán như sau:

$$D_1(x) = \begin{cases} 1, nếu\ x\ có\ kiểu\ dáng\ 1 \\ 0, nếu\ x\ có\ kiểu\ dáng\ khác \end{cases};$$

$$D_2(x) = \begin{cases} 1, nếu\ x\ có\ kiểu\ dáng\ 2 \\ 0, nếu\ x\ có\ kiểu\ dáng\ khác \end{cases}; \ D(x) = \begin{cases} 1, nếu\ x\ bán\ ở\ thành\ thị \\ 0, nếu\ x\ bán\ ở\ nông\ thôn \end{cases}$$

Khi đó mô hình hồi quy có dạng:

$$E(Y|X, D, D_1, D_2) = a + b.X + c.D + b_1.D_1 + b_2 D_2 \qquad (4.3)$$

trong đó:

$E(Y|X, D = 0, D_1 = 0, D_2 = 0) = a + bX$: Doanh số bình quân bán kiểu dáng 3 ở khu vực nông thôn có giá bán X

$E(Y|X, D = 0, D_1 = 1, D_2 = 0) = a + bX + b_1$: Doanh số bình quân bán kiểu dáng 1 ở khu vực nông thôn có giá bán X

$E(Y|X, D = 0, D_1 = 0, D_2 = 1) = a + bX + b_2$: Doanh số bình quân bán kiểu dáng 2 ở khu vực nông thôn có giá bán X

$E(Y|X, D = 1, D_1 = 0, D_2 = 0) = a + bX$: Doanh số bình quân bán kiểu dáng 3 ở khu vực thành thị có giá bán X
$E(Y|X, D = 1, D_1 = 1, D_2 = 0) = a + bX + b_1$: Doanh số bình quân bán kiểu dáng 1 ở khu vực thành thị có giá bán X
$E(Y|X, D = 1, D_1 = 0, D_2 = 1) = a + bX + b_2$: Doanh số bình quân bán kiểu dáng 2 ở khu vực thành thị có giá bán X

4.2. Kỹ thuật sử dụng biến giả

4.2.1. Mô hình có biến giả

Mô hình có biến giả là mô hình phải sử dụng biến giả, đó là mô hình có đưa vào một hay nhiều biến giải thích là biến định tính. Sự xuất hiện của biến định tính trong mô hình là cần thiết khi ta cảm nhận sự phụ thuộc hoặc muốn khảo sát sự phụ thuộc của biến được giải thích vào sự thay đổi thuộc tính của biến định tính.

4.2.2. Kỹ thuật sử dụng biến giả

Đứng trước một mô hình có xét đến ảnh hưởng trực tiếp của một hay nhiều biến định tính đối với biến phụ thuộc (một biến định lượng), để sử dụng biến giả, ta cần lưu ý các bước sau:
- Xác định xem có bao nhiêu biến giải thích là biến định tính
- Xác định số thuộc tính của mỗi biến định tính
- Số biến giả (nhị phân) cần cho một biến định tính bằng số thuộc tính của biến định tính này trừ đi 1.

Gọi m là số biến định tính được đưa vào, trong đó biến thứ j có k_j thuộc tính, j = 1, 2,…, m,. Khi đó biến thứ j này cần đến $(k_j - 1)$ biến giả và tổng số biến giả (nhị phân) trong mô hình là: $k_1 + k_2 + \cdots + k_m - m$.
- Biểu diễn các điều kiện, các tình huống trong mô hình qua các biến giả

Ví dụ sau minh họa cho kỹ thuật này:
Để khảo sát lương Y của giáo viên theo thâm niên giảng dạy X, ta sử dụng mô hình hồi quy sau: $\begin{cases} E(Y|X) = a + b.X \\ Y = a + b.X + U \end{cases}$
Bây giờ ta muốn tìm hiểu về sự chênh lệch tiền lương bình quân giữa giáo viên nam và giáo viên nữ, tức là sự tác động của giới tính đến mức lương. Điều này đòi hỏi ta phải đưa thêm một biến giả D vào mô hình để mô tả sự tác động của giới tính. Đặt:

$$D(x) = \begin{cases} 1, nếu\ giáo\ viên\ x\ là\ nam \\ 0, nếu\ giáo\ viên\ x\ là\ nữ \end{cases}$$

Ta có các tình huống:
* TH1: Lương khởi điểm của giáo viên nam và nữ khác nhau, nhưng tốc độ tăng lương của nam và nữ là như nhau.
* TH2: Lương khởi điểm của giáo viên nam và nữ như nhau, nhưng tốc độ tăng lương của nam và nữ là khác nhau.
* TH3: Lương khởi điểm của giáo viên nam và nữ khác nhau và tốc độ tăng lương của nam và nữ cũng khác nhau.

Ta sử dụng biến giả tương ứng với các tình huống trên như sau:

4.2.2.1. TH1: Dịch chuyển số hạng tung độ gốc

Đặt: $a = a_0 + a_1 D$. *Khi đó hàm PRF có dạng*:

$$\boldsymbol{Y = a_0 + a_1 D + b.X + U} \quad (4.4)$$

Hồi quy mẫu SRF cho mức lương của giáo viên nữ (D = 0) là:

$$\hat{Y} = \hat{a}_0 + \hat{b}.X \quad (4.4a)$$

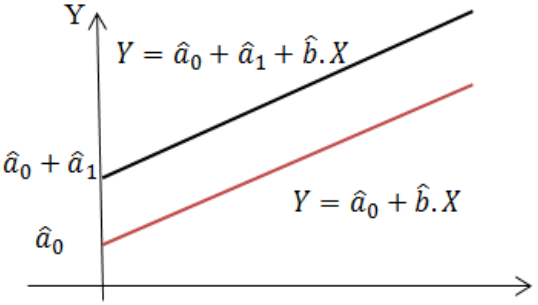
Hồi quy mẫu SRF cho mức lương của giáo viên nam (D = 1) là:

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 + \hat{b}.X \quad (4.4b)$$

Theo đó mức lương khởi điểm của giáo viên nữ là a_0 có ước lượng là $\hat{a}_0$, mức lương khởi điểm của giáo viên nam là $a_0 + a_1$ có ước lượng là $\hat{a}_0 + \hat{a}_1$

Hai đường thẳng (4.4a) và (4.4b) có cùng hệ số góc $\hat{b}$ biểu thị tốc độ tăng lương của giáo viên nữ và giáo viên nam là như nhau.

Việc xác minh giới tính có thực sự ảnh hưởng đến lương khởi điểm của giáo viên hay không chính là bài toán kiểm định hai phía đối với giả thuyết H_0: $a_1 = 0$ (có thể dùng kiểm định Wald hoặc kiểm định t)



Hình 4.1. Đồ thị biểu diễn Lương khởi điểm của giáo viên nam và nữ khác nhau, nhưng tốc độ tăng lương của nam và nữ là như nhau

4.2.2.2. TH2: Dịch chuyển độ dốc

Đặt: $b = b_0 + b_1 . D$. *Mô hình hồi quy tổng thể có dạng*:

$$Y = a + b.X + U = a + b_0.X + b_1.(D.X) + U \quad (4.5)$$

Ta gọi (D.X) là biến tương tác. Khi đó:

Hồi quy SRF cho mức lương của giáo viên nữ (D = 0) là:

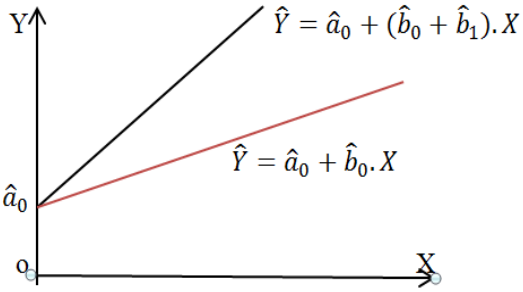
$$\hat{Y} = \hat{a}_0 + \hat{b}_0.X \quad (4.5a)$$

Hồi quy SRF cho mức lương của giáo viên nam (D = 1) là:

$$\hat{Y} = \hat{a}_0 + (\hat{b}_0 + \hat{b}_1).X \quad (4.5b)$$

Đường thẳng (4.5a) và (4.5b) khi bỏ qua yếu tố nhiễu được cho bởi hình 4.2 sau, trong đó lương khởi điểm của giáo viên nam và nữ đều là mức $\hat{a}_0$ như nhau, nhưng tốc độ tăng lương của nữ là $\hat{b}_0$, của nam là $(\hat{b}_0 + \hat{b}_1)$, chênh lệch một lượng $\hat{b}_1$.

Việc xác minh giới tính có thực sự ảnh hưởng đến tốc độ tăng lương của giáo viên hay không chính là bài toán kiểm định hai phía đối với giả thuyết H_0: $b_1 = 0$ (có thể dùng kiểm định Wald hoặc kiểm định t)



Hình 4.2. Đồ thị biểu diễn Lương khởi điểm của giáo viên nam và nữ như nhau, nhưng tốc độ tăng lương của nam, nữ là khác nhau

4.2.2.3. TH3: Dịch chuyển cả tung độ gốc và độ dốc

Đặt: $a = a_0 + a_1 D, b = b_0 + b_1 . D$. *Mô hình hồi quy tổng thể có dạng*:

$$Y = a + b.X + U = a_0 + a_1 D + b_0 . X + b_1 . (D.X) + U \quad (4.6)$$

Khi đó:

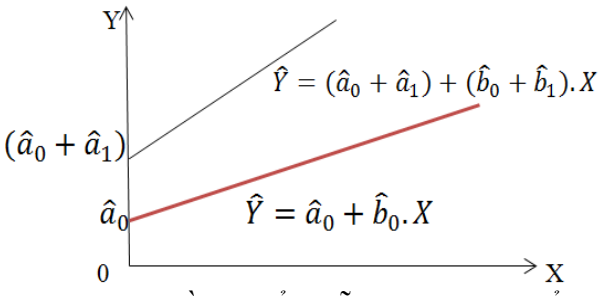
Hồi quy SRF cho mức lương của giáo viên nữ (D = 0) là:

$$\hat{Y} = \hat{a}_0 + \hat{b}_0 . X \quad (4.6a)$$

Hồi quy SRF cho mức lương của giáo viên nam (D = 1) là:

$$\hat{Y} = (\hat{a}_0 + \hat{a}_1) + (\hat{b}_0 + \hat{b}_1).X \quad (4.6b)$$

Đường thẳng (4.6a) và (4.6b) khi bỏ qua yếu tố nhiễu được cho bởi hình 4.3 sau. Trong đó lương khởi điểm của giáo viên nữ là mức $\hat{a}_0$, của giáo viên nam là $(\hat{a}_0 + \hat{a}_1)$, với mức chênh lệch là $\hat{a}_1$; tốc độ tăng lương của nữ là $\hat{b}_0$, của nam là $(\hat{b}_0 + \hat{b}_1)$, chênh lệch một lượng $\hat{b}_1$. Để xác minh giới tính có thực sự ảnh hưởng đến mức lương khởi điểm và tốc độ tăng lương của giáo viên hay không chính là bài toán kiểm định giả thuyết $H_0: a_1 = b_1 = 0$ (dùng kiểm định Wald).



Hình 4.3. Đồ thị biểu diễn lương khởi điểm của giáo viên nam và nữ khác nhau và tốc độ tăng lương của nam, nữ cũng khác nhau.

Lưu ý:

a. Trong thực tế người ta không xác định được trước việc bài toán đặt ra rơi vào tình huống nào. Do đó hoặc ta phải xét lần lượt 3 mô hình hồi quy ứng với 3 tình huống rồi chọn mô hình phù hợp nhất; hoặc ta bắt đầu từ tình huống thứ 3 rồi kiểm định $H_0: a_1 = b_1 = 0$: nếu chấp nhận H_0 thì ta dùng mô hình hồi quy Y theo X (không có tác động của giới tính), nếu bác bỏ H_0 thì tiếp tục kiểm định riêng từng hệ số để xác minh sự ảnh hưởng của giới tính nằm ở mức lương khởi đầu (tung độ gốc) hay tốc độ tăng lương (độ dốc), tức là kiểm định: $H_0: a_1 = 0$ hoặc kiểm định: $H_0: b_1 = 0$.

b. Có thể mở rộng bài toán hồi quy dạng trên đây khi số biến định tính tăng lên hoặc số thuộc tính của các biến định tính tăng lên. Chẳng hạn trong mô hình trên, ngoài thâm niên giảng dạy, giới tính, ta còn xét đến sự ảnh hưởng của trình độ với 3 thuộc tính: cử nhân, thạc sỹ, tiến sỹ đối với lương của giáo viên. Theo đó ta phải đưa vào mô hình 3 biến giả D, K_1, K_2:

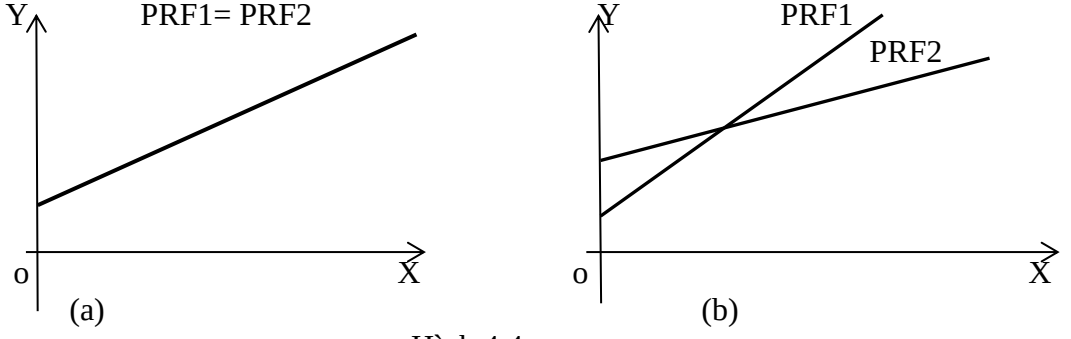
$$D(x) = \begin{cases} 1, \text{nếu giáo viên } x \text{ là nam} \\ 0, \text{nếu giáo viên } x \text{ là nữ} \end{cases}$$

$$K_1(x) = \begin{cases} 1, \text{nếu } x \text{ có học vị thạc sỹ} \\ 0, \text{nếu } x \text{ có học vị khác} \end{cases}; \; K_2 = \begin{cases} 1, \text{nếu } x \text{ có học vị tiến sỹ} \\ 0, \text{nếu } x \text{ có học vị khác} \end{cases}$$

Ta có mô hình PRF: $Y = a + b.X + c.D + d_1.K_1 + d_2.K_2 + U$ (4.7)
Khi đó:
$* E(Y|X, D = 0, K_1 = 0, K_2 = 0) = a + b.X$: lương b.quân của g.viên nữ có trình độ cử nhân
$* E(Y|X, D = 0, K_1 = 1, K_2 = 0) = a + d_1 + b.X$: lương bình quân của giáo viên nữ có trình độ thạc sỹ
$* E(Y|X, D = 0, K_1 = 0, K_2 = 1) = a + d_2 + b.X$: lương bình quân của giáo viên nữ có trình độ tiến sỹ
$* E(Y|X, D = 1, K_1 = 0, K_2 = 0) = a + c + b.X$: lương bình quân của giáo viên nam có trình độ cử nhân
$* E(Y|X, D = 1, K_1 = 1, K_2 = 0) = a + c + d_1 + b.X$: lương bình quân của giáo viên nam có trình độ thạc sỹ
$* E(Y|X, D = 1, K_1 = 0, K_2 = 1) = a + c + d_2 + b.X$: lương bình quân của giáo viên nam có trình độ tiến sỹ
- Ngoài ra, khi tốc độ tăng lương có thể bị chi phối bởi giới tính thì ta cần bổ sung thêm biến tương tác (D.X) vào mô hình hồi quy. Khi đó mô hình hồi quy có dạng:
$Y = a + b.X + \lambda(D.X) + c.D + d_1.K_1 + d_2.K_2 + U$ (4.8)
- Khi tốc độ tăng lương có thể bị chi phối bởi yếu tố trình độ thì ta cần bổ sung thêm 2 biến tương tác: $(K_1.X)$ *và* $(K_2.X)$. Khi đó mô hình hồi quy có dạng:
$Y = a + b.X + \lambda_1(K_1.X) + \lambda_2(K_2.X) + c.D + d_1.K_1 + d_2.K_2 + U$ (4.9)

4.2.3. So sánh cấu trúc của mô hình hồi quy

Xét mô hình hồi quy của biến phụ thuộc Y theo biến giải thích X qua 2 thời kỳ. Ta muốn biết liệu biến Y có sự thay đổi khác nhau ở 2 thời kỳ này hay không. Nếu không có sự khác biệt giữa 2 thời kỳ thì chỉ cần sử dụng một hàm hồi quy. Nếu có sự khác biệt thì cần xây dựng 2 hàm hồi quy: một cho thời kỳ trước là PRF1, một cho thời kỳ sau là PRF2. Hai hình ảnh sau đây minh họa cho các trường hợp không có sự khác biệt và có sự khác biệt này:



Hình 4.4

Để kiểm định có sự khác nhau thật sự của biến phụ thuộc Y giữa hai thời kỳ, tức là có hay không sự thay đổi về mặt cấu trúc của mô hình hồi quy, ta có thể sử dụng hai phương pháp sau đây:

4.2.3.1. Kiểm định Chow: phân cắt mẫu thành nhóm

Bước 1: Hồi quy riêng cho từng tời kỳ với thời kỳ trước có $n_1$quan sát và thời kỳ sau có n_2 quan sát, ta nhận được hai hàm hồi quy:
- Thời kỳ trước: $Y = a_1 + b_1.X + U_1, tính\ RSS_1 với\ (n_1 - k) bậc\ tự\ do$
- Thời kỳ sau: $Y = a_2 + b_2.X + U_2, tính\ RSS_2 với\ (n_2 - k) bậc\ tự\ do,$
(trong đó k là số tham số trong mô hình hồi quy, ở đây k = 2). Đặt:

$$RSS_U = RSS_1 + RSS_2\ , có\ bậc\ tự\ do\ là\ (n_1 + n_2 - 2k)$$

Bước 2: Gộp tất cả quan sát cả hai thời kỳ, ta có mẫu cỡ $n = n_1 + n_2$ và ước lượng mô hình sau: $Y = \delta_1 + \delta_2.X + U$
Tính RSS_R tương ứng với bậc tự do (n – k)

Bước 3: Kiểm định giả thuyết H_0: Hàm hồi quy của cả 2 thời kỳ là như nhau, với đối thuyết H_1: Hàm hồi quy của 2 thời kỳ là khác nhau
Tính giá trị của thống kê F là:

$$F_C = \frac{(RSS_R - RSS_U)/k}{RSS_U/(n-2k)},$$

Dùng phương pháp giá trị tới hạn: nếu $F_C > F_\alpha(k, n-2k)\ thì\ bác\ bỏ\ H_0.$

Lưu ý: *Kiểm định Chow nói trên có thể mở rộng cho nhiều thời kỳ*

4.2.3.2. Phương pháp biến giả
Tất cả các quan sát của cả 2 thời kỳ được gộp lại, với các thuộc tính về thời kỳ được biểu diễn bởi biến giả:

$$D(x) = \begin{cases} 1, nếu\ quan\ sát\ x\ ở\ thời\ kỳ\ trước \\ 0, nếu\ quan\ sát\ x\ ở\ thời\ kỳ\ sau. \end{cases}$$

Ta sử dụng mô hình hồi quy dịch chuyển cả tung độ gốc và độ dốc:

$$Y = a_0 + a_1.D + b_0.X + b_1(D.X) + U$$

với a_1 biểu thị chênh lệch về tung độ gốc, b_1 biểu thị chênh lệch về độ dốc.
Khi đó kiểm định giả thuyết không có sự khác nhau về mặt cấu trúc hồi quy giữa hai thời kỳ thực chất là kiểm định giả thuyết $H_0: a_1 = b_1 = 0.$

4.2.4. Hồi quy tuyến tính từng khúc (*Piecewise linear regression*)
Nếu có sự thay đổi về cấu trúc của hàm hồi quy trên các khoảng giá trị khác nhau của biến giải thích thì ta có thể dùng kỹ thuật biến giả để thiết lập một hàm hồi quy tuyến tính từng khúc để biểu diễn chung cho các cấu trúc khác nhau. Lúc đó ta coi các khoảng giá trị khác nhau nói trên của biến giải thích như là các thuộc tính khác nhau của một biến định tính.
Giả sử biến Y có mức độ phụ thuộc khác nhau đối với biến X ở các khoảng giá trị [0, T_1], (T_1, T_2], (T_2, $+\infty$) (chẳng hạn: Y là mức thuế thu nhập cá nhân, X là thu nhập cá nhân; Y là mức chi trả hoa hồng, X là mức doanh thu;....). Ta thường gọi T_1, T_2 là các ngưỡng của X.
Khi đó ta sử dụng 2 biến giả D_1, D_2 như sau:

$$D_1 = \begin{cases} 1, nếu\ T_1 < X \leq T_2 \\ 0, nếu\ X \notin (T_1, T_2] \end{cases}; D_2 = \begin{cases} 1, nếu\ X > T_2 \\ 0, nếu\ X \leq T_2 \end{cases}$$

ta có mô hình hồi quy tuyến tính từng khúc:

$$Y = a + b_1.X + b_2.(X - T_1).D_1 + b_2.(X - T_2).D_2 + U$$

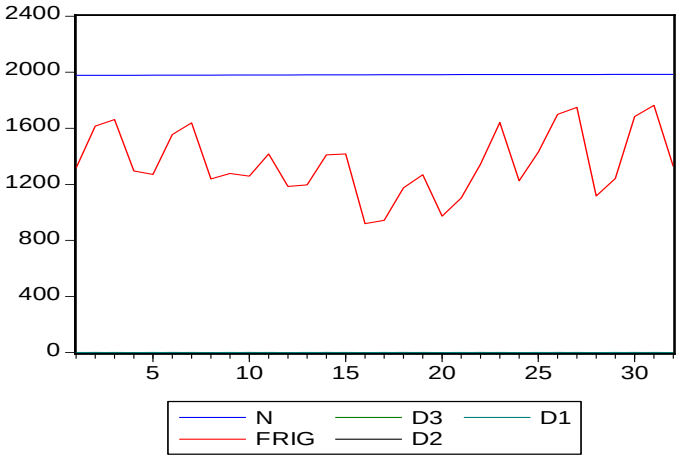
4.2.5. Phân tích mùa

Trong thực tế, đặc biệt là trong kinh tế có nhiều biến quan sát là biến chuỗi thời gian mang tính chất thời vụ, chẳng hạn doanh số bán văn phòng phẩm vào những ngày đầu năm học, doanh số bán hàng giải khát, hàng điện lạnh vào những ngày nóng bức, lượng du khách đến một điểm du lịch trong những ngày lễ hội,.... Để tách biệt tác động của yếu tố thời vụ (hay nhân tố mùa) trong chuỗi thời gian để tập trung vào các thành phần khác của số liệu (như chu kỳ, xu hướng, ngẫu nhiên,...) ta có thể sử dụng phương pháp biến giả.

Ví dụ 4.4: Khảo sát số lượng tủ lạnh Y(ngàn cái) bán được tại Mỹ từ quý 1 năm 1978 đến quý 4 năm 1985, ta có bảng số liệu 4.1 sau đây:

Năm:quý	FRIG	D1	D2	D3	Năm: quý	FRIG	D1	D2	D3
1978-1	1317	1	0	0	1982-1	943	1	0	0
1978-2	1615	0	1	0	1982-2	1175	0	1	0
1978-3	1662	0	0	1	1982-3	1269	0	0	1
1978-4	1295	0	0	0	1982-4	973	0	0	0
1979-1	1271	1	0	0	1983-1	1102	1	0	0
1979-2	1555	0	1	0	1983-2	1344	0	1	0
1979-3	1639	0	0	1	1983-3	1641	0	0	1
1979-4	1238	0	0	0	1983-4	1225	0	0	0
1980-1	1277	1	0	0	1984-1	1429	1	0	0
1980-2	1258	0	1	0	1984-2	1699	0	1	0
1980-3	1417	0	0	1	1984-3	1749	0	0	1
1980-4	1185	0	0	0	1984-4	1117	0	0	0
1981-1	1196	1	0	0	1985-1	1242	1	0	0
1981-2	1410	0	1	0	1985-2	1684	0	1	0
1981-3	1417	0	0	1	1985-3	1764	0	0	1
1981-4	919	0	0	0	1985-4	1328	0	0	0

Bảng 4.1



Hình 4.6. Đồ thị (FRIG) bán ở Mỹ theo các quý từ quý 1-1978 đến quý 4-1985

Các biến giả:

$$D_1(x) = \begin{cases} 1, nếu\ x\ được\ bán\ ở\ quý\ 1 \\ 0, nếu\ x\ được\ bán\ ở\ các\ quý\ khác \end{cases}$$

$$D_2(x) = \begin{cases} 1, nếu\ x\ bán\ ở\ quý\ 2 \\ 0, nếu\ x\ bán\ ở\ các\ quý\ khác \end{cases};\ D_3(x) = \begin{cases} 1, nếu\ x\ bán\ ở\ quý\ 3 \\ 0, nếu\ x\ bán\ ở\ các\ quý\ khác \end{cases}$$

Hàm hồi quy có dạng: $Y = a + b_1.D_1 + b_2.D_2 + b_3.D_3 + U$
Sử dụng phần mềm Eviews ta có bảng kết quả hồi quy 4.2 dưới đây.

Dependent Variable: FRIG
Method: Least Squares
Sample: 1978Q1 1985Q4
Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1160.000	59.99041	19.33642	0.0000
D1	62.12500	84.83926	0.732267	0.4701
D2	307.5000	84.83926	3.624501	0.0011
D3	409.7500	84.83926	4.829722	0.0000

R-squared	0.531797	Mean dependent var	1354.844
Adjusted R-squared	0.481632	S.D. dependent var	235.6719
S.E. of regression	169.6785	Akaike info criterion	13.22216
Sum squared resid	806142.4	Schwarz criterion	13.40537
Log likelihood	-207.5545	Hannan-Quinn criter.	13.28289
F-statistic	10.60102	Durbin-Watson stat	0.392512
Prob(F-statistic)	0.000079		

Bảng 4.2.

Theo đó ta có hàm hồi quy SRF: $\hat{Y} = 1160 + 62{,}125.D_1 + 307{,}5.D_2 + 409{,}75.D_3$
Ngoài ra:
Q1: bình quân số tủ lạnh bán được là: 1160000 + 62125 = 1222125 (cái)
Q2: bình quân số tủ lạnh bán được là: 1160000 + 307500 = 1467500 (cái)
Quý 3: bình quân số tủ lạnh bán được: 1160000 + 409750 = 1569750 (cái)
Quý 4: bình quân số tủ lạnh bán được là: 1160000 (cái)
Các hệ số của các biến giả D_1, D_2, D_3 cho biết lượng chênh lệch của bình quân số tủ lạnh bán được ở quý 1, quý 2, quý 3 so với quý 4.
Nhận thấy rằng: Hệ số hồi quy của D_1 có *p-value* = 0,4701 là khá lớn, tức là giá trị của nó khác 0 không có ý nghĩa thống kê nên ta thừa nhận hệ số hồi quy của D_1 trong PRF bằng 0. Điều này có nghĩa là bình quân số tủ lạnh bán được trong quý 1 và trong quý 4 không có sự khác biệt đáng kể. Trong khi đó các hệ số hồi quy của D_2 và D_3 đều có ý nghĩa thống kê (có p-value khá bé), tức là bình quân số tủ lạnh bán được ở quý 2 và ở quý 3 có sự khác biệt đáng kể so với quý 4. Như vậy ở đây tác động của yếu tố thời vụ (mùa) ảnh hưởng đến quý 2 và quý 3: nhu cầu về tủ lạnh về mùa xuân và mùa hè (ứng với quý 2 và quý 3 ở Mỹ) nhiều hơn về mùa đông và mùa thu (ứng với quý 1 và quý 4).
Việc điều chỉnh thời vụ được thể hiện qua việc điều chỉnh chuỗi số liệu như sau: lấy phần dư (resid) của mô hình hồi quy (là chênh lệch giữa số tủ lạnh thực tế bán được và lượng tủ lạnh trung bình bán được mỗi quý) cộng với giá trị trung bình của biến phụ thuộc Y. Chuỗi số liệu sau khi điều chỉnh có thể thể hiện sự tác động của các thành phần khác trong chuỗi số liệu như chu kỳ, xu hướng,.... Trong các bảng sau đây, bảng 4.3. thao tác quá trình điều chỉnh số liệu.

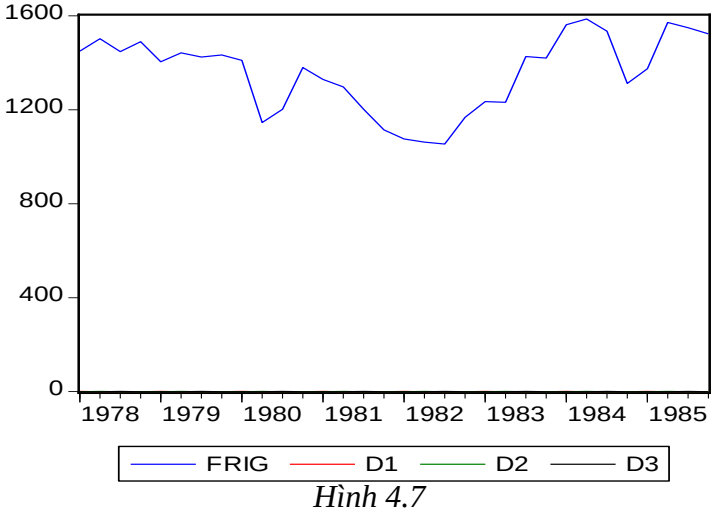
N:quý	Frig	Tb.Quý	Tb.Frig	Frig.đc	N:quý	Frig	Tb.Quý	Tb.Frig	Frig.đc
1978-1	1317	1222.125	1354.844	1449.719	1982-1	943	1222.125	1354.844	1075.719
1978-2	1615	1467.500	1354.844	1502.344	1982-2	1175	1467.500	1354.844	1062.344
1978-3	1662	1569.750	1354.844	1447.094	1982-3	1269	1569.750	1354.844	1054.094
1978-4	1295	1160.000	1354.844	1489.844	1982-4	973	1160.000	1354.844	1167.844
1979-1	1271	1222.125	1354.844	1403.719	1983-1	1102	1222.125	1354.844	1234.719
1979-2	1555	1467.500	1354.844	1442.344	1983-2	1344	1467.500	1354.844	1231.344
1979-3	1639	1569.750	1354.844	1424.094	1983-3	1641	1569.750	1354.844	1426.094
1979-4	1238	1160.000	1354.844	1432.844	1983-4	1225	1160.000	1354.844	1419.844
1980-1	1277	1222.125	1354.844	1409.719	1984-1	1429	1222.125	1354.844	1561.719
1980-2	1258	1467.500	1354.844	1145.344	1984-2	1699	1467.500	1354.844	1586.344
1980-3	1417	1569.750	1354.844	1202.094	1984-3	1749	1569.750	1354.844	1534.094
1980-4	1185	1160.000	1354.844	1379.844	1984-4	1117	1160.000	1354.844	1311.844
1981-1	1196	1222.125	1354.844	1328.719	1985-1	1242	1222.125	1354.844	1374.719
1981-2	1410	1467.500	1354.844	1297.344	1985-2	1684	1467.500	1354.844	1571.344
1981-3	1417	1569.750	1354.844	1202.094	1985-3	1764	1569.750	1354.844	1549.094
1981-4	919	1160.000	1354.844	1113.844	1985-4	1328	1160.000	1354.844	1522.844

Bảng 4.3.

Bảng 4.4 dưới đây là bảng số liệu sau khi đã điều chỉnh, hình 4.7. biểu diễn đồ thị của lượng tủ lạnh bán được theo các quí từ quý 1-1978 đến quý 4-1985 sau khi đã có sự điều chỉnh.

N:quý	Frig.đc	D1	D2	D3	N:quý	Frig.đc	D1	D2	D3
1978-1	1449.719	1	0	0	1982-1	1075.719	1	0	0
1978-2	1502.344	0	1	0	1982-2	1062.344	0	1	0
1978-3	1447.094	0	0	1	1982-3	1054.094	0	0	1
1978-4	1489.844	0	0	0	1982-4	1167.844	0	0	0
1979-1	1403.719	1	0	0	1983-1	1234.719	1	0	0
1979-2	1442.344	0	1	0	1983-2	1231.344	0	1	0
1979-3	1424.094	0	0	1	1983-3	1426.094	0	0	1
1979-4	1432.844	0	0	0	1983-4	1419.844	0	0	0
1980-1	1409.719	1	0	0	1984-1	1561.719	1	0	0
1980-2	1145.344	0	1	0	1984-2	1586.344	0	1	0
1980-3	1202.094	0	0	1	1984-3	1534.094	0	0	1
1980-4	1379.844	0	0	0	1984-4	1311.844	0	0	0
1981-1	1328.719	1	0	0	1985-1	1374.719	1	0	0
1981-2	1297.344	0	1	0	1985-2	1571.344	0	1	0
1981-3	1202.094	0	0	1	1985-3	1549.094	0	0	1
1981-4	1113.844	0	0	0	1985-4	1522.844	0	0	0

Bảng 4.4



Hình 4.7

****

Bài tập.

1. Ta có bảng số liệu về lượng hàng bán được Y(tấn/tháng) của một loại hàng và thu nhập X(triệu đồng/tháng) của người tiêu dùng ở 2 khu vực thành phố và nông thôn:

Y	6	8	10	10	12	7	5	8
X	3	4	6	7	8	4	3	5
Nơi bán	TP	NT	TP	NT	TP	TP	NT	NT

a. Thiết lập hàm PRF tuyến tính biểu thị sự phụ thuộc của lượng hàng bán được theo thu nhập của người tiêu dùng và khu vực bán.
b. Thiết lập hàm hồi quy mẫu từ bảng số liệu
c. Nêu ý nghĩa của các hệ số hồi quy riêng
d. Dựa vào điều tra, với độ tin cậy 95%, hãy ước lượng KTC cho mức tăng bình quân của lượng hàng bán được khi thu nhập tăng 1 triệu đồng, ước lượng KTC cho mức chênh lệch bình quân về lượng hàng bán được giữa khu vực nông thôn so với khu vực thành phố.
e. Dựa vào số liệu trên, theo bạn thì nơi bán có ảnh hưởng tới lượng hàng bán được hay không?
f. Hàm hồi quy mẫu được thiết lập có phù hợp với mẫu hay không?
g. Hãy ước lượng KTC cho phương sai nhiễu với độ tin cậy 90%.
h. Xác định tổng bình phương các độ lệch của Y.

2. Số liệu về lợi nhuận Y(tỷ VNĐ) và doanh thu X(tỷ VNĐ) của một số doanh nghiệp thuộc một ngành dịch vụ ở Tp. Hồ Chí Minh năm 2004 cho ở bảng sau:

Y	15	17	20	21	24	26	27	35
X	120	130	145	149	155	162	165	174
Chủ doanh nghiệp	Nữ	Nam	Nam	Nữ	Nữ	Nam	Nữ	Nam

Bảng 4.9

a/ Hãy thiết lập mô hình hồi quy tuyến tính SRF ngẫu nhiên biểu diễn sự phụ thuộc của lợi nhuận theo doanh thu và giới tính của chủ doanh nghiệp.
b/ Cho biết ý nghĩa của các hệ số hồi quy trong mô hình này.

3. Sự phụ thuộc của tiền lương Y(USD/ tháng) vào số năm giáo dục X_1 vượt quá lớp 8 khi được thuê, số năm làm việc X_2 tại công ty, tuổi X_3 của người lao động, giới tính, chủng tộc (da trắng, không phải da trắng), nhân viên văn phòng hay không là nhân viên văn phòng, làm nghề thủ công hay không làm nghề thủ công, công việc bảo trì hay không phải công việc bảo trì, qua số liệu điều tra có kết quả hồi quy sau đây (trong đó:
$D_1 = 1$: nam, $D_1 = 0$: nữ; $D_2 = 1$: da trắng. $D_2 = 0$:không phải da trắng; $D_3 = 1$: nhân viên vp, $D_3 = 0$: không là nhân viên vp; $D_4 = 1$: việc bảo trì, $D_4 = 0$: không lv bảo trì; $D_5 = 1$: nghề thủ công, $D_5 = 0$: không phải nghề thủ công). Từ các kết quả hồi quy dưới đây:
a/ Hãy thiết lập mô hình hồi quy PRF của Y theo các biến X_1, X_2, X_3, có sự tác động của các yếu tố: giới tính, màu da, các tính chất công việc.

b/ Từ kết quả chạy hồi quy, hãy thiết lập hàm hồi quy ước lượng SRF của Y theo các biến đã chỉ ra. Nêu ý nghĩa của các hệ số hồi quy.
c/ Phân tích kết quả hồi quy ở bảng 4.7. Theo bạn yếu tố giới tính có ảnh hưởng đến tiền lương hay không?
d/ Phân tích kết quả hồi quy ở bảng 4.8.

Dependent Variable: Y
Method: Least Squares
Sample: 1 49
Included observations: 49

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X3	-8.669319	5.778252	-1.500336	0.1414
X2	33.35794	10.18345	3.275702	0.0022
X1	43.17623	27.88269	1.548496	0.1294
D5	-763.3644	177.1858	-4.308271	0.0001
D4	-1074.695	200.8963	-5.349500	0.0000
D3	-938.9372	172.1240	-5.455005	0.0000
D2	241.4220	130.5250	1.849622	0.0718
D1	527.0849	154.3649	3.414540	0.0015
C	1954.029	334.7502	5.837274	0.0000

R-squared	0.751501	Mean dependent var	1820.204
Adjusted R-squared	0.701801	S.D. dependent var	648.2687
S.E. of regression	354.0041	Akaike info criterion	14.74090
Sum squared resid	5012756.	Schwarz criterion	15.08838
Log likelihood	-352.1521	F-statistic	15.12078
Durbin-Watson stat	2.014802	Prob(F-statistic)	0.000000

Bảng 4.7

Dependent Variable: Y
Method: Least Squares
Sample: 1 49
Included observations: 49

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X1	108.0616	32.43961	3.331162	0.0017
D1	549.0727	152.7324	3.594998	0.0008
C	856.2312	227.8354	3.758112	0.0005

R-squared	0.351727	Mean dependent var	1820.204
Adjusted R-squared	0.323541	S.D. dependent var	648.2687
S.E. of regression	533.1824	Akaike info criterion	15.45487
Sum squared resid	13077038	Schwarz criterion	15.57070
Log likelihood	-375.6444	F-statistic	12.47887
Durbin-Watson stat	1.815936	Prob(F-statistic)	0.000047

Bảng 4.8

4. Số liệu về doanh thu Y(triệu VNĐ) của một công ty cho ở bảng sau:

Năm	Quý			
	1	2	3	4
1999	632	794	767	870
2000	905	1255	1394	1488
2001	1828	2006	2443	2260
2002	2685	3212	3230	3118
2003	3096	3412	3618	3470

Bảng 4.10

a/ Đưa vào biến xu thế t (t =1 ứng với quan sát thứ nhất,..., t = 20 ứng với quan sát thứ 20 của mẫu). Dựa vào bảng số liệu, sử dụng phương pháp OLS, hãy thiết lập mô hình hồi quy SRF tuyến tính của Y theo biến xu thế t.

b/ Thiết lập mô hình hồi quy SRF tuyến tính của doanh thu Y theo biến xu thế t và các biến giả biểu diễn các thuộc tính về quý. Cho biết ý nghĩa của các hệ số hồi quy trong mô hình này. Từ đó dự báo cho doanh thu của quý 4/2004

5. Kết quả hồi quy về doanh số bán hàng Y(tỷ VNĐ) của một siêu thị theo các biến t, D1, D2, D3, qua số liệu của các quý từ năm 2001 đến 2004 được cho như sau:

Dependent Variable: Y
Method: Least Squares
Sample: 1 16
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.408562	1.370282	5.406597	0.0002
t	2.001444	0.102135	19.59610	0.0000
D1	8.296581	1.327752	6.248591	0.0001
D2	6.908388	1.307963	5.281789	0.0003
D3	0.261944	1.295945	0.202126	0.8435

R-squared	0.973266	Mean dependent var	28.28756
Adjusted R-squared	0.963545	S.D. dependent var	9.569037
S.E. of regression	1.827043	Akaike info criterion	4.293581
Durbin-Watson stat	0.964741	Prob(F-statistic)	0.000000

Bảng 4.11

trong đó t là biến xu thế (t = 1 ứng với quan sát thứ nhất,..., t = 16 ứng với quan sát thứ 16), D1, D2, D3 là các biến giả: D1 = 1 đối với quý 1, D1 = 0 với quý khác, D2 = 1 đối với quý 2, D2 = 0 với quý khác, D3 = 1 đối với quý 3, D3 = 0 với các quý khác.

a/ Từ kết quả trên, hãy thiết lập mô hình hồi quy ước lượng của doanh số theo t và D1, D2, D3. Cho biết ý nghĩa của các hệ số hồi quy và đánh giá mức độ phù hợp của mô hình này.

b/ Với độ tin cậy 95%, từ kết quả trên, hãy ước lượng khoảng tin cậy cho: chênh lệch doanh số bán bình quân của quý 1, quý 2, quý 3 so với quý 4 của siêu thị này.

c/ Dùng mô hình trên để dự báo doanh số của siêu thị trong quý 1, quý 2, quý 3, quý 4 của năm 2005.

6. Có bảng số liệu sau về: Tiền lương Y (triệu đồng), số năm công tác X (năm), trình độ, hệ số chức vụ (đơn vị: bậc) của một số giáo viên như sau:

Y	5	4.7	6	6.5	6	6.2	7	8	9	8
X	1	2	3	4	5	6	7	8	9	10
Tr.độ	TS	ThS	TS	TS	ThS	ThS	TS	TS	TS	ThS
Z	1	0	2	2	1	1	3	4	4	3

a/ Thiết lập PRF tuyến tính mô tả sự phụ thuộc của tiền lương vào số năm công tác, trình độ, hệ số chức vụ. Nêu ý nghĩa của các hệ số hồi quy.
b/ Chạy hồi quy để thiết lập SRF ước lượng cho PRF nói trên.
c/ Thiết lập PRF tuyến tính mô tả sự phụ thuộc của tiền lương vào số năm công tác và trình độ.
d/ Chạy hồi quy để thiết lập SRF ước lượng cho PRF ở c/
e/ Bạn chọn mô hình nào trong hai mô hình SRF nói trên, tại sao?

7. Với số liệu điều tra về tiền lương Y (triệu đồng), số năm công tác X, trình độ Z (Z = 0, nếu là ThS, Z = 1, nếu là TS) và giới tính S (S = 0, nếu là nữ, S = 1, nếu là nam) của 20 giáo viên, có kết quả hồi quy như sau:
$\hat{Y} = 4{,}303 + 0.391X - 0.434S + 0.674Z;$
$t = \ 6.503 \quad 10.964; \ -1.940; \quad 3.299$
$R^2 = 0.9950$
a/ Viết SRF của một giáo viên nam có trình độ TS
b/ Cho biết SRF của một giáo viên nữ có trình độ ThS
c/ Tìm khoảng tin cậy 95% cho mức chênh lệch bình quân về lương của giáo viên nam so với giáo viên nữ cùng thâm niên và cùng trình độ.

8. Sự phụ thuộc của chi tiêu Y (triệu đồng/tháng) cho mặt hàng A đối với thu nhập X (triệu đồng/tháng) và giới tính S của người tiêu dùng (S = 0, nếu là nữ, S = 1, nếu là nam), qua mẫu điều tra 20 khách hàng, có kết quả hồi quy như sau:

$$\hat{Y} = 6.426 + 0.098.X - 0.025.XS + 2.453.S$$
$$\widehat{se} = 3.628; \quad 0.032; \quad 0.011; \quad 0.988$$

a/ Cho biết ý nghĩa của các hệ số hồi quy trên
b/ Hãy ước lượng khoảng tin cậy 95% cho các hệ số hồi quy
c/ Hãy cho biết chi tiêu về mặt hàng này của nam và nữ có thực sự khác nhau hay không.
HD: Viết lại SRF dưới dạng: $\hat{Y} = \quad (6.426 + 2.453.S) + (0.098 - 0.025.S).X$. Từ đó:
- Hệ số hồi quy của S là 2.453 là mức chênh lệch của tung độ gốc của hai đường thẳng hồi quy SRF của nam và của nữ, nó phản ánh mối quan hệ giữa chi tiêu mặt hàng A đối với thu nhập của nam và nữ.

- Hệ số hồi quy của biến XS là – 0.025 là mức chênh lệch bình quân về chi tiêu mặt hàng A của nam so với nữ khi thu nhập tăng lên 1 triệu đồng, nó phản ánh chênh lệch về tốc độ chi tiêu mặt hàng này của nữ so với nam theo sự gia tăng của thu nhập.

9. Kết quả hồi quy của lượng khách đi xe bus Y(triệu lượt người/năm), X1 là giá vé xe bus(ngàn đồng/lượt), X2 là giá xăng(ngàn đồng/lít), Z = 0: xe bus loại lớn, Z = 1: xe bus loại nhỏ, qua 20 quan sát được cho như sau:

Dependent Variable: Y
Method: Least Squares
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.522383	.....	6.314238	0.0001
X1	.....	0.091549	-6.514679	0.0000
X2	0.159738	0.058358	.....	0.0193
Z	.....	0.311234	3.671297	0.0037

R-squared	0.906713	Mean dependent var	6.800000
Adjusted R-squared	0.881272	S.D. dependent var	1.612452
S.E. of regression	0.555602	Akaike info criterion	1.885650
Durbin-Watson stat	2.097768	Prob(F-statistic)	0.000006

a/ Hãy điền vào chỗ trống các thông tin phù hợp trong bảng hồi quy nói trên và thiết lập mô hình hồi quy tuyến tính SRF của lượng khách đi xe bus Y theo giá vé xe bus , giá xăng và loại xe Z, nêu ý nghĩa của các hệ số hồi quy. Đánh giá mức độ phù hợp của mô hình.
b/ Nêu ý nghĩa của các hệ số hồi quy ước lượng của các biến.
c/ Dựa vào kết quả trên, với độ tin cậy 95%, hãy ước lượng khoảng tin cậy cho mức chênh lệch bình quân của lượng khách đi xe nhỏ và lượng khách đi xe lớn (trong điều kiện các yếu tố khác không đổi).

****

Chương 5.

MỘT SỐ VẤN ĐỀ TRONG MÔ HÌNH HỒI QUY

Chương này đề cập tới ba vấn đề thường xảy ra trong mô hình, vi phạm giả thiết của phương pháp OLS: Đa cộng tuyến, phương sai nhiễu thay đổi, tự tương quan của nhiễu. Đồng thời, trong một chừng mực nào đó, chỉ ra nguyên nhân, phát hiện vấn đề và tìm cách khắc phục, hạn chế những hậu quả không tốt của chúng.

5.1. Đa cộng tuyến

5.1.1. Khái niệm về đa cộng tuyến

a. Xét mô hình hồi quy k biến:

$$Y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_{k-1}X_{k-1} + U \qquad (5.1)$$

Giả thiết 4 của phương pháp OLS là ma trận

$$\mathcal{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k-1,1} \\ 1 & X_{12} & \cdots & X_{k-1,2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{1n} & \cdots & X_{k-1,n} \end{pmatrix}$$

có hạng bằng k, tức là k cột của ma trận này không phải là k véc tơ phụ thuộc tuyến tính.

Khi các biến giải thích không tương quan với nhau, mỗi biến chứa đựng những thông tin riêng về Y, không liên quan đến các biến giải thích khác. Khi đó hệ số hồi quy riêng của mỗi biến giải thích cho biết ảnh hưởng của biến này đối với biến phụ thuộc khi các biến khác không đổi. Trong trường hợp này ta nói mô hình không có hiện tượng đa cộng tuyến.

Ta nói mô hình có hiện tượng đa cộng tuyến (*multicollinearity*) nếu tồn tại các hằng số không đồng thời bằng 0: $\lambda_1, \lambda_2, \dots, \lambda_{k-1}$ *và biến ngẫu nhiên* ε *sao cho*:

$$\lambda_1.X_1 + \lambda_2.X_2 + \cdots + \lambda_{k-1}.X_{k-1} = \varepsilon$$

- Khi $\varepsilon \equiv 0$ thì hiện tượng đa cộng tuyến được gọi là đa cộng tuyến hoàn hảo (*perfect multicollinearity*) (Khi đó rõ ràng giả thiết 4 nói trên bị vi phạm)
- Khi $\varepsilon \not\equiv 0$ thì hiện tượng đa cộng tuyến được gọi là đa cộng tuyến không hoàn hảo, (*imperfect multicollinearity*), hay đơn giản là đa cộng tuyến.

b. Nguyên nhân của hiện tượng đa cộng tuyến:

Những nguyên nhân chính là:

* Khi các biến giải thích có mối quan hệ nhân quả cao, tức là có những quan hệ ràng buộc.

Chẳng hạn: trong mô hình hồi quy của Y là lượng điện năng tiêu thụ theo các biến giải thích là: thu nhập X_1, diện tích nhà ở X_2, thì sẽ xảy ra hiện tượng đa cộng tuyến vì thu nhập cao thường kéo theo diện tích nhà ở lớn hơn.

* Khi các số liệu quá ít thì chúng vừa không đủ tính đại diện cho tổng thể, lại không xác định được duy nhất các hệ số hồi quy.
* Chọn biến giải thích có độ biến thiên nhỏ.

* Phương pháp chọn mẫu không đủ tính đại diện.

5.1.2. Hậu quả của đa cộng tuyến

Trong thực tế hiện tượng đa cộng tuyến là không tránh khỏi, vấn đề là mức độ đa cộng tuyến là cao hay thấp. Khi mô hình có hiện tượng đa cộng tuyến đáng kể giữa các biến giải thích thì mặc dù tính chất BLUE của các hệ số ước lượng vẫn được bảo toàn, nhưng xuất hiện các hậu quả không tốt sau:

1/ Các hệ số ước lượng có phương sai và hiệp phương sai lớn, nghĩa là các ước lượng này có giá trị thay đổi nhiều từ mẫu này qua mẫu khác, khiến độ chính xác của các ước lượng không cao.

Để thấy rõ điều này, xét mô hình SRF ba biến: $\hat{Y} = \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2$, ta có:

$$var(\hat{a}_1) = \frac{\sigma^2}{nS^2(X_1).(1-r_{12}^2)};\ \ var(\hat{a}_2) = \frac{\sigma^2}{nS^2(X_2).(1-r_{12}^2)}; \qquad (*)$$

$$cov(\hat{a}_1, \hat{a}_2) = \frac{-r_{12}\sigma^2}{nS(X_1).S(X_2).(1-r_{12}^2)};$$

trong đó r_{12} là hệ số tương quan mẫu giữa X_1, X_2 . Khi mô hình có hiện tượng đa cộng tuyến cao thì $|r_{12}|$ gần đến 1, do đó giá trị tuyệt đối của các biểu thức trên trở nên rất lớn.

2/ Từ hậu quả trên mà khoảng tin cậy cho các hệ số hồi quy rộng hơn, nghĩa là ước lượng có độ chính xác kém đi.

3/ Khi sử dụng thống kê $t = \frac{\hat{b}_j - b_j^*}{se\hat{b}_j}$ để *kiểm định giả thuyết* $H_0: b_j = b_j^*$, nếu có đa cộng tuyến ở mức độ cao thì các sai số chuẩn của các ước lượng có xu hướng tăng cao, dẫn tới giá trị $|t|$ có xu hướng nhỏ đi, do đó ta có xu hướng chấp nhận giả thuyết H_0.

4/ Trong khi $|t|$ bé đi thì hệ số xác định R^2 có thể rất cao, dẫn tới những kết luận không phù hợp với thực tế.

5/ Dấu của các hệ số hồi quy ước lượng có thể sai

6/ Các ước lượng $\hat{b}_j$ qua OLS cho các hệ số hồi quy và $se(\hat{b}_j)$ trở nên rất nhạy với những thay đổi nhỏ trong số liệu.

7/ Do các hậu quả trên mà khi thêm vào hay bớt đi các biến cộng tuyến với các biến khác thì mô hình sẽ có sự thay đổi về dấu hoặc độ lớn của các ước lượng.

5.1.3. Cách phát hiện đa cộng tuyến

Như đã chỉ ra, hiện tượng đa cộng tuyến là không tránh khỏi. Người ta đưa ra một số quy tắc kinh nghiệm nhằm phát hiện và đánh giá mức độ đa cộng tuyến như sau.

a/ *Hệ số xác định R^2 cao nhưng giá trị $|t|$ thấp*: đây là một điều mâu thuẫn trong mô hình mà mức độ đa cộng tuyến thấp hoặc không có. Khi $R^2 > 0,8$ thì thường giả thuyết về các hệ số hồi quy đồng thời bằng 0 bị bác bỏ, nhưng khi $|t|$ có giá trị bé thì lại có xu hướng chấp nhận giả thuyết nói trên. Hiện tượng này chỉ thể hiện rõ khi có đa cộng tuyến ở mức độ cao.

b/ *Các cặp biến giải thích có hệ số tương quan cao:* Khi thấy hệ số tương quan cặp giữa các biến giải thích > 0,8 thì kinh nghiệm cho thấy hiện tương đa cộng tuyến trở nên nghiêm trọng (tuy nhiên đây chỉ là điều kiện cần nếu mô hình nhiều hơn 2 biến).

c/ *Sử dụng các hồi quy phụ*: Chạy mô hình hồi quy của một biến giải thích X_j với các biến giải thích còn lại (gọi là hồi quy phụ), ta nhận được hệ số xác định của mô hình này,

ký hiệu là R_j^2. Theo quy tắc " *ngón tay cái*" (*Rule of Thumb*) của Klein, hiện tượng đa cộng tuyến là nghiêm trọng chỉ nếu có hệ số xác định R_j^2 của hồi quy phụ nào đó vượt quá hệ số xác định R^2 của mô hình hồi quy chính của biến phụ thuộc.
d/ *Sử dụng nhân tử phóng đại phương sai VIF*: Nhân tử phóng đại của hồi quy phụ của biến X_j là: $$\boldsymbol{VIF_j = \frac{1}{1-R_j^2}}$$
Quy tắc kinh nghiệm là khi $VIF_j > 10$ *hay* $R_j^2 > 0{,}9$ thì dễ có hiện tượng đa cộng tuyến ở mức độ cao.

5.1.4. Biện pháp khắc phục đa cộng tuyến

Nếu mục tiêu của phân tích hồi quy là dự báo thì trong một số trường hợp, ta không cần khắc phục đa cộng tuyến. Nếu mục tiêu của phân tích hồi quy là xét tác động riêng của từng biến giải thích lên biến phụ thuộc để quyết định chính sách thì đa cộng tuyến trở thành một vấn đề nghiêm trọng. Sau đây là một số biện pháp khắc phục.
a/ *Dùng thông tin tiên nghiệm* (*A priori information*) Thông tin tiên nghiệm có thể nhận được từ các nghiên cứu thực nghiệm trước đây, hoặc từ các lý thuyết liên quan đến các biến giải thích ta đang xét.
Chẳng hạn khi nghiên cứu hàm sản xuất Cobb – Douglas ở Mexico giai đoạn 1955-1974 trong chương trước, ta có mối quan hệ giữa sản lượng Y (đầu ra) phụ thuộc vào các yếu tố đầu tư như lao động X_1 và vốn X_2
$$lnY = \beta_0 + \beta_1 lnX_1 + \beta_2 lnX_2 + U \ \ (\beta_0 = ln\alpha)$$
Kết quả thực nghiệm cho thấy giữa vốn và lao động ở Mexico có quan hệ là sản lượng không đổi theo quy mô, tức là: $\beta_1 + \beta_2 = 1$. Nếu sử dụng kết quả thực nghiệm này như là một thông tin tiên nghiệm thì ta có biến đổi mối quan hệ trên về dạng:
$$lnY = \beta_0 + \beta_1 lnX_1 + (1-\beta_1) lnX_2 + U \ ,$$
hay: $\ln\left(\frac{Y}{X_2}\right) = \beta_0 + \beta_1 ln(X_1/X_2) + U$ (là mô hình hồi quy 2 biến)
b/ *Tăng cỡ mẫu (bổ sung thêm số liệu) hoặc lấy thêm mẫu mới*
Với mẫu mới thu được theo cách này, trong nhiều trường hợp người ta hy vọng nó sẽ làm giảm mức độ đa cộng tuyến cao trong mẫu cũ, chẳng hạn trong các hệ thức (*) ở trên, nếu mẫu mới không làm tăng trị tuyệt đối của hệ số tương quan mẫu r_{12}, thì các biểu thức $var(\hat{a}_1);\ var(\hat{a}_2); cov(\hat{a}_1, \hat{a}_2)$ sẽ nhỏ đi do các phương sai mẫu $S^2(X_1), S^2(X_2)$ tăng lên.
c/ Kết hợp số liệu chéo và số liệu chuỗi thời gian
Trong số liệu chuỗi thời gian, thường ẩn chứa vấn đề đa cộng tuyến giữa các biến giải thích. Khi kết hợp thêm số liệu chéo, sẽ khắc phục hay hạn chế bớt mức độ đa cộng tuyến giữa các biến.
d/ Bỏ bớt biến giải thích trong các biến có cộng tuyến với nhau
Đây là biện pháp khắc phục khá đơn giản. Tuy nhiên, khi bỏ bớt biến giải thích có thể tránh được đa cộng tuyến cao, nhưng có thể gây nên hậu quả nghiêm trọng là dẫn đến ước lượng chệch nhiều so với giá trị thực của tham số cần ước lượng (trong khi vấn đề đa cộng tuyến không làm thay đổi tính không chệch của ước lượng)
e/ Sử dụng sai phân cấp 1 (first difference)

Theo diễn biến của thời gian, các biến kinh tế thường chịu ảnh hưởng của xu hướng nên dễ có tương quan với nhau. Để làm giảm sự tương quan đó, ta có thể sử dụng sai phân cấp 1.

Giả sử theo thời gian t, có mô hình: $Y_t = a_0 + a_1X_{1t} + a_2X_{2t} + U_t$

thì tại thời điểm t – 1, ta có: $Y_{t-1} = a_0 + a_1X_{1,t-1} + a_2X_{2,t-1} + U_{t-1}$

Từ đó:

$Y_t - Y_{t-1} = a_1(X_{1t} - X_{1,t-1}) + a_2(X_{2t} - X_{2,t-1}) + (U_t - U_{t-1})$ (5.2)

(5.2) được gọi là mô hình sai phân cấp 1, được sử dụng để ước lượng các tham số hồi quy a_1, a_2. Giữa X_{1t}, X_{2t} nếu có đa cộng tuyến cao thì giữa $(X_{1t} - X_{1,t-1})$, $(X_{2t} - X_{2,t-1})$ có thể không xảy ra đa cộng tuyến cao. Vì thế mô hình sai phân có thể làm giảm mức độ đa cộng tuyến.

Khi sử dụng mô hình sai phân cần lưu ý nhược điểm của nó là bậc tự do giảm đi 1 do giảm đi một quan sát khi chuyển sang mô hình sai phân, nên dễ ảnh hưởng đến kết quả ước lượng khi cỡ mẫu bé; mặc dù U_t có thể không có tự tương quan, nhưng $V_t = (U_t - U_{t-1})$ thì có thể có tự tương quan; hơn nữa việc sử dụng sai phân cấp 1 không thích hợp với số liệu chéo.

f/ Thay đổi dạng hàm hồi quy: Nếu ở dạng hàm hồi quy này, các biến giải thích có hiện tượng đa cộng tuyến, thì chuyển sang dạng khác có thể khắc phục được hiện tượng này.

g/ Một số biện pháp khác: Ngoài các biện pháp nói trên, để khắc phụ vấn đề đa cộng tuyến, tùy vào các trường hợp cụ thể, người ta còn sử dạng các biện pháp khác như: Sử dụng hàm hồi quy độ lệch theo giá trị trung bình trong hồi quy đa thức, hồi quy thành phần chính, hồi quy dạng sóng,…

Khắc phục hiện tượng đa cộng tuyến đòi hỏi các kỹ thuật phức tạp và đôi khi không mang lại hiệu quả như mong muốn. Hơn nữa hầu hết mô hình hồi quy bội đều có tính đa cộng tuyến nhất định nên ta phải thận trọng trong việc xây dựng mô hình và giải thích kết quả.

Ví dụ 5.1: Khi nghiên cứu về quan hệ giữa tiêu dùng nội địa Y(USD), thu nhập X_1 từ lương, thu nhập khác X_2 từ phi nông nghiệp và thu nhập X_3 từ nông nghiệp của nền kinh tế Mỹ từ năm 1928

đến 1950, với số liệu của các năm 1942 đến 1944 bị loại ra khỏi dữ liệu, từ bảng số liệu:

N	Y	X_1	X_2	X_3	N	Y	X_1	X_2	X_3
1928	52.8	39.21	17.73	4.39	1938	63.9	44.16	15.92	4.37
1929	62.2	42.31	20.29	4.6	1939	67.5	47.68	17.59	4.51
1930	58.6	40.37	18.83	3.25	1940	71.3	50.79	18.49	4.9
1931	56.6	39.15	17.44	2.61	1941	76.6	57.78	19.18	6.37
1932	51.6	34	14.76	1.67	1942	86.3	78.97	19.12	8.42
1033	51.1	33.59	13.39	2.44	1946	95.7	73.54	19.76	9.27
1034	54	36.88	13.93	2.39	1947	98.3	74.92	17.55	8.87
1035	57.2	39.27	14.67	5	1948	100.3	74.01	19.17	9.3
1936	62.8	45.51	17.20	3.93	1949	103.2	75.51	20.20	6.95
1937	65	46.06	17.15	5.48	1950	108.9	80.97	22.12	7.15

Bảng 5.1

Klein và Golberger (1995) đã thực hiện hồi quy tiêu dùng Y theo 3 loại thu nhập trên như

sau: $Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + U$

Dependent Variable: Y
Method: Least Squares
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.304002	8.882885	0.822256	0.4230
X1	1.135052	0.172127	6.594285	0.0000
X2	0.405300	0.645026	0.628347	0.5386
X3	-0.405888	1.105135	-0.367274	0.7182
R-squared	0.954028	Mean dependent var		72.19500
Adjusted R-squared	0.945409	S.D. dependent var		19.34671
S.E. of regression	4.520317	Akaike info criterion		6.031898
Sum squared resid	326.9323	Schwarz criterion		6.231044

Bảng 5.2. Kết quả hồi quy của tiêu dùng theo các loại thu nhập

Kết quả này cho thấy mô hình có tính giải thích cao thể hiện qua R^2 = 0,954028 rất cao. Tuy nhiên xuất hiện những vấn đề không phù hợp với ý nghĩa kinh tế, đó là hệ số hồi quy của X3 là – 0,405888 < 0 và hệ số hồi quy của X1 là 1,135052 cho thấy: khi thu nhập từ lương tăng 1 USD thì bình quân tiêu dùng tăng 1,135052 USD ! Đây là những biểu hiện cho thấy hồi quy này gặp phải hiện tượng đa cộng tuyến và điều này là do các loại thu nhập có xu hướng cùng tăng theo sự phát triển của kinh tế.

Xét ma trận tương quan giữa các biến:

	Y	X1	X2	X3
Y	1.000000	0.975908	0.717164	0.887671
X1	0.975908	1.000000	0.709395	0.918613
X2	0.717164	0.709395	1.000000	0.630607
X3	0.887671	0.918613	0.630607	1.000000

Bảng 5.3

ta thấy hệ số tương quan giữa X1 và X3 là 0.918613, rất cao. Đây cũng là một biểu hiện của hiện tượng đa cộng tuyến cao giữa các biến giải thích.

* Nếu bỏ bớt biến X3, ta có hồi quy:

Dependent Variable: Y
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.710141	8.586595	0.897928	0.3818
X1	1.080588	0.085136	12.69245	0.0000
X2	0.423208	0.626601	0.675402	0.5085
R-squared	0.953641	Mean dependent var		72.19500
Adjusted R-squared	0.948187	S.D. dependent var		19.34671

Bảng 5.4

theo đó ta vẫn chưa khắc phục được hiện tượng hệ số hồi quy không phù hợp với lý thuyết kinh tế: hệ số hồi quy của X1 là $\mathbf{1.080588 > 1}$.

* Nếu dùng hồi quy sai phân cấp 1, ta nhận được kết quả từ Eviews như sau:

Dependent Variable: Y-Y(-1)
Method: Least Squares
Sample (adjusted): 2 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X1-X1(-1)	0.339464	0.143755	2.361398	0.0312
X2-X2(-1)	1.515549	0.518401	2.923507	0.0099
X3-X3(-1)	0.728987	0.667778	1.091661	0.2911

R-squared	0.460379	Mean dependent var	2.952632
Adjusted R-squared	0.392926	S.D. dependent var	4.153896
S.E. of regression	3.236505	Akaike info criterion	5.330805
Sum squared resid	167.5995	Schwarz criterion	5.479927
Log likelihood	-47.64264	Hannan-Quinn criter.	5.356042
Durbin-Watson stat	1.014878		

Bảng 5.5

theo đó mô hình ít phù hợp với số liệu (R^2 = 0,460379), mặt khác hệ số hồi quy của (X2-X2(-1)) là 1,515549 > 1.

Do vậy đối với mô hình này, để khắc phục những hiện tượng trên, ta phải kết hợp các biện pháp khác nhau: bổ sung thêm số liệu, kết hợp thêm các số liệu chéo, bỏ bớt biến trong các biến có đa cộng tuyến cao, thay đổi mô hình,.... Để khắc phục hiện tượng này, ta sẽ trở lại ví dụ 5.1 trong phần sau, khi thay đổi dạng hàm hồi quy sang tuyến tính log.

5.2. Phương sai của nhiễu thay đổi

5.2.1. Khái niệm về phương sai thay đổi

Giả thiết 2 của mô hình hồi quy tuyến tính cổ điển yêu cầu phương sai của nhiễu không thay đổi qua các quan sát. Do trung bình của nhiễu bằng 0 nên yêu cầu này có nghĩa là:

$$var(U_i) = E(U_i^2) = \sigma^2$$

Trong thực tế sai số nhiễu có thể tăng, giảm khi giá trị của các biến giải thích thay đổi, tức là:

$$var(U_i) = E(U_i^2) = \sigma_i^2 \qquad (5.3)$$

Khi đó ta nói có hiện tượng phương sai nhiễu thay đổi (*heteroscedasticity*). Hiện tượng phương sai thay đổi thường gặp ở dữ liệu chéo và dữ liệu bảng.

Có thể chỉ ra những lý do sau đây:

* Do việc tích lũy kinh nghiệm hay do học được hành vi trong quá khứ mà sai số theo thời gian ngày càng giảm.Chẳng hạn đối với thợ học việc, khi số giờ thực hành càng nhiều thì số phế phẩm càng nhỏ và càng ít biến động. Trong trường hợp này phương sai nhiễu có xu hướng giảm theo thời gian.

* Do bản chất của mối liên hệ mà có nhiều mối quan hệ kinh tế đã chứa đựng hiện tượng này, khi biến kinh tế tăng kéo theo sai số nhiễu cũng tăng. Chẳng hạn khi thu nhập tăng

người ta có nhiều lựa chọn hơn trong tiêu dùng. Khi đó trong hồi quy của tiết kiệm theo thu nhập thì phương sai nhiễu có xu hướng tăng theo thu nhập.
* Khi cải thiện phương pháp và kỹ thuật thu thập số liệu thì sai số càng giảm.
* Khi trong mẫu có các số liệu vượt trội (quá lớn hoặc quá bé so với tập số liệu) cũng khiến cho phương sai thay đổi.
* Không xác định đúng dạng mô hình, thiếu biến quan trọng.
* Trong mô hình sử dụng số liệu chéo cũng khiến cho phương sai không đồng đều.

5.2.2. Hậu quả của phương sai thay đổi
* Các ước lượng OLS tuy vẫn còn tính chất tuyến tính không chệch, nhưng không còn là ước lượng hiệu quả nữa.
* Phương sai của sai số bị tính sai nên việc dùng thống kê t và thống kê F để kiểm định giả thuyết không còn đáng tin cậy nữa (thống kê t không chắc có phân phối student), các trị của t – stat và sai số chuẩn của hệ số ước lượng do phần mềm cung cấp trở nên vô dụng.
* Kết quả dự báo không hiệu quả khi dựa trên các ước lượng OLS có phương sai không nhỏ nhất.

5.2.3. Cách phát hiện phương sai nhiễu thay đổi
Việc phát hiện ra có hiện tượng này trong thực tế không đơn giản vì ta chỉ có thể dựa vào mẫu chứ không thể có toàn bộ thông tin về tổng thể. Vì thế ta không thể có một phương pháp chắc chắn để phát hiện ra phương sai thay đổi, mà chỉ có thể dựa vào một số công cụ sau đây để chẩn đoán giúp ta phát hiện ra hiện tượng này:
a/ *Bản chất của vấn đề nghiên cứu*: Bản chất của vấn đề nghiên cứu khiến ta phải nghĩ tới khả năng xảy ra hiện tượng này, chẳng hạn khi ta dùng các số liệu chéo liên quan đến các đơn vị không thuần nhất, khác nhau về quy mô.
b/ *Xem xét đồ thị của phần dư*: Đó là đồ thị của sai số của hồi quy (hay phần dư) đối với biến giải thích X nào đó hoặc đối với giá trị ước lượng $\hat{Y}$. Phương sai của phần dư được chỉ ra bằng độ rộng của biểu đồ phân rải của phần dư khi X hoặc $\hat{Y}$ tăng. Nếu độ rộng này tăng hoặc giảm thì giả thiết về phương sai không đổi có thể bị vi phạm.
- Đối với mô hình hồi quy bội, người ta thường khảo sát đồ thị phần dư $\hat{U}^2$ đối với $\hat{Y}$.
c/ Dùng các phương pháp kiểm định:
c1/ Kiểm định Park: Kiểm định Park dựa trên cơ sở giả định rằng phương sai nhiễu thay đổi dưới dạng hàm lũy thừa của biến giải thích X:

$$\sigma_i^2 = \sigma^2 . X_i^{\beta} . e^{V_i} \qquad (5.4)$$

lấy log hai vế ta nhận được:

$$ln\sigma_i^2 = ln\sigma^2 + \beta . lnX_i + V_i \qquad (5.5)$$

Vì σ_i^2 chưa biết nên Park thay σ_i^2 *bởi* $\hat{U}_i^2$(*có được từ hồi quy gốc*) trong (5.5), nhận được:

$$ln\hat{U}_i^2 = \alpha + \beta . lnX_i + V_i \ \ (\alpha = ln\sigma^2) \qquad (5.6)$$

Khi đó kiểm định Park gồm các bước sau:
B1. Thực hiện hồi quy gốc: $Y = a + b.X + U$, *nhận được các ước lượng*: $\hat{Y}_i$ *và* $\hat{U}_i$.
B2. Thực hiện hồi quy: $ln\hat{U}_i^2 = \alpha + \beta . lnX_i + V_i$.

B3. Tiến hành kiểm định giả thuyết

$$H_0: \beta = 0 (phương\ sai\ không\ đổi), H_1: \beta \neq 0 (phương\ sai\ thay\ đổi)$$

Chú ý:

* Đối với mô hình hồi quy bội, các bước tiến hành là tương tự như đối với hồi quy đơn, trong đó có thể hồi quy $ln\widehat{U}_i^2$ theo mỗi biến độc lập hoặc theo $\widehat{Y}_i$.
* Trong kiểm định Park, nhiễu V_i phải thỏa mãn các giả thiết cổ điển.

c2. *Kiểm định White:* Kiểm định White khảo sát phần dư $\widehat{U}_i^2$ theo các biến độc lập. Kiểm định này không đòi hỏi nhiễu U_i phải có phân phối chuẩn.

Giả sử ta đang xét mô hình hồi quy gốc:

$$Y = a_0 + b_1X_1 + b_2X_2 + U \qquad (5.7)$$

Kiểm định White gồm các bước sau:

B1. Hồi quy mô hình gốc (5.7), tìm được các phần dư $\widehat{U}_i$.

B2. Hồi quy mô hình phụ:

$$U_i^2 = \alpha_0 + \beta_1X_{1i} + \beta_2X_{2i} + \beta_3X_{1i}^2 + \beta_4X_{2i}^2 + \beta_5X_{1i}X_{2i} + V_i \qquad (5.8)$$

Từ đó nhận được hệ số xác định của mô hình này, ký hiệu là: R_{aut}^2

Mô hình phụ có thể có số mũ cao hơn và nhất thiết phải có hệ số chặn α_0, bất kể mô hình gốc có hay không có hệ số chặn a_0.

B3. Tiến hành kiểm định

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (phương sai không thay đổi)

Trên cơ sở H_0 đúng thì người ta chỉ ra được rằng: $n.R_{aut}^2$ có phân phối xấp xỉ $\chi^2(df)$, với bậc tự do df = số tham số của mô hình phụ (5.8), không kể hệ số chặn (trong trường hợp này $df = 5$). Vì thế:

- Nếu $n.R_{aut}^2 > \chi_\alpha^2(df)$ thì bác bỏ H_0.

c3. Kiểm định Glejser: Tương tự như kiểm định *Park,* kiểm định *Glejser* coi nhiễu có thể thay đổi theo biến độc lập X, nhưng theo một trong các dạng hàm:

$$\left|\widehat{U}_i\right| = \alpha_0 + \alpha_1X_i + V_i;\ \left|\widehat{U}_i\right| = \alpha_0 + \alpha_1\sqrt{X_i} + V_i; \qquad (5.9)$$

$$\left|\widehat{U}_i\right| = \alpha_0 + \alpha_1\frac{1}{X_i} + V_i;\ \left|\widehat{U}_i\right| = \alpha_0 + \alpha_1\frac{1}{\sqrt{X_i}} + V_i; \qquad (5.10)$$

$$\left|\widehat{U}_i\right| = \sqrt{\alpha_0 + \alpha_1X_i} + V_i;\ \left|\widehat{U}_i\right| = \sqrt{\alpha_0 + \alpha_1X_i^2} + V_i \qquad (5.11)$$

Kiểm định giả thuyết phương sai thay đổi ở đây là kiểm định giả thuyết: $H_0: \alpha_1 = 0, đối\ thuyết\ H_1: \alpha_1 \neq 0$.

Lưu ý:

* Kiểm định *Glejser* yêu cầu nhiễu V_i thỏa mãn các giả thiết cổ điển.
* Các mô hình (5.11) không phải là mô hình tuyến tính nên không dùng được phương pháp OLS.

c4. Kiểm định Goldfeld – Quandt:

Nếu ta phát hiện phương sai nhiễu tương quan thuận với một biến giải thích X nào đó dưới dạng: $\sigma_i^2 = \sigma^2.X_i^2$ (σ^2 *là hằng số*) thì sử dụng kiểm định *Goldfeld – Quandt*, theo các bước sau:

B1. Sắp xếp số liệu theo thứ tự tăng dần của X

B2. Loại bỏ c quan sát nằm ở giữa, (n – c) quan sát còn lại chia làm 2 nhóm, mỗi nhóm có (n – c)/2 quan sát.

B3. Thực hiện hồi quy OLS đối với mô hình gốc: $Y_i = a + bX_i + U$ với *(n – c)/2* quan sát đầu ta được RSS_1 (gọi là nhóm phương sai nhỏ) và với *(n – c)/2* quan sát cuối ta được RSS_2 (gọi là nhóm phương sai lớn) và chúng đều có bậc tự do là *df = (n – c – 2k)/2* (*k* là số tham số trong mô hình)

B4. Để xác minh phương sai của hai nhóm có sự khác biệt đáng kể hay không, ta tiến hành kiểm định *F* với giả thiết *H_0: phương sai không đổi* như sau: Trên cơ sở H_0 là đúng, người ta chỉ ra được đại lượng: $F = \frac{RSS_2/df}{RSS_1/df}$ có phân phối F với các bậc tự do (df, df).

Do đó nếu $F > F_\alpha(df, df)$ thì bác bỏ *H_0*, tức là chấp nhận phương sai có thay đổi.

Lưu ý:

* Mặc dù độ tin cậy của kết luận phụ thuộc vào *c*, nhưng ta lại không có quy tắc nào để xác định giá trị *c* cho tốt nhất. Theo kinh nghiệm, người ta thường chọn *c* như sau:
- Nếu *n* xấp xỉ 30 thi chọn *c = 4* hoặc *c = 8*,
- Nếu *n* xấp xỉ 60 thì chọn *c = 10* hoặc *c = 16*.

* Kiểm định *Goldfeld – Quandt* thích hợp với những mẫu nhỏ.

* Đối với mô hình hồi quy bội, ta có thể sắp xếp các quan sát theo một biến bất kỳ trong các biến giải thích của mô hình. Khi không có thông tin tiên nghiệm để biết biến giải thích nào là thích hợp, ta có thể thực hiện kiểm định Park đối với mỗi biến giải thích.

5.2.4. Biện pháp khắc phục

Do hậu quả của phương sai thay đổi, biện pháp khắc phục là hết sức cần thiết. Việc khắc phục được chia ra hai trường hợp: biết hay chưa biết σ_i^2. Trước khi đi vào các biện pháp khắc phục, ta trình bày các phương pháp bình phương bé nhất có trọng số và phương pháp bình phương bé nhất tổng quát.

1. Phương pháp bình phương bé nhất có trọng số

Xét mô hình hai biến: $Y_i = a + b.X_i + U_i$

Trước đây, để nhận được các ước lượng, phương pháp OLS nhằm cực tiểu tổng bình phương các phần dư: $\sum_{i=1}^n U_i^2 = \sum_{i=1}^n \left(Y_i - \hat{a} - \hat{b}.X_i\right)^2$ (5.12)

Bây giờ ta đặt cho mỗi phần dư U_i^2 một trọng số: $W_i = \frac{1}{\sigma_i^2}$, (trong đó $var(U_i|X_i) = var(Y_i|X_i) = \sigma_i^2$) với lý do là: khi có hiện tượng phương sai nhiễu thay đổi thì ta không thể đặt mức độ tin cậy các quan sát như nhau, quan sát nào ít sai lệch thì mức độ tin cậy sẽ cao hơn.

Để nhận được các ước lượng cho a, b, theo phương pháp bình phương bé nhất có trọng số, ta cực tiểu hóa tổng bình phương các phần dư có trọng số:

$$\sum_{i=1}^n W_i.U_i^2 = \sum_{i=1}^n W_i(Y_i - a^* - b^*.X_i)^2 \to min \qquad (5.13)$$

Vế trái (5.13) là hàm bậc 2 đối với các biến *a*, b** nên việc cực tiểu hóa hàm này cho ta các ước lượng:

$$b^* = \frac{\sum_{i=1}^n W_i.\ \sum_{i=1}^n W_i.X_i.Y_i - \sum_{i=1}^n W_i.X_i.\ \sum_{i=1}^n W_i.Y_i}{\sum_{i=1}^n W_i.\ \sum_{i=1}^n W_i.X_i^2 - \left(\sum_{i=1}^n W_i.X_i\right)^2.}; a^* = \bar{Y}^* - b^*\bar{X}^*; \qquad (5.14)$$

$(trong\ đó: \bar{X}^* = \sum_{i=1}^n W_i.X_i / \sum_{i=1}^n W_i\ ;\ \bar{Y}^* = \sum_{i=1}^n W_i.Y_i / \sum_{i=1}^n W_i\ )$

2. Phương pháp bình phương bé nhất tổng quát GLS (*Generalized Least Squares*)

Xét mô hình hai biến: $Y_i = a + b.X_i + U_i$ (5.15)

trong đó tất cả các giả thiết của mô hình hồi quy tuyến tính cổ điển đều được thỏa mãn, trừ giả thiết phương sai nhiễu không đổi bị vi phạm.

Mục đích của phương pháp GLS biến đổi từ mô hình có phương sai nhiễu thay đổi sang mô hình có phương sai nhiễu không thay đổi.

Đặt $X_{0i} = 1, \forall i$ và chia 2 vế của (5.15) cho σ_i ta nhận được mô hình:

$$Z_i = a.D_{0i} + b.D_i + V_i \qquad (5.16)$$

$$(trong\ đó: Z_i = \frac{Y_i}{\sigma_i}\ ;\ D_{0i} = \frac{X_{0i}}{\sigma_i}\ ;\ D_i = \frac{X_i}{\sigma_i};\ V_i = \frac{U_i}{\sigma_i}\ )$$

Mô hình (5.16) thỏa mãn tất cả các giả thiết của mô hình tuyến tính cổ điển, với phương sai nhiễu không đổi ($var(V_i) = var\left(\frac{U_i}{\sigma_i}\right) = 1$). Vì thế dùng phương pháp OLS cho mô hình (5.16) ta nhận được các ước lượng không chệch tuyến tính tốt nhất cho *a* và *b* là $\hat{a}^*$ *và* $\hat{b}^*$ sau đây:

$$\hat{b}^* = \frac{\sum_{i=1}^n W_i.\ \sum_{i=1}^n W_i.X_i.Y_i - \sum_{i=1}^n W_i.X_i.\ \sum_{i=1}^n W_i.Y_i}{\sum_{i=1}^n W_i.\ \sum_{i=1}^n W_i.X_i^2 - \left(\sum_{i=1}^n W_i.X_i\right)^2.};\ \hat{a}^* = \bar{Z} - \hat{b}^*.\bar{D}. \qquad (5.17)$$

$$(\text{để ý: } var\hat{b}^* = \frac{\sum_{i=1}^n W_i}{\sum_{i=1}^n W_i.\ \sum_{i=1}^n W_i.X_i^2 - \left(\sum_{i=1}^n W_i.X_i\right)^2.})$$

Phương pháp tìm các ước lượng $\hat{a}^*$ *và* $\hat{b}^*$ vừa chỉ ra gọi là *Phương pháp bình phương bé nhất tổng quát (GLS)(phương pháp OLS là trường hợp riêng của GLS)*

3. Biện pháp khắc phục: Ta chia các trường hợp để khắc phục hiện tượng này như sau:

a. Khi biết $\boldsymbol{\sigma_i^2}$: Sử dụng phương pháp GLS nói trên.

b. Khi chưa biết $\boldsymbol{\sigma_i^2}$: Ta vẫn sử dụng phương pháp GLS, nhưng đòi hỏi phải có những giả thiết nhất định sau đây về phương sai tổng thể:

Giả thiết 1: Phương sai tổng thể tỷ lệ với bình phương của biến giải thích:

$$var(U_i) = E(U_i^2) = \sigma^2.X_i^2 \qquad (5.18)$$

Khi đó từ mô hình gốc $Y_i = a + b.X_i + U_i$, ta đưa về mô hình:

$$\frac{Y_i}{X_i} = a.\frac{1}{X_i} + b + \frac{U_i}{X_i} \qquad (5.19)$$

(5.19) có phương sai nhiễu: $var\left(\frac{U_i}{X_i}\right) = \frac{varU_i}{X_i^2} = \sigma^2, \forall i$

(Lưu ý rằng phép lấy mẫu đối với X là không ngẫu nhiên mà xác định trước nên các thành phần mẫu X_i xem là các hằng số). Trong thực tế ta dùng $\boldsymbol{\hat{U}_i}$ để ước lượng cho $\boldsymbol{U_i}$, vì thế người ta thường khảo sát $\hat{U}_i^2\ theo\ X_i$. Đối hồi quy bội, có thể dùng đồ thị biểu diễn $\hat{U}_i^2$ theo từng biến giải thích, hoặc sử dụng hồi quy phụ $\hat{U}_i^2$ theo bình phương của từng biến giải thích, qua đó đánh giá được biến giải thích nào thích hợp với giả thiết 1 nhiều nhất để tiến hành biến đổi trên biến giải thích này. Tuy nhiên cần đề phòng trường hợp biến đổi mô hình gốc theo một biến nào đó dẫn đến vi phạm một giả thiết cổ điển khác.

Giả thiết 2: Phương sai tổng thể tỷ lệ với biến độc lập, tức là:

$$var(U_i) = E(U_i^2) = \sigma^2.X_i \qquad (5.20)$$

Khi đó: từ mô hình gốc $Y_i = a + b.X_i + U_i$, ta đưa về mô hình:

$$\frac{Y_i}{\sqrt{X_i}} = a.\frac{1}{\sqrt{X_i}} + b.\sqrt{X_i} + V_i \qquad (5.21)$$

trong đó $V_i = \frac{U_i}{\sqrt{X_i}}, có \quad varV_i = var\left(\frac{U_i}{\sqrt{X_i}}\right) = \sigma^2$

Mô hình (5.21) có phương sai nhiễu không thay đổi và là mô hình hồi quy tuyến tính qua gốc. Sau khi chạy hồi quy mô hình này, ta có mô hình ước lượng cho mô hình gốc bằng cách nhân 2 vế của mô hình nhận được với $\sqrt{X_i}$.

Giả thiết 3: Phương sai của nhiễu tỷ lệ với bình phương của kỳ vọng của Y, tức là: $var(U_i) = E(U_i^2) = \sigma^2.(EY_i)^2$

Khi đó: từ mô hình gốc $Y_i = a + b.X_i + U_i$, ta đưa về mô hình:

$$\frac{Y_i}{EY_i} = a.\frac{1}{EY_i} + b\frac{X_i}{EY_i} + \frac{U_i}{EY_i} \qquad (5.22)$$

(5.22) là mô hình tuyến tính cổ điển có phương sai nhiễu: $varV_i = var\left(\frac{U_i}{EY_i}\right) = \sigma^2$

Tuy nhiên trong mô hình (5.22) ta chưa biết được EY_i (do a, b chưa biết), ta sẽ thay EY_i bằng một ước lượng của nó. Ta tiến hành theo các bước sau:

B1. Chạy hồi quy mô hình gốc bằng phương pháp OLS, thu được $\hat{Y}_i$ là một ước lượng vững cho EY_i. Dùng $\hat{Y}_i$ đưa mô hình gốc về dạng:

$$\frac{Y_i}{\hat{Y}_i} = a.\frac{1}{\hat{Y}_i} + b\frac{X_i}{\hat{Y}_i} + \frac{U_i}{\hat{Y}_i} \qquad (5.22a)$$

B2. Chạy hồi quy mô hình (5.22a), từ đó nhận được mô hình hồi quy gốc.

Lưu ý: Vì trong (5.22a), ta xấp xỉ EY_i bằng ước lượng vững $\hat{Y}_i$ của nó, nên khi cỡ mẫu khá lớn thì sai số trong xấp xỉ này sẽ bé và mô hình là chấp nhận được.

Giả thiết 4: Dùng mô hình tuyến tính log thay thế:

$$lnY_i = a + b.lnX_i + V_i \qquad (5.23)$$

Ví dụ 5.2: Bảng 5.6 dưới đây cho số liệu về chi phí đầu tư Y(triệu USD) cho việc nghiên cứu và phát triển của 18 ngành công nghiệp ở Mỹ trong năm 1988, trong đó nhóm các ngành công nghiệp được đánh số thứ tự từ 1 đến 18, X_2(triệu USD) là số liệu về doanh thu, X_1(triệu USD) là lợi nhuận.Ta muốn xét tác động của doanh thu đối với đầu tư cho phát triển như thế nào qua việc ước lượng mô hình hồi quy sau:

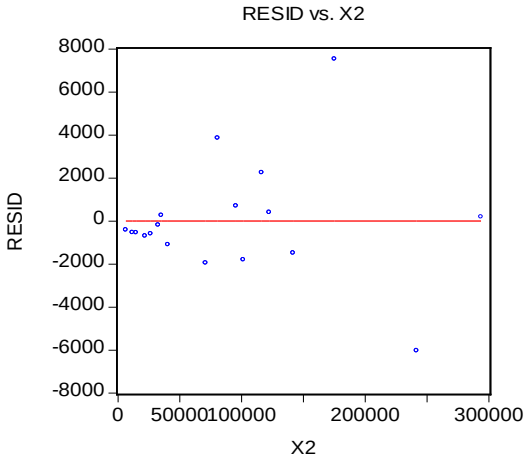
$$Y_i = \alpha + \beta.X_2 + U_i$$

với hy vọng khi doanh thu tăng thì đầu tư cho nghiên cứu và phát triển cũng sẽ tăng, mà việc nghiên cứu và phát triển có ảnh hưởng tích cực đối với các nhóm ngành nên làm tăng lợi nhuận, tức là giữa Y và X_2 có mối quan hệ đồng biến.

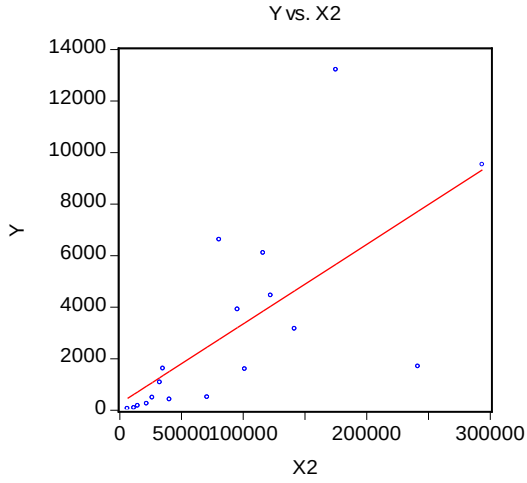
STT	Y	X_2	X_1	STT	Y	X_2	X_1
1	62.5	6375.3	185.1	10	6620.1	80552.8	13869.9
2	92.9	11626.4	1569.5	11	3918.6	95294	4487.8
3	178.3	14655.1	276.8	12	1595.3	101314.1	10278.9
4	258.4	21869.2	2828.1	13	6107.5	116141.3	8787.3
5	494.7	26408.3	2225.9	14	4454.1	122315.7	16438.8
6	1083	32405.6	3751.9	15	3163.8	141649.9	9761.4
7	1620.6	35107.7	2884.1	16	13210.7	175025.8	19774.5
8	421.7	40295.4	4645.7	17	1703.8	241434.8	23168.5
9	509.2	70761.6	5036.4	18	9528.2	293543	18415.4

Bảng 5.6

Khảo sát các biểu đồ phân tán:

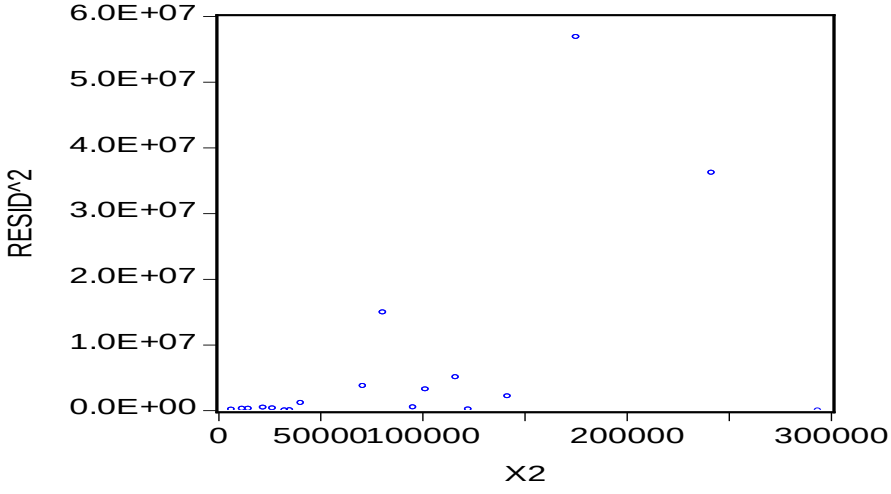


Hình 5.1a



Hình 5.1b

Hình 5.1a cho thấy các điểm phân tán có xu thế đi lên theo chiều tăng của X_2 và khi X_2 càng lớn thì các điểm phân tán càng dãn rộng ra. Điều này cho thấy khi doanh thu tăng thì bình quân đầu tư cũng tăng và phương sai nhiễu (đo mức độ phân tán) cung tăng, tức là phương sai thay đổi. Điều này có thể được lý giải bởi số liệu sử dụng là số liệu chéo, từ các ngành nghề khác nhau với quy mô và đặc điểm khác nhau.



Hình 5.1c

Để thấy rõ hơn hiện tượng này, ta khảo sát hình 5.1b biểu diễn sự biến thiên của phần dư theo doanh thu, hình 5.1c biểu diễn sự biến thiên của bình phương phần dư theo doanh thu.

Chạy hồi quy đầu tư và phát triển theo doanh thu (mô hình 2 biến), nhận được:

Dependent Variable: Y
Method: Least Squares
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	266.1917	1002.961	0.265406	0.7941
X2	0.030878	0.008346	3.699582	0.0019

R-squared	0.461042	Mean dependent var	3056.856
Adjusted R-squared	0.427357	S.D. dependent var	3705.973
S.E. of regression	2804.428	Akaike info criterion	18.82023
Sum squared resid	1.26E+08	Schwarz criterion	18.91916
Log likelihood	-167.3820	Hannan-Quinn criter.	18.83387
F-statistic	13.68690	Durbin-Watson stat	3.020747
Prob(F-statistic)	0.001944		

Bảng 5.7. Kết quả hồi quy đầu tư và phát triển theo doanh thu

Chạy hồi quy của đầu tư và phát triển Y theo X_1 và X_2 (mô hình 3 biến), ta có kết quả sau:

Dependent Variable: Y
Method: Least Squares
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.644362	1013.043	-0.002610	0.9980
X1	0.251118	0.207017	1.213031	0.2439
X2	0.010947	0.018375	0.595769	0.5602

R-squared	0.509189	Mean dependent var	3056.856
Adjusted R-squared	0.443747	S.D. dependent var	3705.973
S.E. of regression	2764.002	Akaike info criterion	18.83776
Sum squared resid	1.15E+08	Schwarz criterion	18.98615
Log likelihood	-166.5398	Hannan-Quinn criter.	18.85822
F-statistic	7.780819	Durbin-Watson stat	3.170338
Prob(F-statistic)	0.004807		

Bảng 5.8. Hồi quy Đầu tư theo doanh thu và lợi nhuận

Tiến hành kiểm định White đối với mô hình 3 biến:

Heteroskedasticity Test: White

F-statistic	*20.18959*	*Prob. F(5,12)*	*0.0000*
*Obs*R-squared*	*16.08761*	*Prob. Chi-Square(5)*	***0.0066***
Scaled explained SS	*23.57634*	*Prob. Chi-Square(5)*	*0.0003*

Bảng 5.9. Kết quả kiểm định White về phương sai thay đổi

Từ bảng này ta có: giá trị *p – value* = 0,0066 < 0,05 nên bác bỏ H_0. Vậy ta chấp nhận có hiện tượng phương sai thay đổi.

- Cách khắc phục: Lần lượt theo một trong các giả thiết:

a/ $E(U_i^2) = \sigma_i^2 = \sigma^2.X_{2i}^2$

Với giả thiết này, chạy hồi quy ước lượng cho mô hình: $\frac{Y}{X_2} = \alpha_0 + \alpha_1.\frac{1}{X_2} + \alpha_2.\frac{X_1}{X_2} + V$ ta nhận được kết quả sau:

Dependent Variable: Y/X2
Method: Least Squares
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.022272	0.015723	1.416528	0.1771
1/X2	-197.8795	142.5285	-1.388351	0.1853
X1/X2	0.141385	0.140051	1.009526	0.3287

R-squared	0.220286	Mean dependent var	0.029063
Adjusted R-squared	0.116324	S.D. dependent var	0.023120
S.E. of regression	0.021734	Akaike info criterion	-4.668853
Sum squared resid	0.007086	Schwarz criterion	-4.520458
Log likelihood	45.01968	Hannan-Quinn criter.	-4.648392
F-statistic	2.118911	Durbin-Watson stat	2.665118
Prob(F-statistic)	0.154709		

Bảng 5.10. Điều chỉnh mô hình để khắc phục

Theo đó nhận được mô hình ước lượng:

$$\widehat{\left(\frac{Y}{X_2}\right)} = 0.022272 - 197.8795.\frac{1}{X_2} + 0.141385.\frac{X_1}{X_2}$$

Đối với mô hình vừa nhận được, dùng kiểm định White có số hạng tích chéo:

Heteroskedasticity Test: White

F-statistic	1.821390	Prob. F(5,12)	0.1830
Obs*R-squared	7.766404	Prob. Chi-Square(5)	0.1696
Scaled explained SS	3.258843	Prob. Chi-Square(5)	0.6601

Bảng 5.11

Ta có: p – value = 0,1696 > 0,05 nên chấp nhận H_0. Vậy trong mô hình vừa nhận được không còn hiện tượng phương sai thay đổi.

b/ $\boldsymbol{E(U_i^2) = \sigma_i^2 = \sigma^2.X_{2i}}$

Đối với trường hợp này, chạy hồi quy ước lượng cho mô hình:

$$\frac{Y}{\sqrt{X_2}} = \frac{\alpha_0}{\sqrt{X_2}} + \alpha_1.\frac{X_1}{\sqrt{X_2}} + \alpha_2.\sqrt{X_2} + V$$

ta nhận được kết quả sau:

Dependent Variable: Y/SQR(X2)
Method: Least Squares
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
1/SQR(X2)	-243.4290	367.5355	-0.662328	0.5178
SQR(X2)	0.011638	0.017296	0.672900	0.5112
X1/SQR(X2)	0.272748	0.174788	1.560450	0.1395

R-squared	0.445020	Mean dependent var	8.850796
Adjusted R-squared	0.371023	S.D. dependent var	8.837239
S.E. of regression	7.008643	Akaike info criterion	6.883177
Sum squared resid	736.8161	Schwarz criterion	7.031572
Log likelihood	-58.94859	Hannan-Quinn criter.	6.903639
Durbin-Watson stat	3.035036		

Bảng 5.12. Điều chỉnh mô hình để khắc phục

Dùng kiểm định White có số hạng tích chéo:

Heteroskedasticity Test: White

F-statistic	5.746517	Prob. F(5,12)	0.0062
Obs*R-squared	12.69712	Prob. Chi-Square(5)	**0.0264**
Scaled explained SS	10.32900	Prob. Chi-Square(5)	0.0664

Bảng 5.13

p – value = 0.0264 < 0.05 nên bác bỏ H_0: vẫn còn hiện tượng phương sai thay đổi trong mô hình ba biến. Điều này cho thấy giả thiết $E(U_i^2) = \sigma_i^2 = \sigma^2.X_{2i}$ là không thích hợp.

c/ $\boldsymbol{var(U_i) = E(U_i^2) = \sigma^2.(EY_i)^2}$

Với giả thiết này, ước lượng cho mô hình: $\frac{Y}{\hat{Y}} = a_0.\frac{1}{\hat{Y}} + a_1.\frac{X_1}{\hat{Y}} + a_2\frac{X_2}{\hat{Y}} + V$,

ta nhận được kết quả sau:

Dependent Variable: Y/YDB
Method: Least Squares
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
1/YDB	-144.6701	92.52261	-1.563619	0.1388
X1/YDB	0.091016	0.123207	0.738726	0.4715
X2/YDB	0.024967	0.011741	2.126469	0.0505

R-squared	0.140061	Mean dependent var	0.869221
Adjusted R-squared	0.025402	S.D. dependent var	0.598553
S.E. of regression	0.590902	Akaike info criterion	1.936678
Sum squared resid	5.237474	Schwarz criterion	2.085073
Log likelihood	-14.43010	Hannan-Quinn criter.	1.957140
Durbin-Watson stat	2.563496		

Bảng 5.14

Theo đó ta có mô hình ước lượng:

$$\frac{Y}{\hat{Y}} = -144.6701.\frac{1}{\hat{Y}} + 0.091016.\frac{X_1}{\hat{Y}} + 0.024967.\frac{X_2}{\hat{Y}} + \hat{V}$$

Dùng kiểm định White đối với mô hình này, nhận được kết quả:

Heteroskedasticity Test: White

F-statistic	0.655352	Prob. F(5,12)	0.6636
Obs*R-squared	3.860875	Prob. Chi-Square(5)	**0.5696**
Scaled explained SS	1.282882	Prob. Chi-Square(5)	0.9367

Bảng 5.15

Kết quả trên cho thấy p – value = 0,5696 > 0,05, vậy ta chấp nhận giả thuyết H_0: mô hình điều chỉnh nhận được không còn hiện tượng phương sai nhiễu thay đổi.

d. Nếu dùng mô hình tuyến tính log thay thế:

$$lnY = a + blnX_1 + c.lnX_2 + V$$

ta nhận được kết quả hồi quy ước lượng:

Dependent Variable: LOG(Y)
Method: Least Squares
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6.553704	2.411367	-2.717838	0.0159
LOG(X1)	0.173952	0.352393	0.493631	0.6287
LOG(X2)	1.113761	0.441872	2.520548	0.0235

R-squared	0.793743	Mean dependent var	7.109987
Adjusted R-squared	0.766242	S.D. dependent var	1.606119
S.E. of regression	0.776535	Akaike info criterion	2.483063
Sum squared resid	9.045107	Schwarz criterion	2.631458
Log likelihood	-19.34756	Hannan-Quinn criter.	2.503524
F-statistic	28.86237	Durbin-Watson stat	2.464834
Prob(F-statistic)	0.000007		

Bảng 5.16

Có SRF ngẫu nhiên:

$$lnY = -6.553704 + 0.173952.lnX_1 + 1.113761.lnX_2 + \hat{V}$$

Dùng kiểm định White đối với mô hình này, ta có:

Heteroskedasticity Test: White

F-statistic	0.699894	Prob. F(5,12)	0.6340
Obs*R-squared	4.064039	Prob. Chi-Square(5)	**0.5402**
Scaled explained SS	2.078399	Prob. Chi-Square(5)	0.8382

Bảng 5.17

Theo đó p – value = 0,5402 > 0,5. Vậy ta chấp nhận giả thuyết H_0, tức là mô hình thay thế này không còn hiện tượng phương sai nhiễu thay đổi.

Nhận xét: Trong 4 trường hợp giả thiết trên, mô hình tuyến tính log không có hiện tượng phương sai nhiễu thay đổi và tỏ ra phù hợp hơn cả vì có hệ số xác định R^2 =0.793743 là cao nhất.

Ví dụ 5.3: Xét tập số liệu trong ví dụ 5.1, bỏ đi biến X_3, thay đổi sang mô hình tuyến tính Lin-log:

$$Y = a_0 + a_1 lnX_1 + a_2 lnX_2 + U$$

ta nhận được kết quả hồi quy:

Dependent Variable: LOG(Y)
Method: Least Squares
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.990484	0.250201	3.958756	0.0010
LOG(X1)	0.810357	0.054504	14.86794	0.0000
LOG(X2)	0.028181	0.125057	0.225343	0.8244

R-squared	0.965967	Mean dependent var	4.247404
Adjusted R-squared	0.961963	S.D. dependent var	0.255259
S.E. of regression	0.049783	Akaike info criterion	-3.024800
Sum squared resid	0.042132	Schwarz criterion	-2.875440
Log likelihood	33.24800	Hannan-Quinn criter.	-2.995643
F-statistic	241.2589	Durbin-Watson stat	1.405147
Prob(F-statistic)	0.000000		

Bảng 5.18. Điều chỉnh mô hình để khắc phục

Theo đó mô hình có hệ số xác định R^2 = 0,965967 là rất cao, hơn nữa các hệ số hồi quy không có dấu hiệu bất thường. Với kết quả nhận được, dùng kiểm định White, ta có:

Heteroskedasticity Test: White

F-statistic	1.324761	Prob. F(5,14)	0.3097
Obs*R-squared	6.423455	Prob. Chi-Square(5)	0.2672
Scaled explained SS	14.89457	Prob. Chi-Square(5)	0.0108

Bảng 5.29

Theo đó: p – value = 0,2672 > 0,05, ta chấp nhận giả thuyết H_0: mô hình tuyến tính log không có hiện tượng phương sai nhiễu thay đổi.

5.3. Tự tương quan của nhiễu

5.3.1. Khái niệm về tự tương quan

Trong mô hình h.quy tuyến tính cổ điển, gỉa thiết 2 còn yêu cầu không có tương quan giữa các phần dư, tức là:

$$cov(U_i, U_j) = 0, \forall i \neq j, hay: E(U_i . U_j) = 0, \forall i \neq j$$
$$(do\ g.thiết: EU_i = 0, \forall i)$$

Ý nghĩa thực tế của yêu cầu này là: Nhiễu của quan sát này không bị ảnh hưởng bởi nhiễu của các quan sát khác.

Tuy nhiên trong thực tế đối với một chuỗi số liệu thì yêu cầu này dễ bị vi phạm. Hiện tượng này được gọi là tự tương quan (*Autocorrelation)*của nhiễu, đó là sự tương quan giữa các thành phần của dãy quan sát theo thời gian hoặc không gian. Như vậy hiện tượng tự tương quan có nghĩa là:

$$\exists i, j, i \neq j \; sao \; cho: cov(U_i, U_j) \neq 0$$

Có thể chỉ ra các nguyên nhân sau đây:

1. Nguyên nhân khách quan:

- *Tính chất quán tính của dãy số liệu*: hầu hết số liệu chuỗi thời gian trong kinh tế đều có tính chất quán tính. Chẳng hạn số liệu theo thời gian về chỉ số giá, tỷ lệ thất nghiệp, GNP,... thường có tính chu kỳ và do đó trong hồi quy chuỗi thời gian thì các quan sát kế tiếp nhau có nhiều khả năng tương quan với nhau;
- *Sự tác động trễ (Lags) trong chuỗi thời gian*: số liệu tại thời điểm t chịu tác động bởi số liệu tại thời điểm t – 1 trước đó
- *Hiện tượng mạng nhện* (*Cobweb phenomenon*): Khi lượng cung của một số mặt hàng phản ứng lại trước sự thay đổi của giá trễ hơn một khoảng thời gian vì các quyết định cung đòi hỏi phải có thời gian để thực hiện.

2. Nguyên nhân chủ quan:

- *Việc xử lý, làm trơn số liệu*: Trước khi sử dụng, số liệu thô thường được xử lý, làm trơn (chẳng hạn dùng phương pháp trung bình di động). Sự làm trơn này có thể dẫn tới sai số hệ thống trong các nhiễu và gây ra tự tương quan giữa chúng.
- *Phép nội suy (interpolation) và ngoại suy (extrapolation)* có thể gây ra sai số có tính chất hệ thống.
- *Định dạng hồi quy chưa phù hợp, đưa không đủ biến hay bỏ sót biến quan trọng trong mô hình.*

5.3.2. Hậu quả của hiện tượng tự tương quan

1. *Các hệ số hồi quy ước lượng theo OLS không chệch nhưng không hiệu quả, tức là không còn tính chất* BLUE.
2. *Ước lượng của phương sai bị chệch nên các kiểm định t, F không còn tin cậy.*
3. *Ước lượng của hệ số R^2 tăng quá cao.*
4. *Các giá trị dự báo không còn đáng tin cậy.*

5.3.3. Cách phát hiện có tự tương quan

1. Dựa vào biểu đồ phân tán

Trong mô hình hồi quy tuyến tính cổ điển, giả thiết không có tự tương quan gắn với các nhiễu U_t không quan sát được. Ta chỉ quan sát được các phần dư $\widehat{U}_t = Y_t - \hat{Y}_t$. Mặc dù $\widehat{U}_t$ không hoàn toàn giống U_t, nhưng nó là ước lượng của U_t nên quan sát các phần dư $\widehat{U}_t$ có thể gợi ý cho ta những nhận xét về U_t. Vì thế để có thông tin về tự tương quan của nhiễu U, ta có thể khảo sát một trong các biểu đồ phân tán sau:

a/ Biểu đồ phân tán $(\widehat{U}_t, t)$ của $\widehat{U}_t$ (hoặc của $\widehat{U}_t^2$) theo thời gian.

b/ Biểu đồ phân tán $(\frac{\widehat{U}_t}{\widehat{\sigma}}, t)$ của phần dư chuẩn hóa $\frac{\widehat{U}_t}{\widehat{\sigma}}$ theo thời gian t

Lưu ý rằng $U_t \sim N(0,\sigma^2)$ *nên* $\frac{U_t}{\sigma} \sim N(0,1)$. *Vì thế khi cỡ mẫu n khá lớn thì* $\frac{\hat{U}_t}{\hat{\sigma}}$ có phân phối xấp xỉ phân phối $N(0,1)$.

c/ Biểu đồ phân tán $(\hat{U}_t, \hat{U}_{t-1})$ *của* $\hat{U}_t$ *theo* $\hat{U}_{t-1}$ *(gọi là lược đồ AR(1))*

- Nếu biểu đồ phân tán có dạng ngẫu nhiên thì không có tự tương quan, nếu biểu đồ phân tán có dạng không ngẫu nhiên, biểu thị xu hướng biến thiên có tính chất hệ thống thì nhận định có tự tương quan.

2/ Kiểm định Durbin - Watson

a. Xét thống kê: $$\boldsymbol{d} = \frac{\sum_{t=2}^{n}(\hat{U}_t - \hat{U}_{t-1})^2}{\sum_{t=1}^{n}\hat{U}_t^2} \quad (5.24)$$

Người ta chỉ ra được rằng khi n đủ lớn thì: $d \approx 2(1-\hat{\rho})$

trong đó: $\hat{\rho} = \frac{\sum_{t=2}^{n}\hat{U}_t.\hat{U}_{t-1}}{\sum_{t=1}^{n}\hat{U}_t^2}$ $(hệ\ số\ tự\ tương\ quan\ mẫu\ bậc\ nhất)$ (5.25)

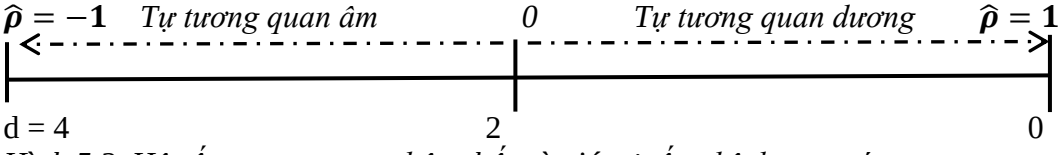
$\hat{\rho}$ là ước lượng của hệ số tự tương quan bậc nhất ρ trong mô hình tự hồi quy bậc nhất (hay tự tương quan bậc nhất):

$$U_t = \rho.U_{t-1} + \varepsilon_t \quad (-1 \le \rho \le 1) \quad (AR(1)) \quad (5.26)$$

với ε_t *là nhiễu ngẫu nhiên thỏa*:

$$E\varepsilon_t = 0, cov(\varepsilon_t, \varepsilon_s) = 0, t \ne s, var\varepsilon_t = \sigma^2, \forall t \quad (5.27)$$

Nhận xét: từ $-1 \le \rho, \hat{\rho} \le 1$ suy ra: $0 \le d \le 4$



Hình 5.2. Hệ số tự tương quan bậc nhất và giá trị tống kê d tương ứng

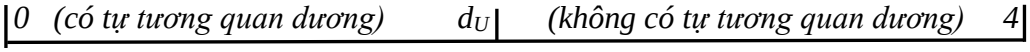
- Khi d = 4 hoặc = 0 , ta có tự tương quan hoàn hảo. Khi d = 2 thì không có tự tương quan.

Bảng thống kê Durbin – Watson chỉ ra các giá trị tới hạn d_U, d_L dựa vào ba tham số: *mức ý nghĩa* α, *số quan sát* n, *số biến độc lập* k'.

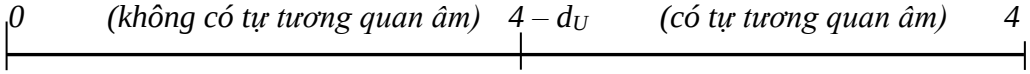
b. Quy tắc kiểm định Durbin – Watson: Kiểm định giả thuyết mô hình có tự tương quan chính là $H_0: \rho = 0$, đối thuyết $H_1: \rho > 0 / \rho < 0 / \rho \ne 0$

Tính giá trị của thống kê d qua số liệu, so sánh d với các giá trị tới hạn để bác bỏ hay chấp nhận H_0 theo quy tắc sau:

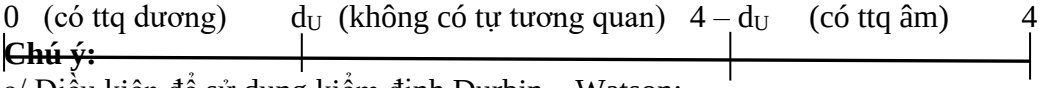
$*$ $H_0: \rho = 0$, $H_1: \rho > 0$, mức ý nghĩa α



$*$ $H_0: \rho = 0$, $H_1: \rho < 0$, mức ý nghĩa α



$*$ $H_0: \rho = 0$, $H_1: \rho \ne 0$, mức ý nghĩa 2α



Chú ý:

a/ Điều kiện để sử dụng kiểm định Durbin – Watson:

* Mô hình hồi quy phải có hệ số bị chặn. Nếu mô hình không có hệ số bị chặn thì phải ước lượng mô hình có hệ số bị chặn để tính $RSS = \sum \widehat{U}_i^2$, sau đó tiến hành kiểm định.
* Việc lấy mẫu các biến độc lập là lấy mẫu xác định (không phải mẫu ngẫu nhiên).
* Các nhiễu có tương quan bậc nhất: $U_t = \rho . U_{t-1} + \varepsilon_t \ \ (AR(1))$
* Mô hình không có dạng tự hồi quy, tức là không xét mô hình dạng:

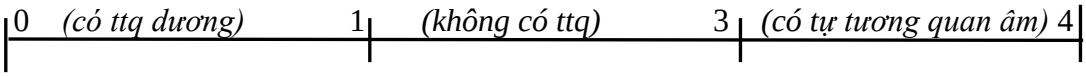
$$Y_t = a + b_1 X_t + b_2 Y_{t-i} + U_t$$

* Không có quan sát bị mất trong dữ liệu.

b/ Nhược điểm của kiểm định Durbin- Watson:

* Khi cỡ mẫu n lớn thì không có trong bảng tra,
* Có một số mâu thuẫn khi tra bảng tìm d_U, d_L (chẳng hạn khi n = 9, k' = 3, $\alpha = 5\%$ *thì* $4 - d_U < d_U$)

c/ Đôi khi người ta sử dụng tiêu chuẩn kiểm định Durbin-Watson theo kinh nghiệm như sau:



3. Kiểm định Breusch – Godfrey (BG)

Xét mô hình hồi quy: $Y_t = a + bX_t + U_t$ (5.28)

với thành phần nhiễu có tự tương quan bậc p (AR(p)):

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + \cdots + \rho_p U_{t-p} + \varepsilon_t \quad (5.29)$$

trong đó ε_t là nhiễu ngẫu nhiên thỏa các giả thiết OLS.

Khi đó giả thuyết không có sự tương quan bậc p là: $H_0: \rho_1 = \rho_2 = \cdots = \rho_p = 0$

Thủ tục kiểm định BG như sau:

B1: Chạy hồi quy mô hình (5.28) theo OLS và tìm được phần dư $\widehat{U}_t$.

B2: Chạy hồi quy mô hình:

$$\widehat{U}_t = \alpha + \beta . X_t + \gamma_1 \widehat{U}_{t-1} + \gamma_2 \widehat{U}_{t-2} + \cdots + \gamma_p \widehat{U}_{t-p} + V_t \quad (5.30)$$

từ đó tính được hệ số xác định của mô hình (5.30), ký hiệu là $R^2_{(5.30)}$.

B3: Nếu $(n-p).R^2_{(5.30)} > \chi^2_\alpha(p)$ *thì bác bỏ* H_0, tức là thừa nhận có tự tương quan bậc p.

Chú ý:

* Kiểm định BG áp dụng cho cỡ mẫu lớn và mở rộng cho mô hình nhiều biến.
* Kiểm định BG có thể áp dụng cho mô hình tự hồi quy (mô hình có biến giải thích Y_{t-1}, Y_{t-2},..., tức là có biến trễ).
* Kiểm định BG áp dụng cho tự tương quan với bậc bất kỳ.
* Kiểm định BG đòi hỏi phải xác định trước bậc của tự tương quan p. Trong thực tế người ta phải kiểm định với nhiều giá trị p khác nhau.
* Kiểm định BG có thể được áp dụng cho mô hình có nhiễu U được tạo ra theo tiến trình trung bình động bậc q (MA(q): $q^{th} - order\ Moving\ Average$),

tức là: $U_t = \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + \cdots + \lambda_q \varepsilon_{t-q}$

trong đó ε là nhiễu ngẫu nhiên với kỳ vọng bằng 0 và phương sai không đổi.

4. Kiểm định các đoạn mạch (hay kiểm định chuỗi dấu(*Runs test*)) **(Tham khảo)**

Số hạng nhiễu có giá trị khi âm, khi dương, do đó nếu sự thay đổi về dấu của số hạng nhiễu diễn ra mang tính hệ thống, theo một xu thế nào đó thì biểu hiện có sự tự tương

quan giữa các nhiễu. Nếu dấu của nhiễu thay đổi một cách ngẫu nhiên thì có thể xem là biểu hiện không có hiện tượng tự tương quan. Kiểm định các đoạn mạch hay kiểm định chuỗi dấu dựa vào sự thay đổi dấu của các phần dư ước lượng từ mô hình hồi quy và được thực hiện theo các bước sau:

Xét mô hình hồi quy gốc: $Y_t = a + bX_t + U_t$

B1: Chạy hồi quy mô hình gốc, có được các phần dư ước lượng: $\widehat{U}_t = Y_t - Y_{db}$

B2: Phần dư $\widehat{U}_t > 0$ được thay bởi dấu +, phần dư $\widehat{U}_t < 0$ được thay bởi dấu – . Như thế ta có được một dãy các dấu + , – đan xen nhau:

+ + + + + – – – + + – – – – + + + + + – –– – – + + + – – – + + + + +

Ta gọi một dãy liên tục các dấu + (hoặc –) mà kề hai đầu mút hoặc không có dấu nào, hoặc có dấu khác, là một đoạn mạch hay một chuỗi dấu.

B3: Xác định các số: n_1 là tổng số dấu + trong dãy; n_2 là tổng số dấu – trong dãy; $n = n_1 + n_2$; N là số đoạn mạch hay số chuỗi dấu.

B4: Thiết lập giả thuyết H_0: không có tự tương quan của các thành phần nhiễu. Nếu H_0 đúng, với giả định: $n_1 \geq 10$; $n_2 \geq 10$ thì đại lượng N là đại lượng ngẫu nhiên có phân phối tiệm cận chuẩn, trong đó kỳ vọng và phương sai của N được tính như sau:

$E(N) = \frac{2n_1n_2}{n_1+n_2} + 1$; $Var(N) = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$; se(N) = $\sqrt{Var(N)}$

và N rơi vào khoảng tin cậy sau với độ tin cậy $\gamma = 1 - \alpha$:

$$(\boldsymbol{E(N) - u_{\frac{\alpha}{2}}.se(N);\ E(N) + u_{\frac{\alpha}{2}}.se(N)})$$

* Nếu $\boldsymbol{N \in (E(N) - u_{\frac{\alpha}{2}}.se(N);\ E(N) + u_{\frac{\alpha}{2}}.se(N))}$ thì chấp nhận H_0

* Nếu $\boldsymbol{N \notin (E(N) - u_{\frac{\alpha}{2}}.se(N);\ E(N) + u_{\frac{\alpha}{2}}.se(N))}$ thì bác bỏ H_0.

Lưu ý: Kiểm định các đoạn mạch là kiểm định phi tham số, tuy nhiên nó yêu cầu $n_1 \geq 10$, $n_2 \geq 10$. Khi yêu cầu này không thỏa, người ta có bảng chuyên dụng cho các giá trị tới hạn đối với số đoạn mạch mà ta kỳ vọng trong một dãy ngẫu nhiên n quan sát.

5.3.4. Cách khắc phục

Khi có sự tương quan chuỗi, các ước lượng OLS trở nên không hiệu quả. Để khắc phục hiện tượng này, ta phân biệt hai trường hợp: Biết cấu trúc tự tương quan và chưa biết cấu trúc tự tương quan.

1. Trường hợp biết cấu trúc tự tương quan

Do nhiễu U_t không quan sát được nên cấu trúc tương quan chuỗi thường là vấn đề suy đoán hoặc

do những đòi hỏi cấp thiết của thực tiễn. Trong thực hành, người ta thường giả sử nhiễu có mô

hình tự hồi quy bậc 1, nghĩa là: $U_t = \rho.U_{t-1} + \varepsilon_t$ (5.31)

với hệ số tự hồi quy ρ đã biết và $-1 \leq \rho \leq 1$, còn ε_t là nhiễu thỏa các giả thiết cổ điển.

Xét mô hình hồi quy gốc: $Y_t = a + b.X_t + U_t$ (5.32)

tại thời điểm t – 1, (5.32) cho ta: $Y_{t-1} = a + b.X_{t-1} + U_{t-1}$ (5.33)

Từ (5.33) suy ra: $\rho.Y_{t-1} = \rho a + b.\rho X_{t-1} + \rho.U_{t-1}$ (5.34)

Trừ vế theo vế hai hệ thức (5.32), (5.34), nhận được:

$$Y_t - \rho.Y_{t-1} = a.(1-\rho) + b.(X_t - \rho X_{t-1}) + (U_t - \rho.U_{t-1})$$

hay: $$Y_t^* = a^* + b^* X_t^* + \varepsilon_t \quad (5.35)$$

$$(trong\ đó: Y_t^* = Y_t - \rho . Y_{t-1}; X_t^* = X_t - \rho X_{t-1})$$

(5.35) là mô hình hồi quy tuyến tính cổ điển (ε_t thỏa mãn các giả thiết cổ điển) nên các ước lượng OLS của mô hình này có tính chất BLUE.

Chú ý:

* (5.35) là phương trình sai phân tổng quát, do việc ghép đuổi hai số liệu liên tiếp thành một nên mô hình bị bớt đi một số liệu (quan sát thứ nhất) so với mô hình gốc. Trong thực nghiệm, theo biến đổi Prais-Winsten, quan sát thứ nhất của (5.35) được tạo như sau:

$$Y_1^* = Y_1 . \sqrt{1-\rho^2}, X_1^* = X_1 . \sqrt{1-\rho^2} \quad (5.36)$$

* Khi $\rho = 1$ thì (5.35) trở thành phương trình sai phân cấp 1:

$$Y_t - Y_{t-1} = b.(X_t - X_{t-1}) + (U_t - U_{t-1}) \quad (5.37a)$$

* Khi $\rho = -1$ thì (5.35) trở thành phương trình hồi quy trung bình trượt:

$$\frac{Y_t + Y_{t-1}}{2} = a + b_1 . \frac{X_t + X_{t-1}}{2} + \frac{U_t + U_{t-1}}{2} \quad (5.37b)$$

2. Trường hợp chưa biết cấu trúc của tự tương quan

Trong thực tế ta chưa biết cấu trúc của tự tương quan do ít khi biết được giá trị ρ. Vậy phải tìm cách ước lượng ρ.

a. Ước lượng bằng thống kê d.

Trong kiểm định Durbin- Watson, khi n đủ lớn ta có: $d \approx 2(1-\hat{\rho})$, vì thế ta nhận được:

$$\hat{\rho} \approx 1 - \frac{d}{2}. \quad (5.38)$$

Khi n nhỏ, Theil và Nagar dùng ước lượng: $\hat{\rho} = \frac{n^2\left(1-\frac{d}{2}\right)+k^2}{n^2-k^2}$ (5.39)

trong đó d là thống kê Durbin-Watson, k là số các hệ số của mô hình (bao gồm cả tung độ gốc).

Khi đã có được $\hat{\rho}$, ta chạy hồi quy ước lượng cho mô hình (5.35) theo phương pháp OLS. Chú ý rằng các ước lượng thu được từ mô hình này cũng chỉ tiệm cận với tính chất BLUE khi n khá lớn, vì trong mô hình ta đã thay ρ bởi ước lượng $\hat{\rho}$ của nó. Vì thế khi cỡ mẫu n bé, ta cần thận trọng khi giải thích các kết quả ước lượng.

b. Ước lượng ρ bởi thủ tục lặp Cochrance – Orcutt (CORC)

Phương pháp này sử dụng các phần dư đã được ước lượng để thu được thông tin về ρ.

Xét mô hình hồi quy gốc: $$Y_t = a + b.X_t + U_t \quad (5.40)$$

với U_t thỏa mãn lược đồ AR(1): $$U_t = \rho . U_{t-1} + \varepsilon_t \quad (5.41)$$

Ước lượng cho ρ được thực hiện theo các bước sau:

B1: Ước lượng mô hình (5.40) bằng phương pháp OLS, thu được phần dư $\widehat{U}_t$.

B2: Sử dụng các phần dư $\widehat{U}_t$ làm số liệu để ước lượng hồi quy cho (5.41), từ đó nhận được $\hat{\rho}$.

B3: Thay ρ bởi $\hat{\rho}$ để ước lượng phương trình sai phân tổng quát:

$$Y_t - \hat{\rho} . Y_{t-1} = a.(1-\hat{\rho}) + b.(X_t - \hat{\rho} X_{t-1}) + (U_t - \hat{\rho} . U_{t-1})$$

hay ước lượng hồi quy: $$Y_t^* = a^* + b^* X_t^* + \varepsilon_t \quad (5.42)$$

$$(trong\ đó: Y_t^* = Y_t - \hat{\rho} . Y_{t-1}; X_t^* = X_t - \hat{\rho} X_{t-1})$$

B4: Để cải thiện chất lượng của ước lượng $\hat{\rho}$ nhận được từ B2, ta thay giá trị $\hat{a}^*, \hat{b}^*$ là các ước lượng của a^*, b^* tìm được trong B3 vào hồi quy gốc (5.40) và nhận được các phần dư mới:

$$\widehat{U}_t^* = Y_t - (\hat{a}^* + \hat{b}^*.X_t) \quad (5.43)$$

Sử dụng phần dư mới $\widehat{U}_t^*$ làm số liệu để ước lượng cho hồi quy:

$$U_t^* = \rho.U_{t-1}^* + V_t \quad (5.44)$$

từ đó nhận được $\hat{\rho}$ là ước lượng vòng 2 cho ρ. Các vòng lặp này được tiếp tục cho đến khi hai ước lượng kế tiếp nhau của ρ sai khác nhau rất bé (chẳng hạn sai khác dưới 0,05 hoặc 0,005) (Thực tế cho thấy dùng tới 3 – 4 bước lặp là đủ).

c. Ước lượng ρ bởi phương pháp Durbin-Watson 2 bước.

Phương trình sai phân tổng quát được viết lại dưới dạng:

$$Y_t = a.(1-\rho) + b.X_t - b.\rho X_{t-1} + \rho.Y_{t-1} + \varepsilon_t \quad (5.45)$$

B1: Chạy hồi quy mô hình (5.45) theo OLS, nhận được ước lượng $\hat{\rho}$ của ρ.

B2: Chạy hồi quy mô hình:

$$Y_t - \hat{\rho}.Y_{t-1} = a.(1-\hat{\rho}) + b.(X_t - \hat{\rho}X_{t-1}) + (U_t - \hat{\rho}.U_{t-1})$$

từ đó nhận được các ước lượng $\hat{a}^* \; cho \; a^* = a.(1-\hat{\rho}), \hat{b}^* \; cho \; b^* = b$, do đó có thể ước lượng $a \; bởi \; \hat{a} = \frac{\hat{a}^*}{1-\hat{\rho}}$.

Ví dụ 5.4: Tỷ lệ Y(%) về lực lượng lao động dân thường tham gia ở Mỹ, tỷ lệ X_1(%) về dân thường thất nghiệp, số tiền trung bình X_2(USD) kiếm được thực tế theo giờ, theo số liệu thu được từ 1980 – 2002 có kết quả sau:

Năm	Y	X_1	X_2	Năm	Y	X_1	X_2	Năm	Y	X_1	X_2
1980	63.8	7.1	7.78	1988	65.9	5.5	7.69	1996	66.8	5.4	7.43
1981	63.9	7.6	7.69	1989	66.5	5.3	7.64	1997	67.1	4.9	7.55
1982	64.0	9.7	7.68	1990	66.5	5.6	7.52	1998	67.1	4.5	7.75
1983	64.0	9.6	7.79	1991	66.2	6.8	7.45	1999	67.1	4.2	7.86
1984	64.4	7.5	7.80	1992	66.4	7.5	7.41	2000	67.2	4.0	7.89
1985	64.8	7.2	7.77	1993	66.3	6.9	7.39	2001	66.9	4.8	7.99
1986	65.3	7.0	7.81	1994	66.6	6.1	7.40	2002	66.6	5.8	8.14
1987	65.6	6.2	7.73	1995	66.6	5.6	7.40				

Bảng 5.30

Chạy hồi quy của Y theo X_1 và X_2 ta có bảng kết quả:

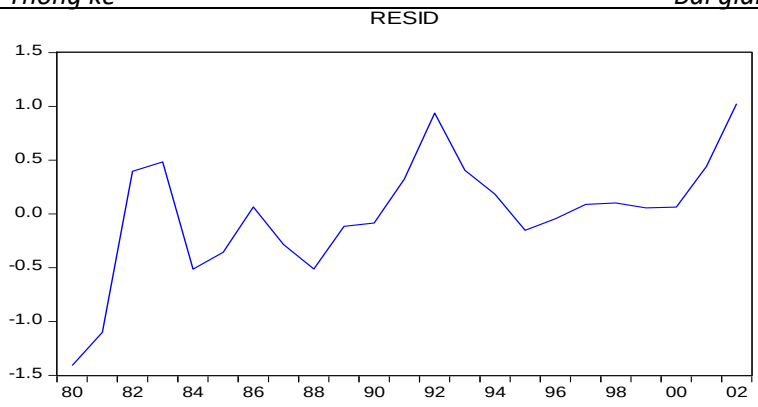
Dependent Variable: Y
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	80.95122	4.770337	16.96971	0.0000
X1	-0.671631	0.082705	-8.120845	0.0000
X2	-1.410432	0.610348	-2.310867	0.0316
R-squared	0.772914	Mean dependent var		65.89565

Bảng 5.31

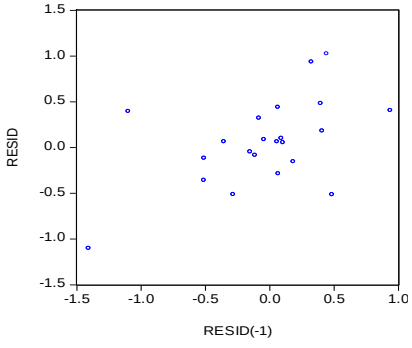
***1/ Cách phát hiện*:**

* Xét đồ thị phần dư $\widehat{U}_t$ theo thời gian

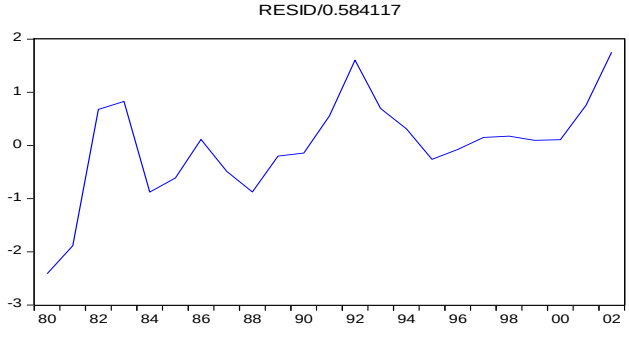


Hình 5.2

Vẽ biểu đồ $\hat{U}_t$ theo $\hat{U}_{t-1}$ (hay lược đồ AR(1)) và đồ thị chuẩn hóa của $\hat{U}_t/\hat{\sigma}$ theo thời gian



Hình 5.3



Hình 5.4

Các đồ thị và biểu đồ về resid đều có xu hướng tăng nên ta nhận định có tự tương quan. Ta có thể xác minh điều này qua các kiểm định.

* *Kiểm định Durbin-Watson:*

Từ bảng kết quả hồi quy ta có giá trị thống kê d = 0.787065 < 1 nên ta kết luận có tự tương quan dương bậc 1.

* Kiểm định BG: Với tự tương quan bậc 1

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	4.318934	Prob. F(1,19)	0.0515
Obs*R-squared	4.259864	Prob. Chi-Square(1)	**0.0390**

Bảng 5.32

có p – value = 0.0390 < 0.05 nên kết luận có tự tương quan bậc 1.

2/ Các biện pháp khắc phục:

a. Dùng thống kê d: có $\hat{\rho} = 1 - \frac{d}{2} = 1 - 0.787065/2 = 0.6064675$

Hồi quy sai phân cấp 1 tổng quát:

Dependent Variable: Y-0.6064675*Y(-1)
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	28.14487	1.738930	16.18517	0.0000
X1-0.6064675*X1(-1)	-0.349864	0.072465	-4.828032	0.0001
X2-0.6064675*X2(-1)	-0.412303	0.564268	-0.730685	0.4739
R-squared	0.551313	Mean dependent var		26.04675

Bảng 5.33. Điều chỉnh mô hình để khắc phục

* Dùng kiểm định BG ta có:

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	1.257971	Prob. F(1,18)	0.2768
Obs*R-squared	1.437087	Prob. Chi-Square(1)	**0.2306**

Bảng 5.34

Từ đó nhận được: p – value = 0.2306 > 0.05 nên ta chấp nhận giả thuyết H_0, tức là không còn tự tương quan. Vậy mô hình SRF sau khi khắc phục là mô hình hồi quy sai phân cấp 1 tổng quát:

$$\mathbf{Y_t - 0.6064675*Y_{t-1} = 28.1448656078 - 0.349864468794*(X_{1t}-0.6064675*X_{1t-1}) - 0.412302674777*(X_{2t}-0.6064675*X_{2,t-1}) + \widehat{U}_t}$$

b. Dùng Durbin-Watson 2 bước:

* Chạy hồi quy ước lượng cho mô hình (5.45):

$Y_t = a.(1-\rho) + b_1.X_{1t} - b_2.\rho X_{1t-1} + c_1.X_{2t} - c_2.\rho X_{2t-1} + \rho.Y_{t-1} + \varepsilon_t$

có kết quả:

Dependent Variable: Y
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	23.66925	8.181627	2.892976	0.0106
X1	-0.205990	0.052832	-3.898985	0.0013
X1(-1)	0.042676	0.063112	0.676191	0.5086
X2	-0.163399	0.528324	-0.309278	0.7611
X2(-1)	-0.352127	0.680114	-0.517747	0.6117
Y(-1)	0.718008	0.093132	7.709597	0.0000
R-squared	0.979506	Mean dependent var		65.99091

Bảng 5.35

Ta nhận được: $\hat{\rho}$ =0.718008 (hệ số của Y(-1)).

* Hồi quy sai phân cấp 1 tổng quát :

$$Y_t - \hat{\rho}Y_{t-1} = a(1-\hat{\rho}) + b_1(X_{1t} - \hat{\rho}X_{1t-1}) + b_2(X_{2t} - \hat{\rho}X_{2t-1}) + (U_t - \hat{\rho}U_{t-1})$$

ta nhận được kết quả:

Dependent Variable: Y-0.718008*Y(-1)
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	19.79754	1.149986	17.21545	0.0000
X1-0.718008*X1(-1)	-0.258695	0.063126	-4.098081	0.0006
X2-0.718008*X2(-1)	-0.299577	0.520838	-0.575183	0.5719
R-squared	0.469613	Mean dependent var		18.70029

Bảng 5.36

* Dùng kiểm định BG:

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.392812	Prob. F(1,18)	0.5387
Obs*R-squared	0.469850	Prob. Chi-Square(1)	**0.4931**

Bảng 5.37

Từ bảng kết quả kiểm định nhận được: p – value = 0.4931 > 0.05 nên ta chấp nhận không còn tự tương quan trong mô hình ước lượng:

Y-0.718008.Y(-1) = 19.79754 - 0.258695.(X1-0.718008.X1(-1)) – 0.299577.(X2-0.718008.X2(-1)) + $\hat{U}$.

c. Dùng CORC 2 bước: chạy hồi quy ước lượng cho mô hình: $U_t = \rho . U_{t-1} + \varepsilon_t$

(Với lưu ý là U_t = Y – Ydb, U_{t-1} = Y(−1) − Ydb(−1))

Dependent Variable: Y-YDB
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Y(-1)-YDB(-1)	**0.453785**	0.173479	2.615783	0.0161
R-squared	0.231404	Mean dependent var		0.064067

Bảng 5.38

Từ đó nhận được: $\hat{\rho} = 0.453785$ (hệ số hồi quy của {Y(-1)-Ydb(-1)})

Hồi quy sai phân cấp 1 tổng quát:

Dependent Variable: Y-0.453785*Y(-1)
Method: Least Squares
Sample (adjusted): 1981 2002
Included observations: 22 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	40.43469	2.469916	16.37088	0.0000
X1-0.453785*X1(-1)	-0.467019	0.077385	-6.034979	0.0000
X2-0.453785*X2(-1)	-0.654515	0.577216	-1.133918	0.2709
R-squared	0.658179	Mean dependent var		36.10298

Bảng 5.39

Nhận được mô hình ước lượng:

Y-0.453785*Y(-1) = 40.4346880342 - 0.467019486845*(X1-0.453785*X1(-1)) - 0.654515242431*(X2-0.453785*X2(-1)) + $\hat{U}$

* Dùng kiểm định BG:

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	1.452444	Prob. F(1,18)	0.2437
Obs*R-squared	1.642661	Prob. Chi-Square(1)	0.2000

Bảng 5.40

Ta nhận được: p – value = 0.2000 > 0.05 nên ta thừa nhận không còn tự tương quan trong mô hình vừa nhận được.

d. Dùng biến trễ $\boldsymbol{Y_{t-1}}$: chạy hồi quy ước lượng cho mô hình:

$$Y_t = a_0 + a_1X_{1t} + a_2X_{2t} + a_3Y_{t-1}$$

Dependent Variable: Y
Method: Least Squares
Sample (adjusted): 1981 2002
Included observations: 22 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24.46852	4.989686	4.903820	0.0001
X1	-0.183419	0.047358	-3.873028	0.0011
X2	-0.486649	0.196506	-2.476506	0.0234
Y(-1)	0.704545	0.061236	11.50535	0.0000

R-squared	0.977967	Mean dependent var	65.99091
Adjusted R-squared	0.974295	S.D. dependent var	1.101042
S.E. of regression	0.176528	Akaike info criterion	-0.467707
Sum squared resid	0.560919	Schwarz criterion	-0.269336
Log likelihood	9.144780	Hannan-Quinn criter.	-0.420977
F-statistic	266.3192	Durbin-Watson stat	2.258316
Prob(F-statistic)	0.000000		

Bảng 5.41

Nhận được mô hình:

Y = 24.4685200341- 0.183418712076*X1-0.486649020366*X2 + 0.704545160497*Y(-1) + $\widehat{U}$

Bằng kiểm định BG:

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	1.474721	Prob. F(1,17)	0.2412
Obs*R-squared	1.756122	Prob. Chi-Square(1)	0.1851

Bảng 5.42

Kết quả trên cho thấy p – value = 0.1851 > 0.05 nên ta có thể cho rằng không còn tự tương quan trong mô hình vừa nhận được.

Bài tập.

1. Ta có bảng số liệu sau đây về chi tiêu Y và thu nhập X hàng tháng của 20 hộ gia đình ở một vùng nông thôn:

Y	X	Y	X	Y	X	Y	X	Y	X
19.9	2.3	40.7	42.3	10.3	10.3	33.5	38.0	29.3	30.1
31.2	32.3	6.1	6.2	38.8	40.2	13.1	14.1	25.0	28.3
31.8	33.6	38.6	44.7	8.0	8.1	14.8	16.4	17.9	18.2
12.1	12.1	25.5	26.1	33.1	34.5	21.6	24.1	19.8	20.1

a. Tính các đặc trưng mẫu cho các biến và tìm ma trận tương quan mẫu của véc tơ (X, Y).
b. Hồi quy ước lượng cho các mô hình:
b1. Y = a + b.X + U,
b2. LnY = a' + b'.lnX + V
c. Vẽ Line Graph giữa các giá trị dự báo điểm và giá trị quan sát của Y từ việc ước lượng cho mô hình b1/
d. Hãy xem xét vấn đề phương sai thay đổi trong các mô hình trên, và khắc phục, nếu có.
2. Tiến hành khảo sát giá bán X_1(ngàn đồng/kg), chi phí quảng cáo X_2 (triệu đồng/tháng) và lượng hàng bán được Y(tấn/tháng), ở 20 khu vực có số liệu sau đây, trong đó Z = 0 nếu khu vực khảo sát ở nông thôn, Z = 1 nếu khu vực khảo sát ở thành thị.

Y	X_1	Z	Z	Y	X_1	X_2	Z	Y	X_1	X_2	Z
20	2.5	10	1	16	4.7	7.1	1	12	7.7	7.5	0
19	3.1	9.2	0	15	5.3	6.9	1	15	5.9	6.9	1
18	3.5	8.8	1	15	5.8	6.5	1	16	4.8	6.7	0
18	4.2	8.4	0	14	5.9	6.8	0	12	7.2	6.5	1
17	4.6	8	1	14	6.4	6.6	1	10	8.3	7.2	0
17	3.8	7.6	1	13	6.8	7.0	0	11	8.5	8.3	1
16	4.2	7.2	0	12	7.2	7.8	1				

a. Hãy ước lượng mô hình: $Y_i = a + b_1X_{1i} + b_2X_{2i} + b_3Z_i + U_i$
b. Tính giá trị các dự báo điểm đối với mô hình trên. Vẽ đồ thị Line Graph giữa các giá trị dự báo điểm của Y với giá trị thực tế của Y.
c. Kiểm định xem mô hình trên có hiện tượng đa cộng tuyến, phương sai thay đổi và tự tương quan hay không. Nếu có hãy tìm cách khắc phục.
3. Từ số liệu về chi tiêu Y ($) cho tiêu dùng, thu nhập X ($) và sự giàu có Z ($) qua khảo sát 12 hộ gia đình, chạy hồi quy ước lượng cho mô hình: $Y = a_0 + a_1X + a_2Z + U$, nhận được:

Dependent Variable: Y
Method: Least Squares
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	29.30834	5.286306	5.544200	0.0004
X	0.265502	0.382741	0.693686	0.5054
Z	0.020920	0.037858	0.552581	0.5940
R-squared	0.968220	Mean dependent var		110.4167

Ma trận tương quan mẫu của véc tơ (Y, X, Z) là:

	Y	X	Z
Y	1.000000	0.983434	0.983118
X	0.983434	1.000000	0.997168
Z	0.983118	0.997168	1.000000

Từ đó hãy nhận định xem mô SRF thu được có hiện tượng đa cộng tuyến hay không.

4. Cũng với SRF nhận được từ bài tập 3, nhận định xem mô hình này có vấn đề về phương sai nhiễu thay đổi hay không qua kết quả kiểm định dưới đây:

Heteroskedasticity Test: White

F-statistic	0.560802	Prob. F(5,6)	0.7289
Obs*R-squared	3.821907	Prob. Chi-Square(5)	0.5753
Scaled explained SS	1.719122	Prob. Chi-Square(5)	0.8865

5. Với mô hình SRF nhận được từ bài tập 3, có nhận xét gì về mô hình này từ kết quả kiểm định BG sau:

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	1.614981	Prob. F(2,7)	0.2650
Obs*R-squared	3.788826	Prob. Chi-Square(2)	0.1504

6. Có số liệu về chi tiêu Y ($) cho tiêu dùng, thu nhập X ($) và sự giàu có Z ($) qua khảo sát 10 hộ gia đình như sau:

Y	X	Z	Y	X	Z	Y	X	Z
70	80	810	95	140	1425	140	220	2201
65	100	1009	110	160	1633	155	240	2435
90	120	1273	115	180	1876	150	260	2686
			120	200	2052			

a/ Chạy hồi quy ước lượng cho mô hình: $Y = a_0 + a_1X + a_2Z + U$
b/ Mô hình SRF nhận được từ a/ có hiện tượng đa cộng tuyến hay không? Nếu có hãy tìm cách khắc phục.

7. Với 10 doanh nghiệp ở Tp.HCM được chọn ngẫu nhiên để điều tra về doanh thu Y (tỷ đồng) và chi phí sản xuất X (tỷ đồng) có bảng số liệu sau:

Y	5,5	7	7,5	7,8	8	8,5	8,9	9,5	10	10,4
X	8	8,5	9	9,5	10	10,5	11	11,5	12	12,5

a/ Thiết lập SRF ước lượng cho mô hình: $Y = a_0 + a_1X + U$
b/ Dùng kiểm định White và kiểm định Glejser để xác minh xem mô hình SRF ở a/ có vấn đề về phương sai nhiễu thay đổi hay không.
c/ Bằng cách chọn trọng số: $W_j = 1/X_j$, hãy dùng phương pháp OLS có trọng số để tìm SRF cho hồi quy của Y theo X.
d/ Kiểm tra xem mô hình ở a/ có vấn đề về đa cộng tuyến hay không.

Chương 6.

PHÂN TÍCH ĐẶC TRƯNG VÀ LỰA CHỌN MÔ HÌNH

Chương này trình bày những vấn đề chính sau đây: (1)Phân tích đặc trưng mô hình (Các thuộc tính của một mô hình tốt, các loại sai lầm chỉ định, cách tiếp cận để lựa chọn mô hình); (2) Các kiểm định về sai lầm chỉ định; (3) Ứng dụng hồi quy trong phân tích, dự báo.

6.1. Phân tích đặc trưng mô hình

6.1.1. Các thuộc tính của một mô hình tốt.

Trong các chương trước, khi xét một mô hình, ta giả định rằng mô hình đang xét là mô hình thích hợp, nghĩa là vấn đề nghiên cứu được mô hình hóa phù hợp với bản chất của vấn đề. Tuy nhiên trong thực tế, nói chúng ta không thể tìm được mô hình chính xác hoàn toàn, mà chỉ hy vọng tìm được mô hình mô tả thực tế vấn đề một cách gần đúng có thể chấp nhận được. Theo quan điểm của A. V. Harvey các tiêu chuẩn để đánh giá một mô hình tốt là:

* *Tính tiết kiệm* (*parsimony*): mô hình càng đơn giản (nhưng phải chứa biến chính ảnh hưởng đến biến phụ thuộc) càng tốt.

* *Tính đồng nhất* (*identifiability*): với mỗi tập dữ liệu đã cho thì các tham số ước lượng được phải có giá trị thống nhất.

* *Tính thích hợp* (*goodness of fit*): Mục đích của phân tích hồi quy là giải thích sự biến động của biến phụ thuộc bằng các biến giải thích của mô hình. Mô hình càng thích hợp nếu các biến giải thích càng giải thích được nhiều sự thay đổi của biến phụ thuộc, tức là hệ số R^2 *hoặc* $\bar{R}^2$ càng lớn càng tốt (tuy nhiên không nên chỉ căn cứ vào hệ số xác định hoặc hệ số xác định điều chỉnh).

* *Tính vững về mặt lý thuyết* (*theoretical consistency*): mô hình phải phù hợp với cơ sở lý thuyết nền tảng của lĩnh vực đang xét. Nếu có hệ số xác định cao nhưng dấu của hệ số hồi quy sai thì mô hình không thể được đánh giá là tốt.

* *Khả năng dự báo tốt* (*predictiv power*): mô hình có khả năng dự báo càng chính xác, càng phù hợp với thực tế càng tốt.

6.1.2. Các loại sai lầm chỉ định

Trong mục này ta xem xét các khả năng dẫn tới một mô hình không phù hợp mà ta gọi là những sai lầm trong chỉ định.

1/ Chọn dạng hàm không thích hợp

Sai lầm này có thể dẫn đến các hậu quả sau:

- Làm sai dấu hoặc ước lượng chệch các hệ số hồi quy.
- Các ước lượng có thể không có ý nghĩa thống kê.
- Hệ số xác định R^2 không cao.
- Phần dư của các quan sát có trị tuyệt đối cao.

2/ Bỏ sót biến thích hợp

Biến thích hợp là biến có nhiều ảnh hưởng đến biến phụ thuộc. Việc bỏ sót biến thích hợp trong mô hình có thể dẫn đến các hậu quả sau:

- Các ước lượng bị chệch, khoảng tin cậy rộng ra, kém hiệu quả và do đó dễ có xu hướng chấp nhận giả thuyết.
- Hệ số xác định không cao và như vậy mức độ phù hợp của mô hình không cao.

3/ Thừa biến

Đây là việc đưa vào mô hình biến không có hoặc có ít ảnh hưởng đến biến phụ thuộc. Sai lầm này có thể không ảnh hưởng đến tính vững và không chệch của các ước lượng, nhưng các ước lượng có thể không còn tính hiệu quả ở chỗ phương sai của chúng không phải là nhỏ nhất và vì thế mà khoảng tin cậy rộng ra (kém chính xác)

Theo quan điểm của nhiều nhà kinh tế lượng thì đối với hậu quả của việc bỏ sót biến hay thừa biến, tính chất không chệch của các ước lượng được chú trọng hơn. Do vậy người ta thường chọn cách tiếp cận đi từ tổng quát đến đơn giản, chấp nhận tình huống ban đầu thừa biến hơn là thiếu biến.

6.1.3. Cách tiếp cận để lựa chọn mô hình

B1: Xác định số biến giải thích có trong mô hình.

Có 2 hướng tiếp cận:

***a/ Từ đơn giản đến tổng quát*:** Từ mô hình đơn giản, từng bước bổ sung biến giải thích vào mô hình.

Quá trình này được thực hiện thông qua kiểm định bỏ sót biến (*Omitted variables Test*).

***b/ Từ tổng quát đến đơn giản*:** Từ mô hình có đầy đủ các biến giải thích đã được xác định, từng bước loại ra những biến không quan trọng.

Quá trình này được thực hiện thông qua kiểm định thừa biến (*Redundant variables Test*).

Thường thì biến được xem xét để loại ra là biến không có cơ sở lý thuyết để cho là biến quan trọng cần giữ lại, *p – value* tương ứng của biến này trong mô hình hồi quy có giá trị không nhỏ,hệ số tương quan riêng phần của biến này với biến phụ thuộc có trị tuyệt đối nhỏ. Hướng tiếp cận thứ hai, đi từ tổng quát đến đơn giản, được nhiều nhà kinh tế lượng quan tâm hơn.

B2: Kiểm tra các vi phạm giả thiết (Kiểm định các vấn đề: đa cộng tuyến, phương sai thay đổi, tự tương quan) và khắc phục các giả thiết bị vi phạm.

B3: Chọn dạng hàm: Cơ sở để chọn dạng hàm là dựa vào cơ sở lý thuyết kinh tế, dựa vào kết quả thực nghiệm, so sánh các dạng hàm khác nhau.

B4: Căn cứ vào các tiêu chuẩn thông dụng để chọn mô hình:

1- Xem xét giá trị R^2 *hoặc* $\bar{R}^2$.

2- Giá trị của hàm hợp lý *log – likelihood* (***L***):

$$\boldsymbol{L} = -\frac{n}{2} ln\sigma^2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum U_i^2 \qquad (6.1)$$

Giá trị L càng lớn thì mô hình càng phù hợp.

Trong thực hành sử dụng *Eviews*, giá trị của hàm *log – likelihood* được ước lượng bởi công thức:

$$L = -\frac{n}{2}(1 + \log(2\pi) + \log\left(\frac{RSS}{n}\right)) \qquad (6.1a)$$

3- Tiêu chuẩn *Akaike(Akaie info criterion)* (**AIC**):

$$\boldsymbol{AIC} = \left(\frac{RSS}{n}\right).e^{2k/n} \quad (6.2)$$

trong đó *k* là số tham số trong mô hình hồi quy.

*Giá trị **AIC** càng nhỏ thì mô hình hồi quy càng phù hợp*.

Phần mềm *Eviews* ước lượng giá trị *AIC* bằng biểu thức:

$$AIC = -\frac{2L}{n} + \frac{2k}{n} \quad (6.2a)$$

4-Tiêu chuẩn *Schwarz* (*Schwarz criterion*):

$$\boldsymbol{SC} = \left(\frac{RSS}{n}\right).n^{k/n} \quad (6.3)$$

Giá trị SC càng nhỏ thì mô hình càng phù hợp.

Trong *Eviews, SC* được ước lượng bởi:

$$SC = -\frac{2L}{n} + \frac{klogn}{n} \quad (6.3a)$$

Lưu ý:

- Trong một số trường hợp, một mô hình tốt hơn theo tiêu chuẩn này thì cũng tốt hơn theo tiêu chuẩn khác. Tuy nhiên trong trường hợp tổng quát thì một mô hình có thể tốt hơn theo tiêu chuẩn này nhưng lại không tốt hơn tiêu chuẩn khác. Nếu chú ý đến độ phức tạp của mô hình thì người ta thường sử dụng tiêu chuẩn ***SC***, nếu trong phân tích chuỗi thời gian thì người ta hay sử dụng tiêu chuẩn ***AIC***.
- Việc so sánh các tiêu chuẩn giữa các mô hình yêu cầu các biến phụ thuộc phải có cùng dạng trong mô hình hồi quy. Nếu các biến phụ thuộc xuất hiện dưới các dạng khác nhau thì phải thực hiện quy đổi về dạng tương đương mới được so sánh.

6.2. Các kiểm định về sai lầm chỉ định.

6.2.1. Kiểm định bỏ sót biến.

Giả sử mô hình hồi quy ban đầu là: $Y_i = a + b.X_i + U_i$ (6.4)

Vấn đề đặt ra là liệu còn có biến giải thích nào khác nữa có ảnh hưởng quan trọng đến Y mà chưa được đưa vào mô hình hay không ? Làm thế nào để phát hiện được một biến Z có bị bỏ sót hay không? Ta phân biệt các trường hợp sau:

1. Khi có số liệu về biến Z:

Cách 1: *Dùng kiểm định t* (và $\bar{R}^2$):

- Tiến hành hồi quy mô hình (6.4) và mô hình:

$$Y_t = a_0 + b_1.X_t + b_2.Z_i + V_i \quad (6.5)$$

- Kiểm định $H_0: b_2 = 0$, đồng thời kết hợp với việc so sánh giá trị $\bar{R}^2$ của hai mô hình. Nếu biến Z là biến quan trọng bị bỏ sót thì thông thường có xu hướng bác bỏ giả thuyết H_0 và làm tăng đáng kể giá trị $\bar{R}^2$.

Nếu nghi ngờ bỏ sót nhiều biến giải thích, ta có thể áp dụng cách làm trên bằng việc xét lần lượt bổ sung từng biến một.

Cách 2*: Dùng kiểm định Wald*

Cách 3: *Dùng phương pháp nhân tử Lagrange* (*LM – Lagrange multiplier*)

Ký hiệu mô hình ban đầu (6.4) là (*R*):

$$Y_i = a + b.X_i + U_i \text{ (mô hình bị ràng buộc)} \quad (6.6)$$

Mô hình (U):

$$Y_i = a_0 + b_1.X_i + b_2.Z_i + V_i \text{ (mô hình không bị ràng buộc)} \quad (6.7)$$

Khi đó giả thuyết $H_0: b_2 = 0$ chính là: không bỏ sót biến Z. Thực hiện kiểm định giả thuyết H_0 theo các bước:

B1: Hồi quy mô hình (R), nhận được phần dư: $\widehat{U_R} = Y - \hat{a} - \hat{b}.X_i$

B2: Nếu biến Z bị bỏ sót thì ảnh hưởng của nó được quan sát qua phần dư $\widehat{U_R}$. Do đó $\widehat{U_R}$ được xem như có liên hệ với biến bị bỏ sót ($U = b_2Z + V$), ngoài ra $\widehat{U_R}$ có liên hệ với X_i . Từ đó tiến hành hồi quy $\widehat{U_R}$ theo tất cả các biến giải thích:

$$\widehat{U_R} = \alpha + \beta_1 X_i + \beta_2 Z_i + \varepsilon \quad \text{(hồi quy phụ)} \quad (6.8)$$

qua đó tính được hệ số xác định của mô hình hồi quy phụ mà ta ký hiệu là R^2_{aux}.

B3: Kiểm định giả thuyết $H_0: b_2 = 0$

Vì trên cơ sở H_0 đúng thì biến ngẫu nhiên $n.R^2_{aux} \sim \chi^2(1)$, nên với mức ý nghĩa α, nếu giá trị $n.R^2_{aux} > \chi^2_\alpha(1)$ thì ta bác bỏ H_0, nghĩa là thừa nhận biến Z bị bỏ sót.

Lưu ý: Trong mô hình hồi quy bội, nếu nghi ngờ bỏ sót một số biến giải thích, ta tiến hành kiểm định tương tự:

- Mô hình ban đầu (R): $Y = a + b_1X_1 + \cdots + b_mX_m + U$
- Mô hình (U): $Y = a + b_1X_1 + \cdots + b_mX_m + b_{m+1}X_{m+1} + \cdots + b_kX_k + V$

Giả thiết: không bỏ sót các biến $X_{m+1}, \dots, X_k$ chính là: $H_0: b_{m+1} = \cdots = b_k = 0$

đối thuyết H_1: bỏ sót ít nhất một trong các biến $X_{m+1}, \dots, X_k$, tức là có ít nhất một trong các hệ số $b_{m+1}, \dots, b_k$ khác không.

- Hồi quy mô hình (R), thu được phần dư $\widehat{U_R}$.
- Tiến hành hồi quy phụ: $\widehat{U_R} = \alpha + \beta_1X_1 + \cdots + \beta_kX_k + \varepsilon$

từ đó tính được hệ số xác định R^2_{aux} của mô hình hồi quy phụ.

- Nếu $n.R^2_{aux} > \chi^2_\alpha(k-m)$ thì bác bỏ H_0.

Cách 3: *Dùng tỷ số hàm hợp lý* (*Likelihood ratio – LR*)

Giả thuyết H_0: không bỏ sót biến

- Hồi quy mô hình ban đầu (R):

$$Y = a + b_1X_1 + \cdots + b_mX_m + U \quad (6.9)$$

- Hồi quy mô hình (U):

$$Y = a + b_1X_1 + \cdots + b_mX_m + b_{m+1}X_{m+1} + \cdots + b_kX_k + V. \quad (6.10)$$

Ký hiệu l_R và l_U là giá trị lớn nhất của logarit hàm hợp lý ứng với mô hình (R) và mô hình (U) tương ứng. Xét thống kê: $LR = -2(l_R - l_U)$ ($(k-m)$ là số biến giải thích nghi ngờ bị bỏ sót).

Từ kết quả hồi quy, nếu giá trị $LR > \chi^2_\alpha(k-m)$ thì bác bỏ giả thuyết H_0.

2. Khi không có số liệu về biến Z

* Cách 1: *Kiểm định của Ramsey* (hay kiểm định *RESET: Regression Specification Error Test*): *Ramsey* đã đề xuất sử dụng tổ hợp tuyến tính của $\hat{Y}^2, \hat{Y}^3, \hat{Y}^4$xấp xỉ cho Z.

Thủ tục kiểm định như sau:

B1: Hồi quy mô hình gốc:

$Y_i = a + b.X_i + U_i$ $(\boldsymbol{old})$, thu được $\hat{Y}_i$

B2: Hồi quy mô hình: $Y_i = \alpha + \beta_1X_i + \beta_2\hat{Y}_i^2 + \beta_3\hat{Y}_i^3 + \beta_4\hat{Y}_i^4 + V_i$ $(\boldsymbol{new})$

B3: Kiểm định giả thuyết $H_0: \beta_2 = \beta_3 = \beta_4 = 0$ (không bỏ sót biến)

- Xét thống kê:

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

trong đó *m* là số biến giải thích mới được thêm vào trong mô hình (*new*) (cụ thể ở đây *m* = 3); *k* là số hệ số của mô hình (*new*) (cụ thể ở đây *k* = 5)

- Nếu giá trị $F > F_\alpha(m, n-k)$ thì bác bỏ H_0.

* Cách 2: Kiểm định nhân tử *Lagrange*

B1: Hồi quy mô hình (old), thu được phần dư: $\hat{U}_i$.

B2: Hồi quy: $\hat{U}_i = \alpha + \beta_1 X_i + \beta_2 \hat{Y}_i^2 + \beta_3 \hat{Y}_i^3 + \beta_4 \hat{Y}_i^4 + \varepsilon_i$ (6.11)

tính được hệ số xác định R^2_{aux} *của mô hình này.*

B3: Kiểm định giả thuyết $H_0: \beta_2 = \beta_3 = \beta_4 = 0$ (không bỏ sót biến) như sau:

- Nếu $n.R^2_{aux} > \chi^2_\alpha(m)$ thì bác bỏ H_0 (m là số biến giải thích mới thêm vào mo hình, cụ thể ở đây: m = 3)

Lưu ý: Có thể dựa trên khảo sát dạng đồ thị của phần dư $\hat{U}_i$ *theo* $\hat{Y}_i$ để xác định bậc của $\hat{Y}$. Thông thường bậc của $\hat{Y}$ càng cao thì kết quả kiểm định càng chính xác.

* Cách 3: *Kiểm định Durbin – Watson*

B1: Hồi quy mô hình (old) và nhận được $Y = \hat{a} + \hat{b}.X + \hat{U}$

B2: Giả sử biến Z là biến được nghi ngờ bị bỏ sót. Sắp xếp phần dư $\hat{U}_i$ theo thứ tự tăng của Z hoặc theo thứ tự tăng của một biến giải thích nếu số liệu của Z chưa có.

B3: Tính $d = \frac{\sum_{i=2}^n (\hat{U}_i - \hat{U}_{i-1})^2}{\sum_{i=1}^n \hat{U}_i^2}$ (chú ý rằng ở **B2**, dãy phần dư đã được sắp xếp lại nên biểu thức này không phải là thống kê *Durbin – Watson*)

B4: Giả thuyết H_0: Dạng hàm đúng (không có Z),
Đối thuyết H_1: Dạng hàm sai (biến Z bị bỏ sót)

Dựa vào bảng *Durbin – Watson* và mức ý nghĩa để bác bỏ hay chấp nhận H_0.

6.2.2. Kiểm định thừa biến

Đối với kiểm định thừa biến, giả thuyết H_0 là thừa biến, có thể dùng các kiểm định sau:

1. *Kiểm định t thông thường* (để xét bỏ một biến)

2. *Kiểm định Wald* (để xét bỏ một hoặc nhiều biến)

Giả sử có mô hình hồi quy: $Y = a_0 + a_1 X_1 + \cdots + a_m X_m + \cdots + a_k X_k + U$

Trước hết, nếu về mặt lý thuyết cho rằng tất cả các biến $X_{m+1}, \ldots, X_k$ đều quan trọng đối với Y thì ta phải giữ lại chúng trong mô hình, cho dù trong hồi quy ước lượng hệ số của một trong chúng không có ý nghĩa thống kê. Nếu không chắc chúng có thực sự cần thiết trong mô hình thì ta dùng kiểm định Wald, theo các bước sau:

B1: Chạy hồi quy các mô hình:

(U): $Y = a_0 + a_1 X_1 + \cdots + a_m X_m + \cdots + a_k X_k + U$ (gọi là mô hình không giới hạn), có được tổng RSS của mô hình này, ký hiệu là RSS_U.

(R): $Y = a_0 + a_1 X_1 + \cdots + a_m X_m + V$(gọi là mô hình giới hạn), có được tổng RSS của mô hình này, ký hiệu là RSS_R.

- Thiết lập thống kê: $F = \frac{(RSS_R - RSS_U)/(k-m)}{RSS_U/(n-k)}$

B2: Với mức ý nghĩa α cho trước (hoặc mặc định), tìm giá trị tới hạn: $F_\alpha(k-m, n-k)$.

B3: Quy tắc bác bỏ giả thuyết: H_0 là: $F > F_\alpha(k-m, n-k)$.

3. *Kiểm định tỷ số hàm hợp lý* (như trong kiểm định bỏ sót biến)

Lưu ý: Trong thực hành có thể thực hiện kiểm định *Wald*, kiểm định tỷ số hàm hợp lý nhờ vào Eviews.

Ví dụ 6.1(Hàm sản xuất của *Đài Loan*): Số liệu về Y(GNP: đơn vị: triệu $ Đài Loan), lượng lao động X_1 (ngàn người), lượng vốn thực X_2 (triệu $ Đài Loan) và biến xu hướng thời gian X_3 xếp thứ tự từ năm 1958 đến năm 1972 của Đài Loan được cho bởi bảng sau:

Năm	Y	X_1	X_2	X_3	Năm	Y	X_1	X_2	X_3
1958	8911.4	281.5	120753	1	1966	23052	*616.7*	153714	9
1959	10873.2	284.4	122242	2	1967	26128.2	695.7	164783	10
1960	11132.5	289	125263	3	1968	29563.7	730.3	176864	11
1961	12086.5	375.8	128539	4	1969	33373.6	816	188146	12
1962	12767.5	375.2	131427	5	1970	38354.3	848.4	205841	13
1963	16347.1	402.5	134267	6	1971	46868.3	873.1	221748	14
1964	19542.7	478	139038	7	1972	54308	999.2	239715	15
1965	21075.9	553.4	146450	8					

Bảng 6.1. Số liệu về GNP, lượng lao động và vốn ở Đài Loan

(Nguồn: Thomas Pei-Fan Chen, "Economic Growth and Structural Change in Taiwan 1952-1972, A Production Approach". (D.N. Gujarati))

Giả sử hàm sản xuất đúng, theo lý thuyết thì mô hình Cobb-Douglas có dạng:

$$lnY_t = a + b_1 lnX_{1t} + b_2 lnX_{2t} + U_t \qquad \text{(a1)}$$

Giả sử ta không đưa biến X_2 vào và tiến hành hồi quy lnY theo lnX_1,tức là hồi quy mô hình:

$$lnY_t = a + b_1 lnX_{1t} + U_t \qquad \text{(a2)}$$

thì kết quả hồi quy được cho bởi bảng sau:

Dependent Variable: LOG(Y)
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.069560	0.417743	4.954143	0.0003
LOG(X1)	1.257567	0.066516	18.90615	0.0000
R-squared	0.964907	Mean dependent var		9.949171
Adjusted R-squared	0.962207	S.D. dependent var		0.566287

Bảng 6.2. Kết quả hồi quy GNP theo lượng lao động

Qua đó thấy mức độ phù hợp của mô hình rất cao (R^2 = 0.964907), hệ số hồi quy có ý nghĩa thống kê (đó là hệ số co giãn của sản lượng đối với lao động). Theo lý thuyết kinh tế thì lao động không phải là yếu tố duy nhất tác động đến GNP nên giả thuyết có biến bị bỏ sót là có cơ sở.

- Tiến hành kiểm định bỏ sót biến (biến nghi ngờ bỏ sót: log(X_2)), ta có kết quả:

Omitted Variables Test
Equation: UNTITLED
Specification: LOG(Y) C LOG(X1)
Omitted Variables: LOG(X2)

	Value	df	Probability
t-statistic	3.722069	12	0.0029
F-statistic	13.85380	(1, 12)	**0.0029**
Likelihood ratio	11.51326	1	**0.0007**

Bảng 6.3

nhận được *p – value* của thống kê F là **0.0029** và của tỷ số hợp lý *log-likelihood* là 0.0007, đều rất bé (< 0.01 < 0.05) nên ta bác bỏ giả thuyết H_0 và thừa nhận đã bỏ sót biến quan trọng log(X_2).

- Nếu không có số liệu về X_2, ta dùng kiểm định *Ramsey's RESET*

Ramsey RESET Test
Equation: UNTITLED
Specification: LOG(Y) C LOG(X1)
Omitted Variables: Squares of fitted values

	Value	df	Probability
t-statistic	2.223393	12	0.0462
F-statistic	4.943477	(1, 12)	0.0462
Likelihood ratio	5.174644	1	0.0229

Bảng 6.4. Kết quả kiểm định Ramsey's RESET

Có *p – value* của thống kê F là 0.0462 và của tỷ số hợp lý *log-likelihood* là 0.0229, đều bé (< 0.05) nên ta bác bỏ giả thuyết H_0 và thừa nhận đã bỏ sót biến quan trọng log(X_2).

Bây giờ đưa biến bị bỏ sót log(X_2) bổ sung vào mô hình (a2), ta đi tới hồi quy mô hình (a1) và nhận được kết quả:

Dependent Variable: LOG(Y)
Method: Least Squares
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7.843851	2.679836	-2.926989	0.0127
LOG(X1)	0.714779	0.153268	4.663602	0.0005
LOG(X2)	1.113474	0.299154	3.722069	0.0029

R-squared	0.983712	Mean dependent var	9.949171
Adjusted R-squared	0.980997	S.D. dependent var	0.566287
S.E. of regression	0.078064	Akaike info criterion	-2.085725
Sum squared resid	0.073127	Schwarz criterion	-1.944115
Log likelihood	18.64294	Hannan-Quinn criter.	-2.087234
F-statistic	362.3594	Durbin-Watson stat	1.416754
Prob(F-statistic)	0.000000		

Bảng 6.5. Hồi quy GNP theo lượng lao động và vốn

Tiếp tục kiểm định bỏ sót biến xu hướng X_3

Omitted Variables Test
Equation: UNTITLED
Specification: LOG(Y) C LOG(X1) LOG(X2)
Omitted Variables: X3

	Value	df	Probability
t-statistic	3.602328	11	0.0042
F-statistic	12.97677	(1, 11)	0.0042
Likelihood ratio	11.68785	1	0.0006

Bảng 6.6

Tiếp tục bổ sung biến xu hướng X_3 vào mô hình (a1), nhận được kết quả:

Dependent Variable: LOG(Y)
Method: Least Squares
Sample: 1958 1972
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.944253	4.024478	1.228545	0.2449
LOG(X1)	-0.121815	0.256302	-0.475277	0.6439
LOG(X2)	0.403372	0.289219	1.394693	0.1906
X3	0.118107	0.032786	3.602328	0.0042

R-squared	0.992527	Mean dependent var	9.949171
Adjusted R-squared	0.990489	S.D. dependent var	0.566287
S.E. of regression	0.055226	Akaike info criterion	-2.731582
Sum squared resid	0.033549	Schwarz criterion	-2.542769
Log likelihood	24.48686	Hannan-Quinn criter.	-2.733593
F-statistic	487.0038	Durbin-Watson stat	1.496272
Prob(F-statistic)	0.000000		

Bảng 6.7. Hồi quy GNP theo lao động, vốn và biến xu hướng

Chú ý: Thực ra trong bảng kiểm định bỏ sót biến đã có kết quả hồi quy có bổ sung biến này ở phần cuối của bảng.

Nhận xét: Qua các kết quả trên, nhận thấy rằng:

- Giá trị của các hệ số hồi quy thay đổi đáng kể khi sử dụng các mô hình khác nhau. Hệ số hồi quy trong mô hình tuyến tính log đóng vai trò là hệ số co giãn. Theo đó trong mô hình 2 biến, hệ số co giãn biểu thị cho tác động của lượng lao động đối với GNP nhận được ước lượng khoảng 1,257%, tuy nhiên trong mô hình 3 biến thì tác động của lượng lao động đối với GNP chỉ nhận được ước lượng khoảng 0.714%. Điều này cho thấy hậu quả của việc bỏ sót biến quan trọng đã làm cho ta ước lượng quá cao tác động của các biến có trong mô hình.

- Việc biến xu hướng có mặt trong mô hình đã làm thay đổi dấu của hệ số hồi quy, trái với tác động đồng biến như đã phân tích trong lý thuyết kinh tế giữa Y và X_1. Trong khi đó hệ số xác định của mô hình này rất cao. Điều này khiến ta nghi ngờ có hiện tượng đa cộng tuyến rất cao giữa các biến giải thích, tức là các biến này chịu tác động lớn của biến xu hướng.
- Trong mô hình hồi quy ở bảng 6.7, ta tiến hành kiểm định giả thuyết:

$$H_0: \textit{không có đa cộng tuyến trong mô hình.}$$

Hồi quy mô hình phụ của X_3 theo log(X_1) và log(X_2), ta nhận được:

Dependent Variable: X3
Method: Least Squares
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-108.2759	16.69252	-6.486493	0.0000
LOG(X1)	7.083379	0.954694	7.419528	0.0000
LOG(X2)	6.012380	1.863413	3.226542	0.0073
R-squared	0.989867	Mean dependent var		8.000000
Adjusted R-squared	0.988178	S.D. dependent var		4.472136
F-statistic	586.1090	Durbin-Watson stat		1.414549
Prob(F-statistic)	0.000000			

Bảng 6.8. Kết quả hồi quy phụ của biến X_3 theo log(X1) và log(X2)

Từ đó nhận được: p-value (của F) ≈ 0.000000 nên ta bác bỏ giả thuyết H_0, thừa nhận có đa cộng tuyến giữa các biến giải thích, với mức độ cao, vì giá trị VIF rất lớn:

$$VIF = \frac{1}{1-R^2} = \frac{1}{1-0.989867} \approx 98.6875$$

- Trong các mô hình 2, 3 biến, các hệ số hồi quy có các *p – value* rất bé, tức là có ý nghĩa thống kê. Tuy nhiên trong mô hình có biến xu hướng, các hệ số hồi quy lại không có ý nghĩa thống kê. Đó là do hậu quả của hiện tượng đa cộng tuyến cao giữa các biến X_1 và X_3, X_2 và X_3.
- Nói về mức độ phù hợp của các mô hình: Trong các mô hình trên, do số biến giải thích không bằng nhau nên để đánh giá mức độ phù hợp của mô hình, ngoài hệ số xác định, ta còn phải dựa vào hệ số xác định điều chỉnh $\bar{R}^2$. Mặc dù có $\bar{R}^2$ cao nhất nhưng mô hình có biến xu hướng sẽ không được chọn vì có hiện tượng đa cộng tuyến nghiêm trọng (dấu của hệ số hồi quy của X_1 sai, hệ số hồi quy của X_1, X_2 không có ý nghĩa thống kê, dạng của mô hình không phù hợp với cơ sở lý thuyết, hệ số $\bar{R}^2$ không vượt quá nhiều so với các mô hình khác). Nếu dựa vào tiêu chuẩn *log – likelihood* hay *AIC* hoặc *Schwarz* thì mô hình có biến xu hướng có kết quả tốt hơn, nhưng kết quả này lại không đáng tin cậy do ảnh hưởng của đa cộng tuyến cao trong mô hình. Mô hình 3 biến sẽ là sự lựa chọn thích hợp, vì nó phù hợp với cơ sở lý thuyết, các hệ số hồi quy đều có ý nghĩa th.kê, $\bar{R}^2 = 0.983712$, còn mô hình 2 biến do bỏ sót biến quan trọng nên có ước lượng về ảnh hưởng của biến giải thích quá cao.

Sau khi đã lựa chọn được mô hình thích hợp là mô hình 3 biến, ta tiến hành kiểm định các vấn đề đa cộng tuyến, phương sai thay đổi và tự tương quan cho mô hình này.

* *Kiểm định về đa cộng tuyến*:
Sử dụng hồi quy phụ *log(X_1)* theo *log(X_2)*: Kiểm định giả thuyết: H_0: không có đa cộng tuyến:

Dependent Variable: LOG(X1)
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-15.94080	1.992390	-8.000841	0.0000
LOG(X2)	1.857120	0.166594	11.14755	0.0000
R-squared	0.905295	Mean dependent var		6.265758
Prob(F-statistic)	0.000000			

Bảng 6.9. Hồi quy phụ log(X1) theo log(X2)

Từ kết quả hồi quy phụ (R^2 = 0.905295, Prob(F-statistic) ≈ 0.000000) ta bác bỏ H_0, tức là có đa cộng tuyến giữa log(X_1) và log(X_2), tuy nhiên $VIF = \frac{1}{1-R^2} \approx 10.5591$ nên đa cộng tuyến ở đây không nghiêm trọng như trong mô hình bốn biến.

* *Kiểm định phương sai thay đổi*:
Sử dụng kiểm định White (có số hạng tích chéo), nhận được:

Heteroskedasticity Test: White

F-statistic	1.823062	Prob. F(4,10)	0.2010
Obs*R-squared	6.325594	Prob. Chi-Square(4)	0.1761
Scaled explained SS	2.356287	Prob. Chi-Square(4)	0.6705

Bảng 6.10. Kiểm định phương sai thay đổi cho mô hình 3 biến

Kết quả kiểm định (*p-value* = 0.1761 > 0.05) cho thấy không có hiện tượng phương sai thay đổi trong mô hình ba biến.

* Kiểm định *Breusch-Pagan-Godfrey*:

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	2.070338	Prob. F(2,12)	0.1689
Obs*R-squared	3.848050	Prob. Chi-Square(2)	0.1460
Scaled explained SS	1.433401	Prob. Chi-Square(2)	0.4884

Bảng 6.11. *Kiểm định Breusch-Pagan-Godfrey*

cho cùng kết luận: không có hiện tượng phương sai thay đổi trong mô hình ba biến (vì *p-value* = 0.1460 > 0.05)

* Kiểm định tự tương quan: sử dụng kiểm định BG:

- Với lags = 1, ta nhận được:

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.463262	Prob. F(1,11)	0.5102
Obs*R-squared	0.606191	Prob. Chi-Square(1)	0.4362

Bảng 6.12. *Kiểm định BG với tự tương quan bậc 1*

-Với lags = 2, ta nhận được:

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	1.935312	Prob. F(2,10)	0.1948
Obs*R-squared	4.185779	Prob. Chi-Square(2)	0.1233

Bảng 6.13. *Kiểm định BG với tự tương quan bậc 2*

- Với lags = 3, ta nhận được:

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	1.161208	Prob. F(3,9)	0.3769
Obs*R-squared	4.185831	Prob. Chi-Square(3)	0.2421

Bảng 6.13. *Kiểm định BG với tự tương quan bậc 3*

Các kết quả kiểm định đều cho kết luận như nhau là chấp nhận giả thiết H_0 (*Prob. F > 0.05; Prob. Chi-Square > 0.05*), tức là không có tự tương quan trong mô hình 3 biến. Vậy việc chọn mô hình ba biến là thích hợp nhất.

Ví dụ 6.2: Lượng hàng bán được Y(kg/tháng) của mặt hàng A, giá bán X_1 của mặt hàng A, giá bán X_2 của mặt hàng B, qua điều tra ở các khu vực nông thôn và thành phố, thu được số liệu sau đây, trong đó: Z = 1, nếu khu vực bán là thành phố, Z = 0, nếu khu vực bán là nông thôn:

Y	X1	X2	Z	Y	X1	X2	Z
14	5	14	0	15	5	12	1
14	6	15	1	15	5	13	1
13	6	13	0	16	4	17	1
12	7	14	1	16	4	10	0
12	7	12	0	17	3	16	1
15	5	16	1	17	4	11	1
16	4	15	0	18	4	16	0
12	7	18	1	18	3	15	1
10	8	16	0	19	3	13	0
11	8	20	1	20	2	14	1

Bảng 6.14

* Hồi quy Y theo X1, X2, Z theo mô hình: $Y = a_0 + a_1X_1 + a_2X_2 + a_3.Z + U$,
ta có bảng kết quả:

Dependent Variable: Y
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	22.39719	0.962214	23.27674	0.0000
X1	-1.541265	0.095086	-16.20923	0.0000
X2	0.018480	0.072000	0.256667	0.8007
Z	0.068620	0.329596	0.208193	0.8377
R-squared	0.948921	Mean dependent var		15.00000

Bảng 6.15

Từ bảng trên, nhận thấy hệ số hồi quy ước lượng của X2 và Z khác không không có ý nghĩa. Để có cơ sở kết luận có nên để lại hai biến X2 và Z trong mô hình hay không, ta thực hiện kiểm định giả thuyết: $H_0: a_2 = a_3 = 0$.

- Sử dụng kiểm định *Wald, Eviews* cho ta kết quả:

Wald Test:
Equation: EQ01

Test Statistic	Value	df	Probability
F-statistic	0.082219	(2, 16)	0.9215
Chi-square	0.164439	2	0.9211

Null Hypothesis: C(3)=C(4)=0
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(3)	0.018480	0.072000
C(4)	0.068620	0.329596

Restrictions are linear in coefficients.

Bảng 6.16

Theo đó thống kê F có: p – value = 0,9215 > 0,05, nên ta chấp nhận giả thuyết H_0, và cho rằng: cả hai biến: X2 và Z đều không cần thiết đưa vào mô hình.

- Sử dụng kiểm định tỷ số hợp lý, Eviews cho kết quả:

Redundant Variables Test
Equation: EQ01
Specification: Y C X1 X2 Z
Redundant Variables: X2 Z

	Value	df	Probability
F-statistic	0.082219	(2, 16)	0.9215
Likelihood ratio	0.204499	2	0.9028

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	0.075595	2	0.037797
Restricted SSR	7.431034	18	0.412835
Unrestricted SSR	7.355440	16	0.459715
Unrestricted SSR	7.355440	16	0.459715

LR test summary:

	Value	df
Restricted LogL	-18.47810	18
Unrestricted LogL	-18.37585	16

Restricted Test Equation:
Dependent Variable: Y
Method: Least Squares
Date: 08/16/13 Time: 22:48
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	22.67241	0.445632	50.87697	0.0000
X1	-1.534483	0.084367	-18.18811	0.0000

R-squared	0.948396	Mean dependent var	15.00000
Adjusted R-squared	0.945529	S.D. dependent var	2.752989
S.E. of regression	0.642523	Akaike info criterion	2.047810
Sum squared resid	7.431034	Schwarz criterion	2.147383
Log likelihood	-18.47810	Hannan-Quinn criter.	2.067248
F-statistic	330.8074	Durbin-Watson stat	1.808905
Prob(F-statistic)	0.000000		

Bảng 6.17

Theo đó, p – value các thống kê F và tỷ số hợp lý đều lớn hơn 0,05 nên ta chấp nhận giả thuyết H_0, và cho rằng: cả hai biến: X2 và Z đều không cần thiết đưa vào mô hình.

Lưu ý: Trong kiểm định *Likelihood ratio*, nửa cuối bảng kết quả, Eviews đã cung cấp cả kết quả hồi quy sau khi đã loại các biến không cần thiết.

6.3. Ứng dụng hồi quy trong phân tích, dự báo.

Trong mục này, ta đề cập đến vấn đề dự báo thông qua mô hình kinh tế lượng. Như đã chỉ ra, mối quan hệ giữa biến phụ thuộc và các biến giải thích là quan hệ phụ thuộc thống kê: Ứng với mỗi giá trị của biến giải thích, có thể có nhiều giá trị của biến phụ thuộc mà người ta gọi là những giá trị cá biệt. Tuy nhiên trong thực tế, người ta thường quan tâm đến giá trị trung bình của những giá trị cá biệt, đó chính là giá trị trung bình có điều kiện của biến phụ thuộc với điều kiện cho trước giá trị của biến giải thích. Trong phần này ta sẽ đề cập hai loại dự báo: Dự báo giá trị cá biệt và dự báo giá trị trung bình, với hai phương pháp: dự báo điểm và dự báo khoảng (hay ước lượng điểm và ước lượng khoảng tin cậy).

6.3.1. Dự báo với mô hình hai biến

Giả sử mô hình hồi quy hai biến sau đây đã được xác định là phù hợp tốt:

$$PRF: \begin{cases} E(Y|X) = a + b.X \\ Y = a + b.X + U \end{cases} \quad ; \quad SRF: \begin{cases} \hat{Y} = \hat{a} + \hat{b}.X \\ Y = \hat{a} + \hat{b}.X + \hat{U} \end{cases}$$

Bây giờ ta sử dụng mô hình này để dự báo giá trị trung bình có điều kiện $E(Y|X)$ và giá trị cá biệt của Y khi cho biến giải thích X nhận giá trị x_0.

6.3.1.1. Dự báo điểm (*Point Prediction*)

Trước hết cần lưu ý rằng: các hệ số hồi quy ước lượng $\hat{a},\ \hat{b}$ phụ thuộc vào mẫu nên chúng là các đại lượng ngẫu nhiên mà ứng với một mẫu cụ thể chúng có giá trị xác định. Vì thế khi thay $X = x_0$ vào SRF, ta nhận được:

$$\hat{Y}_0 = \hat{a} + \hat{b}.x_0$$

là một đại lượng ngẫu nhiên (do $\hat{a},\ \hat{b}$ là các đại lượng ngẫu nhiên). Theo định lý *Gauss – Markov*, $\hat{Y}_0$ là ước lượng tuyến tính không chệch tốt nhất của $E(Y|X = x_0)$. Vì vậy ta dùng $\hat{Y}_0$ để ước lượng điểm cho cả giá trị trung bình và giá trị cá biệt của biến phụ thuộc Y_{x_0}.

Cũng cần lưu ý rằng Y_{x_0} là những giá trị của biến phụ thuộc Y ứng với $X = x_0$ (Quan hệ giữa Y và X là quan hệ phụ thuộc thống kê), vậy nên Y_{x_0} cũng là một biến ngẫu nhiên mà kỳ vọng $E(Y_{x_0}) = E(Y|X = x_0)$.

6.3.1.2. Dự báo khoảng (*Interval Prediction*)

Với điều kiện nhiễu $U \sim N(0;\ \sigma^2)$ thì $\hat{Y}_0 = \hat{a} + \hat{b}.x_0$ có phân phối chuẩn, với:
Kỳ vọng:

$$E\hat{Y}_0 = E(\hat{a} + \hat{b}.x_0) = a + b.x_0 = E(Y|X = x_0)$$

và phương sai:

$$var(\hat{Y}_0) = \sigma^2.\left\{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{nS^2(X)}\right\} \qquad (6.12)$$

Người ta chỉ ra được rằng :

$$t = \frac{\hat{Y}_0 - E(Y|X = x_0)}{se(\hat{Y}_0)}$$

là biến ngẫu nhiên có phân phối Student với (n – 2) bậc tự do

a. Dự báo khoảng cho giá trị trung bình $E(Y|X = x_0)$:
Với độ tin cậy $\gamma = 1 - \alpha$, khoảng tin cậy cho giá trị trung bình $E(Y|X = x_0)$ là:

$$\left(\hat{Y}_0 - t_{n-2}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{Y}_0);\ \hat{Y}_0 + t_{n-2}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{Y}_0)\right) \qquad (6.13)$$

b. Dự báo khoảng cho giá trị cá biệt Y_{x_0}:
Khi dùng $\hat{Y}_0$ để dự báo cho Y_{x_0} thì sai số dự báo là: $\hat{U}_0 = Y_{x_0} - \hat{Y}_0$ là một đại lượng ngẫu nhiên. Người ta chỉ ra được rằng:
$var(\hat{U}_0) = var(\hat{Y}_0) + \sigma^2$,
và đại lượng: $t = \frac{\hat{U}_0}{se(\hat{U}_0)}$ có phân phối Student với (n – 2) bậc tự do.
Vì thế: Với độ tin cậy $\gamma = 1 - \alpha$, khoảng tin cậy cho giá trị cá biệt của biến phụ thuộc là:

$$\left(\hat{Y}_0 - t_{n-2}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{U}_0);\ \hat{Y}_0 + t_{n-2}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{U}_0)\right) \qquad (6.14)$$

Nhận xét: Do: $var(\hat{U}_0) = var(\hat{Y}_0) + \sigma^2 > var(\hat{Y}_0)$, nên khoảng tin cậy cho giá trị cá biệt sẽ rộng hơn và bao hàm cả khoảng dự báo cho giá trị trung bình.

Chú ý: Để tiến hành dự báo, ta có thể chạy trên phần mềm Eviews để nhận được kết quả. Tuy nhiên, nếu trường hợp không chạy trên Eviews hay các phần mềm khác, ta tiến hành tính toán trực tiếp như sau:
- Với mô hình SRF đã có, thay $X := x_0, tìm\ được\ dự\ báo\ điểm\ \hat{Y}_0$.
- Tính giá trị:

$$\widehat{se}(\hat{Y}_0) = \sqrt{\widehat{var}(\hat{Y}_0)} = \hat{\sigma}.\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{nS^2(X)}} \qquad (6.15)$$

$$\widehat{se}(\hat{U}_0) = \sqrt{\widehat{var}(\hat{U}_0)} = \sqrt{\widehat{var}(\hat{Y}_0) + \hat{\sigma}^2} \qquad (6.16)$$

- Tra bảng tìm $t_{n-2}\left(\frac{\alpha}{2}\right)$
- Suy ra khoảng tin cậy cần tìm.

Ví dụ 6.3: Y là GDP bình quân của Việt Nam từ 1998 – 2006 có số liệu như sau:

Năm	1998	1999	2000	2001	2002	2003	2004	2005	2006
Y	360	374	401	413	440	489	553	618	655

a/ Thiết lập SRF ước lượng cho mô hình: $Y = a + bT + U$

trong đó T là biến xu thế (T = 1, ứng với năm 1998, T = 2, ứng với năm 1999,…, T = 9, ứng với năm 2006)
b/ Sử dụng mô hình được thiết lập để:
- Dự báo GDP cho năm 2007.
- Dự báo khoảng cho giá trị cá biệt và cho giá trị trung bình của GDP năm 2007 với độ tin cậy 95%.
Giải:
▪ Nếu không sử dụng Eviews, tính trực tiếp, ta có:

$$\hat{b} = 38{,}2;\ \hat{a} = 287{,}1111; RSS = 5458{,}489;\ \hat{\sigma} = 27{,}92462;\ R^2 = 0{,}941315;$$
$$\bar{T} = 5;\ S^2(T) = 6{,}66667$$

a/ SRF nhận được: $\hat{Y} = 287{,}1111 + 38{,}2.T$
b/ Thay T = 10, tính được:
b1- Dự báo GDP cho năm 2007 là: $\hat{Y}_0 = 287{,}1111 + 38{,}2 * 10 = 669{,}1111$
b2- Tính:

$$\widehat{se}(\hat{Y}_0) = \sqrt{\widehat{var}(\hat{Y}_0)} = \hat{\sigma}.\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{T})^2}{nS^2(T)}} = 90{,}60570;$$

$$\widehat{se}(\hat{U}_0) = \sqrt{\widehat{var}(\hat{Y}_0) + \hat{\sigma}^2} = 94{,}81127$$

- Với độ tin cậy: $\gamma = 1 - \alpha = 0{,}95;\ \alpha = 0{,}05;\ t_{n-2}\left(\frac{\alpha}{2}\right) = t_7(0{,}025) = 2{,}365$
- Suy ra khoảng dự báo cho giá trị trung bình của GDP năm 2007 là:

$$\left(\hat{Y}_0 - t_{n-2}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{Y}_0);\ \hat{Y}_0 + t_{n-2}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{Y}_0)\right) = (454{,}82862; 883{,}33936)$$

- Suy ra khoảng dự báo cho giá trị cá biệt của GDP năm 2007 là:

$$\left(\hat{Y}_0 - t_{n-2}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{U}_0);\ \hat{Y}_0 + t_{n-2}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{U}_0)\right) = (444{,}88245; 893{,}33975)$$

6.3.2. Dự báo với mô hình nhiều biến

Xét mô hình hồi quy nhiều biến (k biến) dạng ma trận:
PRF ngẫu nhiên : $\mathcal{Y} = \mathcal{X}.a + \mathcal{U}$;
SRF ngẫu nhiên : $\mathcal{Y} = \mathcal{X}.\hat{a} + \hat{\mathcal{U}}$

6.3.2.1. Dự báo điểm:

Cho các biến giải thích X = (X_1, X_2,..., X_{k-1}) nhận giá trị: (X_{10}, X_{20},..., $X_{k-1,0}$),
ký hiệu: $X^0 = \left(1, X_{10}, X_{20}, \dots, X_{(k-1),0}\right)^T$, khi đó dựa vào SRF, ta nhận được:

$$\hat{Y}_0 = (X^0)^T.\hat{a} \qquad (6.17)$$

là ước lượng điểm tuyến tính, không chệch tốt nhất cho giá trị trung bình có điều kiện: $E(Y|X = X^0)$ và cho các giá trị cá biệt Y_{X^0} (là các giá trị của Y ứng với $X = X^0$).
$\hat{Y}_0$ là dự báo điểm.

6.3.2.2. Dự báo khoảng:

Tương tự như trong mô hình hai biến, với độ tin cậy $\gamma = 1 - \alpha$, ta có:

a. Dự báo khoảng cho giá trị trung bình $E(Y|X = X^0)$:

$$\left(\hat{Y}_0 - t_{n-k}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{Y}_0);\ \hat{Y}_0 + t_{n-k}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{Y}_0)\right) \qquad (6.18)$$

(Trong đó: $\widehat{se}(\hat{Y}_0) = \sqrt{\widehat{var}(\hat{Y}_0)} = \sqrt{\hat{\sigma}^2.(X^0)^T.(X^T.X)^{-1}.X^0}$)

b. Dự báo khoảng cho các giá trị cá biệt Y_{X^0}:

$$\left(\hat{Y}_0 - t_{n-k}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{U}_0);\ \hat{Y}_0 + t_{n-k}\left(\frac{\alpha}{2}\right).\widehat{se}(\hat{U}_0)\right) \qquad (6.19)$$

(Trong đó: $\widehat{se}(\hat{U}_0) = \sqrt{\widehat{var}(\hat{U}_0)} = \sqrt{\hat{\sigma}^2.[(X^0)^T.(X^T.X)^{-1}.X^0 + 1]}$)

Chú ý: Để tìm các khoảng dự báo (6.18), (6.19), nói chung ta phải chạy trên phần mềm ứng dụng, như Eviews. Nếu tính trực tiếp có sự kết hợp với bảng hồi quy, ta cần phải tính toán khá phức tạp, trong đó cần tìm ma trận nghịch đảo: $(X^T.X)^{-1}$(với mô hình gồm k biến thì $X^T.X$ là nhân 2 ma trận vuông cấp k).

6.3.3. Đánh giá độ chính xác của dự báo

6.3.3.1. *Mẫu khởi động và mẫu kiểm tra*:

Một mô hình tốt phải là một mô hình có khả năng dự báo với độ chính xác cao. Để đánh giá mức độ chính xác về dự báo của mô hình đòi hỏi phải có các số liệu theo hai hướng: hoặc đưa thêm số liệu mới qua điều tra điều tra bổ sung để làm mẫu kiểm tra; hoặc phân chia mẫu hiện có thành hai mẫu con, mẫu con thứ nhất dùng để ước lượng mô hình hồi quy – gọi là mẫu khởi động (*initialization set*), mẫu con thứ hai – gọi là mẫu kiểm tra (*test set*), dùng để kiểm tra độ chính xác của các giá trị dự báo từ mô hình hồi quy có được nhờ mẫu khởi động. Việc tách mẫu phải đảm bảo: một mặt không làm thay đổi nhiều đến kết quả hồi quy dựa trên mẫu khởi động, mặt khác có đủ số quan sát cho mẫu kiểm tra để đánh giá được khả năng dự báo của mô hình.

6.3.3.2. Tiêu chuẩn đánh giá mức độ chính xác của dự báo

Giả sử mẫu kiểm tra có kích thước m. Kí hiệu:

Y_i: là giá trị quan sát của biến phụ thuộc Y

$\hat{Y}_i$: là giá trị dự báo điểm của mô hình hồi quy

$\hat{U}_i = Y_i - \hat{Y}_i$: là sai số của dự báo

Việc đánh giá khả năng dự báo của mô hình phải dựa trên các sai số dự báo trong mẫu kiểm tra chứ không phải trên mẫu khởi động (vì khi xây dựng mô hình ước lượng SRF, phương pháp OLS đã cực tiểu tổng bình phương các phần dư trong mẫu khởi động). Điều đáng quan tâm là kết quả ước lượng còn phù hợp với các quan sát ngoài mẫu khởi động hay không?

Người ta đưa ra các tiêu chuẩn đo lường thống kê sau đây:

• $\boldsymbol{ME} = \frac{1}{m}.\sum_{i=1}^{m} \widehat{U}_i$: Sai số trung bình (*mean error*): *Cho biết có hay không khuynh hướng dự báo thấp hơn* ($ME > 0$) *hay cao hơn* ($ME < 0$)*giá trị thực tế.*

• $\boldsymbol{MAE} = \frac{1}{m}.\sum_{i=1}^{m} |\widehat{U}_i|$: Sai số tuyệt đối trung bình (*mean absolute error*)

• $\boldsymbol{MSE} = \frac{1}{m}.\sum_{i=1}^{m} \widehat{U}_i^2$: Sai số bình phương trung bình (*mean squared error*): có tác dụng phóng đại các sai số dự báo có trị tuyệt đối lớn, do đó chú trọng tới các quan sát đặc biệt (vượt trội) trong mẫu, tuy nhiên nó không cùng đơn vị đo của các quan sát.

• $\boldsymbol{RMSE} = \sqrt{MSE}$: Căn bậc hai của sai số bình phương trung bình (*root mean squared error*): nó cùng đơn vị đo với các quan sát, nên có thể sử dụng $RMSE$ để thay thế cho MSE khi so sánh mức độ chính xác.

Nhận xét: Các chỉ số: $\boldsymbol{ME, MAE, MSE, RMSE}$ đều phụ thuộc vào đơn vị đo của biến quan sát, nên khi dùng chúng để đánh giá, so sánh mức độ chính xác của dự báo cần lưu ý điều này.

• $\boldsymbol{PE_i} = \left(\frac{Y_i - \hat{Y}_i}{Y_i}\right).100\%$: Sai số phần trăm (tương đối) PE (*percentage error*)

• $\mathbf{MPE} = \frac{1}{m}\sum_{i=1}^{m} PE_i$: Sai số phần trăm trung bình (*mean percentage error*)

• $\mathbf{MAPE} = \frac{1}{m}\sum_{i=1}^{m} |PE_i|$: Sai số phần trăm tuyệt đối trung bình (*mean absolute percentage error*)

• $\boldsymbol{TIC} = \frac{\sqrt{\sum_{i=1}^{m} \frac{(\hat{Y}_i - Y_i)^2}{m}}}{\sqrt{\frac{1}{m}\sum_{i=1}^{m} \hat{Y}_i^2} + \frac{1}{m}\sum_{i=1}^{m} Y_i^2}$: Hệ số bất đẳng thức *Theil* (*Theil Inequality Coefficient*), đánh giá mức sai lệch giữa giá trị dự báo điểm với giá trị thực tế quan sát. $0 \leq TIC \leq 1$;

TIC càng gần 0 thì dự báo điểm càng chính xác. Khi $\boldsymbol{TIC = 0}$ ***thì hàm hồi quy dự báo chính xác hoàn toàn.***

• ***Tỷ lệ chệch***: $\frac{(\bar{\hat{Y}} - \bar{Y})^2}{\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i - Y_i)^2}$ $(Bias\ Proportion)$: Đánh giá sự khác biệt của trung bình các giá trị dự báo so với trung bình các giá trị thực tế quan sát.

($\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$: giá trị trung bình của các quan sát thực tế trong mẫu kiểm tra;

$\bar{\hat{Y}} = \frac{1}{m}\sum_{i=1}^{m} \hat{Y}_i$: giá trị trung bình của các dự báo điểm trong mẫu kiểm tra)

• ***Tỷ lệ phương sai***: $\frac{(S_{\hat{Y}} - S_Y)^2}{\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i - Y_i)^2}$ $(Variance\ Proportion)$: Đánh giá sự khác biệt về mức độ biến thiên của các giá trị dự báo so với mức độ biến thiên của các giá trị thực tế quan sát.

($S_{\hat{Y}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2}$ là độ lệch chuẩn của các giá trị dự báo trong mẫu kiểm tra; $S_Y = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y})^2}$ là độ lệch chuẩn của các giá trị thực tế quan sát trong mẫu kiểm tra)

• ***Tỷ lệ hiệp phương sai:*** $\frac{2(1-R_{\hat{Y}Y}).S_{\hat{Y}}.S_Y}{\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i-Y_i)^2}$ (*Covariance Proportion*):

Đánh giá sự khác biệt về mức độ biến thiên của các giá trị dự báo so với mức độ biến thiên của các giá trị thực tế quan sát.

($R_{\hat{Y}Y} = \frac{\sum(\hat{Y}_i-\bar{\hat{Y}})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{m}(\hat{Y}_i-\bar{\hat{Y}})^2.\sum_{i=1}^{m}(Y_i-\bar{Y})^2}}$: là hệ số tương quan giữa các giá trị dự báo và các giá trị thực tế trong mẫu kiểm tra)

Chú ý:

1. Có thể chỉ ra hệ thức:

$$\frac{1}{m}\sum_{i=1}^{m}\left(\hat{Y}_i - Y_i\right)^2 = \left(\bar{\hat{Y}} - \bar{Y}\right)^2 + (S_{\hat{Y}} - S_Y)^2 + 2(1 - R_{\hat{Y}Y}).S_{\hat{Y}}.S_Y$$

Hệ thức trên cho thấy:

Tỷ lệ chệch + Tỷ lệ phương sai + Tỷ lệ hiệp phương sai = 1

Do đó nếu dự báo là tốt thì tỷ lệ chệch và tỷ lệ phương sai có xu hướng nhỏ, nên phần lớn sai số trong dự báo sẽ nằm trong tỷ lệ hiệp phương sai là phần đo lường thể hiện tính chất không có quy luật (không hệ thống).

2. Việc phân tích độc lập các giá trị của các chỉ số: ME, MAE, MSE, RMSE ít có ý nghĩa. Người ta thường dùng các chỉ số này để đối chiếu, so sánh giữa các mô hình hồi quy có cùng dạng biến phụ thuộc và cùng cỡ mẫu. Các chỉ số: PE, M$PE, TIC,$ Tỷ lệ chệch, Tỷ lệ phương sai, Tỷ lệ hiệp phương sai có thể được sử dụng để phân tích, đánh giá khả năng dự báo của một mô hình. Để lựa chọn được mô hình có khả năng dự báo tốt nhất, người ta có thể so sánh các tiêu chuẩn đo lường thống kê giữa các mô hình hồi quy.

3. Nếu mục đích dự báo là để kiểm tra khả năng dự báo của mô hình thì giá trị của các biến giải thích được sử dụng để dự báo được lấy từ trong mẫu kiểm tra. Ứng dụng của phân tích hồi quy là sử dụng mô hình hồi quy để dự báo cho biến phụ thuộc – dự báo ngoài phạm vi mẫu phân tích.

Sau khi dự báo trong mẫu nhằm đánh giá khả năng dự báo chính xác của mô hình, thì mô hình hồi quy có thể được sử dụng để dự báo ngoài mẫu. Đối với dự báo ngoài mẫu, người ta thường tiến hành dự báo khoảng. Sai số của dự báo sẽ càng nhỏ nếu giá trị của biến giải thích dùng để dự báo nằm trong khoảng biến thiên của mẫu và càng gần với giá trị trung bình mẫu.

Ví dụ 6.4: Số liệu về doanh thu Y, chi phí quảng cáo X1, tiền lương của nhân viên tiếp thị X2, của một số nhân viên được cho trong bảng dưới đây. Yêu cầu:

a. Chạy hồi quy SRF tuyến tính của Y theo X1, X2.

b. Dùng mô hình SRF nhận được ở a/ để tìm khoảng dự báo giá trị trung bình và giá trị cá

biệt của Y khi X1 = 20, X2 = 16, với độ tin cậy 95%.

Y	X1	X2	Y	X1	X2
127	18	10	161	25	14
140	25	11	128	16	12
106	19	6	139	17	12
163	24	16	144	23	12
102	15	7	159	22	14
180	26	17	138	15	15

a. Chạy hồi quy SRF tuyến tính của Y theo X1, X2.
b. Dùng mô hình SRF nhận được ở a/ để tìm khoảng dự báo giá trị trung bình và giá trị cá biệt của Y khi X1 = 20, X2 = 16, với độ tin cậy 95%.

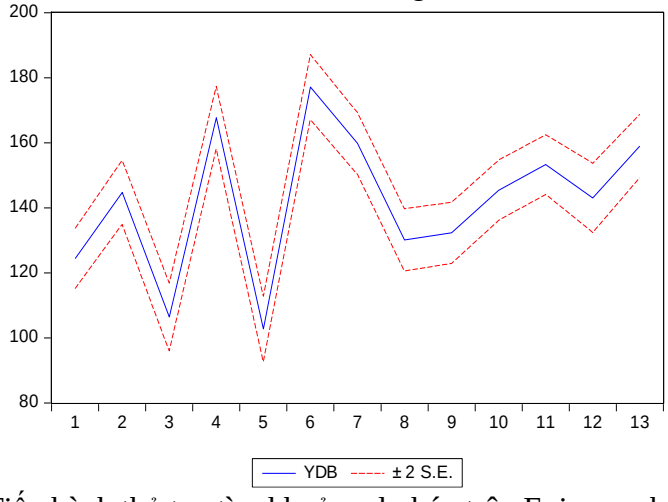
Giải:
a. Từ mẫu điều tra, ta nhận được kết quả hồi quy:

Dependent Variable: Y
Method: Least Squares
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	34.69682	6.811656	5.093742	0.0007
X1	2.185409	0.357924	6.105796	0.0002
X2	5.035705	0.447043	11.26448	0.0000

R-squared	0.970799	Mean dependent var	140.5833
Adjusted R-squared	0.964310	S.D. dependent var	23.08269

b. Nhập thêm số liệu X1 = 20, X2 = 16 vào bảng Group để dự báo, ta có:
* Đồ thị dự báo và các chỉ số đánh giá



Forecast: YDB
Actual: Y
Forecast sample: 1 13
Included observations: 12

Root Mean Squared Error	3.776520
Mean Absolute Error	3.183247
Mean Abs. Percent Error	2.203199
Theil Inequality Coefficient	0.013271
Bias Proportion	0.000000
Variance Proportion	0.007409
Covariance Proportion	0.992591

* Tiến hành thủ tục tìm khoảng dự báo trên Eviews, nhận được:

obs	CANDUOICB	CANDUOITB	CANTRENCB	CANTRENTB
1	113.9063	120.8385	134.8762	127.9439
2	133.5539	139.4829	155.8958	149.9667
3	94.65425	99.99596	118.2134	112.8717
4	156.8521	163.1623	178.5837	172.2735
5	91.32833	97.01495	114.1275	108.4408
6	165.7874	171.5372	188.4615	182.7117
7	149.0605	155.5063	170.6033	164.1575
8	119.2447	125.5810	140.9389	134.6026
9	121.6633	128.3602	142.8912	136.1942
11	142.8728	149.9730	163.6786	156.5784
12	130.9746	136.1126	155.0524	149.9144
13	147.9391	154.0258	170.0135	163.9267

Vậy khoảng dự báo giá trị trung bình và giá trị cá biệt của Y là:

CANDUOITB	CANTRENTB	CANDUOICABIET	CANTRENCABIET
154.0258	***163.9267***	**147.9391**	**170.0135**

Ví dụ 6.5: Sử dụng số liệu và kết quả hồi quy trong 1/,ví dụ 3.4 của chương 3, tiến hành dự báo khoảng cho giá trị trung bình, giá trị cá biệt về lượng hàng bán được khi giá bán là 7 ngàn đồng/kg, chi phí quảng cáo là 9 triệu đồng/tháng với độ tin cậy 95%, có kết quả được in đậm ở dòng cuối cùng của bảng:

obs	CANDUOICABIET	CANTRENCABIET	CANDUOITRUNGBINH	CANTRENTRUNGBINH
1	18.64852	21.45079	19.30068	20.79864
2	17.66743	20.28501	18.41878	19.53366
3	17.01769	19.56326	17.82381	18.75714
4	15.91724	18.39805	16.78833	17.52696
5	15.24851	17.69529	16.16453	16.77927
6	16.32315	18.82582	17.16996	17.97901
7	15.63494	18.14254	16.47666	17.30081
8	14.88055	17.36183	15.75106	16.49132
9	13.94606	16.41835	14.82746	15.53695
10	13.09244	15.60241	13.93176	14.76310
11	13.03281	15.49845	13.92264	14.60862
12	12.23338	14.71796	13.10006	13.85128
13	11.73798	14.20041	12.63201	13.30638
14	11.28785	13.81686	12.10877	12.99594
15	10.46554	13.01443	11.26882	12.21115
16	13.06036	15.51571	13.96403	14.61204
17	14.61572	17.14934	15.43241	16.33264
18	11.00577	13.51631	11.84450	12.67758
19	9.481321	12.07586	10.24877	11.30841
20	9.324743	12.12932	9.975893	11.47817
21	***11.73978***	***14.49885***	***12.41156***	***13.82707***

Bảng 3.11

Bài tập.

1. Số liệu về doanh thu Y (triệu đồng/tháng), chi phí quảng cáo X_1 (triệu đồng/tháng) và tiền lương X_2 (triệu đồng/tháng) của nhân viên tiếp thị của 12 công ty tư nhân được cho bởi bảng sau:

Y	X1	X2	Y	X1	X2	Y	X1	X2
158	18	11	130	15	9	158	16	12
176	23	14	195	24	17	173	21	14
142	18	10	180	23	15	188	22	15
180	22	16	167	15	11	157	14	13

a. Giả sử tương quan giữa Y với X_1 và X_2 có thể biểu diễn bởi hàm hồi quy tuyến tính. Hãy ước lượng hàm này.
b. Với mức ý nghĩa 5%, hãy kiểm định giả thuyết về hệ số hồi quy của X_1 và X_2 trong mô hình hồi quy tổng thể bằng 0. Cho biết ý nghĩa của kết quả kiểm định.
c. Hãy ước lượng các mô hình sau:
(1) $Y_i = a + a_1X_{1i} + U_i$;
(2) $Y_i = b + b_1X_{2i} + U_i$;
(3) $Y_i = \gamma + \gamma_1X_{1i} + \gamma_2X_{2i} + U_i$;
(4) $Y_i = \delta + \delta_1X_{1i} + \delta_2X_{1i}(-1) + \delta_3X_{2i} + U_i$;
d. Để dự báo lượng hàng bán được thì trong các mô hình ước lượng cho 4 mô hình trên, nên chọn mô hình nào? Vì sao?
e. Từ mô hình đã được lựa chọn ở d/ hãy dự báo doanh thu bình quân của một công ty tư nhân với độ tin cậy 95%.

2. Một nhà phân tích công nghiệp điện ảnh muốn ước lượng doanh thu (Y) của một bộ phim được phát hành đối với các biến là: chi phí sản xuất phim X1, chi phí quảng cáo X2 (các biến này cùng đơn vị tiền tệ) và biến Z xác định như sau: Z = 1, nếu bộ phim đã được giới thiệu ít nhất trên một tạp chí trước khi phát hành; Z = 0, nếu ngược lại.Nhà phân tích này đã thu được thông tin trên qua mẫu ngẫu nhiên gồm 20 bộ phim đã phát hành trong vòng 6 năm như sau:

Bộ phim	Y	X1	X2	Z	Bộ phim	Y	X1	X2	Z
1	27	4.0	1	0	11	48	10.7	1	1
2	36	6.1	3	1	12	82	11.0	15	1
3	51	5.5	6	1	13	25	3.5	4	0
4	20	3.3	1	0	14	50	6.9	10	0
5	75	12.5	11	1	15	58	7.8	9	1
6	62	9.7	8	1	16	62	10	10	0
7	15	2.5	5	0	17	30	5.0	1	1
8	45	10.8	5	0	18	37	7.5	5	0
9	50	8.4	3	1	19	45	6.4	7	1
10	35	6.6	2	0	20	73	10.0	12	1

a. Hãy ước lượng mô hình hồi quy: Y = a + b.X1 + c.X2 + e.Z + U qua số liệu nói trên.
b. Cho nhận xét từ các hệ số hồi quy ước lượng. Việc giới thiệu trên tạp chí có ý nghĩa đ/v doanh thu hay không? Hãy kiểm định điều này.
c. Chạy hồi quy ước lượng cho mô hình:

$$Y = \alpha_0 + \alpha_1 lnX1 + \alpha_2 lnX2 + \alpha_3 Z + V$$

c. Hãy tìm các đặc trưng thống kê và ma trận tương quan mẫu của véc tơ quan sát (Y, X1, X2).
d. Kiểm tra các vấn đề đa cộng tuyến, phương sai nhiễu thay đổi, tự tương quan và khắc phục nếu có trong các mô hinh SRF được thiết lập.
e. Trong hai mô hình trên, bạn chọn mô hình nào, tại sao?
h. Sử dụng mô hình đã chọn để dự báo khoảng cho giá trị cá biệt và giá trị trung bình của doanh thu của bộ phim, nếu chi phí sản xuất là 10, chi phí quảng cáo là 6 và được giới thiệu ít nhất trên một tạp chí trước khi phát hành.

3. Chi tiêu Y (tỷ đô, năm 1982) cho nhập khẩu hàng hóa và thu nhập cá nhân X (tỷ đô, năm 1982) sau thuế ở Mỹ từ 1970 -1987 như sau:

Năm	X	Y	Năm	X	Y	Năm	X	Y
1970	1668.1	150.9	1976	2001.0	229.3	1982	2261.5	249.5
1971	1728.4	166.2	1977	2066.6	259.4	1983	2331.9	282.2
1972	1797.4	190.7	1978	2167.4	274.1	1984	2469.8	351.1
1973	1916.3	218.2	1979	2112.6	277.9	1985	2542.8	367.9
1974	1896.6	211.8	1980	2214.3	253.6	1986	2640.9	412.3
1975	1931.7	187.9	1981	2248.6	258.7	1987	2686.3	439.0

a/ Định mẫu từ 1970 – 1983 (mẫu khởi động) chạy hồi quy để ước lượng cho mô hình: Y = a + b. X + U.
b/ Dùng SRF nhận được ở trên để dự báo điểm và dự báo khoảng với độ tin cậy 95% cho giá trị trung bình cà giá trị cá biệt của chi tiêu cho nhập khẩu hàng hóa ở Mỹ trong các năm 1984, 1985, 1986, 1987. Trình bày đồ thị dự báo cùng các chỉ số đánh giá tính chính xác của dự báo.

4. Với bảng số liệu của bài tập 3, có các kết quả hồi quy sau:
a/

$$\begin{cases} \hat{Y} = -284{,}1252 + 0{,}255801.X \\ R^2 = 0{,}925020;\ \bar{R}^2 = 0{,}920334; AIC = 9{,}200923; L = -80{,}80831; SC = 9{,}299853 \end{cases}$$

b/

$$\begin{cases} \hat{Y} = -807{,}4096 + 0{,}582688.X - 18{,}86305.T \\ R^2 = 0{,}960638;\ \bar{R}^2 = 0{,}955390; AIC = 8{,}667607; L = -75{,}00847; SC = 8{,}816003 \end{cases}$$

c/

$$\begin{cases} \hat{Y} = -3892{,}362 + 542{,}5807.lnX \\ R^2 = 0{,}900890;\ \bar{R}^2 = 0{,}894695; AIC = 9{,}479943; L = -83{,}31948; SC = 9{,}578873 \end{cases}$$

Hãy dùng các tiêu chuẩn : $R^2, \bar{R}^2$, Akaike, Schwarz, Log likelihood để phân tích và lựa

chọn mô hình thích hợp.

5. Với bảng số liệu của bài tập 3, có các kết quả hồi quy sau:
a/

$$\begin{cases} \widehat{lnY} = 3{,}510084 + 0{,}000945.X \\ R^2 = 0{,}942158;\ \bar{R}^2 = 0{,}938543; AIC = -2{,}279745; L = 22{,}51770; SC = -2{,}180814 \end{cases}$$

b/

$$\begin{cases} \widehat{lnY} = -9{,}993454 + 2{,}026996.lnX \\ R^2 = 0{,}934163;\ \bar{R}^2 = 0{,}935361; AIC = -2{,}229269; L = 22{,}06342; SC = -2{,}130339 \end{cases}$$

c/

$$\begin{cases} \widehat{lnY} = -26{,}12156 + 4{,}203836.lnX - 0{,}058280.T \\ R^2 = 0{,}958494;\ \bar{R}^2 = 0{,}952959; AIC = -2{,}500506; L = 25{,}50455; SC = -2{,}35211 \end{cases}$$

Hãy dùng các tiêu chuẩn : $R^2, \bar{R}^2$, Akaike, Schwarz, Log likelihood để phân tích và lựa chọn mô hình thích hợp.

6. Hãy dùng các tiêu chuẩn : $R^2, \bar{R}^2$, Akaike, Schwarz, Log likelihood để phân tích và lựa chọn mô hình thích hợp trong các bài tập 4, 5. Dựa vào mô hình đã lựa chọn, hãy dự báo chi tiêu cho nhập khẩu hàng hóa trong năm 1988, nếu năm này thu nhập cá nhân là 2670,5.

7. Tổng sản lượng Y (triệu USD), ngày lao động X (triệu ngày) và lượng vốn K(triệu USD) của khu vực nông nghiệp ở một quốc gia trong giai đoạn 1958-1972 có dữ liệu sau:

Năm	Y	X	K	Năm	Y	X	K
1958	16807.7	275.5	17803.7	1966	27403.0	307.5	24939.0
1959	17711.3	274.4	18096.8	1967	28628.7	303.7	26713.7
1960	20471.2	269.7	18271.8	1968	29904.5	304.7	29957.8
1961	21032.9	267.0	19167.3	1969	28508.2	298.6	31585.9
1962	21406.0	267.8	19647.6	1970	29305.5	295.5	33474.5
1963	22531.6	275.0	20803.5	1971	30821.5	299.0	34821.8
1964	24806.3	283.0	22076.6	1972	31535.8	288.1	41794.3
1965	26465.8	300.7	23445.2				

a/ Chạy hồi quy ước lượng cho mô hình: $Y = a + bX + c.K + U$
b/ Kiểm định xem mô hình SRF ở a/ có xảy ra vấn đề đa cộng tuyến, phương sai nhiễu thay đổi, tự tương quan của nhiễu hay không. Nếu có, hãy tìm cách khắc phục.
c/ Kiểm tra xem mô hình SRF ở a/ có bị bỏ sót biến hay thừa biến không.

8. Với bảng số liệu ở bài 7:
a/ Chạy hồi quy ước lượng cho mô hình: $lnY = a + blnX + c.lnK + V$
b/ Kiểm định xem mô hình SRF ở a/ có xảy ra vấn đề đa cộng tuyến, phương sai nhiễu thay đổi, tự tương quan của nhiễu hay không. Nếu có, hãy tìm cách khắc phục.
c/ Kiểm tra xem mô hình SRF ở a/ có bị bỏ sót biến hay thừa biến không.

9. Với các mô hình SRF ở phần a/ của bài tập 7, 8, bạn chọn mô hình nào? Với mô hình được chọn, hãy kiểm định xem mô hình này có xảy ra các vấn đề: Đa cộng tuyến, Phương sai nhiễu thay đổi, Tự tương quan của nhiễu. Nếu có, hãy tìm cách khắc phục.

10. Với bảng số liệu bài tập 7:
a/ Định mẫu 1958-1967 (mẫu khởi động), chạy hồi quy ước lượng cho mô hình:

$$Y = a + bX + c.K + U$$

b/ Dùng SRF ở phần a/ để dự báo điểm và dự báo khoảng với độ tin cậy 95% cho giá trị trung bình cà giá trị cá biệt của tổng sản lượng các năm 1968, 1969, 1970, 1971, 1972.
c/ Trình bày đồ thị dự báo cùng các chỉ số đánh giá tính chính xác của dự báo ở b/

11. Với bảng số liệu bài tập 7:
a/ Định mẫu 1958-1967 (mẫu khởi động), chạy hồi quy ước lượng cho mô hình:

$$lnY = a + b.lnX + c.lnK + U$$

b/ Dùng SRF ở phần a/ để dự báo điểm và dự báo khoảng với độ tin cậy 95% cho giá trị trung bình cà giá trị cá biệt của tổng sản lượng các năm 1968, 1969, 1970, 1971, 1972.
c/ Trình bày đồ thị dự báo cùng các chỉ số đánh giá tính chính xác của dự báo ở b/

12. Có các giá trị quan sát về thu nhập Y (USD/người), tỷ lệ lao động nông nghiệp X (%) và Z là số năm trung bình được đào tạo đối với những người trên 25 tuổi như sau:

Y	X	Z	Y	X	Z	Y	X	Z
7	9	8	12	4	14	10	7	11
8	10	12	9	5	9	11	4	14
9	8	10	8	5	10	9	9	12
8	7	9	9	6	12	10	5	10
10	10	11	10	8	13	11	8	11

a/ Chạy hồi quy ước lượng các mô hình sau:
(1) $\hat{Y} = a_0 + a_1.X + a_2.Z$
(1) $\hat{Y} = b_0 + b_1.X$
(1) $\hat{Y} = \alpha_0 + \alpha_1.Z$
b/ Để dự báo, nên chọn mô hình nào trong 3 mô hình trên, tại sao?
c/ Từ mô hình SRF đã được chọn ở b/, hãy tìm các dự báo điểm và vẽ đồ thị Line Graph để so sánh các giá trị dự báo điểm với giá trị thực tế và nêu nhận xét.
d/ Dùng mô hình SRF ở c/ để dự báo khoảng cho giá trị trung bình và giá trị cá biệt của thu nhập với độ tin cậy 95% khi tỷ lệ lao động nông nghiệp là 11 và số năm trung bình được đào tạo là 13 năm.

Chương phụ lục.
MỘT SỐ VẤN ĐỀ CẦN THIẾT TRONG LÝ THUYẾT XÁC SUẤT VÀ THỐNG KÊ TOÁN

Chương này nhắc lại và nhấn mạnh tới một số khái niệm và các bài toán cơ bản của Lý thuyết xác suất và thống kê toán sẽ được sử dụng trong Kinh tế lượng như: Các khái niệm về ước lượng và bài toán ước lượng tham ẩn; các khái niệm về kiểm định giả thuyết thống kê và bài toán kiểm định giả thuyết thống kê; khái niệm về kỳ vọng có điều kiện và hàm hồi quy.

Vấn đề 1. Bài toán ước lượng tham ẩn

1. Đặt vấn đề

Giả sử θ là một giá trị chân thực(một hằng số nào đó) mà ta chưa biết, nhưng cần biết mà lại không thể biết được chính xác. Khi đó ta phải tìm một đại lượng $\hat{\theta}$ để xấp xỉ cho θ, ta nói $\hat{\theta}$ là một ước lượng cho θ. Vậy tìm ước lượng $\hat{\theta}$ như thế nào? Rõ ràng yêu cầu trong mọi trường hợp là ước lượng phải phù hợp với vai trò và ý nghĩa của θ cũng như những thông tin có được về θ.

Nói chung θ gắn liền với một biến quan sát nào đó mà nó đóng vai trò là một tham số của phân phối xác suất. Vì thế θ được gọi là tham ẩn.

2. Phương pháp chung giải quyết vấn đề

Trước hết ta cần phải có thông tin về θ, do đó ta cần chỉ ra biến quan sát, ta ký hiệu là X, có liên quan đến θ mà trong đó θ đóng vai trò là tham số của phân phối xác suất. Việc lấy thông tin ở đây có nghĩa là lập mẫu ngẫu nhiên điều tra về biến quan sát X, những thông tin này cũng là những thông tin về θ. Trên mẫu ta có đại lượng lượng $\hat{\theta}$ tương ứng với θ (theo nghĩa: Vai trò của $\hat{\theta}$ trên mẫu tương tự như vai trò của θ đối với biến quan sát X). Ta dùng đại lượng $\hat{\theta}$ để ước lượng cho θ.

Chú ý: Ước lượng $\hat{\theta}$ phụ thuộc vào mẫu: $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ nên nó là biến ngẫu nhiên phụ thuộc cỡ mẫu n.

3. Ước lượng không chệch và ước lượng vững

Người ta thường quan tâm tới những tính chất tốt sau đây, nếu có, của một ước lượng:

a/ Tính không chệch:

Ước lượng $\hat{\theta}$ được gọi là ước lượng không chệch cho θ nếu: $E\hat{\theta} = \theta$

- Tính không chệch của ước lượng $\hat{\theta}$ được hiểu là: mặc dù $\hat{\theta}$ thay đổi giá trị tùy theo mẫu, nhưng trung bình của các giá trị này vẫn là θ.

b/ Tính vững:

Ước lượng $\hat{\theta}$ được gọi là ước lượng vững cho θ nếu:

$$\forall \varepsilon > 0, \lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1$$

(Như vậy ước lượng $\hat{\theta}$ hội tụ theo xác suất về θ: $\hat{\theta} \xrightarrow{P} \theta$ $(n \to \infty)$)

- Tính vững của ước lượng $\hat{\theta}$ có nghĩa là khi cỡ mẫu n càng lớn thì hầu hết các giá trị của ước lượng càng gần với θ, tức là ước lượng càng chính xác.
- Từ các tính chất của các đặc trưng mẫu ta suy ra:

* Trung bình mẫu $\bar{X}$ là ước lượng vững và không chệch cho trung bình tổng thể: $m = EX$.
* Phương sai mẫu $S^2(X)$ là ước lượng vững và chệch cho phương sai tổng thể : $\sigma^2 = VarX$
* Tần suất mẫu $f = f(A)$ là ước lượng vững và không chệch cho tần suất tổng quát hay xác suất $p = P(A)$.

Chú ý: Như đã chỉ ra ở trên, phương sai mẫu $S^2(X)$ tuy là ước lượng vững cho phương sai tổng thể, nhưng lại là ước lượng chệch, vì: $ES^2(X) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$.

Tuy nhiên, nếu xét đại lượng:

$$S'^2(X) = \frac{n}{n-1}S^2(X) = \frac{n}{n-1}\left\{\overline{X^2} - (\bar{X})^2\right\} = \frac{1}{n-1}\sum_{j=1}^{n} X_j^2 - \frac{n}{n-1}(\bar{X})^2$$

thì ta có $S'^2(X)$ là một ước lượng vững và không chệch cho phương sai tổng thể $\sigma^2 = VarX$.

Người ta gọi $S'^2(X) = \frac{n}{n-1}S^2(X)$ là *phương sai mẫu điều chỉnh* của biến quan sát X.

* Lưu ý, có một số tài liệu về xác suất thống kê lại định nghĩa phương sai mẫu của biến quan sát X là $S^2(X) = \frac{n}{n-1}\left\{\overline{X^2} - (\bar{X})^2\right\}$, tuy nhiên đại lượng này lại không có tính chất tương tự như phương sai tổng thể.

4. Một số phương pháp ước lượng

Chúng ta khảo sát ở đây hai phương pháp ước lượng tham ẩn. Đó là phương pháp ước lượng điểm và phương pháp ước lượng khoảng tin cậy.

a/ Phương pháp ước lượng điểm

Giả sử trên mẫu ta thu được ước lượng $\hat{\theta}$ về θ. Khi đó với một mẫu cụ thể thì ước lượng $\hat{\theta}$ nhận một giá trị xác định là $\hat{\theta} \coloneqq \hat{\theta}_0$. Ta dùng giá trị $\hat{\theta}_0$ để ước lượng cho θ và khi đó giá trị $\hat{\theta}_0$ được gọi là một ước lượng điểm cho θ.

Ví dụ2: Mức thu nhập X(triệu đồng/tháng) của các hộ gia đình ở địa phương A qua điều tra 250 hộ cho thấy mức thu nhập bình quân của các hộ này là 1,8 triệu đồng. Ta cần ước lượng cho mức thu nhập bình quân mỗi hộ là: $\theta = EX$. Trên mẫu ngẫu nhiên ta dùng đại lượng tương ứng là trung bình mẫu: $\hat{\theta} = \bar{X}$ để ước lượng. Với mẫu cụ thể 250 hộ nói trên, ta có: $\hat{\theta} = \bar{X} \coloneqq \hat{\theta}_0 = 1{,}8$, giá trị này là một ước lượng điểm cho mức thu nhập bình quân của một hộ trong toàn địa phương A.

b/ Phương pháp ước lượng khoảng tin cậy

Phương pháp này nhằm chỉ ra một khoảng ngẫu nhiên (khoảng có các đầu mút là các biến ngẫu nhiên): $(\hat{\theta}',\ \hat{\theta}'')$ mà giá trị θ có thể rơi vào với xác suất γ đủ lớn, tức là:

$P(\hat{\theta}' \leq \theta \leq \hat{\theta}'') = \gamma$ (đủ lớn)

Khi đó: Khoảng: $(\hat{\theta}',\ \hat{\theta}'')$ gọi là khoảng tin cậy cho θ; γ được gọi là độ tin cậy của ước lượng; $\varepsilon = \frac{1}{2}(\hat{\theta}'' - \hat{\theta}')$ được gọi là độ chính xác của ước lượng.

• Khoảng tin cậy đối xứng là khoảng tin cậy có dạng: $(\hat{\theta} - \varepsilon;\ \hat{\theta} + \varepsilon)$.

• Khoảng tin cậy một phía là khoảng tin cậy có dạng:$(-\infty; \hat{\theta})$ $hoặc$: $(\hat{\theta};\ +\infty)$

Vấn đề 2. Các khái niệm về kiểm định giả thuyết thống kê

1. Đặt vấn đề

Đứng trước một vấn đề hay một hiện tượng nào đó, dựa vào những thông tin nhất định, người ta đưa ra hai nhận định khác nhau về vấn đề hay hiện tượng nói trên. Yêu cầu đặt ra là chúng ta phải đánh giá lựa chọn xem nhận định nào là phù hợp hơn, xác đáng hơn.
- Các nhận định được nêu ra được gọi là các giả thuyết thống kê
- Việc đánh giá lựa chọn nói trên gọi là kiểm định giả thuyết thống kê.

2. Giải quyết vấn đề

- Cơ sở để giải quyết vấn đề này là lập luận thường gặp sau đây trong thống kê mà người ta thường gọi là nguyên lý biến cố hiếm, phát biểu như sau:

Giả sử A là một biến cố hiếm trong phép thử (tức là biến cố có xác suất rất bé- thông thường mức xác suất không quá 5% được coi là rất bé). Khi đó nếu ta chỉ tiến hành một lần thử thôi để quan sát biến cố A thì nói chung ta sẽ thấy A không xuất hiện.

- Trên cơ sở nói đó, chúng ta phải thu thập thông tin về vấn đề hay hiện tượng đang xét, muốn vậy ta đưa ra biến quan sát hoặc tính chất cần quan sát liên quan đến vấn đề, hiện tượng đang xét sao cho những thông tin về nó cũng là những thông tin về vấn đề hay hiện tượng này.
- Chọn một trong hai nhận định đưa ra, gọi nó là giả thuyết – ký hiệu là H_0 (*Null Hypothesis*); nhận định còn lại được gọi là đối thuyết- ký hiệu là H_1 (còn gọi là giả thuyết thay thế: *Altenative Hypothesis*). Nếu ta chấp nhận giả thuyết H_0 thì có nghĩa là bác bỏ đối thuyết H_1 và ngược lại: nếu ta bác bỏ giả thuyết H_0 thì có nghĩa là chấp nhận đối thuyết H_1.
- Với mẫu ngẫu nhiên quan sát, trên cơ sở giả thuyết H_0 đúng, ta xây dựng một biên cố W có xác suất rất bé, tức là: $P(W|H_0) \leq \alpha$ (rất bé).

(Như vậy nếu H_0 đúng thì W là một biến cố hiếm)

- Lập mẫu cụ thể và quan sát biến cố W:

* Nếu W xảy ra thì ta bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1
* Nếu W không xảy ra thì ta tạm chấp nhận giả thuyết H_0 (hay nói: chưa có cơ sở để bác bỏ H_0) và bác bỏ H_1 (hay nói: chưa có cơ sở để chấp nhận H_1)

Giải thích: Như trên đã lưu ý: Nếu H_0 đúng thì W là một biến cố hiếm. Khi đó việc lập một mẫu cụ thể thực chất là tiến hành một lần thử để quan sát W.

- Nếu W xảy ra thì điều này mâu thuẫn với nguyên lý biến cố hiếm, mâu thuẫn này là do ta giả thiết H_0 đúng, vậy phải bác bỏ H_0.
- Nếu W không xảy ra thì có nghĩa là biến cố hiếm mới thử một lần chưa thấy xảy ra, điều này phù hợp với nguyên lý biến cố hiếm. Vậy giả thiết H_0 đúng là phù hợp, hay ta chấp nhận H_0.

• W được gọi là miền bác bỏ giả thuyết H_0 hay miền tiêu chuẩn.
• α được gọi là mức ý nghĩa của việc kiểm định.

3. Sai lầm loại 1 và sai lầm loại 2

Ta biết rằng một biến cố có xác suất rất bé, kể cả bằng 0 vẫn có thể xảy ra trong phép thử. Vì thế khi ta dựa vào nguyên lý biến cố hiếm để bác bỏ (hay chấp nhận) một nhận định thì không có nghĩa là nhận định này sai (hay đúng) hoàn toàn. Khi ta bác bỏ H_0 nhưng H_0 vẫn có thể xảy ra, khi ta chấp nhận H_0 thì H_0 vẫn có thể không xảy ra. Tức là khi ta kiểm định, không tránh khỏi những sai lầm nhất định.

- Sai lầm loại 1: Bác bỏ giả thuyết H_0 trong khi thực tế H_0 đúng.

$P(W|H_0)$ là xác suất sai lầm loại 1.

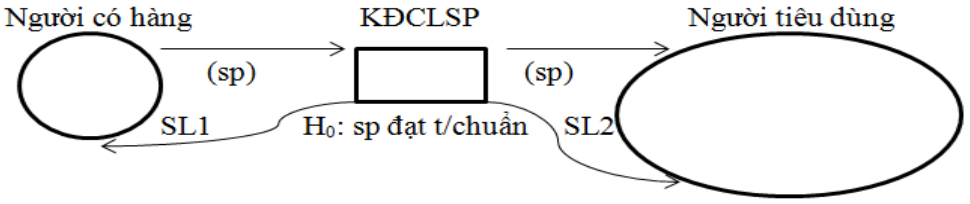
- Sai lầm loại 2: Chấp nhận giả thuyết H_0 trong khi thực tế H_0 sai.

$P(\overline{W}|H_1)$ là xác suất sai lầm loại 2.

Như vậy mức ý nghĩa α chính là mức khống chế xác suất sai lầm loại 1.

Một quy tắc kiểm định là lý tưởng nếu nó cực tiểu hóa được đồng thời xác suất sai lầm loại 1 và xác suất sai lầm loại 2. Tuy nhiên điều này là không thể và trong một chừng mực nào đó nó tựa như việc ấn xuống đồng thời cả hai đầu một cái bập bênh. Cả hai loại

sai lầm đều có hậu quả không hay và thực tế thì sai lầm loại 2 thường nghiêm trọng hơn. Chúng ta có thể thấy rõ điều này qua mô hình kiểm tra chất lượng sản phẩm như sau:



Sai lầm loại 1 ở đây là trả lại một sản phẩm đạt tiêu chuẩn cho người có hàng.

Sai lầm loại 2 ở đây là chuyển đến người tiêu dùng một sản phẩm không đạt tiêu chuẩn.

Rõ ràng ở đây sai lầm loại 2 nghiêm trọng hơn cả về quy mô lẫn chiều sâu. Hơn nữa bộ phận KĐCLSP cũng không thể đồng thời đảm bảo tối đa quyền lợi của cả hai phía: người có hàng và người tiêu dùng.

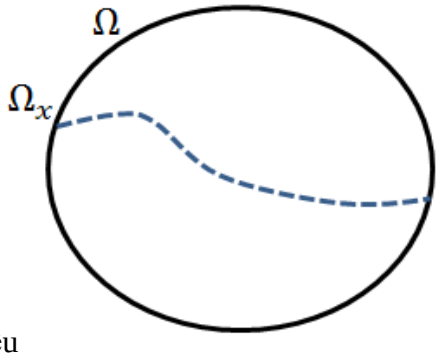
Vì thế người ta xử lý vấn đề này bằng cách khống chế xác suất sai lầm loại 1 ở mức α đủ bé (mức ý nghĩa), trên cơ sở đó tìm miền tiêu chuẩn W cực tiểu hóa xác suất sai lầm loại 2, tức là giải quyết bài toán tối ưu sau đây:

$$\begin{cases} P(W|H_0) \leq \alpha \\ P(\overline{W}|H_1) \rightarrow min \end{cases}$$

Vấn đề 3. Hàm hồi quy

a. Các khái niệm chung

Giả sử trên mỗi cá thể của tập hợp Ω, chúng ta quan sát hai tiêu chuẩn về số lượng là X, Y, hay nói cách khác, ta xét véc tơ ngẫu nhiên hai chiều: (X, Y). Khi đó trong tập hợp Ω_x gồm các cá thể cùng tiêu chuẩn thứ nhất $X = x$, ta quan sát tiêu chuẩn thứ hai thì giá trị của tiêu chuẩn thứ hai (nói chung lúc này phụ thuộc vào x) ta ký hiệu là Y_x, cũng là một biến ngẫu nhiên.



* Ta gọi phân phối xác suất của biến Y_x là phân phối có điều kiện của biến ngẫu nhiên Y với điều kiện $X = x$.

* Kỳ vọng hay giá trị trung bình của biến ngẫu nhiên Y_x là $EY_x = f(x)$ được gọi là kỳ vọng có điều kiện (hay trung bình có điều kiện) của biến Y, với điều kiện $X = x$ và được ký hiệu là:

$E(Y|X = x)$: $E(Y|X = x) = EY_x = f(x)$.

Khi đó biến ngẫu nhiên $f(X)$ được gọi là kỳ vọng có điều kiện (hay trung bình có điều kiện) của biến Y với điều kiện biến X và được ký hiệu là: $E(Y|X)$

* Hàm biến thực: $f(x) = E(Y|X = x), x \in \mathbb{R}$ được gọi là hàm hồi quy của Y theo X, có ý nghĩa quan trọng trong các bài toán thống kê để nghiên cứu sự phụ thuộc giữa các biến

quan sát. Hoàn toàn tương tự, ta có khái niệm về kỳ vọng có điều kiện của một biến ngẫu nhiên với điều kiện một véc tơ ngẫu nhiên và hàm hồi quy nhiều chiều, bằng cách thay biến ngẫu nhiên X bởi véc tơ ngẫu nhiên X.

Ví dụ 1: Điều tra chiều cao X(cm) và cân nặng Y(kg) của thanh niên ở một khu vực dân cư A, ta có véc tơ ngẫu nhiên hai chiều: (X, Y). Khi đó Y_x là cân nặng của thanh niên có chiều cao x , phân phối có điều kiện của Y với điều kiện $X = x$ là phân phối xác suất của biến Y_x , đó là phân phối về cân nặng của những thanh niên có cùng chiều cao x. Kỳ vọng có điều kiện $E(Y|X = x)$ chính là cân nặng trung bình của những thanh niên có cùng chiều cao x.

Ví dụ 2: Véc tơ ngẫu nhiên rời rạc hai chiều (X, Y) có bảng phân phối xác suất:

Y / X	-1	0	3	$p_{i\bullet}$
1	0,1	0,3	0,2	0,6
2	0,1	0,2	0,1	0,4
$p_{\bullet j}$	0,2	0,5	0,3	1

a/ Tìm bảng phân phối xác suất có điều kiện của Y với điều kiện X = 2, X = 2.

b/ Tính kỳ vọng có điều kiện $E(Y|X = 1) = EY_1;\ E(Y|X = 2) = EY_2$.

c/ Tìm $E(Y|X)$

a/ Phân phối có điều kiện của Y với điều kiện X = 1 là phân phối xác suất của biến Y_1, như vậy:

$$P(Y_1 = -1) = P(Y = -1|X = 1) = \frac{P(X = 1, Y = -1)}{P(X = 1)} = \frac{0,1}{0,6} = \frac{1}{6}$$

$$P(Y_1 = 0) = P(Y = 0|X = 1) = \frac{P(X = 1, Y = 0)}{P(X = 1)} = \frac{0,3}{0,6} = \frac{3}{6}$$

$$P(Y_1 = 3) = P(Y = 3|X = 1) = \frac{P(X = 1, Y = 3)}{P(X = 1)} = \frac{0,2}{0,6} = \frac{2}{6}$$

Do đó phân phối có điều kiện của Y với điều kiện X = 2 là bảng sau:

Y_1	- 1	0	3	∑
P	1/6	3/6	2/6	1

Phân phối có điều kiện của Y với điều kiện X = 2 là phân phối xác suất của biến Y_2 , như vậy:

$$P(Y_2 = -1) = P(Y = -1|X = 2) = \frac{P(X = 2, Y = -1)}{P(X = 2)} = \frac{0,1}{0,4} = \frac{1}{4}$$

$$P(Y_2 = 0) = P(Y = 0|X = 2) = \frac{P(X = 2, Y = 0)}{P(X = 2)} = \frac{0,2}{0,4} = \frac{2}{4}$$

$$P(Y_2 = 3) = P(Y = 3|X = 2) = \frac{P(X = 2, Y = 3)}{P(X = 2)} = \frac{0,1}{0,4} = \frac{1}{4}$$

Do đó phân phối có điều kiện của Y với điều kiện X = 2 là bảng sau:

Y_2	- 1	0	3	Σ
P	1/4	2/4	1/4	1

b/ $$E(Y|X=1) = EY_1 = -1.\frac{1}{6} + 0.\frac{3}{6} + 3.\frac{2}{6} = \frac{5}{6};$$

$$E(Y|X=2) = EY_2 = -1.\frac{1}{4} + 0.\frac{2}{4} + 3.\frac{1}{4} = 0{,}5$$

c/ $$E(Y|X) = \begin{cases} \frac{5}{6}, nếu\ X = 1 \\ 0{,}5, nếu\ X = 2 \end{cases} = f(X)$$

Đây là biến ngẫu nhiên rời rạc có bảng phân phối xác suất là:

$E(Y\|X)$	$1/2$	$5/6$	Σ
P	0,4	0,6	1

****

PHỤ LỤC: CÁC BẢNG THỐNG KÊ

Bảng phụ lục I: Bảng giá trị tới hạn của phân phối t (student):

$$P\big(t > t_\alpha(k)\big) = \alpha \; hay: P\big(|t| > t_\alpha(k)\big) = 2\alpha$$

α / k	0.001	0.002	0.005	0.01	0.02	0.025	0.05	0.1	0.2
1	318.309	159.153	63.657	31.821	15.895	12.706	6.314	3.078	1.376
2	22.327	15.764	9.925	6.965	4.849	4.303	2.920	1.886	1.061
3	10.215	8.053	5.841	4.541	3.482	3.182	2.353	1.638	0.978
4	7.173	5.951	4.604	3.474	2.999	2.776	2.132	1.533	0.941
5	5.893	5.030	4.032	3.365	2.757	2.571	2.015	1.476	0.920
6	5.208	4.524	3.707	3.143	2.612	2.447	1.943	1.440	0.906
7	4.785	4.207	3.499	2.998	2.517	2.365	1.895	1.415	0.896
8	4.501	3.991	3.355	2.896	2.449	2.306	1.860	1.397	0.889
9	4.297	3.835	3.250	2.821	2.398	2.262	1.833	1.383	0.883
10	4.144	3.716	3.169	2.764	2.359	2.228	1.812	1.372	0.879
11	4.025	3.624	3.106	2.718	2.328	2.201	1.796	1.363	0.876
12	3.930	3.550	3.055	2.681	2.303	2.179	1.782	1.356	0.873
13	3.852	3.489	3.012	2.650	2.282	2.160	1.771	1.350	0.870
14	3.787	3.438	2.977	2.624	2.264	2.145	1.761	1.345	0.868
15	3.733	3.395	2.947	2.602	2.249	2.131	1.753	1.341	0.866
16	3.686	3.358	2.921	2.583	2.235	2.120	1.746	1.337	0.865
17	3.646	3.326	2.898	2.567	2.224	2.110	1.740	1.333	0.863
18	3.610	3.298	2.878	2.552	2.214	2.101	1.734	1.330	0.862
19	3.579	3.273	2.861	2.539	2.205	2.093	1.729	1.328	0.861
20	3.552	3.251	2.845	2.528	2.197	2.086	1.725	1.325	0.860
21	3.527	3.231	2.831	2.518	2.189	2.080	1.721	1.323	0.859
22	3.505	3.214	2.819	2.508	2.183	2.074	1.717	1.321	0.858
23	3.485	3.198	2.807	2.500	2.177	2.069	1.714	1.319	0.858
24	3.467	3.183	2.797	2.492	2.172	2.064	1.711	1.318	0.857
25	3.450	3.170	2.787	2.485	2.167	2.060	1.708	1.316	0.856
26	3.435	3.158	2.779	2.479	2.162	2.056	1.706	1.315	0.856
27	3.421	3.147	2.771	2.473	2.158	2.052	1.703	1.314	0.855
28	3.408	3.136	2.763	2.467	2.154	2.048	1.701	1.313	0.855
29	3.396	3.127	2.756	2.462	2.150	2.045	1.699	1.311	0.854
30	3.385	3.118	2.750	2.457	2.147	2.042	1.697	1.310	0.854
31	3.375	3.109	2.744	2.453	2.144	2.040	1.696	1.309	0.853
32	3.365	3.102	2.738	2.449	2.141	2.037	1.694	1.309	0.853
33	3.356	3.094	2.733	2.445	2.138	2.035	1.692	1.308	0.853
34	3.348	3.088	2.728	2.441	2.136	2.032	1.691	1.307	0.852
35	3.340	3.081	2.724	2.438	2.133	2.030	1.690	1.306	0.852
36	3.333	3.075	2.719	2.434	2.131	2.028	1.688	1.306	0.852
37	3.326	3.070	2.715	2.431	2.129	2.026	1.687	1.305	0.851
38	3.319	3.064	2.712	2.429	2.127	2.024	1.686	1.304	0.851
39	3.313	3.059	2.708	2.426	2.125	2.023	1.685	1.304	0.851
40	3.307	3.055	2.704	2.423	2.123	2.021	1.684	1.303	0.851
41	3.301	3.050	2.701	2.421	2.121	2.020	1.683	1.303	0.850
42	3.296	3.046	2.698	2.418	2.120	2.018	1.682	1.302	0.850
43	3.291	3.042	2.695	2.416	2.118	2.017	1.681	1.302	0.850
44	3.286	3.038	2.692	2.414	2.116	2.015	1.680	1.301	0.850
45	3.281	3.034	2.690	2.412	2.115	2.014	1.679	1.301	0.850
46	3.277	3.030	2.687	2.410	2.114	2.013	1.679	1.300	0.850
47	3.273	3.027	2.685	2.408	2.112	2.012	1.678	1.300	0.849
48	3.269	3.024	2.682	2.407	2.211	2.011	1.677	1.299	0.849
49	3.265	3.021	2.680	2.405	2.110	2.010	1.677	1.299	0.849

Bảng giá trị tới hạn $t_\alpha(k)$ của phân phối Student (tiếp theo)

α / k	0.001	0.002	0.005	0.01	0.02	0.025	0.05	0.1	0.2
50	3.261	3.018	2.678	2.403	2.109	2.009	1.676	1.299	0.849
51	3.258	3.015	2.676	2.402	2.108	2.008	1.675	1.298	0.849
52	3.255	3.012	2.674	2.400	2.107	2.007	1.675	1.298	0.849
53	3.251	3.009	2.672	2.399	2.106	2.006	1.674	1.298	0.848
54	3.248	3.007	2.670	2.397	2.105	2.005	1.674	1.297	0.848
55	3.245	3.005	2.668	2.396	2.104	2.004	1.673	1.297	0.848
56	3.242	3.002	2.667	2.395	2.103	2.003	1.673	1.297	0.848
57	3.239	3.000	2.665	2.394	2.102	2.002	1.672	1.297	0.848
58	3.237	2.998	2.663	2.392	2.101	2.002	1.672	1.296	0.848
59	3.234	2.996	2.662	2.391	2.100	2.001	1.671	1.296	0.848
60	3.232	2.994	2.660	2.390	2.099	2.000	1.671	1.296	0.848
61	3.229	2.992	2.659	2.389	2.099	2.000	1.670	1.296	0.848

Bảng phụ lục II: Bảng giá trị tới hạn $F_\alpha(k,m)$ của phân phối F

$P(F(k,m) > F_\alpha(k,m)) = \alpha$, với $\alpha = 0.01$

m / k	1	2	3	4	5	6	7	8
2	98,503	99,000	99,166	99,249	99,299	99,333	99,356	99,374
3	34,116	30,817	29,457	28,710	28,237	27,911	27,672	27,489
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289
6	13,745	10,925	9,780	9,148	8,746	8466	8,260	8,102
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029
9	10,561	8,022	6,992	6,422	6,057	5,802	5,613	5,467
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890
17	8,400	6,112	5,185	4,669	4,336	4,102	3,927	3,791
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841	3,705
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564
21	8,017	5,780	4,874	4,369	4,042	3,812	3,640	3,506
22	7,945	5,719	4,817	4,313	3,988	3,758	3,587	3,453
23	7,881	5,664	4,765	4,264	3,939	3,710	3,539	3,406
24	7,823	5,614	4,718	4,218	3,895	3,667	3,496	3,363
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457	3,324
26	7,721	5,526	4,637	4,140	3,818	3,591	3,421	3,288
27	7,677	5,488	4,601	4,106	3,785	3,558	3,388	3,256
28	7,636	5,453	4,568	4,074	3,754	3,528	3,358	3,226
29	7,598	5,420	4,538	4,045	3,725	3,499	3,330	3,198
30	7,562	5,390	4,510	4,018	3,699	3,473	3,304	3,173

Bảng giá trị tới hạn $F_\alpha(k,m)$ của phân phối F, với $\alpha = 0.01$ (tiếp theo)

m \ k	9	10	11	12	13	14	15	16
2	99,388	99,399	99,408	99,416	99,422	99,428	99,433	99,437
3	27,345	27,229	27,133	27,052	26,983	26,924	26,872	26,827
4	14,659	14,546	14,452	14,374	14,307	14,249	14,198	14,154
5	10,158	10,051	9,963	9,888	9,825	9,770	9,722	9,680
6	7,976	7,874	7,790	7,718	7,657	7,605	7,559	7,519
7	6,719	6,620	6,538	6,469	6,410	6,359	6,314	6,275
8	5,911	5,814	5,734	5,667	5,609	5,559	5,515	5,477
9	5,351	5,257	5,178	5,111	5,055	5,005	4,962	4,924
10	4,942	4,849	4,772	4,706	4,650	4,601	4,558	4,520
11	4,632	4,539	4,462	4,397	4,342	4,293	4,251	4,213
12	4,388	4,296	4,220	4,155	4,100	4,052	4,010	3,972
13	4,191	4,100	4,025	3,960	3,905	3,857	3,815	3,778
14	4,030	3,939	3,864	3,800	3,745	3,698	3,656	3,619
15	3,895	3,805	3,730	3,666	3,612	3,564	3,522	3,485
16	3,780	3,691	3,616	3,553	3,498	3,451	3,409	3,372
17	3,682	3,593	3,519	3,455	3,401	3,353	3,312	3,275
18	3,597	3,508	3,434	3,371	3,316	3,269	3,227	3,190
19	3,523	3,434	3,360	3,297	3,242	3,195	3,153	3,116
20	3,457	3,368	3,294	3,231	3,177	3,130	3,088	3,051
21	3,398	3,310	3,236	3,173	3,119	3,072	3,030	2,993
22	3,346	3,258	3,184	3,121	3,067	3,019	2,978	2,941
23	3,299	3,211	3,137	3,074	3,020	2,973	2,931	2,894
24	3,256	3,168	3,094	3,032	2,977	2,930	2,889	2,852
25	3,217	3,129	3,056	2,993	2,939	2,892	2,850	2,813
26	*3,182*	*3,094*	*3,021*	*2,958*	*2,904*	*2,857*	*2,815*	*2,778*
27	*3,149*	*3,062*	*2,988*	*2,926*	*2,871*	*2,824*	*2,783*	*2,746*
28	*3,120*	*3,032*	*2,959*	*2,896*	*2,842*	*2,795*	*2,753*	*2,716*
29	*3,092*	*3,005*	*2,931*	*2,868*	*2,814*	*2,767*	*2,726*	*2,689*
30	*3,067*	*2,979*	*2,906*	*2,843*	*2,789*	*2,742*	*2,700*	*2,663*
31	*3,043*	*2,955*	*2,882*	*2,820*	*2,765*	*2,718*	*2,677*	*2,640*
32	*3,021*	*2,934*	*2,860*	*2,798*	*2,744*	*2,696*	*2,655*	*2,618*
33	*3,000*	*2,913*	*2,840*	*2,777*	*2,723*	*2,676*	*2,634*	*2,597*

Bảng giá trị tới hạn $F_\alpha(k,m)$ của phân phối F, với $\alpha = 0.05$

m \ k	1	2	3	4	5	6	7	8
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447
21	4,325	3,467	3,072	2,840	2,685	2,573	2,488	2,420
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464	2,397
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442	2,375
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388	2,321
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373	2,305
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359	2,291
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346	2,278
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266

Bảng giá trị tới hạn $F_\alpha(k,m)$ của phân phối F, với $\alpha = 0.05$ (tiếp theo)

m \ k	9	10	11	12	13	14	15	16
2	19,385	19,396	19,405	19,413	19,419	19,424	19,429	19,433
3	8,812	8,786	8,763	8,745	8,729	8,715	8,703	8,692
4	5,999	5,964	5,936	5,912	5,891	5,873	5,858	5,844
5	4,772	4,735	4,704	4,678	4,655	4,636	4,619	4,604
6	4,099	4,060	4,027	4,000	3,976	3,956	3,938	3,922
7	3,677	3,637	3,603	3,575	3,550	3,529	3,511	3,494
8	3,388	3,347	3,313	3,284	3,259	3,237	3,218	3,202
9	3,179	3,137	3,102	3,073	3,048	3,025	3,006	2,989
10	3,020	2,978	2,943	2,913	2,887	2,865	2,845	2,828
11	2,896	2,854	2,818	2,788	2,761	2,739	2,719	2,701
12	2,796	2,753	2,717	2,687	2,660	2,637	2,617	2,599
13	2,714	2,671	2,635	2,604	2,577	2,554	2,533	2,515
14	2,646	2,602	2,565	2,534	2,507	2,484	2,463	2,445
15	2,588	2,544	2,507	2,475	2,448	2,424	2,403	2,385
16	2,538	2,494	2,456	2,425	2,397	2,373	2,352	2,333
17	2,494	2,450	2,413	2,381	2,353	2,329	2,308	2,289
18	2,456	2,412	2,374	2,342	2,314	2,290	2,269	2,250
19	2,423	2,378	2,340	2,308	2,280	2,256	2,234	2,215
20	2,393	2,348	2,310	2,278	2,250	2,225	2,203	2,184
21	2,366	2,321	2,283	2,250	2,222	2,197	2,176	2,156
22	2,342	2,297	2,259	2,226	2,198	2,173	2,151	2,131
23	2,320	2,275	2,236	2,204	2,175	2,150	2,128	2,109
24	2,300	2,255	2,216	2,183	2,155	2,130	2,108	2,088
25	2,282	2,236	2,198	2,165	2,136	2,111	2,089	2,069
26	2,265	2,220	2,181	2,148	2,119	2,094	2,072	2,052
27	2,250	2,204	2,166	2,132	2,103	2,078	2,056	2,036
28	2,236	2,190	2,151	2,118	2,089	2,064	2,041	2,021
29	2,223	2,177	2,138	2,104	2,075	2,050	2,027	2,007
30	2,211	2,165	2,126	2,092	2,063	2,037	2,015	1,995
31	2,199	2,153	2,114	2,080	2,051	2,026	2,003	1,983
32	2,189	2,142	2,103	2,070	2,040	2,015	1,992	1,972
33	2,179	2,133	2,093	2,060	2,030	2,004	1,982	1,961

Bảng phụ lục III: Giá trị tới hạn $\chi^2_\alpha(k)$ của phân phối Chi – square:

$$P\left(\chi^2(k) > \chi^2_\alpha(k)\right) = \alpha$$

k \ α	**0,01**	**0,99**	**0,025**	**0,975**	**0,05**	**0,95**
1	6.6349	0.0002	5.0239	0.0010	3.8415	0.0039
2	9.2103	0.0201	7.3778	0.0506	5.9915	0.1026
3	11.3449	0.1148	9.3484	0.2158	7.8147	0.3518
4	13.2767	0.2971	11.1433	0.4844	9.4877	0.7107
5	15.0863	0.5543	12.8325	0.8312	11.0705	1.1455
6	16.8119	0.8721	14.4494	1.2373	12.5916	1.6354
7	18.4753	1.2390	16.0128	1.6899	14.0671	2.1673
8	20.0902	1.6465	17.5345	2.1797	15.5073	2.7326
9	21.6660	2.0879	19.0228	2.7004	16.9190	3.3251
10	23.2093	2.5582	20.4832	3.2470	18.3070	3.9403
11	24.7250	3.0535	21.9200	3.8157	19.6751	4.5748
12	26.2170	3.5706	23.3367	4.4038	21.0261	5.2260
13	27.6882	4.1069	24.7356	5.0088	22.3620	5.8919
14	29.1412	4.6604	26.1189	5.6287	23.6848	6.5706
15	30.5779	5.2293	27.4884	6.2621	24.9958	7.2609
16	31.9999	5.8122	28.8454	6.9077	26.2962	7.9616
17	33.4087	6.4078	30.1910	7.5642	27.5871	8.6718
18	34.8053	7.0149	31.5264	8.2307	28.8693	9.3905
19	36.1909	7.6327	32.8523	8.9065	30.1435	10.1170
20	37.5662	8.2604	34.1696	9.5908	31.4104	10.8508
21	38.9322	8.8972	35.4789	10.2829	32.6706	11.5913
22	40.2894	9.5425	36.7807	10.9823	33.9244	12.3380
23	41.6384	10.1957	38.0756	11.6886	35.1725	13.0905
24	42.9798	10.8564	39.3641	12.4012	36.4150	13.8484
25	44.3141	11.5240	40.6465	13.1197	37.6525	14.6114
26	45.6417	12.1981	41.9232	13.8439	38.8851	15.3792
27	46.9629	12.8785	43.1945	14.5734	40.1133	16.1514
28	48.2782	13.5647	44.4608	15.3079	41.3371	16.9279

Giá trị tới hạn $\chi^2_\alpha(k)$ của phân phối Chi – square (tiếp theo)

k \ α	0.01	0.99	0.025	0.975	0.05	0.95
29	49.5879	14.2565	45.7223	16.0471	42.5570	17.7084
30	50.8922	14.9535	46.9792	16.7908	43.7730	18.4927
31	52.1914	15.6555	48.2319	17.5387	44.9853	19.2806
32	53.4858	16.3622	49.4804	18.2908	46.1943	20.0719
33	54.7755	17.0735	50.7251	19.0467	47.3999	20.8665
34	56.0609	17.7891	51.9660	19.8063	48.6024	21.6643
35	57.3421	18.5089	53.2033	20.5694	49.8018	22.4650
36	58.6192	19.2327	54.4373	21.3359	50.9985	23.2686
37	59.8925	19.9602	55.6680	22.1056	52.1923	24.0749
38	61.1621	20.6914	56.8955	22.8785	53.3835	24.8839
39	62.4281	21.4262	58.1201	23.6543	54.5722	25.6954
40	63.6907	22.1643	59.3417	24.4330	55.7585	26.5093
41	64.9501	22.9056	60.5606	25.2145	56.9424	27.3256
42	66.2062	23.6501	61.7768	25.9987	58.1240	28.1440
43	67.4593	24.3976	62.9904	26.7854	59.3035	28.9647
44	68.7095	25.1480	64.2015	27.5746	60.4809	29.7875
45	69.9568	25.9013	65.4102	28.3662	61.6562	30.6123
46	71.2014	26.6572	66.6165	29.1601	62.8296	31.4390
47	72.4433	27.4158	67.8206	29.9562	64.0011	32.2676
48	73.6826	28.1770	69.0226	30.7545	65.1708	33.0981
49	74.9195	28.9406	70.2224	31.5549	66.3386	33.9303
50	76.1539	29.7067	71.4202	32.3574	67.5048	34.7643
51	77.3860	30.4750	72.6160	33.1618	68.6693	35.5999
52	78.6158	31.2457	73.8099	33.9681	69.8322	36.4371
53	79.8433	32.0185	75.0019	34.7763	70.9935	37.2759
54	81.0688	32.7934	76.1920	35.5863	72.1532	38.1162
55	82.2921	33.5705	77.3805	36.3981	73.3115	38.9580
56	83.5134	34.3495	78.5672	37.2116	74.4683	39.0813
57	84.7328	35.1305	79.7522	38.0267	75.6237	40.6459
58	85.9502	35.9135	80.9356	38.8435	76.7778	41.4920
59	87.1657	36.6982	82.1174	39.6619	77.9305	42.3393
60	88.3794	37.4849	83.2977	40.4817	79.0819	43.1880

Bảng phụ lục IV: Bảng thống kê d (*Durbin – Watson*) với $\alpha = 0.05$

(n: cỡ mẫu, k': số biến giải thích)

	k' = 1		**k' = 2**		**k' = 3**		**k' = 4**		**k' = 5**	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.610	1,400								
7	0.700	1,356	0.467	1,896						
8	0.763	1,332	0.559	1,777	0.368	2,287				
9	0.824	1,320	0.629	1,699	0.455	2,128	0.296	2,588		
10	0.879	1,320	0.697	1,641	0.525	2,016	0.376	2,414	0.243	2,822
11	0.927	1,324	0.658	1,604	0.595	1,928	0.444	2,283	0.316	2,645
12	0.971	1,331	0.812	1,579	0.658	1,864	0.512	2,177	0.379	2,506
13	1,010	1,340	0.861	1,562	0.715	1,816	0.574	2,094	0.445	2,390
14	1,045	1,350	0.905	1,551	0.767	1,779	0.632	2,030	0.505	2,296
15	1,077	1,361	0.946	1,543	0.814	1,750	0.685	1,977	0.562	2,220
16	1,106	1,371	0.982	1,539	0.857	1,728	0.734	1,935	0.615	2,157
17	1,133	1,381	1,015	1,536	0.897	1,710	0.779	1,900	0.664	2,104
18	1,158	1,391	1,046	1,535	0.933	1,696	0.820	1,872	0.710	2,060
19	1,180	1,401	1,074	1,536	0.967	1,685	0.859	1,848	0.752	2,023
20	1,201	1,411	1,100	1,537	0.998	1,676	0.894	1,828	0.792	1,991
21	1,221	1,420	1,125	1,535	1,026	1,669	0.927	1,812	0.829	1,964
22	1,239	1,429	1,147	1,541	1,053	1,664	0.958	1,797	0.863	1,940
23	1,257	1,437	1,168	1,543	1,078	1,660	0.986	1,785	0.895	1,920
24	1,273	1,446	1,188	1,546	1,101	1,656	1,013	1,775	0.925	1,902
25	1,288	1,454	1,206	1,550	1,123	1,654	1,038	1,767	0.953	1,886
26	1,302	1,461	1,224	1,553	1,143	1,652	1,062	1,759	0.979	1,873
27	1,316	1,469	1,240	1,556	1,162	1,651	1,084	1,753	1,004	1,861
28	1,328	1,476	1,255	1,560	1,181	1,650	1,104	1,747	1,028	1,850
29	1,341	1,483	1,270	1,563	1,198	1,650	1,124	1,743	1,053	1,841
30	1,352	1,489	1,284	1,567	1,214	1,650	1,143	1,749	1,071	1,833

Bảng thống kê d (*Durbin – Watson*) với $\alpha = 0.05$ (tiếp theo)

	k' = 1		**k' = 2**		**k' = 3**		**k' = 4**		**k' = 5**	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
31	1,363	1,496	1,297	1,570	1,229	1,650	1,160	1,735	1,090	1,825
32	1,373	1,502	1,309	1,574	1,244	1,650	1,177	1,732	1,109	1,819
33	1,383	1,508	1,321	1,577	1,258	1,651	1,193	1,730	1,127	1,813
34	1,393	1,514	1,333	1,580	1,271	1,652	1,208	1,728	1,144	1,808
35	1,402	1,519	1,343	1,584	1,283	1,653	1,222	1,726	1,160	1,803
36	1,411	1,525	1,354	1,587	1,295	1,654	1,236	1,724	1,175	1,799
37	1,419	1,530	1,364	1,590	1,307	1,655	1,249	1,723	1,190	1,795
38	1,427	1,535	1,373	1,594	1,318	1,656	1,261	1,722	1,204	1,792
39	1,435	1,540	1,382	1,597	1,328	1,658	1,273	1,722	1,218	1,789
40	1,442	1,544	1,391	1,600	1,338	1,659	1,285	1,721	1,230	1,786
45	1,475	1,556	1,430	1,615	1,383	1,666	1,336	1,720	1,287	1,776
50	1,503	1,585	1,462	1,628	1,421	1,674	1,378	1,721	1,335	1,771
55	1,528	1,601	1,490	1,641	1,452	1,681	1,414	1,724	1,374	1,768
60	1,549	1,616	1,514	1,652	1,480	1,689	1,444	1,727	1,408	1,767
65	1,567	1,629	1,536	1,662	1,503	1,696	1,471	1,731	1,438	1,767
70	1,583	1,641	1,554	1,672	1,525	1,703	1,494	1,735	1,464	1,768
75	1,598	1,652	1,571	1,680	1,543	1,709	1,515	1,739	1,487	1,770
80	1,611	1,662	1,586	1,688	1,560	1,715	1,534	1,743	1,507	1,772
85	1,624	1,671	1,600	1,696	1,575	1,721	1,550	1,747	1,525	1,774
90	1,635	1,679	1,612	1,703	1,589	1,726	1,566	1,751	1,542	1,776
95	1,645	1,687	1,623	1,709	1,602	1,732	1,579	1,755	1,557	1,778
100	1,654	1,694	1,634	1,715	1,613	1,736	1,592	1,758	1,571	1,780
150	1,720	1,746	1,706	1,760	1,693	1,774	1,679	1,788	1,665	1,802
200	1,758	1,778	1,748	1,789	1,738	1,799	1,728	1,810	1,718	1,820

Bảng thống kê d (*Durbin – Watson*) với $\alpha = 0.05$ (tiếp theo)

	k'= 6		k'= 7		k'= 8		k'= 9		k'= 10	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6										
7										
8										
9										
10										
11	0.203	3,005								
12	0.268	2,832	0.171	3,149						
13	0.328	2,692	0.230	2,985	0.147	3,226				
14	0.389	2,572	0.286	2,848	0.200	3,111	0.127	3,360		
15	0.447	2,472	0.343	2,727	0.251	2,979	0.175	3,216	0.111	3,438
16	0.502	2,388	0.398	3,624	0.304	2,860	0.222	3,090	0.155	3,304
17	0.554	2,318	0.451	2,537	0.356	2,757	0.272	2,975	0.198	3,184
18	0.603	2,257	0.502	2,461	0.407	2,667	0.321	2,873	0.244	3,073
19	0.649	2,206	0.549	2,396	0.456	2,589	0.369	2,783	0.290	3,974
20	0.692	2,162	0.595	2,339	0.502	2,521	0.416	2,704	0.336	2,885
21	0.732	2,124	0.637	2,290	0.547	2,460	0.461	2,633	0.380	2,806
22	0.769	2,090	0.677	2,246	0.588	2,407	0.504	2,571	0.424	2,734
23	0.804	2,061	0.715	2,208	0.628	2,360	0.545	2,514	0.465	2,670
24	0.837	2,035	0.751	2,174	0.666	2,318	0.584	2,464	0.506	2,613
25	0.868	2,012	0.784	2,144	0.702	2,280	0.621	2,419	0.544	2,560
26	0.897	1,992	0.816	2,117	0.735	2,246	0.657	2,379	0.581	2,513
27	0.925	1,974	0.845	2,093	0.767	2,216	0.691	2,342	0.616	2,470
28	0.951	1,958	0.874	2,071	0.798	2,188	0.723	2,309	0.650	2,431
29	0.975	1,944	0.900	2,055	0.826	2,164	0.753	2,278	0.682	2,396
30	0.998	1,931	0.926	2,034	0.854	2,141	0.782	2,251	0.712	2,363

Bảng thống kê d (*Durbin – Watson*) với $\alpha = 0.05$ (tiếp theo)

	k'= 6		**k'= 7**		**k'= 8**		**k'= 9**		**k'= 10**	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
31	1,020	1,920	0.950	2,018	0.879	2,120	0.810	2,226	0.741	2,333
32	1,041	1,909	0.972	2,004	0.904	2,122	0.836	2,203	0.769	2,306
33	1,061	1,900	0.994	1,991	0.927	2,085	0.861	2,181	0.795	2,281
34	1,080	1,891	1,015	1,979	0.950	2,069	0.885	2,162	0.821	2,257
35	1,097	1,884	1,034	1,967	0.971	2,054	0.908	2,144	0.845	2,236
36	1,114	1,877	1,053	1,957	0.991	2,041	0.930	2,127	0.868	2,216
37	1,131	1,870	1,071	1,948	1,011	2,029	0.951	2,112	0.891	2,198
38	1,146	1,864	1,088	1,939	1,029	2,017	0.970	2,098	0.912	2,180
39	1,161	1,859	1,104	1,932	1,047	2,007	0.990	2,085	0.932	2,164
40	1,175	1,854	1,120	1,924	1,064	1,997	1,008	2,072	0.952	2,149
45	1,238	1,835	1,189	1,895	1,139	1,958	1,089	2,022	1,038	2,088
50	1,291	1,822	1,247	1,875	1,201	1,930	1,156	1,986	1,110	2,044
55	1,334	1,814	1,294	1,861	1,253	1,909	1,212	1,959	1,170	2,010
60	1,372	1,808	1,335	1,850	1,298	1,894	1,260	1,939	1,222	1,984
65	1,404	1,805	1,370	1,843	1,336	1,882	1,305	1,923	1,266	1,964
70	1,433	1,802	1,401	1,837	1,369	1,883	1,337	1,910	1,305	1,948
75	1,458	1,801	1,428	1,834	1,399	1,867	1,369	1,901	1,339	1,935
80	1,480	1,801	1,453	1,831	1,425	1,861	1,397	1,893	1,369	1,925
85	1,500	1,801	1,474	1,829	1,448	1,857	1,422	1,886	1,396	1,916
90	1,518	1,801	1,494	1,827	1,469	1,854	1,445	1,881	1,420	1,909
95	1,535	1,802	1,512	1,827	1,489	1,852	1,465	1,877	1,442	1,903
100	1,550	1,803	1,528	1,826	1,506	1,850	1,484	1,874	1,462	1,898
150	1,651	1,817	1,637	1,832	1,622	1,847	1,608	1,862	1,594	1,877
200	1,707	1,831	1,697	1,841	1,686	1,852	1,675	1,863	1,665	1,874

Bảng thống kê d (*Durbin – Watson*) với $\alpha = 0.05$ (tiếp theo)

	k'= 11		**k'= 12**		**k'= 13**		**k'= 14**		**k'= 15**	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
16	0.098	3,503								
17	0.138	3,378	0.087	3,557						
18	0.177	3,265	0.123	3,441	0.078	3,603				
19	0.220	3,159	0.160	3,335	0.111	3,496	0.070	3,642		
20	0.263	3,063	0.200	3,234	0.145	3,395	0.100	3,542	0.063	3,676
21	0.307	2,976	0.240	3,141	0.182	3,300	0.132	3,448	0.091	3,583
22	0.349	2,897	0.281	3,057	0.220	3,211	0.166	3,358	0.120	3,495
23	0.391	2,826	0.322	2,979	0.259	3,128	0.202	3,272	0.153	3,409
24	0.431	2,761	0.362	2,908	0.297	3,053	0.239	3,193	0.186	3,327
25	0.470	2,702	0.400	2,844	0.335	2,983	0.275	3,119	0.221	3,251
26	0.508	2,649	0.438	2,784	0.373	2,919	0.312	3,051	0.256	3,179
27	0.544	2,600	0.475	2,730	0.409	2,859	0.348	2,987	0.291	3,112
28	0.578	2,555	0.510	2,680	0.445	2,805	0.388	2,928	0.325	3,050
29	0.612	2,515	0.544	2,634	0.479	2,755	0.418	2,874	0.359	2,922
30	0.643	2,477	0.577	2,592	0.512	2,708	0.451	2,823	0.392	2,937
31	0.674	2,443	0.608	2,553	0.545	2,665	0.484	2,776	0.425	2,887
32	0.703	2,411	0.638	2,517	0.576	2,625	0.515	2,733	0.457	2,840
33	0.731	2,382	0.668	2,484	0.606	2,588	0.546	2,692	0.488	2,976
34	0.758	2,355	0.695	2,454	0.634	2,554	0.575	2,654	0.518	2,754
35	0.783	2,330	0.722	2,425	0.662	2,521	0.604	2,619	0.547	2,716
36	0.808	2,306	0.748	2,398	0.689	2,492	0.631	2,586	0.575	2,680
37	0.831	2,285	0.772	2,374	0.714	2,464	0.657	2,555	0.602	2,646
38	0.854	2,265	0.796	2,351	0.739	2,438	0.683	2,526	0.628	2,614
39	0.875	2,246	0.819	2,329	0.763	2,413	0.707	2,499	0.653	2,585
40	0.896	2,228	0.840	2,309	0.785	2,391	0.731	2,473	0.678	2,557
45	0.988	2,156	0.938	2,225	0.887	2,296	0.838	2,367	0.788	2,439
50	1,064	2,103	1,019	2,163	0.973	2,225	0.927	2,287	0.882	2,350
55	1,129	2,062	1,087	2,116	1,045	2,170	1,003	2,225	0.961	2,281
60	1,184	2,031	1,145	2,079	1,106	2,127	1,068	2,177	1,029	2,227
65	1,231	2,006	1,195	2,049	1,160	2,093	1,124	2,138	1,088	2,183
70	1,272	1,986	1,239	2,026	1,206	2,066	1,172	2,106	1,139	2,148
75	1,308	1,970	1,277	2,006	1,247	2,043	1,215	2,080	1,184	2,148
80	1,340	1,957	1,311	1,991	1,283	2,024	1,253	2,059	1,224	2,093
85	1,369	1,946	1,342	1,977	1,315	2,009	1,287	2,040	1,260	2,073
90	1,395	1,937	1,369	1,966	1,344	1,995	1,318	2,025	1,292	2,055
95	1,418	1,929	1,394	1,956	1,370	1,984	1,345	2,012	1,321	2,040
100	1,439	1,923	1,416	1,948	1,393	1,974	1,371	2,000	1,347	2,026
150	1,579	1,892	1,564	1,908	1,550	1,924	1,535	1,940	1,519	1,956
200	1,654	1,885	1,643	1,896	1,632	1,908	1,621	1,919	1,610	1,931

Bảng thống kê d (*Durbin – Watson*) với $\alpha = 0.05$ (tiếp theo)

	k'= 16		**k'= 17**		**k'= 18**		**k'= 19**		**k'= 20**	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
16										
17										
18										
19										
20										
21	0.058	3,705								
22	0.083	3,619	0.052	3,731						
23	0.110	3,535	0.076	3,650	0.048	3,753				
24	0.141	3,454	0.101	3,572	0.070	3,678	0.044	3,773		
25	0.172	3,376	0.130	3,494	0.094	3,604	0.065	3,702	0.041	3,790
26	0.205	3,303	0.160	3,420	0.120	3,531	0.087	3,632	0.060	3,724
27	0.238	3,233	0.191	3,349	0.149	3,460	0.112	3,653	0.081	3,658
28	0.271	3,168	0.222	3,283	0.178	3,392	0.138	3,495	0.104	3,592
29	0.305	3,107	0.254	3,219	0.208	3,327	0.166	3,431	0.129	3,528
30	0.337	3,050	0.286	3,160	0.238	3,266	0.195	3,368	0.156	3,465
31	0.370	2,996	0.317	3,103	0.269	3,208	0.224	3,309	0.183	3,406
32	0.401	2,946	0.349	3,050	0.299	3,153	0.253	3,252	0.211	3,348
33	0.432	2,899	0.379	3,000	0.329	3,100	0.283	3,198	0.239	3,293
34	0.462	2,854	0.409	2,954	0.359	3,051	0.312	3,147	0.267	3,240
35	0.492	2,813	0.439	2,910	0.388	3,005	0.340	3,099	0.295	3,190
36	0.520	2,774	0.467	2,868	0.417	2,961	0.369	3,053	0.323	3,142
37	0.548	2,738	0.495	2,829	0.445	2,920	0.397	3,009	0.351	3,097
38	0.675	2,703	0.522	2,792	0.472	2,880	0.424	2,968	0.378	3,054
39	0.600	2,671	0.549	2,757	0.499	2,843	0.451	2,929	0.404	3,013
40	0.626	2,641	0.575	2,724	0.525	2,808	0.477	2,892	0.430	2,974
45	0.740	2,512	0.692	2,586	0.644	2,659	0.598	2,733	0.553	2,807
50	0.836	2,414	0.792	2,479	0.747	2,544	0.703	2,610	0.660	2,675
55	0.919	2,338	0.877	2,396	0.836	2,454	0.795	2,512	0.754	2,571
60	0.990	2,278	0.951	2,330	0.913	2,382	0.874	2,434	0.836	2,487
65	1,052	2,229	1,016	2,276	0.980	2,323	0.944	2,371	0.908	2,419
70	1,105	2,189	1,072	2,232	1,038	2,275	1,005	2,318	0.971	2,362
75	1,153	2,156	1,121	2,195	1,090	2,235	1,058	2,275	1,027	2,315
80	1,195	2,129	1,165	2,165	1,136	2,201	1,106	2,238	1,076	2,275
85	1,232	2,105	1,205	2,139	1,177	2,172	1,149	2,206	1,121	2,241
90	1,266	2,085	1,240	2,116	1,213	2,148	1,187	2,179	1,160	2,211
95	1,296	2,068	1,271	2,097	1,247	2,126	1,222	2,156	1,197	2,186
100	1,324	2,053	1,301	2,080	1,277	2,108	1,253	2,135	1,229	2,164
150	1,504	1,972	1,489	1,989	1,474	2,006	1,458	2,023	1,443	2,040
200	1,599	1,943	1,588	1,955	1,576	1,967	1,565	1,979	1,554	1,991

TÀI LIỆU THAM KHẢO

[1]. Phạm Chí Cao, Vũ Minh Châu. *Kinh tế lượng ứng dụng*. NXB Thống kê TP. Hồ Chí Minh – 2009

[2]. Nguyễn Quang Đông, *Bài giảng Kinh tế lượng,* NXB Thống kê – 2003

[3]. Nguyễn Khắc Minh. *Các phương pháp phân tích và dự báo trong kinh tế,* NXB. Khoa học kỹ thuật- 2002.

[4]. Đào Hữu Hồ, Nguyễn Văn Hữu, Hoàng Hữu Như, *Thống kê toán học*, NXB Đại học Quốc gia Hà Nội, 2004

[5]. Damodar N. Gujarati, *Basic Econometrics*, Mc Graw-Hill Inc, Third Ed. 1995.

[6]. Amemiya, Takeshi. *Introduction to Statistics and Econometrics*, Harvard University Press, 1994.

[7]. William H. Greene, *Econometric Analysis*, MacMilan Publshing Company, New York, 1991.

[8]. Ramu Ramanathan, *Introductory Econometrics with Applications*, The Dryden Press-Harcourt Brace College Publisher, 1978.

MỤC LỤC