BỘ MÔN KINH TẾ

BÀI GIẢNG

KINH TẾ LƯỢNG

MỤC LỤC **Trang**

CHƯƠNG 1 GIỚI THIỆU

1.1. Kinh tế lượng là gì?

Thuật ngữ tiếng Anh "Econometrics" có nghĩa là đo lường kinh tế[1]. Thật ra phạm vi của kinh tế lượng rộng hơn đo lường kinh tế. Chúng ta sẽ thấy điều đó qua một định nghĩa về kinh tế lượng như sau:

"Không giống như thống kê kinh tế có nội dung chính là số liệu thống kê, kinh tế lượng là một môn độc lập với sự kết hợp của lý thuyết kinh tế, công cụ toán học và phương pháp luận thống kê. Nói rộng hơn, kinh tế lượng liên quan đến: (1) Ước lượng các quan hệ kinh tế, (2) Kiểm chứng lý thuyết kinh tế bằng dữ liệu thực tế và kiểm định giả thiết của kinh tế học về hành vi, và (3) Dự báo hành vi của biến số kinh tế."[2]

Sau đây là một số ví dụ về ứng dụng kinh tế lượng.

Ước lượng quan hệ kinh tế

(1) Đo lường mức độ tác động của việc hạ lãi suất lên tăng trưởng kinh tế.

(2) Ước lượng nhu cầu của một mặt hàng cụ thể, ví dụ nhu cầu xe hơi tại thị trường Việt Nam.

(3) Phân tích tác động của quảng cáo và khuyến mãi lên doanh số của một công ty.

Kiểm định giả thiết

(1) Kiểm định giả thiết về tác động của chương trình khuyến nông làm tăng năng suất lúa.

(2) Kiểm chứng nhận định độ co dãn theo giá của cầu về cá basa dạng fillet ở thị trường nội địa.

(3) Có sự phân biệt đối xử về mức lương giữa nam và nữ hay không?

Dự báo

(1) Doanh nghiệp dự báo doanh thu, chi phí sản xuất, lợi nhuận, nhu cầu tồn kho…

(2) Chính phủ dự báo mức thâm hụt ngân sách, thâm hụt thương mại, lạm phát…

(3) Dự báo chỉ số VN Index hoặc giá một loại cổ phiếu cụ thể như REE.

---

1. A.Koutsoyiannis, Theory of Econometrics-Second Edition, ELBS with Macmillan-1996, trang 3

2. Ramu Ramanathan, Introductory Econometrics with Applications, Harcourt College Publishers-2002, trang 2.

1.2. Phương pháp luận của kinh tế lượng

Theo phương pháp luận truyền thống, còn gọi là phương pháp luận cổ điển, một nghiên cứu sử dụng kinh tế lượng bao gồm các bước như sau[3]:

(1) Phát biểu lý thuyết hoặc giả thiết.

(2) Xác định đặc trưng của mô hình toán kinh tế cho lý thuyết hoặc giả thiết.

(3) Xác định đặc trưng của mô hình kinh tế lượng cho lý thuyết hoặc giả thiết.

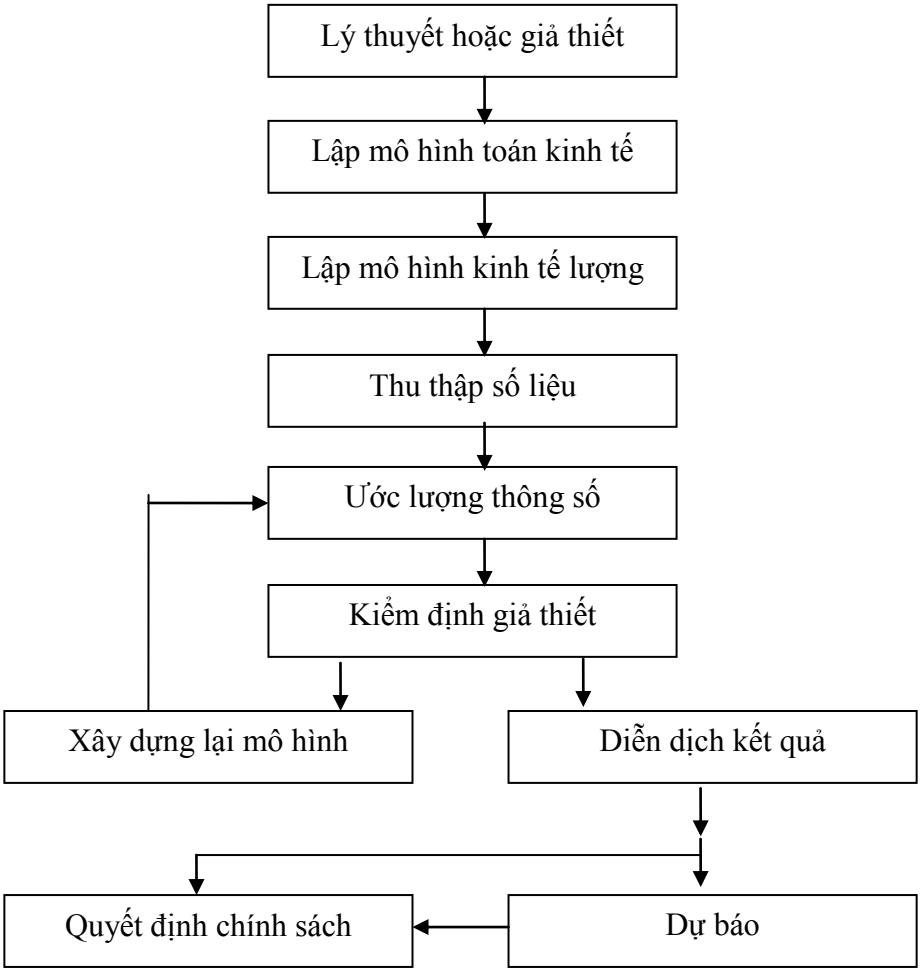
(4) Thu thập dữ liệu.

(5) Ước lượng tham số của mô hình kinh tế lượng.

(6) Kiểm định giả thiết.

(7) Diễn giải kết quả

(8) Dự báo và sử dụng mô hình để quyết định chính sách



Hình 1.1 Phương pháp luận của kinh tế lượng

[3] Theo Ramu Ramanathan, Introductory Econometrics with Applications, Harcourt College Publishers-2002

Ví dụ 1: Các bước tiến hành nghiên cứu một vấn đề kinh tế sử dụng kinh tế lượng với đề tài nghiên cứu xu hướng tiêu dùng biên của nền kinh tế Việt Nam.

(1) Phát biểu lý thuyết hoặc giả thiết

Keynes cho rằng:

Qui luật tâm lý cơ sở ... là đàn ông (đàn bà) muốn, như một qui tắc và về trung bình, tăng tiêu dùng của họ khi thu nhập của họ tăng lên, nhưng không nhiều như là gia tăng trong thu nhập của họ.[4]

Vậy Keynes cho rằng xu hướng tiêu dùng biên(marginal propensity to consume-MPC), tức tiêu dùng tăng lên khi thu nhập tăng 1 đơn vị tiền tệ lớn hơn 0 nhưng nhỏ hơn 1.

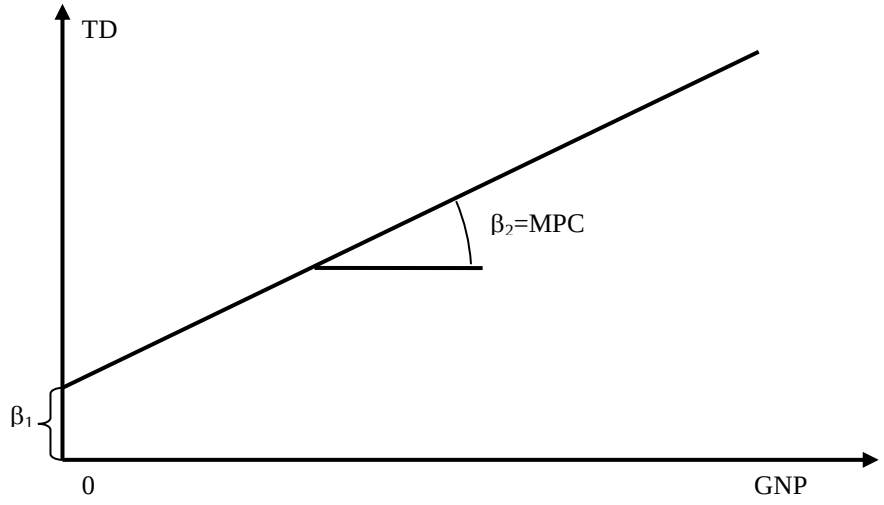
(2) Xây dựng mô hình toán cho lý thuyết hoặc giả thiết

Dạng hàm đơn giản nhất thể hiện ý tưởng của Keynes là dạng hàm tuyến tính.

$$TD = \beta_1 + \beta_2 GNP \qquad (1.1)$$

Trong đó : $0 < \beta_2 < 1$.

Biểu diễn dưới dạng đồ thị của dạng hàm này như sau:



β_1 : Tung độ gốc

β_2: Độ dốc

TD : Biến phụ thuộc hay biến được giải thích

GNP: Biến độc lập hay biến giải thích

Hình 1. 2. Hàm tiêu dùng theo thu nhập.

(3) Xây dựng mô hình kinh tế lượng

[4] John Maynard Keynes, 1936, theo D.N.Gujarati, Basic Economics, 3rd , 1995, trang 3.

Mô hình toán với dạng hàm (1.1) thể hiện mối quan hệ tất định(deterministic relationship) giữa tiêu dùng và thu nhập trong khi quan hệ của các biến số kinh tế thường mang tính không chính xác. Để biểu diễn mối quan hệ không chính xác giữa tiêu dùng và thu nhập chúng ta đưa vào thành phần sai số:

$$TD = \beta_1 + \beta_2 GNP + \varepsilon \qquad (1.2)$$

Trong đó ε là sai số, ε là một biến ngẫu nhiên đại diện cho các nhân tố khác cũng tác động lên tiêu dùng mà chưa được đưa vào mô hình.

Phương trình (1.2) là một mô hình kinh tế lượng. Mô hình trên được gọi là mô hình hồi quy tuyến tính. Hồi quy tuyến tính là nội dung chính của học phần này.

(4) Thu thập số liệu

Số liệu về tiêu dùng và thu nhập của nền kinh tế Việt Nam từ 1986 đến 1998 tính theo đơn vị tiền tệ hiện hành như sau:

Năm	Tiêu dùng TD, đồng hiện hành	Tổng thu nhập GNP, đồng hiện hành	Hệ số khử lạm phát
1986	526.442.004.480	553.099.984.896	2,302
1987	2.530.537.897.984	2.667.299.995.648	10,717
1988	13.285.535.514.624	14.331.699.789.824	54,772
1989	26.849.899.970.560	28.092.999.401.472	100
1990	39.446.699.311.104	41.954.997.960.704	142,095
1991	64.036.997.693.440	76.707.000.221.696	245,18
1992	88.203.000.283.136	110.535.001.505.792	325,189
1993	114.704.005.464.064	136.571.000.979.456	371,774
1994	139.822.006.009.856	170.258.006.540.288	425,837
1995	186.418.693.406.720	222.839.999.299.584	508,802
1996	222.439.040.614.400	258.609.007.034.368	540,029
1997	250.394.999.521.280	313.623.008.247.808	605,557
1998	284.492.996.542.464	361.468.004.401.152	659,676

Bảng 1.1. Số liệu về tổng tiêu dùng và GNP của Việt Nam

Nguồn : World Development Indicator CD-ROM 2000, WorldBank.

TD: Tổng tiêu dùng của nền kinh tế Việt Nam, đồng hiện hành.

GNP: Thu nhập quốc nội của Việt Nam, đồng hiện hành.

Do trong thời kỳ khảo sát có lạm phát rất cao nên chúng ta cần chuyển dạng số liệu về tiêu dùng và thu nhập thực với năm gốc là 1989.

Năm	Tiêu dùng TD, đồng-giá cố định 1989	Tổng thu nhập GNP, đồng-giá cố định 1989
1986	22.868.960.302.145	24.026.999.156.721
1987	23.611.903.339.515	24.888.000.975.960
1988	24.255.972.171.640	26.165.999.171.928
1989	26.849.899.970.560	28.092.999.401.472
1990	27.760.775.225.362	29.526.000.611.153
1991	26.118.365.110.163	31.285.998.882.813
1992	27.123.609.120.801	33.990.999.913.679
1993	30.853.195.807.667	36.735.001.692.581
1994	32.834.660.781.138	39.982.003.187.889
1995	36.638.754.378.646	43.797.002.601.354
1996	41.190.217.461.479	47.888.002.069.333
1997	41.349.567.191.335	51.790.873.128.795
1998	43.126.144.904.439	54.794.746.182.076

Bảng 1.2. Tiêu dùng và thu nhập của Việt Nam, giá cố định 1989

(5) Ước lượng mô hình (Ước lượng các hệ số của mô hình)

Sử dụng phương pháp tổng bình phương tối thiểu thông thường (Ordinary Least Squares)[5] chúng ta thu được kết quả hồi quy như sau:

$TD = 6.375.007.667 + 0,680GNP$

t [4,77] [19,23]

$R^2 = 0,97$

Ước lượng cho hệ số β_1 là $\hat{\beta}_1 = 6.375.007.667$

Ước lượng cho hệ số β_2 là $\hat{\beta}_2 = 0,68$

Xu hướng tiêu dùng biên của nền kinh tế Việt Nam là MPC = 0,68.

(6) Kiểm định giả thiết thống kê

Trị số xu hướng tiêu dùng biên được tính toán là MPC = 0,68 đúng theo phát biểu của Keynes. Tuy nhiên chúng ta cần xác định MPC tính toán như trên có lớn hơn 0 và nhỏ hơn 1 với ý nghĩa thống kê hay không. Phép kiểm định này cũng được trình bày trong chương 2.

(7) Diễn giải kết quả

Dựa theo ý nghĩa kinh tế của MPC chúng ta diễn giải kết quả hồi quy như sau:

Tiêu dùng tăng 0,68 ngàn tỷ đồng nếu GNP tăng 1 ngàn tỷ đồng.

[5] Sẽ được giới thiệu trong chương 2.

(8) Sử dụng kết quả hồi quy

Dựa vào kết quả hồi quy chúng ta có thể dự báo hoặc phân tích tác động của chính sách. Ví dụ nếu dự báo được GNP của Việt Nam năm 2004 thì chúng ta có thể dự báo tiêu dùng của Việt Nam trong năm 2004. Ngoài ra khi biết MPC chúng ta có thể ước lượng số nhân của nền kinh tế theo lý thuyết kinh tế vĩ mô như sau:

$$M = 1/(1-MPC) = 1/(1-0,68) = 3,125$$

Vậy kết quả hồi quy này hữu ích cho phân tích chính sách đầu tư, chính sách kích cầu…

1.3. Những câu hỏi đặt ra cho một nhà kinh tế lượng

1. Mô hình có ý nghĩa kinh tế không?
2. Dữ liệu có đáng tin cậy không?
3. Phương pháp ước lượng có phù hợp không?
4. Kết quả thu được so với kết quả từ mô hình khác hay phương pháp khác như thế nào?

1.4. Dữ liệu cho nghiên cứu kinh tế lượng

Có ba dạng dữ liệu kinh tế cơ bản: dữ liệu chéo, dữ liệu chuỗi thời gian và dữ liệu bảng.

Dữ liệu chéo bao gồm quan sát cho nhiều đơn vị kinh tế ở một thời điểm cho trước. Các đơn vị kinh tế bao gồm các các nhân, các hộ gia đình, các công ty, các tỉnh thành, các quốc gia…

Dữ liệu chuỗi thời gian bao gồm các quan sát trên một đơn vị kinh tế cho trước tại nhiều thời điểm. Ví dụ ta quan sát doanh thu, chi phí quảng cáo, mức lương nhân viên, tốc độ đổi mới công nghệ… ở một công ty trong khoảng thời gian 1990 đến 2002.

Dữ liệu bảng là sự kết hợp giữa dữ liệu chéo và dữ liệu chuỗi thời gian. Ví dụ với cùng bộ biến số về công ty như ở ví dụ trên, chúng ta thu thập số liệu của nhiều công ty trong cùng một khoảng thời gian.

Biến rời rạc hay liên tục

Biến rời rạc là một biến có tập hợp các kết quả có thể đếm được.Ví dụ biến Quy mô hộ gia đình ở ví dụ mục 1.2 là một biến rời rạc.

Biến liên tục là biến nhận kết quả một số vô hạn các kết quả. Ví dụ lượng lượng mưa trong một năm ở một địa điểm.

Dữ liệu có thể thu thập từ một thí nghiệm có kiểm soát, nói cách khác chúng ta có thể thay đổi một biến số trong điều kiện các biến số khác giữ không đổi. Đây chính là cách bố trí thí nghiệm trong nông học, y khoa và một số ngành khoa học tự nhiên.

Đối với kinh tế học nói riêng và khoa học xã hội nói chung, chúng ta rất khó bố trí thí nghiệm có kiểm soát, và sự thực dường như tất cả mọi thứ đều thay đổi nên chúng ta chỉ có thể quan sát hay điều tra để thu thập dữ liệu.

1.5. Vai trò của máy vi tính và phầm mềm chuyên dụng

Vì kinh tế lượng liên quan đến việc xử lý một khối lượng số liệu rất lớn nên chúng ta cần đến sự trợ giúp của máy vi tính và một chương trình hỗ trợ tính toán kinh tế lượng. Hiện nay có rất nhiều phần mềm chuyên dùng cho kinh tế lượng hoặc hỗ trợ xử lý kinh tế lượng.

Excel

Nói chung các phần mềm bảng tính(spreadsheet) đều có một số chức năng tính toán kinh tế lượng. Phần mềm bảng tính thông dụng nhất hiện nay là Excel nằm trong bộ Office của hãng Microsoft. Do tính thông dụng của Excel nên mặc dù có một số hạn chế trong việc ứng dụng tính toán kinh tế lượng, giáo trình này có sử dụng Excel trong tính toán ở ví dụ minh hoạ và hướng dẫn giải bài tập.

Phần mềm chuyên dùng cho kinh tế lượng

Hướng đến việc ứng dụng các mô hình kinh tế lượng và các kiểm định giả thiết một cách nhanh chóng và hiệu quả chúng ta phải quen thuộc với ít nhất một phần mềm chuyên dùng cho kinh tế lượng. Hiện nay có rất nhiều phần mềm kinh tế lượng như:

Phần mềm	Công ty phát triển
AREMOS/PC	Wharton Econometric Forcasting Associate
BASSTAL	BASS Institute Inc
BMDP/PC	BMDP Statistics Software Inc
DATA-FIT	Oxford Electronic Publishing
ECONOMIST WORKSTATION	Data Resources, MC Graw-Hill
ESP	Economic Software Package
ET	New York University
EVIEWS	**Quantitative Micro Software**
GAUSS	Aptech System Inc
LIMDEP	New York University
MATLAB	MathWorks Inc
PC-TSP	TSP International

P-STAT	P-Stat Inc
SAS/STAT	VAR Econometrics
SCA SYSTEM	SAS Institute Inc
SHAZAM	University of British Columbia
SORITEC	The Soritec Group Inc
SPSS	**SPSS Inc**
STATPRO	Penton Sofware Inc

Trong số này có hai phần mềm được sử dụng tương đối phổ biến ở các trường đại học và viện nghiên cứu ở Việt Nam là SPSS và EVIEWS. SPSS rất phù hợp cho nghiên cứu thống kê và cũng tương đối thuận tiện cho tính toán kinh tế lượng trong khi EVIEWS được thiết kế chuyên cho phân tích kinh tế lượng.

CHƯƠNG 2

ÔN TẬP VỀ XÁC SUẤT VÀ THỐNG KÊ

Biến ngẫu nhiên.

Một biến mà giá trị của nó được xác định bởi một phép thử ngẫu nhiên được gọi là một biến ngẫu nhiên. Nói cách khác ta chưa thể xác định giá trị của biến ngẫu nhiên nếu phép thử chưa diễn ra. Biến ngẫu nhiên được ký hiệu bằng ký tự hoa X, Y, Z…. Các giá trị của biến ngẫu nhiên tương ứng được biểu thị bằng ký tự thường x, y, z…

Biến ngẫu nhiên có thể rời rạc hay liên tục. Một biến ngẫu nhiên rời rạc nhận một số hữu hạn(hoặc vô hạn đếm được) các giá trị. Một biến ngẫu nhiên liên tục nhận vô số giá trị trong khoảng giá trị của nó.

Ví dụ 2.1. Gọi X là số chấm xuất hiện khi tung một con súc sắc (xí ngầu). X là một biến ngẫu nhiên rời rạc vì nó chỉ có thể nhận các kết quả 1,2,3,4,5 và 6.

Ví dụ 2.2. Gọi Y là chiều cao của một người được chọn ngẫu nhiên trong một nhóm người. Y cũng là một biến ngẫu nhiên vì chúng ta chỉ có nhận được sau khi đo đạc chiều cao của người đó. Trên một người cụ thể chúng ta đo được chiều cao 167 cm. Con số này tạo cho chúng ta cảm giác chiều cao là một biến ngẫu nhiên rời rạc, nhưng không phải thế, Y thực sự có thể nhận được bất cứ giá trị nào trong khoảng cho trước thí dụ từ 160 cm đến 170 cm tuỳ thuộc vào độ chính xác của phép đo. Y là một biến ngẫu nhiên liên tục.

2.1. Xác suất

2.1.1 Xác suất biến ngẫu nhiên nhận được một giá trị cụ thể

Chúng ta thường quan tâm đến xác suất biến ngẫu nhiên nhận được một giá trị xác định. Ví dụ khi ta sắp tung một súc sắc và ta muốn biết xác suất xuất hiện Xi = 4 là bao nhiêu.

Do con súc sắc có 6 mặt và nếu không có gian lận thì khả năng xuất hiện của mỗi mặt đều như nhau nên chúng ta có thể suy ra ngay xác suất để X= 4 là: P(X=4) = 1/6.

Nguyên tắc lý do không đầy đủ(the principle of insufficient reason): Nếu có K kết quả có khả năng xảy ra như nhau thì xác suất xảy ra một kết quả là 1/K.

Không gian mẫu: Một không gian mẫu là một tập hợp tất cả các khả năng xảy ra của một phép thử, ký hiệu cho không gian mẫu là S. Mỗi khả năng xảy ra là một điểm mẫu.

Biến cố : Biến cố là một tập con của không gian mẫu.

Ví dụ 2.3. Gọi Z là tổng số điểm phép thử tung hai con súc sắc.

Không gian mẫu là S = {2;3;4;5;6;7;8;9;10;11;12}

A = {7;11} Tổng số điểm là 7 hoặc 11

B = {2;3;12} Tổng số điểm là 2 hoặc 3 hoặc 12

C = {4;5;6;8;9;10}

D = {4;5;6;7}

Là các biến cố.

Hợp của các biến cố

E = A hoặc B = $A \cup B$ = {2;3;7;11;12}

Giao của các biến cố:

F = C và D = $C \cap D$ = {4;5;6}

Các tính chất của xác suất

$$P(S) = 1$$

$$0 \le P(A) \le 1$$

$$P(E) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Tần suất

Khảo sát biến X là số điểm khi tung súc sắc. Giả sử chúng ta tung n lần thì số lần xuất hiện giá trị xi là ni. Tần suất xuất hiện kết quả xi là

$$f_i = \frac{n_i}{n}$$

Nếu số phép thử đủ lớn thì tần suất xuất hiện xi tiến đến xác suất xuất hiện xi.

Định nghĩa xác suất

Xác suất biến X nhận giá trị xi là

$$P(X = xi) = \lim_{n \to \infty} \frac{n_i}{n}$$

2.1.2. Hàm mật độ xác suất (phân phối xác suất)

Hàm mật độ xác suất-Biến ngẫu nhiên rời rạc

X nhận các giá trị xi riêng rẽ $x_1, x_2, \ldots, x_n$. Hàm số

$$f(x) = P(X=xi) \quad , \text{với } i = 1;2;..;n$$

$$= 0 \quad , \text{với } x \neq xi$$

được gọi là hàm mật độ xác suất rời rạc của X. P(X=xi) là xác suất biến X nhận giá trị xi.

Xét biến ngẫu nhiên X là số điểm của phép thử tung một con súc sắc. Hàm mật độ xác suất được biểu diễn dạng bảng như sau.

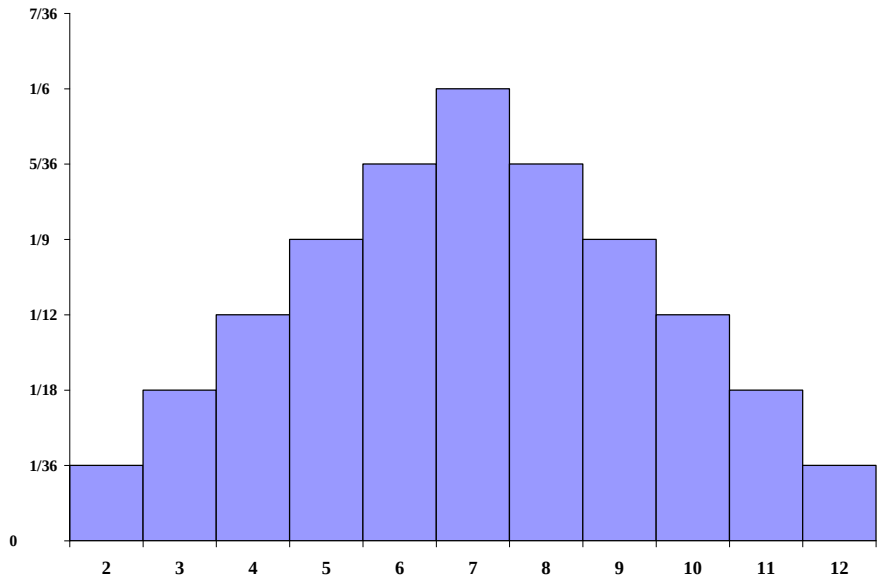
X	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6

Bảng 2.1. Mật độ xác suất của biến ngẫu nhiên rời rạc X

Xét biến Z là tổng số điểm của phép thử tung 2 con súc sắc. Hàm mật độ xác suất được biểu diễn dưới dạng bảng như sau.

z	2	3	4	5	6	7	8	9	10	11	12
P(Z=z)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Bảng 2.2. Mật độ xác suất của biến ngẫu nhiên rời rạc Z



Hình 2.1. Biểu đồ tần suất của biến ngẫu nhiên Z.

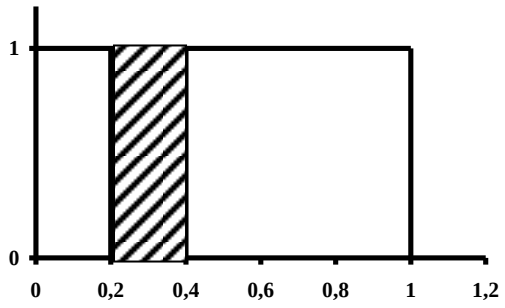
Hàm mật độ xác suất(pdf)-Biến ngẫu nhiên liên tục.

Ví dụ 2.4. Chúng ta xét biến R là con số xuất hiện khi bấm nút Rand trên máy tính cầm tay dạng tiêu biểu như Casio fx-500. R là một biến ngẫu nhiên liên tục nhận giá trị bất kỳ từ 0 đến 1. Các nhà sản xuất máy tính cam kết rằng khả năng xảy ra một giá trị cụ thể là như nhau. Chúng ta có một dạng phân phối xác suất có mật độ xác suất đều.

Hàm mật độ xác suất đều được định nghĩa như sau: $f(r) = \frac{1}{U - L}$

Với L : Giá trị thấp nhất của phân phối

U: Giá trị cao nhất của phân phối



Hình 2.2. Hàm mật độ xác suất đều R.

Xác suất để R rơi vào khoảng (a; b) là P(a <r<b) = $\frac{b-a}{U-L}$.

Cụ thể xác suất để R nhận giá trị trong khoảng (0,2; 0,4) là:

P(0,2 < r < 0,4) = $\frac{0,4-0,2}{1-0} = 20\%$, đây chính là diện tích được gạch chéo trên hình 2.1.

Tổng quát, hàm mật độ xác suất của một biến ngẫu nhiên liên tục có tính chất như sau:

(1) $f(x) \geq 0$

(2) P(a<X<b) = Diện tích nằm dưới đường pdf

$$P(a<X<b) = \int_a^b f(x)dx$$

(3) $$\int_S f(x)dx = 1$$

Hàm đồng mật độ xác suất -Biến ngẫu nhiên rời rạc

Ví dụ 2.5. Xét hai biến ngẫu nhiên rời rạc X và Y có xác suất đồng xảy ra X = xi và Y = yi như sau.

		X		
		2	3	P(Y)
Y	1	0,2	0,4	0,6
	2	0,3	0,1	0,4
	P(X)	0,5	0,5	1,0

Bảng 2.3. Phân phối đồng mật độ xác xuất của X và Y.

Định nghĩa :Gọi X và Y là hai biến ngẫu nhiên rời rạc. Hàm số

$f(x,y)$ = P(X=x và Y=y)

= 0 khi $X \neq$ x và $Y \neq$ y

được gọi là hàm đồng mật độ xác suất, nó cho ta xác xuất đồng thời xảy ra X=x và Y=y.

Hàm mật độ xác suất biên

$f(x) = \sum_y f(x,y)$ hàm mật độ xác suất biên của X

$f(y) = \sum_x f(x,y)$ hàm mật độ xác suất biên của Y

Ví dụ 2.6. Ta tính hàm mật độ xác suất biên đối với số liệu cho ở ví dụ 2.5.

$f(x=2) = \sum_y f(x=2,y) = 0,3 + 0,3 = 0,5$

$f(x=3) = \sum_y f(x=3,y) = 0,1 + 0,4 = 0,5$

$f(y=1) = \sum_x f(x,y=1) = 0,2 + 0,4 = 0,6$

$f(y=2) = \sum_x f(x,y=2) = 0,3 + 0,1 = 0,4$

Xác suất có điều kiện

Hàm số

$f(x \mid y) = P(X=x \mid Y=y)$, xác suất X nhận giá trị x với điều kiện Y nhận giá trị y, được gọi là xác suất có điều kiện của X.

Hàm số

$f(y \mid x) = P(Y=y \mid X=x)$, xác suất Y nhận giá trị y với điều kiện X nhận giá trị x, được gọi là xác suất có điều kiện của Y.

Xác suất có điều kiện được tính như sau

$f(x|y) = \frac{f(x,y)}{f(y)}$, hàm mật độ xác suất có điều kiện của X

$f(y|x) = \frac{f(x,y)}{f(x)}$, hàm mật độ xác suất có điều kiện của Y

Như vậy hàm mật độ xác suất có điều kiện của một biến có thể tính được từ hàm đồng mật độ xác suất và hàm mật độ xác suất biên của biến kia.

Ví dụ 2.7. Tiếp tục ví dụ 2.5 và ví dụ 2.6.

$$f(X=2|Y=1)=\frac{f(X=2,Y=1)}{f(Y=1)}=\frac{0,2}{0,6}=\frac{1}{3}$$

$$f(Y=2|X=3)=\frac{f(X=3,Y=2)}{f(X=3)}=\frac{0,1}{0,5}=\frac{1}{5}$$

Độc lập về thống kê

Hai biến ngẫu nhiên X và Y độc lập về thống kê khi và chỉ khi

$f(x,y)=f(x)f(y)$

tức là hàm đồng mật độ xác suất bằng tích của các hàm mật độ xác suất biên.

Hàm đồng mật độ xác suất cho biến ngẫu nhiên liên tục

Hàm đồng mật độ xác suất của biến ngẫu nhiên liên tục X và Y là f(x,y) thỏa mãn

$f(x,y) \geq 0$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

$$\int_{a}^{b}\int_{c}^{d} f(x,y)dxdy = P(a \leq x \leq b; c \leq y \leq d)$$

Hàm mật độ xác suất biên được tính như sau

$f(x)=\int_{-\infty}^{\infty} f(x,y)dy$, hàm mật độ xác suất biên của X

$f(y)=\int_{-\infty}^{\infty} f(x,y)dx$, hàm mật độ xác suất biên của Y

2.1.3. Một số đặc trưng của phân phối xác suất

Giá trị kỳ vọng hay giá trị trung bình

Giá trị kỳ vọng của một biến ngẫu nhiên rời rạc

$$E(X)=\sum_{X} xf(x)$$

Giá trị kỳ vọng của một biến ngẫu nhiên liên tục

$$E(X) = \int_x xf(x)dx$$

Ví dụ 2.8. Tính giá trị kỳ vọng biến X là số điểm của phép thử tung 1 con súc sắc

$$E(X) = 1*\frac{1}{6} + 2*\frac{1}{6} + 3*\frac{1}{6} + 4*\frac{1}{6} + 5*\frac{1}{6} + 6*\frac{1}{6} = 3,5$$

Một số tính chất của giá trị kỳ vọng

(1) E(a) = a với a là hằng số

(2) E(a+bX) = a + bE(X) với a và b là hằng số

(3) Nếu X và Y là độc lập thống kê thì E(XY) = E(X)E(Y)

(4) Nếu X là một biến ngẫu nhiên có hàm mật độ xác suất f(x) thì

$$E[g(X)] = \sum_x g(X)f(x)$$, nếu X rời rạc

$$E[g(X)] = \int_{-\infty}^{\infty} g(X)f(x)dx$$, nếu X liên tục

Người ta thường ký hiệu kỳ vọng là μ : $\mu = E(X)$

Phương sai

X là một biến ngẫu nhiên và $\mu = E(X)$. Độ phân tán của dữ liệu xung quanh giá trị trung bình được thể hiện bằng phương sai theo định nghĩa như sau:

$$var(X) = \sigma_X^2 = E(X-\mu)^2$$

Độ lệch chuẩn của X là căn bậc hai dương của σ_X^2, ký hiệu là σ_X.

Ta có thể tính phương sai theo định nghĩa như sau

$$var(X) = \sum_x (X-\mu)^2 f(x)$$, nếu X là biến ngẫu nhiên rời rạc

$$= \int_{-\infty}^{\infty} (X-\mu)^2 f(x)dx$$, nếu X là biến ngẫu nhiên liên tục

Trong tính toán chúng ta sử dụng công thức sau

$var(X)=E(X^2)-[E(X)]^2$

Ví dụ 2.9. Tiếp tục ví dụ 2.8. Tính var(X)

Ta đã có E(X) = 3,5

Tính $E(X^2)$ bằng cách áp dụng tính chất (4).

$$E(X^2) = 1^2 * \frac{1}{6} + 2^2 * \frac{1}{6} + 3^2 * \frac{1}{6} + 4^2 * \frac{1}{6} + 5^2 * \frac{1}{6} + 6^2 * \frac{1}{6} = 15{,}17$$

$var(X)=E(X^2)-[E(X)]^2 = 15{,}17 - 3{,}5^2 = 2{,}92$

Các tính chất của phương sai

(1) $E(X-\mu)^2 = E(X^2) - \mu^2$

(2) var(a) = 0 với a là hằng số

(3) $var(a+bX) = b^2 var(X)$ với a và b là hằng số

(4) Nếu X và Y là các biến ngẫu nhiên độc lập thì

var(X+Y) = var(X) + var(Y)

var(X-Y) = var(X) + var(Y)

(5) Nếu X và Y là các biến độc lập, a và b là hằng số thì

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

Hiệp phương sai

X và Y là hai biến ngẫu nhiên với kỳ vọng tương ứng là μ_x và μ_y. Hiệp phương sai của hai biến là

$cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = E(XY) - \mu_x\mu_y$

Chúng ta có thể tính toán trực tiếp hiệp phương sai như sau

Đối với biến ngẫu nhiên rời rạc

$$cov(X,Y) = \sum_y \sum_x (X-\mu_x)(Y-\mu_y) f(x,y)$$

$$= \sum_y \sum_x XY f(x,y) - \mu_x\mu_y$$

Đối với biến ngẫu nhiên liên tục

$$cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (X-\mu_x)(Y-\mu_y) f(x,y)\,dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} XY f(x,y)\,dxdy - \mu_x\mu_y$$

Tính chất của hiệp phương sai

(1) Nếu X và Y độc lập thống kê thì hiệp phương sai của chúng bằng 0.

$$\begin{aligned}\text{cov}(X,Y) &= E(XY) - \mu_x\mu_y \\ &= \mu_x\mu_y - \mu_x\mu_y \\ &= 0\end{aligned}$$

(2) $\text{cov}(a+bX, c+dY) = bd\,\text{cov}(X,Y)$ với a,b,c,d là các hằng số

Nhược điểm của hiệp phương sai là nó phụ thuộc đơn vị đo lường.

Hệ số tương quan

Để khắc phục nhược điểm của hiệp phương sai là phụ thuộc vào đơn vị đo lường, người ta sử dụng hệ số tương quan được định nghĩa như sau:

$$\rho_{xy} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}$$

Hệ số tương quan đo lường mối quan hệ tuyến tính giữa hai biến. ρ sẽ nhận giá trị nằm giữa -1 và 1. Nếu ρ=-1 thì mối quan hệ là nghịch biến hoàn hảo, nếu ρ=1 thì mối quan hệ là đồng biến hoàn hảo.

Từ định nghĩa ta có

$$\text{cov}(X,Y) = \rho\sigma_x\sigma_y$$

2.1.4. Tính chất của biến tương quan

Gọi X và Y là hai biến có tương quan

$$\begin{aligned}\text{var}(X+Y) &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y) \\ &= \text{var}(X) + \text{var}(Y) + 2\rho\sigma_x\sigma_y \\ \text{var}(X-Y) &= \text{var}(X) + \text{var}(Y) - 2\text{cov}(X,Y) \\ &= \text{var}(X) + \text{var}(Y) - 2\rho\sigma_x\sigma_y\end{aligned}$$

Mô men của phân phối xác suất

Phương sai của biến ngẫu nhiên X là mô men bậc 2 của phân phối xác suất của X.

Tổng quát mô men bậc k của phân phối xác suất của X là

$$E(X-\mu)^k$$

Mô men bậc 3 và bậc 4 của phân phối được sử dụng trong hai số đo hình dạng của phân phối xác suất là skewness(độ bất cân xứng) và kurtosis(độ nhọn) mà chúng ta sẽ xem xét ở phần sau.

2.1.5. Một số phân phối xác suất quan trọng

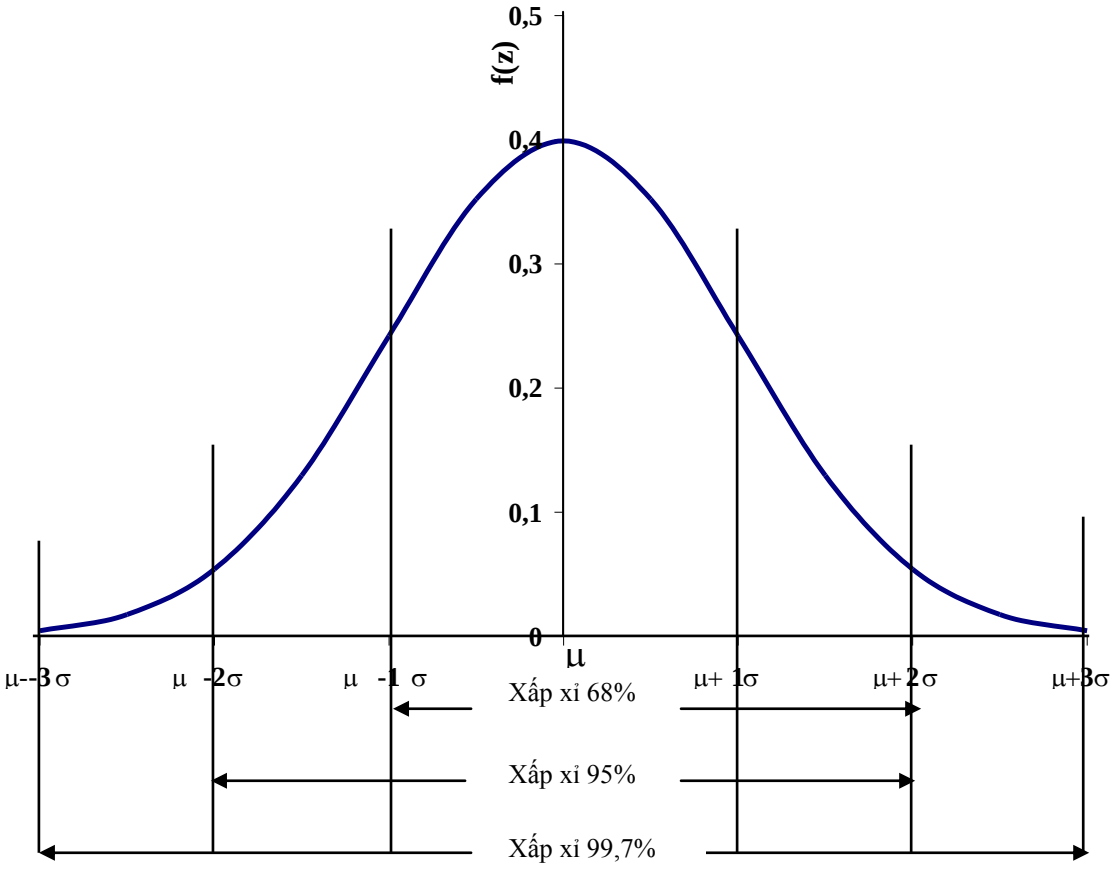
Phân phối chuẩn

Biến ngẫu nhiên X có kỳ vọng là μ, phương sai là σ^2. Nếu X có phân phối chuẩn thì nó được ký hiệu như sau

$$X \sim N(\mu, \sigma^2)$$

Dạng hàm mật độ xác xuất của phân phối chuẩn như sau

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$



Hình 2.3. Hàm mật độ xác suất phân phối chuẩn

Tính chất của phân phối chuẩn

(1) Hàm mật độ xác suất của đối xứng quanh giá trị trung bình.

(2) Xấp xỉ 68% diện tích dưới đường pdf nằm trong khoảng $\mu\pm\sigma$, xấp xỉ 95% diện tích nằm dưới đường pdf nằm trong khoảng $\mu\pm2\sigma$, và xấp xỉ 99,7% diện tích nằm dưới đường pdf nằm trong khoảng $\mu\pm3\sigma$.

(3) Nếu đặt $Z = (X-\mu)/\sigma$ thì ta có $Z\sim N(0,1)$. Z gọi là biến chuẩn hoá và N(0,1) được gọi là phân phối chuẩn hoá.

(4) Định lý giới hạn trung tâm 1: Một kết hợp tuyến tính các biến có phân phối chuẩn,, trong một số điều kiện xác định cũng là một phân phối chuẩn. Ví dụ $X_1 \sim N(\mu_1, \sigma_1^2)$ và $X_2 \sim N(\mu_2, \sigma_2^2)$ thì $Y = aX_1+bX_2$ với a và b là hằng số có phân phối $Y\sim N[(a\mu_1+b\mu_2),(a^2\sigma_1^2 + b^2\sigma_2^2)]$.

(5) Định lý giới hạn trung tâm 2: Dưới một số điều kiện xác định, giá trị trung bình mẫu của các một biến ngẫu nhiên sẽ gần như tuân theo phân phối chuẩn.

(6) Mô men của phân phối chuẩn

Mô men bậc ba: $E[(X-\mu)^3]=0$

Mô men bậc bốn : $E[(X-\mu)^4]=3\sigma^4$

Đối với một phân phối chuẩn

Độ trôi (skewness):

$$S = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = 0$$

Độ nhọn(kurtosis):

$$K = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = 3$$

(7) Dựa vào kết quả ở mục (6), người có thể kiểm định xem một biến ngẫu nhiên có tuân theo phân phối chuẩn hay không bằng cách kiểm định xem S có gần 0 và K có gần 3 hay không. Đây là nguyên tắc xây dựng kiểm định quy luật chuẩn Jarque-Bera.

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

JB tuân theo phân phối χ^2 với hai bậc tự do(df =2).

Phân phối χ^2

Định lý : Nếu $X_1, X_2,\ldots, X_k$ là các biến ngẫu nhiên độc lập có phân phối chuẩn hoá thì $\chi^2_k = \sum_{i=1}^{k} X^2_i$ tuân theo phân phối Chi-bình phương với k bậc tự do.

Tính chất của χ^2

(1) Phân phối χ^2 là phân phối lệch về bên trái, khi bậc tự do tăng dần thì phân phối χ^2 tiến gần đến phân phối chuẩn.

(2) $\mu = k$ và $\sigma^2 = 2k$

(3) $\chi^2_{k1} + \chi^2_{k2} = \chi^2_{k1+k2}$, hay tổng của hai biến có phân phối χ^2cũng có phân phối χ^2 với số bậc tự do bằng tổng các bậc tự do.

Phân phối Student t

Định lý: Nếu Z~N(0,1) và χ^2_k là độc lập thống kê thì $t_{(k)} = \frac{Z}{\sqrt{\chi^2_k / k}}$ tuân theo phân phối Student hay nói gọn là phân phối t với k bậc tự do.

Tính chất của phân phối t

(1) Phân phối t cũng đối xứng quanh 0 như phân phối chuẩn hoá nhưng thấp hơn. Khi bậc tự do càng lớn thì phân phối t tiệm cận đến phân phối chuẩn hoá. Trong thực hành. Khi bậc tự do lớn hơn 30 người ta thay phân phối t bằng phân phối chuẩn hoá.

(2) $\mu = 0$ và $\sigma = k/(k-2)$

Phân phối F

Định lý : Nếu χ^2_{k1} và χ^2_{k2} là độc lập thống kê thì $F_{(K1,k2)} = \frac{\chi^2_{k1}/k_1}{\chi^2_{k2}/k_2}$ tuân theo phân phối F với (k_1,k_2) bậc tự do.

Tính chất của phân phối F

(1) Phân phối F lệch về bên trái, khi bậc tự do k_1 và k_2 đủ lớn, phân phối F tiến đến phân phối chuẩn.

(2) $\mu = k_2/(k_2-2)$ với điều kiện $k_2>2$ và $\sigma^2 = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$ với điều kiện $k_2>4$.

(3) Bình phương của một phân phối t với k bậc tự do là một phân phối F với 1 và k bậc tự do $t_k^2 = F_{(1,k)}$

(4) Nếu bậc tự do mẫu k_2 khá lớn thì $k_1 F_{(k_1,k_2)} = \chi^2_{k_1}$.

Lưu ý : Khi bậc tự do đủ lớn thì các phân phối χ^2, phân phối t và phân phối F tiến đến phân phối chuẩn. Các phân phối này được gọi là phân phối có liên quan đến phân phối chuẩn

2.2. Thống kê mô tả

Mô tả dữ liệu thống kê(Descriptive Statistic)

Có bốn tính chất mô tả phân phối xác suất của một biến ngẫu nhiên như sau:

- Xu hướng trung tâm hay "điểm giữa" của phân phối.
- Mức độ phân tán của dữ liệu quanh vị trí "điểm giữa".
- Độ trôi(skewness) của phân phối.
- Độ nhọn(kurtosis) của phân phối.

Mối quan hệ thống kê giữa hai biến số được mô tả bằng hệ số tương quan.

2.2.1. Xu hướng trung tâm của dữ liệu

Trung bình tổng thể (giá trị kỳ vọng) $\mu_x = E[X]$

Trung bình mẫu $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$

Trung vị của tổng thể : X là một biến ngẫu nhiên liên tục, Md là trung vị của tổng thể khi $P(X<Md) = 0,5$.

Trung vị mẫu : Nếu số phân tử của mẫu là lẻ thì trung vị là số "ở giữa" của mẫu sắp theo thứ tự tăng dần hoặc giảm dần.

Nếu số phần tử của mẫu chẵn thì trung vị là trung bình cộng của hai số "ở giữa".

Trong kinh tế lượng hầu như chúng ta chỉ quan tâm đến trung bình mà không tính toán trên trung vị.

2.2.2. Độ phân tán của dữ liệu

Phương sai

Phương sai của tổng thể : $\sigma_x^2 = E[(X - \mu_x)^2]$

Phương sai mẫu: $S_x^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$

hoặc $\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$

Độ lệch chuẩn

Độ lệch chuẩn tổng thể : $\sigma_x = \sqrt{\sigma_x^2}$

Độ lệch chuẩn mẫu : $S_x = \sqrt{S_x^2}$

hoặc : $\hat{\sigma}_x = \sqrt{\hat{\sigma}_x^2}$

2.2.3. Độ trôi S

Độ trôi tổng thể : $E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$

Độ trôi mẫu : $S = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{\hat{\sigma}}\right)^3$

Đối với phân phối chuẩn độ trôi bằng 0.

2.2.4. Độ nhọn K

Độ nhọn của tổng thể $E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$

Độ nhọn mẫu $K = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{\hat{\sigma}}\right)^4$

Đối với phân phối chuẩn độ nhọn bằng 3. Một phân phối có K lớn hơn 3 là là nhọn, nhỏ hơn 3 là phẳng.

2.2.5. Quan hệ giữa hai biến-Hệ số tương quan

Hệ số tương quan tổng thể $\rho_{XY} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$

Hệ số tương quan mẫu $r_{XY} = \frac{S_{XY}}{S_X S_Y}$

với $$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

2.3. Thống kê suy diễn - vấn đề ước lượng

2.3.1. Ước lượng

Chúng ta tìm hiểu bản chất, đặc trưng và yêu cầu của ước lượng thống kê thông qua một ví dụ đơn giản là ước lượng giá trị trung bình của tổng thể.

Ví dụ 11. Giả sử chúng ta muốn khảo sát chi phí cho học tập của học sinh tiểu học tại trường tiểu học Y. Chúng ta muốn biết trung bình chi phí cho học tập của một học sinh tiểu học là bao nhiêu. Gọi X là biến ngẫu nhiên ứng với chi phí cho học tập của một học sinh tiểu học (X tính bằng ngàn đồng/học sinh/tháng). Giả sử chúng ta biết phương sai của X là σ_x^2=100. Trung bình thực của X là μ là một số chưa biết. Chúng ta tìm cách ước lượng μ dựa trên một mẫu gồm n=100 học sinh được lựa chọn một cách ngẫu nhiên.

2.3.2. Hàm ước lượng cho μ

Chúng ta dùng giá trị trung bình mẫu $\overline{X}$ để ước lượng cho giá trị trung bình của tổng thể μ. Hàm ước lượng như sau

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$\overline{X}$ là một biến ngẫu nhiên. Ứng với một mẫu cụ thể thì $\overline{X}$ nhận một giá trị xác định.

Ước lượng điểm

Ứng với một mẫu cụ thể, giả sử chúng ta tính được $\overline{x}$ = 105 (ngàn đồng/học sinh). Đây là một ước lượng điểm.

Xác suất để một ước lượng điểm như trên đúng bằng trung bình thực là bao nhiêu? Rất thấp hay có thể nói hầu như bằng 0.

Ước lượng khoảng

Ước lượng khoảng cung cấp một khoảng giá trị có thể chứa giá trị chi phí trung bình cho học tập của một học sinh tiểu học. Ví dụ chúng ta tìm được $\overline{x}$ = 105. Chúng ta có thể nói μ có thể nằm trong khoảng $\overline{X} \pm 10$ hay $95 \le \mu \le 115$.

Khoảng ước lượng càng rộng thì càng có khả năng chứa giá trị trung bình thực nhưng một khoảng ước lượng quá rộng như khoảng $\overline{X} \pm 100$ hay $5 \le \mu \le 205$ thì hầu như không giúp ích được gì cho chúng ta trong việc xác định μ. Như vậy có một sự đánh đổi trong ước lượng khoảng với cùng một phương pháp ước lượng nhất định: khoảng càng hẹp thì mức độ tin cậy càng nhỏ.

2.3.3. Phân phối của $\overline{X}$

Theo định lý giới hạn trung tâm 1 thì $\overline{X}$ là một biến ngẫu nhiên có phân phối chuẩn. Vì $\overline{X}$ có phân phối chuẩn nên chúng ta chỉ cần tìm hai đặc trưng của nó là kỳ vọng và phương sai.

Kỳ vọng của $\overline{X}$

$$E(\overline{X}) = E\left(\frac{1}{n}(X_1 + X_2 + ... + X_n)\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} * n\mu = \mu$$

Phương sai của $\overline{X}$

$$\text{var}(\overline{X}) = \text{var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n^2}\text{var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} n\sigma_x^2 = \frac{\sigma_x^2}{n}$$

Vậy độ lệch chuẩn của $\overline{X}$ là $\frac{\sigma_x}{\sqrt{n}}$.

Từ thông tin này, áp dụng quy tắc 2σ thì xác suất khoảng $\overline{X} \pm 2\frac{\sigma_x}{\sqrt{n}}$ chứa μ sẽ xấp xỉ 95%. Ước lượng khoảng với độ tin cậy 95% cho μ là

$$\overline{X} - 2\frac{\sigma_x}{\sqrt{n}} \le \mu \le \overline{X} + 2\frac{\sigma_x}{\sqrt{n}}$$

$$105 - 2\frac{10}{\sqrt{100}} \le \mu \le 105 + 2\frac{10}{\sqrt{100}}$$

$$\hat{\theta}_1 = 103 \le \mu \le 107 = \hat{\theta}_2$$

Lưu ý: Mặc dù về mặt kỹ thuật ta nói khoảng $\overline{X} \pm 2\frac{\sigma_x}{\sqrt{n}}$ chứa μ với xác suất 95% nhưng không thể nói một khoảng cụ thể như (103; 107) có xác suất chứa μ là 95%. Khoảng (103;107) chỉ có thể hoặc chứa μ hoặc không chứa μ.

Ý nghĩa chính xác của độ tin cậy 95% cho ước lượng khoảng cho μ như sau: Với quy tắc xây dựng khoảng là $\overline{X} \pm 2\frac{\sigma_x}{\sqrt{n}}$ và chúng ta tiến hành lấy một mẫu với cỡ mẫu n và tính được một khoảng ước lượng. Chúng ta cứ lặp đi lặp lại quá trình lấy mẫu và ước lượng khoảng như trên thì khoảng 95% khoảng ước lượng chúng ta tìm được sẽ chứa μ.

Tổng quát hơn, nếu trị thống kê cần ước lượng là θ và ta tính được hai ước lượng $\hat{\theta}_1$ và $\hat{\theta}_2$ sao cho

$$P(\hat{\theta}_1 \le \mu \le \hat{\theta}_1) = 1 - \alpha \quad \text{với } 0 < \alpha < 1$$

hay xác suất khoảng từ $\hat{\theta}_1$ đến $\hat{\theta}_2$ chứa giá trị thật θ là 1-α thì 1-α được gọi là độ tin cậy của ước lượng, α được gọi là mức ý nghĩa của ước lượng và cũng là xác suất mắc sai lầm loại I.

Nếu α = 5% thì 1-α là 95%. Mức ý nghĩa 5% hay độ tin cậy 95% thường được sử dụng trong thống kê và trong kinh tế lượng.

Các tính chất đáng mong đợi của một ước lượng được chia thành hai nhóm, nhóm tính chất của ước lượng trên cỡ mẫu nhỏ và nhóm tính chất ước lượng trên cỡ mẫu lớn.

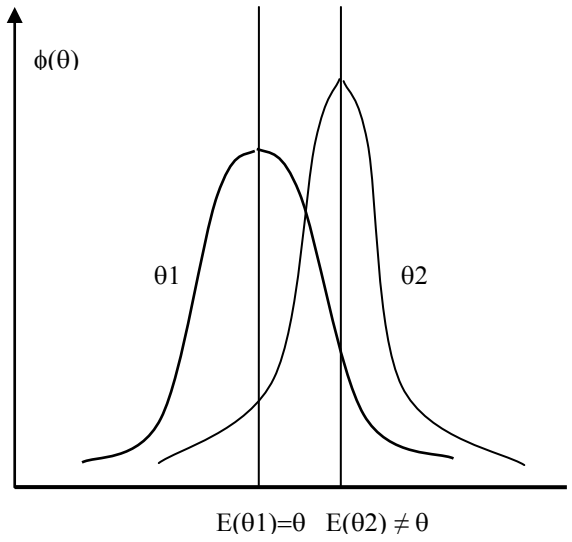
2.3.4. Các tính chất ứng với mẫu nhỏ

Không thiên lệch(không chệch)

Một ước lượng là không thiên lệch nếu kỳ vọng của $\hat{\theta}$ đúng bằng θ.

$$E(\hat{\theta}) = \theta$$

Như đã chứng minh ở phần trên, $\overline{x}$ là ước lượng không thiên lệch của μ.



Hình 2.4. Tính không thiên lệch của ước lượng.

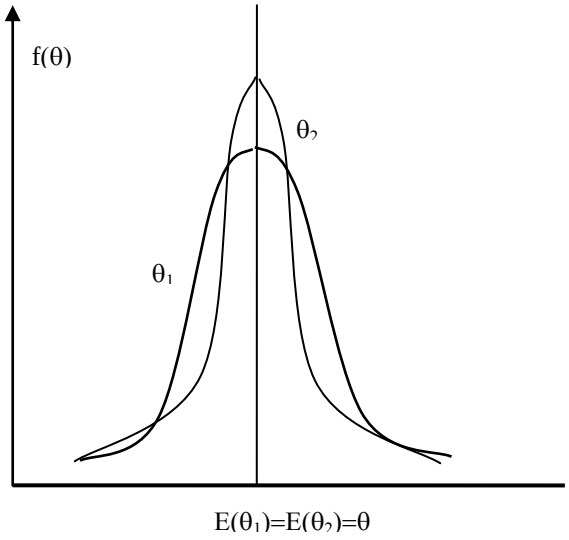
θ_1 là ước lượng không thiên lệch của θ trong khi θ_2 là ước lượng thiên lệch của θ.

Phương sai nhỏ nhất

Hàm ước lượng $\hat{\theta}_1$ có phương sai nhỏ nhất khi với bất cứ hàm ước lượng $\hat{\theta}_2$ nào ta cũng có $var(\hat{\theta}_1) \leq var(\hat{\theta}_2)$.

Không thiên lệch tốt nhất hay hiệu quả

Một ước lượng là hiệu quả nếu nó là ước lượng không thiên lệch và có phương sai nhỏ nhất.



Hình 2.5. Ước lượng hiệu quả. Hàm ước lượng θ_2 hiệu quả hơn θ_1.

Tuyến tính

Một ước lượng $\hat{\theta}$ của θ được gọi là ước lượng tuyến tính nếu nó là một hàm số tuyến tính của các quan sát mẫu.

Ta có $\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$

Vậy $\overline{x}$ là ước lượng tuyến tính cho μ.

Ước lượng không thiên lệch tuyến tính tốt nhất (Best Linear Unbiased Estimator-BLUE)

Một ước lượng $\hat{\theta}$ được gọi là BLUE nếu nó là ước lượng tuyến tính, không thiên lệch và có phương sai nhỏ nhất trong lớp các ước lượng tuyến tính không thiên lệch của θ. Có thể chứng minh được $\overline{x}$ là BLUE.

Sai số bình phương trung bình nhỏ nhất

Sai số bình phương trung bình: $MSE(\hat{\theta})=E(\hat{\theta}-\theta)^2$

Sau khi biến đổi chúng ta nhận được: $MSE(\hat{\theta})=var(\hat{\theta})+E[E(\hat{\theta})-\theta]^2$

$MSE(\hat{\theta})=var(\hat{\theta})+bias(\hat{\theta})$

Sai số bình phương trung bình bằng phương sai của ước lượng cộng với thiên lệch của ước lượng. Chúng ta muốn ước lượng ít thiên lệch đồng thời có phương sai nhỏ. Người ta sử dụng tính chất sai số bình phương trung bình nhỏ khi không thể chọn ước lượng không thiên lệch tốt nhất.

2.3.5. Tính chất của mẫu lớn

Một số ước lượng không thoả mãn các tính chất thống kê mong muốn khi cỡ mẫu nhỏ nhưng khi cỡ mẫu lớn đến vô hạn thì lại có một số tính chất thống kê mong muốn. Các tính chất thống kê này được gọi là tính chất của mẫu lớn hay tính tiệm cận.

Tính không thiên lệch tiệm cận

Ước lượng $\hat{\theta}$ được gọi là không thiên lệch tiệm cận của θ nếu $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$

Ví dụ 2.12. Xét phương sai mẫu của biến ngẫu nhiên X:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n-1}$$

$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n}$$

Có thể chứng minh được

$$E[s_x^2] = \sigma_x^2$$

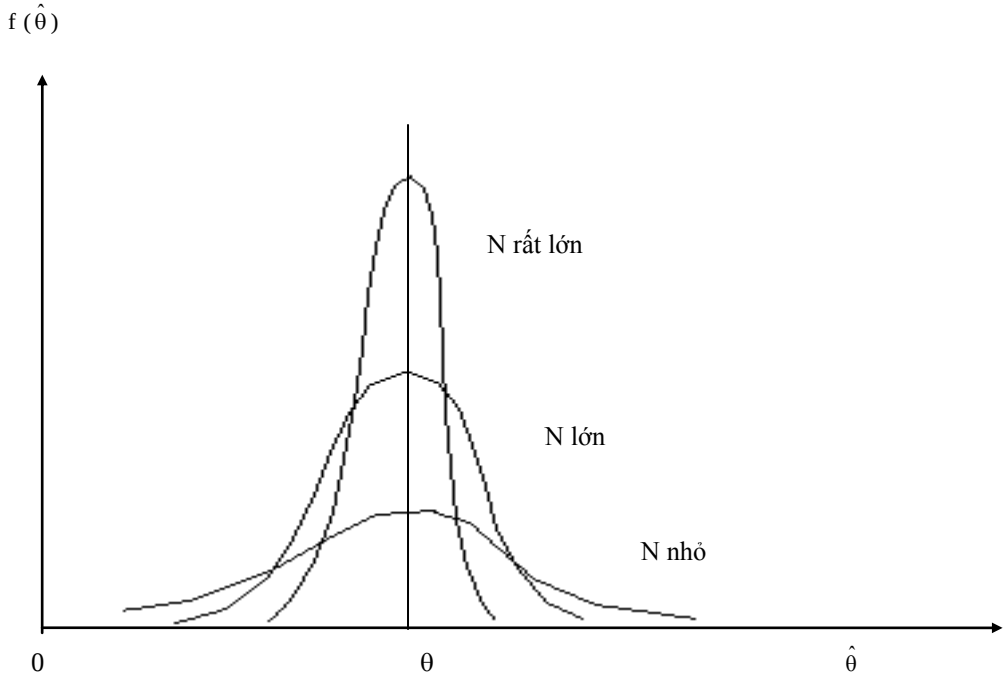
$$E[\hat{\sigma}_x^2] = \sigma_x^2\left(1 - \frac{1}{n}\right)$$

Vậy s_x^2 là ước lượng không thiên lệch của σ_x^2, trong khi $\hat{\sigma}_x^2$ là ước lượng không thiên lệch tiệm cận của σ_x^2.

Nhất quán

Một ước lượng $\hat{\theta}$ được gọi là nhất quán nếu xác suất nếu nó tiến đến giá trị đúng của θ khi cỡ mẫu ngày càng lớn.

$\hat{\theta}$ là nhất quán thì $\lim_{n\to\infty} \left\{\left|\hat{\theta} - \theta\right| < \delta\right\} = 1$ với δ là một số dương nhỏ tuỳ ý.



Hình 2.6. Ước lượng nhất quán

Quy luật chuẩn tiệm cận

Một ước lượng $\hat{\theta}$ được gọi là phân phối chuẩn tiệm cận khi phân phối mẫu của nó tiến đến phân phối chuẩn khi cỡ mẫu *n* tiến đến vô cùng.

Trong phần trên chúng ta đã thấy biến X có phân phối chuẩn với trung bình μ và phương sai σ^2 thì $\overline{X}$ có phân phối chuẩn với trung bình μ và phương sai σ^2/n với cả cỡ mẫu nhỏ và lớn.

Nếu X là biến ngẫu nhiên có trung bình μ và phương sai σ^2 nhưng không theo phân phân phối chuẩn thì $\overline{X}$ cũng sẽ có phân phối chuẩn với trung bình μ và phương sai σ^2/n khi n tiến đến vô cùng. Đây chính là định lý giới hạn trung tâm 2.

2.4. Thống kê suy diễn - Kiểm định giả thiết thống kê

2.4.1. Giả thiết

Giả thiết không là một phát biểu về giá trị của tham số hoặc về giá trị của một tập hợp các tham số. Giả thiết ngược phát biểu về giá trị của tham số hoặc một tập hợp tham số khi giả thiết không sai. Giả thiết không thường được ký hiệu là H_0 và giả thiết ngược thường được ký hiệu là H_1.

2.4.2. Kiểm định hai đuôi

Ví dụ 13. Quay lại ví dụ 11 về biến X là chi phí cho học tập của học sinh tiểu học. Chúng ta biết phương sai của X là σ_x^2=100. Với một mẫu với cỡ mẫu n=100 chúng ta đã tính được $\overline{X}_1$=105 ngàn đồng/học sinh/tháng. Chúng ta xem xét khả năng bác bỏ phát biểu cho rằng chi phí cho học tập trung bình của học sinh tiểu học là 106 ngàn đồng/tháng.

Giả thiết

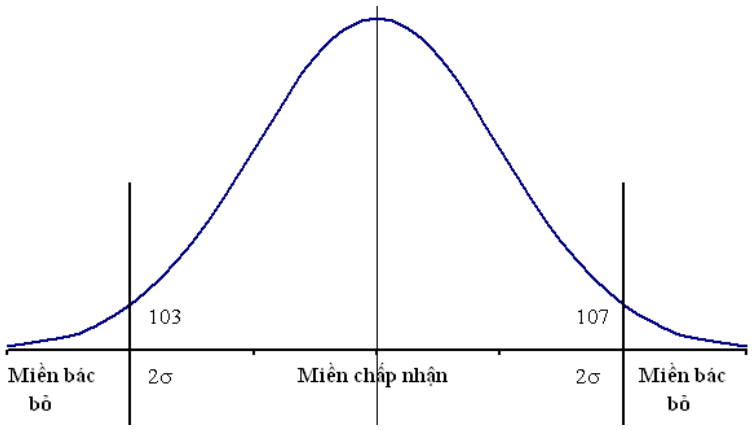
H_0: $\mu = 106 = \mu_0$

H_1: $\mu \neq 106 = \mu_0$

Chúng ta đã biết $\overline{X} \sim N(\mu, \sigma_x^2/n)$, với độ tin cậy 95% hay mức ý nghĩa a = 5% chúng ta đã xây dựng được ước lượng khoảng của μ là $\overline{X}_1 \pm 2\frac{\sigma_x}{\sqrt{n}}$. Nếu khoảng này không chứa μ thì ta bác bỏ giả thiết không với độ tin cậy 95%, ngược lại ta không đủ cơ sở để bác bỏ giả thiết H_0.

Ở phần trên chúng ta đã tính được ước lượng khoảng của μ dựa theo $\overline{X}_1$ là (103;107). Khoảng này chứa $\mu_0 = 106$. Vậy ta không thể bác bỏ được giả thiết H_0.

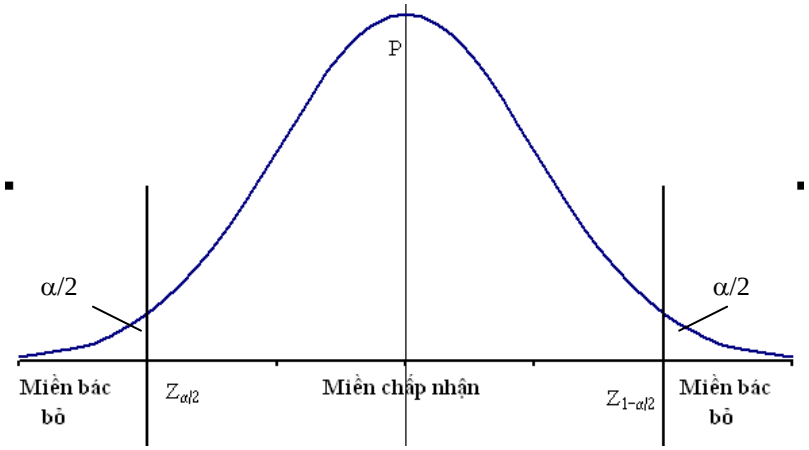
Khoảng tin cậy mà ta thiết lập được được gọi là miền chấp nhận, miền giá trị nằm ngoài miền chấp nhận được gọi là miền bác bỏ.



Hình 2.7. Miền bác bỏ và miền chấp nhận H_0.

Tổng quát hơn ta có

$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ hay Z tuân theo phân phối chuẩn hoá.



Hình 2.8. Miền chấp nhận và miền bác bỏ theo α của trị thống kê Z

Ta có tất cả hai miền bác bỏ và do tính chất đối xứng của phân phối chuẩn, nếu mức ý nghĩa là α thì xác suất để Z nằm ở miền bác bỏ bên trái là α/2 và xác suất để Z nằm ở miền bác bỏ bên trái cũng là α/2. Chúng ta đặt giá trị tới hạn bên trái là $Z_{\alpha/2}$ và giá trị tới hạn bên phải là $Z_{1-\alpha/2}$. Do tính đối xứng ta lại có $Z_{\alpha/2} = - Z_{1-\alpha/2}$.

Xác suất để Z nằm trong hai khoảng tới hạn là

$$P\left(Z_{\alpha/2} \le Z \le Z_{1-\alpha/2}\right) = 1 - \alpha \qquad (2.1)$$

hay

$$P\left(-Z_{1-\alpha/2} \le Z \le Z_{1-\alpha/2}\right) = 1 - \alpha$$

Thay $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ và biến đổi một chút chúng ta nhận được

$$P\left(\overline{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \qquad (2)$$

Các mệnh đề (2.1) và (2.2) là những mệnh đề xác suất.

Kiểm định giả thiết thống kê theo phương pháp truyền thống

Phát biểu mệnh đề xác suất

$$P\left(\overline{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \middle| \mu = \mu_0\right) = 1 - \alpha$$

Nguyên tắc ra quyết định

- Nếu $\overline{X}_1 - Z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}} > \mu_0$ hoặc $\overline{X}_1 + Z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}} < \mu_0$ thì ta bác bỏ H_0 với độ tin cậy 1-α hay xác suất mắc sai lầm là α.

- Nếu $\overline{X}_1 - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \overline{X}_1 + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ thì ta không thể bác bỏ H_0.

Với mức ý nghĩa α =5% thì $Z_{1-\alpha/2} = Z_{97,5\%} = 1,96 \approx 2$

Ta có $\overline{X}_1 - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 105 - 2\frac{10}{10} = 103$

$\overline{X}_1 + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 105 + 2\frac{10}{10} = 107$

Vậy ta không thể bác bỏ giả thiết Ho.

Kiểm định giả thiết thống kê theo trị thống kê Z

Phát biểu mệnh đề xác suất

$$P\left(Z_{\alpha/2} \le Z \le Z_{1-\alpha/2}\right) = 1-\alpha$$

Quy tắc quyết định

- Nếu $Z_{tt} = \frac{\overline{X}_1 - \mu_0}{\sigma^2/\sqrt{n}} < Z_{\alpha/2}$ hoặc $Z_{tt} = \frac{\overline{X}_1 - \mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha/2}$ thì ta bác bỏ H_0 với độ tin cậy 1-α hay xác suất mắc sai lầm là α.
- Nếu $Z_{\alpha/2} \le Z_{tt} \le Z_{1-\alpha/2}$ thì ta không thể bác bỏ H_0.

Với mức ý nghĩa α =5% ta có

$Z_{1-\alpha/2} = Z_{97,5\%} = 1,96 \approx 2$

và $Z_{\alpha/2} = Z_{2,5\%} = -1,96 \approx -2$

$Z_{tt} = \frac{\overline{X}_1 - \mu_0}{\sigma/\sqrt{n}} = \frac{105 - 106}{10/\sqrt{100}} = -1$

Vậy ta không thể bác bỏ Ho.

Kiểm định giả thiết thống kê theo giá trị p

Đối với kiểm định hai đuôi giá trị p được tính như sau:

$$p = 2P\left(\left|Z_{tt}\right| < Z\right)$$

Với Ztt = -1 ta có P(1<Z) = 0,16, vậy giá trị p = 0,32.

Quy tắc quyết định

- Nếu $p < \alpha$: Bác bỏ Ho.
- Nếu $p \geq \alpha$: Không thể bác bỏ Ho.

Trong ví dụ trên $p = 0,32 > \alpha = 5\%$. Vậy ta không thể bác bỏ Ho.

Ba cách tiếp cận trên cho cùng một kết quả vì thực ra chỉ từ những biến đổi của cùng một mệnh đề xác suất. Trong kinh tế lượng người ta cũng thường hay sử dụng giá trị p.

2.4.3. Kiểm định một đuôi

Kiểm định đuôi trái

Ví dụ 14. Tiếp tục ví dụ 13. Kiểm định phát biểu : "Chi cho học tập trung bình của học sinh tiểu học lớn hơn 108 ngàn đồng/học sinh/tháng".

Giả thiết

H_0: $\mu > 108 = \mu_0$

H_1: $\mu \leq 108 = \mu_0$

Phát biểu mệnh đề xác suất

$P(Z_\alpha < Z) = 1-\alpha$

Quy tắc quyết định

- Nếu $Z_{tt} < Z_\alpha$: Bác bỏ Ho.
- Nếu $Z_{tt} \geq Z_\alpha$: Không thể bác bỏ Ho.

Với $\alpha = 5\%$ ta có $Z_{5\%} = -1,644$

Ta có $Ztt = \dfrac{\overline{X}_1 - \mu_0}{\sigma/\sqrt{n}} = \dfrac{105 - 108}{10/\sqrt{100}} = -3 < Z_{5\%} = -1,644$ vậy ta bác bỏ Ho.

Kiểm định đuôi phải

Ví dụ 15. Tiếp tục ví dụ 13. Kiểm định phát biểu : "Chi tiêu cho học tập trung bình của học sinh tiểu học nhỏ hơn 108 ngàn đồng/học sinh/tháng".

Giả thiết

H_0: $\mu < 107 = \mu_0$

H_1: $\mu \geq 107 = \mu_0$

Phát biểu mệnh đề xác suất

$P(Z<Z_{1-\alpha}) = 1-\alpha$

Quy tắc quyết định

- Nếu $Z_{tt} > Z_\alpha$: Bác bỏ Ho.
- Nếu $Z_{tt} \leq Z_\alpha$: Không thể bác bỏ Ho.

Ta có Ztt = $\frac{\overline{X}_1 - \mu_0}{\sigma/\sqrt{n}} = \frac{105 - 107}{10/\sqrt{100}} = -2 < Z_{5\%} = -1{,}644$ vậy ta không thể bác bỏ Ho.

2.4.4. Một số trường hợp đặc biệt cho ước lượng giá trị trung bình của tổng thể

- Tổng thể có phân phối chuẩn, cỡ mẫu lớn, phương sai chưa biết. Chiến lược kiểm định giống như trên nhưng thay phương sai tổng thể bằng phương sai mẫu.
- Tổng thể có phân phối chuẩn, phương sai chưa biết, cỡ mẫu nhỏ:

$$\frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim \text{t-stat} \sim t_{(n-1)}$$

Kiểm định trên trị thống kê t cũng tương tự như đối với trị thống kê Z, ta chỉ việc tra t thay cho Z. Khi cỡ mẫu đủ lớn trị thống kê t tương tự trị thống kê Z.

- Tổng thể không tuân theo phân phối chuẩn, áp dụng định lý giới hạn trung tâm. Khi cỡ mẫu đủ lớn thì trị thống kê t tính toán như phần trên có phân phối gần với phân phối Z.

Ngoài ra chúng ta còn có thể kiểm định các giả thiết về phương sai, kiểm định sự bằng nhau giữa các phương sai của hai tổng thể và kiểm định sự bằng nhau giữa các trung bình tổng thể. Chúng ta xét kiểm định giả thiết về phương sai vì giả định về phương sai không đổi là một giả định quan trọng trong phân tích hồi quy.

Kiểm định giả thiết về phương sai

Xét giả thiết

Ho : $\sigma^2 = \sigma_0^2$

H1 : $\sigma^2 \neq \sigma_0^2$

Có thể chứng minh được

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Mệnh đề xác suất

$$P\left(\chi^2_{(n-1,\alpha/2)} \le (n-1)\frac{s^2}{\sigma_0^2} \le \chi^2_{(n-1,1-\alpha/2)} \right) = 1-\alpha$$

Quy tắc quyết định

Nếu $(n-1)\frac{s^2}{\sigma_0^2} < \chi^2_{(n-1,\alpha/2)}$ hoặc $(n-1)\frac{s^2}{\sigma_0^2} > \chi^2_{(n-1,\alpha/2)}$, thì bác bỏ H_0.

Nếu $\chi^2_{(n-1,\alpha/2)} \le (n-1)\frac{s^2}{\sigma_0^2} \le \chi^2_{(n-1,1-\alpha/2)}$, thì không bác bỏ H_0.

Kiểm định sự bằng nhau của phương sai hai tổng thể

Chúng ta có mẫu cỡ n_1 từ tổng thể 1 và mẫu cỡ n_2 từ tổng thể 2.

Xét giả thiết

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$

$H_1 : \sigma_1^2 \ne \sigma_2^2$

Chúng ta đã có $(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{(n-1)}$

Vậy $$\frac{(n_1-1)\frac{s_1^2}{\sigma^2} \Big/ (n_1-1)}{(n_2-1)\frac{s_2^2}{\sigma^2} \Big/ (n_2-1)} \sim \frac{\chi^2_{(n_1-1)} \Big/ (n_1-1)}{\chi^2_{(n_2-1)} \Big/ (n_2-1)} \sim F_{(n_1-1,n_2-1)}$$

Hay $\frac{s_1^2}{s_2^2} \sim F_{(n_1-1,n_2-1)}$

Phát biểu mệnh đề xác suất

$$P\left(F_{(n_1-1,n_2-1,\alpha/2)} \le \frac{s_1^2}{s_2^2} \le F_{(n_1-1,n_2-1,1-\alpha/2)} \right) = 1-\alpha$$

Quy tắc quyết định

- Nếu $\frac{s_1^2}{s_2^2} < F_{(n_1-1,n_2-1,\alpha/2)}$ hoặc $\frac{s_1^2}{s_2^2} > F_{(n_1-1,n_2-1,1-\alpha/2)}$ thì ta bác bỏ H_0.

➢ Nếu $F_{(n_1-1,n_2-1,\alpha/2)} \le \frac{s_1^2}{s_2^2} \le F_{(n_1-1,n_2-1,1-\alpha/2)}$ thì không bác bỏ H_0.

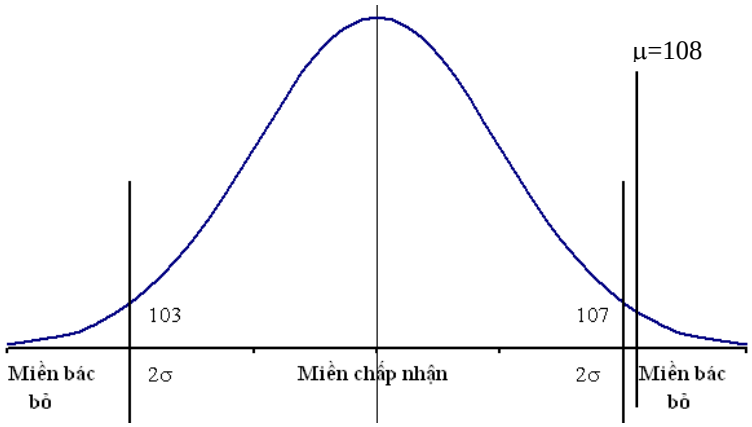
2.4.5. Sai lầm loại I và sai lầm loại II

Khi ta dựa vào một mẫu để bác bỏ một giả thiết, ta có thể mắc phải một trong hai sai lầm như sau:

Sai lầm loại I: Bác bỏ Ho khi thực tế Ho đúng.

Sai lầm loại II : Không bác bỏ Ho khi thực tế nó sai.

	Tính chất	
Quyết định	H_0 đúng	H_0 sai
Bác bỏ	**Sai lầm loại I**	Không mắc sai lầm
Không bác bỏ	Không mắc sai lầm	**Sai lầm loại II**



Hình 2.7. Sai lầm loại I-Bác bỏ H_0: μ=108 trong khi thực tế H_0 đúng.

Xác suất mắc sai lầm loại I

Ví dụ 16. Tiếp tục ví dụ 13. Kiểm định phát biểu : "Chi cho học tập trung bình của học sinh tiểu học là 108 ngàn đồng/học sinh/tháng". Trung bình thực $\mu = \mu_0=108$.

Giả thiết

H_0: $\mu = 108 = \mu_0$

H_1: $\mu \neq 108 = \mu_0$

Giả sử giá trị μ thực là μ=108. Với ước lượng khoảng cho μ là (103;107) với độ tin cậy 95% chúng ta bác bỏ H_0 trong khi thực sự H_0 là đúng. Xác suất chúng ta mắc sai lầm loại này là $\alpha = 5\%$.

Xác suất mắc sai lầm loại II

Ví dụ 17. Tiếp tục ví dụ 13. Kiểm định phát biểu : "Chi tiêu cho học tập trung bình của học sinh tiểu học là 108 ngàn đồng/học sinh/tháng". Trung bình thực $\mu = \mu_0=104$.

Giả thiết

H_0: $\mu = 108 = \mu_0$

H_1: $\mu \neq 108 = \mu_0$

Giả sử giá trị μ thực là $\mu=104$. Với ước lượng khoảng cho μ là (103;107) với độ tin cậy 95% chúng ta không bác bỏ H_0 trong khi H_0 sai. Xác suất chúng ta mắc sai lầm loại II này là β.

Lý tưởng nhất là chúng ta tối thiểu hoá cả hai loại sai lầm. Nhưng nếu chúng ta muốn hạn chế sai lầm loại I, tức là chọn mức ý nghĩa α nhỏ thì khoảng ước lượng càng lớn và xác suất mắc phải sai lầm loại II càng lớn. Nghiên cứu của Newman và Pearson[6] cho rằng sai lầm loại I là nghiêm trọng hơn sai lầm loại II. Do đó, trong thống kê suy diễn cổ điển cũng như trong kinh tế lượng cổ điển, người ta chọn mức ý nghĩa α hay xác suất mắc sai lầm loại I nhỏ, thông thường nhất là 5% mà không quan tâm nhiều đến β.

2.4.6. Tóm tắt các bước của kiểm định giả thiết thống kê

Bước 1. Phát biểu giả thiết H_0 và giả thiết ngược H_1.

Bước 2. Lựa chọn trị thống kê kiểm định

Bước 3. Xác định phân phối thống kê của kiểm định

Bước 4. Lựa chọn mức ý nghĩa α hay xác suất mắc sai lầm loại I.

Bước 5. Sử dụng phân phối xác suất của thống kê kiểm định, thiết lập một khoảng tin cậy 1-α, khoảng này còn được gọi là miền chấp nhận. Nếu trị thống kê ứng với H_0 nằm trong miền chấp nhận thì ta không bác bỏ H_0, nếu trị thông kê ứng với H_0 nằm ngoài miền chấp nhận thì ta bác bỏ H_0. Lưu ý là khi bác bỏ H_0 chúng ta chấp nhận mức độ sai lầm là α.

[6] Damodar N. Gujarati, Basic Econometrics-Third Edition, McGraw-Hill Inc -1995, p 787.

CHƯƠNG 3
HỒI QUY HAI BIẾN

3.1. Giới thiệu

3.1.1. Khái niệm về hồi quy

Phân tích hồi quy là tìm quan hệ phụ thuộc của một biến, được gọi là biến phụ thuộc vào một hoặc nhiều biến khác, được gọi là biến độc lập nhằm mục đích ước lượng hoặc tiên đoán giá trị kỳ vọng của biến phụ thuộc khi biết trước giá trị của biến độc lập.[7]

Một số tên gọi khác của biến phụ thuộc và biến độc lập như sau:

Biến phụ thuộc: biến được giải thích, biến được dự báo, biến được hồi quy, biến phản ứng, biến nội sinh.

Biến độc lập: biến giải thích, biến dự báo, biến hồi quy, biến tác nhân hay biến kiểm soát, biến ngoại sinh.

Sau đây là một và ví dụ về phân tích hồi quy

(1) Ngân hàng XYZ muốn tăng lượng tiền huy động. Ngân hàng này muốn biết mối quan hệ giữa lượng tiền gửi và lãi suất tiên gửi, cụ thể hơn họ muốn biết khi tăng lãi suất thêm 0,1% thì lượng tiền gửi sẽ tăng trung bình là bao nhiêu.

(2) Một nhà nghiên cứu nông nghiệp muốn biết năng suất tôm sú nuôi trong hệ thống thâm canh phụ thuộc thế nào vào diện tích ao nuôi, mật độ thả tôm giống, chi phí hoá chất xử lý môi trường, trình độ nhân công. Từ phân tích hồi quy này ông ta đề ra các chỉ tiêu kỹ thuật phù hợp cho loại hình này.

3.1.2. Sự khác nhau giữa các dạng quan hệ

Quan hệ tất định và quan hệ thống kê

Quan hệ tất định là loại quan hệ có thể biểu diễn bằng môt hàm số toán học. Một số quan hệ trong vật lý, hoá học và một số ngành khoa học tự nhiên khác là quan hệ tất định.

Ví dụ định luật Ohm trong vật lý : gọi U là điện áp, R là điện trở của mạch điện thì dòng điện I sẽ là $I = \frac{U}{R}$, nói cách khác khi điện áp và điện trở được cố định trước thì chúng ta chỉ nhận được một và chỉ một giá trị dòng điện.

Đa số các biến số kinh tế không có quan hệ tất định. Thí dụ ta không thể nói với diện tích nuôi tôm cho trước và kỹ thuật nuôi được chọn thì năng suất sẽ là bao nhiêu. Lý do là có rất nhiều biến số được kể đến trong mô hình cũng tác động lên năng suất, ngoài ra trong số các biến số vắng mặt này có những biến không thể kiểm soát được như thời tiết, dịch bệnh… Nhà nghiên cứu nông nghiệp kể trên chỉ có thể tiên đoán một giá trị trung bình của năng suất ứng với kỹ thuật nuôi đã chọn. Quan hệ giữa các biến số kinh tế có tính chất quan hệ thống kê.

[7] Theo Damodar N.Gujarati, Basic Econometrics-Third Edition, McGraw-Hill-1995, p16.

Hồi quy và quan hệ nhân quả

Mặc dù phân tích hồi quy dựa trên ý tưởng sự phụ thuộc của một biến số kinh tế vào biến số kinh tế khác nhưng bản thân kỹ thuật phân tích hồi quy không bao hàm quan hệ nhân quả. Một ví dụ điển hình của sự nhầm lẫn hai khái niệm này tiến hành hồi quy số vụ trộm ở một thành phố với số nhân viên cảnh sát của thành phố. Gọi Y là số vụ trộm trong một năm và X là số nhân viên cảnh sát. Khi chúng ta hồi quy Y theo X, nếu chúng ta tìm được mối quan hệ đồng biến của Y và X có ý nghĩa thống kê thì phân tích hồi quy này cho kết luận: "Tăng số lượng nhân viên cảnh sát sẽ làm tăng số vụ trộm". Rõ ràng phân tích này sai lầm trong việc nhận định mối quan hệ nhân quả. Số cảnh sát tăng lên là do sự tăng cường của lực lượng cảnh sát trong bối cảnh số vụ trộm tăng lên. Vậy đúng ra chúng ta phải hồi quy số cảnh sát theo số vụ trộm hay X theo Y.Vậy trước khi phân tích hồi quy chúng ta phải nhận định chính xác mối quan hệ nhân quả.[8]

Một sai lầm phổ biến nữa trong phân tích kinh tế lượng là quy kết mối quan hệ nhân quả giữa hai biến số trong khi trong thực tế chúng đều là hệ quả của một nguyên nhân khác. Ví dụ chúng ta phân tích hồi quy giữa số giáo viên và số phòng học trong toàn ngành giáo dục. Sự thực là cả số giáo viên và số phòng học đều phụ thuộc vào số học sinh. Như vậy phân tích mối quan hệ nhân quả dựa vào kiến thức và phương pháp luận của môn khác chứ không từ phân tích hồi quy.

Hồi quy và tương quan

Phân tích tương quan chỉ cho thấy độ mạnh yếu của mối quan hệ tuyến tính giữa hai biến số. Phân tích tương quan cũng không thể hiện mối quan hệ nhân quả.Ví dụ chúng ta xét quan hệ giữa hai biến số X là số bệnh nhân bị xơ gan và Y là số lít rượu được tiêu thụ của một nước. Chúng ta có thể nhận được hệ số tương quan cao giữa X và Y. Hệ số tương quan được xác định như sau:

$$r_{XY} = \frac{\text{cov}(X, Y)}{S_X S_Y} = \frac{\text{cov}(Y, X)}{S_Y S_X} = r_{YX}$$

Qua đẳng thức này chúng ta cũng thấy trong phân tích tương quan vai trò của hai biến là như nhau và hai biến đều là ngẫu nhiên.

Phân tích hồi quy của X theo Y cho ta biết trung bình số bệnh nhân bị xơ gan là bao nhiêu ứng với lượng tiêu dùng rượu cho trước. Chúng ta không thể đảo ngược hồi quy thành Y theo X. Phân tích hồi quy dựa trên giả định biến độc lập là xác định trong khi biến phụ thuộc là ngẫu nhiên. Chúng ta tìm giá trị kỳ vọng của biến phụ thuộc dựa vào giá trị cho trước của của biến độc lập.

[8] Ramu Ramanathan, Introductory Econometrics with Applications, Harcourt College Publishers-2002, trang 113.

3.2.Hàm hồi quy tổng thể và hồi quy mẫu

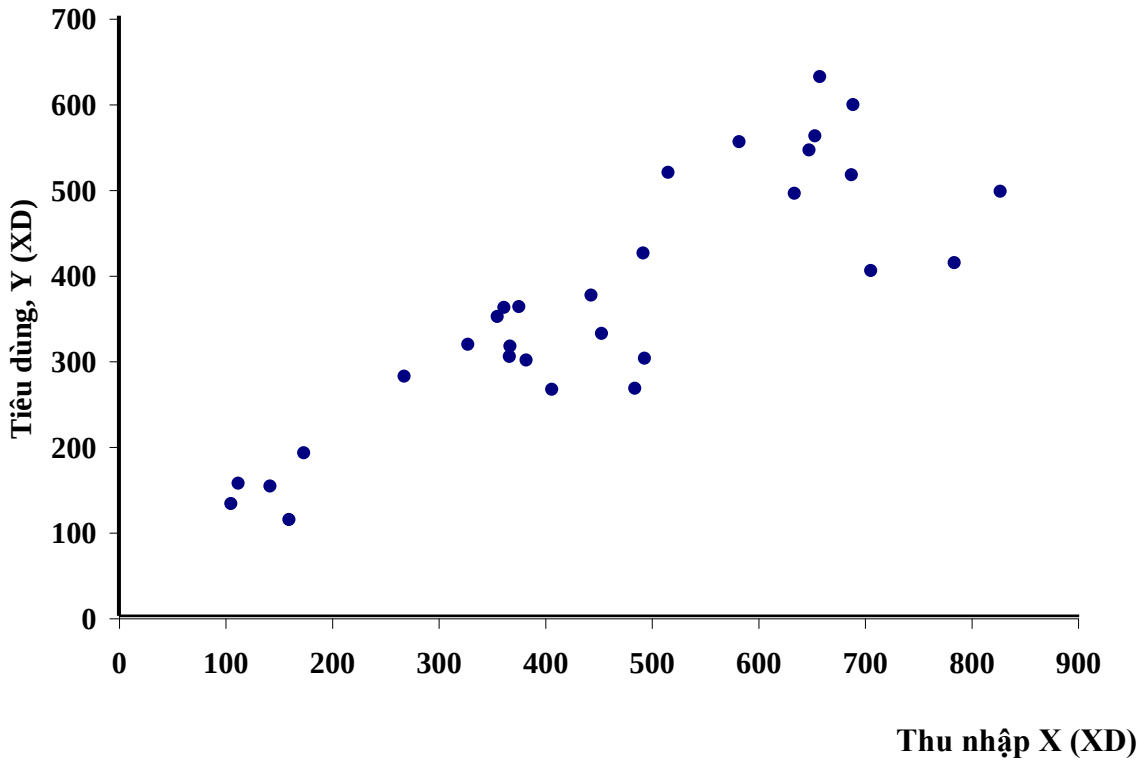
3.2.1.Hàm hồi quy tổng thể (PRF)

Ví dụ 3.1. Hồi quy tiêu dùng Y theo thu nhậpX.

Theo Keynes thì hàm tiêu dùng như sau [9]:

$$Y = \beta_1 + \beta_2 X \text{ , với } \beta_2 \text{ là xu hướng tiêu dùng biên, } 0<\beta_2<1. \qquad (3.1)$$

Chúng ta kiểm chứng giả thiết trên với số liệu từ một nước giả định Z có dân số 30 người với số liệu tiêu dùng và thu nhậpcủa từng người như đồ thị phân tán sau.[10]



Hình 3.1. Đồ thị phân tán quan hệ giữa tiêu dùng và thu nhập khả dụng.

Đồ thị 3.1. cho thấy có mối quan hệ đồng biến giữa tiêu dùng và thu nhập khả dụng, hay là thu nhậptăng sẽ làm tiêu dùng tăng. Tuy quan hệ giữa Y và X không chính xác như hàm bậc nhất (3.1).

Trong phân tích hồi quy chúng ta xem biến độc lập X có giá trị xác định trong khi biến phụ thuộc Y là biến ngẫu nhiên. Điều này tưởng như bất hợp lý. Khi chúng ta chọn ngẫu nhiên người thứ i thì chúng ta thu được đồng thời hai giá trị: Xi là thu nhậpvà Yi là tiêu dùng của người đó. Vậy tại sao lại xem Yi là ngẫu nhiên? Câu trả như sau : Xét một mức thu nhậpX_i xác định, cách lấy mẫu của chúng ta là chọn ngẫu nhiên trong số những người có thu nhậplà Xi. Thu nhậpgóp phần chính yếu quyết định tiêu dùng như thể hiện ở hàm số (1.3), tuy nhiên còn nhiều yếu tố khác cũng tác động lên tiêu dùng nên ứng với một

[9] Damodar N Gujarati, Basic Economics-3rd Edidtion,p4.

[10] Số liệu ở phụ lục 3.1.PL cuối chương 3.

cách lấy mẫu thì với nhiều lần lấy mẫu với tiêu chí X = Xi ta nhận được các giá trị Yi khác nhau. Vậy chính xác hơn biến phụ thuộc Y là một biến ngẫu nhiên có điều kiện theo biến độc lập X. Ước lượng tốt nhất cho Y trong trường hợp này là giá trị kỳ vọng của Y ứng với điều kiện X nhận giá trị Xi xác định.

Hàm hồi quy tổng thể (PRF):

$$E(Y/X=Xi) = \beta_1 + \beta_2 X \qquad (3.2)$$

Đối với một quan sát cụ thể thì giá trị biến phụ thuộc lệch khỏi kỳ vọng toán, vậy:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \qquad (3.3)$$

β_1 và β_2 : các tham số của mô hình

β_1 : tung độ gốc

β_2: độ dốc

Giá trị ước lượng của Y_i

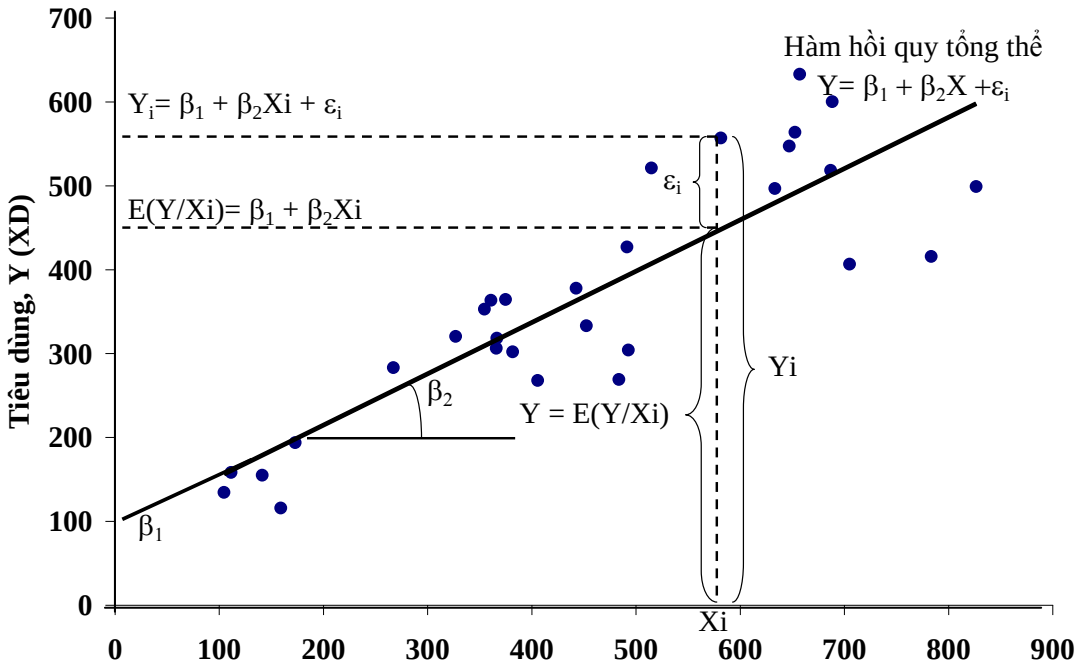
$$\hat{Y}_i = \beta_1 + \beta_2 X_i$$

ε_i : Sai số của hồi quy hay còn được gọi là nhiễu ngẫu nhiên

Nhiễu ngẫu nhiên hình thành từ nhiều nguyên nhân:

- Bỏ sót biến giải thích.
- Sai số khi đo lường biến phụ thuộc.
- Các tác động không tiên đoán được.
- Dạng hàm hồi quy không phù hợp.

Dạng hàm hồi quy (3.2) được gọi là hồi quy tổng thể tuyến tính. Chúng ta sẽ thảo luận chi tiết về thuật ngữ hồi quy tuyến tính ở cuối chương. Hình 3.2 cho ta cái nhìn trực quan về hồi quy tổng thể tuyến tính và sai số của hồi quy.



Hình 3.2. Hàm hồi quy tổng thể tuyến tính

3.2.2.Hàm hồi quy mẫu (SRF)

Trong thực tế hiếm khi chúng có số liệu của tổng thể mà chỉ có số liệu mẫu. Chúng ta phải sử dụng dữ liệu mẫu để ước lượng hàm hồi quy tổng thể.

Hàm hồi quy mẫu:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad (3.4)$$

Trong đó

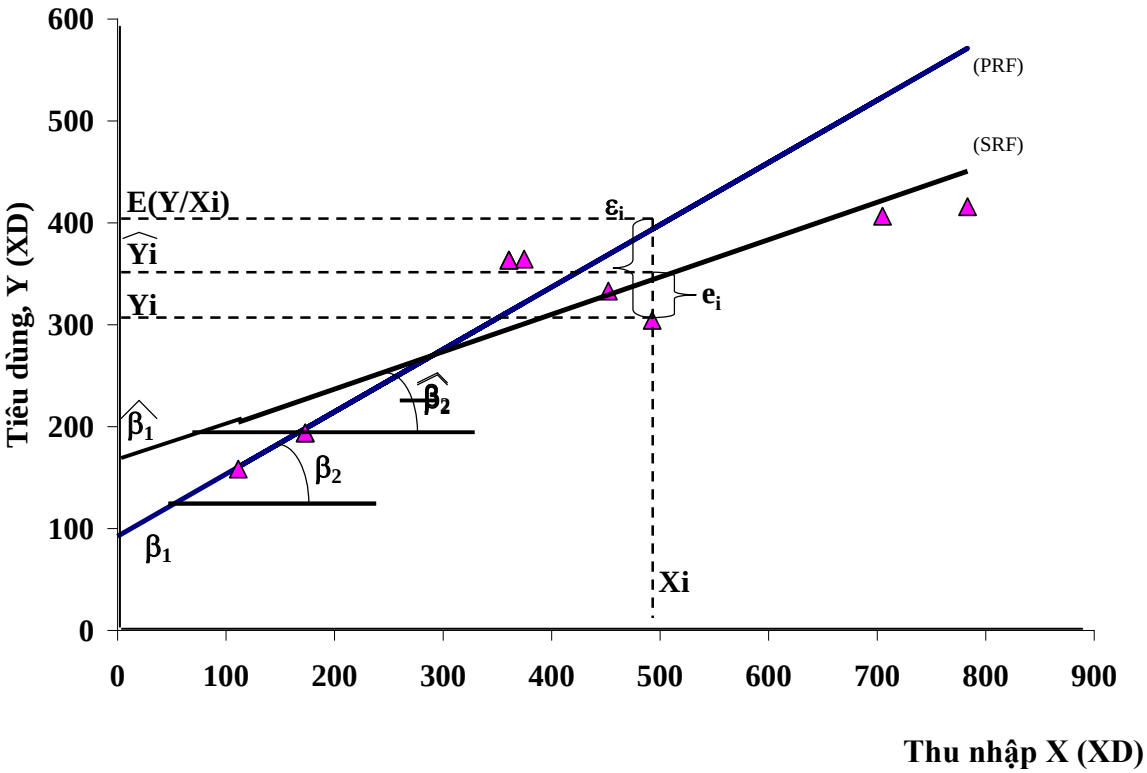
$\hat{\beta}_1$: ước lượng cho β_1.

$\hat{\beta}_2$: Ước lượng cho β_2.

Đối với quan sát thứ i :

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \qquad (3.5)$$

Hình 3.3 cho thấy sự xấp xỉ của hàm hồi quy mẫu (SRF) và hàm hồi quy tổng thể (PRF).



Hình 3.3. Hồi quy mẫu và hồi quy tổng thể

3.3.Ước lượng các hệ số của mô hình hồi quy theo phương pháp bình phương tối thiểu-OLS[11]

3.3.1.Các giả định của mô hình hồi quy tuyến tính cổ điển

Các giả định về sai số hồi quy như sau đảm bảo cho các ước lượng hệ số hàm hồi quy tổng thể dựa trên mẫu theo phương pháp bình phương tối thiểu là ước lượng tuyến tính không chệch tốt nhất(BLUE).

Giá trị kỳ vọng bằng 0: $E[\varepsilon_i | X_i] = 0$

Phương sai không đổi: $\text{var}[\varepsilon_i | X_i] = E[\varepsilon_i^2 | X_i] = \sigma^2$

Không tự tương quan: $\text{cov}[\varepsilon_i \varepsilon_j | X_i, X_j] = E[\varepsilon_i \varepsilon_j | X_i, X_j] = 0$

Không tương quan với X: $\text{cov}[\varepsilon_i X_j | X_i, X_j] = E[\varepsilon_i X_j | X_i, X_j] = 0$

Có phân phối chuẩn: $\varepsilon_i = N(0, \sigma^2)$

Ở chương 5 chúng ta sẽ khảo sát hậu quả khi các giả thiết trên bị vi phạm.

[11] OLS-Ordinary Least Square

3.3.2.Phương pháp bình phương tối thiểu:

Ý tưởng của phương pháp bình phương tối thiểu là tìm $\hat{\beta}_1$ và $\hat{\beta}_2$ sao cho tổng bình phương phần dư có giá trị nhỏ nhất.

Từ hàm hồi quy (3.5)

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

Vậy $$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2 \qquad (3.6)$$

Điều kiện để (3.6) đạt cực trị là:

(1) $$\frac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) = -2\sum_{i=1}^{n} e_i = 0 \qquad (3.7)$$

(2) $$\frac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)X_i = -2\sum_{i=1}^{n} e_i X_i = 0 \qquad (3.8)$$

Từ (3.7) và (3.8) chúng ta rút ra

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \qquad (3.9)$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \qquad (3.10)$$

Các phương trình (3.9) và (3.10) được gọi là các phương trình chuẩn. Giải hệ phương trình chuẩn ta được

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} \qquad (3.11)$$

Thay (3.9) vào (3.8) và biến đổi đại số chúng ta có

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \qquad (3.12)$$

Đặt $x_i = X_i - \overline{X}$ và $y_i = Y_i - \overline{Y}$ ta nhận được

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \qquad (3.13)$$

3.3.3.Tính chất của hàm hồi quy mẫu theo OLS

Tính chất của tham số ước lượng

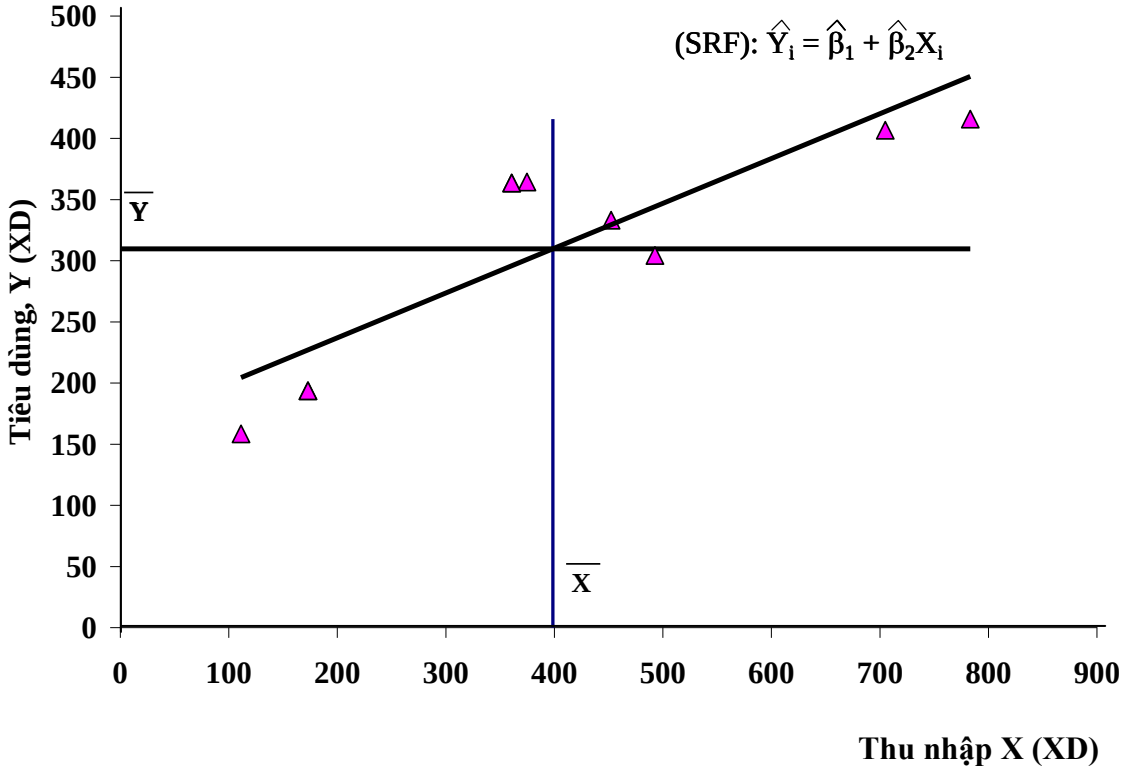
(1) $\hat{\beta}_1$ và $\hat{\beta}_2$ là duy nhất ứng với một mẫu xác định gồm n quan sát (Xi,Yi).

(2) $\hat{\beta}_1$ và $\hat{\beta}_2$ là các ước lượng điểm của β_1 và β_2 . Giá trị của $\hat{\beta}_1$ và $\hat{\beta}_2$ thay đổi theo mẫu dùng để ước lượng.

Tính chất của hàm hồi quy mẫu[12]

(1) Hàm hồi quy mẫu đi qua giá trị trung bình của dữ liệu

Thật vậy, từ (3.11) ta có $\overline{Y} = \hat{\beta}_1 - \hat{\beta}_2 \overline{X}$



Hình 3.4. Đường hồi quy mẫu đi qua giá trị trung bình của dữ liệu

[12] Phần chứng minh các tính chất ở phần này có thể tìm đọc ở Gujarati, Basic Econometrics,3rd Edition, p56-59.

(2) Giá trị trung bình của ước lượng bằng giá trị trung bình của quan sát đối với biến phụ thuộc: $E(\hat{Y}) = \overline{Y}$.

(3) Giá trị trung bình của phần dư bằng 0: $E(e_i) = 0$

(4) Các phần dư e_i và Y_i không tương quan với nhau: $\sum_{i=1}^{n} e_i Y_i = 0$

(5) Các phần dư e_i và X_i không tương quan với nhau: $\sum_{i=1}^{n} e_i X_i = 0$

3.3.4.Phân phối của $\hat{\beta}_1$ và $\hat{\beta}_2$ [13]

Ước lượng	$\hat{\beta}_1$	$\hat{\beta}_2$
Kỳ vọng	$E(\hat{\beta}_1) = \beta_1$	$E(\hat{\beta}_2) = \beta_2$
Phương sai	$\text{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\sigma^2$	$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$
Sai số chuẩn	$\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}}\sigma$	$\sigma_{\hat{\beta}_2} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} x_i^2}}$
Phân phối	$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\sigma^2\right)$	$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$

Hiệp phương sai của hai hệ số ước lượng

$$\text{cov}(\hat{\beta}, \hat{\beta}_2) = -\overline{X}\,\text{var}(\hat{\beta}_2) = -\overline{X}\left(\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$$

Trong các biểu thức trên $\sigma^2 = \text{var}(\varepsilon_i)$ với giả định $\varepsilon_i \sim N(0, \sigma^2)$

[13] Có thể tính toán chứng minh các biểu thức này dựa vào các định nghĩa và định lý về kỳ vọng và phương sai. Tham khảoVũ Thiếu và đồng sự, Kinh tế lượng, PL chương 2, trang 61.

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

3.4.1. Khoảng tin cậy cho các hệ số hồi quy

Thực sự chúng ta không biết σ^2 nên ta dùng ước lượng không chệch của nó là

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Sai số chuẩn của hệ số hồi quy cho độ dốc $se(\beta_2) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} x_i^2}}$

Từ $\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$ với $\sigma_{\hat{\beta}_2} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}}$ ta có

$$Z = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\beta_2}} \sim N(0,1) \tag{3.14}$$

Từ tính chất của phương sai mẫu ta có

$$(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)} \tag{3.15}$$

Từ (3.14) và (3.15) Ta xây dựng trị thống kê

$$\frac{\frac{\hat{\beta}_2 - \beta_2}{\sigma_{\beta_2}}}{\sqrt{\frac{(n-2)\frac{\hat{\sigma}^2}{\sigma^2}}{n-2}}} \sim \frac{Z}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}} \sim t_{(n-2)} \tag{3.16}$$

Biến đổi vế trái chúng ta được

$$\frac{\frac{\hat{\beta}_2 - \beta_2}{\sigma_{\beta_2}}}{\sqrt{\frac{(n-2)\frac{\hat{\sigma}^2}{\sigma^2}}{n-2}}} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}\sigma^2_{\beta_2}}} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2} * \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}}} = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$$

Thay vào (3.16) ta được

$$\frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \sim t_{(n-2)} \tag{3.17}$$

Chứng minh tương tự ta có

$$\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{(n-2)} \tag{3.18}$$

Ước lượng khoảng cho hệ số hồi quy với mức ý nghĩa α như sau

$$\hat{\beta}_1 - t_{(n-2,1-\alpha/2)} se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{(n-2,1-\alpha/2)} se(\hat{\beta}_1) \tag{3.19}$$

$$\hat{\beta}_2 - t_{(n-2,1-\alpha/2)} se(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{(n-2,1-\alpha/2)} se(\hat{\beta}_2) \tag{3.20}$$

3.4.2. Kiểm định giả thiết về hệ số hồi quy

Chúng ta quan tâm nhiều đến ý nghĩa thống kê độ dốc (β_2) của phương trình hồi quy hơn là tung độ gốc (β_1). Cho nên từ đây đến cuối chương chủ yếu chúng ta kiểm định giả thiết thống kê về độ dốc.

Giả thiết

$$H_0 : \beta_2 = \beta_2^*$$

$$H_1 : \beta_2 \neq \beta_2^*$$

Phát biểu mệnh đề xác suất

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

Quy tắc quyết định

- Nếu $\frac{\hat{\beta}_2 - \beta_2^*}{se(\hat{\beta}_2)} < t_{(n-2,\alpha/2)}$ hoặc $\frac{\hat{\beta}_2 - \beta_2^*}{se(\hat{\beta}_2)} > t_{(n-2,1-\alpha/2)}$ thì bác bỏ H_0.

- Nếu $t_{(n-2,\alpha/2)} \le \frac{\hat{\beta}_2 - \beta_2^*}{se(\hat{\beta}_2)} \le t_{(n-2,1-\alpha/2)}$ thì ta không thể bác bỏ H_0.

Quy tắc thực hành-Trị thống kê t trong các phần mềm kinh tế lượng

Trong thực tế chúng ta thường xét xem biến độc lập X có tác động lên biến phụ thuộc Y hay không. Vậy khi thực hiện hồi quy chúng ta kỳ vọng $\beta_2 \neq 0$. Mức ý nghĩa hay được dùng trong phân tích hồi quy là α=5%.

Giả thiết

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Trị thống kê trở thành

$$\text{t-stat} = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)}$$

Quy tắc quyết định

- Nếu /t-stat/ > $t_{(n-2,97,5\%)}$ thì bác bỏ H_0.
- Nếu /t-stat/ ≤ $t_{(n-2,97,5\%)}$ thì không thể bác bỏ H_0.

Tra bảng phân phối Student chúng ta thấy khi bậc tự do n trên 20 thì trị thống kê $t_{97,5\%}$ thì xấp xỉ 2.

Quy tắc thực hành

- Nếu /t-stat/ > 2 thì bác bỏ giả thiết $\beta_2 = 0$.
- Nếu /t-stat/≤ 2 thì ta không thể bác bỏ giả thiết β_2=0.

Trong các phần mềm bảng tính có tính toán hồi quy, người ta mặc định mức ý nghĩa α=5% và giả thiết H_0: β_i=0. Thủ tục tính toán hồi quy của Excel cung cấp cho ta các hệ số hồi quy, trị thống kê t, ước lượng khoảng của hệ số hồi quy và giá trị p[14].Sau đây là kết quả hồi quy được tính toán bằng thủ tục hồi quy của một vài phần mềm thông dụng.

Excel

Kết quả Regresstion cho dữ liệu của ví dụ 3.1. (Chỉ trích phần hệ số hồi quy)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	92,24091128	33,61088673	2,744376012	0,010462	23,39205354	161,089769
X	0,611539034	0,067713437	9,031280327	8,68E-10	0,472834189	0,750243878

Intercept: Tung độ gốc

Coefficients : Hệ số hồi quy

Standard Error : Sai số chuẩn của ước lượng hệ số

t Stat : Trị thống kê $t_{(n-2)}$

P-value : Giá trị p

[14] Ở chương 2 chúng ta đã biết ước kiểm định trên ước lượng khoảng, trị thống kê và giá trị p là tương đương nhau.

Lower95%: Giá trị tới hạn dưới của khoảng ước lượng với độ tin cậy 95%.

Upper95% : Giá trị tới hạn trên của khoảng ước lượng với độ tin cậy 95%.

Bác bỏ H_0 khi /t-stat/ > 2 hoặc p-value < 0,05 hoặc khoảng (Lower;Upper) không chứa 0.[15]

Eviews

Thủ tục Make Equation cho kết quả như sau(chỉ trích phần hệ số hồi quy):

Dependent Variable: Y

Method: Least Squares

Included observations: 30 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	92.24091	33.61089	2.744376	0.0105
X	0.611539	0.067713	9.031280	0.0000

C : Tung độ gốc

Coefficient : Hệ số hồi quy

Std. Error : Sai số chuẩn của ước lượng hệ số

t – Statistic : Trị thống kê $t_{(n-2)}$

Prob: Giá trị p. Bác bỏ H_0 khi /t-Statistic/ > 2 hoặc Prob < 0,05.

SPSS

Thủ tục Regression->Linear. (Chỉ trích phần hệ số hồi quy).

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1(Constant)	92,241	33,611		2,744	,010
X	,612	,068	,863	9,031	,000

Constant: Tung độ gốc

Unstandardized Coefficients: Các hệ số hồi quy

[15] Như đã trình bày ở chương 2, đây thực ra là 3 cách diễn đạt từ một mệnh đề xác suất nên kết luận từ 3 trị thống kê t, p và ước lượng khoảng là tương đương nhau.

Standardized Coefficients: Các hệ số hồi quy chuẩn hoá[16].

t: t-Stat Sig: Giá trị p.

Bác bỏ H_0 khi /t/ >2 hoặc Sig < 0,05

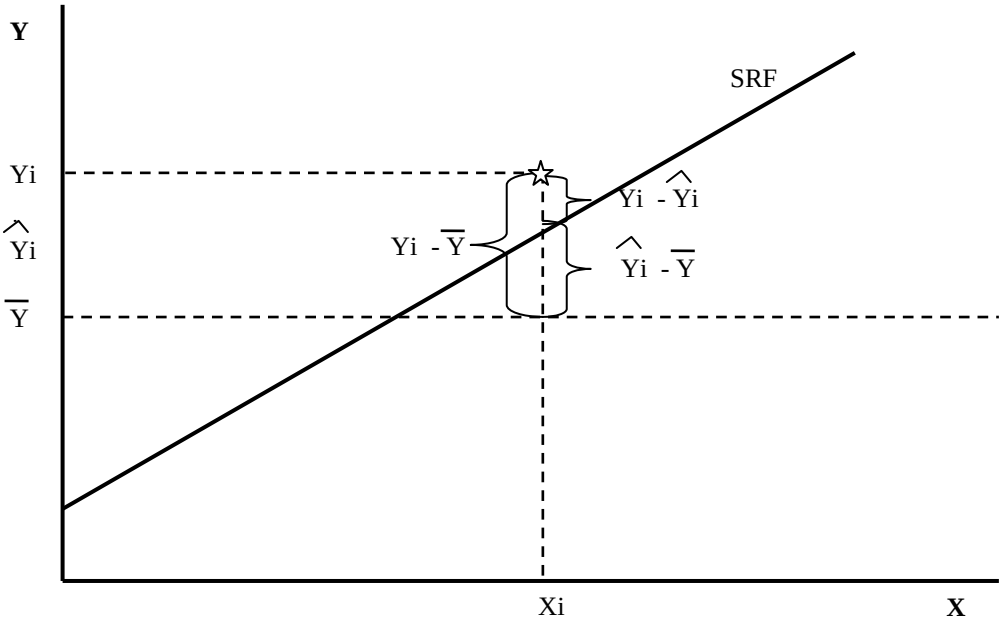
3.5.Định lý Gauss-Markov

Với các giả định của mô hình hồi quy tuyến tính cổ điển, hàm hồi quy tuyến tính theo phương pháp bình phương tối thiểu là ước lượng tuyến tính không thiên lệch tốt nhất.

Chúng ta sẽ không chứng minh đinh lý này.[17]

3.6.Độ thích hợp của hàm hồi quy – R^2

Làm thế nào chúng ta đo lường mức độ phù hợp của hàm hồi quy tìm được cho dữ liệu mẫu. Thước đo độ phù hợp của mô hình đối với dữ liệu là R^2. Để có cái nhìn trực quan về R^2, chúng ta xem xét đồ thị sau



Hình 3.5. Phân tích độ thích hợp của hồi quy

$Y_i - \overline{Y}$: biến thiên của biến phụ thuộc Y, đo lường độ lệch của giá trị Yi so với giá trị trung bình $\overline{Y}$.

$\hat{Y}_i - \overline{Y}$: biến thiên của Y được giải thích bởi hàm hồi quy

[16] Khái niệm này nằm ngoài khuôn khổ của giáo trình.

[17] Phần chứng minh các tính chất ở phần này có ở Gujarati, Basic Econometrics-3rd Edition, trang 97-98.

$e_i = Y_i - \hat{Y}_i$: biến thiên của Y không giải thích được bởi hàm hồi quy hay sai số hồi quy.

Trên mỗi Xi chúng ta kỳ vọng e_i nhỏ nhất, hay phần lớn biến thiên của biến phụ thuộc được giải thích bởi biến độc lập. Nhưng một hàm hồi quy tốt phải có tính chất mang tính tổng quát hơn. Trong hồi quy tuyến tính cổ điển, người ta chọn tính chất tổng bình phương biến thiên không giải thích được là nhỏ nhất.

Ta có

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

Với $y_i = Y_i - \overline{Y}$ và $\hat{y}_i = \hat{Y} - \overline{Y}$

$$\text{Vậy} \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} \hat{y}_i e_i \qquad (3.21)$$

Số hạng cuối cùng của (3.21) bằng 0.

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

Đặt $TSS = \sum_{i=1}^{n} y_i^2$, $ESS = \sum_{i=1}^{n} \hat{y}_i^2$ và $RSS = \sum_{i=1}^{n} e_i^2$

TSS(Total Sum of Squares): Tổng bình phương biến thiên của Y.

ESS(Explained Sum of Squares): Tổng bình phương phần biến thiên giải thích được bằng hàm hồi quy của Y.

RSS(Residual Sum of Squares) : Tổng bình phương phần biến thiên không giải thích được bằng hàm hồi quy của Y hay tổng bình phương phần dư.Ta có:

$$TSS = ESS + RSS$$

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

$$R^2 = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \hat{\beta}_2^2 \frac{\left(\sum_{i=1}^{n} x_i^2 \Big/ n-1 \right)}{\left(\sum_{i=1}^{n} y_i^2 \Big/ n-1 \right)} = \hat{\beta}_2^2 \frac{S_x^2}{S_y^2}$$

Mặt khác ta có $\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$ Vậy

$$R^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i \right)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} = r_{X,Y}^2 \qquad (3.22)$$

Vậy đối với hồi quy hai biến R^2 là bình phương của hệ số tương quan.

Tính chất của R^2

(1) $0 \leq R^2 \leq 1$. Với $R^2=0$ thể hiện X và Y độc lập thống kê. $R^2 =1$ thể hiện X và Y phụ thuộc tuyến tính hoàn hảo.

(2) R^2 không xét đến quan hệ nhân quả.

3.7.Dự báo bằng mô hình hồi quy hai biến

Dựa trên X_0 xác định chúng ta dự báo Y_0.

Ước lượng điểm cho Y_0 là : $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$.

Để ước lượng khoảng chúng ta phải tìm phân phối xác suất của $\hat{Y}_i$.

Dự báo giá trị trung bình $E(Y_o | X = X_0)$

Từ $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$

Suy ra $\text{var}(\hat{Y}_0) = \text{var}(\hat{\beta}_1 + \hat{\beta}_2 X_0) = \text{var}(\hat{\beta}_1) + X_0^2 \text{var}(\hat{\beta}_2) + 2X_0 \text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ (3.23)

Thay biểu thức của $\text{var}(\hat{\beta}_1)$, $\text{var}(\hat{\beta}_2)$ và $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ ở mục 3.3.4 vào (3.23) và rút gọn

$$\text{var}\left(\hat{Y}_0\right) = \sigma^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} x_i^2}\right]$$

Dự báo giá trị cụ thể của Y_0

Từ $Y_0 - \hat{Y}_0 = (\beta_1 - \hat{\beta}_1) + (\beta - \hat{\beta}_2)X_0 + e_0$

Ta có $E(Y_0 - \hat{Y}_0) = E(\beta_1 - \hat{\beta}_1) + X_0 E(\beta - \hat{\beta}_2) + E(e_0) = 0$

và $\text{var}(Y_0 - \hat{Y}_0) = \text{var}(\hat{\beta}_1) + X_0^2 \text{var}(\hat{\beta}_2) + 2X_0 \text{cov}(\hat{\beta}_1, \hat{\beta}_2) + \text{var}(e_0)$ (3.25)

Số hạng cuối cùng $\text{var}(e_0) = \sigma^2$. Vậy

$$\text{var}\left(Y_0 - \hat{Y}_0\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} x_i^2}\right] \quad (3.26)$$

Sai số chuẩn của dự báo

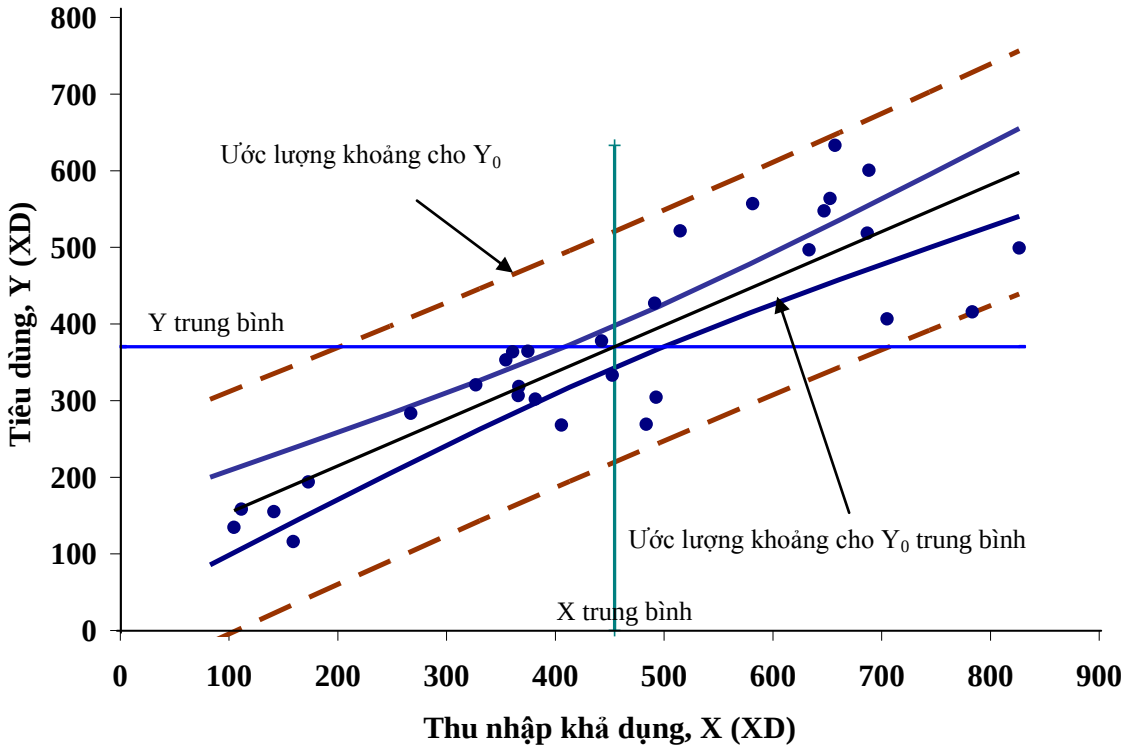
Cho giá trị của Y_0

$$\text{se}\left(\hat{Y}_0\right) = \sigma\left(1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} x_i^2}\right)^{1/2}$$

Khoảng tin cậy cho dự báo

$$\hat{Y}_o \pm t_{(n-2,1-\alpha/2)}\,\text{se}(\hat{Y}_o)$$

Nhận xét: X_0 càng lệch ra khỏi giá trị trung bình thì dự sai số của dự báo càng lớn. Chúng ta sẽ thấy rõ điều này qua đồ thị sau.



Hình 3.6. Ước lượng khoảng cho Y_0.

3.8. Ý nghĩa của hồi quy tuyến tính và một số dạng hàm thường được sử dụng

3.8.1. Tuyến tính trong tham số

Trong mục 3.2.1 chúng ta đã đặt yêu cầu là để ước lượng theo phương pháp bình phương tối thiểu thì mô hình hồi quy phải tuyến tính. Sử dụng tính chất hàm tuyến tính của các phân phối chuẩn cũng là phân phối chuẩn, dựa vào các giả định chặt chẽ và phương pháp bình phương tối thiểu, người ta rút ra các hàm ước lượng tham số hiệu quả và các trị thống kê kiểm định.

Hồi quy tuyến tính chỉ yêu cầu tuyến tính trong các tham số, không yêu cầu tuyến tính trong biến số.

Mô hình $$Y = \beta_1 + \beta_2 \frac{1}{X} + \varepsilon \qquad (3.27)$$

là mô hình tuyến tính trong các tham số nhưng phi tuyến theo biến số.

Mô hình $$Y = \beta_1 + (1 - \beta_1^2)X \qquad (3.28)$$

là mô hình phi tuyến trong các tham số nhưng tuyến tính trong biến số.

Hồi quy tuyến tính theo OLS chấp nhận dạng mô hình tuyến tính trong tham số như (3.27) mà không chấp nhận dạng mô hình phi tuyến trong tham số như (3.28).

3.8.2. Một số mô hình thông dụng

Mô hình Logarit kép

Mô hình logarit kép phù hợp với dữ liệu ở nhiều lĩnh vực khác nhau. Ví dụ đường cầu với độ co dãn không đổi hoặc hàm sản xuất Cobb-Douglas.

Mô hình đường cầu : $Y = \beta_1 X^{\beta_2} e^{\varepsilon}$ (3.29)

Không thể ước lượng mô hình (3.29) theo OLS vì nó phi tuyến trong tham số. Tuy nhiên nếu chúng ta lấy logarit hai vế thì ta được mô hình

$$\ln(Y) = \ln(\beta_1) + \beta_2 X + \varepsilon \quad (3.30)$$

Đặt $Y^* = \ln(Y)$ và $\beta_1^* = \ln(\beta_1)$ ta được mô hình

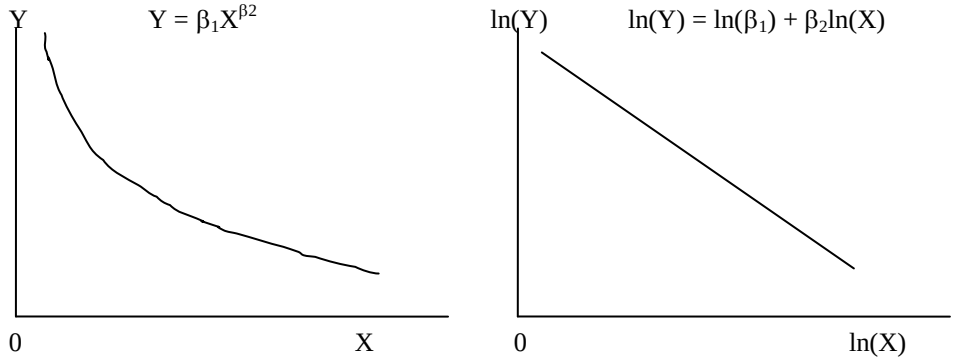
$$Y^* = \beta_1^* + \beta_2 X + \varepsilon \quad (3.31)$$

Mô hình này tuyến tính theo tham số nên có thể ước lượng theo OLS.

Chúng ta sẽ chứng minh đặc tính đáng lưu ý của mô hình này là độ co dãn cầu theo giá không đổi. Định nghĩa độ co dãn: $\eta_D = \dfrac{\partial Y / Y}{\partial X / X} = \dfrac{\partial Y}{\partial X} * \dfrac{X}{Y}$

Lấy vi phân hai vế của (3.30) ta có $\dfrac{\partial Y}{Y} = \beta_2 \dfrac{\partial X}{X} \Rightarrow \eta_D = \dfrac{\partial Y}{\partial X}\dfrac{X}{Y} = \beta_2$

Vậy độ co dãn của cầu theo giá không đổi.



Hình 3.8. Chuyển dạng Log-log

Tổng quát, đối với mô hình logarit kép, hệ số ứng với ln của một biến số độc lập là độ co dãn của biến phụ thuộc vào biến độc lập đó.

Mô hình Logarit-tuyến tính hay mô hình tăng trưởng

Gọi g là tốc độ tăng trưởng, t chỉ thời kỳ. Mô hình tăng trưởng như sau

$$Y_t = (1 + g)^t Y_0 \quad (3.32)$$

Lấy logarit hai vế của (3.32)

$$\ln(Y_t) = t \ln(1 + g) + \ln(Y_0) \quad (3.33)$$

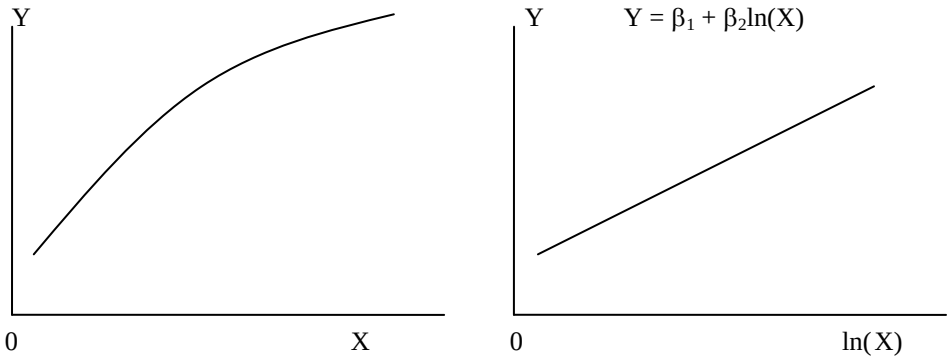
Đặt $Y_t^* = \ln(Y_t)$, $\beta_1 = \ln(Y_0)$ và $\beta_2 = \ln(1 + g)$ ta được mô hình hồi quy

$$Y_t^* = \beta_1 + \beta_2 t + \varepsilon \quad (3.34)$$

Mô hình tuyến tính-Logarit (Lin-log)

$$Y = \beta_1 + \beta_2 \ln(X) + \varepsilon \quad (3.35)$$

Mô hình này phù hợp với quan hệ thu nhập và tiêu dùng của một hàng hoá thông thường với Y là chi tiêu cho hàng hoá đó và X là thu nhập. Quan hệ này cho thấy Y tăng theo X nhưng tốc độ tăng chậm dần.

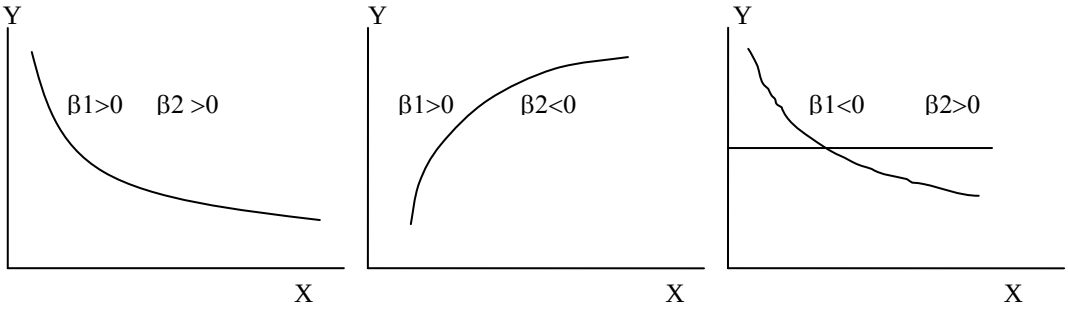


Hình 3.9. Chuyển dạng Lin-log

Mô hình nghịch đảo hay mô hình Hyperbol

$$Y = \beta_1 + \beta_2 \frac{1}{X} + \varepsilon \quad (3.36)$$

Mô hình này phù hợp cho nghiên cứu đường chi phí đơn vị, đường tiêu dùng theo thu nhập Engel hoặc đường cong Philip.



Đường chi phí đơn vị Đường tiêu dùng Đường Philip

Hình 3.10. Dạng hàm nghịch đảo

Phụ lục 3.1.PL Số liệu về thu nhập và tiêu dùng, XD.

STT	Thu nhập khả dụng X	Tiêu dùng Y
1	173	194
2	361	363
3	355	353
4	366	306
5	581	557
6	382	302
7	633	497
8	406	268
9	375	364
10	267	283
11	783	416
12	515	521
13	705	407
14	493	304
15	367	318
16	159	116
17	492	427
18	827	499
19	111	158
20	452	333
21	688	600
22	327	320
23	647	547
24	687	518
25	443	378
26	657	633
27	105	134
28	484	269
29	653	564
30	141	155

CHƯƠNG 4

MÔ HÌNH HỒI QUY TUYẾN TÍNH BỘI

4.1. Xây dựng mô hình

4.1.1. Giới thiệu

Mô hình hồi quy hai biến mà chúng ta đã nghiên cứu ở chương 3 thường không đủ khả năng giải thích hành vi của biến phụ thuộc. Ở chương 3 chúng ta nói tiêu dùng phụ thuộc vào thu nhập khả dụng, tuy nhiên có nhiều yếu tố khác cũng tác động lên tiêu dùng, ví dụ độ tuổi, mức độ lạc quan vào nền kinh tế, nghề nghiệp… Vì thế chúng ta cần bổ sung thêm biến giải thích(biến độc lập) vào mô hình hồi quy. Mô hình với một biến phụ thuộc với hai hoặc nhiều biến độc lập được gọi là hồi quy bội.

Chúng ta chỉ xem xét hồi quy tuyến tính bội với mô hình tuyến tính với trong tham số, không nhất thiết tuyến tính trong biến số.

Mô hình hồi quy bội cho tổng thể

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + ... + \beta_k X_{k,i} + \varepsilon_i \qquad (4.1)$$

Với $X_{2,i}, X_{3,i},\ldots,X_{k,i}$ là giá trị các biến độc lập ứng với quan sát i

$\beta_2, \beta_2, \beta_3,\ldots, \beta_k$ là các tham số của hồi quy

ε_i là sai số của hồi quy

Với một quan sát i, chúng ta xác định giá trị kỳ vọng của Yi

$$E\left[Y \middle| X's\right] = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + ... + \beta_k X_{k,i} \qquad (4.2)$$

4.1.2. Ý nghĩa của tham số

Các hệ số β được gọi là các hệ số hồi quy riêng

$$\frac{\partial\left[Y \middle| X's\right]}{\partial X_m} = \beta_m \qquad (4.3)$$

β_k đo lường tác động riêng phần của biến X_m lên Y với điều kiện các biến số khác trong mô hình không đổi. Cụ thể hơn nếu các biến khác trong mô hình không đổi, giá trị kỳ vọng của Y sẽ tăng β_m đơn vị nếu X_m tăng 1 đơn vị.

4.1.3. Giả định của mô hình

Sử dụng các giả định của mô hình hồi quy hai biến, chúng ta bổ sung thêm giả định sau:

(1) Các biến độc lập của mô hình không có sự phụ thuộc tuyến tính hoàn hảo, nghĩa là không thể tìm được bộ số thực $(\lambda_1,\lambda_2,...,\lambda_k)$ sao cho

$$\lambda_1 + \lambda_2 X_{2,i} + \lambda_3 X_{3,i} + ... + \lambda_k X_{k,i} = 0 \quad \text{với mọi i.}$$

Giả định này còn được được phát biểu là " không có sự đa cộng tuyến hoàn hảo trong mô hình".

(2) Số quan sát n phải lớn hơn số tham số cần ước lượng k.

(3) Biến độc lập X_i phải có sự biến thiên từ quan sát này qua quan sát khác hay $Var(X_i)>0$.

4.2. Ước lượng tham số của mô hình hồi quy bội

4.2.1. Hàm hồi quy mẫu và ước lượng tham số theo phương pháp bình phương tối thiểu

Trong thực tế chúng ta thường chỉ có dữ liệu từ mẫu. Từ số liệu mẫu chúng ta ước lượng hồi quy tổng thể.

Hàm hồi quy mẫu

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2,i} + \hat{\beta}_3 X_{3,i} + ... + \hat{\beta}_k X_{k,i} + e_i \qquad (4.4)$$

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2,i} - \hat{\beta}_3 X_{3,i} - ... - \hat{\beta}_k X_{k,i}$$

Với các $\hat{\beta}_m$ là ước lượng của tham số β_m. Chúng ta trông đợi $\hat{\beta}_m$ là ước lượng không chệch của β_m, hơn nữa phải là một ước lượng hiệu quả. Với một số giả định chặt chẽ như ở mục 3.3.1 chương 3 và phần bổ sung ở 4.1, thì phương pháp tối thiểu tổng bình phương phần dư cho kết quả ước lượng hiệu quả β_m.

Phương pháp bình phương tối thiểu

Chọn $\beta_1, \beta_2, \ldots, \beta_k$ sao cho

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2,i} - \hat{\beta}_3 X_{3,i} - ... - \hat{\beta}_k X_{k,i}\right)^2 \qquad (4.5)$$

đạt cực tiểu.

Điều kiện cực trị của (4.5)

$$\left.\begin{aligned}
&\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_1} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2,i} - \hat{\beta}_3 X_{3,i} - ... - \hat{\beta}_K X_{K,i}\right) = 0 \\
&\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_2} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2,i} - \hat{\beta}_3 X_{3,i} - ... - \hat{\beta}_K X_{K,i}\right)X_{2,i} = 0 \\
&... \\
&\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_k} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2,i} - \hat{\beta}_3 X_{3,i} - ... - \hat{\beta}_K X_{K,i}\right)X_{k,i} = 0
\end{aligned}\right\} \quad (4.6)$$

Hệ phương trình (4.6) được gọi là hệ phương trình chuẩn của hồi quy mẫu (4.4).

Cách giải hệ phương trình (4.4) gọn gàng nhất là dùng ma trận. Do giới hạn của chương trình, bài giảng này không trình bày thuật toán ma trận mà chỉ trình bày kết quả tính toán cho hồi quy bội đơn giản nhất là hồi quy ba biến với hai biến độc lập. Một số tính chất của hồi quy ta thấy được ở hồi quy hai biến độc lập có thể áp dụng cho hồi quy bội tổng quát.

4.2.2. Ước lượng tham số cho mô hình hồi quy ba biến

Hàm hồi quy tổng thể

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i \quad (4.7)$$

Hàm hồi quy mẫu

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2,i} + \hat{\beta}_3 X_{3,i} + e_i \quad (4.8)$$

Nhắc lại các giả định

(1) Kỳ vọng của sai số hồi quy bằng 0: $E\left(e_i \middle| X_{2,i}, X_{3,i}\right) = 0$

(2) Không tự tương quan: $\text{cov}\left(e_i, e_j\right) = 0$, i≠j

(3) Phương sai đồng nhất: $\text{var}\left(e_i\right) = \sigma^2$

(4) Không có tương quan giữa sai số và từng X_m: $\text{cov}\left(e_i, X_{2,i}\right) = \text{cov}\left(e_i, X_{3,i}\right) = 0$

(5) Không có sự đa cộng tuyến hoàn hảo giữa X_2 và X_3.

(6) Dạng hàm của mô hình được xác định một cách đúng đắn.

Với các giả định này, dùng phương pháp bình phương tối thiểu ta nhận được ước lượng các hệ số như sau.

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3 \tag{4.10}$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} y_i x_{2,i}\right)\left(\sum_{i=1}^{n} x_{3,i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{3,i}\right)\left(\sum_{i=1}^{n} x_{2,i} x_{3,i}\right)}{\left(\sum_{i=1}^{n} x_{2,i}^2\right)\left(\sum_{i=1}^{n} x_{3,i}^2\right) - \left(\sum_{i=1}^{n} x_{2,i} x_{3,i}\right)^2} \tag{4.11}$$

$$\hat{\beta}_3 = \frac{\left(\sum_{i=1}^{n} y_i x_{3,i}\right)\left(\sum_{i=1}^{n} x_{2,i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{2,i}\right)\left(\sum_{i=1}^{n} x_{2,i} x_{3,i}\right)}{\left(\sum_{i=1}^{n} x_{2,i}^2\right)\left(\sum_{i=1}^{n} x_{3,i}^2\right) - \left(\sum_{i=1}^{n} x_{2,i} x_{3,i}\right)^2} \tag{4.12}$$

4.2.3. Phân phối của ước lượng tham số

Trong phần này chúng ta chỉ quan tâm đến phân phối của các hệ số ước lượng $\hat{\beta}_2$ và $\hat{\beta}_3$. Hơn nữa vì sự tương tự trong công thức xác định các hệ số ước lượng nên chúng ta chỉ khảo sát $\hat{\beta}_2$. Ở đây chỉ trình bày kết quả[18].

$\hat{\beta}_2$ là một ước lượng không chệch : $E(\hat{\beta}_2) = \beta_2$ (4.13)

$$\text{var}(\hat{\beta}_2) = \frac{\sum_{i=1}^{n} x_{3,i}^2}{\left(\sum_{i=1}^{n} x_{2,i}^2\right)\left(\sum_{i=1}^{n} x_{3,i}^2\right) - \left(\sum_{i=1}^{n} x_{2,i} x_{3,i}\right)^2} \sigma^2 \tag{4.14}$$

Nhắc lại hệ số tương quan giữa X_2 và X_3 : $r_{X_2X_3} = \dfrac{\sum_{i=1}^{n} x_{2,i} x_{3,i}}{\sqrt{\left(\sum_{i=1}^{n} x_{2,i}^2\right)}\sqrt{\left(\sum_{i=1}^{n} x_{3,i}^2\right)}}$

Đặt $r_{X_2X_3} = r_{23}$ biến đổi đại số (4.14) ta được

$$\text{var}(\hat{\beta}_2) = \frac{1}{\sum_{i=1}^{n} x_{2,i}^2 (1 - r_{23}^2)} \sigma^2 \tag{4.15}$$

[18] Các thao tác chứng minh khá phức tạp, để tự chứng minh độc giả hãy nhớ lại các định nghĩa và tính chất của giá trị kỳ vọng, phương sai và hiệp phương sai của biến ngẫu nhiên.

Từ các biểu thức (4.13) và (4.15) chúng ta có thể rút ra một số kết luận như sau:

(1) Nếu X2 và X3 có tương quan tuyến tính hoàn hảo thì r_{23}^2 =1. Hệ quả là $\text{var}(\hat{\beta}_2)$ vô cùng lớn hay ta không thể xác định được hệ số của mô hình hồi quy.

(2) Nếu X2 và X3 không tương quan tuyến tính hoàn hảo nhưng có tương quan tuyến tính cao thì ước lượng $\hat{\beta}_2$ vẫn không chệch nhưng không hiệu quả.

Những nhận định trên đúng cho cả hồi quy nhiều hơn ba biến.

4.3. R^2 và R^2 hiệu chỉnh

Nhắc lại khái niệm về R^2: $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$

Một mô hình có R^2 lớn thì tổng bình phương sai số dự báo nhỏ hay nói cách khác độ phù hợp của mô hình đối với dữ liệu càng lớn. Tuy nhiên một tính chất đặc trưng quan trọng của là nó có xu hướng tăng khi số biến giải thích trong mô hình tăng lên. Nếu chỉ đơn thuần chọn tiêu chí là chọn mô hình có R^2 cao, người ta có xu hướng đưa rất nhiều biến độc lập vào mô hình trong khi tác động riêng phần của các biến đưa vào đối với biến phụ thuộc không có ý nghĩa thống kê.

Để hiệu chỉnh phạt việc đưa thêm biến vào mô hình, người ra đưa ra trị thống kê R^2 hiệu chỉnh(Adjusted R^2)[19]

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \qquad (4.16)$$

Với n là số quan sát và k là số hệ số cần ước lượng trong mô hình.

Qua thao tác hiệu chỉnh này thì chỉ những biến thực sự làm tăng khả năng giải thích của mô hình mới xứng đáng được đưa vào mô hình.

4.4. Kiểm định mức ý nghĩa chung của mô hình

Trong hồi quy bội, mô hình được cho là không có sức mạnh giải thích khi toàn bộ các hệ số hồi quy riêng phần đều bằng không.

Giả thiết

H_0: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$

H_1: Không phải tất cả các hệ số đồng thời bằng không.

[19] Công thức của Theil, được sử dụng ở đa số các phần mềm kinh tế lượng. Một công thức khác do Goldberger đề xuất là Modified $R^2 = \left(1 - \frac{k}{n}\right)R^2$. (Theo Gujarati, Basic Econometrics-3rd, trang 208).

Trị thống kê kiểm định H_0:

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

Quy tắc quyết định

- Nếu $F_{tt} > F_{(k-1,n-k,\alpha)}$ thì bác bỏ H_0.
- Nếu $F_{tt} \leq F_{(k-1,n-k,\alpha)}$ thì không thể bác bỏ H_0.

4.5. **Quan hệ giữa R^2 và F**

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

$$= \frac{(n-k)ESS/TSS}{(k-1)(1-ESS/TSS)} = \frac{(n-k)R^2}{(k-1)(1-R^2)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

4.6. Ước lượng khoảng và kiểm định giả thiết thống kê cho hệ số hồi quy

Ước lượng phương sai của sai số

$$s_\varepsilon^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k} \qquad (4.17)$$

Người ta chứng minh được s_ε^2 là ước lượng không chệch của σ^2, hay $E(s_\varepsilon^2) = \sigma^2$.

Nếu các sai số tuân theo phân phối chuẩn thì $\frac{(n-k)s_\varepsilon^2}{\sigma^2} \sim \chi^2_{(n-k)}$.

Ký hiệu $s.e(\hat{\beta}_m) = s_{\hat{\beta}_m} = \hat{\sigma}_{\hat{\beta}_m}$. Ta có trị thống kê $\frac{\hat{\beta}_m - \beta_m}{s.e(\hat{\beta}_m)} \sim t_{(n-k)}$

Ước lượng khoảng cho β_m với mức ý nghĩa α là

$$\hat{\beta}_m - t_{(n-k,1-\alpha/2)}s.e(\hat{\beta}_m) \leq \beta_m \leq \hat{\beta}_m + t_{(n-k,1-\alpha/2)}s.e(\hat{\beta}_m) \qquad (4.18)$$

Thông thường chúng ta muốn kiểm định giả thiết H_0 là biến X_m không có tác động riêng phần lên Y.

H_0: $\beta_m = 0$

H_1: $\beta_m \neq 0$

Quy tắc quyết định

- Nếu /t-stat/ > $t_{(n-k,\alpha/2)}$ thì ta bác bỏ H_0.
- Nếu /t-stat/≤ $t_{(n-k,\alpha/2)}$ thì ta không thể bác bỏ H_0.

4.7. Biến phân loại (Biến giả-Dummy variable)

Trong các mô hình hồi quy mà chúng ta đã khảo sát từ đầu chương 3 đến đây đều dựa trên biến độc lập và biến phụ thuộc đều là biến định lượng. Thực ra mô hình hồi quy cho phép sử dụng biến độc lập và cả biến phụ thuộc là biến định tính. Trong giới hạn chương trình chúng ta chỉ xét biến phụ thuộc là biến định lượng. Trong phần này chúng ta khảo sát mô hình hồi quy có biến định tính.

Đối với biến định tính chỉ có thể phân lớp, một quan sát chỉ có thể rơi vào một lớp. Một số biến định tính có hai lớp như:

Biến định tính	Lớp 1	Lớp 2
Giới tính	Nữ	Nam
Vùng	Thành thị	Nông thôn
Tôn giáo	Có	Không
Tốt nghiệp đại học	Đã	Chưa

Bảng 4.1. Biến nhị phân

Người ta thường gán giá trị 1 cho một lớp và giá trị 0 cho lớp còn lại. Ví dụ ta ký hiệu S là giới tính với S =1 nếu là nữ và S = 0 nếu là nam.

Các biến định tính được gán giá trị 0 và 1 như trên được gọi là biến giả(dummy variable), biến nhị phân, biến phân loại hay biến định tính.

4.7.1. Hồi quy với một biến định lượng và một biến phân loại

Ví dụ 4.1. Ở ví dụ này chúng ta hồi quy tiêu dùng cho gạo theo quy mô hộ có xem xét hộ đó ở thành thị hay nông thôn.

Mô hình kinh tế lượng như sau:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \varepsilon i \qquad (4.19)$$

Y: Chi tiêu cho gạo, ngàn đồng/năm

X : Quy mô hộ gia đình, người

D: Biến phân loại, D = 1 nếu hộ ở thành thị, bằng D = 0 nếu hộ ở nông thôn.

Chúng ta muốn xem xét xem có sự khác biệt trong tiêu dùng gạo giữa thành thị và nông thôn hay không ứng với một quy mô hộ gia đình Xi xác định.

Đối với hộ ở nông thôn

$$E\left[Y_i \middle| X_i, D_i = 0\right] = \beta_1 + \beta_2 X_i \qquad (4.20)$$

Đối với hộ ở thành thị

$$E\left[Y_i \middle| X_i, D_i = 1\right] = (\beta_1 + \beta_3) + \beta_2 X_i \qquad (4.21)$$

Vậy sự chênh lệch trong tiêu dùng gạo giữa thành thị và nông thôn như sau

$$E\left[Y_i \middle| X_i, D_i = 1\right] - E\left[Y_i \middle| X_i, D_i = 0\right] = \beta_3 \qquad (4.22)$$

Sự khác biệt trong tiêu dùng gạo giữa thành thị và nông thôn chỉ có ý nghĩa thống kê khi β_3 khác không có ý nghĩa thống kê.

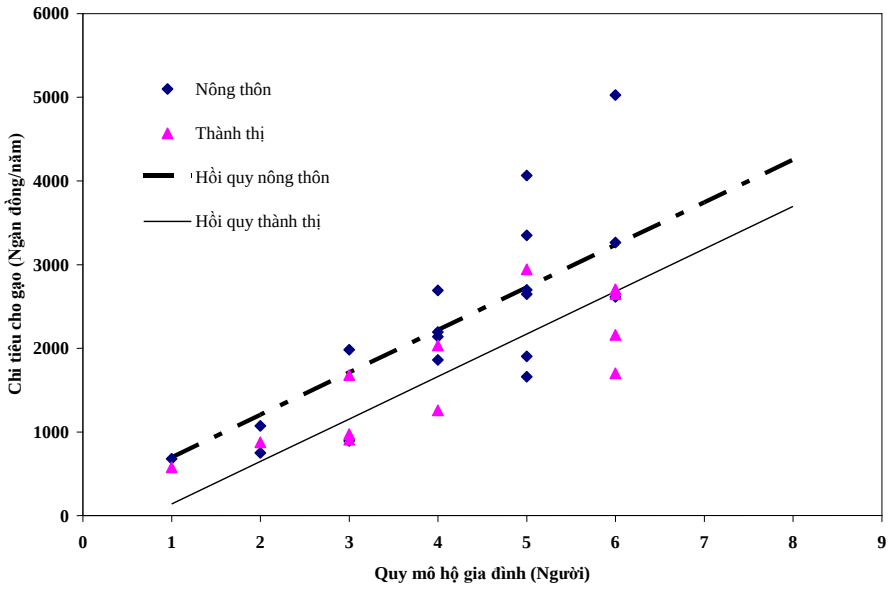
Chúng ta đã có phương trình hồi quy như sau

Y	=	187	+	508*X - 557*D	(4.23)
t-stat		[0,5]		[6,4] [-2,2]	

R^2 hiệu chỉnh = 0,61

Hệ số hồi quy $\hat{\beta}_3 = -557$ khác không với độ tin cậy 95%. Vậy chúng ta không thể bác bỏ được sự khác biệt trong tiêu dùng gạo giữa thành thị và nông thôn.

Chúng ta sẽ thấy tác động của làm cho tung độ gốc của phương trình hồi quy của thành thị và nông thôn sai biệt nhau một khoảng β_3 = -557 ngàn đồng/năm. Cụ thể ứng với một quy mô hộ gia đình thì hộ ở thành thị tiêu dùng gạo ít hơn hộ ở nông thôn 557 ngàn đồng/năm.Chúng ta sẽ thấy điều này một cách trực quan qua đồ thị sau:



Hình 4.1. Hồi quy với một biến định lượng và một biến phân loại.

4.7.2. Hồi quy với một biến định lượng và một biến phân loại có nhiều hơn hai phân lớp

Ví dụ 4.2. Giả sử chúng ta muốn ước lượng tiền lương được quyết định bởi số năm kinh nghiệm công tác và trình độ học vấn như thế nào.

Gọi Y : Tiền lương

X : Số năm kinh nghiệm

D: Học vấn. Giả sử chúng ta phân loại học vấn như sau : chưa tốt nghiệp đại học, đại học và sau đại học.

Phương án 1:

Di = 0 nếu chưa tốt nghiệp đại học

Di = 1 nếu tốt nghiệp đại học

Di =2 nếu có trình độ sau đại học

Cách đặt biến này đưa ra giả định quá mạnh là phần đóng góp của học vấn vào tiền lương của người có trình độ sau đại học lớn gấp hai lần đóng góp của học vấn đối với người có trình độ đại học. Mục tiêu của chúng ta khi đưa ra biến D chỉ là phân loại nên ta không chọn phương án này.

Phương án 2: Đặt bộ biến giả

D_{1i}	D_{2i}	Học vấn
0	0	Chưa đại học
1	0	Đại học

0	1	Sau đại học

Mô hình hồi quy

$$Y_i = \beta_1 + \beta_2 X + \beta_3 D_{1i} + \beta_4 D_{2i} + \varepsilon_i \quad (4.24)$$

Khai triển của mô hình (4.24) như sau

Đối với người chưa tốt nghiệp đại học

$$E(Y_i) = \beta_1 + \beta_2 X \quad (4.25)$$

Đối với người có trình độ đại học

$$E(Y_i) = (\beta_1 + \beta_3) + \beta_2 X_3 \quad (4.26)$$

Đối với người có trình độ sau đại học

$$E(Y_i) = (\beta_1 + \beta_3 + \beta_4) + \beta_2 X \quad (4.27)$$

4.7.3. Cái bẫy của biến giả

	Số lớp của biến phân loại	Số biến giả
Trong ví dụ 4.1.	2	1
Trong ví dụ 4.2	3	2

Điều gì xảy ra nếu chúng ta xây dựng số biến giả đúng bằng số phân lớp?

Ví dụ 4.3. Xét lại ví dụ 4.1.

Giả sử chúng ta đặt biến giả như sau

D_{1i}	D_{2i}	Vùng
1	0	Thành thị
0	1	Nông thôn

Mô hình hồi quy là

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_{1i} + \beta_4 D_{2i} + \varepsilon i \quad (4.28)$$

Chúng ta hãy xem kết quả hồi quy bằng Excel

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2235,533	0	65535	#NUM!
X	508,1297	80,36980143	6,322396	1,08E-06
D1	-2605,52	0	65535	#NUM!
D2	-2048	0	65535	#NUM!

Kết quả hồi quy rất bất thường và hoàn toàn không có ý nghĩa kinh tế.

Lý do là có sự đa cộng tuyến hoàn hảo giữa D1, D2 và một biến hằng X2 =-1.

$$D_{1i} + D_{2i} + X_2 = 0 \quad \forall i .$$

Hiện tượng đa cộng tuyến hoàn hảo này làm cho hệ phương trình chuẩn không có lời giải. Thực tế sai số chuẩn tiến đến vô cùng chứ không phải tiến đến 0 như kết quả tính toán của Excel. Hiện tượng này được gọi là cái bẩy của biến giả.

Quy tắc: Nếu một biến phân loại có k lớp thì chỉ sử dụng (k-1) biến giả.

4.7.4. Hồi quy với nhiều biến phân loại

Ví dụ 4.4. Tiếp tục ví dụ 4.2. Chúng ta muốn khảo sát thêm có sự phân biệt đối xử trong mức lương giữa nam và nữ hay không.

Đặt thêm biến và đặt lại tên biến

GT_i: Giới tính, 0 cho nữ và 1 cho nam.

TL : Tiền lương

KN: Số năm kinh nghiệm làm việc

ĐH: Bằng 1 nếu tốt nghiệp đại học và 0 cho chưa tốt nghiệp đại học

SĐH: Bằng 1 nếu có trình độ sau đại học và 0 cho chưa.

Mô hình hồi quy $TL_i = \beta_1 + \beta_2 KN_i + \beta_3 ĐH_i + \beta_4 SĐH_i + \beta_5 GT_i + \varepsilon_i$ (4.29)

Chúng ta xét tiền lương của nữ có trình độ sau đại học

$$E(TL_i / SĐH=1 \cap GT=0) = (\beta_1 + \beta_4) + \beta_2 KN_i$$

4.7.5. Biến tương tác

Xét lại ví dụ 4.1. Xét quan hệ giữa tiêu dùng gạo và quy mô hộ gia đình.Để cho đơn giản trong trình bày chúng ta sử dụng hàm toán như sau.

Nông thôn: $Y = \alpha_1 + \beta_1 X$

Thành thị: $Y = \alpha_2 + \beta_2 X$

D : Biến phân loại, bằng 1 nếu hộ ở thành thị và bằng 0 nếu hộ ở nông thôn.

Có bốn trường hợp có thể xảy ra như sau

(1) $\alpha_1=\alpha_2$ và $\beta_1= \beta_2$, hay không có sự khác biệt trong tiêu dùng gạo giữa thành thị và nông thôn.

Mô hình : Y = a + b X

Trong đó $\alpha_1=\alpha_2 = a$ và $\beta_1= \beta_2 = b$.

(2) $\alpha_1\neq\alpha_2$ và $\beta_1= \beta_2$, hay có sự khác biệt về tung độ gốc

Mô hình: $Y = a + bX + cD$

Trong đó $\alpha_1 = a$, $\alpha_2 = a + c$ và $\beta_1 = \beta_2 = b$.

(3) $\alpha_1=\alpha_2$ và $\beta_1\neq \beta_2$, hay có sự khác biệt về độ dốc

Mô hình: $Y = a + bX + c(DX)$

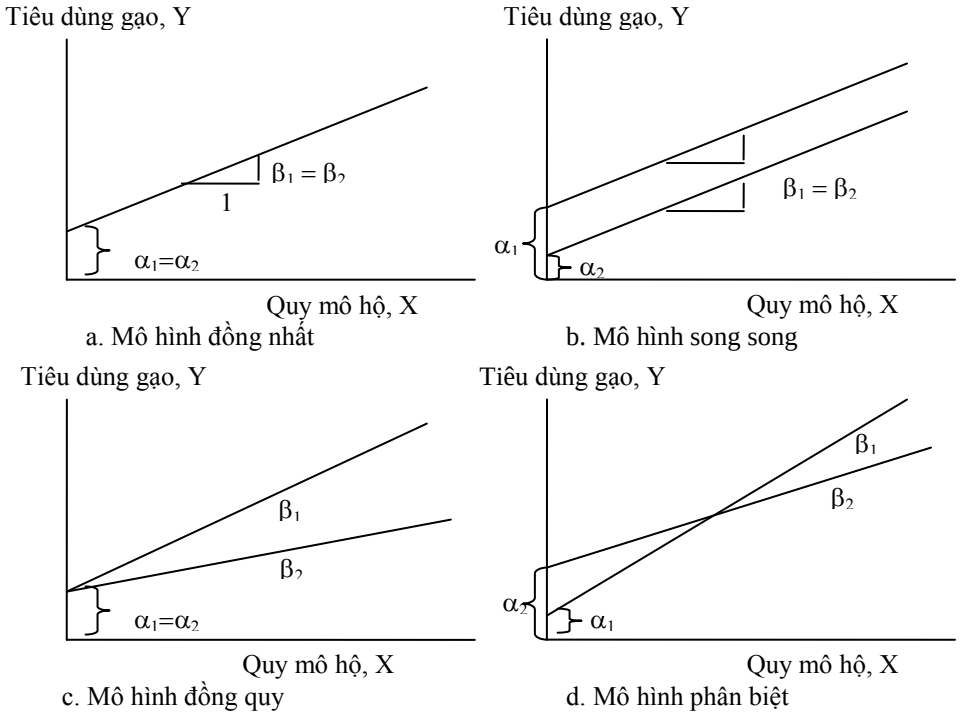
Trong đó DX = X nếu nếu D =1 và DX = 0 nếu D = 0

$\alpha_1 = \alpha_2 = a$, $\beta_1 = b$ và $\beta_2 = b + c$.

(4) $\alpha_1\neq\alpha_2$ và $\beta_1\neq \beta_2$, hay có sự khác biệt hoàn toàn về cả tung độ gốc và độ dốc.

Mô hình: $Y = a + bX + cD + d(DX)$

$\alpha_1 = a$, $\alpha_2 = a + c$, $\beta_1 = b$ và $\beta_2 = b + d$.



Hình 4.2. Các mô hình hồi quy

Biến DX được xây dựng như trên được gọi là biến tương tác. Tổng quát nếu X_p là một biến định lượng và D_q là một biến giả thì X_pD_q là một biến tương tác. Một mô hình hồi quy tuyến tổng quát có thể có nhiều biến định lượng, nhiều biến định tính và một số biến tương tác.

CHƯƠNG 5

GIỚI THIỆU MỘT SỐ VẤN ĐỀ LIÊN QUAN ĐẾN MÔ HÌNH HỒI QUY

5.1. Đa cộng tuyến

5.1.1. Bản chất của đa cộng tuyến

Đa cộng tuyến hoàn hảo: Các biến $X_1, X_2,\ldots,X_k$ được gọi là đa cộng tuyến hoàn hảo nếu tồn tại $\lambda_1, \lambda_2, \ldots,\lambda_k$ không đồng thời bằng không sao cho

$$\lambda_1X_1 + \lambda_2X_2 + \cdots + \lambda_kX_k = 0 \qquad (5.1)$$

Hiện tượng đa cộng tuyến hoàn hảo thường xảy do nhầm lẫn của nhà kinh tế lượng như trường hợp cái bẫy của biến giả mà chúng ta đã xem xét ở mục 4.7.3 chương 4.

Hiện tượng đa cộng tuyến mà chúng ta xét trong kinh tế lượng được hiểu với nghĩa rộng hơn đa cộng tuyến hoàn hảo như điều kiện (5.1). Các biến $X_1, X_2,\ldots,X_k$ được gọi là đa cộng tuyến không hoàn hảo nếu tồn tại $\lambda_1, \lambda_2, \ldots,\lambda_k$ sao cho

$$\lambda_1X_1 + \lambda_2X_2 + \cdots + \lambda_kX_k + \varepsilon = 0 \qquad (5.2)$$

với ε là sai số ngẫu nhiên.

Chúng ta có thể biểu diễn biến X_i theo các biến còn lại như sau

$$X_i = -\frac{\lambda_1}{\lambda_i}X_2 - \frac{\lambda_2}{\lambda_i}X_3 - \cdots - \frac{\lambda_k}{\lambda_i}X_k - \frac{\varepsilon}{\lambda_i} \text{ với } \lambda_i \neq 0. \qquad (5.3)$$

Vậy hiện tượng đa cộng tuyến xảy ra khi một biến là sự kết hợp tuyến tính của các biến còn lại và một nhiễu ngẫu nhiên.

Một số nguyên nhân gây ra hiện tượng đa cộng tuyến

(1) Khi chọn các biến độc lập mối quan có quan hệ nhân quả hay có tương quan cao vì đồng phụ thuộc vào một điều kiện khác. Ví dụ số giường bệnh và số bác sĩ nếu đồng thời là biến độc lập của một hồi quy thì sẽ gây ra hiện tượng đa cộng tuyến gần hoàn hảo.

(2) Khi số quan sát nhỏ hơn số biến độc lập. Một ví dụ điển hình là một nghiên cứu y khoa trên một số lượng nhỏ bệnh nhân nhưng lại khảo sát quá nhiều nhân tố tác động lên hiệu quả điều trị.

(3) Cách thu thập mẫu. Ví dụ chỉ thu thập mẫu trên một số lớp giới hạn của tổng thể.

(4) Chọn biến X_i có độ biến thiên nhỏ.

5.1.2. Hệ quả của đa cộng tuyến

Ví dụ 5.1[20]. Nghiên cứu của Klein và Golberger(1995) về quan hệ giữa tiêu dùng nội địa C, thu nhập từ lương W, thu nhập khác phi nông nghiệp P và thu nhập từ nông nghiệp A của nền kinh tế Hoa Kỳ từ năm 1928 đến 1950, với số liệu của các năm 1942 đến 1944 bị loại ra khỏi dữ liệu. Klein và Golberger thực hiện hồi quy tiêu dùng nội địa theo ba loại thu nhập như sau

$$C_t = \beta_1 + \beta_2 W_t + \beta_3 P_t + \beta_4 A + \varepsilon_t \qquad (5.4)$$

Hồi quy này có thể gặp phải hiện tượng đa cộng tuyến vì các loại thu nhập có xu hướng cùng tăng theo sự phát triển của nền kinh tế.

Năm	C	W	P	A
1928	52,8	39,21	17,73	4,39
1929	62,2	42,31	20,29	4,60
1930	58,6	40,37	18,83	3,25
1931	56,6	39,15	17,44	2,61
1932	51,6	34,00	14,76	1,67
1933	51,1	33,59	13,39	2,44
1934	54	36,88	13,93	2,39
1935	57,2	39,27	14,67	5,00
1936	62,8	45,51	17,20	3,93
1937	65	46,06	17,15	5,48
1938	63,9	44,16	15,92	4,37
1939	67,5	47,68	17,59	4,51
1940	71,3	50,79	18,49	4,90
1941	76,6	57,78	19,18	6,37
1945	86,3	78,97	19,12	8,42
1946	95,7	73,54	19,76	9,27
1947	98,3	74,92	17,55	8,87
1948	100,3	74,01	19,17	9,30
1949	103,2	75,51	20,20	6,95
1950	108,9	80,97	22,12	7,15

Bảng 5.1. Số liệu thu nhập và tiêu dùng của nền kinh tế Hoa Kỳ

Kết quả hồi quy như sau

$\hat{C}$ =	8,133 +	1,059W +	0,452P +	0,121A	(5.5)
t-Stat	(0,91)	(6,10)	(0,69)	(0,11)	
Khoảng 95%	(-10,78;27,04)	(0,69;1,73)	(-0,94;1,84)	(-2,18;2,43)	
$R^2 = 0,95$	$F = 107,07 > F(3,16,99\%) = 5,29.$				

[20] Ví dụ này lấy từ William E.Griffiths et al, Learning and Practicing Econometrics, John Wiley&Sons Inc, 1998, trang 433.

Mô hình này có tính giải thích cao thể hiện qua R^2 rất cao và thống kê F cao. Tuy nhiên một số hệ số lại không khác không với ý nghĩa thống kê thể hiện qua t-stat thấp, nghĩa là ước lượng khoảng cho các hệ số chứa 0. W với hệ số có t-stat lớn thì ý nghĩa kinh tế lại rất lạ: nếu thu nhập từ lương tăng 1 USD thì tiêu dùng tăng 1,059 USD. Để tìm hiểu lý do gây ra hiện tượng trên chúng ta phải dùng lý thuyết của đại số ma trận, ở đây chỉ minh hoạ bằng mô hình hồi quy ba biến. Phương sai của ước lượng hệ số β2 là

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

Khi X_2 và X_3 có hiện tượng cộng tuyến thì r_{23}^2 cao làm cho phương sai của ước lượng β_2 cao. Ước lượng b2 theo phương pháp bình phương tối thiểu trở nên không hiệu quả.

Hệ quả của đa cộng tuyến

(1) Ước lượng các hệ số không hiệu quả do phương sai của ước lượng lớn. Mô hình có đa cộng tuyến có t-stat nhỏ và một số hệ số của thể có dấu trái với lý thuyết hay có giá trị không phù hợp. R^2 thể hiện độ phù hợp của dữ liệu và F thể hiện ý nghĩa chung của các hệ số có thể rất cao.

(2) Giá trị ước lượng của các hệ số rất nhạy cảm đối với việc tăng hoặc bớt một hoặc quan sát hoặc loại bỏ biến có mức ý nghĩa thấp.

(3) Mặc dù việc phân tích tác động riêng phần của một biến khó khăn nhưng tính chính xác của dự báo có thể vẫn cao khi bản chất của đa cộng tuyến vẫn không đổi đối với quan sát mới.

5.1.3 Biện pháp khắc phục

Nếu mục tiêu của phân tích hồi quy là dự báo thì trong một số trường hợp chúng ta không cần khắc phục hiện tượng đa cộng tuyến.

Nếu mục tiêu của phân tích là xét tác động riêng phần của từng biến số 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.

(1) Dùng thông tin tiên nghiệm. Ví dụ khi hồi quy hàm sản xuất Cobb-Douglas

$$\text{Ln}(Y_i)=\beta_1 + \beta_2\ln(Ki)+ \beta_3\ln(Li) + \varepsilon_i \qquad (5.6)$$

Chúng ta có thể gặp hiện tượng đa cộng tuyến do K và L cùng tăng theo quy mô sản xuất. Nếu ta biết là hiệu suất không đổi theo quy mô thì ta có thêm thông tin $\beta_2+\beta_3=1$. Với thông tin tiên nghiệm này chúng ta chuyển mô hình hồi quy (5.6) thành

$$\text{Ln}(Y_i)=\beta_1 + \beta_2\ln(Ki)+ (1-\beta_2)\ln(Li) + \varepsilon_i \qquad (5.7)$$

(2) Bỏ đi một biến có đa cộng tuyến. Đây là cách làm đơn giản nhất. Ví dụ trong mô hình có biến giải thích là số bác sĩ và số giường bệnh thì ta có thể bỏ đi biến số giường bệnh. Nếu biến bị bỏ đi thực sự cần phải có trong mô hình thì chúng ta lại gặp phải một vấn đề khác, đó là ước lượng chệch đối với các hệ số còn lại. Vấn đề này chúng ta sẽ tiếp tục xem xét ở cuối chương.

(3) Chuyển dạng dữ liệu

Giả sử chúng ta hồi quy trên dữ liệu chuỗi thời gian

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_t \qquad (5.8)$$

Và chúng ta gặp phải hiện tượng đa cộng tuyến do X_{1t} và X_{3t} có thể cùng tăng hoặc giảm theo từng năm. Ta có thể tối thiểu tác động đa cộng tuyến này bằng kỹ thuật hồi quy trên sai phân bậc nhất như sau:

Ta có

$$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + \varepsilon_{t-1} \qquad (5.9)$$

Từ (5.8) và (5.9) ta xây dựng mô hình hồi quy

$$(Y_t - Y_{t-1}) = \beta 2(X_{2t} - X_{2,t-1}) + \beta_3(X_{3t} - {}_3X_{3,t-1}) + \nu_t \qquad (5.10)$$

Với $\nu_t = \varepsilon_t - \varepsilon_{t-1}$.

Một vấn đề mới nảy sinh là ν_t có thể có tính tương quan chuỗi, và như thế không tuân theo giả định của mô hình hồi quy tuyến tính cổ điển. Nếu hiện tượng tương quan chuỗi là nghiêm trọng thì mô hình (5.10) còn kém hơn cả mô hình (5.8).

(4) Tăng thêm quan sát. Giải pháp này thích hợp cho hiện tượng đa cộng tuyến do cỡ mẫu nhỏ. Đôi khi chỉ cần tăng thêm một số quan sát là ta khắc phục được hiện tượng đa cộng tuyến. Một lần nữa chúng ta lại có sự đánh đổi. Tăng dữ liệu đôi khi đồng nghĩa với việc tăng chi phí, nhất là đối với dữ liệu sơ cấp. Mặt khác nếu là dữ liệu không có kiểm soát, chúng ta phải biết chắc rằng các điều kiện khác tương tự với khi ta thu thập dữ liệu gốc.

Khắc phục hiện tượng đa cộng tuyến đòi hỏi các kỹ thuật phức tạp và đôi khi cũng không mang lại hiệu quả như ta mong muốn. Mặt khác, hầu hết các mô hình hồi quy bội đều có tính cộng tuyến nhất định nên chúng ta phải cẩn thận trong việc xây dựng mô hình và giải thích kết quả. Chúng ta sẽ nghiên cứu nguyên tắc xây dựng mô hình ở cuối chương.

5.2.Phương sai của sai số thay đổi - HETEROSKEDASTICITY

5.2.1. Bản chất của phương sai của sai số thay đổi

Giả định của mô hình hồi quy tuyến tính cổ điển là phương sai của sai số hồi quy không đổi qua các quan sát. Trong thực tế sai số hồi quy có thể tăng lên hoặc giảm đi khi giá trị biến độc lập X tăng lên. Tổng quát, thay cho giả định

$$E(e_i^2) = \sigma^2$$

chúng ta giả định

$$E(e_i^2) = \sigma_i^2 \qquad (5.11)$$

Thường gặp phương sai không đồng nhất ở dữ liệu chéo và dữ liệu bảng. Nguyên nhân phương sai không đồng nhất rất đa dạng, sau đây là một số trường hợp điển hình:

(1) Gọi Y là số phế phẩm trong 100 sản phẩm của một thợ học việc, X là số giờ thực hành. Khi số giờ thực hành càng lớn thì số phế phẩm càng nhỏ và càng ít biến động. Chúng ta có trường hợp phương sai giảm dần khi X tăng dần.

(2) Khi thu nhập(X) tăng thì chi tiêu cho các mặt hàng xa xỉ tăng và mức biến động càng lớn. Chúng ta có trường hợp phương sai tăng dần khi X tăng dần.

(3) Khi cải thiện phương pháp thu thập số liệu thì phương sai giảm.

(4) Phương sai của sai số tăng do sự xuất hiện của điểm nằm ngoài, đó là các trường hợp bất thường với dữ liệu rất khác biệt(rất lớn hoặc rất nhỏ so với các quan sát khác).

(5) Phương sai thay đổi khi không xác đúng dạng mô hình, nếu một biến quan trọng bị bỏ sót thì phương sai của sai số lớn và thay đổi. Tình trạng này giảm hẳn khi đưa biến bị bỏ sót vào mô hình.

5.2.2. Hệ quả của phương sai thay đổi khi sử dụng ước lượng OLS

Xét hồi quy

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon i \qquad (5.12)$$

với $E(e_i^2) = \sigma_i^2$

Sử dụng phương pháp bình phương tối thiểu thông thường (OLS) chúng ta có

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} = \beta_2 + \frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2} \qquad (5.13)$$

$$E\left(\hat{\beta}_2\right) = \beta_2 + \frac{\sum_{i=1}^{n} x_i E(\varepsilon_i)}{\sum_{i=1}^{n} x_i^2} = \beta_2$$

vậy ước lượng theo OLS không chệch.

$$\text{var}\left(\hat{\beta}_2\right) = \frac{\sum_{i=1}^{n} x_i^2 \sigma_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2}$$

Chúng ta không chưa rõ là OLS có cho ước lượng hiệu quả hay không.

Ước lượng bình phương tối thiểu có trọng số (WLS)

Đặt $\sigma_i^2 = w_i^2 \sigma^2$, chia hai vế của (5,12) cho w_i chúng ta có mô hình hồi quy

$$\frac{Y_i}{w_i} = \beta_1 \frac{1}{w_i} + \beta_2 \frac{X_i}{w_i} + \frac{\varepsilon_i}{w_i} \qquad (5.14)$$

Ta viết lại mô hình (5.13) như sau

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \varepsilon_i^* \qquad (5.15)$$

Mô hình (5.14) không có tung độ gốc và phương sai đồng nhất.

$$\text{var}(\varepsilon_i^*) = \text{var}\left(\frac{\varepsilon_i}{w_i}\right) = \frac{w_i^2 \sigma^2}{w_i^2} = \sigma^2$$

Vậy ước lượng hệ số của (5.15) theo OLS là ước lượng hiệu quả(BLUE).

Kết quả ước lượng β_2 của (5.15) theo OLS như sau

$$\hat{\beta}_{2,WLS} = \frac{\sum_{i=1}^{n}\left(\frac{X_i Y_i}{w_i^2}\right)\sum_{i=1}^{n}\left(\frac{1}{w_i^2}\right) - \sum_{i=1}^{n}\left(\frac{Y_i}{w_i^2}\right)\sum_{i=1}^{n}\left(\frac{X_i}{w_i^2}\right)}{\sum_{i=1}^{n}\left(\frac{X_i^2}{w_i^2}\right)\sum_{i=1}^{n}\left(\frac{1}{w_i^2}\right) - \left(\sum_{i=1}^{n}\left(\frac{X_i}{w_i^2}\right)\right)^2} \qquad (5.16)$$

Ước lượng (5.16) hoàn toàn khác với (5.13). Chúng ta biết ước lượng theo WLS (5.16) là ước lượng hiệu quả vậy ước lượng theo OLS (5.13) là không hiệu quả.

Phương sai đúng của hệ số ước lượng β_2 là $\text{var}\left(\hat{\beta}_2\right) = \frac{\sum_{i=1}^{n} x_i^2 \sigma_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2}$ nhưng các phần mềm máy tính báo cáo phương sai là $\text{var}\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$.

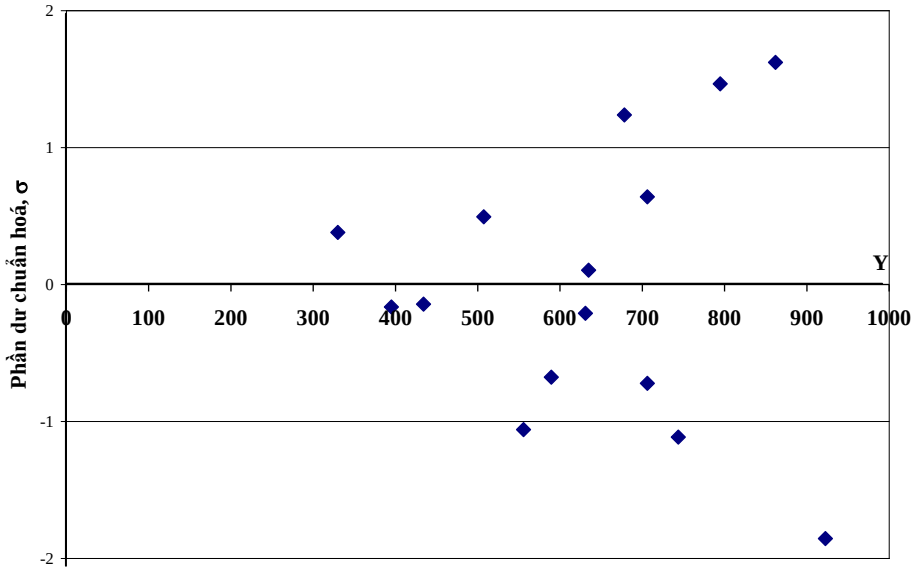
Từ phương sai của sai số bị tính sai này các trị thống kê t-stat và sai số chuẩn của hệ số ước lượng phần mềm cung cấp là vô dụng.

Tóm lại, với sự hiện diện của phương sai của sai số thay đổi mặc dù ước lượng các hệ số theo OLS vẫn không chệch nhưng ước lượng không hiệu quả và các trị thống kê như t-stat không chính xác.

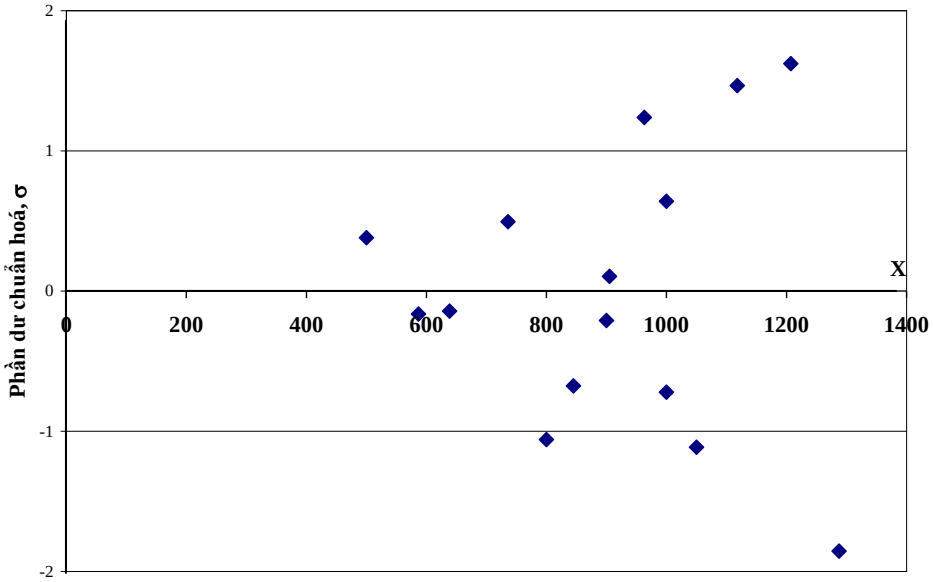
5.2.3. Phát hiện và khắc phục

Phát hiện phương sai của sai số thay đổi.

Phương pháp đồ thị. Xét đồ thị của phần dư theo giá trị Y và X.



Hình 5.1. Đồ thị phân tán phần dư e_i theo $\hat{Y}_i$.



Hình 5.2. Đồ thị phân tán phần dư e_i theo X_i

Theo các đồ thị trên thì khi giá trị dự báo Y tăng (hoặc khi X tăng) thì phần dư có xu hướng tăng, hay mô hình có phương sai của sai số thay đổi.

Các phép thử chính thức

Xét hồi quy bội

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + ... + \beta_k X_{k,i} + \varepsilon_i \quad (5.17)$$

Trong (k-1) biến độc lập trên ta trích ra (p-1) biến làm biến độc lập cho một hồi quy phụ. Trong hồi quy phụ này phần dư từ hồi quy mô hình(5.17) làm hồi quy biến phụ thuộc.

Các dạng hồi quy phụ thường sử dụng là

$$e_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_p Z_{pi} + \delta_i \quad (5.18)$$

$$|e_i| = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_p Z_{pi} + \delta_i \quad (5.19)$$

$$\ln(e_i^2) = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_p Z_{pi} + \delta_i \quad (5.20)$$

Kiểm định Breusch-Pagan căn cứ vào hồi quy phụ (5.18), kiểm định Glejser căn cứ vào (5.19) và kiểm định Harvey-Godfrey căn cứ vào (5.20).

Giả thiết không là không có phương sai không đồng nhất

$H_0 : \alpha_2 = \alpha_3 = \ldots = \alpha_p = 0$

H_1 : Không phải tất cả các hệ số trên đều bằng 0.

R^2 xác định từ hồi quy phụ, n là cỡ mẫu dùng để xây dựng hồi quy phụ, với cỡ mẫu lớn thì nR^2 tuân theo phân phối Chi bình phương với (p-1) bậc tự do.

Quy tắc quyết định

Nếu $\chi^2_{(p-1,1-\alpha)} \leq nR^2$ thì bác bỏ H_0.

Nếu bác bỏ được H_0 thì chúng ta chấp nhận mô hình có phương sai của sai số thay đổi và thực hiện kỹ thuật ước lượng mô hình như sau:

Đối với kiểm định Breusch-Pagan

$$\hat{w}_i^2 = \hat{\alpha}_1 + \hat{\alpha}_2 Z_{2i} + \cdots + \hat{\alpha}_p Z_{pi}$$

Đối với kiểm định Glejser

$$\hat{w}_i^2 = (\hat{\alpha}_1 + \hat{\alpha}_2 Z_{2i} + \cdots + \hat{\alpha}_p Z_{pi})^2$$

Đối với kiểm định Harvey-Godfrey

$$\hat{w}_i^2 = \exp(\hat{\alpha}_1 + \hat{\alpha}_2 Z_{2i} + \cdots + \hat{\alpha}_p Z_{pi})$$

Ta có $\hat{w}_i = \sqrt{\hat{w}_i^2}$. Đến đây chúng ta có thể chuyển dạng hồi quy theo OLS thông thường sang hồi quy theo bình phương tối thiểu có trọng số WLS.

5.3. Tự tương quan (tương quan chuỗi)

Trong mô hình hồi quy tuyến tính cổ điển chúng ta giả định không có tương quan giữa các phần dư hay $E(\varepsilon_i\varepsilon_j) = 0$ với mọi i, j.

Trong thực tế đối với dữ liệu chuỗi thời gian, giả định này hay bị vi phạm. Một lý do nôm na là biến số kinh tế có một quán tính(sức ỳ) nhất định. Ví dụ sự tăng cầu một loại hàng hóa của năm nay sẽ làm tăng lượng cung nội địa của hàng hoá đó vào năm sau, đây là tác động trễ của biến độc lập hay biến phụ thuộc thời kỳ t chịu tác động của biến độc lập ở thời kỳ t-1.

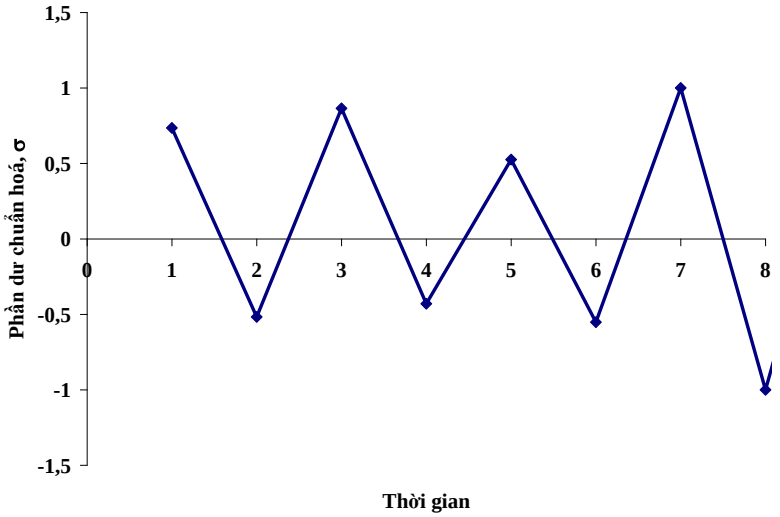
Đôi khi nền kinh tế lại phản ứng quá nhạy với sự thay đổi. Ví dụ giá mía cao ở năm nay sẽ làm cho nông dân đổ xô trồng mía, sản lượng mía năm sau tăng vọt làm giảm giá mía ở năm sau, đây là tác động trễ của biến phụ thuộc hay giá trị biến phụ thuộc thời kỳ t chịu ảnh hưởng của giá trị biến phụ thuộc thời kỳ t-1.

Hiện tượng tự tương quan làm cho $E(\varepsilon_i\varepsilon_j) \neq 0$ và gây ra các hậu quả sau

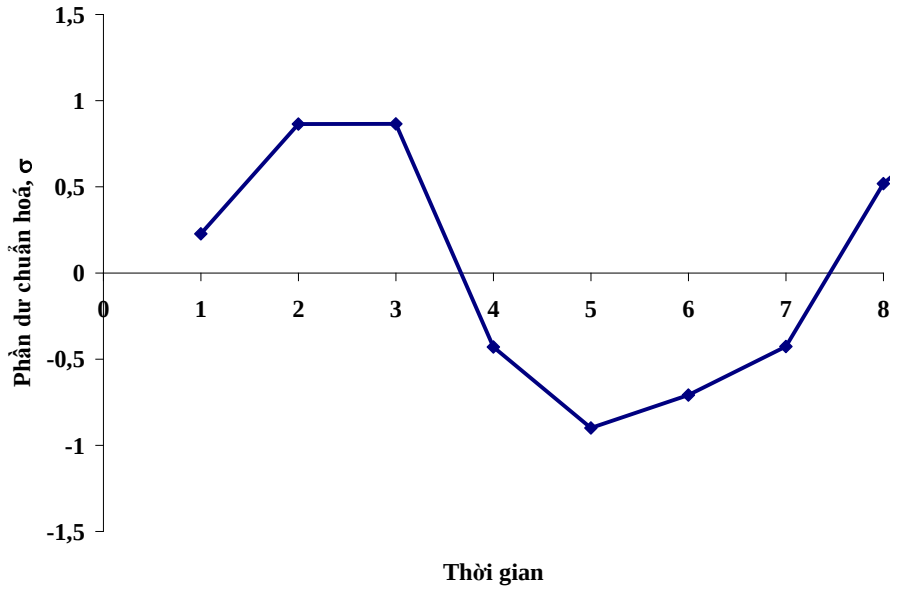
(1) Ước lượng theo OLS không chệch nhưng không hiệu quả

(2) Các trị thống kê tính theo OLS không hữu ích trong việc nhận định mô hình.

Chúng ta có thể phát hiện hiện tượng tự tương quan bằng cách quan sát đồ thị phần dư của mô hình trên dữ liệu chuỗi thời gian.



Hình 5.3. Tương quan chuỗi nghịch



Hình 5.4. Tương quan chuỗi thuận

Chúng ta sẽ tiếp tục làm việc với dữ liệu chuỗi và xử lý hiện tượng tự tương quan ở phần sau của giáo trình liên quan đến các mô hình dự báo.

5.4.Lựa chọn mô hình

Một yếu tố quan trọng đầu tiên để chọn đúng mô hình hồi quy là chọn đúng dạng hàm. Để chọn đúng dạng hàm chúng ta phải hiểu ý nghĩa và mối quan hệ kinh tế của các biến số. Ý nghĩa của một số loại hàm thông dụng đã được trình bày ở mục 3.8.2 chương 3. Ở phần này chúng ta xét hậu quả của một số dạng xây dựng mô hình sai và chiến lược xây dựng mô hình kinh tế lượng. Chúng ta cũng không đi sâu vào chứng minh các kết quả.

5.4.1. Thiếu biến có liên quan và chứa biến không liên quan.

Xét hai hồi quy sau

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \xi_i \qquad (5.21)$$

và

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \beta_{(K+1)} X_{K+1,i} + \cdots \beta_{(K+L)} X_{K+L,i} + \varepsilon_i \qquad (5.22)$$

Mô hình (5.21) có các trị thông kê tương ứng có ký hiệu R và mô hình (5.22) có các trị thống kê tương ứng có ký hiệu U.

Có hai trường hợp xảy ra:

- Trường hợp 1: Nếu mô hình (5.22) là đúng nhưng chúng ta chọn mô hình (5.21) nghĩa là chúng ta bỏ sót L biến quan trọng ($X_{K+1},..X_{K+L}$). Hậu quả là ước lượng các hệ số cho K-1 biến độc lập còn lại bị chệch, mô hình kém tính giải thích cho cả mục tiêu dự báo vào phân tích chính sách.

- Trường hợp 2: Nếu mô hình (5.21) là đúng nhưng chúng ta chọn mô hình (5.22), nghĩa là chúng ta đưa vào mô hình các biến không liên quan. Hậu quả là ước lượng hệ số cho các biến quan trọng vẫn không chệch nhưng không hiệu quả.

5.4.2. Kiểm định so sánh mô hình (5.21) và (5.22) - Kiểm định Wald

Chúng ta muốn kiểm định xem L biến ($X_{K+1},..X_{K+L}$) có đáng được đưa vào mô hình hay không.

$$H_0: \beta_{K+1} = \beta_{K+2} = \cdots = \beta_{K+L} = 0$$

Trị thống kê

$$\frac{(RSS_R - RSS_U)/L}{RSS_U/(n-K-L)} \sim F^* \sim F_{(L,n-K-L)}$$

Quy tắc quyết dịnh: Nếu $F^* > F_{((L,n-K-L),1-\alpha)}$ thì ta bác bỏ H_0 hay chấp nhận L biến ($X_{K+1},..X_{K+L}$) xứng đáng được đưa vào mô hình.

5.4.3. Hai chiến lược xây dựng mô hình

Có hai chiến lược xây dựng mô hình kinh tế lượng là:

- Xây dựng mô hình từ đơn giản đến tổng quát: chứa tất cả các biến có liên quan trong mô hình và loại bỏ dần những biến ít ý nghĩa thống kê nhất cho đến khi nhận được mô hình "tốt nhất".

- Xây dựng mô hình tổng quát đến đơn giản : Xuất phát từ biến độc lập có quan hệ kinh tế trực tiếp nhất với biến phụ thuộc, tiếp tục bổ sung biến mới cho đến khi nhận được mô hình "tốt nhất".

Mỗi cách làm đều có những ưu và nhược điểm. Hiện nay với công cụ máy vi tính, người ta không còn ngại tính toán trên mô hình lớn và nhiều nhà kinh tế lượng cho rằng xây dựng mô hình từ tổng quát đến đơn giản thì hiệu quả hơn từ đơn giản đến tổng quát. Nét chung của cả hai chiến lược này là ở từng bước đều phải thực hiện kiểm định Wald.

CHƯƠNG 6

DỰ BÁO VỚI MÔ HÌNH HỒI QUY (Đọc thêm)

PHÂN LOẠI CÁC PHƯƠNG PHÁP DỰ BÁO

Có hai nhóm phương pháp dự báo chính là nhóm định tính và nhóm định lượng. Trong giáo trình này chúng ta chủ yếu sử dụng phương pháp định lượng có kết hợp với các phán đoán định tính để dự báo.

Các phương pháp dự báo định tính

Các phương pháp dự báo định tính dựa vào phán đoán chủ quan và trực giác để đưa ra dự báo thay cho vì dựa vào các số liệu quá khứ. Phương pháp dự báo định tính hữu ích cho việc dự báo toàn cục và một số trường hợp mà số liệu quá khứ không hữu ích cho dự báo.

Các phương pháp dự báo định lượng

Các kỹ thuật dự báo định lượng dựa vào việc phân tích số liệu quá khứ để đưa ra dự báo. Giả định của phương pháp này là các nhân tố từng tác động lên biến được dự báo trong quá khứ vẫn tiếp tục ảnh hưởng đến biến này trong tương lai. Vậy dựa vào diễn biến dữ liệu trong quá khứ ta có thể dự báo cho tương lai. Các phương pháp dự báo định lượng lại được chia thành hai nhóm chính: dự báo định lượng mang tính nhân quả và dự báo định lượng mang tính thống kê.

Các phương pháp dự báo định lượng mang tính nhân quả

Đại diện của nhóm phương pháp này là phân tích hồi quy. Mô hình dự báo có hai nhóm biến số: các biến số được dự báo được gọi là biến độc lập, các biến số dùng để dự báo được gọi là biến phụ thuộc. Chúng ta đã nghiên cứu mô hình hồi quy ở phần 1, nay chúng ta tiếp tục nghiên cứu việc áp dụng mô hình hồi quy cho dự báo và một số kỹ thuật phân tích hồi quy với dữ liệu chuỗi thời gian.

Các phương pháp dự báo định lượng mang tính thống kê

Nhóm các phương pháp dự báo mang tính thống kê chỉ quan tâm đến quy luật biến thiên của biến cần dự báo trong quá khứ để dưa ra dự báo. Biến thiên của một biến số kinh tế được chia thành các thành phần: xu hướng, chu kỳ, thời vụ và ngẫu nhiên.

Nhóm các phương pháp dự báo mang tính thống kê lại chia thành hai nhóm chính.

- Nhóm thứ nhất phân tích một thành phần hoặc kết hợp một số thành phần riêng biệt nêu trên như: đường xu hướng, san bằng số mũ, trung bình động.

- Nhóm thứ hai sử dụng các khái niệm thống kê về dữ liệu chuỗi thời gian mà không chia biến động của dữ liệu thành các thành phần riêng biệt như ở phương pháp luận Box-Jenkins.

6.1. Dự báo với mô hình hồi quy thông thường

Mô hình hồi quy

$$Y_t = \beta_1 + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \varepsilon_t \qquad (6.1)$$

Chỉ số t chỉ thời kỳ thứ t.

Giả sử mô hình này thoả mãn các điều kiện của phương pháp ước lượng theo bình phương tối thiểu. Các tham số ước lượng từ mô hình tương ứng là $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$.

Ước đoán tốt nhất cho Y_{t+1} khi biết các $X_{i,t+1}$ là:

$$\hat{Y}_{t+1} = E\left(\hat{\beta}_1 + \hat{\beta}_2 X_{2,t+1} + \cdots + \hat{\beta}_k X_{k,t+1}\right) \qquad (6.2)$$

Độ lệch chuẩn của ước lượng là

Đối với hồi quy hai biến

$$se\left(\hat{Y}_{t+1}\right) = \sigma\left(1 + \frac{1}{n} + \frac{(X_{t+1} - \overline{X})^2}{\sum_{i=1}^{n} x_i^2}\right)^{1/2} \qquad (6.3)$$

Đối với hồi quy bội: công thức rất phức tạp và nằm ngoài phạm vi giáo trình này.

6.2. Tính chất "trễ" của dữ liệu chuỗi thời gian và hệ quả của nó đến mô hình

Khi chúng ta sử dụng mô hình (6.1) chúng ta giả định rằng các biến độc lập tác động tức thì lên biến phụ thuộc và biến phụ thuộc chỉ chịu tác động của biến độc lập. Đối với các biến số kinh tế các giả định này thường không đúng. Tác động của biến độc lập có thành phần tác động tức thời và có thành phần tác động trễ. Mặt khác, đôi khi bản thân biến phụ thuộc cũng có "quán tính" hay "sức ỳ" của nó. Có ba nguyên nhân gây ra "độ trễ" hay "sức ỳ" trong kinh tế là

(1) Nguyên nhân tâm lý

Khi thu nhập của một người giảm tiêu dùng của người đó có thể không giảm ngay lập tức do thói quen duy trì mức sống cao. Nếu tình hình thu nhập vẫn không phục hồi trong thời gian dài, anh ta phải học cách chi tiêu tiết kiệm hơn.

(2) Nguyên nhân kỹ thuật

Giả sử cầu nội địa đối với một mặt hàng tăng lên làm giá một mặt hàng này tăng. Sản lượng nội địa có thể không tăng tức thời vì để tăng sản lượng cần phải có thời gian xây dựng nhà máy, đầu tư máy móc thiết bị và đào tạo công nhân. Doanh nghiệp còn phải phân tích xem sự tăng cầu nội địa này có mang tính chất lâu dài hay chỉ là tức thời.

(3) Nguyên nhân định chế

Các ràng buộc pháp lý là nguyên nhân của một số hiện tượng tác động trễ. Ví dụ nếu hợp đồng tài trợ Giải bóng đá chuyên nghiệp Việt Nam đã được ký kết có hiệu lực 2 năm thì Liên đoàn Bóng đá Việt Nam không thể huỷ hợp đồng để ký lại với một đối tác khác có số tiền tài trợ cao hơn. Giả sử số tiền tài trợ phụ thuộc tầm ảnh hưởng của giải đấu lên công chúng thể hiện qua số lượt khán giả đến sân và số lượt khán giả theo dõi qua truyền hình. Số khán giả đến sân tăng lên chỉ có thể tác động làm tăng số tiền tài trợ của lần ký kết ở 2 năm sau.

Khi có tính chất "trễ" nêu trên của dữ liệu chuỗi thời gian, mô hình (6.1) có sai số hồi quy không thỏa mãn các điều kiện của mô hình hồi quy tuyến tính cổ điển.(Tại sao?). Từ đó dự báo theo (6.2) sẽ không chính xác.

6.3. Mô hình tự hồi quy

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \gamma_t \qquad (6.4)$$

Mô hình (6.4) còn được gọi là mô hình động vì nó thể hiện mối liên hệ giữa giá trị của biến phụ thuộc với giá trị quá khứ của nó.

6.4. Mô hình có độ trễ phân phối

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_k X_{t-k} + \varepsilon_t \qquad (6.5)$$

Trong mô hình này k được gọi là độ trễ. Chúng ta phải xác định độ trễ k.

6.4.1. Cách tiếp cận của Alt và Tinberger[21]:

Vì X_t là xác định và không tương quan với ε_t nên $X_{t-1}, X_{t-2}, \ldots, X_{t-k}$ đều xác định và không tương quan với ε_t. Do đó chúng ta có thể áp dụng OLS để ước lượng tham số cho mô hình (6.5). Chúng ta sẽ xác định k bằng cách tăng dần độ trễ như sau:

(1) Hồi quy Y_t theo X_t

(2) Hồi quy Y_t theo X_t và X_{t-1} …

(k) Hồi quy Y_t theo $X_t, X_{t-1}, \ldots, X_{t-k}$

(k+1) Hồi quy Y_t theo $X_t, X_{t-1}, \ldots, X_{t-(k+1)}$

Quá trình này dừng ở độ trễ (k+1) hoặc (k+2) khi chúng ta nhận thấy các hệ số ứng với các biến trễ không có ý nghĩa thống kê hoặc đổi dấu.

Quá trình trên vướng phải bốn nhược điểm như sau:

(1) Không có tiên liệu trước là độ trễ sẽ là bao nhiêu.

[21] F.F.Alt, "Distribution Lags", Economitrica, quyển 10,1942, trang 113-128. (Theo D.N.Gujarati, Basis Econometrics 3rd Edition, 1995, trang 591).

(2) Mô hình có thêm một độ trễ thì mất đi một bậc tự do, nếu dữ liệu chuỗi thời gian không đủ dài thì ý nghĩa thống kê của mô hình ngày càng kém.

(3) Các biến giải thích thực chất là giá trị của một biến X theo thời gian, điều này gây ra sự tương quan giữa các biến giải thích trong mô hình, tức là có hiện tượng đa cộng tuyến. Ước lượng các tham số của mô hình trong trường hợp có đa cộng tuyến sẽ cho kết quả kém chính xác.

(4) Việc xác định độ trễ k của mô hình (6.5) theo cách thức trên là một dạng của "đào mỏ dữ liệu".

6.4.2. Mô hình Koyck

Giả định:

(1) Tất cả các hệ số ứng với biến trễ có cùng dấu

(2) Các hệ số tuân theo cấp số nhân giảm dần: $\beta_k = \beta_0 \lambda^k$ với $0 < \lambda < 1$.

Chúng ta viết lại mô hình (6.5) như sau

$$Y_t = \alpha + \beta_0 X_t + \beta_0 \lambda X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \cdots + \varepsilon_t \quad (6.6)$$

Tương tự

$$Y_{t-1} = \alpha + \beta_0 X_{t-1} + \beta_0 \lambda X_{t-2} + \beta_0 \lambda^2 X_{t-3} + \cdots + \varepsilon_{t-1} \quad (6.7)$$

Nhân (6.7) với λ

$$\lambda Y_{t-1} = \alpha\lambda + \beta_0 \lambda X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \beta_0 \lambda^3 X_{t-3} + \cdots + \varepsilon_{t-1} \quad (6.8)$$

Lấy (6.6) trừ (6.7)

$$Y_t - \lambda Y_{t-1} = \alpha(1-\lambda) + \beta_0 X_t + (\varepsilon_t - \lambda \varepsilon_{t-1}) \quad (6.9)$$

Kết quả cuối cùng

$$Y_t = \alpha(1-\lambda) + \beta_0 X_t + \lambda Y_{t-1} + \gamma_t \quad (6.10)$$

Với $\gamma_t = \varepsilon_t - \lambda \varepsilon_{t-1}$, γ_t còn được gọi là trung bình trượt của ε_t và ε_{t-1}.

Mô hình (6.10) được gọi là mô hình chuyển dạng Koyck. Chúng ta đã chuyển mô hình trễ phân phối thành mô hình tự hồi quy.

6.4.3. Mô hình kỳ vọng thích nghi

Giả sử mô hình xác định cầu tiền có dạng như sau[22]

[22] P.Cagan, "The Monetary Dynamics of Hyperinflations", in M.Friedman (ed.), "*Studies in the Quantity Theory of Money"*, University of Chicago Press, 1956.

$$Y_t = \beta_0 + \beta_1 X_t^* + \varepsilon_t \qquad (6.11)$$

Y : Cầu tiền

X*: Giá trị kỳ vọng[23] của lãi suất danh nghĩa

ε: Sai số hồi quy

Lãi suất kỳ vọng của năm nay(năm t) không thể quan sát được một cách trực tiếp mà được xác định như sau

$$X_t^* - X_{t-1}^* = \gamma(X_t - X_{t-1}^*) \text{ với } 0 < \lambda \leq 1.$$

Biểu thức này hàm ý kỳ vọng của người ta thay đổi(thích hợp) theo lãi suất thực tế, hay nói cách khác người ta học hỏi từ sai lầm.

$$X_t^* = \gamma X_t + (1-\gamma) X_{t-1}^* \qquad (6.12)$$

Thay (6.12) vào (6.11)

$$Y_t = \beta_0 + \beta_1 [\gamma X_t + (1-\gamma) X_{t-1}^*] + \varepsilon_t$$

Qua một số phép biến đổi tương tự như mô hình Koyck ta có

$$Y_t = \gamma\beta_0 + \gamma\beta_1 X_t + (1-\gamma) Y_{t-1} + \gamma_t \qquad (6.13)$$

Với $\gamma_t = \varepsilon_t - (1-\gamma)\varepsilon_{t-1}$

6.4.4. Mô hình hiệu chỉnh từng phần

Mô hình hiệu chỉnh từng phần phù hợp với phân tích hồi quy có độ trễ do lý do kỹ thuật và định chế.

Giả sử mức đầu tư tư bản tối ưu ứng với một mức sản lượng X cho trước là Y*. Mô hình hồi quy đơn giản Y* theo X như sau:

$$Y_t^* = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad (6.14)$$

Thực tế chúng ta không trực tiếp quan sát được Y_t^*.

Giả định Y_t^* được xác định như sau:

$$Y_t - Y_{t-1} = \delta(Y_t^* - Y_{t-1}) \qquad \text{với } 0 < \delta \leq 1. \qquad (6.15)$$

Trong đó

$Y_t - Y_{t-1} = I$: Thay đổi lượng tư bản thực tế, cũng chính là đầu tư trong kỳ

[23] Giá trị kỳ vọng ở đây mang ý nghĩa giá trị được mong đợi, không mang ý nghĩa giá trị trung bình thực.

$Y_t^* - Y_{t-1}$: Thay đổi lượng tư bản mong muốn

Từ (6.14) và(6.15) sau một vài phép biến đổi chúng ta nhận được

$$Y_t = \delta\beta_0 + \delta\beta_1 X_t + (1-\delta)Y_{t-1} + \delta\varepsilon_t \qquad (6.17)$$

Một lần nữa chúng ta lại nhận được mô hình tự hồi quy.

6.5. Ước lượng mô hình tự hồi quy

Trong cả ba mô hình vừa xét, chúng ta đều nhận được mô hình cuối cùng có dạng tự hồi quy.

Koyck:

$$Y_t = \alpha(1-\lambda) + \beta_0 X_t + \lambda Y_{t-1} + (\varepsilon_t - \lambda\varepsilon_{t-1}) \qquad (6.18)$$

Kỳ vọng thích nghi

$$Y_t = \gamma\beta_0 + \gamma\beta_1 X_t + (1-\gamma)Y_{t-1} + [\varepsilon_t - (1-\gamma)\varepsilon_{t-1}] \qquad (6.19)$$

Hiệu chỉnh từng phần

$$Y_t = \delta\beta_0 + \delta\beta_1 X_t + (1-\delta)Y_{t-1} + \delta\varepsilon_t \qquad (6.20)$$

Dạng chung của ba mô hình này là

$$Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 Y_{t-1} + \gamma_t \qquad (6.21)$$

Có hai vấn đề cần lưu tâm đối với mô hình (6.21):

(1) Thứ nhất, có sự hiện diện của biến ngẫu nhiên trong các biến độc lập, đó là Y_{t-1}. Điều này vi phạm điều kiện của mô hình hồi quy tuyến tính cổ điển.

(2) Thứ hai, có khả năng xảy ra hiện tượng tương quan chuỗi.

Để tránh các hệ quả bất lợi do Y_{t-1} gây ra người ta sử dụng một biến thay thế cho Y_{t-1} với đặc tính biến này tương quan mạnh với Y_{t-1} nhưng không tương quan với X_t. Biến độc lập có đặc tính vừa kể được gọi là biến công cụ[24].

6.6. Phát hiện tự tương quan trong mô hình tự hồi quy

Trị thống kê h

$$h = \hat{\rho}\sqrt{\frac{n}{1 - n[\text{var}(\hat{\alpha}_2)]}} \qquad (6.22)$$

[24] N.Levitan có đề xuất dùng X_{t-1} làm biến công cụ cho Y_{t-1} và dề xuất một hệ phương trình chuẩn đặc biệt cho ước lượng hệ số, nhưng vấn đề đa cộng tuyến của mô hình cũng không được khắc phục triệt để. (Theo Gujarati, Basic Econometrics, 3rd Edition,Mc Graw-Hill Inc,1995, trang 604-605).

Trong đó: n = cỡ mẫu; $\text{var}(\hat{\alpha}_2)$ = phương sai hệ số ước lượng của Y_{t-1}.

$\hat{\rho}$ là hệ số tự tương quan mẫu bậc nhất được xác định từ công thức

$$\hat{\rho} = \frac{\sum_{t=1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=}^{n} \hat{\varepsilon}_t^2} \qquad (6.23)$$

h có phân phối chuẩn hoá tiệm cận. Từ phân phối chuẩn hoá chúng ta có

$P(-1,96 < h < 1,96) = 0,95$

Quy tắc quyết định:

- √ Nếu $h < -1,96$, chúng ta bác bỏ H_0 cho rằng mô hình không có tự tương quan bậc 1 nghịch.
- √ Nếu $h > 1,96$, chúng ta bác bỏ H_0 cho rằng mô hình không có tự tương quan bậc 1 thuận.
- √ Nếu $-1,96 < h < 1,96$: chúng ta không thể bác bỏ H_0 cho rằng không có tự tương quan bậc nhất.

CHƯƠNG 7

CÁC MÔ HÌNH DỰ BÁO MANG TÍNH THỐNG KÊ (Tham khảo)

7.1. Các thành phần của dữ liệu chuỗi thời gian

Các thành phần chính của dữ liệu chuỗi thời gian là

a. Xu hướng
b. Chu kỳ
c. Thời vụ
d. Ngẫu nhiên

7.1.1. Xu hướng dài hạn

Xu hướng dài hạn thể hiện sự tăng trưởng hoặc giảm sút của một biến số theo thời gian với khoảng thời gian đủ dài. Một số biến số kinh tế có xu hướng tăng giảm dài hạn như

e. Tốc độ tăng dân số của Việt Nam có xu hướng giảm.
f. Tỷ trọng nông nghiệp trong GDP của Việt Nam có xu hướng giảm.
g. Mức giá có xu hướng tăng.

7.1.2. Chu kỳ

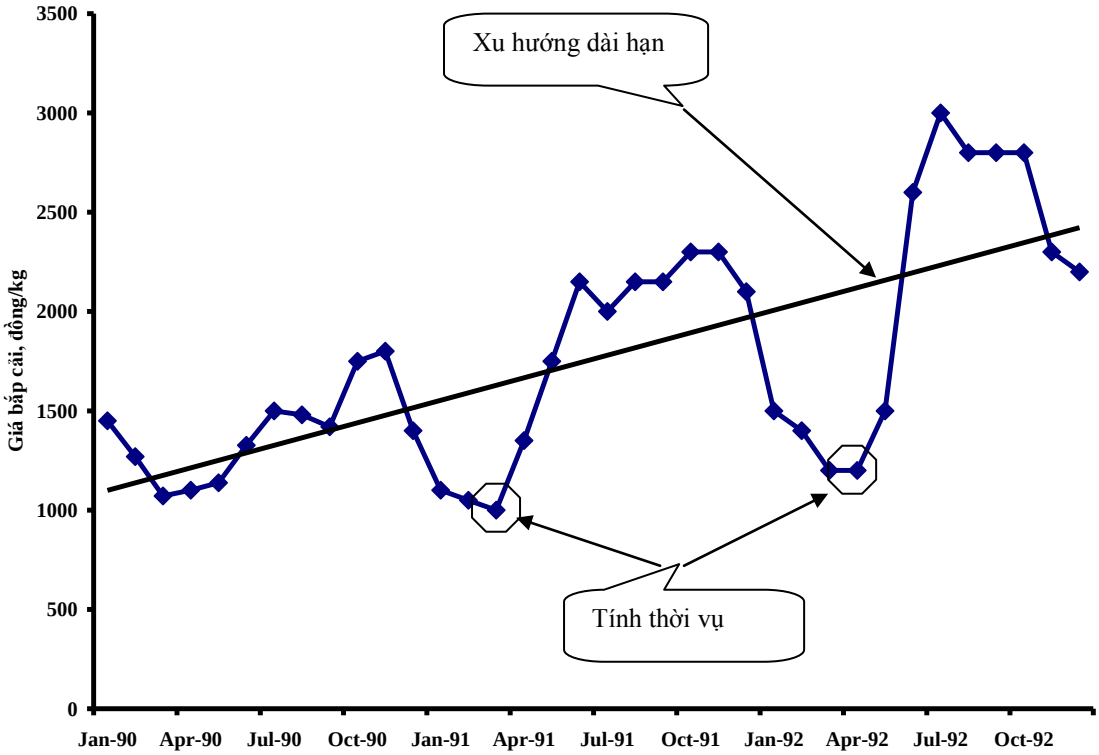
Các số liệu kinh tế vĩ mô thường có sự tăng giảm có quy luật theo chu kỳ kinh tế. Sau một thời kỳ suy thoái kinh tế sẽ là thời kỳ phục hồi và bùng nổ kinh tế, kế tiếp tăng trưởng kinh tế sẽ chựng lại và khởi đầu cho một cuộc suy thoái mới. Tuỳ theo nền kinh tế mà chu kỳ kinh tế có thời hạn là 5 năm, 7 năm hay 10 năm.

7.1.3. Thời vụ

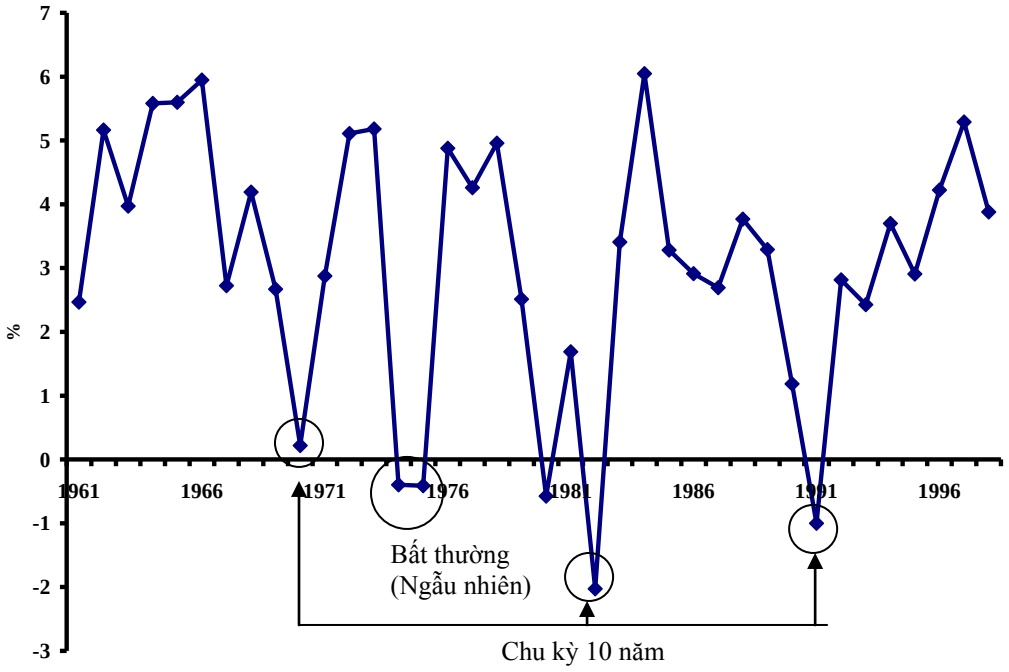
Biến động thời vụ của biến số kinh tế là sự thay đổi lặp đi lặp lại từ năm này sang năm khác theo mùa vụ. Biến động thời vụ xảy ra do khí hậu, ngày lễ, phong tục tập quán…Biến động thời vụ có tính ngắn hạn với chu kỳ lặp lại thường là 1 năm.

7.1.4. Ngẫu nhiên

Những dao động không thuộc ba loại trên được xếp vào dao động ngẫu nhiên. Các nguyên nhân gây ra biến động ngẫu nhiên có thể là thời tiết bất thường, chiến tranh, khủng hoảng năng lượng, biến động chính trị…



Hình 7.1. Xu hướng và thời vụ[25]



Hình 7.2. Chu kỳ và ngẫu nhiên-Tăng trưởng kinh tế của Hoa Kỳ giai đoạn 1961-1999.

Nguồn : World Development Indicator CD-Rom 2000, World Bank.

[25] Nguồn: Problem set 7, Analytic method for Policy Making, Chương trình Giảng dạy Kinh tế Fulbright Việt Nam 2000.

7.2. Dự báo theo đường xu hướng dài hạn

7.2.1. Mô hình xu hướng tuyến tính

Chúng ta sử dụng mô hình xu hướng tuyến tính nếu tin rằng biến Y tăng một lượng không đổi trong một đơn vị thời gian.

$$\hat{Y}_t = \beta_1 + \beta_2 t \qquad (7.1)$$

hoặc dạng

$$\hat{Y}_{n+k} = Y_n + \beta_2 k \qquad (7.2)$$

Ứng với dữ liệu ở hình 7.2, phương trình đường xu hướng là

$$g_t = 3{,}6544 - 0{,}029t$$

Với g_t = tốc độ tăng trưởng GDP của Hoa Kỳ, tính bằng %.

t = năm đang xét- 1991.

Dự báo tốc độ tăng trưởng kinh tế cho năm 2000 là

$$g_{2000} = 3{,}6544 - 0{,}029*(2000 - 1961) = 2{,}52\ \%$$

7.2.2. Mô hình xu hướng dạng mũ

Chúng ta sử dụng hàm mũ khi cho rằng có tỷ lệ tăng trưởng cố định trong một đơn vị thời gian.

$$\hat{Y}_t = \alpha e^{\beta t} \qquad (7.3)$$

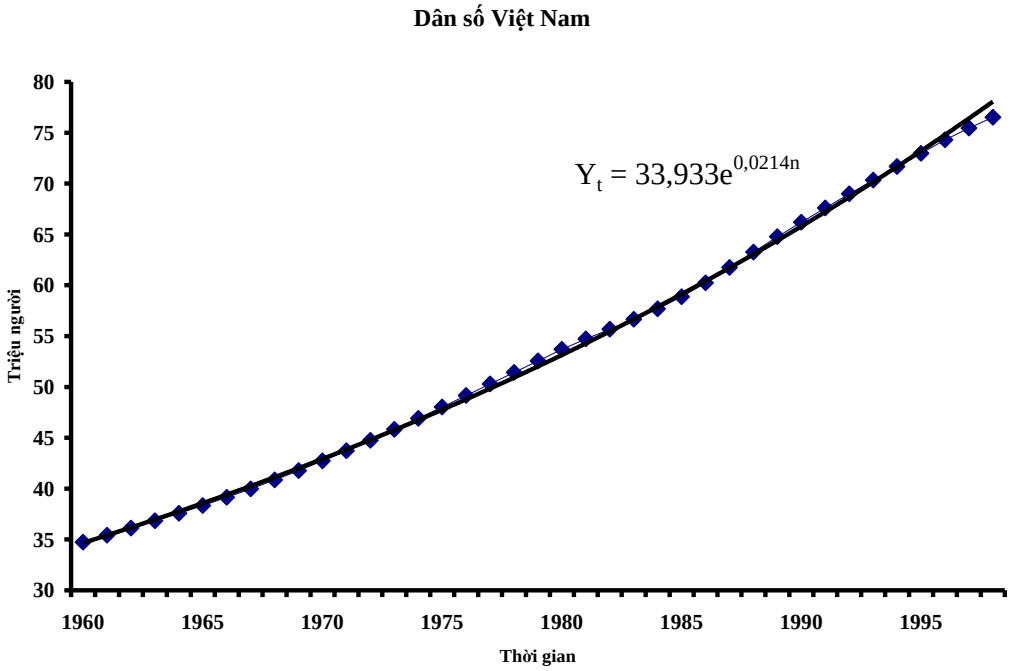
chuyển dạng

$$\ln(\hat{Y}_t) = \ln(\alpha) + \beta \ln t \qquad (7.4)$$

Mô hình xu hướng dạng mũ dùng để dự báo dân số, sản lượng, nhu cầu năng lượng…Hình 7.3 cho thấy dân số của Việt Nam có dạng hàm mũ với phương trình ước lượng như sau:

$$Y_t = 33{,}933e^{0{,}0214n}$$

Từ dạng hàm (7.3), kết quả (7.4) cho thấy tốc độ tăng dân số của Việt Nam trong thời kỳ 1960-1999 khoảng 2,14 %.



Hình 7.3. Dân số Việt Nam giai đoạn 1960-1999

Nguồn : World Development Indicator CD-Rom 2000, World Bank.

7.2.3. Mô hình xu hướng dạng bậc hai

$$\hat{Y}_t = \beta_1 + \beta_2 t + \beta_3 t^2 \qquad (7.5)$$

Dấu của các tham số quyết định dạng đường xu hướng như sau:

- Nếu β_2 và β_3 đều dương: Y tăng nhanh dần theo thời gian.
- Nếu β_2 âm và β_3 dương: Y giảm sau đó tăng
- Nếu β_2 dương và β_3 âm: Y tăng nhưng tốc độ tăng giảm dần sau đó đạt cực trị và bắt đầu giảm.

7.3. Một số kỹ thuật dự báo đơn giản

7.3.1. Trung bình trượt (Moving Average)

Giá trị dự báo bằng trung bình của m giá trị trước đó

$$\hat{Y}_t = \frac{1}{m}(Y_{t-1} + Y_{t-2} + \cdots + Y_{t-m}) \qquad (7.6)$$

Một lưu ý là khi làm trơn chuỗi dữ liệu bằng kỹ thuật trung bình trượt như trên mô hình giảm (m-1) bậc tự do. Chúng ta tạm gác lại việc thảo luận về số số hạng m của mô hình trung bình trượt (7.6).

7.3.2. San bằng số mũ (Exponential Smoothing Method)[26]

Ý tưởng của mô hình san bằng số mũ tương tự mô hình kỳ vọng thích nghi mà chúng ta đã xét ở chương 6. Giá trị dự báo mới không chỉ phụ thuộc vào giá trị giai đoạn trước mà còn phụ thuộc giá trị dự báo của giai đoạn trước.

$$\hat{Y}_t = \alpha Y_{t-1} + (1-\alpha)\hat{Y}_{t-1} \qquad (7.7.a)$$

hoặc

$$\hat{Y}_t = \hat{Y}_{t-1} + \alpha(Y_{t-1} - \hat{Y}_{t-1}) \qquad (7.7.b)$$

- α càng gần 1 thì dự báo mới càng gần với giá trị gần nhất, nếu α càng gần 0 thì dự báo mới càng gần với dự báo gần nhất. Trong thực tế người ta sẽ thử với các giá trị α khác nhau, giá trị được chọn là giá trị làm cho sai số dự báo bình phương trung bình(MSE) của mô hình nhỏ nhất.
- Có thể dùng trung bình của 5 đến 6 số đầu tiên để làm giá trị dự báo đầu tiên[27].

7.3.3. Tự hồi quy (Autoregression)

Giá trị dự báo được xác định từ mô hình tự hồi quy với m độ trễ.

$$\hat{Y}_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_n Y_{t-m} \qquad (7.8)$$

Trong mô hình (7.7) có thể có số β_0 hoặc không có β_0. Trường hợp có β_0 ứng với dữ liệu có xu hướng dài hạn tăng hoặc giảm, trường hợp không có β_0 ứng với dữ liệu có tính dừng[28].

7.4. Tiêu chuẩn đánh giá mô hình dự báo

Gọi $\hat{Y}_t$ là giá trị dự báo cho Y_t. Sai số của dự báo là $\varepsilon_t = Y_t - \hat{Y}_t$.

Hai tiêu chuẩn thường được sử dụng để đánh giá và so sánh các mô hình dự báo là

Sai số dự báo tuyệt đối trung bình(Mean absolute deviation-MAD)

$$MAD = \frac{\sum_{t=1}^{n} \left| Y_t - \hat{Y}_t \right|}{n} \qquad (7.9)$$

[26] Phương pháp dự báo này còn được gọi là phương pháp Holt.

[27] Theo Loan Lê, Hệ thống dự báo điều khiển kế hoạch ra quyết định, NXB Thống Kê-2001, trang 307-308.

[28] Chúng ta sẽ thảo luận về tính dừng khi nghiên cứu mô hình ARIMA.

Sai số dự báo bình phương trung bình(Mean squared error-MSE)

$$MSE = \frac{\sum_{t=1}^{n} \left(Y_t - \hat{Y}_t\right)^2}{n} \qquad (7.10)$$

Mô hình tốt là mô hình có MAD và MSE nhỏ.

7.5. Một ví dụ bằng số

Sử dụng số liệu giá bắp cải đến tháng 12/1992(hình7.1), chúng ta lập mô hình dự báo giá bắp cải và dự báo cho các tháng của năm 1993.

Mô hình 1: Lin

Xu hướng tuyến tính: $\hat{Y}_t = \alpha_0 + \alpha_1 k$ với k là số thứ tự của thời kỳ t.

Mô hình 2: MA

Trung bình trượt: $\hat{Y}_t = \frac{Y_{t-1} + Y_{t-2}}{2}$

Mô hình 3: Holt

Phương pháp Holt: $\hat{Y}_t = \hat{Y}_{t-1} + \alpha(Y_{t-1} - \hat{Y}_{t-1})$ với $\alpha = 0{,}6$.

Mô hình 4: AR

Tự hồi quy: $\hat{Y}_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2}$

Sau khi ước lượng các hệ số của mô hình 1 và 4 dựa trên số liệu đến hết 1992(trong mẫu), chúng ta ước lượng cho cả giai đoạn trước 1993(trong mẫu) và 1993(ngoài mẫu). Chúng ta vẽ đồ thị các dãy số liệu dự báo và số liệu gốc như ở hình 7.5.

Kết quả tính toán sai số của các mô hình như sau:

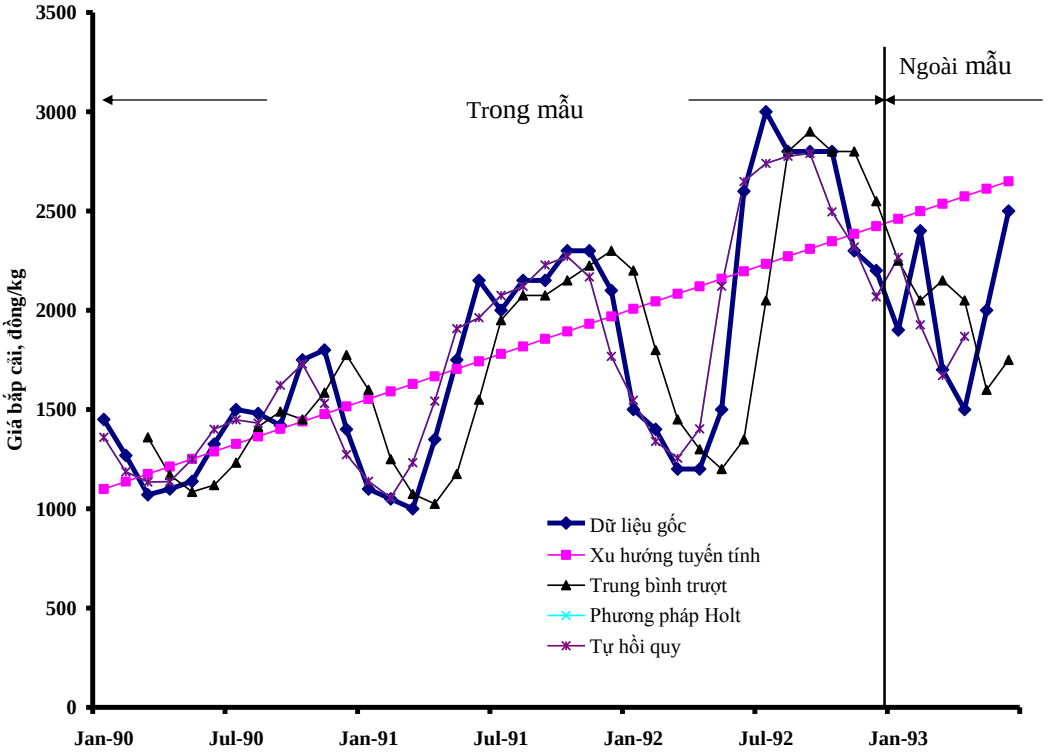
Trong mẫu:

Mô hình	Lin	MA	Holt	AR
MSE trong mẫu, đồng^2	2.733	157	2.216	59.629

Ngoài mẫu

Mô hình	Lin	MA	Holt	AR
MSE dự báo, đồng^2	429.043	245.417	216.134	260.392

Trong trường hợp cụ thể của ví dụ này mô trung bình trượt(MA) cho MSE trong mẫu nhỏ nhất nhưng phương pháp Holt lại cho MSE nhỏ nhất ngoài mẫu.



Hình 7.4. Các phương pháp dự báo đơn giản

7.6. Giới thiệu mô hình ARIMA

7.6.1. Tính dừng của dữ liệu

Quá trình ngẫu nhiên(Stochastic process)

Bất cứ dữ liệu chuỗi thời gian nào cũng được tạo ra bằng một quá trình ngẫu nhiên. Một dãy số liệu thực tế cụ thể như giá bắp cải từng tháng ở hình 7.1 là kết quả của một quá trình ngẫu nhiên. Đối với dữ liệu chuỗi thời gian, chúng ta có những khái niệm về tổng thể và mẫu như sau:

- Quá trình ngẫu nhiên là một tổng thể.
- Số liệu thực tế sinh ra từ quá trình ngẫu nhiên là mẫu.

Tính dừng(Stationary)

Một quá trình ngẫu nhiên được gọi là có tính dừng khi nó có các tính chất sau:

- Kỳ vọng không đổi theo thời gian, $E(Y_t) = \mu$.
- Phương sai không đổi theo thời gian, $Var(Y_t) = E(Y_t-\mu) = \sigma^2$.
- Đồng phương sai chỉ phụ thuộc khoảng cách của độ trễ mà không phụ thuộc thời điểm tính đồng phương sai đó, $\nu_k = E[(Y_t-\mu)(Y_{t-k}-\mu)]$ không phụ thuộc t.

Lưu ý: Chúng ta có thể biến dữ liệu chuỗi thời gian từ không có tính dừng thành có tính dừng bằng cách lấy sai phân của nó.

$w_t = Y_t - Y_{t-1}$: Sai phân bậc nhất

$w_t^2 = w_t - w_{t-1}$ **: Sai phân bậc hai…**

7.6.2. Hàm tự tương quan và hàm tự tương quan mẫu

Hàm tự tương quan(ACF) ở độ trễ k được ký hiệu là ρ_k được định nghĩa như sau:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{E[(Y_t - \mu)(Y_{t-k} - \mu)]}{E[(Y_t - \mu)^2]} \qquad (7.11)$$

Tính chất của ACF

- ρ_k không có thứ nguyên.
- Giá trị của ρ_k nằm giữa -1 và 1.

Trong thực tế chúng ta chỉ có thể có số liệu thực tế là kết quả của quá trình ngẫu nhiên, do đó chúng chỉ có thể tính toán được hàm tự tương quan mẫu(SAC), ký hiệu là r_k.

$$r_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} \quad \text{với}$$

$$\hat{\gamma}_k = \frac{\sum (Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{n} \quad \text{và} \quad \hat{\gamma}_0 = \frac{\sum (Y_t - \overline{Y})^2}{n}$$

Độ lệch chuẩn hệ số tự tương quan mẫu

$$s(r_j) = \frac{\sqrt{1 + 2\sum_{i=1}^{j-1} r_i^2}}{\sqrt{n}} \qquad (7.12)$$

Trị thống kê t

$$t_k = \frac{r_k}{s(r_k)} \qquad (7.13)$$

Với cỡ mẫu lớn thì $t_k \sim Z$ nên với $t > 1{,}96$ thì r_k khác không có ý nghĩa thống kê, khi đó người ta gọi r_k là 1 đỉnh.

Các phần mềm kinh tế lượng sẽ tính toán cho chúng ta kết quả của SAC và các giá trị đến hạn(hoặc trị thống kê t) của nó ứng với mức ý nghĩa $\alpha = 5\%$.

Thống kê Ljung-Box

$$LB = n(n+2)\sum_{k=1}^{m}\left(\frac{r_k^2}{n-k}\right) \sim \chi^2_m \qquad (7.14)$$

n là cỡ mẫu

m là chiều dài của độ trễ

H_0: Tất cả các r_k đều bằng 0.

H_1: Không phải tất cả các r_k đều bằng 0.

Nếu LB > $\chi^2_{m,1-\alpha}$ thì ta bác bỏ H_0.

Một số phần mềm kinh tế lượng có tính toán trị thống kê LB.

7.6.3. Hàm tự tương quan riêng phần (PACF)

Hệ số tự tương quan riêng phần với độ trễ k đo lường tương quan của Y_{t-k} với Y_t sau khi loại trừ tác động tương quan của tất các các độ trễ trung gian. Công thức tính PACF như sau

$$r_{kk} = \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-j,j} r_j} \qquad (7.15)$$

Độ lệch chuẩn của r_{kk} [29]

$$s(r_{kk}) = \frac{1}{\sqrt{n}} \qquad (7.16)$$

Trị thống kê t

$$t_{kk} = \frac{r_{kk}}{s(r_{kk})} \qquad (7.17)$$

Với cỡ mẫu lớn thì $t_{kk} \sim Z$ nên với $t_{kk} > 1{,}96$ thì r_{kk} khác không có ý nghĩa thống kê, khi đó người ta gọi r_{kk} là 1 đỉnh.

Các chương trình kinh tế lượng có thể tính toán cho chúng ta các giá trị PACF, các giá trị tới hạn hay trị thống kê t.

[29] Công thức tính độ lệch chuẩn của r_{kk} phụ thuộc vào bậc của sai phân. Công thức trình bày ở trên là công thức gần đúng với số quan sát đủ lớn.

7.6.4. Mô hình AR, MA và ARMA

Xét quá trình ngẫu nhiên có tính dừng với dữ liệu chuỗi thời gian Y_t có $E(Y_t) = \mu$ và sai số ngẫu nhiên ε_t có trung bình bằng 0 và phương sai σ^2(nhiễu trắng).

Mô hình tự hồi quy (AR-Autoregressive Model)

Mô hình tự hồi quy bậc p được ký hiệu là AR(p) có dạng

$$(Y_t - \mu) = \alpha_1(Y_{t-1} - \mu) + \alpha_2(Y_{t-2} - \mu) + \cdots + \alpha_p(Y_{t-p} - \mu) + \varepsilon_t$$

$$Y_t = \mu(1 - \alpha_1 - \alpha_2 - \cdots - \alpha_p) + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t \quad (7.17)$$

Nhận dạng mô hình AR(p): PACF có đỉnh đến độ trễ p và SAC suy giảm nhanh ngay sau độ trễ thứ nhất thì mô hình dự báo có dạng tự hồi quy bậc p.

Mô hình trung bình trượt(MA-Moving average Model)

Mô hình trung bình trượt bậc q được ký hiệu là MA(q) có dạng

$$Y_t = \mu + \varepsilon_t + \beta_1\varepsilon_{t-1} + \cdots + \beta_q\varepsilon_{t-q} \quad (7.18)$$

với μ là hằng số, ε_t là nhiễu trắng.

Nhận dạng mô hình MA(q): SAC có đỉnh đến độ trễ q và SPAC suy giảm nhanh ngay sau độ trễ thứ nhất.

Mô hình kết hợp tự hồi quy kết hợp trung bình trượt(ARMA)

Mô hình có tự hồi quy bậc p và trung bình trượt bậc q được ký hiệu là ARMA(p,q) có dạng

$$Y_t = \delta + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t + \beta_1\varepsilon_{t-1} + \cdots + \beta_q\varepsilon_{t-q} \quad (7.19)$$

Nhận dạng mô hình ARMA(p,q): cả SAC và SPAC đều có giá trị giảm dần theo hàm mũ. Nhận dạng đúng p và q đòi hỏi phải có nhiều kinh nghiệm. Trong thực hành người ta chọn một vài mô hình ARMA và lựa chọn mô hình tốt nhất.

7.6.5. Mô hình ARIMA và SARIMA

ARIMA

Đa số dữ liệu kinh tế theo chuỗi thời gian không có tính dừng(stationary) mà có tính kết hợp(integrated). Để nhận được dữ liệu có tính dừng, chúng ta phải sử dụng sai phân của dữ liệu.

Các bậc sai phân

Sai phân bậc 0 là I(0): chính là dữ liệu gốc Y_t.

Sai phân bậc 1 là I(1): $w_t = Y_t - Y_{t-1}$.

Sai phân bậc 2 là I(2): $w^2_t = w_t - w_{t-1}$ …

Sai phân bậc d ký hiệu I(d).

Mô hình ARMA(p,q) áp dụng cho I(d) được gọi là mô hình ARIMA(p,d,q).

SARIMA

Trong mô hình ARIMA nếu chúng ta tính toán sai phân bậc nhất với độ trễ lớn hơn 1 để khử tính mùa vụ như sau $w_t = Y_t - Y_{t-s}$, với s là số kỳ giữa các mùa thì mô hình được gọi là SARIMA hay ARIMA có tính mùa vụ.

7.6.6. Phương pháp luận Box-Jenkins

Phương pháp luận Box-Jenkins cho mô hình ARIMA có bốn bước như sau:

Bước 1: Xác lập mô hình ARIMA(p,d,q)

- Dùng các đồ thị để xác định bậc sai phân cần thiết để đồ thị có tính dừng. Giả sử dữ liệu dùng ở I(d). Dùng đồ thị SAC và SPAC của I(d) để xác định p và q.

- Triển khai dạng của mô hình.

Bước 2: Tính toán các tham số của mô hình.

Trong một số dạng ARIMA đơn giản chúng ta có thể dùng phương pháp bình phương tối thiểu. Một số dạng ARIMA phức tạp đòi hỏi phải sử dụng các ước lượng phi tuyến. Chúng ta không phải lo lắng về việc ước lượng tham số vì các phần mềm kinh tế lượng sẽ tính giúp chúng ta. Quay lại bước 1 xây dựng mô hình với cặp (p,q) khác dường như cũng phù hợp. Giả sử chúng ta ước lượng được m mô hình ARIMA.

Bước 3: Kiểm tra chẩn đoán

So sánh các mô hình ARIMA đã ước lượng với các mô hình truyền thống(tuyến tính, đường xu hướng, san bằng số mũ,…) và giữa các mô hình ARIMA với nhau để chọn mô hình tốt nhất.

Bước 4: Dự báo

Trong đa số trường hợp mô hình ARIMA cho kết quả dự báo ngắn hạn đáng tin cậy nhất trong các phương pháp dự báo. Tuy nhiên giới hạn của của ARIMA là:

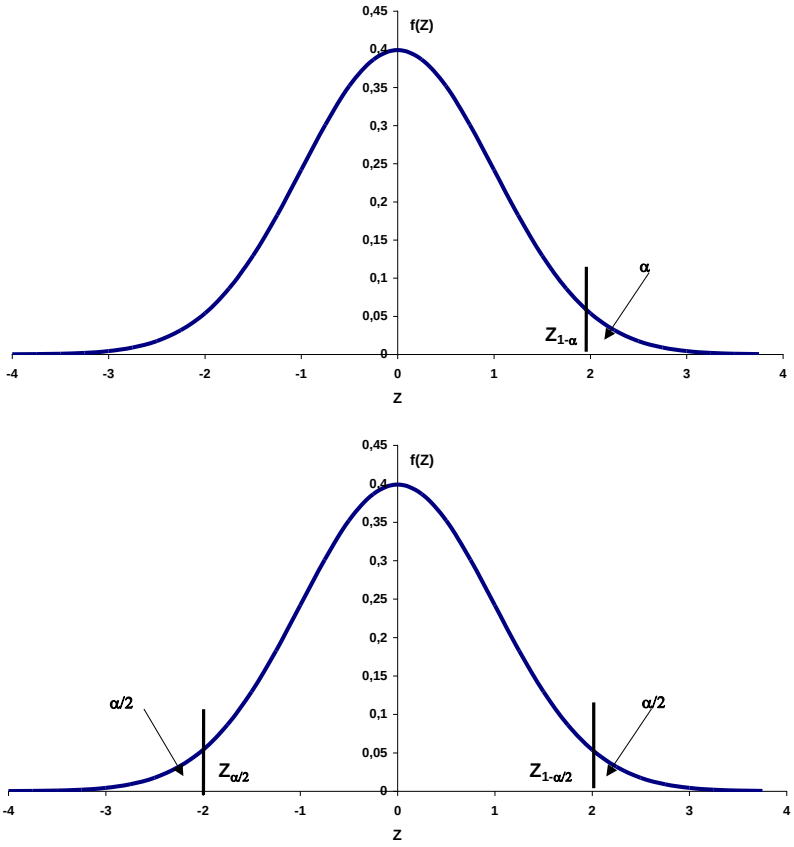
- Số quan sát cần cho dự báo phải lớn.

- Chỉ dùng để dự báo ngắn hạn

- Không thể đưa các yếu tố thay đổi có ảnh hưởng đến biến số cần dự báo của thời kỳ cần dự báo vào mô hình.

Xây dựng mô hình ARIMA theo phương pháp luận Box-Jenkins có tính chất nghệ thuật hơn là khoa học, hơn nữa kỹ thuật và khối lượng tính toán khá lớn nên đòi hỏi phải có phần mềm kinh tế lượng chuyên dùng.

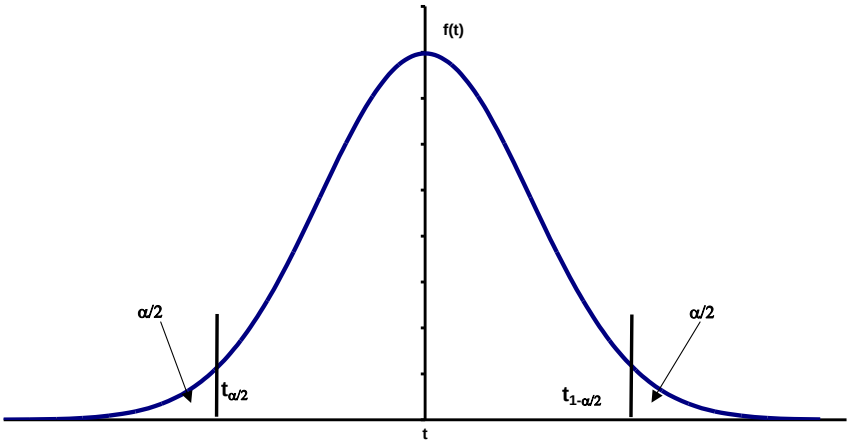
MỘT SỐ GIÁ TRỊ Z THƯỜNG ĐƯỢC SỬ DỤNG



Mức ý nghĩa	Kiểm định 1 đuôi	Kiểm định 2 đuôi
α	$Z_{1-\alpha}$	$Z_{1-\alpha/2}$
1%	2,326	2,576
5%	1,645	1,960
10%	1,282	1,645
20%	0,842	1,282

Nguồn: hàm Normsinv của Excel.

MỘT SỐ GIÁ TRỊ t THƯỜNG ĐƯỢC SỬ DỤNG

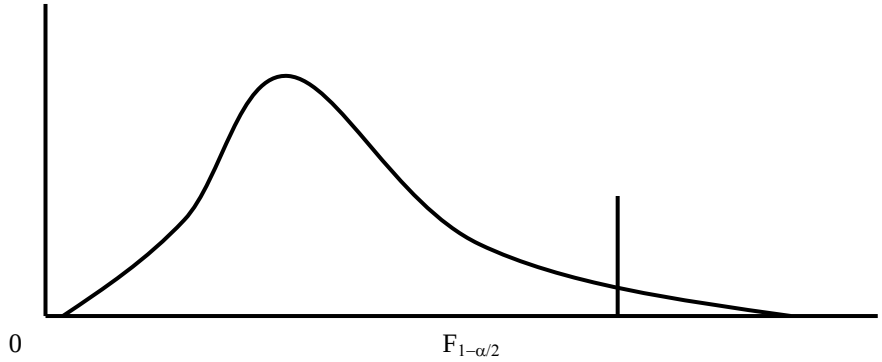


Bậc tự do	Mức ý nghĩa α			
	1%	5%	10%	20%
1	63,656	12,706	6,314	3,078
2	9,925	4,303	2,920	1,886
3	5,841	3,182	2,353	1,638
4	4,604	2,776	2,132	1,533
5	4,032	2,571	2,015	1,476
6	3,707	2,447	1,943	1,440
7	3,499	2,365	1,895	1,415
8	3,355	2,306	1,860	1,397
9	3,250	2,262	1,833	1,383
10	3,169	2,228	1,812	1,372
11	3,106	2,201	1,796	1,363
12	3,055	2,179	1,782	1,356
13	3,012	2,160	1,771	1,350
14	2,977	2,145	1,761	1,345
15	2,947	2,131	1,753	1,341
16	2,921	2,120	1,746	1,337
17	2,898	2,110	1,740	1,333
18	2,878	2,101	1,734	1,330
19	2,861	2,093	1,729	1,328
20	2,845	2,086	1,725	1,325
21	2,831	2,080	1,721	1,323
22	2,819	2,074	1,717	1,321
23	2,807	2,069	1,714	1,319
24	2,797	2,064	1,711	1,318
25	2,787	2,060	1,708	1,316
26	2,779	2,056	1,706	1,315
27	2,771	2,052	1,703	1,314
28	2,763	2,048	1,701	1,313
29	2,756	2,045	1,699	1,311
30	2,750	2,042	1,697	1,310
>30	2,576	1,960	1,645	1,282

Nguồn: hàm Tinv của Excel.

MỘT SỐ GIÁ TRỊ F TỚI HẠN TRÊN THƯỜNG ĐƯỢC SỬ DỤNG

Mức ý nghĩa $\alpha = 5\%$

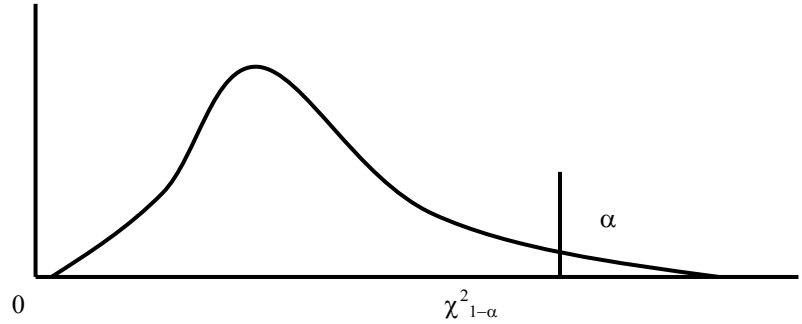


	df1									
df2	1	2	3	4	5	6	7	8	9	10
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
31	4,16	3,30	2,91	2,68	2,52	2,41	2,32	2,25	2,20	2,15
32	4,15	3,29	2,90	2,67	2,51	2,40	2,31	2,24	2,19	2,14
33	4,14	3,28	2,89	2,66	2,50	2,39	2,30	2,23	2,18	2,13
34	4,13	3,28	2,88	2,65	2,49	2,38	2,29	2,23	2,17	2,12
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11
36	4,11	3,26	2,87	2,63	2,48	2,36	2,28	2,21	2,15	2,11
37	4,11	3,25	2,86	2,63	2,47	2,36	2,27	2,20	2,14	2,10
38	4,10	3,24	2,85	2,62	2,46	2,35	2,26	2,19	2,14	2,09
39	4,09	3,24	2,85	2,61	2,46	2,34	2,26	2,19	2,13	2,08
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08

Nguồn: hàm Finv của Excel.

MỘT SỐ GIÁ TRỊ χ^2 TỚI HẠN TRÊN THƯỜNG ĐƯỢC SỬ DỤNG

Mức ý nghĩa $\alpha = 5\%$



	α			
df	1%	5%	10%	20%
2	9,21	5,99	4,61	3,22
3	11,34	7,81	6,25	4,64
4	13,28	9,49	7,78	5,99
5	15,09	11,07	9,24	7,29
6	16,81	12,59	10,64	8,56
7	18,48	14,07	12,02	9,80
8	20,09	15,51	13,36	11,03
9	21,67	16,92	14,68	12,24
10	23,21	18,31	15,99	13,44
11	24,73	19,68	17,28	14,63
12	26,22	21,03	18,55	15,81
13	27,69	22,36	19,81	16,98
14	29,14	23,68	21,06	18,15
15	30,58	25,00	22,31	19,31
16	32,00	26,30	23,54	20,47
17	33,41	27,59	24,77	21,61
18	34,81	28,87	25,99	22,76
19	36,19	30,14	27,20	23,90
20	37,57	31,41	28,41	25,04
21	38,93	32,67	29,62	26,17
22	40,29	33,92	30,81	27,30
23	41,64	35,17	32,01	28,43
24	42,98	36,42	33,20	29,55
25	44,31	37,65	34,38	30,68
26	45,64	38,89	35,56	31,79
27	46,96	40,11	36,74	32,91
28	48,28	41,34	37,92	34,03
29	49,59	42,56	39,09	35,14
30	50,89	43,77	40,26	36,25
31	52,19	44,99	41,42	37,36
32	53,49	46,19	42,58	38,47
33	54,78	47,40	43,75	39,57
34	56,06	48,60	44,90	40,68
35	57,34	49,80	46,06	41,78
36	58,62	51,00	47,21	42,88
37	59,89	52,19	48,36	43,98
38	61,16	53,38	49,51	45,08
39	62,43	54,57	50,66	46,17
40	63,69	55,76	51,81	47,27

Nguồn: Hàm Chiinv của Excel

TÀI LIỆU THAM KHẢO

1) PGS.TS. Vũ Thiếu, TS. Nguyễn Quang Dong, TS. Nguyễn Khắc Minh
Kinh tế lượng
NXB Khoa học và Kỹ thuật Hà nội-1996

2) TS. Bùi Phúc Trung
Giáo trình Kinh tế lượng
Trường Đại học Kinh tế TP Hồ Chí Minh-2001

3) TS. Nguyễn Thống
Kinh tế lượng ứng dụng
NXB Đại học Quốc gia TP Hồ Chí Minh-2000

4) TS. Nguyễn Quang Dong
Bài tập Kinh tế lượng với sự trợ giúp của phần mềm Eviews
NXB Khoa học và kỹ thuật-2002

5) TS. Nguyễn Quang Dong
Kinh tế lượng nâng cao
NXB Khoa học và kỹ thuật-2002

6) Loan Lê
Hệ thống dự báo điều khiển kế hoạch ra quyết định
NXB Thống Kê-2001

7) Lê Thanh Phong
Hướng dẫn sử dụng SPSS for Windows V.10
Đại học Cần Thơ-2001

8) PGS. Đặng Hấn
Xác suất thống kê
NXB Thống kê-1996

9) PGS. Đặng Hấn
Bài tập xác suất thống kê
NXB Thống kê-1996

10) Nguyễn Đình Trí, Tạ Văn Dĩnh và Nguyễn Hồ Quỳnh
Toán học cao cấp
NXB Giáo Dục-1998

11) Đỗ Công Khanh
Giải tích một biến
Tủ sách Đại học đại cương TP Hồ Chí Minh-1997

12) Đỗ Công Khanh
Giải tích nhiều biến
Tủ sách Đại học đại cương TP Hồ Chí Minh-1997

13) Bùi Văn Mưa
Logic học
Đại học Kinh tế TP Hồ Chí Minh-1998

14) Cao Hào Thi, Lê Nguyễn Hậu, Tạ Trí Nhân, Võ Văn Huy và Nguyễn Quỳnh Mai
Crystal Ball- Dự báo và phân tích rủi ro cho những người sử dụng bảng tính
Chương trình giảng dạy kinh tế Fulbright Việt nam-1995

15) Đoàn Văn Xê
Kinh tế lượng
Đại học Cần thơ 1993

16) Ban biên dịch First News
EXCEL toàn tập

Nhà Xuất Bản Trẻ-2001
17) TS.Phan Hiếu Hiền
Phương pháp bố trí thí nghiệm và xử lý số liệu(Thống kê thực nghiệm)
NXB Nông Nghiệp 2001.
18) Chris Brooks
Introductory Econometrics for Finance
Cambridge University Press-2002
19) A.Koutsoyiannis
Theory of Econometrics-Second Edition
ELBS with Macmillan-1996
20) Damodar N. Gujarati
Basic Econometrics-Second Edition
McGraw-Hill Inc -1988
21) Damodar N. Gujarati
Basic Econometrics-Third Edition
McGraw-Hill Inc -1995
22) Damodar N. Gujarati
Basic Econometrics-Student solutions manual to accompany
McGraw-Hill Inc-1988
23) Ernst R. Berndt
The Practice of Econometrics: Classic and Contemporary
MIT-1991
24) William E. Griffiths, R. Carter Hill, George G.Judge
Learning and Practicing Econometrics
John Wiley & Sons-1993
25) Daniel Westbrook
Applied Econometrics with Eviews
Fulbright Economics Teaching Program-2002
26) Ramu Ramanathan
Introductory Econometrics with Applications
Harcourt College Publishers-2002
27) Robert S.Pindyck and Daniel L.Rubinfeld
Econometric Models and Economics Forcasts-Third Edition
McGraw-Hill Inc-1991
28) Kwangchai A.Gomez and Arturo A.Gomez
Statistical Procedures for Agricultural Research
John Wiley & Sons-1983
29) Chandan Mukherjee, Howard White and Marc Wuyts
Data Analysis in Development Economics
Draft -1995
30) Aswath Damodaran
Corporate Finance-Theory and Practice
John Willey & Sons, Inc - 1997