CHƯƠNG 1: ÔN TẬP

1.1. Trung bình mẫu – Phương sai mẫu

1.1.1. Trung bình mẫu

Trong phân tích dữ liệu, cũng như trong cuộc sống hàng ngày, chúng ta thường nói đến chiều cao trung bình, thu nhập trung bình, vân vân. Đó chính là trung bình mẫu. Hãy xét ví dụ sau:

Ví dụ 1.1: Bảng quan sát nhiệt độ ở Đà Lạt

Thứ 2 (x_1)	Thứ 3 (x_2)	Thứ 4 (x_3)	Thứ 5 (x_4)
19^o	21^o	20^o	18^o

$$\Rightarrow \bar{x} = \frac{1}{4}(19+21+20+18) = 19.5^o$$

Một cách khái quát, *trung bình mẫu* được tính bằng công thức sau:

$$\bar{x} = \frac{1}{N}(x_1 + x_2 + x_3 + ... + x_N)$$

$$\textit{Hay:} \quad \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

1.1.2. Phương sai mẫu

Phương sai mẫu **[**ký hiệu s_X^2**]** bằng trung bình của tổng bình phương độ lệch giữa giá trị quan sát so với giá trị trung bình:

$$s_X^2 = \frac{1}{N}\left[(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + ...(x_N-\bar{x})^2\right]$$

$$\textit{Hay:} \quad s_X^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})^2$$

Chẳng hạn, về trung bình mà nói thì khí hậu ở sa mạc rất nóng. Hơn nữa nhiệt độ giao động rất lớn giữa ngày và đêm. Để thể hiện được sự khắc nghiệt của khí hậu sa mạc, chúng ta không những chỉ sử dụng trung bình (mẫu) về nhiệt độ, mà cả sự giao

động của nhiệt độ theo từng thời điểm so với trung bình. Đó chính là khái niệm về phương sai mẫu nói trên.

1.2. Hàm mật độ xác suất, hàm phân bố xác suất

1.2.1. Tần suất và xác suất

Để có sự hình dung về tần suất, hãy xét ví dụ sau:

Ví dụ 1.2: Xếp hạng tốc độ gia tăng giá cổ phiếu trên thị trường chứng khoán Việt Nam.

Gọi X là tỉ lệ phần trăm mức tăng giá cổ phiếu trung bình trong 3 tháng đầu tiên sau khi "lên sàn"; gọi P là phần trăm các công ty có mức tăng giá cổ phiếu tương ứng với giá trị của X

	X	Y
(x_1)	50%	10%
(x_2)	40%	20%
(x_3)	30%	35%
(x_4)	20%	25%

Con số P= 10%, X= 50% có nghĩa là có 10% trong tổng số các công ty có mức tăng giá trong 3 tháng đầu sau khi phát hành cổ phiếu ra công chúng là 50%. Đó chính là ví dụ về tần suất

Ví dụ 1.3: Trò chơi tung đồng xu.

Giả sử bạn tham gia cuộc chơi tung đồng xu tại hội chợ. Nếu là mặt sấp, bạn sẽ được $100. Ngược lại, nếu là mặt ngửa, bạn được $0. Với thể lệ đó, bạn sẵn sàng trả bao nhiêu đôla để tham gia trò chơi?

Để cho tiện, hãy kí hiệu mặt sấp là 1, mặt ngửa là 0. Giả sử kết quả tung xu sau 10 lần là như sau:

X	P
1	3/10
0	7/10

Con số 3/10 chính là tần suất xuất hiện mặt sấp (X = 1). Nghĩa là, trong 10 lần tung xu, có 3 lần xuất hiện mặt sấp. Và do đó, có 7 lần xuất hiện mặt ngửa.

Số tiền bạn bỏ ra cho việc tham dự 10 lần tung xu là: $50 x 10 = $500.
Số tiền nhận được trong cuộc chơi: $100 x 3 + $0 x 7 = $300.

→ Do vậy, cuộc chơi không hứng thú đối với bạn ($500 > $300).

Tuy nhiên, nếu giả sử rằng bạn tham dự cuộc chơi vô hạn lần. Khi đó, số lần xuất hiện mặt sấp và mặt ngửa là như nhau, và bằng ½. Khi đó, kỳ vọng đượccuộc sẽ là: $100x1/2 + $0x1/2 = $50; và bằng chính số tiền lớn nhất bạn sẵn sàng trả để tham dự cuộc chơi.

Điều chúng ta cần phân biệt là con số P = 3/10 trong ví dụ nêu trên là *tần suất* xuất hiện mặt sấp trong 10 lần thử. Và con số ½ là *xác suất* xuất hiện mặt sấp (hoặc ngửa). Khái niệm tần suất ứng với từng mẫu thử; còn xác suất tương ứng với tổng thể.

1.2.2. Biến ngẫu nhiên rời rạc và liên tục

2.2.1. Biến ngẫu nhiên rời rạc:

Một biến ngẫu nhiên là **rời rạc** nếu các giá trị có thể có của nó lập nên một tập hợp hữu hạn hoặc đếm được, nghĩa là có thể liệt kê được tất cả các giá trị có thể có của nó.

Cuộc chơi tung xu nêu trên là ví dụ về biến ngẫu nhiên rời rạc.

Một cách hình thức hóa, ta có thể nói như sau. Giả sử đối tượng quan sát X có thể xuất hiện trong K sự kiện khác nhau [trong ví dụ tung xu, K = 2]. Ta ký hiệu các sự kiện đó là $x_1, x_2, ..., x_K$.

Tần suất xuất hiện một biến cố x_k trong N phép thử, ký hiệu là p_k, là tỉ số giữa số lần xuất hiện biến cố cụ thể đó so với N phép thử được thực hiện.

Với mọi chỉ số, $k = 1,2,3,..., K$, ta có thể viết như sau:

X	x_1	x_2	x_3	…	x_K
P	p_1	p_2	p_3	…	p_K

$p_1, p_2, p_3, \ldots p_K > 0$, và

$p_1 + p_2 + p_3 + \ldots\ldots + p_K = 1$, hay cũng vậy,

$$\sum_{k=1}^{K} p_k = 1$$

Nếu số mẫu N là đủ lớn (tiến đến vô hạn), khái niệm tần suất xuất hiện một biến cố được thay bằng khái niệm **xác suất** xuất hiện biến cố, ký hiệu bởi: $f_k = f(x_k), k = 1,2,.., K.$ Trong đó, $f(x_k)$ là hàm mật độ xác suất của $x_k, k = 1,2...K.$

Ta cũng có,

$f_1, f_2, f_3, \ldots f_K > 0$, và

$$\sum_{k=1}^{K} f_k = 1$$

2.2.2. Biến ngẫu nhiên liên tục

Một biến ngẫu nhiên là liên tục nếu các giá trị có thể có của nó lắp đầy một khoảng trên trục số, nghĩa là không thể liệt kê và đếm được tất cả các giá trị có thể có của nó.

Tương tự với trường hợp phân bố xác suất rời rạc, nếu gọi X là một biến ngẫu nhiên liên tục; và f(x) là hàm mật độ xác suất của X. Khi đó:

$$f(x) \geq 0$$
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Ta định nghĩa hàm phân bố xác suất của X là:

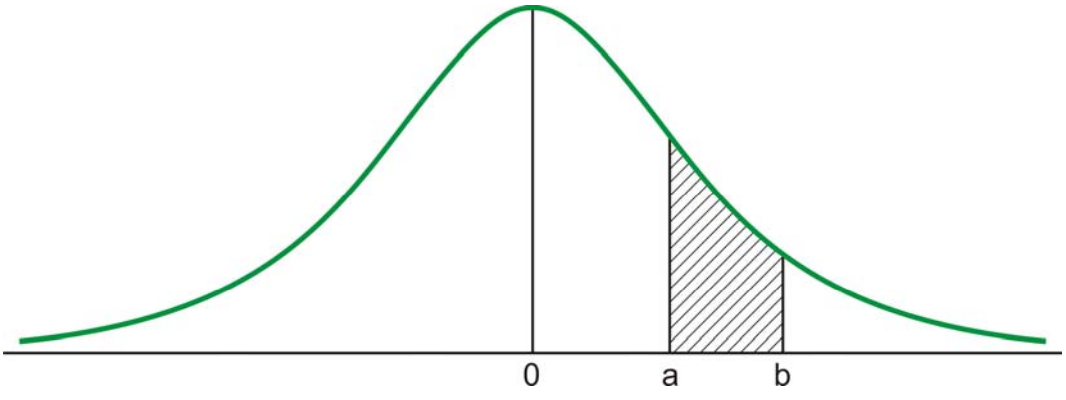
$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Điều đó có nghĩa là, xác suất của biến ngẫu nhiên X nhận giá trị trong khoảng $[a,b]$ sẽ là:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

Ví dụ, trong phân bố chuẩn, về đồ thị ta có thể biểu diễn công thức tính xác suất này như sau:

Đồ thị 1.1: Phân bố xác suất



Phần tô đậm chính là xác suất $P(a \le X \le b)$, được tính bởi tích phân:

$\int_a^b f(x)dx = F(b) - F(a)$.

1.3. Phân bố xác suất đồng thời

Nhiều khi chúng ta muốn đưa ra một đánh giá xác suất đồng thời cho một số biến lượng ngẫu nhiên. Ví dụ, bảng thống kê có ghi lại dữ kiện về thất nghiệp (u) và lạm phát (π). Cả hai biến lượng này đều là biến ngẫu nhiên, rất nhiều khả năng là chính phủ muốn hỏi những nhà kinh tế câu hỏi sau đây: *"Liệu khả năng lạm phát thấp hơn 8% và mức độ thất nghiệp nhỏ hơn 6% vào năm sau là bao nhiêu?"*. Điều đó có nghĩa là, ta cần phải xác định xác suất đồng thời:

$$P(\pi < 8, u < 6) = ?$$

Để trả lời được những câu hỏi như vậy, chúng ta cần phải xác định *hàm mật độ xác suất đồng thời* [joint probability density function].

1.3.1. Hàm mật độ xác suất đồng thời

Định nghĩa: Giả sử X và Y là 2 biến ngẫu nhiên. Hàm mật độ xác suất đồng thời của *x* và *y* là:

$$f(x,y) = P(X = x, Y = y)$$

Hàm số đó cần thỏa mãn điều kiện:

$f(x,y) \ge 0$, và

$\sum_x \sum_y f(x,y) = 1$ nếu X, Y rời rạc

$\int_x \int_y f(x,y)dy.dx$ nếu X, Y liên tục

Khi đó,

$P(a \le x \le b, c \le y \le d) = \sum_{a \le x \le b} \sum_{c \le y \le d} f(x,y)$, nếu X, Y là biến ngẫu nhiên rời rạc, và

$P(a \le x \le b, c \le y \le d) = \int_a^b \left[\int_c^d f(x,y)dy \right] dx$, nếu X,Y là biến ngẫu nhiên liên tục.

1.3.2. Hàm phân bố xác suất đồng thời F(x,y)

Tương tự như trường hợp biến ngẫu nhiên một biến, ta đưa ra định nghĩa sau về hàm phân bố xác suất đồng thời:

Định nghĩa: Gọi F(x,y) là hàm phân bố xác suất đồng thời của biến ngẫu nhiên x và y. Khi đó:

$$F(x,y) = \Pr ob(X \le x, Y \le y) = \sum_{X \le x} \sum_{Y \le y} f(x,y)$$, nếu X, Y rời rạc

$$F(x,y) = \Pr ob(X \le x, Y \le y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(t,s)ds.dt$$, nếu X, Y liên tục

1.3.3. Phân phối xác suất cận biên

Hãy xét ví dụ sau:

Ví dụ 4: Xét một tổng thể, gồm có 1000 người. [Vì vậy ta nói về mật độ xác suất chứ không phải là tần suất]. Giả sử họ được phân loại theo 2 tiêu chuẩn:

Theo giới tính:

G = 1 nếu người đó là nam
G = 0 nếu người đó là nữ

Và theo trình độ học vấn:

D = 0 học xong trung học
D = 1 học xong đại học
D = 2 học xong cao học

Giả sử kết quả thống kê trên tổng thể 1000 người đó là như sau:

	Nam	Nữ	Học vị (tổng số)
Trung học	200	270	470
Đại học	300	100	400
Cao học	60	70	130
Giới tính(tổng số)	560	440	1000

Dựa trên bảng thống kê này, chúng ta có thể thấy xác suất 1 cá nhân là nữ, học xong đại học: f(0,1)= 100/1000 = 0.1. Một cách khái quát, chúng ta có thể viết hàm mật độ xác suất đồng thời $f(G,D)$ như sau:

		G		Tổng
		1	2	
D	0	0.2	0.27	0.47
	1	0.3	0.1	0.40
	2	0.06	0.07	0.13
Tổng		0.56	0.44	1

Bảng phân bố xác suất trên cho thấy, xác suất một cá nhân là nam trong tổng thể những người có học là: Prob(G=1) = 0.56. Tương tự, xác suất một cá nhân là nữ: Prob(G=0) = 440/1000 = 0.44.

Như vậy, ta có thể lập một biến ngẫu nhiên, thể hiện phân bố mật độ xác suất theo giới tính của tổng thể:

G	f(g)
1	0.56
0	0.44

Hàm f(G) được gọi là hàm mật độ xác suất cận biên. Hàm mật độ này được tính bằng cách cộng dồn theo cột qua tất cả mọi trình độ học vấn:

$f(g)=\sum_d f(g,d), \quad g=\overline{0,1,2}$. Tức là:

$$\begin{cases} f_G(1)=\sum_d f(1,d)=0.56 \\ f_G(0)=\sum_d f(0,d)=0.44 \end{cases}$$

Tương tự như vậy, ta cũng có thể tính được hàm mật độ xác suất cận biên theo học vấn:

$$f_D(d) = \sum_g f(g,d) \qquad d = \overline{0,1,2}$$

Hay cũng vậy,

$$\begin{cases} f_D(0) = \sum_g f(g,0) = 0.47 \\ f_D(1) = \sum_g f(g,1) = 0.4 \\ f_D(2) = \sum_g f(g,2) = 0.13 \end{cases}$$

Một cách tổng quát, gọi f(x,y) là hàm mật độ xác suất đồng thời của X và Y. Khi đó, hàm mật độ xác suất cận biên của X được xác định như sau:

$$f_X(x) = \sum_y f(x,y) \qquad \text{nếu X rời rạc}$$
$$f_X(x) = \int_y f(x,y)dy \qquad \text{nếu X liên tục}$$

Tương tự, ta xác định $f_Y(y)$

1.3.4. Các biến ngẫu nhiên độc lập

Định nghĩa: Hai biến ngẫu nhiên là độc lập khi và chỉ khi:

$$f(x,y) = f_X(x) \cdot f_Y(y)$$

$$\leftrightarrow F(x,y) = F_X(x) \cdot F_Y(y)$$

$$\leftrightarrow \Pr ob(X \le x, Y \le y) = \Pr ob(X \le x) \cdot \Pr ob(Y \le y)$$

1.4. Kỳ vọng – Phương sai

1.4.1. Khái niệm về Kỳ vọng của biến ngẫu nhiên:

Gọi X là một biến ngẫu nhiên rời rạc, nhận một trong các giá trị có thể có x_1, x_2, x_3,… x_K với xác suất tương ứng f_1, f_2, f_3,… f_K. Giá trị kỳ vọng của X được định nghĩa như sau:

$E(X) = x_1 f_1 + x_2 f_2 + x_3 f_3 + ... + x_K f_K$, hay cũng vậy:

$$E(X) = \sum_{k=1}^{K} x_k f_k$$

Tương tự, đối với biến ngẫu nhiên liên tục, giá trị kỳ vọng được định nghĩa như sau:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

Các tính chất của kỳ vọng:

1. $E(a) = a$, với a là hằng số
2. $E(a + bX) = a + bE(X)$
3. $E(XY) = E(X)E(Y)$

Định lý 1.1: Giả sử X là một biến ngẫu nhiên với hàm mật độ xác suất *f(x)* và *g(X)* là một hàm liên tục của X. Khi đó:

$$E[g(X)] = \sum_k g(x_k) f_k \qquad \text{nếu X rời rạc}$$

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

1.4.2. Phương sai

Gọi X là một biến ngẫu nhiên với kỳ vọng EX. Để đo lường sự tán xạ của X so với giá trị trung bình (hay kỳ vọng) của nó, ta sử dụng phương sai, ký hiệu Var(X), được định nghĩa như sau:

$$Var(X) = \sigma_x^2 = E\big(X - E(X)\big)^2$$

Với độ lệch chuẩn:

$$\sigma_x = \sqrt{\sigma_x^2}$$

Sử dụng Định lý 1.1, phương sai của X được tính như sau:

$$Var(X) = \sum_k (x_k - EX)^2 f_k \qquad \text{nếu X rời rạc}$$

$$Var(X) = \int_{-\infty}^{+\infty} \big(X - E(X)\big)^2 f(x)dx \qquad \text{nếu X liên tục}$$

Các tính chất của phương sai:

1. $VarX = E\big(X - E(X)\big)^2 = E(X^2) - \big(E(X)\big)^2$

2. $Var(a) = 0$, với a là hằng số
3. $Var(a + bX) = b^2 \cdot Var(X)$
4. $Var(X + Y) = Var(X) + Var(Y)$
   $Var(X - Y) = Var(X) + Var(Y)$
5. $Var\big(X - E(X)\big) = Var(X)$

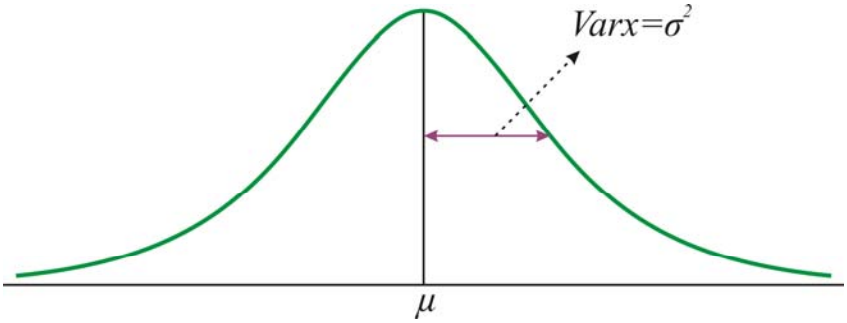
1.5. Hàm phân phối chuẩn

Biến ngẫu nhiên liên tục X nhận các giá trị trong khoảng $(-\infty, +\infty)$ có phân phối chuẩn với các tham số μ và σ^2, ký hiệu là: $X \sim N(\mu, \sigma^2)$, nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

với $\mu = E(X)$ và $\sigma^2 = Var(X)$

Đồ thị 1.2: Hàm phân phối chuẩn



Định lý 1.2: Giả sử X là biến ngẫu nhiên với phân bố chuẩn: $X \sim N(\mu, \sigma^2)$. Gọi $Z = (a + bx)$ là một biến đổi tuyến tính của X. Khi đó, Z cũng là hàm phân bố chuẩn: $Z \sim N(a + b\mu, b^2\sigma^2)$.

Hệ quả: Đặt $Z = \frac{x - \mu}{\sigma}$. Khi đó, $Z \sim N(0,1)$

Địnhlý 1.3: Cho trước chuỗi các biến ngẫu nhiên: $(x_1, x_2, x_3, ..., x_n) \sim N(\mu_n, \sigma_n^2)$
Khi đó, tổ hợp tuyến tính của chúng, cũng có phân bố chuẩn:

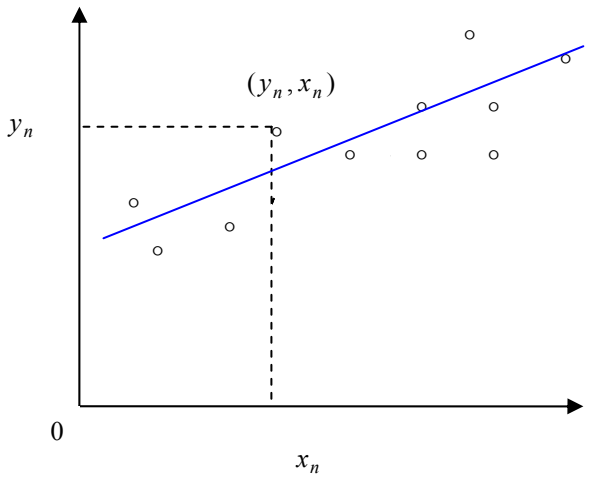
$$c_1x_1 + c_2x_2 + ... + c_nx_n \sim N\left(\sum \mu_n, \sum c_n^2\sigma_n^2\right)$$

1.6. Phân tích Covariance

Trong phần trên, chúng ta đã nói đến việc tồn tại hay không tính độc lập, hay quan hệ phụ thuộc giữa hai biến ngẫu nhiên *X* và *Y*. Nhưng nếu tồn tại quan hệ phụ thuộc lẫn nhau, thì quan hệ đó có thể mạnh hay yếu. Trong phần này, chúng ta sẽ đề cập tới hai thước đo mức độ liên quan giữa hai biến ngẫu nhiên**, tương quan** (hay **covariance**), và **hệ số tương quan** (hay **correlation**, ký hiệu là ρ_{XY}).

Để minh họa, giả sử *X* là trọng lượng của một mẫu nước lấy từ giếng lên, và *Y* là khối lượng của nó. Hiển nhiên là mối quan hệ rất chặt giữa *X* và *Y*. Nếu ta ký hiệu $\{x_n, y_n\}_{n=1}^{N}$ là các cặp đo lường với *N* mẫu thử; và vẽ chúng lên đồ thị, thì các quan sát dữ liệu này sẽ tạo thành một đường thẳng tuyến, thể hiện mối quan hệ vật lý của chúng. Nhưng chúng không rơi đúng vào các điểm dọc theo đường tuyến tính thể hiện quy luật liên hệ giữa khối lượng và trọng lượng nước. Chúng chỉ "bám" xung quanh cái trục tuyến tính đó, vì có sai số đo lường, hoặc các tạp chất trong nước làm các quan sát lệch khỏi quy luật vật lý, mô tả mối quan hệ ổn định giữa *X* và *Y*.

Đồ thị 1.3: Mối quan hệ giữa trọng lượng nước X và khối lượng nước Y



Câu hỏi đặt ra là làm sao chúng ta có thể đo lường mức độ tương quan mạnh hay yếu giữa hai biến X và Y này. Làm sao thể hiện mối quan hệ đó là đồng biến hay nghịch biến?

1.6.1. Covariance

Định nghĩa: *Covariance* giữa hai biến X và Y là hệ số đo:

$$Cov(X,Y) = E\left[(X - EX)(y - EY)\right]$$

Nếu như những giá trị lớn hơn trung bình của X được quan sát với những giá trị lớn hơn trung bình của Y; và những giá trị nhỏ của X cũng đi kèm với những giá trị nhỏ của Y, thì $Cov(X,Y)>0$. Nói khác đi, nếu $(X-EX)>0$ có xu hướng đi kèm với $(Y-EY)>0$; hay ngược lại, khi $(X-EX)<0$, thì $(Y-EY)<0$, thì quan hệ đó có xu hướng tạo ra tích $(X-EX)(Y-EY)>0$. Điều đó có nghĩa là $Cov(X,Y)>0$, thể hiện rằng *X và Y* có mối quan hệ **đồng biến**. Ví dụ như quan hệ giữa khối lượng và trọng lượng các mẫu nước vừa nêu.

Nhiều khi, mối tương quan là **nghịch biến**, chứ không thuận. Chẳng hạn như chúng ta quan sát mối quan hệ giữa điều kiện bảo trợ quá dễ dàng cho một cá nhân, hay doanh nghiệp (ký hiệu là X); và nỗ lực tự vươn lên, tính khởi nghiệp của cá nhân, hay doanh nghiệp đó (ký hiệu là Y). Khi đó, mối quan hệ này thường là nghịch biến. Hỗ trợ nhiều làm chết tính tự chủ, tự vươn lên, tự chịu trách nhiệm của cá nhân. Nói khác đi, giá trị X rất lớn [được nâng đỡ, bảo trợ nhiều] thường đi với giá trị Y rất nhỏ [thiếu nỗ lực bản thân, hay ỉ lại]. Và giá trị X rất nhỏ [không được nâng đỡ] thường đi với giá trị Y rất lớn [tính tự lập, tự chủ cao]. Do vậy, $(X-EX)>0$ thường đi kèm với $(Y-EY)<0$, và $(X-EX)<0$ thường xẩy ra với $(Y-EY)>0$. Kết cục lại, chúng thường tạo ra tích $(X-EX)(Y-EY)<0$. Hay cũng vậy, $Cov(X,Y)<0$, thể hiện mối quan hệ nghịch biến giữa X và Y.

Chúng ta cũng nhận xét luôn rằng, mối quan hệ giữa việc được hỗ trợ, bảo trợ, với tính tự chủ, tự chịu trách nhiệm, ký hiệu là X và Y là nghịch biến. Nhưng về mức độ, nó có thể không mạnh như quan hệ vật lý giữa khối lượng và trọng lượng nước. Nếu chúng ta vẽ đồ thị các quan sát, mối quan hệ giữa việc được hỗ trợ với tính tự vươn lên sẽ **dốc** xuống, thể hiện mối quan hệ nghịch biến. Nhưng không nhất thiết nằm xung quanh một đường thẳng, trải dọc theo một đường cong phi tuyến, thể hiện mối quan hệ đó là yếu hơn so với quan hệ vật lý ở ví dụ đầu. Để đo lường sự khác biệt đó ta dùng hệ số tương quan.

1.6.2. Hệ số tương quan:

Định nghĩa: Hệ số tương quan giữa X và Y là hệ số đo $\rho(X,Y)$:

$$\rho(X,Y)=\frac{Cov(X,Y)}{\sqrt{VarX\cdot VarY}} \qquad (-1\le\rho(X,Y)\le 1)$$

Ta có thể nói rằng, covariance cho phép xác định có mối quan hệ hay không giữa X và Y, và đó là quan hệ nghịch biến hay đồng biến. Hệ số tương quan lại cho phép đo lường mối quan hệ đó là mạnh tới mức nào. Nếu X và Y có quan hệ tuyến tính: $X=\alpha\pm\beta Y$, thì quan hệ đó là mạnh nhất. Và $|\rho(X,Y)|=1$. Nếu đó là quan hệ phi tuyến, thì $|\rho(X,Y)|<1$. Khi X và Y không có quan hệ tương quan: $Cov(X,Y)=0$, khi đó, hệ số tương quan $\rho(X,Y)=0$.

1.6.3. Hai đẳng thức với tương quan mẫu

Hai đẳng thức sau là hai đẳng thức thường sử dụng trong các chương tiếp theo.

1/ $\sum_n (x_n - \bar{x}) \cdot c = 0$, với c: const

2/ $\sum_n (x_n - \bar{x}) \cdot y_n = \sum_n [(x_n - \bar{x}) \cdot (y_n - \bar{y})]$

Chứng minh:

1/ $\sum_n (x_n - \bar{x}) \cdot c = c \cdot \sum_n (x_n - \bar{x}) = c \cdot (\sum_n x_n - \sum_n \bar{x}) = c \cdot (n \cdot \bar{x} - n \cdot \bar{x}) = 0$

2/ Vì $\bar{y}$ là hằng số nên theo chứng minh trên $\sum_n (x_n - \bar{x}) \cdot y_n = 0$, vì vậy:

$$\sum_n (x_n - \bar{x}) \cdot y_n = \sum_n (x_n - \bar{x}) \cdot y_n - \sum_n (x_n - \bar{x}) \cdot \bar{y}$$
$$= \sum_n [(x_n - \bar{x}) \cdot y_n - (x_n - \bar{x}) \cdot \bar{y}]$$
$$= \sum_n [(x_n - \bar{x}) \cdot (y_n - \bar{y})]$$

Chú ý rằng, dòng cuối cùng được gọi là tương quan mẫu giữa X và Y.

CHƯƠNG 2: HỒI QUI ĐƠN BIẾN

Ở bài trước, ta nêu lên ví dụ về mối quan hệ giữa khối lượng và trọng lượng của các mẫu nước. Dựa trên việc lấy các mẫu thử $\{x_n, y_n\}_{n=1}^{N}$, chúng ta có thể **ước lượng**, hay tìm lại mối quan hệ tuyến tính $Y = \alpha + \beta X$, mà nó thể hiện quy luật vật lý, hay tính xu thế, ổn định giữa hai đại lượng ngẫu nhiên là trọng lượng và khối lượng nước.

Trong chương này, chúng ta sẽ giới thiệu việc ước lượng các quy luật tự nhiên, kinh tế, hay xã hội kiểu như vậy thông qua phương pháp **hồi quy đơn** (**simple regression**). Chúng ta sẽ sử dụng học thuyết Keynes về tiêu dùng như là ví dụ điển hình cho việc giới thiệu phương pháp xây dựng và ước lượng mô hình hồi qui đơn biến.

2.1 Học thuyết Keynes về tiêu dùng

Chúng ta hãy trích định luật sau, nêu ra bởi Keynes (1936) trong *Lý thuyết tổng quát* (*general Theory*) của ông:

Chúng ta sẽ xác định quy luật mà ta gọi là khuynh hướng tiêu dùng theo thu nhập như là một mối quan hệ phụ thuộc f giữa X, được gọi là mức thu nhập khả dụng, và Y là chi tiêu cho tiêu dùng từ thu nhập đó, và vì vậy: $Y = f(X)$.

- Số tiền mà từng hộ gia đình chi tiêu cho tiêu dùng phụ thuộc (i) một phần vào thu nhập của hộ đó, (ii) vào những yếu tố khách quan khác của hoàn cảnh sống, và (iii) một phần vào đòi hỏi có tính thiết yếu, thói quen và những yếu tố tâm lý của các cá nhân trong hộ gia đình đó….

- Luật tâm sinh lý cơ bản mà chúng ta dựa vào một cách rất tin cậy, được kiểm chứng bởi tri thức của chúng ta về loài người, và bởi kinh nghiệm, rằng con người có xu hướng tăng tiêu dùng khi thu nhập của họ tăng, nhưng tăng không nhanh bằng thu nhập. Tức là $\frac{dY}{dX}$ là dương và nhỏ hơn 1.

- Về trung bình, nếu thu nhập tăng lên thì khoảng cách giữa tiêu dùng và thu nhập ngày càng mở rộng, nghĩa là có một tỉ lệ lớn hơn trong thu nhập được đưa vào tiết kiệm khi thu nhập tăng lên.

Lý thuyết của Keynes đã đặt một mối quan hệ ổn định giữa tiêu dùng và thu nhập $Y = f(X)$. Chúng ta muốn xác định cụ thể mối quan hệ này là như thế nào, tìm cách đo lường quan hệ đó, và kiểm định lại tính đích thức của học thuyết Keynes.

2.2 Cơ sở vi mô cho học thuyết Keynes về tiêu dùng

Gọi X là mức thu nhập dùng để chi cho tiêu dùng và tiết kiệm (nhằm tăng tiêu dùng cho tương lai). Gọi Y là mức tiêu dùng hiện tại; và S là tiêu dùng trong tương lai.

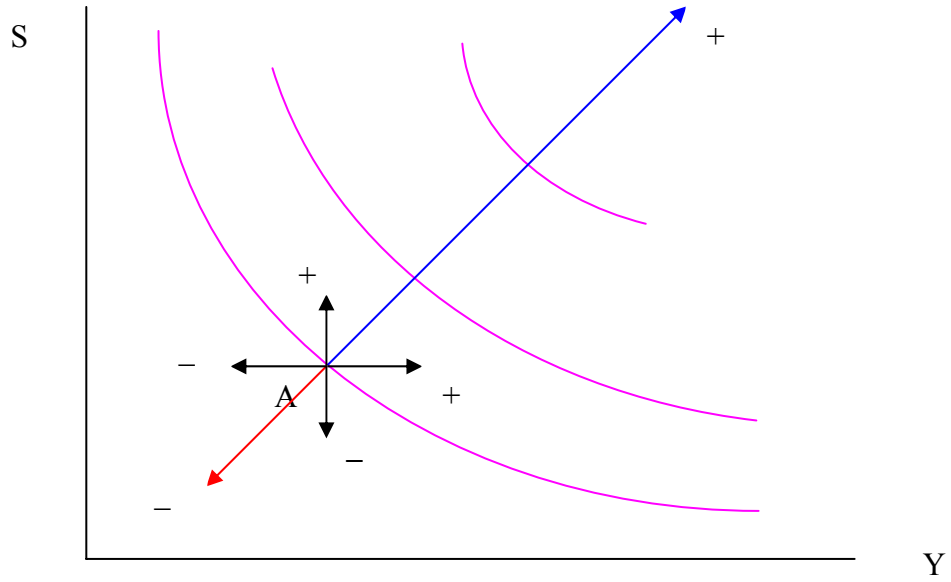
Khi đó, ta có **ràng buộc ngân sách** (budget constraint):

$$Y + \frac{1}{1+r} S = X \qquad (2.1)$$

Thành phần thứ hai trong vế trái $\frac{1}{1+r} S$ là khoản tiết kiệm. Nó thể hiện giá trị hiện tại (present value) của thu nhập cho tiêu dùng trong tương lai *S,* được chiết khấu bởi $\frac{1}{1+r}$. Trong đó, r là lãi suất tiền gửi tiết kiệm.

Về thực chất, 1 đồng tiền ngày hôm nay có thể sinh ra $(1+r)$ đồng thu nhập cho tiêu dùng ngày mai, nếu được gửi vào tiết kiệm. Vì vậy, 1 đồng tiền tiêu trong tương lai chỉ có giá bằng $\frac{1}{1+r}$ đồng tiền ngày hôm nay. Đó chính là khái niệm về **hệ số chiết khấu** (*discount rate*). Nó thể hiện rằng, nếu tiêu dùng bị trì hoãn đi tới một thời điểm trong tương lai, thì nó không thể có giá trị bằng việc được tiêu dùng ngay lập tức vào ngày hôm nay.

Tiếp theo, chúng ta hãy đo lường mức độ thỏa dụng của cá nhân với các lựa chọn khác nhau về tiêu dùng cho hiện tại và cho tương lai (Y, S).



Đồ thị 2.1: Đường bàng quan (indifference curve)

Trong đồ thị 2.1, điểm A thể hiện mức thỏa dụng hiện tại của cá nhân ứng với mức tiêu dùng tại điểm đó. Giả sử có một sự gia tăng về tiêu dùng hiện tại, trong khi tiêu dùng trong

tương lai vẫn giữ nguyên. Khi đó ta dịch chuyển từ điểm A sang bên phải và song song với trục hoành ($\rightarrow^+$). Dấu cộngthể hiện rằng độ thỏa dụng của cá nhân được nâng lên.

Ngược lại, giả sử ta giữ nguyên mức tiêu dùng hiện tại, nhưng tiêu dùng tương lai được tăng lên ($\uparrow^+$). Khi đó, sự cảm nhận về an toàn của cá nhân về cuộc sống tương lai cũng tăng, tức là độ thỏa dụng của cá nhân đó tăng.

Vì vậy, ¼ không gian, được xác định bởi sự gia tăng của tiêu dùng hiện tại ($\rightarrow^+$), hoặc tiêu dùng trong tương lai ($\uparrow^+$), hoặc sự gia tăng đồng thời của cả hai yếu tố đó, thể hiện độ thỏa dụng ngày càng tăng lên (+). Cá nhân cảm thấy giàu lên, sung sướng và an toàn hơn về vật chất.

Phân tích tương tự cho trường hợp ngược lại, khi độ thỏa dụng ngày càng giảm (-).

Trong ngắn hạn, mức thu nhập là không đổi. Do đó, sự gia tăng mức tiêu dùng hiện tại thường phải bị đánh đổi (hay trả giá) bằng việc giảm tiêu dùng trong tương lai. Tuy nhiên, cá nhân chỉ làm sự đánh đổi như vậy một khi độ thỏa dụng mới ít ra là không kém đi so với trạng thái đã có. Trong kinh tế học vi mô, người ta thể hiện các lựa chọn như vậy bằng đường bàng quan **(***indifference curve***)**. Nó có chiều dốc xuống mô tả sự đánh đổi. Nghĩa là, nếu muốn tăng mức tiêu dùng trong hiện tại thì phải giảm mức tiêu dùng trong tương lai, sao cho lợi ích hay độ thỏa dụng vẫn giữ nguyên.

Bây giờ, hãy đưa đường ràng buộc ngân sách vào đồ thị 2.1. Điểm tiếp xúc giữa đường ràng buộc ngân sách với đường bàng quan thể hiện sự lựa chọn tốt nhất của cá nhân về tiêu dùng ứng với mỗi mức thu nhập [xem đồ thị 2.2].

Ví dụ 2.1: Giả sử thu nhập (X) và tiêu dùng (Y) của 3 cá nhân có giá trị cụ thể như sau:

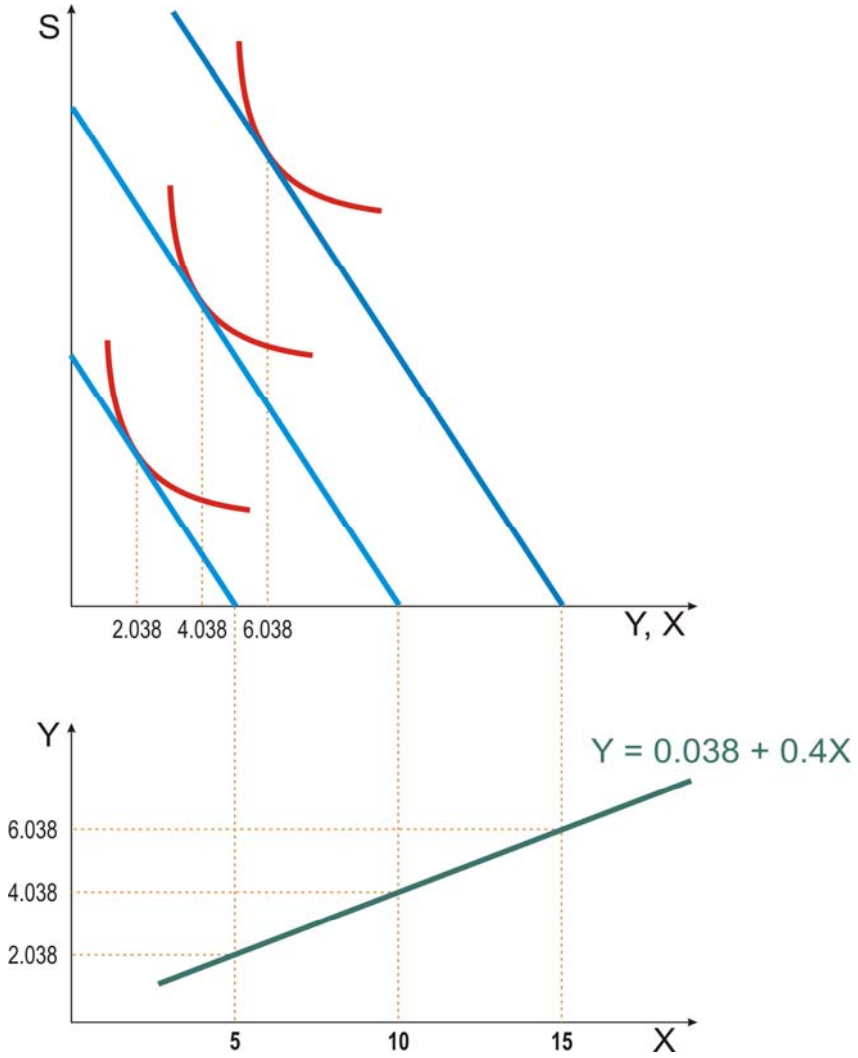
***X* [thu nhập]**	***Y* [tiêu dùng]**
5	*2.038*
10	*4.038*
15	*6.038*

Bảng 2.1: Quan hệ giữa thu nhập và tiêu dùng

Sử dụng phương pháp phân tích nêu trên, chúng ta có thể biểu diễn sự lựa chọn của mỗi cá nhân như sau:

Trong đồ thị 2.2, hình vẽ thứ nhất, ta thể hiện sự lựa chọn của cá nhân về tiêu dùng ứng với mỗi mức thu nhập khả dụng. Khi họ có 5 triệu đồng thu nhập, họ giành cho tiêu dùng hiện tại Y là 2.038 triệu. Phần còn lại được đưa vào tiêu dùng trong tương lai S. Tương tự cho các mức tiêu dùng 4.08 và 6.038 ứng với các mức thu nhập khác là 10 và 15 triệu.

Tiếp theo, trong hình vẽ thứ hai, ta chỉ ra mối quan hệ giữa tiêu dùng hiện tại Y với từng mức thu nhập khả dụng X. Đó chính là mối quan hệ cơ bản, mô tả bởi học thuyết Keynes về tiêu dùng.



Đồ thị 2.2: Sự lựa chọn tiêu dùng theo thu nhập của cá nhân.

Như chỉ ra trên hình vẽ thứ hai, quan hệ giữa tiêu dùng và thu nhập: $Y = f(X)$, là mối quan hệ tuyến tính. Trong ví dụ vừa nêu, quan hệ đó có dạng cụ thể là:

$$Y = 0.038 + 0.40\ X$$

Ý nghĩa của phương trình này như sau:

- Nếu X = 0 thì Y = 0.038, điều này có nghĩa rằng người không có thu nhập vẫn tiêu dùng ở mức tối thiểu là 0.038 triệu đồng một tháng.
- Hệ số 0.40 (hay khuynh hướng tiêu dùng theo thu nhập) cho biết, nếu thu nhập tăng lên 1 triệu thì tiêu dùng tăng lên 0.40 triệu. Tức là, mức tăng tiêu dùng không nhanh bằng mức tăng thu nhập.

- Về trung bình, khi thu nhập tăng thì tỉ lệ giữa thu nhập và tiêu dùng (X/Y) ngày càng giảm: $\frac{2.038}{5} > \frac{4.038}{10} > \frac{6.038}{15}$. Điều đó kiểm chứng lại điều mà Keynes nói là, có một tỷ lệ lớn hơn của thu nhập được đưa vào tiết kiệm khi người ta giàu lên.

Kết quả nghiên cứu trên phù hợp với những nhận định của Keynes về tiêu dùng.

Một cách tổng quát, dạng hàm mô tả tốt nhất khuynh hướng tiêu dùng theo thu nhập của Keynes có dạng tuyến tính:

$$Y = \alpha + \beta X \qquad (\alpha > 0, \beta \in (0,1)) \qquad (2.2)$$

Như đã chỉ ra qua ví dụ, dạng hàm này thỏa mãn mọi nhận định của Keynes về tiêu dùng.

Bây giờ, chúng ta hãy sử dụng các dữ liệu điều tra thực tế để nghiên cứu về nhu cầu tiêu dùng theo thu nhập thông qua lăng kính của học thuyết Keynes.

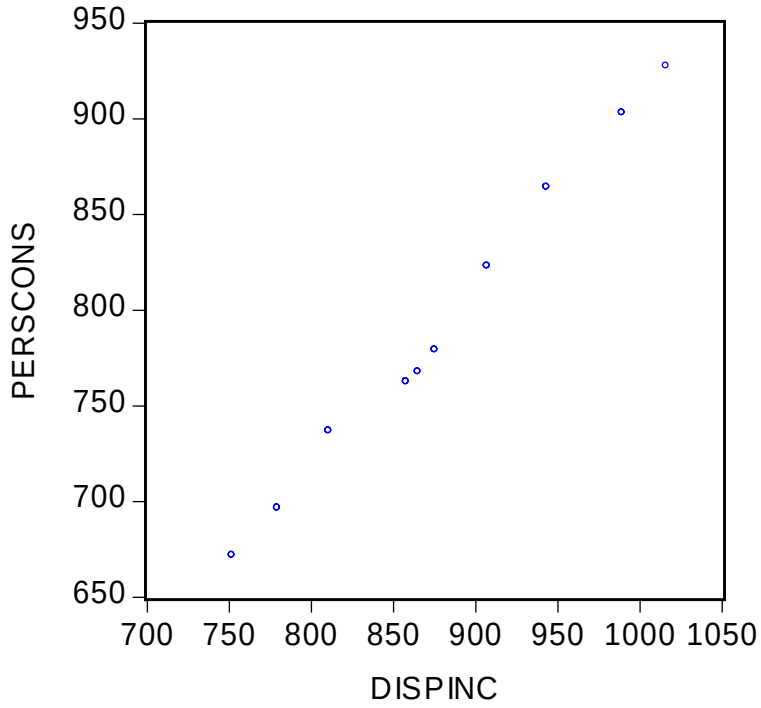
Ví dụ 2.2: Số liệu về tiêu dùng trung bình (PERCONS) và thu nhập khả dụng (DISPINC) theo giá cố định theo năm 1972 của nền kinh tế Mỹ trong 10 năm 1970 – 1979:

ĐVT: tỷ dollars

Năm	DISPINC	PERSCONS
1970	751.6	672.1
1971	779.2	696.8
1972	810.3	737.1
1973	864.7	767.9
1974	857.5	762.8
1975	874.9	779.4
1976	906.8	823.1
1977	942.9	864.3
1978	988.8	903.2
1979	1015.7	927.6

Bảng 2.2: Số liệu gộp về thu nhập và tiêu dùng tại Mỹ (1970-79)
(Nguồn: *Economic Report of the President*)

Đồ thị mô tả mối quan hệ giữa thu nhập và tiêu dùng của Mỹ được chỉ ra dưới đây:



Đồ thị 2.3: Mối quan hệ giữa thu nhập và tiêu dùng của nền kinh tế Mỹ từ 1970 đến 1979.

Mặc dù dữ liệu xem ra thể hiện khá tốt qui luật tuyến tính nêu ở trên nhưng rõ ràng mối quan hệ có tính xác định đó là **không đủ** để mô tả thực tiễn, vì còn rất nhiều yếu tố khác ảnh hưởng đến tiêu dùng (giới tính, tuổi tác, tâm lý,…).

Nói chung, chúng ta không có tham vọng đưa hết tất cả mọi yếu tố ảnh hưởng tới tiêu dùng vào mô hình, mà chỉ những yếu tố quan trọng, thiết yếu nhất.

Vì vậy, để có thể biểu diễn qui luật tiêu dùng trên thế giới thực, ta cần đưa thêm vào mô hình tuyến tính (2.2) một thành phần khác nữa, mang tính ngẫu nhiên, thể hiện sự tác động tổng gộp của các nhân tố nhỏ, không ổn định, tới tiêu dùng. Tức là, những yếu tố làm cho quan sát thật về tiêu dùng và thu nhập bị lệch khỏi xu thế ổn định, tuyến tính (2.2) nêu trên. Tức là, ta muốn biểu diễn mối quan hệ thực giữa các cặp dữ liệu quan sát được về thu nhập và tiêu dùng $\{x_n, y_n\}_{n=1}^{N}$ như sau:

$$y_n = \alpha + \beta x_n + \varepsilon_n, \qquad n = 1,2,3...,N. \tag{2.3}$$

Trong đó, $(X,Y) = (x_n, y_n)$: tiêu dùng và thu nhập thực tế của mẫu quan sát thứ *n*. Xét vế phải của phương trình (2.3), thành phần thứ nhất, $\alpha + \beta x_n$, là quy luật xác định [*deterministic part*], mà ta cần ước lượng; phần thứ hai, ε_n, là nhiễu [*random part*]. (Tức là, ε_n bao gồm sự tác động tổng hợp của mọi yếu tố khác của hoàn cảnh, có tính ngẫu nhiên, làm quan sát bị lệch khỏi khuynh hướng, hay qui luật ổn định). Cả hai phần này – tính xu thế, xác định; *và* yếu tố ngẫu nhiên - được gộp lại trong phương trình (2.3) để mô tả lý thuyết tiêu dùng của Keynes.

Do tác động của yếu tố ngẫu nhiên, trên đồ thị 2.3, chúng ta không quan sát thấy một đường thẳng thể hiện mối quan hệ tuyến tính $Y = \alpha + \beta X$ giữa tiêu dùng và thu nhập, như trên đồ thị 2.2 với số liệu giả định. Với dữ liệu điều tra thực tế, ta chỉ thấy một đám mây dữ liệu, dường như đang "bám" xung quanh một xu thế nào đó mà ta muốn ước lượng.

Ví dụ 2.3: Dữ liệu điều tra 44 nhân khẩu của nhóm gồm 5 sinh viên K04 khoa Kinh tế về thu nhập và tiêu dùng đầu người hộ gia đình tại TP HCM, Bình Dương, Thủ Dầu Một, Bà Rịa - Vũng Tàu, Mỹ Tho, và Nghệ An được ghi lại như sau[1]:

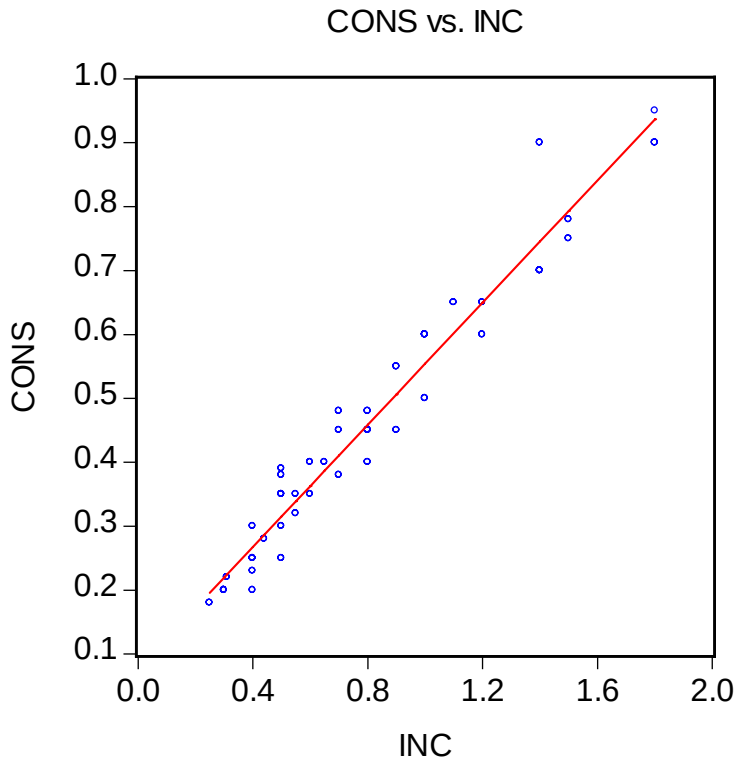
ĐVT: triệu đồng

Obs	INC	CONS	Obs	INC	CONS
1	1 .00	0.60	**23**	0.50	0.35
2	1.10	0.65	**24**	0.70	0.38
3	0.70	0.48	**25**	0.40	0.20
4	1.40	0.90	**26**	0.55	0.35
5	0.50	0.38	**27**	0.50	0.35
6	0.40	0.23	**28**	0.90	0.55
7	0.55	0.32	**29**	0.40	0.30
8	0.80	0.48	**30**	0.31	0.22
9	0.70	0.45	**31**	1 .20	0.65
10	0.25	0.18	**32**	0.60	0.40
11	0.65	0.40	**33**	0.30	0.20
12	0.40	0.25	**34**	0.80	0.40
13	1.80	0.95	**35**	0.44	0.28
14	0.40	0.25	**36**	0.50	0.39
15	0.50	0.30	**37**	1.00	0.60
16	0.30	0.20	**38**	1.80	0.90
17	1.00	0.50	**39**	1.40	0.70
18	0.50	0.25	**40**	1.50	0.75
19	0.80	0.45	**41**	1 .20	0.60
20	1.40	0.70	**42**	0.80	0.45
21	0.80	0.45	**43**	0.90	0.45
22	0.60	0.35	**44**	1.50	0.78

Bảng 2.3: Điều tra về thu nhập và tiêu dùng đầu người hộ gia đình tại một số tỉnh Việt nam

(*Ghi chú*: INC và CONS là thu nhập và tiêu dùng đầu người, đơn vị triệu đồng, tính tại thời điểm tháng 6, 2006.)

[1] Trưởng nhóm nghiên cứu này có mã số sinh viên là K 04 406 0975



Đồ thị 2.4: Thu nhập và tiêu dùng đầu người hộ gia đình tại một số tỉnh ở Việt Nam, năm 2006.

Như chỉ ra trên đồ thị, dữ liệu điều tra về tiêu dùng và thu nhập đầu người của hộ gia đình Việt nam tại một số tỉnh được điều tra cho thấy học thuyết tiêu dùng của Keynes phản ánh khá đúng về quy luật tiêu dùng của hộ gia đình tại các địa phương này.

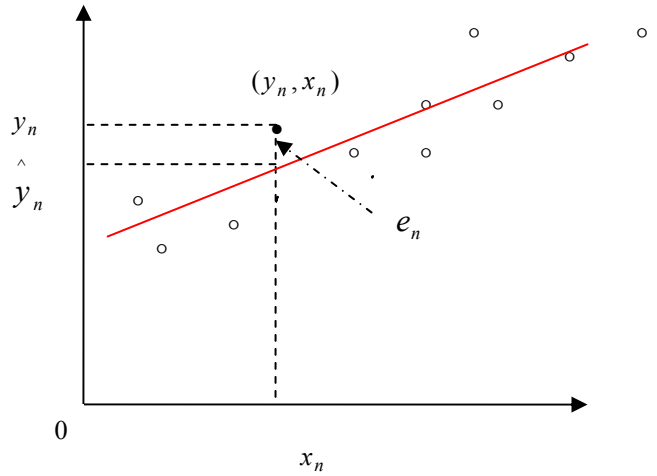
Bước tiếp sau là chúng ta hãy sử dụng những dữ liệu quan sát được này để xác định trở lại các tham số α, β trong mô hình hồi quy tuyến tính (2.2) và (2.3).

2.3. Ước lượng qui luật tiêu dùng:

Ta hãy vẽ các cặp quan sát về thu nhập và tiêu dùng $\{x_n, y_n\}_{n=1}^{N}$ lên đồ thị. Giả sử vạch đỏ trên đồ thị 2.5 dưới đây mô tả đường ước lượng quy luật tiêu dùng theo thu nhập. Nói khác đi, ta muốn ước lượng xu thế tiêu dùng bằng qui luật **tuyến tính**:

$$\hat{y}_n = \hat{\alpha} + \hat{\beta} x_n \qquad (2.4)$$

Trong đó, $\hat{y}_n$ là *ước lượng* về tiêu dùng, khi cho trước quan sát thu nhập x_n. Tương ứng, $\hat{\alpha}, \hat{\beta}$: các tham số ước lượng của các tham số tổng thể, chưa biết α, β.



Đồ thị 2.5: Ước lượng quy luật tiêu dùng qua các quan sát $(x_n, y_n), n = \overline{1, N}$

Mức độ tốt của việc ước lượng có thể được đo lường qua số dư (*residual*):

$$e_n = y_n - \hat{y} \qquad (2.5)$$

Như đã nói, y_n là giá trị quan sát thực tế về tiêu dùng ứng với thu nhập x_n. Và $\hat{y}_n$: giá trị ước lượng về tiêu dùng.

Về mặt toán học, ta có thể viết tổng bình phương của sai số ước lượng (2.5) như sau:

$$\sum_n e_n^2 = \sum_n (y_n - \hat{y}_n)^2 \qquad (2.6)$$

Sử dụng quan hệ (2.4), ta viết lại tổng bình phương sai số [*error sum of squares*], ký hiệu là ESS, ghi trong (2.6) như sau:

$$\sum_n e_n^2 = \sum_n (y_n - \hat{\alpha} - \hat{\beta} x_n)^2 \qquad (2.7)$$

Một cách tự nhiên, chúng ta muốn rằng tổng bình phương sai số phần dư là nhỏ nhất**.** Vì vậy phương pháp có tên gọi là **bình phương cực tiểu** [*Least Squares*]:

$$S(\hat{\alpha},\hat{\beta}) = \sum_n (y_n - \hat{\alpha} - \hat{\beta} x_n)^2 \to \min \qquad (2.8)$$

Lưu ý rằng ở bài toán (2.8), chúng ta muốn chọn các tham số ước lượng $\hat{\alpha},\hat{\beta}$ sao cho tổng bình phương các sai số ước lượng, ESS, là nhỏ nhất.

Sử dụng điều kiện tìm điểm cực trị, (first order condition, FOC), chúng ta thấy rằng:

$$\frac{\partial S(\hat{\alpha},\hat{\beta})}{\partial \hat{\alpha}} = 2\sum_n (y_n - \hat{\alpha} - \hat{\beta}\cdot x_n)(-1) = 0 \qquad (2.10)$$

$$\frac{\partial S(\hat{\alpha},\hat{\beta})}{\partial \hat{\beta}} = 2\sum_n (y_n - \hat{\alpha} - \hat{\beta} x_n)(-x_n) = 0 \qquad (2.11)$$

Từ (2.10) ta có:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\cdot\bar{x} \Rightarrow \bar{y} = \hat{\alpha} + \hat{\beta}\cdot\bar{x} \qquad (2.12)$$

Nói khác đi, điểm $(\bar{x},\bar{y})$ nằm trên đường hồi qui $\hat{y}_n = \hat{\alpha} + \hat{\beta} x_n$.

Tiếp theo, từ phương trình (2.11), ta cũng có:

$$\sum_n y_n x_n = \hat{\alpha}\sum_n x_n + \hat{\beta}\sum_n x_n^2$$

Thay thế $\hat{\alpha} = \bar{y} - \hat{\beta}\cdot\bar{x}$ trong (2.12) vào biểu thức trên, sắp xếp lại các vế, ta tìm ra:

$$\sum_n (y_n - \bar{y}) x_n = \hat{\beta}\sum_n (x_n^2 - n\cdot\bar{x}^2)$$

Hay cũng vậy,

$$\hat{\beta} = \frac{\sum_n (x_n - \bar{x})(y_n - \bar{y})}{\sum_n (x_n - \bar{x})^2} \qquad (2.13)$$

Tóm lại, kết quả ước lượng $\hat{\alpha}, \hat{\beta}$ theo phương pháp bình phương cực tiểu như sau:

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$
$$\hat{\beta} = \frac{\sum_n (x_n - \bar{x}) \cdot (y_n - \bar{y})}{\sum_n (x_n - \bar{x})^2} = \frac{S_{XY}}{S_{XX}} \qquad (2.14)$$

Trong đó, S_{XY} là Covariance mẫu, và S_{XX} : Variance mẫu của X.

2.5 Đo lường mức độ phù hợp của Ước lượng

Công thức (2.14) thể hiện hai điều: Thứ nhất, đường hồi quy đi qua điểm trung bình $(\bar{x}, \bar{y})$. Thứ hai, hệ số góc $\hat{\beta}$ là covariance mẫu của X và Y, cho phép đánh giá những biến động trong thu nhập X có tác động thế nào tới biến động trong tiêu dùng Y. Nếu mô hình phân tích và dự báo là tốt, thì một sự tăng (giảm) mạnh của thu nhập so với trung bình sẽ dẫn tới một sự tăng (giảm) mạnh tương của tiêu dùng so với trung bình.

Câu hỏi đặt ra là: liệu ta có thể sử dụng mô hình ước lượng để dự báo không? Liệu sự giao động của thu nhập so với trung bình $(x - \bar{x})$ có phải là dự đoán tốt cho sự giao động của tiêu dùng so với trung bình $(y - \bar{y})$ hay không?

Hãy lấy một quan sát cụ thể về tiêu dùng và thu nhập (x_n, y_n). Khi đó, sự khác biệt của thu nhập cá nhân thứ *n* so với trung bình $(y_n - \bar{y})$ có thể được viết lại như sau:

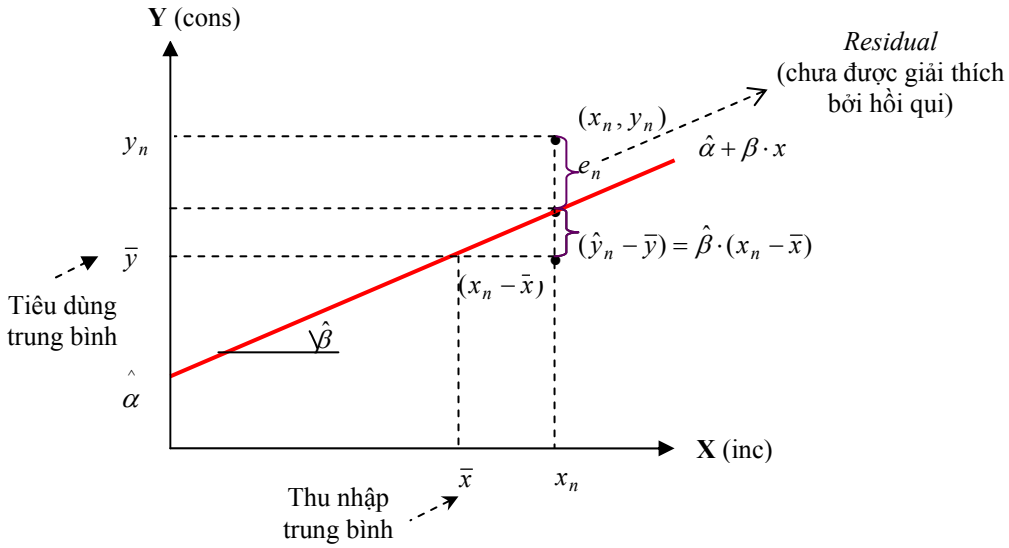
$$(y_n - \bar{y}) = \hat{y}_n - \bar{y} + e_n \qquad (2.15)$$

Hay cũng vậy,

$$(y_n - \bar{y}) = \hat{\beta}(x_n - \bar{x}) + e_n \qquad (2.16)$$

Vế trái là giao động của tiêu dùng so với mức trung bình; thành phần thứ nhất của vế phải là phần mà giao động đó đã được giải thích bởi mô hình hồi quy; và phần cuối cùng là sai

số ước lượng, thể hiện những giao động trong tiêu dùng chưa được giải thích bởi mô hình. Nói khác đi, đó là sai số dự báo từ mô hình.



Đồ thị 2.6: Phân tách các thành phần của $(y_n - \bar{y})$

Sử dụng các điều kiện tìm cực trị FOC (2.10) và (2.11), quan hệ (2.15) có thể viết lại như sau:

$$\sum_n (y_n - \bar{y})^2 = \sum_n (\hat{y}_n - \bar{y})^2 + \sum_n e_n^2 \qquad (2.17)$$

Vế trái là tổng bình phương các giao động trong tiêu dùng, ký hiệu là TSS (*total sum of squares*). Vế phải phân ra thành tổng bình phương phần đã được giải thích bởi mô hình hồi quy RSS (*regression sum of squares*), cộng với tổng sai số ước lượng ESS (*error sum of squares*).

Nói khác đi, ta có:

$$\text{TSS} = \text{RSS} + \text{ESS} \qquad (2.18)$$

Vì vậy,

$$1 = \frac{RSS}{TSS} + \frac{ESS}{TSS}$$

Hay cũng thế,

$$\frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} \qquad (2.19)$$

Vế phải của (2.19) được ký hiệu là $R^2 = 1 - \frac{ESS}{TSS}$. Ta thấy $0 \le R^2 \le 1$.

Ví dụ 2.4: Ước lượng khuynh hướng tiêu dùng cho một số tỉnh thành ở Việt Nam, sử dụng dữ liệu điều tra trong Ví dụ 2.3.

Kết quả ước lượng theo phương pháp bình phương cực tiểu được ghi lại dưới đây [các tham số $\hat{\alpha}, \hat{\beta}$ được tính theo công thức (2.14), và hệ số đo lường mức phù hợp R^2 theo (2.19)]:

Dependent Variable: CONS
Method: Least Squares
Date: 06/24/06 Time: 21:39
Sample: 1 44
Included observations: 44
Weighting series: INC

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.038254	0.004082	9.370437	0.0000
INC/HHSIZE	0.401771	0.014340	28.01749	0.0000

Weighted Statistics

R-squared	0.989838	Mean dependent var	0.137000
Adjusted R-squared	0.989596	S.D. dependent var	0.083538
S.E. of regression	0.008521	Akaike info criterion	-6.648174
Sum squared resid	0.003050	Schwarz criterion	-6.567075
Log likelihood	148.2598	F-statistic	784.9795
Durbin-Watson stat	2.221397	Prob(F-statistic)	0.000000

Bảng 2.4: Kết quả hồi quy mô hình tiêu dùng với dữ liệu điều tra tại Việt nam

Để có một hình dung rõ ràng về độ tốt của mô hình, ta dùng 40 quan sát đầu tiên để ước lượng mô hình. Sau đó dùng 4 quan sát cuối để kiểm tra độ tốt của dự báo(*ex post forecasting*). Kết quả dự báo cho 4 mẫu quan sát cuối cùng trong dữ liệu điều tra là như sau:

Obs	CONS	CONSF
41	0.600000	0.638548
42	0.450000	0.438385
43	0.450000	0.478911
44	0.780000	0.798185

Bảng 2.5: Kết quả dự báo

Ở đây, CONS là dữ liệu thu thập được về tiêu dùng của mẫu quan sát, [tương ứng với ký hiệu $y_n, n = 41,..,44.$]; và CONSF là kết quả dự báo từ mô hình, [tương ứng với ký hiệu $\hat{y}_n, n = 41,..,44.$] Như đã thấy, kết quả dự báo là khá phù hợp với dữ liệu thực có được từ điều tra.

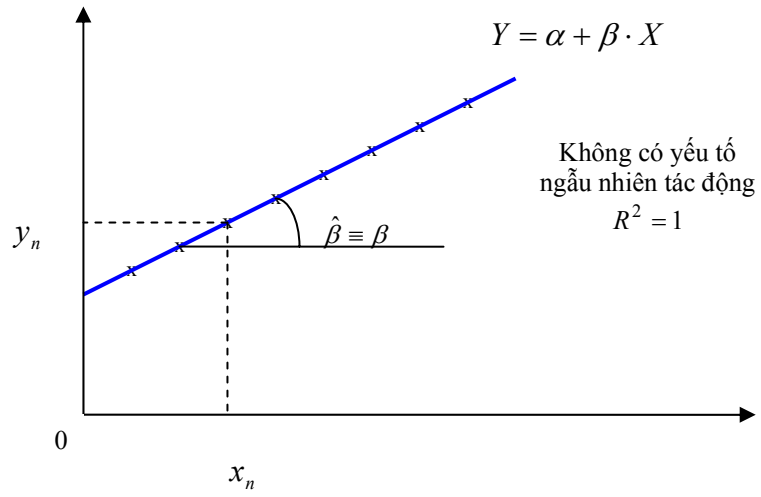
Trong mục tiếp sau, chúng ta sẽ đánh giá độ tốt của ước lượng theo các tiêu chuẩn thống kê.

CHƯƠNG 3: HỒI QUI ĐƠN BIẾN

3.1 Bản chất thống kê của mô hình hồi quy đơn biến

Phương pháp ước lượng *LS*, về thực chất, chỉ là vẽ một đường hồi quy đi xuyên qua "đám bụi" dữ liệu, sao cho tổng bình phương các phần dư [hay sai số] ESS là nhỏ nhất. Nhưng việc đo lường mang tính thuần túy đại số đó chưa có gì bảo đảm chắc chắn rằng nó sẽ cho ra những ước lượng $\hat{\alpha}, \hat{\beta}$ tốt nhất của các tham số tổng thể α, β theo những tiêu chuẩn xác định về mặt thống kê. Để có thể những đánh giá cụ thể hơn về độ tốt của ước lượng, chúng ta cần xem xét sâu hơn **bản chất thống kê** của mô hình hồi quy.

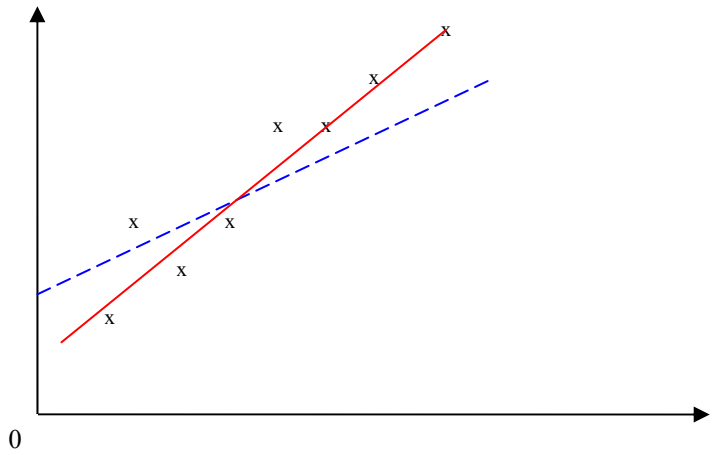
Để dễ hình dung, chúng ta bắt đầu bằng sự giả định phi thực rằng, quan hệ giữa biến X và Y [chẳng hạn như giữa thu nhập và tiêu dùng] chỉ tuân theo quy luật xác định, và hoàn toàn không bị chi phối bởi các yếu tố ngẫu nhiên. Khi đó, các quan sát $\{x_n, y_n\}_{n=1}^{N}$ sẽ nằm gọn trên một đường thẳng mô tả xu thế thực của tổng thể:



Đồ thị 3.1a: quy luật xác định giữa X và Y.

Khi đó, việc ước lượng trở nên tầm thường, vì ta luôn có $\hat{\alpha} = \alpha, \hat{\beta} = \beta$, và $R^2 = 1$.

Bây giờ, chúng ta cho phép các yếu tố ngẫu nhiên tác động lên quan hệ giữa X, Y. Như đã nêu, các nhân tố này khiến cho các quan sát $\{x_n, y_n\}_{n=1}^{N}$ bị lệch một cách ngẫu nhiên khỏi đường xu thế tổng thể. Vì vậy, thay vì nhìn thấy một đường xu thẳng tuyến tính như trên hình 3.1a, ta chỉ nhìn thấy một đám bụi dữ liệu bám xung quanh một xu thế nào đó mà ta muốn ước lượng.



Đồ thị 3.1b: Quan hệ giữa X và Y bị nhiễu bởi các yếu tố ngẫu nhiên

Trên Đồ thị 3.1b, ta thấy các điểm quan sát $\{x_n, y_n\}_{n=1}^{N}$, trước đây nằm trên cùng một đường thẳng trên hình 3.1a, nay bị "thổi bay" lên thành một "đám bụi" dữ liệu, mà việc "chụp ảnh" chúng [tức là đi thu thập dữ liệu], rồi vẽ một đường hồi quy chạy xuyên qua chúng sẽ không nhất thiết là trùng với quy luật tổng thể (mô tả bởi gạch chấm). Điều này gợi ý rằng mỗi ước lượng $\hat{\beta}$ chịu sự quy định bởi tham số tổng thể β, nhưng bị lái đi bởi các biến ngẫu nhiên. [Tương tự, ta có thể nói như vậy về $\hat{\alpha}$]. Vì vậy, $\hat{\beta}$ cũng là một biến ngẫu nhiên. Vấn đề đặt ra là, về trung bình mà nói [tức là sau rất nhiều lần chụp ảnh các đám bụi dữ liệu], liệu ước lượng $\hat{\beta}$ có thể hiện đúng β hay không? Và liệu phương pháp ước lượng bình phương cực tiểu có là hiệu quả nhất hay không?

Về mặt toán học, phương pháp bình phương cực tiểu cho ta ước lượng sau:

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} = \frac{\sum (x_n - \bar{x})(y_n - \bar{y})}{S_{XX}} \quad (3.1)$$

Hay cũng vậy,

$$\hat{\beta} = \frac{\sum (x_n - \bar{x}) y_n}{S_{XX}} \quad (3.2)$$

[điều này là do $\sum_n (x_n - \bar{x})\,\bar{y} = 0$, như đã chỉ ra ở chương 1, phần ôn tập].

Trong (3.2), ta đặt $c_n = \frac{(x_n - \bar{x})}{S_{XX}}$, và nhận xét rằng, tham số đó chỉ phụ thuộc vào các quan sát $\{x_n\}_{n=1}^{N}$. Do vậy, nó không chịu ảnh hưởng bởi các yếu tố ngẫu nhiên. Khi đó, công thức (3.2) có thể viết lại như sau:

$$\begin{aligned}\hat{\beta} &= \sum_n c_n y_n \\ &= \sum_n c_n [\alpha + \beta x_n + \varepsilon_n] \\ &= \alpha \sum c_n + \beta \sum c_n x_n + \sum c_n \varepsilon_n\end{aligned}$$

Chúng ta có thể dễ dàng chỉ ra rằng, $\sum_n c_n = 0$ và $\sum_n c_n x_n = 1$. Và do vậy:

$$\hat{\beta} = \beta + \sum c_n \varepsilon_n \quad (3.3)$$

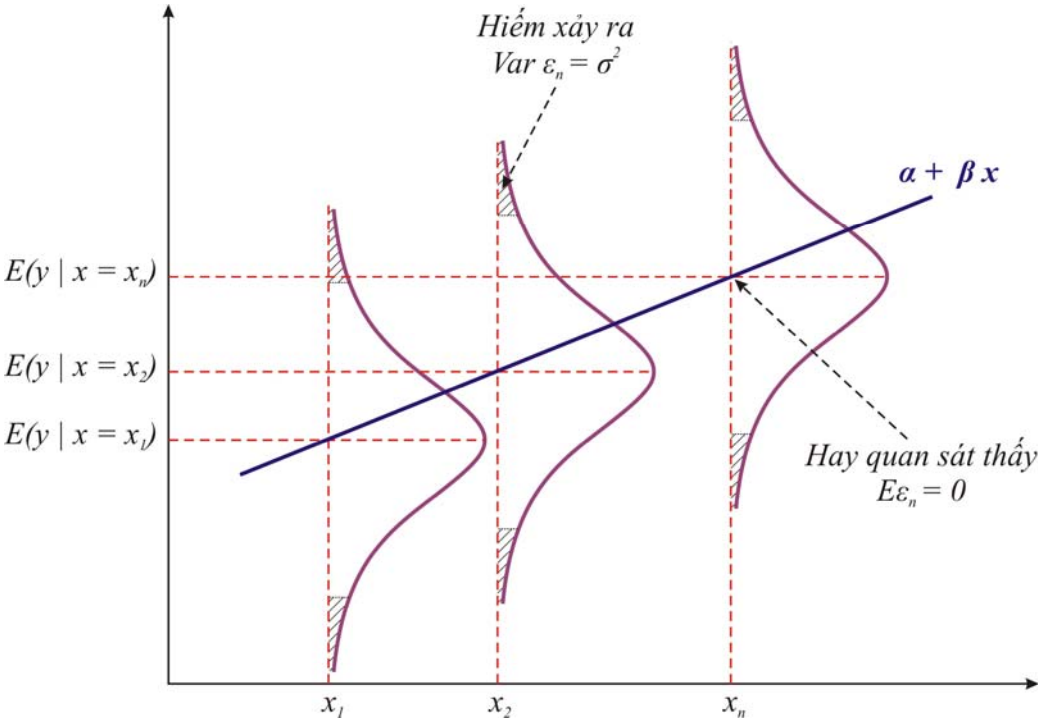
Phương trình (3.3) khẳng định nhận định trước đây về $\hat{\beta}$ là đúng: Ước lượng $\hat{\beta}$ bị ảnh hưởng bởi các yếu tố ngẫu nhiên ε_n, làm giá trị của nó không trùng khít với β tổng thể. Và vì vậy, $\hat{\beta}$ cũng là một biến ngẫu nhiên.

Chúng ta gọi $\hat{\beta}$ là **ước lượng không chệch**, nếu $E\hat{\beta} = \beta$. Và gọi nó là **ước lượng hiệu quả nhất**, nếu sai số ước lượng $Var\hat{\beta} = E(\hat{\beta} - \beta)^2$ là nhỏ nhất trong lớp tất cả các ước lượng tuyến tính, không chệch.

Để trả lời xem $\hat{\beta}$ có phải là ước lượng không chệch và hiệu quả hay không, ta phải xét đến bản chất thống kê của các quá trình ngẫu nhiên $\{\varepsilon_n\}_{n=1}^N$ [mà ta đã ví chúng như những "cơn gió", ngẫu nhiên "thổi bay" các quan sát khỏi đường xu thế xác định của tổng thể].

3.2 Các yếu tố ngẫu nhiên

Chúng ta hãy nêu lên giả định về các quá trình ngẫu nhiên. Hãy nhìn vào đồ thị sau:



Đồ thị 3.2: Quy luật phân phối xác suất của các nhiễu $\{\varepsilon_n\}_{n=1}^N$

Như đã nhận xét từ các Đồ thị 3.1a và 3.1b, khi không có các tác động ngẫu nhiên, hay $\varepsilon_n = 0$, các quan sát $\{x_n, y_n\}_{n=1}^N$ nằm ngay trên đường xu thế của tổng thể. Dưới tác động của yếu tố ngẫu nhiên, các quan sát $\{x_n, y_n\}_{n=1}^N$ nằm rải ra, nhưng "bám" xung quanh đường xu thế. Rất hiếm khi có quan sát bị "thổi" mạnh tới nỗi "bay" quá xa so với đường xu thế. Điều đó dẫn đến hai giả thiết sau:

A1 $E\varepsilon_n = 0,$ với mọi *n*. [Bụi giữ liệu không thể bay quá xa, mà bám xung quanh đường tổng thể]

A2 $Var\varepsilon_n = \sigma^2$, với mọi *n*. [Độ tán xạ của đám bụi dữ liệu được thể hiện bởi độ lớn của σ^2].

Chúng ta cũng coi rằng quy luật tác động của "cơn gió", tức là phân bố xác suất của yếu tố ngẫu nhiên ε_n là như nhau (*identical)*, và theo phân bố chuẩn. Hơn nữa, các yếu tố ngẫu nhiên đó là độc lập (*independent*). Vì vậy, kết hợp với các giả thiết A1 và A2, ta có:

A3 $\varepsilon_n \overset{iid}{\sim} N(0,\sigma^2)$ với mọi *n*.

Cuối cùng, ta coi ta coi x_n là xác định trước. Từ giả thiết A1 và dạng mô hình $y_n = \alpha + \beta x_n + \varepsilon_n$, điều đó bao hàm rằng:

A4 $E(y_n \mid x_n) = \alpha + \beta x_n,$ với mọi *n*.

Hai giả thiết cuối là quan trọng nhất. A3 tóm tắt mọi đặc trưng thống kê của nhiễu ngẫu nhiên, và A4 mô tả xu thế của tổng thể, mà ta ước lượng nó theo phương pháp bình phương cực tiểu.

3.3 Những đặc trưng thống kê của ước lượng bình phương cực tiểu

Bây giờ ta có thể nói đến tính tốt của các ước lượng theo các tiêu chuẩn thống kê .

Từ phương trình (3.3), ta đã có: $\hat{\beta} = \beta + \sum c_n \varepsilon_n$. Bây giờ, hãy áp dụng toán tử kỳ vọng vào hai vế của (3.3):

$$\begin{aligned} E\hat{\beta} &= E(\beta + \sum c_n \varepsilon_n) \\ &= \beta + \sum c_n E\varepsilon_n \\ &= \beta \end{aligned}$$

[ở đây, ta sử dụng giả thiết A1: $E\varepsilon_n = 0$].Ta đi đến kết luận rằng, ước lượng $\hat{\beta}$ là không chệch:

$$E\hat{\beta} = \beta \tag{3.4}$$

Tiếp theo, sử dụng công thức: $Var(x) = Var(x - Ex)$ [xem chương 1, phần ôn tập], và lưu ý (3.3), (3.4), ta có:

$$\begin{aligned} Var\hat{\beta} &= Var(\hat{\beta} - \beta) \\ &= Var(\sum c_n \varepsilon_n) \end{aligned}$$

Sử dụng giả thiết A3 về tính độc lập của các yếu tố ngẫu nhiên, cuối cùng ta nhận được:

$$\begin{aligned} Var\hat{\beta} &= \sum c_n^2 Var\varepsilon_n \\ &= \sigma^2 \sum c_n^2 \text{ , hay} \\ Var\hat{\beta} &= \frac{\sigma^2}{S_{XX}} \end{aligned} \tag{3.5}$$

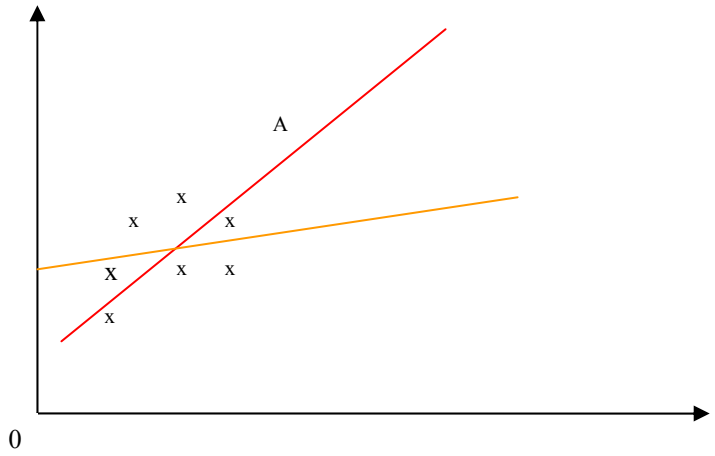
(ở đây, ta sử dụng cái điều là $\sum c_n^2 = \sum \left[\frac{(x_n - \bar{x})}{S_{XX}} \right]^2 = \frac{S_{XX}}{S_{XX}^2} = \frac{1}{S_{XX}}$)

Định Lý Gauss - Markov: Phương pháp bình phương cực tiểu có sai số ước lượng, đo lường bởi $Var\hat{\beta}$, là nhỏ nhất trong lớp tất cả các ước lượng tuyến tính và không chệch.

Định lý Gauss-Markov là hết sức quan trọng. Nó nêu lên rằng, chúng ta có được những tính chất rất tốt cho ước lượng theo phương pháp bình phương cực tiểu, mà chỉ đòi hỏi có trung bình bằng zero, tính độc lập, và phương sai giống nhau của các yếu tố ngẫu nhiên – tức là giả thiết A3.

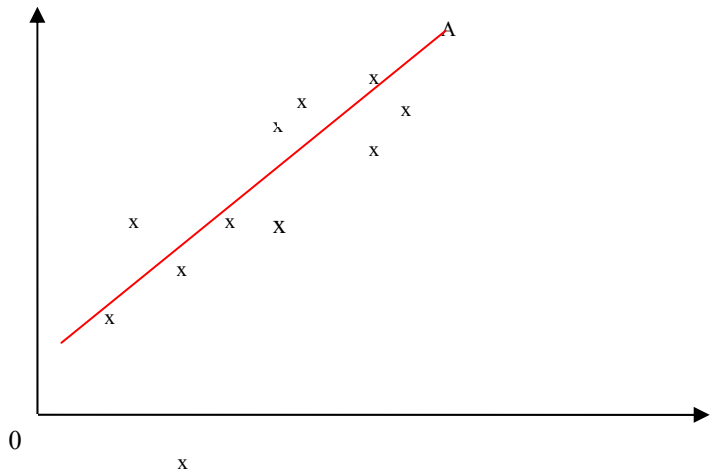
Chúng ta cũng nên nói thêm là, phương trình (3.5) có một ý nghĩa thực tiễn đáng lưu ý. Nó nói rằng sai số của ước lượng $Var\hat{\beta}$ sẽ nhỏ đi, hay hiệu quả ước lượng sẽ tăng lên, nếu độ đa dạng của thông tin quan sát, đo bởi S_{XX}, tăng lên. Điều đó bao hàm rằng, khi làm nghiên cứu, ta không cứ nhất thiết phải tăng rất lớn số quan sát (*sample size*) *N*. Nếu giả thiết về tính tuyến tính của đường hồi quy là đúng, thì việc tăng độ đa dạng của thông tin quan sát,

hay biên độ giao động của biến giải thích, $S_{XX} = \sum_n (x_n - \bar{x})^2$, sẽ làm cho ước lượng có độ chính xác cao hơn. Hãy xét các ví dụ sau:



Đồ thị 3.3a: Ước lượng có độ chính xác thấp, do S_{XX} nhỏ.

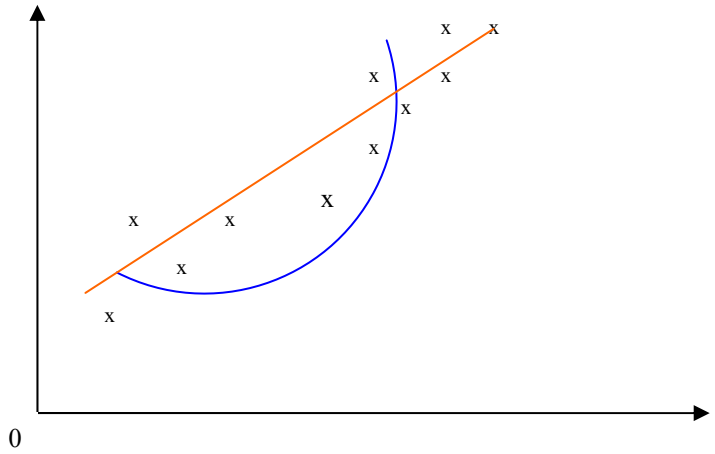
Trên Đồ thị 3.3a, giả sử ta có số quan sát N rất lớn, nhưng với biên độ giao động S_{XX} nhỏ. Khi đó, chỉ cần bỏ đi một quan sát như ứng với điểm A thôi, thì cũng đủ làm các hệ số ước lượng $\{\hat{\alpha}, \hat{\beta}\}$ thay đổi rất mạnh [từ đường mầu đỏ chuyển sang đường tô mầu da cam]. Điều đó chứng tỏ sai số ước lượng, đo bởi $Var\,\hat{\beta}$, là lớn. Ta sẽ xét kỹ hơn vấn đề này trong chương 7 về **đa cộng tuyến** (*multicollinearity*).



Đồ thị 3.3b: Ước lượng có độ chính xác cao hơn, ứng với S_{XX} lớn hơn.

Trên Đồ thị 3.3b, việc loại bỏ đi một vài quan sát, như điểm A, sẽ ít làm thay đổi các hệ số ước lượng. Kết quả ước lượng có độ ổn định cao hơn và chính xác hơn.

Tuy nhiên, những nhận xét trên chỉ đúng, khi giả thuyết **tuyến tính** của đường hồi quy là đúng. Đôi khi, giá trị rất lớn của S_{XX} lại hàm ý rằng giả thuyết tuyến tính là đáng nghi vấn:



Đồ thị 3.3c: Quy luật tổng thể không phải là tuyến tính (gây nên S_{XX} lớn)

Đồ thị 3.3c thể hiện rằng, việc hiểu sai về bản chất kinh tế đã gây nên việc áp dụng sai mô hình hồi quy tuyến tính. Những sai lầm kiểu như vậy dẫn đến yêu cầu phải kiểm định giả thuyết thống kê về tính **có ý nghĩa** của các tham số của mô hình. Đó là chủ đề của phần 3.4.2 của chương này. Việc sử dụng các dạng hàm khác nhau (*functional forms*) để mô tả quy luật chi phối các dữ liệu quan sát $\{x_n, y_n\}_{n=1}^{N}$ là một chủ đề khác nữa, mà nó cũng sẽ được đề cập trong chương 6.

3.4 Kiểm định giả thuyết thống kê

Để có màu sắc kinh tế, ta hãy xét vấn đề kiểm định thông qua một ví dụ cụ thể.

Ví dụ 3.5: Một công ty bảo hiểm ở Mỹ muốn kinh doanh bảo hiểm nhân thọ. Họ tiến hành nghiên cứu tiềm năng của thị trường sở tại. Lý luận kinh tế đã chỉ ra rằng, yêu cầu về mua bảo hiểm tăng lên cùng với khả năng xảy ra rủi ro, với quy mô về tổn thất tài chính khi rủi ro xảy ra, và với tâm lý ngại rủi ro của cá nhân. Họ nhận định rằng, gia đình càng giầu có nhờ kinh doanh, thì người chủ gia đình càng chịu nhiều stress. Tức là, những người lệ thuộc càng ngại rủi ro gây nên bởi stress cho người chủ gia đình, hơn là tại những gia đình có thu

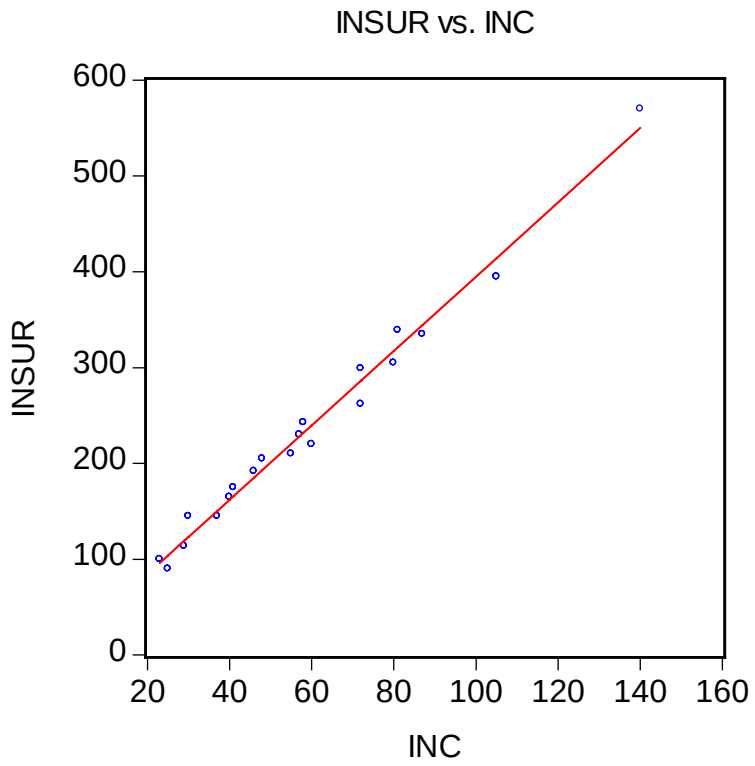
nhập thấp, ít tham dự vào kinh doanh. Vì vậy, ban nghiên cứu thị trường của công ty bảo hiểm này đề xuất mô hình sau:

$$INS = \alpha + \beta INC$$

Trong đó, *INS* là giá trị hợp đồng bảo hiểm, được trả cho bên mua bảo hiểm, nếu xảy ra rủi ro. Và *INC* là thu nhập. Cả hai biến lượng đều tính bằng nghìn dollars. Dữ liệu điều tra và kết quả ước lượng được ghi lại trong các bảng dưới đây

obs	INSUR	INC	obs	INSUR	INC
1	90	25	11	230	57
2	165	40	12	262	72
3	220	60	13	570	140
4	145	30	14	100	23
5	114	29	15	210	55
6	175	41	16	243	58
7	145	37	17	335	87
8	192	46	18	299	72
9	395	105	19	305	80
10	339	81	20	205	48

Bảng 3.1: Số liệu điều tra về nhu cầu mua bảo hiểm



Đồ thị 3.4: Nhu cầu mua bảo hiểm

Sử dụng eviews, chúng ta nhận được kết quả hồi quy dưới đây:

Dependent Variable: INSUR
Method: Least Squares
Date: 04/21/07 Time: 21:41
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.854991	7.383473	0.928424	0.3655
INC	3.880186	0.112125	34.60601	0.0000

R-squared	0.985192	Mean dependent var	236.9500
Adjusted R-squared	0.984370	S.D. dependent var	114.8383
S.E. of regression	14.35730	Akaike info criterion	8.261033
Sum squared resid	3710.375	Schwarz criterion	8.360606
Log likelihood	-80.61033	F-statistic	1197.576
Durbin-Watson stat	3.175965	Prob(F-statistic)	0.000000

Bảng 3.2: Kết quả ước lượng các tham số của mô hình

Kết quả ước lượng được tóm tắt lại như sau:

$$INS = 6.85 + 3.88INC \qquad (3.6)$$
$$(7.38)\ \ (0.11)$$
$$N = 20, \qquad R^2 = 0.98, \qquad ESS = 3710$$

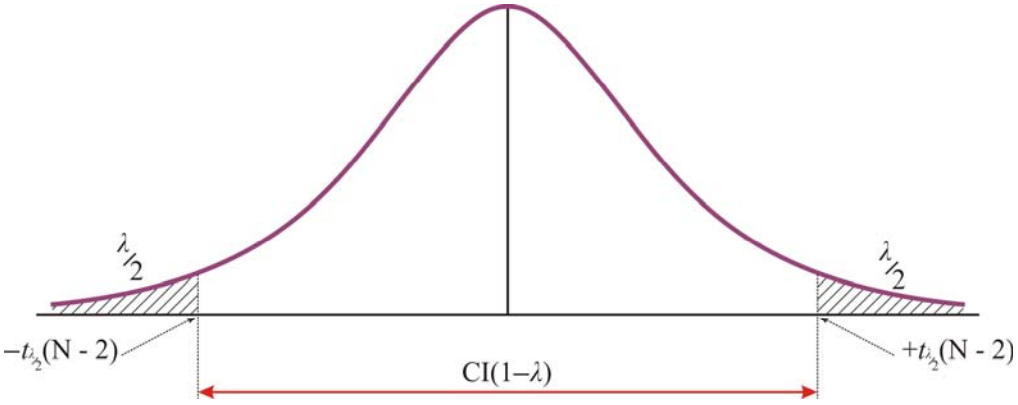
Vấn đề tiếp theo của các nhà hoạch định chiến lược của công ty là liệu họ có thể nói gì về sức mua bảo hiểm tương ứng với từng lớp thu nhập. Điều đó sẽ giúp cho công ty ra quyết định kinh doanh. Ví dụ, nếu thu nhập gia đình tăng thêm một ngàn dollars, thì chi cho bảo hiểm sẽ tăng lên trong khoảng từ 3 ngàn tới 5 ngàn dollars với độ tin cậy là bao nhiêu? Nghĩa là công ty cần xác định được khoảng tin cậy của β tổng thể.

3.4.1 Khoảng tin cậy

Chúng ta sẽ sử dụng các đặc trưng thống kê của ước lượng $\hat{\alpha}, \hat{\beta}$ để đánh giá về các tính chất của tham số thực (tổng thể) α, β.

Từ quan hệ (3.3), $\hat{\beta} = \beta + \sum c_n \varepsilon_n$, và giả thuyết A3 về tính phân bố chuẩn của các yếu tố ngẫu nhiên ε_n, ta đã biết rằng $\hat{\beta}$ có phân bố chuẩn. Hơn thế nữa, từ các đánh giá về trung bình và phương sai của $\hat{\beta}$, ghi trong phương trình (3.4) và (3.5), ta có thể viết lại rằng: $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{XX}})$. Điều đó có nghĩa là, sau khi chuẩn hóa, $Z = \frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / S_{XX}}} \sim N(0,1)$.

Để công thức này có ý nghĩa ứng dụng, ta thay thế σ^2, bởi ức lượng không trệch của nó là $s^2 = \frac{1}{N-2}\sum_n e_n^2 = \frac{1}{N-2} ESS$. Khi đó, thống kê Z chuyển thành thống kê $t = \frac{\hat{\beta} - \beta}{\sqrt{s^2 / S_{XX}}} = \frac{\hat{\beta} - \beta}{se(\hat{\beta})} \sim t(N-2)$. Đồ thị phân bố của thống kê t, trông rất tương tự như thống kê Z:



Đồ thị 3.5: Phân bố $t = \dfrac{\hat{\beta} - \beta}{se(\hat{\beta})} \sim t(N-2)$

Như đã chỉ ra trên Đồ thị 3.5, khoảng tin cậy (Confidence interval) $(1-\lambda)\%$ của thống kê $t = \dfrac{\hat{\beta} - \beta}{se(\hat{\beta})}$ là vùng mà t sẽ rơi vào khoảng đó với xác suất là $(1-\lambda)$. Tức là:

$$\Pr ob\{-t_{\lambda/2}(N-2) \le \frac{\hat{\beta} - \beta}{se(\hat{\beta})} \le t_{\lambda/2}(N-2)\} = (1-\lambda).$$

Nói khác đi, ta có:

$$\beta \in \{\hat{\beta} \pm se(\hat{\beta}) t_{\lambda/2}(N-2)\} \quad \text{với độ tin cậy } (1-\lambda)\% \qquad (3.7)$$

Chẳng hạn, trong ví dụ về công ty bảo hiểm (3.6), ta có: $\hat{\beta} = 3.88$; $se(\hat{\beta}) = 0.112$. Lưu ý rằng $t_{0.025}[18] = 2.101$, độ tin cậy 95% của β tổng thể là:

$$\beta \in \{3.88 \pm 0.112 \times 2.101\} \qquad (3.8)$$

3.4.2 Kiểm định giả thuyết thống kê

Thông thường, kết quả ước lượng mô hình (3.6) và đánh giá độ tin cậy (3.8) sẽ được đính kèm trong bản báo cáo đưa lên cho ban giám đốc công ty để ra quyết định về chiến lược kinh doanh. Tuy nhiên, công việc nghiên cứu thị trường không chỉ dừng lại tại đó. Chúng ta

tiếp tục ví dụ bảo hiểm bằng việc nói rằng, ban giám đốc công ty họp để đánh giá bản báo cáo này. Sau đây là những ghi chép được từ cuộc họp:

Nhà quản lý M1 nói rằng, theo kinh nghiệm của ông, thu nhập đã được thể chế hóa qua các tài sản tài chính, như cổ phiếu, địa ốc, vân vân. Và ảnh hưởng của thu nhập bằng tiền mặt tới chi tiêu cho bảo hiểm nhân thọ là rất yếu.

Thành viên khác của ban giám đốc, nhà quản lý M2 lại cho rằng, thu nhập bằng tiền có ảnh hưởng rất mạnh tới nhu cầu mua bảo hiểm nhân thọ. Kinh nghiệm làm ăn của ông cho thấy, cứ 1000 dollars tăng thêm về thu nhập sẽ kéo theo giá trị gói bảo hiểm mua bởi hộ gia đình tăng lên 5000 dollars.

Cuối cùng, ông M3 nêu lại rằng, thu nhập bằng tiền đúng là có ảnh hưởng, nhưng không mạnh tới như vậy. Cứ 1000 dollars tăng thêm về thu nhập chỉ kéo theo nhu cầu về bảo hiểm tăng lên 4000 dollars.

Vậy ai trong số họ là đúng? Và nếu nhận định của nhà quản lý M1 đúng, thì thật là rất đáng tiếc. Vì vậy, chúng ta cần tiến hành kiểm định lại những nhận định này.

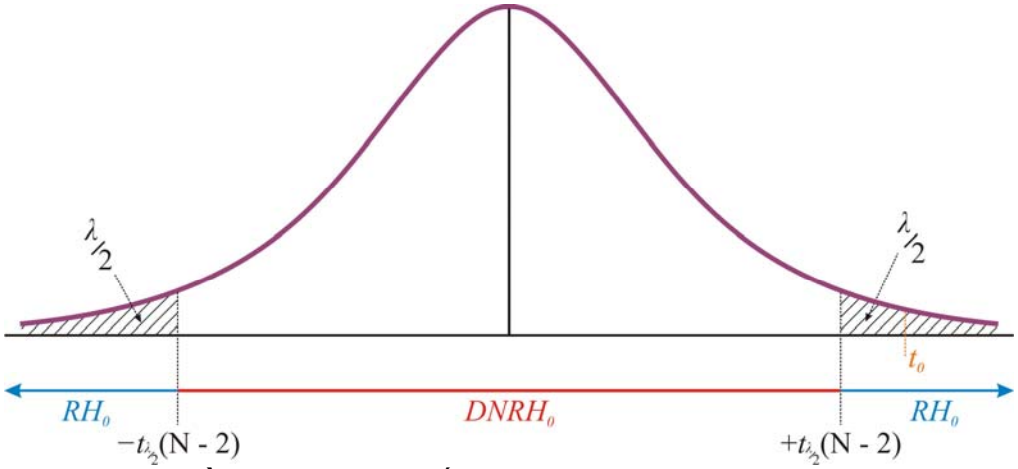
Một cách tổng quát, ta tiến hành kiểm định giả thiết thống kê như sau:

$$H_0 : \beta = \beta_0 \quad \text{.vs.} \quad H_1 : \beta \neq \beta_0$$

Ví dụ, theo nhận định của nhà quản lý công ty M1, ta có:

$$H_0 : \beta = 0 \quad \text{.vs.} \quad H_1 : \beta \neq 0$$

Logic chung của vấn đề kiểm định giả thuyết là như sau: Nếu như nhận định của anh là đúng, thì nó phải phù hợp với phần lớn trường hợp quan sát thấy trên thực tế. Tức là, giá trị thống kê $t_0 = \dfrac{\hat{\beta} - \beta_0}{se(\hat{\beta})}$ phải rơi vào khoảng tin cậy, chẳng hạn là 95%. Trong trường hợp đó, ta không bác bỏ giả thuyết H_0 (hay ký hiệu bằng tiếng Anh: $DNRH_0$). Nếu giá trị $t_0 = \dfrac{\hat{\beta} - \beta_0}{se(\hat{\beta})}$ nằm ngoài khoảng tin cậy, tức là rơi vào vùng hiếm quan sát thấy trên thực tế, khi đó ta bác bỏ H_0 (hay ký hiệu là RH_0).



Đồ thị 3.6: Vùng chấp nhận và bác bỏ H_0

Đồ thị 3.6 thể hiện rằng, chúng ta sẽ bác bỏ (RH_0), nếu $|t_0| = \left|\frac{\hat{\beta} - \beta}{se(\hat{\beta})}\right| \geq t_{\lambda/2}(N-2)$, và chúng ta sẽ không bác bỏ ($DNRH_0$), nếu $\left|\frac{\hat{\beta} - \beta}{se(\hat{\beta})}\right| \leq t_{\lambda/2}(N-2)$.

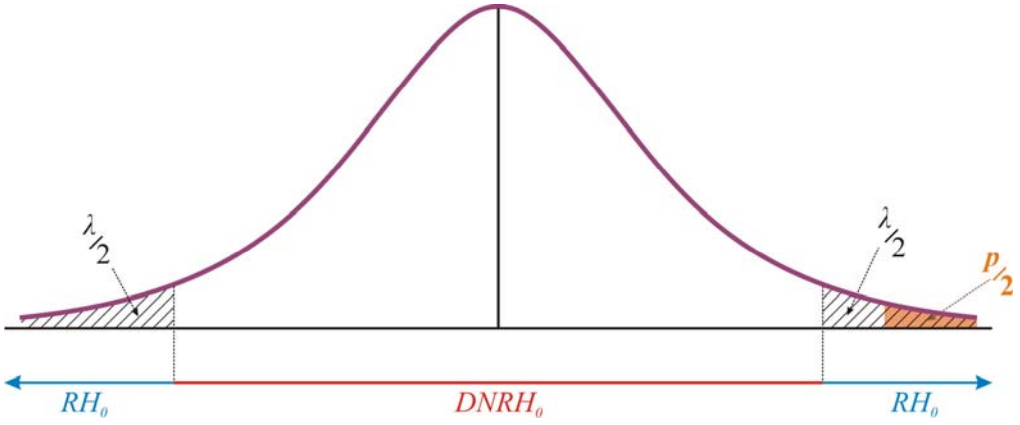
Trong ví dụ nêu trên, đối với nhận định của nhà quản lý M1, ta tiến hành kiểm định như sau:

$$|t_0| = \left|\frac{3.88}{0.112}\right| = 34.6 \geq 2.01 = t_{0.025}[18]$$

Như vậy, dựa trên kết quả kiểm định, ta có thể bác bỏ mạnh mẽ giả định của nhà quản lý M1. Bây giờ chúng ta hãy thử tự kiểm định xem nhận định của các nhà quản lý M2 và M3 có đúng không.

Cuối cùng, để cho tiện sử dụng, trong các software ứng dụng như eviews, người ta thường cho biết giá trị **p-value,** được định nghĩa như sau:

$$P - value = \Pr ob\{|t(N-2)| \geq |t_0|\}$$



Đồ thị 3.7: biểu diễn của p-value

Vì vậy, chúng ta sẽ bác bỏ giả thuyết (RH_0), nếu: $p-value \le \lambda$, [như chỉ ra trên đồ thị 3.7]. Và chúng ta sẽ không bác bỏ giả thuyết đó ($DNRH_0$), nếu $p-value \ge \lambda$.

CHƯƠNG 4: HỒI QUI ĐA BIẾN

Mô hình hồi quy đơn đã trình bày ở các chương 2 và 3 là khá hữu dụng cho rất nhiều trường hợp khác nhau. Mặc dù vậy, nó trở nên không còn phù hợp nữa khi có nhiều hơn một yếu tố tác động đến biến cần được giải thích. **Hồi quy đa biến** cho phép chúng ta nghiên cứu những trường hợp như vậy. Hãy xét các ví dụ sau:

4.1 Giới thiệu về hồi quy đa biến

Ví dụ 4.1: Rất nhiều các nghiên cứu trên thế giới quan tâm tới mối quan hệ giữa thu nhập và trình độ học vấn. Chúng ta kỳ vọng rằng, ít ra về trung bình mà nói, học vấn càng cao, thì thu nhập càng cao. Vì vậy, chúng ta có thể lập phương trình hồi quy sau:

$$Thu\ nhập = \beta_1 + \beta_2 Học\ vấn + \varepsilon$$

Tuy nhiên, mô hình này đã bỏ qua một yếu tố khá quan trọng là mọi người thường có mức thu nhập cao hơn khi họ làm việc lâu năm hơn, bất kể trình độ học vấn của họ thế nào. Vậy nên, mô hình tốt hơn cho mục đích nghiên cứu của chúng ta sẽ là:

$$Thu\ nhập = \beta_1 + \beta_2 Học\ vấn + \beta_3 Tuổi + \varepsilon$$

Nhưng người ta cũng thường quan sát thấy, thu nhập có xu hướng tăng chậm dần khi người ta càng nhiều tuổi hơn so với thời trẻ. Để thể hiện điều đó, chúng ta mở rộng mô hình như sau:

$$Thu\ nhập = \beta_1 + \beta_2 Học\ vấn + \beta_3 Tuổi + \beta_4 Tuổi^2 + \varepsilon$$

Và chúng ta sẽ kỳ vọng rằng, β_3 mang dấu dương, và β_4 mang dấu âm.

Như vậy, chúng ta đã rời bỏ thế giới của hồi quy đơn và bước sang hồi quy đa biến.

Ví dụ 4.2: Nghiên cứu về nhu cầu đầu tư ở Mỹ trong khoảng thời gian từ năm 1968 – 1982.

Ở Mỹ, thời kỳ này mang dấu ấn lịch sử là cuộc chiến tranh Việt Nam kéo dài, dẫn đến bội chi ngân sách và lạm phát. Một năm sau khi chiến tranh kết thúc, lạm phát ở Mỹ đã đạt tới mức kỷ lục là 9.31% vào năm1976. Điều đó dẫn đến việc ngân hàng trung ương phải áp dụng mạnh mẽ chính sách tiền tệ chặt, vốn đã được áp dụng trong vài năm trước, và đưa

mức lãi suất lên tới mức cao kỷ lục là 7.83%. Khi sự dính líu của Mỹ về quân sự tại Việt Nam đã hoàn toàn chấm dứt, nguồn nhân lực trước đây phục vụ cho chiến tranh nay chuyển ào ạt sang khu vực thương mại. Và điều này lại lại làm dấy lên một đợt lạm phát mới, đạt tới 9.44% vào năm 1981, sau đó được đưa về mức 5.99% vào năm 1982 nhờ vào việc nâng lãi suất lên tới 13.42%. Như vậy, lịch sử kinh tế Mỹ trong thời kỳ này được đặc trưng bởi chính sách tiền tệ chặt, kéo theo xu hướng cắt giảm liên tục về đầu tư qua các năm.

Chính vì vậy, các nhà nghiên cứu Mỹ đã đề xuất mô hình nghiên cứu sau về cầu đầu tư vào giai đoạn này:

$$INV = \beta_1 + \beta_2 T + \beta_3 G + \beta_4 INT + \varepsilon$$

Trong đó, *INV* và *G* lần lượt là cầu về đầu tư và GNP thực tế, đơn vị trillions dollars; *INT* là lãi suất; và *T* là biến xu thế, tính theo thời gian đã trôi qua, kể từ năm 1968. Từ lý luận kinh tế vĩ mô, chúng ta kỳ vọng rằng, β_3 mang dấu dương, và β_4 mang dấu âm. Và vì đây là thời kỳ đầu tư ở Mỹ có xu thế bị co hẹp, chúng ta cũng kỳ vọng rằng β_2 mang dấu âm.

Sử dụng dữ liệu thống kê vĩ mô của nền kinh tế Mỹ, từ năm 1968 - 1982 [xem bảng dữ liệu 4.2 phía dưới], kết quả ước lượng của mô hình hồi quy này như sau:

Bảng** Error! No text of specified style in document.**.1: Bảng kết xuất mô hình hồi qui các yếu tố ảnh hưởng đến cầu về đầu tư của Mỹ trong giai đoạn từ 1968 - 1982

Dependent Variable: INV
Method: Least Squares
Date: 04/09/07 Time: 16:14
Sample: 1 15
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.509237	0.052526	-9.694973	0.0000
T	-0.016583	0.001880	-8.819528	0.0000
G	0.670266	0.052426	12.78506	0.0000
INT	-0.002365	0.001034	-2.287282	0.0430

R-squared	0.972420	Mean dependent var	0.203333
Adjusted R-squared	0.964898	S.D. dependent var	0.034177
S.E. of regression	0.006403	Akaike info criterion	-7.040816
Sum squared resid	0.000451	Schwarz criterion	-6.852003
Log likelihood	56.80612	F-statistic	129.2784
Durbin-Watson stat	1.958353	Prob(F-statistic)	0.000000

Dưới dạng báo cáo, kết quả đó có thể được viết tóm tắt như dưới đây:

$$INV = \underset{(0.0525)}{-0.5092} - \underset{(0.0018)}{0.0165T} + \underset{(0.052)}{0.67G} - \underset{(0.001)}{0.0023\,INT}$$

$$R^2 = 0.972\,,\ N = 15,\ ESS = 0.00045$$

Nếu viết dưới dạng sai phân, ta có:

$$\Delta INV = -\,0.0165\,\Delta T + 0.67\,\Delta G - 0.0023\,\Delta INT$$

Nói khác đi, **nếu các yếu tố khác được giữ không đổi**, cứ sau mỗi một năm, kể từ năm 1968 (tức là $\Delta T = 1$), nhu cầu đầu tư sẽ bị giảm là -0.0165 trillions dollars. Cũng như vậy, nếu bỏ qua yếu tố xu thế và lãi suất, tác động riêng phần của việc tăng GNP lên 0.1 trillions dollars ($\Delta G = 0.1$), sẽ làm cầu về đầu tư tăng lên thêm 0.067 trillions; và nếu đẩy lãi suất lên thêm 1% ($\Delta INT = 1$), trong khi giữ nguyên các yếu tố còn lại, thì sẽ làm đầu tư giảm đi là -0.0023 trillions dollars.

Những tính toán trên đây cho thấy có sự tương đồng rõ rệt về cách diễn giải ý nghĩa của các hệ số ước lượng trong mô hình hồi quy đa biến so với trường hợp đơn biến. Điều đó gợi ý rằng, về mặt bản chất, mô hình hồi quy đa biến sẽ chỉ là sự mở rộng của hồi quy đơn biến. Ta sẽ thấy rõ hơn điều đó ở các phần sau.

4.2Biểu diễn đại số của mô hình hồi quy đa biến

Chúng ta hãy đưa ra bảng so sánh về dạng hàm của mô hình hồi quy đa biến so với trường hợp đơn biến:

	Hồi quy đơn biến	**Hồi quy đa biến**
Ví dụ	$CONS = \beta_1 + \beta_2 INC + \varepsilon$	$INV = \beta_1 + \beta_2 T + \beta_3 G + \beta_4 INT + \varepsilon$
Dạng mô hình	$Y = \beta_1 + \beta_2 X + \varepsilon$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$
Với mỗi quan sát	$y_n = \beta_1 + \beta_2 x_n + \varepsilon_n$	$y_n = \beta_1 + \beta_2 x_{n2} + \beta_3 x_{n3} + \beta_4 x_{n4} + \varepsilon_n$

Như vậy, hồi quy đa biến là một sự mở rộng tự nhiên của trường hợp đơn biến, khi số biến giải thích lớn hơn 2, kể cả hằng số. Để cho tiện lợi, chúng ta sẽ đưa vào các ký hiệu vector:

Gọi **vector hàng** $x_n^{'} = (1, x_{n2,} x_{n3,} x_{n4})$ là vector các quan sát thứ $n = 1,2..., N$ của các biến giải thích. [Lưu ý, **dấu phẩy** ở bên phải, phía trên vector x_n là dấu chuyển vị. Như vậy, theo mặc định, mọi vector (mà không có dấu chuyển vị) đều được coi là vector cột].

Từng "cặp" quan sát dữ liệu do vậy, sẽ là $\{y_n, x_n^{'}\}_{n=1}^{N}$.

Để minh họa, trong ví dụ 4.2 về cầu về đầu tư ở Mỹ (1968 – 82), những cặp $(y_5, x_5^{'})$ và $(y_{11}, x_{11}^{'})$ được tô màu:

***Bảng** Error! No text of specified style in document..2: Dữ liệu vĩ mô về đầu tư và các biến giải thích của nền kinh tế Mỹ (1968 – 82).*

Obs (n)	INV (Y)	C (X1)	T (X2)	G (X3)	INT (X4)	
1	0.161	1	1	1.058	5.16	
2	0.172	1	2	1.088	5.87	
3	0.158	1	3	1.086	5.95	
4	0.173	1	4	1.122	4.88	
5	0.195	1	5	1.186	4.5	$(y_5, x_5^{'})$
6	0.217	1	6	1.254	6.44	
7	0.199	1	7	1.246	7.83	
8	0.163	1	8	1.232	6.25	
9	0.195	1	9	1.298	5.5	
10	0.231	1	10	1.37	5.46	
11	0.257	1	11	1.439	7.46	$(y_{11}, x_{11}^{'})$
12	0.259	1	12	1.479	10.28	
13	0.225	1	13	1.474	11.77	
14	0.241	1	14	1.503	13.42	
15	0.204	1	15	1.475	11.02	

Nguồn: Economic Report of the President. Government, Printing Office, Washington D.C., 1983.

Tiếp theo, ta gọi **vector cột** $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$ là vector các tham số tổng thể, cần được ước lượng

Lưu ý rằng, **tích vô hướng** giữa hai vector $x_n^{'}$ và β sẽ tạo lại **phần xu thế** trong vế phải của phương trình hồi quy (4.2):

$$\underset{1\times4}{x_n'}\ \underset{4\times1}{\beta} = (1, x_{n2,} x_{n3,} x_{n4}) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \beta_1 + \beta_2 x_{n2} + \beta_3 x_{n3} + \beta_4 x_{n4}, \qquad n = 1,2..N$$

Vì vậy, ứng với từng "cặp" quan sát $\{x_n', y_n\}_{n=1}^{N}$, ta có thể viết lại phương trình hồi quy đó như sau:

$$y_n = x_n'\beta + \varepsilon_n \qquad n = 1,2,..,N \qquad (4.3)$$

Như vậy, mọi ký hiệu ta đã sử dụng trong ước lượng mô hình hồi quy đơn biến, nay có thể được sử dụng lại cho mô hình hồi quy đa biến. Cụ thể là:

$$\hat{y}_n = x_n'\hat{\beta} \qquad n = 1,2,..,N \qquad (4.3)$$

Và sai số ước lượng hay số dư (residual) sẽ có dạng:

$$e_n = y_n - \hat{y}_n \qquad (4.4)$$

Việc tiến hành ước lượng các tham số của mô hình bằng phương pháp bình phương cực tiểu tương đương với việc giải bài toán sau:

$$S(\hat{\beta}) = \sum\nolimits_n e_n^2 = \sum\ (y_n - x_n'\hat{\beta})^2 \to \min_{\hat{\beta}} \qquad (4.5)$$

Tương tự như trong hồi quy đơn, ở đây, ta sử dụng điều kiện cực trị, (first order condition, FOC), để tìm các tham số ước lượng $\hat{\beta}_k, k = 1,2,3,4..$ Nói khác đi, ta đi giải hệ phương trình sau:

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}_1} = 0$$
$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}_2} = 0$$
$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}_3} = 0$$

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}_4} = 0 \qquad (4.6)$$

Đây là hệ gồm 4 phương trình với 4 ẩn số, mà việc giải nó cho chúng ta tham số ước lượng $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)'$. Sử dụng phần mềm Eviews, kết quả tính toán các tham số này đã được nêu trong bảng báo cáo 4.1 ở trên.

Mặc dù dạng biểu diễn giải tích của vector $\underset{4\times1}{\hat{\beta}}$ là khá phức tạp. Tuy nhiên, về bản chất chúng vẫn không khác gì trường hợp đơn biến. Cụ thể là, tương tự như $\hat{\alpha}$, phương trình đầu tiên của hệ (4.6) để ước lượng $\hat{\beta}_1$ dẫn đến cái điều là, đường hồi quy đi qua điểm trung bình $(\bar{y}_n, \bar{x}_n')$. Và vì vậy, ta cũng có thể nói đến tiêu chuẩn đo lường độ phù hợp của đường hồi quy R^2. Cụ thể là từ mối quan hệ (4.4):

$$y_n = \hat{y}_n + e_n$$

Hay cũng hệt như thế:

$$(y_n - \bar{y}) = \hat{y}_n - \bar{y} + e_n$$

Người ta có thể viết lại nó như sau:

$$(y_n - \bar{y}) = (x_n' - \bar{x}')\hat{\beta} + e_n$$

Tức là, sự giao động so với trung bình của biến *Y* được giải thích một phần bởi mô hình, và phần còn lại là sai số e_n, chưa được giải thích bởi mô hình. Sử dụng các điều kiện tìm cực trị (4.6), ta cũng có thể viết lại quan hệ đó như sau:

$$\sum_n (y_n - \bar{y})^2 = \sum_n (\hat{y}_n - \bar{y})^2 + \sum_n e_n^2$$

Hay cũng vậy,

$$\text{TSS} = \text{RSS} + \text{ESS}$$

Vì thế, chúng ta có thể đưa ra định nghĩa:

$$R^2 = 1 - \frac{ESS}{TSS} \qquad (0 \le R^2 \le 1).$$

và sử dụng nó làm thước đo mức độ phù hợp của đường hồi quy với dữ liệu quan sát.

Phần tiếp sau sẽ đề cập đến bản chất thống kê của mô hình hồi quy đa biến.

4.3 Bản chất thống kê của mô hình hồi quy đa biến

Từ bây giờ, chúng ta sẽ sử dụng dạng tổng quát của mô hình hồi quy đa biến:

$$\begin{aligned} Y &= \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + ... + \beta_K X_K + \varepsilon \\ y_n &= \beta_1 + \beta_2 x_{n2} + ... + \beta_k x_{nk} + ... + \beta_K x_{nK} + \varepsilon_n \\ &= x_n' \beta + \varepsilon_n, \qquad n = 1,2,3,..., N \end{aligned} \tag{4.7}$$

Việc hồi quy mô hình (4.7) sẽ cho ta biểu diễn sau:

$$\begin{aligned} Y &= \hat{\beta}_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k + ... + \hat{\beta}_K X_K + e \\ y_n &= \hat{\beta}_1 + \hat{\beta}_2 x_{n2} + ... + \hat{\beta}_k x_{nk} + ... + \hat{\beta}_K x_{nK} + e_n \\ &= x_n' \hat{\beta} + e_n, \qquad n = 1,2,3,..., N \end{aligned} \tag{4.8}$$

Trong đó, N là số quan sát, và K là số biến giải thích.

Ta phát biểu định lý sau[1]:

Định lý 4.1: Phương pháp bình phương cực tiểu, áp dụng cho mô hình hồi quy đa biến, sẽ cho ta các tham số ước lượng dưới dạng sau:

$$\hat{\beta}_k = \beta_k + \sum_n c_{kn} \varepsilon_n, k = 1,2,.., K \tag{4.9}$$

Cũng như trường hợp đơn biến, phương trình (4.9) chỉ ra rằng: $\hat{\beta}_k$ bị tác động bởi các yếu tố ngẫu nhiên ε_n, làm giá trị của nó không trùng khít với β_k tổng thể. Và vì bị tác động

[1] Xem chứng minh chi tiết ở chương 8, phần Maximum likelihood.

bởi các yếu tố ngẫu nhiên, $\hat{\beta}_k$ cũng là một biến ngẫu nhiên. Do đó, độ tốt của ước lượng sẽ phụ thuộc trực tiếp vào bản chất của các quá trình ngẫu nhiên $\{\varepsilon_n\}_{n=1}^{N}$.

Điều này dẫn đến việc cần phải khắc họa bản chất thống kê của mô hình hồi quy, như chúng ta đã làm cho trường hợp đơn biến. Ta sẽ tiếp tục sử dụng các giả thuyết đã đưa ra về ε_n. Cụ thể là:

A1 $E\varepsilon_n = 0,$ với mọi *n*.

A2 $Var\varepsilon_n = \sigma^2$, với mọi *n*.

A3 $\varepsilon_n \overset{iid}{\sim} N(0,\sigma^2)$, với mọi *n*. Và:

A4 $E(y_n \mid x_n^{'}) = x_n^{'}\beta$, với mọi *n*.

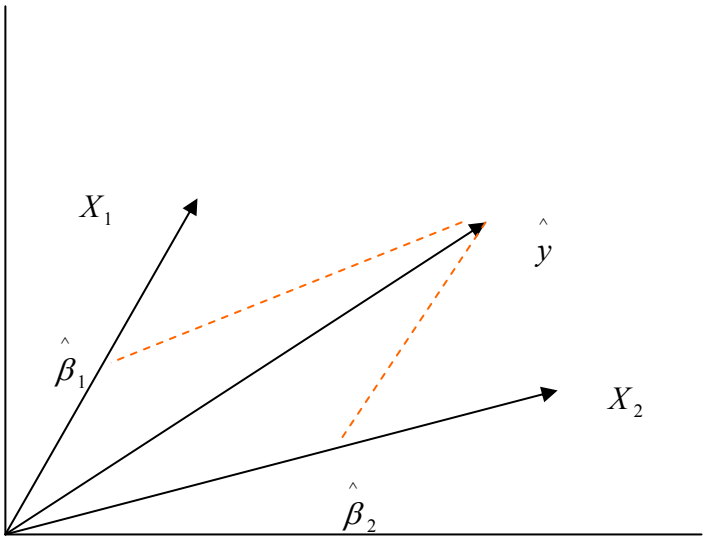
Đối với trường hợp đa biến, chúng ta đưa thêm đòi hỏi sau:

Gọi $\underset{N\times K}{X} = [X_1, X_2, ..., X_k, ..., X_K]$ là ma trận tạo bởi các vector cột của *K* biến giải thích [xem lại ví dụ minh họa về ma trận *X* ở bảng 4.2 về dữ liệu của mô hình đầu tư]. Khi đó, ta đòi hỏi rằng:

A5 Các cột $\{X_1, X_2, ..., X_k, ..., X_K\}$ là độc lập tuyến tính. Hay cũng vậy, rank *X* = *K*.

Về mặt hình học, giả thuyết này có ý nghĩa như sau. Hãy xét trường hợp *K = 2*, phương pháp bình phương cực tiểu có thể được biểu diễn bởi lược đồ dưới đây:

Đồ thị** Error! No text of specified style in document..**1: Biểu diễn hình học của hồi quy



Việc ước lượng tham số $\hat{\beta}$ cũng giống như là tìm các hệ số $\underset{2\times1}{\hat{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ sao cho $\hat{y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$. Để làm được điều đó, điều kiện cần là các vector X_1, X_2 không được trùng khít với nhau. Hay cũng vậy, X_1, X_2 phải độc lập tuyến tính. Đây được gọi là **điều kiện xác định** (identification condition). Trong trường hợp tổng quát, khi $K \geq 2$, điều kiện đó được phát biểu dưới dạng giả thuyết A5. Chúng ta sẽ sử dụng giả thuyết này khi bàn tới vấn đề đa cộng tuyến (multicollinearity) trong chương 7.

4.4 Kiểm định các giả thuyết thống kê

Bây giờ hãy chỉ chú ý đến giả thuyết đầu tiên A1 – A3, và sử dụng chúng để đánh giá tính tốt của ước lượng theo các tiêu chuẩn thống kê.

Từ phương trình (4.9), ta đã có: $\hat{\beta}_k = \beta_k + \sum c_{kn}\varepsilon_n$. Bây giờ, hãy áp dụng toán tử kỳ vọng vào hai vế của (4.9). Ta có:

$$\begin{aligned} E\hat{\beta}_k &= E(\beta_k + \sum c_{kn}\varepsilon_n) \\ &= \beta_k + \sum c_{kn} E\varepsilon_n \\ &= \beta_k \end{aligned} \qquad (4.10)$$

[ở đây, ta sử dụng giả thiết A1: $E\varepsilon_n = 0$]. Do vậy, $\hat{\beta}_k$ là ước lượng không chệch của β_k.

Tiếp theo, sử dụng lại công thức: $Var(x) = Var(x - Ex)$ [xem chương 1, phần ôn tập], và lưu ý (4.9), (4.10), ta có:

$$\begin{aligned} Var\hat{\beta}_k &= Var(\hat{\beta}_k - \beta_k) \\ &= Var(\sum c_{kn}\varepsilon_n) \end{aligned}$$

Sử dụng giả thiết A3 về tính độc lập của các yếu tố ngẫu nhiên, cuối cùng ta nhận được:

$$\begin{aligned} Var\hat{\beta}_k &= \sum c_{kn}^2 Var\varepsilon_n \\ &= \sigma^2 \sum c_{kn}^2 \text{ , hay} \\ Var\hat{\beta}_k &= \frac{\sigma^2}{S_{kk}}, k = 1,2,..,K \end{aligned} \qquad (4.11)$$

(ở đây, mặc dù ta không đưa ra được tính toán trực tiếp; nhưng về cơ bản S_{kk} cũng là phương sai mẫu của biến X_k, tương tự như S_{XX} trong trường hợp đơn biến).

Định Lý 4.2 [Gauss – Markov]: Phương pháp bình phương cực tiểu có sai số ước lượng, đo lường bởi $Var\hat{\beta}_k, k = 1,2,..,K$, là nhỏ nhất trong lớp tất cả các ước lượng tuyến tính và không chệch.

Ta cũng nên nhấn mạnh lại rằng, chúng ta có được những tính chất rất tốt: **không chệch** và **hiệu quả** của ước lượng bình phương cực tiểu, mà chỉ đòi hỏi có trung bình bằng zero, tính độc lập, và phương sai giống nhau của các yếu tố ngẫu nhiên – tức là giả thiết A3.

Sử dụng (4.9) – (4.11), chúng ta đi đến kết luận rằng: $\hat{\beta}_k \sim N(\beta_k, \frac{\sigma^2}{S_{kk}})$. Điều đó có nghĩa là, sau khi chuẩn hóa, $Z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 / S_{kk}}} \sim N(0,1)$. Thay thế σ^2 bởi ước lượng không chệch

của nó là $s^2 = \frac{1}{N-K}\sum_n e_n^2$, ta có thống kê $t_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{s^2/S_{kk}}} = \frac{\hat{\beta}_k - \beta_k}{se(\hat{\beta}_k)} \sim t(N-K)$. Chúng ta bây giờ có thể xây dựng khoảng tin cậy cho $\beta_k, k = 1,2,..,K$, và tiến hành kiểm định các giả thuyết thống kê về tham số của tổng thể. Chẳng hạn như nếu muốn kiểm định tính có ý nghĩa của biến giải thích $X_k, k = 1,2,..,K$, chúng ta lập giả thuyết sau:

$$H_0 : \beta_k = 0 \quad .\text{vs.} \quad H_1 : \beta_k \neq 0$$

Việc kiểm định bao gồm các bước sau:

Bước 1: Xác định thống kê – t (t-stat):

$$t_k = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \sim t(N-K)$$

Bước 2: Tra bảng thống kê t-student với *(N-K)* bậc tự do $t(N-K)$ để tìm giá trị t-tra bảng (t-critical) $t_{\lambda/2}(N-K)$, ứng với mỗi mức ý nghĩa (significance) λ [Chẳng hạn, 0.05 (5%); hay 0.1 (10%)]

Bước 3: Bác bỏ giả thuyết H_0 (viết tắt là RH_0), nếu $|t_k| = \left|\frac{\hat{\beta}_k}{se(\hat{\beta}_k)}\right| \geq t_{\lambda/2}(N-K)$, và không bác bỏ giả thuyết đó ($DNRH_0$), nếu $\left|\frac{\hat{\beta}_k}{se(\hat{\beta}_k)}\right| \leq t_{\lambda/2}(N-K)$.

Cũng như trong trường hợp đơn biến, người ta thường hay sử dụng *p-value*, hơn là phải tính toán và tra bảng qua các bước 1 đến 3 như trên.

Cụ thể, ứng với từng biến giải thích $X_k, k = 1,2,..,K$, ta cũng đặt:

$$p - value = \Pr ob\{|t(N-K)| \geq |t_k|\}$$

Cũng hệt như ở đồ thị 3-9, chúng ta sẽ bác bỏ giả thuyết $H_0 : \beta_k = 0$, nếu: $p - value \leq \lambda$, [trong trường hợp đó, ta nói X_k là có ý nghĩa ở mức $\lambda\%$]. Và chúng ta sẽ không bác bỏ giả thuyết đó, nếu $p - value \geq \lambda$.

Trong ví dụ 4.1 về đầu tư ở Mỹ (1968-82), *p-value* của cả 3 biến giải thích: *T, G, INT*, đều nhỏ hơn 5%. Vì vậy, ta nói rằng tất cả các biến này là có ý nghĩa ở mức $\lambda = 5\%$.

CHƯƠNG 5: LỰA CHỌN MÔ HÌNH VÀ VẤN ĐỀ KIỂM ĐỊNH

Trên thực tế, việc lập mô hình và ước lượng không phải là một vấn đề đơn giản. Chẳng hạn như trong ví dụ 4.2 về nhu cầu đầu tư ở Mỹ (1968 – 82). Cho dù lý thuyết kinh tế vĩ mô đã gợi ý rằng, cầu về đầu tư chịu ảnh hưởng bởi hai yếu tố chính là GNP và lãi suất. Tuy nhiên, việc Ngân hàng trung ương Mỹ sử dụng chính sách tiền tệ chặt trong thời kỳ đó đã đòi hỏi ta phải đưa thêm biến xu thế vào mô hình để giải thích cho cầu về đầu tư. Việc thêm hoặc bớt biến giải thích như vậy làm nẩy sinh một loạt các câu hỏi: Liệu ta nên thêm hoặc bớt những biến nào trong phương trình hồi quy? Chẳng hạn, liệu việc chỉ đưa thêm biến xu thế vào mô hình như vậy là đã đủ chưa? Hay cần phải đưa thêm nhiều biến giải thích khác nữa, như tỷ lệ lạm phát, số lượng quân nhân giải ngũ, vân vân? Trong rất nhiều sự lựa chọn như vậy, mô hình nào là tốt nhất? Và dựa trên tiêu chuẩn đánh giá nào? Ngược lại, nếu giả sử ta áp dụng một cách máy móc lý thuyết ghi trong sách giáo khoa, và bỏ quên, không đưa biến xu thế vào mô hình, thì hậu quả gì sẽ xẩy ra cho ước lượng và dự báo? Đó là những câu hỏi chúng ta muốn trả lời trong chương này.

5.1 Phân tích kết quả hồi quy

Chúng ta hãy bắt đầu bằng ví dụ phân tích một kết quả hồi quy đưa ra trong Ramanathan (1989):

Ví dụ 5.1: Một công ty bất động sản nghiên cứu giá các căn hộ cho những gia đình trẻ. Họ lập mô hình hồi quy như sau:

$$PRICE = \beta_1 + \beta_2 SQFT + \beta_3 BEDRMS + \beta_4 BATHS + \varepsilon \qquad (5.1)$$

Ở đó, *PRICE* là giá căn hộ tính theo nghìn dollars; bên cạnh diện tích sử dụng *SQFT*, (tính theo đơn vị tương tự như mét vuông), giá căn hộ còn chịu ảnh hưởng bởi số lượng phòng ngủ *BEDRMS*, và số nhà tắm *BATHS*. Vì đây đều là các đặc trưng về tính tốt của căn hộ, ta kỳ vọng rằng các hệ số $\beta_2, \beta_3, \beta_4$ đều dương.

Một trong ích lợi cơ bản của phương pháp hồi quy đa biến là nó cho phép đánh giá **tác động riêng phần** của từng yếu tố giải thích lên biến được giải thích. Chẳng hạn, nếu ta có hai căn hộ giống hệt nhau về diện tích sử dụng (*SQFT*) và số nhà tắm (*BATHS*). Nhưng chúng khác nhau về số phòng ngủ (*BEDRMS*). Khi đó, hệ số ước lượng $\hat{\beta}_3$ sẽ cho phép

chúng ta đánh giá liệu giá căn hộ có thêm một phòng ngủ sẽ đắt hơn là bao nhiêu so với căn hộ còn lại.

Để làm những so sánh đó, ta cần tiến hành ước lượng mô hình hồi quy (5.1). Dữ liệu điều tra cho việc ước lượng được ghi ở bảng 5.1 dưới đây:

Bảng 5.1: Dữ liệu điều tra về giá cả các căn hộ

obs	PRICE	CONSTANT	SQFT	BEDRMS	BATHS
	Y	X1	X2	X3	X4
1	199.9	1	1065	3	1.75
2	228	1	1254	3	2
3	235	1	1300	3	2
4	285	1	1577	4	2.5
5	239	1	1600	3	2
6	293	1	1750	4	2
7	285	1	1800	4	2.75
8	365	1	1870	4	2
9	295	1	1935	4	2.5
10	290	1	1948	4	2
11	385	1	2254	4	3
12	505	1	2600	3	2.5
13	425	1	2800	4	3
14	415	1	3000	4	3

Sau đây là kết quả ước lượng mô hình hồi quy mô hình (5.1):

$$PRICE = 129.062 + 0.1548 SQFT - 21.588 BEDRMS - 12.193 BATHS$$

Điều chúng ta nhận thấy ngay là dấu của các hệ số đi kèm với *BEDRMS* và *BATHS* là không giống với kỳ vọng. Thông thường, ta sẽ nghĩ rằng, nếu tăng thêm số lượng phòng ngủ hoặc nhà tắm, thì giá trị căn hộ phải đắt lên. Liệu kết quả ước lượng trên đây có phải là một điều bất hợp lý hay không?

Nhìn kỹ hơn, chúng ta vẫn có thể tìm được một cách diễn giải hợp lý, nếu xét đến **tác động riêng phần** của từng biến giải thích lên giá cả. Giả sử ta giữ nguyên diện tích sử dụng (*SQFT*) và số lượng phòng tắm (*BATHS*). Kết quả ước lượng nói lên rằng, nếu tăng thêm một phòng ngủ, thì về trung bình, giá của căn hộ sẽ giảm đi là 21,588 (21 nghìn 588) dollars. Vấn đề là, cũng vẫn cùng một diện tích sử dụng như vậy, nhưng nay bị chia nhỏ ra để có thêm phòng ngủ. Do vậy, từng phòng ngủ sẽ trở nên chật trội hơn. Và người tiêu dùng không thích việc làm như vậy. Họ chỉ sẵn sàng chi trả ở mức thấp hơn.

Tương tự như vậy, nếu số lượng nhà tắm tăng thêm một, mà diện tích và số phòng ngủ vẫn giữ nguyên, thì giá trị căn hộ sẽ giảm đi là 12,193 (12 nghìn 193) dollars.

Những phân tích trên đây về tác động riêng phần của các nhân tố cho thấy, những điều mà xem ra có vẻ là không hợp lý, thì bây giờ lại là có lý.

Bây giờ nếu giả sử chúng ta đồng thời tăng thêm một phòng ngủ và diện tích sử dụng lên 300. Khi đó, tác động đồng thời của những thay đổi đó lên giá cả sẽ là:

$$\begin{aligned}\Delta PRICE &= 0.1548\Delta SQFT - 21.588\Delta BEDRMS \\ &= 0.1548\times 300 - 21.588\times 1 = 24.852\end{aligned}$$

Nói khác đi, về trung bình, giá căn hộ sẽ tăng thêm là 24, 852 (24 nghìn 852) dollars.

Chúng ta cũng có thể tiến hành dự báo cho giá của một căn hộ, chẳng hạn có 4 phòng ngủ (*BEDRMS*), 3 nhà tắm (*BATHS*), với diện tích (*SQFT*) là 2500:

$$\begin{aligned}PRICE &= 129.062 + 0.1548\times 2500 - 21.588\times 4 - 12.193\times 3 \\ &= 391{,}163 \text{ (391 nghìn 163) dollars.}\end{aligned}$$

Như chúng ta thấy, kết quả dự báo là không tồi so với dữ liệu điều tra (rất gần với mẫu quan sát thứ11).

5.2 Lựa chọn mô hình

Bây giờ chúng ta hãy đưa thêm yếu tố tâm lý của người mua vào việc phân tích. Việc người tiêu dùng không thích căn hộ có phòng ngủ hoặc nhà tắm quá chật hẹp thể hiện rằng họ có những đòi hỏi về tiện nghi. Tức là họ yêu cầu phải có một sự phù hợp giữa diện tích sử dụng với số lượng phòng ngủ và phòng tắm trong căn hộ. Khi những đòi hỏi về tính phù hợp đó được chấp nhận bởi số đông, nó trở thành chuẩn mực chi phối cách thiết kế các căn hộ. Vì vậy, thông tin về diện tích có thể là đủ để cho người tiêu dùng đánh giá được giá trị của căn hộ. Điều đó đặt ra vấn đề là, ngoài mô hình đã xét, ta cần phải thử nghiệm nhiều mô hình khác nữa, và chọn ra đâu là cái tốt nhất.

Trong bảng 5.2 có 3 mô hình khác nhau. Mô hình C giống hệt như cái đã phân tích. Ta đưa thêm vào mô hình A và B, theo đó, mô hình A chỉ còn mỗi biến giải thích là diện tích (*SQFT*); trong khi mô hình B vẫn còn giữ lại số phòng ngủ (*BEDRMS*).

Ta quan tâm trước tiên tới độ phù hợp của từng mô hình với dữ liệu điều tra. Nhắc lại là từ chương 4, chúng ta đo mức độ phù hợp đó bởi quan hệ sau:

$$\sum_n (y_n - \bar{y})^2 = \sum_n (\hat{y}_n - \bar{y})^2 + \sum_n e_n^2$$

$$TSS = RSS + ESS$$

Bảng 5.2 đưa ra các con số so sánh giữa các mô hình. Nhìn từ A sang B và C, ta nhận thấy việc đưa thêm biến giải thích vào mô hình làm **tăng** mức độ giải thích của mô hình, thể hiện bởi tổng bình phương các sai số ước lượng (*ESS*) **giảm xuống**. Một cách trực quan, ta có thể lý giải việc *ESS* giảm như sau: Thay vì chỉ có yếu tố diện tích, việc đưa thêm những tính chất tốt khác của căn hộ vào (như số lượng phòng ngủ, nhà tắm, độ dịu của mầu vôi, độ thoáng gió, vân vân) sẽ làm cho việc diễn giải độ khác biệt của giá căn hộ so với trung bình sẽ tốt hơn lên. Vì vậy, việc tăng số biến giải thích trong mô hình luôn làm cho tổng bình phương sai số *ESS* giảm. Và vì vậy, hệ số đánh giá độ phù hợp của mô hình hồi quy là $R^2 = 1 - \frac{ESS}{TSS}$ **luôn luôn tăng**. [Xem hàng thứ nhất và thứ hai ở sau vạch ngang đầu tiên trong bảng 5.2].

Bảng 5.2: Những mô hình ước lượng cho giá các căn hộ

Variable	**model A**	**model B**	**model C**
Constant	52.351	121.179	129.062
	(38.28)	(80.17)	(88.3)
SQFT	0.13875 ***	0.14831 ***	0.1548 ***
	(0.018)	(0.021)	(0.031)
BEDRMS		-23.911	-21.588
		(24.64)	(27.029)
BATHS			-12.193
			(4.25)
ESS	18,274	16,833	16,700
R^2	0.821	0.835	0.836
$\bar{R}^2$	0.806	0.805	0.787
F-STAT	54.861	27.767	16.989
d.f (N-K)	12	11	10
AIC	1,737	1,846	2,112
SCHWAR	1,903	2,177	2,535

Chú thích: số trong ngoặc là standard error. * là ở mức ý nghĩa 0.1; ** là ở mức ý nghĩa 0.05; *** là ở mức ý nghĩa 0.001.

Tuy nhiên việc làm phức tạp hóa mô hình như vậy, nói chung là không được khuyến khích, bởi vì logic của việc lập mô hình là chỉ quan tâm đến việc đánh giá cái chính, chủ yếu, và lọc bỏ những cái không quan trọng ra khỏi phân tích. Ta không muốn đưa vào bức tranh phân tích tất cả mọi thứ trên đời, vì nó sẽ làm mờ đi yếu tố chính mà ta muốn đánh giá.

Về mặt kỹ thuật, việc đưa thêm các biến giải thích ít có ý nghĩa vào mô hình sẽ làm giảm mức độ chính xác của ước lượng, như chỉ ra vắn tắt dưới đây:

Như đã nêu, đi kèm với ước lượng tham số $\hat{\beta}_k$ là thống kê $t_k = \dfrac{\hat{\beta}_k - \beta_k}{\sqrt{s^2/S_{kk}}} \sim t(N-K)$, [tuân theo phân bố *t-student* với (*N-K*) bậc tự do].

Lưu ý là ở mẫu số của thống kê t_k, độ lớn của $s^2 = \dfrac{1}{N-K}\sum_n e_n^2 = \dfrac{ESS}{N-K}$ sẽ có ảnh hưởng trực tiếp tới giá trị của thống kê t_k. Việc tăng thêm số biến giải thích (*K* tăng) sẽ làm số bậc tự do (*N-K*) giảm, tức là làm s^2 có xu hướng bị đẩy lên. Ước lượng do vậy trở nên kém chính xác, vì sai số của ước lượng: $se(\hat{\beta}_k) = \sqrt{s^2 / S_{kk}}$ bị tăng lên. Hệ quả là, giá trị thống kê t_k sẽ trở nên nhỏ đi. Do đó, t_k dễ bị rơi vào vùng không bác bỏ giả thuyết ($DNRH_0$). Và ta dễ bị mắc phải sai lầm là chấp nhận một giả thuyết sai, mà đáng ra ta cần phải bác bỏ nó.

Nhìn chung, việc thêm biến giải thích vào mô hình có cái lợi là làm giảm tổng bình phương sai số, hay phần chưa được giải thích bởi mô hình, *ESS*. Nhưng cái thiệt là nó cũng làm giảm bậc tự do *(N-K)* [tức là làm cho việc phân tích có độ chính xác kém đi, như vừa nêu ở trên]. Nói một cách ẩn dụ, với việc đưa thêm các yếu tố mới vào mô hình, ta sẽ có cái nhìn đầy đủ hơn về mọi chi tiết, nhưng với cái giá là bức tranh không có điểm nhấn (thiếu *focus*). Chính vì vậy, thay vì sử dụng R^2, người ta thường dùng hệ số hiệu chỉnh của nó: $\bar{R}^2 = 1 - \dfrac{ESS/(N-K)}{TSS/(N-1)}$. Việc hiệu chỉnh như vậy là để tránh khuynh hướng đưa quá nhiều biến giải thích không cần thiết vào mô hình. Cụ thể là, nếu việc đưa thêm biến giải thích **có ý nghĩa** vào mô hình, thì phần lợi [tức là làm giảm *ESS*] phải vượt quá phần thiệt [tức là làm giảm bậc tự do *(N-K)*]. Khi đó, $\bar{R}^2$ tăng lên, thể hiện rằng đó là việc nên làm. Trong hoàn cảnh ngược lại, lợi không đủ bù phần mất mát, thì $\bar{R}^2$ bị giảm xuống, thể hiện rằng ta không nên đưa thêm biến giải thích đó vào mô hình, vì đấy là việc làm ít có ý nghĩa.

Ví dụ, trong bảng 5.2, dòng thứ 3, sau vạch ngang thứ nhất, ta thấy việc đưa thêm biến giải thích là số phòng ngủ và số nhà tắm vào làm giảm $\bar{R}^2$. Theo tiêu chuẩn này, mô hình tốt nhất sẽ là mô hình A: chỉ có duy nhất biến diện tích căn hộ (*SQFT*) là có ý nghĩa giải thích cho giá cả của căn hộ đó.

Người ta có thể chỉ ra rằng $\bar{R}^2$ không phạt đủ nặng việc đưa thêm các biến giải thích ít có ý nghĩa vào mô hình. Vì vậy, bên cạnh tiêu chuẩn đó, người ta còn sử dụng một số đánh giá khác, chẳng hạn như $AIC = \left(\dfrac{ESS}{N}\right)e^{2K/N}$ và $SCHWARZ = \left(\dfrac{ESS}{N}\right)N^{K/N}$. Nhìn chung, khi biến giải thích không có ý nghĩa được đưa vào mô hình, thì các tiêu chuẩn này bị đẩy lên. Vì vậy, mô hình lý tưởng nhất là mô hình có $\bar{R}^2$ cao hơn, và các tiêu chuẩn *AIC* và

SCHWARZ thấp hơn so với những mô hình khác. Ví dụ, trong bảng 5.2, mô hình A là tốt nhất theo mọi tiêu chuẩn đánh giá, bao gồm cả $\bar{R}^2$, *AIC* và *SCHWARZ*.

Trên thực tế, không phải bao giờ ta cũng may mắn như vậy. Rất có thể ta thấy một mô hình tốt hơn các cái còn lại về tiêu chuẩn này, nhưng lại tồi hơn về tiêu chuẩn khác. Khi đó, mô hình có nhiều tiêu chuẩn tốt nhất sẽ được lựa chọn.

5.3 Kiểm định các giả thuyết thống kê

Nhận xét vừa nêu cho thấy, việc chọn ra mô hình tốt nhất không phải lúc nào cũng thuyết phục cho lắm, nếu các tiêu chuẩn $\bar{R}^2$, *AIC* và *SCHWARZ* không đồng thời chỉ ra đâu là mô hình ưu việt nhất. Chính vì vậy, ta cần phải kiểm định lại xem quyết định của chúng ta có phù hợp về mặt thống kê hay không. Chẳng hạn, việc chọn mô hình A thay vì mô hình B hàm ý rằng, ta đã coi giả thuyết $H_0 : \beta_3 = 0$ là đúng. Trong khi việc loại bỏ mô hình C lại bao hàm rằng, ta coi giả thuyết đồng thời: $H_0 : \beta_3 = \beta_4 = 0$ là đúng. Việc kiểm định mức độ có ý nghĩa của từng tham số mô hình, như đã đề cập, được tiến hành bởi t-test. Trong khi đó, việc kiểm định giả thuyết đồng thời lại được thực hiện bởi Wald-test, như sẽ chỉ ra dưới đây.

Trong chương 4, chúng ta đã nói rằng, kiểm định t-test về mức độ có ý nghĩa của tham số ước lượng có thể được làm đơn giản bởi việc sử dụng *p-value*. Trong mô hình phân tích ở đây, ta thấy, chỉ có hệ số hồi quy của *SQFT* là **có ý nghĩa** giải thích trong cả 3 mô hình. Trong khi *p-value* của *BEDRMS* và *BATHS* trong cả hai mô hình B và C đều quá cao: $p - value > 0.05$. Tức là các hệ số hồi quy đi kèm với các biến giải thích này là **không có ý nghĩa** ở mức $\lambda = 5\%$. Vì vậy, xét một cách riêng biệt, ta nên loại từng biến này ra khỏi mô hình. Nhưng liệu ta có nên loại cả hai biến đó ra, và chỉ giữ lại duy nhất biến giải thích là diện tích căn hộ (*SQFT*) hay không? Điều đó dẫn đến vấn đề kiểm định giả dưới đây.

Việc loại bỏ cùng một lúc hai biến *BEDRMS* và *BATHS* là tương đương với việc chấp nhận giả thuyết đồng thời: $H_0 : \beta_3 = \beta_4 = 0$. Ta muốn nhấn mạnh rằng, giả thuyết đó là hoàn toàn **khác** với việc, cùng một lúc, xảy ra hai giả thuyết riêng biệt: $H_0 : \beta_3 = 0$ và $H_0 : \beta_4 = 0$. Ví dụ, nếu xét riêng biệt, từng yếu tố nhỏ như mầu vôi, cách bố cục căn bếp, nhà tắm, vân vân, có thể là không có ý nghĩa quyết định tới sự sẵn lòng chi trả của người đi mua nhà. Nhưng một cách **đồng thời**, chúng vẫn có thể ảnh hưởng tới cái giá mà người mua sẵn lòng bỏ ra. Nói khác đi, từng giả thuyết riêng biệt đúng, không có nghĩa là giả thuyết đồng thời cũng đúng.

Bây giờ, ta hãy xét xem làm thế nào để kiểm định giả thuyết đồng thời $H_0 : \beta_3 = \beta_4 = 0$. Hãy nhìn lại hai mô hình sau:

$$\text{(U):} \quad PRICE = \beta_1 + \beta_2 SQFT + \beta_3 BEDRMS + \beta_4 BATHS + \varepsilon \qquad (5.1)$$

(R): $PRICE = \beta_1 + \beta_2 SQFT + \varepsilon$ (5.2)

Mô hình (U) [tức là mô hình C trong bảng 5.2] được gọi là **mô hình không bị ràng buộc** (*unrestricted model*). Mô hình (R) [tức là mô hình A trong bảng 5.2] được gọi là **mô hình bị ràng buộc** (*restricted model*). Sở dĩ như vậy là vì mô hình (R) chính là mô hình (U), nhưng chịu ràng buộc là $H_0 : \beta_3 = \beta_4 = 0$. Việc lựa chọn xem mô hình nào là đúng, về thực chất quy về việc kiểm định giả thuyết kép sau:

$$H_0 : \begin{cases} \beta_2 = 0 \\ \beta_3 = 0 \end{cases} \quad \text{.vs.} \quad H_1 : \text{không phải là } H_0$$

Chúng ta đã nhận xét rằng, việc đưa thêm biến giải thích vào mô hình luôn làm tăng mức độ giải thích của mô hình, tức là làm giảm tổng bình phương sai số *ESS*. Vì vậy, ta luôn có: $ESS_R > ESS_U$. Trong đó, ESS_R là tổng bình phương sai số ước lượng của mô hình (R), và ESS_U là của mô hình (U).

Chúng ta nhận xét thêm rằng, nếu các biến *BEDRMS* và *BATHS* là không có ý nghĩa lắm cho việc giải thích giá căn hộ (*PRICE*), thì việc đưa chúng vào mô hình sẽ làm tổng bình phương sai số ước lượng giảm đi, nhưng **không nhiều**. Nói khác đi, nếu giả thuyết H_0 là đúng, thì hiệu $(ESS_R - ESS_U)$ là dương, nhưng với độ lớn không đáng kể. Ngược lại, nếu H_0 là sai, thì việc đưa thêm các biến *BEDRMS* và *BATHS* sẽ cải thiện đáng kể mức độ giải thích của mô hình. Do vậy, độ lệch $(ESS_R - ESS_U)$ sẽ rất lớn. Như vậy, chúng ta có thể đi đến nhận định rằng, khi hiệu $(ESS_R - ESS_U)$ là lớn, thì ta sẽ bác bỏ giả thuyết H_0 (RH_0). Tuy nhiên, như thế nào thì hiệu $(ESS_R - ESS_U)$ được coi là lớn? Điều đó dẫn đến việc lập thống kê *F*, mà ta sẽ trình bày dưới dạng tổng quát như sau. Xét hai lựa chọn về mô hình khác nhau:

(U): $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + \varepsilon$ (5.3)

(R): $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_{K-J} X_{K-J} + \varepsilon$ (5.4)

Mô hình (5.4) chính là mô hình (5.3), với *J* ràng buộc: $\beta_{K-J+1} = \beta_{K-J+2} = ... = \beta_K = 0$. Nói khác đi, ta muốn kiểm định giả thuyết đồng thời sau:

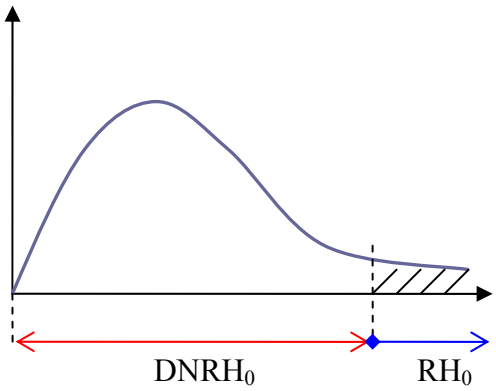
$$H_0 : \begin{cases} \beta_{K-J+1} = 0 \\ \beta_{K-J+2} = 0 \\ ..... \\ \beta_K = 0 \end{cases} \quad \text{.vs.} \quad H_1 : \text{không phải là } H_0$$

Người ta có thể chứng minh được rằng, đại lượng sau có phân bố F với J và $(N\text{-}K)$ bậc tự do:

$$F_c = \frac{(ESS_R - ESS_U)/J}{ESS_U/(N-K)} \sim F(J, N-K) \qquad (5.5)$$

Từ lập luận nêu trên, nếu F_c lớn hơn giá trị F-tra bảng: $F_c > F_\lambda(J, N-K)$, khi đó ta bác bỏ giả thuyết (RH_0). Ngược lại, nếu $F_c < F_\lambda(J, N-K)$ thì ta sẽ không bác bỏ giả thuyết đó $(DNRH_0)$.

Đồ thị 5.1: kiểm định giả thuyết với F-test.



Ví dụ 5.1 (tiếp theo): trong ví dụ về giá căn hộ, với việc chọn giữa mô hình (5.1) và (5.2), ta có $ESS_R = 18{,}274$, $ESS_U = 16{,}700$ [xem bảng 5.2], $J = 2, (N-K) = 10$. Vì vậy:

$$F_c = \frac{(18{,}274 - 16{,}700)/2}{16{,}700/10} = 0.471$$

Ta có thể tra bảng F-statistic: $F_{0.05}(2,10) = 4.1$. Vì vậy, ta có: $F_c < F_{0.05}(2,10)$. Tức là ta sẽ không bác bỏ giả thuyết H_0. Khi đó, mô hình với chỉ một biến giải thích là diện tích sử dụng *(SQFT)* được coi là mô hình đúng nhất theo kiểm định Wald –test.

5.4 Kiểm định tính có ý nghĩa của cả mô hình (overall significance test)

Một trường hợp đặc biệt của Wald test (hay F-test) vừa nêu trên là đánh giá hai mô hình sau:

(U): $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + \varepsilon$ (5.6)

(R): $Y = \beta_1 + \varepsilon$ (5.7)

Trong mô hình bị ràng buộc (R), tất cả các biến giải thích, ngoại trừ hằng số (constant term), bị loại bỏ. Tức là chúng ta muốn kiểm định giả thuyết H_0:

$H_0 : \beta_2 = \beta_3 = ... = \beta_K = 0$ .vs. H_1 :không phải là H_0

Nói khác đi, ta muốn kiểm tra nhận định là: "không có bất cứ một biến giải thích nào trong mô hình, ngoại trừ *constant term*, là có ý nghĩa cả". Wald-test cho kiểm định như vậy có dạng đơn giản như sau:

$$F_c = \frac{R^2 / (K-1)}{(1-R^2)/(N-K)} \sim F(K-1, N-K)$$

Trong đó, R^2 là độ phù hợp của mô hình (5.6).

Nếu ta **không** bác bỏ giả thuyết H_0, thì không có biến giải thích nào, ngoại trừ *constant term* trong (5.6) là có ý nghĩa cả. Chúng ta có một mô hình tồi và cần phải xây dựng lại mô hình hồi quy.

Thông thường các *sofware* như eviews sẽ cho ra thông báo về việc kiểm định giả thuyết về tính có ý nghĩa chung của cả mô hình (*overall significance*). Giá trị của F_c, tính theo công thức (5.5), lúc này được gọi là *F-stat*. Đi kèm theo nó, eviews cũng cho ra *p-value* của *F-stat*.

Ví dụ 5.1 (tiếp theo): ứng với mô hình (C), máy tính sẽ kiểm định giả thuyết: $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$. Và cho ra thông báo *F-stat* = 16.98, [*p-value* =0.000]. Nhìn vào bảng 5.2, các giá trị của *F-stat* cho mô hình (A) và (B) lần lượt là 54.86 và 27.7. Tức là trong cả 3 mô hình, một cách đồng thời, các biến là có ý nghĩa cho việc giải thích những biến động của giá căn hộ *PRICE*.

5.5 Những ứng dụng khác của Wald test

Ứng dụng của Wald test là khá rộng và đa dạng hơn nhiều so với những ví dụ đã nêu ở trên. Nhưng nhìn chung, chúng có cùng chung một cách tiếp cận là so sánh độ tốt về mặt thống kê giữa hai dạng mô hình: bị ràng buộc và không bị ràng buộc. Chúng ta xem lại một số cải biên của ví dụ đơn giản về nhu cầu đầu tư ở Mỹ (1968 -82):

(U): $$INV = \beta_1 + \beta_2 T + \beta_3 G + \beta_4 INT + \beta_5 INF + \varepsilon \quad (5.8)$$

Mô hình này giả định rằng các nhà đầu tư nhậy cảm với lãi suất (*INT)* và lạm phát (*INF)*. Một giả định khác là các nhà đầu tư chỉ nhậy cảm với lãi suất thực. Mô hình biểu diễn sẽ là:

(R): $$INV = \beta_1 + \beta_2 T + \beta_3 G + \beta_4 (INT - INF) + \varepsilon \quad (5.9)$$

Chúng ta nhận xét rằng mô hình (5.9) là bị ràng buộc (restricted) so với mô hình (5.8) bởi giả định là: $H_0 : \beta_4 + \beta_5 = 0$. Hay cũng vậy, ta kiểm định:

$$H_0 : \beta_4 = -\beta_5 \quad \text{.vs.} \quad H_1 : \text{không phải là } H_0 \quad (5.10)$$

Các bước tiến hành kiểm định (Wald test) là như sau:

Bước 1: Xác định rõ đâu là mô hình bị ràng buộc (restricted model: R) , bằng cách nhận dạng yêu cầu cần kiểm định là gì, hay cũng vậy, giả thuyết H_0 bao gồm những ràng buộc gì.

Bước 2: Tiến hành chạy hồi quy mô hình không bị ràng buộc (U) và mô hình bị ràng buộc (R).

Bước 3: Tính thống kê F_c, sử dụng phương trình (5.5), với các bậc tự do *J* [là số các ràng buộc nêu bởi H_0] và (*N-K*).

Bước 4: Từ bảng thống kê *F*, tìm giá trị F-tra bảng [tức là tìm critical value: $F(J, N-K)_\lambda$]. Một cách khác nữa, ta có thể tính $p - value = \Pr ob(F(J, N-K) > F_c)$

Bước 5: Loại bỏ giả thuyết (RH_0) , nếu $F_c > F(J, N-K)_\lambda$, hoặc $p - value < \lambda$.

5.6 Lỗi lầm trong việc lập mô hình (Specification errors)

Chúng ta đã nêu lên dạng tổng quát của mô hình hồi quy như sau:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + \varepsilon$$

Tuy nhiên, người lập mô hình có thể phạm phải rất nhiều loại sai lầm trong việc xác định dạng mô hình cụ thể. Có lẽ hai loại lỗi phổ biến nhất là: bỏ qua những biến có ý nghĩa, không đưa chúng vào mô hình; và ngược lại, đưa quá nhiều biến giải thích không có ý nghĩa vào mô hình. Lỗi sau cùng đã được bàn ở trên. Nó làm ước lượng trở nên mất chính xác. Mặc dù ước lượng vẫn là không chệch. Việc né tránh lỗi lầm đó, như đã nói, được thực hiện dựa trên xem xét các chỉ tiêu đo lường $\bar{R}^2$, *AIC*, và *SCHWARZ*, cũng như sử dụng các kiểm định F-test và t-test. Đối với lỗi lầm thứ nhất, việc phát hiện trở nên khó khăn hơn.

Chủ yếu là do người lập mô hình thường sử dụng cách tiếp cận máy móc, học từ sách giáo khoa (text-book approach), mà không có những phân tích thấu đáo về đối tượng nghiên cứu. Như trong ví dụ về đầu tư của Mỹ (1968-82), nếu chỉ dựa vào sách giáo khoa, người lập mô hình có thể sẽ bỏ quên, không đưa biến xu thế vào phân tích. Ở đây ta muốn nêu lên hậu quả khá tai hại của cách tiếp cận text-book đó là như thế nào.

Để đơn giản, chúng ta giả sử mô hình **đúng** là như sau:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (5.11)$$

Nhưng chúng ta phạm **sai lầm**, và chạy hồi quy mô hình sau:

$$Y = \beta_1 X_1 + \tilde{\varepsilon} \qquad (5.12)$$

Trong (5.12), ta bỏ quên X_2, nên về thực chất, hay về mặt tổng thể, $\tilde{\varepsilon} = \beta_2 X_2 + \varepsilon$.

Bây giờ, do sự lầm tưởng, ta chạy hồi quy (5.12), thay vì chạy mô hình đúng (5.11). Sử dụng phương trình (3.9), định lý 3.1, ta có ước lượng sau:

$$\hat{\beta}_1 = \beta_1 + \sum c_{1n} \tilde{\varepsilon}_n \qquad (5.13)$$

Thế giá trị $\tilde{\varepsilon} = \beta_2 X_2 + \varepsilon$ vào (5.13), và lấy kỳ vọng cả hai vế, ta có:

$$E\hat{\beta}_1 = E[\beta_1 + \sum c_{1n} \tilde{\varepsilon}_n]$$

$$= E[\beta_1 + \sum c_{1n}(\beta_2 x_{n2} + \varepsilon_n)]$$
$$= \beta_1 + \beta_2 \sum c_{1n} x_{n2} + \sum c_{n1} E\varepsilon_n$$

Trong đó, x_{n2} là quan sát thứ *n* của biến giải thích X_2. Sử dụng giả thuyết A1, ta có:

$$E\hat{\beta}_1 = \beta_1 + \beta_2 \sum c_{1n} x_{n2} \neq \beta_1 \qquad (5.14)$$

Phương trình (5.14) nói lên rằng, nhìn chung, việc bỏ quên biến giải thích có ý nghĩa sẽ làm ước lượng **bị chệch (**biased estimation**)**. Vì vậy, mọi kiểm định thống kê trở nên vô giá trị, và việc dự báo trở nên vô nghĩa. Đây có lẽ là lời cảnh tỉnh nghiêm khắc nhất với những nghiên cứu máy móc, dựa trên *text books*. Để tránh tình huống này, việc đánh giá thực tiễn kỹ lưỡng trước khi lập mô hình là một việc làm không thể thiếu.