THỐNG KÊ TOÁN

1 Mẫu ngẫu nhiên và phân bố mẫu

Xét một mẫu ngẫu nhiên

$$(X_1, X_2, ..., X_n)$$

tương ứng với đại lượng ngẫu nhiên X

$$E(X) = m, \quad D(X) = \sigma^2.$$

Gọi ξ là đại lượng ngẫu nhiên:

$$P(\xi = x_i) = \frac{1}{n} \quad \text{với mọi} \quad i = 1, 2, ..., n.$$

Khi đó $E(\xi), D(\xi)$ được gọi là các *đặc trưng mẫu*. Người ta kí hiệu $\overline{X} = E(\xi)$ là *kì vọng mẫu* và $S^2 = D(\xi)$ là *phương sai mẫu*. Hiển nhiên

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

và

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = m, D(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{\sigma^2}{n}.$$

Để tính kì vọng của phương sai mẫu, ta sử dụng

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

Suy ra

$$E(S^2) = \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(m^2 + \sigma^2) - \left(m^2 + \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2.$$

Kí hiệu

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Khi đó

$$E(S^{*2}) = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2.$$

S^{*2} được gọi là là *phương sai mẫu điều chỉnh*.

$$E(\overline{X}) = m = E(X), E(S^{*2}) = \sigma^2 = D(X),$$

Nhận xét 4

1. *$\overline{X}$ không những hội tụ theo xác suất mà hội tụ hầu chắc chắn tới $m = E(X)$.*
2. *S^2, S^{*2} hội tụ hầu chắc chắn (suy ra cũng hội tụ theo xác suất) tới σ^2 khi $n \to \infty$.*

2 Các hàm phân bố thường gặp trong thống kê

Hàm Gamma, Beta và tính chất hàm Gamma, Beta

A. Tích phân sau hội tụ với mọi $x > 0, y > 0$

$$\Gamma(x) = \int_0^{+\infty} e^{-t}t^{x-1}dt, \quad B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt.$$

Tách $\Gamma(x)$ thành hai tích phân

$$\Gamma(x) = \int_0^{+\infty} e^{-t}t^{x-1}dt = \int_0^1 e^{-t}t^{x-1}dt + \int_1^{+\infty} e^{-t}t^{x-1}dt = I_1 + I_2.$$

Tích phân I_1 hội tụ vì với $0 < x < 1, 0 < t \leqslant 1$, ta có $e^{-t}t^{x-1} < \frac{1}{t^{1-x}}$.
Tích phân I_2 hội tụ vì $\lim_{t\to+\infty} e^{-t}t^{x+1} = 0$, suy ra với t đủ lớn $e^{-t}t^{x-1} < \frac{1}{t^2}$.

B. Tích phân sau hội tụ với mọi $x > 0, y > 0$.

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt.$$

Tách $\Gamma(x)$ thành hai tích phân

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = \int_0^c t^{x-1}(1-t)^{y-1}dt + \int_c^1 t^{x-1}(1-t)^{y-1}dt.$$

1. $\Gamma(1) = 1$.
2. $\Gamma(x+1) = x\Gamma(x)$. Thật vậy với $x > 0$, xét

$$\Gamma(x+1) = \int_0^{+\infty} e^{-t}t^x dt = -\int_0^{+\infty} t^x de^{-t} = -t^x e^{-t}|_0^{+\infty} + \int_0^{+\infty} xt^{x-1}e^{-t}dt = x\Gamma(x)$$

3. $\lim_{x\to 0+}\Gamma(x) = \lim_{x\to 0+}\frac{\Gamma(x+1)}{x} = +\infty$.
4. Với $x - k > 0$, k là số tự nhiên bất kì

$$\Gamma(x) = (x-1)(x-2)\cdots(x-k)\Gamma(x-k) \Rightarrow \text{suy ra } \Gamma(n) = (n-1)!$$

5. Chú ý rằng $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, suy ra

$$\Gamma(n+\frac{1}{2}) = \frac{1\cdot 3\cdots(2n-1)}{2^n}\sqrt{\pi} = \frac{(2n-1)!!}{2^n}\sqrt{\pi}$$

6. Ta công nhận kết quả sau đúng với mọi số thực $x > 0, y > 0$

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Phân bố Gamma, Beta

1. Nếu $X_i \in N(m_i, \sigma_i^2), \quad i = 1, 2, ..., n$ độc lập, khi đó trung bình mẫu

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \in N(m, \sigma^2)$$

trong đó

$$m = \frac{m_1 + m_2 + \cdots + m_n}{n}, \quad \sigma_i^2 = \frac{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}{n}.$$

2. Phân bố của $Y = X^2$ với $X \in N(m, \sigma^2)$. Hàm mật độ của Y

$$g(y) = (2\sigma\sqrt{2\pi y})^{-1} e^{\frac{-(y+m^2)}{2\sigma^2}} \left(e^{m\frac{\sqrt{y}}{\sigma^2}} + e^{-m\frac{\sqrt{y}}{\sigma^2}}\right).$$

Nếu $m = 0$

$$g(y) = \frac{1}{2\sigma\sqrt{2\pi}} e^{\frac{-y}{2\sigma^2}} y^{\frac{-1}{2}}.$$

Phân bố của $Y = X^2$ là trường hợp đặc biệt của phân bố Gamma: $G(y, \alpha, p) = const \cdot e^{-\alpha y} y^{p-1}$.

3. Phân bố Gamma là phân bố có hàm mật độ

$$G(x, \alpha, p) = \frac{\alpha^p}{\Gamma(p)} \cdot e^{-\alpha x} x^{p-1}, \quad \alpha > 0, p > 0, x > 0.$$

Mô men cấp k của phân bố Gamma

$$m_k = \int_0^{+\infty} x^k \frac{\alpha^p}{\Gamma(p)} \cdot e^{-\alpha x} x^{p-1} dx = \int_0^{+\infty} \frac{\alpha^p}{\Gamma(p)} \cdot e^{-\alpha x} x^{k+p-1} dx = \frac{\Gamma(p+k)}{\alpha^k \Gamma(p)}.$$

Vì vậy kì vọng và phương sai của phân bố Gamma lần lượt bằng

$$m = \frac{p}{\alpha}, \quad \sigma^2 = m_2 - m_1^2 = \frac{\Gamma(p+2)}{\alpha^2 \Gamma(p)} - \frac{p^2}{\alpha^2} = \frac{p}{\alpha^2}. \tag{1}$$

Bài tập Giả sử X phân bố đều trên đoạn $[0, 1]$. Chứng minh rằng $Y = -\ln X$ có phân bố Gamma với các tham số $\alpha = 1, p = 1$.

4. Phân bố Beta là phân bố có hàm mật độ

$$B(x, \alpha, \beta) = [B(\alpha, \beta)]^{-1} \cdot x^{\alpha-1}(1-x)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1.$$

Đặc biệt $B(x, 1, 1) = x$ là hàm mật độ của phân bố đều trên đoạn $[0, 1]$.

Bài tập 1. Hãy tính các mô men cấp k của phân bố Beta. ($\frac{B(\alpha+k,\beta)}{B(\alpha,\beta)}$).

Từ đó suy ra kì vọng và phương sai của nó. ($m = \frac{\alpha}{\alpha+\beta}, \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$).

Bài tập 2. Giả sử X và Y độc lập có phân bố Beta với các tham số (α_1, β_1) và (α_2, β_2) tương ứng. Chứng minh rằng XY cũng có có phân bố Beta với các tham số $(\alpha_2, \beta_1 + \beta_2)$, nếu $\alpha_1 = \alpha_2 + \beta_2$.

Hướng dẫn: Xét phép biến đổi $u = xy, v = x$. Khi đó Jacôbiên bằng $\frac{1}{v}$. Tích phân hàm mật độ chung của (U, V) theo v từ u đến 1 ta được mật độ của XY.

Bài tập 3. Giả sử $X \in G(\alpha_1, 1)$ và $Y \in G(\alpha_2, 1)$ độc lập có phân bố Gamma. Khi đó $u = \frac{X}{X+Y}$ có phân bố Beta với các tham số (α_1, α_2).

Hướng dẫn: Xét phép biến đổi $u = \frac{x}{x+y}, v = y$. Tích phân hàm mật độ chung theo v từ 0 đến ∞.

Định lí 9 *Nếu $X \in G(\alpha, p_1), Y \in G(\alpha, p_2)$ độc lập, khi đó $r = X + Y$ và $f = \frac{X}{Y}$ cũng độc lập. Ngoài ra $r \in G(\alpha, p_1 + p_2)$ và hàm mật độ của f bằng*

$$\frac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)} \cdot \frac{f^{p_1-1}}{(1+f)^{p_1+p_2}}.$$

Chứng minh. Hàm mật độ của (X, Y) bằng

$$c \cdot e^{-\alpha x - \alpha y} x^{p_1-1} y^{p_2-1}.$$

Đổi biến $x = r\sin^2\varphi, y = r\cos^2\varphi, \quad 0 < r < +\infty, 0 < \varphi < \frac{\pi}{2}$, khi đó Jacobien của (x, y) bằng $J(r, \varphi) = r\sin 2\varphi$. Mật độ của (r, φ) bằng

$$c' \cdot e^{-\alpha r} r^{p_1+p_2-1} (\sin\varphi)^{2p_1-1} (\cos\varphi)^{2p_2-1}, \tag{2}$$

điều đó chứng tỏ r và φ độc lập. Suy ra $r = X + Y$ và $f = \frac{X}{Y} = tg^2\varphi$ cũng độc lập. Từ biểu thức (2) hiển nhiên $r \in G(\alpha, p_1 + p_2)$.

Để xác định hàm mật độ của f, ta sử dụng phép đổi biến $\varphi = arctg\sqrt{f}$, ta thu được kết quả

$$\frac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)} \cdot \frac{f^{p_1-1}}{(1+f)^{p_1+p_2}}.$$

Chú ý rằng với phép biến đổi $u = \frac{1}{1+f}$, khi đó $\int_0^1 u^{p_2-1}(1-u)^{p_1-1} du = \int_0^\infty \frac{f^{p_1-1}}{(1+f)^{p_1+p_2}} df$.

1. **Phân bố χ^2.**

Nếu $X_i \in N(0,1), \quad i = 1, 2, ..., n$ độc lập, khi đó phân bố của $X_1^2 + X_2^2 + \cdots + X_n^2$ được gọi là phân bố χ^2 với n bậc tự do. Người ta thường kí hiệu $\chi^2(n)$ là lớp các đại lượng ngẫu nhiên có phân bố χ^2 với n bậc tự do. Đây là trường hợp đặc biệt của phân bố Gamma ($\alpha = \frac{1}{2}, p = \frac{n}{2}$) với hàm mật độ

$$G(x, \frac{1}{2}, \frac{n}{2}) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \cdot e^{-\frac{x}{2}} x^{\frac{n}{2}-1}, \quad x > 0.$$

Do đẳng thức (1), kì vọng và phương sai của phân bố $\chi^2(n)$ lần lượt bằng

$$m = n, \quad \sigma^2 = 2n.$$

2. **Phân bố F.**

Nếu $X_1 \in \chi^2(m), X_2 \in \chi^2(n)$ độc lập, khi đó phân bố của

$$F = \frac{\frac{1}{m}X_1}{\frac{1}{n}X_2}$$

được gọi là phân bố F với (m, n) bậc tự do.

Mật độ của $\frac{X_1}{X_2}$ bằng

$$\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{f^{\frac{m}{2}-1}}{(1+f)^{\frac{m+n}{2}}}.$$

Mật độ của phân bố F với (m, n) bậc tự do bằng

$$\left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{x^{\frac{m}{2}-1}}{(1+\frac{mx}{n})^{\frac{m+n}{2}}}.$$

3. **Phân bố Student (hay còn gọi là phân bố t).**

Nếu $X \in \chi^2(n)$ và $Y \in N(0,1)$ độc lập, khi đó phân bố của

$$T = \frac{Y}{\sqrt{X}}\sqrt{n}$$

được gọi là phân bố T (hay phân bố Student) với n bậc tự do. Phân bố đồng thời của (Y, X) bằng

$$c \cdot e^{-\frac{y^2}{2}} e^{-\frac{x}{2}} x^{\frac{n}{2}-1}.$$

Đổi biến $y = r\sin\varphi, x = r^2\cos^2\varphi, \quad 0 < r < +\infty, -\frac{\pi}{2} < \varphi < \frac{\pi}{2}$, khi đó Jacobien của (x, y) bằng $J(r, \varphi) = 2r^2\cos\varphi$. Mật độ của (r, φ) bằng

$$c' \cdot e^{-\frac{r^2}{2}} r^n (\cos\varphi)^{n-1},$$

điều đó chứng tỏ r và φ độc lập. *Chú ý rằng hệ số c của* $c(\cos\varphi)^{n-1}$ *bằng* $c = [B(\frac{1}{2}, \frac{n}{2})]^{-1}$. Để xác định hàm mật độ của T, ta sử dụng phép đổi biến

$$t = \frac{\sqrt{n}y}{\sqrt{x}} = \sqrt{n}tg\varphi \quad \text{hay} \quad \varphi = arctg\frac{t}{\sqrt{n}},$$

ta được hàm mật độ của phân bố T với n bậc tự do

$$S(t, n) = \left[\sqrt{n}B\left(\frac{1}{2}, \frac{n}{2}\right)\right]^{-1}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

Nếu $\frac{X}{\sigma^2} \in \chi^2(n)$ và $Y \in N(m, \sigma^2)$ độc lập, khi đó

$$T = \frac{Y - m}{\sqrt{X}}\sqrt{n}$$

có phân bố Student với n bậc tự do.

Kí hiệu $S(n)$ là lớp các đại lượng ngẫu nhiên có phân bố Student với n bậc tự do.

4. **Phân bố của trung bình mẫu và phương sai mẫu.**

Nếu $X_i \in N(m, \sigma^2), i = 1, 2, ..., n$ độc lập, khi đó

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \in N\left(m, \frac{\sigma^2}{n}\right) \quad \text{và} \quad \frac{n}{\sigma^2}S^2 = \frac{n-1}{\sigma^2}S^{*2} \in \chi^2(n-1).$$

Thật vậy, kí hiệu $\mathbf{X} = (X_1, ..., X_n)^T$ và xét phép biến đổi trực giao $\mathbf{Y} = \mathbf{AX}$ với $(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}})$ là hàng thứ nhất của $\mathbf{A}$. Khi đó

(a) $$Y_1 = \overline{X}\sqrt{n}$$

(b) $Y_1^2 + \cdots + Y_n^2 = X_1^2 + \cdots + X_n^2 = \sum(X_i - \overline{X})^2 + n\overline{X}^2 \Leftrightarrow Y_2^2 + \cdots + Y_n^2 = (n-1)S^{*2}$

(c) Với véc tơ $\mathbf{m} = (m, m, ..., m)$, ta có $\mathbf{A}(\mathbf{X} - \mathbf{m}) = \mathbf{Y} - (m\sqrt{n}, 0, ..., 0) = (Y_1 - m\sqrt{n}, Y_2, ..., Y_n)$. Suy ra

$$(Y_1 - m\sqrt{n})^2 + Y_2 + \cdots + Y_n^2 = (X_1 - m)^2 + (X_2 - m)^2 + \cdots + (X_n - m)^2.$$

Biết hàm mật độ của $\mathbf{X}$ bằng

$$c \cdot e^{-\frac{\sum(x_i - m)^2}{2\sigma^2}}.$$

Vậy mật độ của $\mathbf{Y}$ bằng

$$c \cdot e^{-\frac{(y_1 - m\sqrt{n})^2 + y_2 + \cdots + y_n^2}{2\sigma^2}}.$$

Điều đó chứng tỏ $Y_1 = \overline{X}\sqrt{n} \in N(m\sqrt{n}, \sigma^2), Y_i \in N(0, \sigma^2), i = 2, ..., n$ độc lập và

$$\frac{(n-1)S^{*2}}{\sigma^2} = \frac{Y_2^2 + \cdots + Y_n^2}{\sigma^2} \in \chi^2(n-1).$$

Bây giờ ta suy ra hệ quả quan trọng: T có phân bố Student với $n-1$ bậc tự do, với

$$T = \frac{\overline{X} - m}{S^*}\sqrt{n} = \frac{\overline{X} - m}{S}\sqrt{n-1}.$$

Thật vậy T bằng thương của 2 đại lượng ngẫu nhiên

$$T = \sqrt{n-1}\frac{\overline{X} - m}{\sigma}\sqrt{n} : \frac{S\sqrt{n}}{\sigma}$$

trong đó $\frac{\overline{X} - m}{\sigma}\sqrt{n} \in N(0, 1)$ và $\frac{nS^2}{\sigma^2} = \frac{(n-1)S^{*2}}{\sigma^2} \in \chi^2(n-1)$.

3 Khoảng tin cậy cho giá trị trung bình

(a) Mẫu có phân bố chuẩn với phương sai σ^2 đã cho. Khoảng tin cậy cho giá trị trung bình, với độ tin cậy $1-\alpha$

$$\overline{X} - u_\alpha \frac{\sigma}{\sqrt{n}} < m < \overline{X} + u_\alpha \frac{\sigma}{\sqrt{n}},$$

trong đó u_α được xác định từ hệ thức $P(|u| \geq u_\alpha) = \alpha, \quad u \in N(0,1)$.

(b) Mẫu có phân bố chuẩn với phương sai chưa biết. Khoảng tin cậy cho giá trị trung bình, với độ tin cậy $1-\alpha$

$$\overline{X} - t_\alpha \frac{S^*}{\sqrt{n}} < m < \overline{X} + t_\alpha \frac{S^*}{\sqrt{n}},$$

trong đó t_α được xác định từ hệ thức $P(|t| \geq t_\alpha) = \alpha$

(t có phân bố Student với $n-1$ bậc tự do.)

Nếu kích thước mẫu đủ lớn ($n \geq 30$), mặc dù phân bố mẫu có thể không là phân bố chuẩn, tuy nhiên áp dụng luật giới hạn trung tâm ta có thể sử dụng công thức sau để tính khoảng tin cậy cho giá trị trung bình, độ tin cậy $1-\alpha$

$$\overline{X} - u_\alpha \frac{S^*}{\sqrt{n}} < m < \overline{X} + u_\alpha \frac{S^*}{\sqrt{n}},$$

trong đó u_α được xác định từ hệ thức $P(|u| \geq u_\alpha) = \alpha, \quad u \in N(0,1)$.

4 Khoảng tin cậy cho xác suất

Cho biến cố ngẫu nhiên với xác suất p cần phải ước lượng. Giả thiết $\widehat{p} = \frac{k}{n}$ là tần suất xuất hiện của biến cố đó. (Kích thước mẫu đủ lớn - thông thường $n \geq 40$). Khi đó với độ tin cậy $1-\alpha$, khoảng tin cậy cho xác suất

$$\widehat{p} - \frac{u_\alpha}{\sqrt{n}}\sqrt{\widehat{p}(1-\widehat{p})} < p < \widehat{p} + \frac{u_\alpha}{\sqrt{n}}\sqrt{\widehat{p}(1-\widehat{p})},$$

trong đó u_α được xác định từ hệ thức $P(|u| \geq u_\alpha) = \alpha, \quad u \in N(0,1)$.

5 Khoảng tin cậy cho phương sai của phân bố chuẩn

Mẫu có phân bố chuẩn với phương sai σ^2 cần phải ước lượng. Với độ tin cậy $1-\alpha$, khoảng tin cậy cho σ^2

$$\frac{nS^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

trong đó χ^2_α được xác định từ hệ thức $P(\chi^2 > \chi^2_\alpha) = \alpha$,

(χ^2là đại lượng ngẫu nhiên có phân bố χ^2 với $(n-1)$ bậc tự do).

6 Khoảng tin cậy cho hiệu các giá trị trung bình của phân bố chuẩn

6.1 Trường hợp phương sai đã biết

Gọi $(X_1, X_2, ..., X_m)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $X \in N(m_1, \sigma_1^2)$, $(Y_1, Y_2, ..., Y_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $Y \in N(m_2, \sigma_2^2)$. Các tham số m_1, m_2 chưa biết và σ_1^2, σ_2^2 là các tham số đã biết. Giả thiết tiếp các đại lượng ngẫu nhiên

$$X_1, X_2, ..., X_m, \quad Y_1, Y_2, ..., Y_n$$

độc lập nhau.

Dễ dàng nhận thấy

$$E(\overline{X} - \overline{Y}) = m_1 - m_2$$

$$D(\overline{X} - \overline{Y}) = D(\overline{X}) + D(\overline{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

Suy ra

$$u = \frac{(\overline{X} - \overline{Y}) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

có phân bố chuẩn, thuộc lớp N(0,1).

Khoảng tin cậy cho hiệu các giá trị trung bình $m_1 - m_2$ với độ tin cậy $1 - \alpha$

$$(\overline{X} - \overline{Y}) - u_\alpha \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < m_1 - m_2 < (\overline{X} - \overline{Y}) + u_\alpha \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}},$$

trong đó u_α được xác định từ hệ thức $P((|u| \geq u_\alpha) = \alpha, \quad u \in N(0,1)$.

Nếu n_1, n_2 đủ lớn (≥ 30), ta xấp xỉ công thức trên cho hiệu các giá trị trung bình $m_1 - m_2$ cả trong trường hợp các mẫu đã cho không tuân theo phân bố chuẩn, sử dụng S_1^* và S_2^* thay cho σ_1, σ_2 tương ứng trong công thức trên.

6.2 Trường hợp các phương sai chưa biết và bằng nhau

Gọi $(X_1, X_2, ..., X_m)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $X \in N(m_1, \sigma^2)$, $(Y_1, Y_2, ..., Y_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $Y \in N(m_2, \sigma^2)$. (Chúng có phương sai bằng nhau). Các tham số m_1, m_2, σ^2 chưa biết và giả thiết rằng các đại lượng ngẫu nhiên

$$X_1, X_2, ..., X_m, \quad Y_1, Y_2, ..., Y_n$$

độc lập nhau.

Dễ dàng nhận thấy

$$E(\overline{X} - \overline{Y}) = m_1 - m_2$$

$$D(\overline{X} - \overline{Y}) = D(\overline{X}) + D(\overline{Y}) = \frac{\sigma^2}{m} + \frac{\sigma^2}{n} = \left(\sigma \sqrt{\frac{m+n}{mn}}\right)^2$$

Suy ra

$$u = \frac{(\overline{X} - \overline{Y}) - (m_1 - m_2)}{\sigma \sqrt{\frac{m+n}{mn}}}$$

có phân bố chuẩn, thuộc lớp N(0,1). Dễ dàng chứng minh được

$$\frac{mS_X^2 + nS_Y^2}{m + n - 2}$$

là ước lượng không chệch của σ^2. Người ta chứng minh được rằng (thay σ^2 trong thống kê trên bằng ước lượng của nó)

$$t = \frac{(\overline{X}-\overline{Y})-(m_1-m_2)}{\sqrt{\frac{mS_X^2+nS_Y^2}{m+n-2}}\sqrt{\frac{m+n}{mn}}} \quad \left(= \sqrt{\frac{mn(m+n-2)}{m+n}} \cdot \frac{(\overline{X}-\overline{Y})-(m_1-m_2)}{\sqrt{mS_X^2+nS_Y^2}}\right)$$

có phân bố Student với $m+n-2$ bậc tự do.

Đặc biệt khi hai giá trị trung bình bằng nhau $m_1 = m_2$

$$t = \sqrt{\frac{mn(m+n-2)}{m+n}} \cdot \frac{\overline{X}-\overline{Y}}{\sqrt{mS_X^2+nS_Y^2}}$$

cũng có phân bố Student với $m+n-2$ bậc tự do.

Khoảng tin cậy cho hiệu các giá trị trung bình $m_1 - m_2$ với độ tin cậy $1-\alpha$ bằng

Mẫu $\{X_i\}_{i=1}^m \in N(m_1,\sigma^2)$ $\{Y_i\}_{i=1}^n \in N(m_2,\sigma^2)$, có phân bố chuẩn với phương sai σ^2 chưa biết. Giả thiết các phần tử mẫu đó độc lập nhau.

$$(\overline{X}-\overline{Y}) - S.t_\alpha\sqrt{\frac{m+n}{mn}} < m_1 - m_2 < (\overline{X}-\overline{Y}) + S.t_\alpha\sqrt{\frac{m+n}{mn}},$$

trong đó kí hiệu $S^2 = \dfrac{mS_X^2+nS_Y^2}{m+n-2}$ và t_α được xác định từ hệ thức

$P(|t| \geq t_\alpha) = \alpha$ (t có phân bố Student với $m+n-2$ bậc tự do.)

7 Kiểm định giả thiết về giá trị trung bình (trường hợp σ^2 đã biết)

Bài toán 1 và **quy tắc kiểm định**

Mẫu có phân bố chuẩn với phương sai σ^2 đã cho. Kiểm định giả thiết về kì vọng mẫu, mức ý nghĩa α

$$(H): m = m_0,$$

với đối thiết

$$(K): m \neq m_0.$$

Quy tắc: Bác bỏ (H) nếu $\left|\dfrac{\overline{X}-m_0}{\sigma}\sqrt{n}\right| = |u_{qs}| > u_\alpha,$

trong đó u_α được xác định từ hệ thức $P(|u| \geq u_\alpha) = \alpha, \quad u \in N(0,1).$

Bài toán 2 và **quy tắc kiểm định**

Mẫu có phân bố chuẩn với phương sai σ^2 đã cho. Kiểm định giả thiết về kì vọng mẫu, mức ý nghĩa α

$$(H): m = m_0,$$

với đối thiết

$$(K): m > m_0.$$

Quy tắc: Bác bỏ (H) nếu $\dfrac{\overline{X}-m_0}{\sigma}\sqrt{n} = u_{qs} > u_\alpha,$

trong đó u_α được xác định từ hệ thức $P((u \geq u_\alpha) = \alpha, \quad u \in N(0,1).$

Mẫu có phân bố chuẩn với phương sai σ^2 đã cho. Kiểm định giả thiết về kì vọng mẫu, mức ý nghĩa α

$$(H) : m \leqslant m_0,$$

với đối thiết

$$(K) : m > m_0.$$

Quy tắc: Bác bỏ (H) nếu $\dfrac{\overline{X} - m_0}{\sigma}\sqrt{n} = u_{qs} > u_\alpha,$

trong đó u_α được xác định từ hệ thức $P((u \geq u_\alpha) = \alpha, \quad u \in N(0,1).$

Mẫu có phân bố chuẩn với phương sai σ^2 đã cho. Kiểm định giả thiết về kì vọng mẫu, mức ý nghĩa α

$$(H) : m = m_0 \quad \text{hoặc} \quad (H) : m \leqslant m_0$$

với đối thiết

$$(K) : m > m_0.$$

Quy tắc: Bác bỏ (H) nếu $\dfrac{\overline{X} - m_0}{\sigma}\sqrt{n} = u_{qs} > u_\alpha,$

trong đó u_α được xác định từ hệ thức $P((u \geq u_\alpha) = \alpha, \quad u \in N(0,1).$

Hoàn toàn tương tự, chúng ta sẽ xét bài toán kiểm định 1 phía nữa

Bài toán 3

Mẫu có phân bố chuẩn với phương sai σ^2 đã cho. Kiểm định giả thiết về kì vọng mẫu, mức ý nghĩa α

$$(H) : m = m_0 \quad \text{hoặc} \quad (H) : m \geq m_0$$

với đối thiết

$$(K) : m < m_0.$$

Quy tắc: Bác bỏ (H) nếu $\dfrac{\overline{X} - m_0}{\sigma}\sqrt{n} = u_{qs} < -u_\alpha,$

trong đó u_α được xác định từ hệ thức $P((u \geq u_\alpha) = \alpha, \quad u \in N(0,1).$

8 Kiểm định giả thiết về giá trị trung bình (trường hợp σ^2 chưa biết)

Mẫu có phân bố chuẩn với phương sai σ^2 chưa biết. Kiểm định giả thiết về kì vọng mẫu, mức ý nghĩa α

(a) **Bài toán 1**

$$(H): m = m_0$$

với đối thiết

$$(K): m \neq m_0.$$

Quy tắc: Bác bỏ (H) nếu $\left|\dfrac{\overline{X} - m_0}{S^*}\sqrt{n}\right| > t_\alpha,$

trong đó t_α được xác định từ hệ thức $P(|t| \geq t_\alpha) = \alpha$

(t có phân bố Student với $n-1$ bậc tự do.)

(b) **Bài toán 2**

$$(H): m = m_0 \quad \text{hoặc} \quad (H): m \leqslant m_0$$

với đối thiết

$$(K): m > m_0.$$

Quy tắc: Bác bỏ (H) nếu $t_{qs} = \dfrac{\overline{X} - m_0}{S^*}\sqrt{n} > t_\alpha,$

trong đó t_α được xác định từ hệ thức $P(t \geq t_\alpha) = \alpha$

(t có phân bố Student với $n-1$ bậc tự do.)

(c) **Bài toán 3**

$$(H): m = m_0 \quad \text{hoặc} \quad (H): m \geq m_0$$

với đối thiết

$$(K): m < m_0.$$

Quy tắc: Bác bỏ (H) nếu $t_{qs} = \dfrac{\overline{X} - m_0}{S^*}\sqrt{n} < -t_\alpha,$

trong đó t_α được xác định từ hệ thức $P(t \geq t_\alpha) = \alpha$

(t có phân bố Student với $n-1$ bậc tự do.)

9 Kiểm định giả thiết về sự bằng nhau của các giá trị trung bình

9.1 Trường hợp phương sai đã biết

Mẫu $\{X_i\}_{i=1}^m \in N(m_1, \sigma_1^2)$ $\{Y_i\}_{i=1}^n \in N(m_2, \sigma_2^2)$, có phân bố chuẩn với phương sai σ_1^2, σ_2^2 đã biết. Kiểm định giả thiết về kì vọng mẫu, mức ý nghĩa α

(a) **Bài toán 1**

$$(H) : m_1 = m_2$$

với đối thiết

$$(K) : m_1 \neq m_2.$$

Quy tắc: Bác bỏ (H) nếu $\left| \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \right| > u_\alpha,$

trong đó u_α được xác định từ hệ thức $P((|u| \geq u_\alpha) = \alpha, \quad u \in N(0,1).$

(b) **Bài toán 2**

$$(H) : m_1 = m_2 \quad \text{hoặc} \quad (H) : m_1 \leqslant m_2$$

với đối thiết

$$(K) : m_1 > m_2.$$

Quy tắc: Bác bỏ (H) nếu $\dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} > u_\alpha,$

trong đó u_α được xác định từ hệ thức $P((u \geq u_\alpha) = \alpha, \quad u \in N(0,1).$

(c) **Bài toán 3**

$$(H) : m_1 = m_2 \quad \text{hoặc} \quad (H) : m_1 \geq m_2$$

với đối thiết

$$(K) : m_1 < m_2.$$

Quy tắc: Bác bỏ (H) nếu $\dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} < -u_\alpha,$

trong đó u_α được xác định từ hệ thức $P((u \geq u_\alpha) = \alpha, \quad u \in N(0,1).$

Nếu mẫu có kích thước đủ lớn ($m, n > 30$), một cách xấp xỉ khá tốt là áp dụng quy tắc nêu trên để kiểm định giả thiết không, kể cả trường hợp phân bố mẫu không có phân bố chuẩn, thay các phương sai σ_1^2, σ_2^2 trong thống kê u bằng các phương sai mẫu điều chỉnh S_X^{*2} và S_Y^{*2}.

9.2 Trường hợp các phương sai chưa biết và bằng nhau

Mẫu $\{X_i\}_{i=1}^m \in N(m_1, \sigma^2)$ $\{Y_i\}_{i=1}^n \in N(m_2, \sigma^2)$, có phân bố chuẩn với phương sai σ^2 chưa biết. Kiểm định giả thiết về kì vọng mẫu, mức ý nghĩa α

(a) **Bài toán 1**

$$(H): m_1 = m_2$$

với đối thiết

$$(K): m_1 \neq m_2.$$

Quy tắc: Bác bỏ (H) nếu $\left| \sqrt{\frac{mn(m+n-2)}{m+n}} \cdot \frac{\overline{X} - \overline{Y}}{\sqrt{mS_X^2 + nS_Y^2}} \right| > t_\alpha,$

trong đó t_α được xác định từ hệ thức $P(|t| \geq t_\alpha) = \alpha$

(t có phân bố Student với $m+n-2$ bậc tự do.)

(b) **Bài toán 2**

$$(H): m_1 = m_2 \quad \text{hoặc} \quad (H): m_1 \leqslant m_2$$

với đối thiết

$$(K): m_1 > m_2.$$

Quy tắc: Bác bỏ (H) nếu $\sqrt{\frac{mn(m+n-2)}{m+n}} \cdot \frac{\overline{X} - \overline{Y}}{\sqrt{mS_X^2 + nS_Y^2}} > t_\alpha,$

trong đó t_α được xác định từ hệ thức $P(t \geq t_\alpha) = \alpha$

(t có phân bố Student với $m+n-2$ bậc tự do.)

(c) **Bài toán 3**

$$(H): m_1 = m_2 \quad \text{hoặc} \quad (H): m_1 \geq m_2$$

với đối thiết

$$(K): m_1 < m_2.$$

Quy tắc: Bác bỏ (H) nếu $\sqrt{\frac{mn(m+n-2)}{m+n}} \cdot \frac{\overline{X} - \overline{Y}}{\sqrt{mS_X^2 + nS_Y^2}} < -t_\alpha,$

trong đó t_α được xác định từ hệ thức $P(t \geq t_\alpha) = \alpha$

(t có phân bố Student với $m+n-2$ bậc tự do.)

10 Kiểm định giả thiết về sự bằng nhau của các phương sai

Giả sử $\{X_i\}_{i=1}^m \in N(m_1, \sigma_X^2)$ $\{Y_i\}_{i=1}^n \in N(m_2, \sigma_Y^2)$ là các mẫu hoàn toàn độc lập, có phân bố chuẩn. Kiểm định giả thiết về các phương sai, với mức ý nghĩa α. Ta sắp xếp sao cho ${S_X^*}^2 > {S_Y^*}^2$

(a) **Bài toán 1**

$$(H) : \sigma_X^2 = \sigma_Y^2$$

với đối thiết

$$(K) : \sigma_X^2 \neq \sigma_Y^2.$$

Quy tắc: Bác bỏ (H) nếu $\dfrac{{S_X^*}^2}{{S_Y^*}^2} > F_{\alpha/2}$,

trong đó $F_{\alpha/2}$ được xác định từ hệ thức $P(F \geq F_{\alpha/2}) = \dfrac{\alpha}{2}$

(F là đại lượng ngẫu nhiên phân bố F với $m-1, n-1$ bậc tự do.)

(b) **Bài toán 2**

$$(H) : \sigma_X^2 = \sigma_Y^2 \quad \text{hoặc} \quad (H) : \sigma_X^2 \leqslant \sigma_Y^2$$

với đối thiết

$$(K) : \sigma_X^2 > \sigma_Y^2.$$

Quy tắc: Bác bỏ (H) nếu $\dfrac{{S_X^*}^2}{{S_Y^*}^2} > F_{\alpha}$,

trong đó F_{α} được xác định từ hệ thức $P(F \geq F_{\alpha}) = \alpha$

(F là đại lượng ngẫu nhiên phân bố F với $m-1, n-1$ bậc tự do.)

11 Kiểm định giả thiết về xác suất của biến cố ngẫu nhiên

Giả sửA là biến cố ngẫu nhiên có xác suất $P(A) = p$ chưa biết. Ta sử dụng ước lượng

$$\widehat{p} = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

trong đó X_i bằng 1 hoặc 0 tùy theo biến cố A xảy ra hoặc không xảy ra ở phép thử ngẫu nhiên thứ $i, i = 1, 2, ..., n$. ($\widehat{p}$ thực chất là tần suất xuất hiện của biến cố A). Khi đó $n\widehat{p}$ có phân bố nhị thức với

$$E(n\widehat{p}) = np, \quad D(n\widehat{p}) = npq, q = 1 - p$$

với mức ý nghĩa α cho trước

Ta đã biết, theo định lí giới hạn trung tâm

$$\frac{n\widehat{p} - np}{\sqrt{npq}} = \sqrt{n}\frac{\widehat{p} - p}{\sqrt{pq}}$$

có phân bố xấp xỉ chuẩn ($\approx N(0,1)$) khi n đủ lớn. Vì vậy sử dụng thống kê

$$u = u_{qs} = \sqrt{n}\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)}},$$

u có phân bố xấp xỉ chuẩn N(0,1), khi giả thiết (H): $p = p_0$ đúng.

Kiểm định giả thiết về xác suất của biến cố ngẫu nhiên.

Giả thiết kích thước mẫu n đủ lớn ($n \geq 40$). Kiểm định giả thiết về xác suất, mức ý nghĩa α

(a) **Bài toán 1**

$$(H) : p = p_0$$

với đối thiết

$$(K) : p \neq p_0.$$

Quy tắc: Bác bỏ (H) nếu $\left| \sqrt{n} \dfrac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)}} \right| > u_\alpha,$

trong đó u_α được xác định từ hệ thức $P(|u| \geq u_\alpha) = \alpha$

(u có phân bố chuẩn $u \in N(0,1)$.)

(b) **Bài toán 2**

$$(H) : p = p_0 \quad \text{hoặc} \quad (H) : p \leqslant p_0$$

với đối thiết

$$(K) : p > p_0.$$

Quy tắc: Bác bỏ (H) nếu $\sqrt{n} \dfrac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)}} > u_\alpha,$

trong đó u_α được xác định từ hệ thức $P(u \geq u_\alpha) = \alpha$

(u có phân bố chuẩn $u \in N(0,1)$.)

(c) **Bài toán 3**

$$(H) : p = p_0 \quad \text{hoặc} \quad (H) : p \geq p_0$$

với đối thiết

$$(K) : p < p_0.$$

Quy tắc: Bác bỏ (H) nếu $\sqrt{n} \dfrac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)}} < -u_\alpha,$

trong đó u_α được xác định từ hệ thức $P(u \geq u_\alpha) = \alpha$

(u có phân bố chuẩn $u \in N(0,1)$.)

Trong bài toán 2, bài toán 3, u_α được xác định từ hệ thức

$$P(u > u_\alpha) = \alpha$$

trong khi đó ở bài toán 1, u_α được xác định từ hệ thức

$$P(|u| > u_\alpha) = \alpha$$

12 Kiểm định giả thiết về tính phù hợp của hàm phân bố

Giả thiết mẫu ngẫu nhiên gồm n phần tử mẫu. Các phần tử mẫu được phân loại thành r nhóm: mỗi nhóm chứa n_i phần tử mẫu, mỗi phần tử mẫu chỉ thuộc một nhóm duy nhất

$$n = n_1 + n_2 + ... + n_r = \sum_{i=1}^{r} n_i.$$

Xét bài toán kiểm định mức ý nghĩa α, giả thiết không sau đây:

(H) : Xác suất để mỗi phần tử mẫu thuộc nhóm thứ i bằng p_i

$$\text{với mọi } i = 1, 2, ..., r \quad (\sum_{i=1}^{r} p_i = 1).$$

Quy tắc: Bác bỏ (H) nếu $Q^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} > \chi^2_\alpha,$

trong đó χ^2_α được xác định từ hệ thức $P(\chi^2 > \chi^2_\alpha) = \alpha,$

(χ^2là đại lượng ngẫu nhiên có phân bố χ^2 với $r-1$ bậc tự do).

Người ta cũng sử dụng phân bố χ^2 để kiểm định các bài toán về tính phù hợp của hàm phân bố. Xét bài toán kiểm định giả thiết:

(H): Một đại lượng ngẫu nhiên X nào đó có phân bố dạng $F(x, \Theta)$ với đối thiết ngược lại.

Giả sử tham số $\Theta = (\Theta_1, \Theta_2, ..., \Theta_k)$ là véc tơ, gồm k tham số tạo thành (chẳng hạn như dạng phân bố chuẩn $F(x, \Theta) = F(x, m, \sigma^2) \in N(m, \sigma^2)$ gồm 2 tham số thành phần).

Để giải bài toán đó, người ta chọn một mẫu ngẫu nhiên

$$(X_1, X_2, ..., X_n)$$

tương ứng với đại lượng ngẫu nhiên X và chia các phần tử mẫu vào r nhóm: mỗi nhóm chứa n_i phần tử mẫu, mỗi phần tử mẫu chỉ thuộc một nhóm duy nhất

$$n = n_1 + n_2 + ... + n_r = \sum_{i=1}^{r} n_i.$$

Giả sử p_i là xác suất để đại lượng ngẫu nhiên X nhận các giá trị thuộc nhóm thứ $i, i = 1, 2, ..., r$ với điều kiện giả thiết (H) đúng. Khi đó

$$1 = p_1 + p_2 + ... + p_r$$

Hiển nhiên n_i là đại lượng ngẫu nhiên có phân bố nhị thức với kì vọng $E(n_i) = np_i$. Xét thống kê

$$Q^2 = \sum_{i=1}^{r} \frac{(n_i - n\widehat{p_i})^2}{n\widehat{p_i}}$$

trong đó $\widehat{p_i}, i = 1, 2, ..., r$ là xác suất để X nhận các giá trị thuộc nhóm thứ i, xác suất đó được tính thông qua hàm phân bố $F(x, \widehat{\Theta})$ mà $\widehat{\Theta} = (\widehat{\Theta_1}, \widehat{\Theta_2}, ..., \widehat{\Theta_k})$ là các ước lượng hợp lí cực đại của các tham số $\Theta_1, \Theta_2, ..., \Theta_k$.

Người ta đã chứng minh được rằng với n đủ lớn và giả thiết (H) là đúng khi đó Q^2 sẽ có phân bố xấp xỉ phân bố χ^2 với $r - k - 1$ bậc tự do, k là số tham số của phân bố $F(x, \Theta)$ trong giả thiết (H).

(Giả sử phân bố $F(x, \Theta)$ là phân bố chuẩn $N(m, \sigma^2)$, Θ được coi như véc tơ (m, σ^2) và số tham số của phân bố bằng $k = 2$, trường hợp $F(x, \lambda)$ là phân bố mũ chẳng hạn số tham số của phân bố là $k = 1$,...)

Miền bác bỏ của kiểm định do vậy là

$$W = \{(X_1, X_2, ..., X_n) \in R^n / \sum_{i=1}^{r} \frac{(n_i - n\widehat{p_i})^2}{n\widehat{p_i}} > \chi^2_\alpha\}.$$

trong đó χ^2_α được xác định từ hệ thức $P(\chi^2 > \chi^2_\alpha) = \alpha$, ($\chi^2$ là đại lượng ngẫu nhiên có phân bố χ^2 với $r-k-1$ bậc tự do). Ta tóm tắt quy tắc trên trong bảng sau

Kiểm định sự phù hợp với hàm phân bố chứa tham số chưa biết.

Giả thiết mẫu ngẫu nhiên gồm n phần tử mẫu. Các phần tử mẫu được phân loại thành r nhóm: mỗi nhóm chứa n_i phần tử mẫu, mỗi phần tử mẫu chỉ thuộc một nhóm duy nhất

$$n = n_1 + n_2 + ... + n_r = \sum_{i=1}^{r} n_i.$$

Xét bài toán kiểm định mức ý nghĩa α, giả thiết không sau đây:

$$(H): \text{ Mẫu ngẫu nhiên có phân bố dạng } F(x, \Theta)$$

Quy tắc: Bác bỏ (H) nếu $\quad Q^2 = \sum_{i=1}^{r} \frac{(n_i - n\widehat{p_i})^2}{n\widehat{p_i}} > \chi^2_\alpha,$

trong đó $\widehat{p_i}, i = 1, 2, ..., r$ là xác suất để X nhận các giá trị thuộc nhóm thứ i, xác suất đó được tính thông qua hàm phân bố $F(x, \widehat{\Theta})$ mà $\widehat{\Theta} = (\widehat{\Theta_1}, \widehat{\Theta_2}, ..., \widehat{\Theta_k})$ là các **ước lượng hợp lí cực đại** của các tham số $\Theta_1, \Theta_2, ..., \Theta_k$.

Phân vị χ^2_α được xác định từ hệ thức $P(\chi^2 > \chi^2_\alpha) = \alpha,$

(χ^2là đại lượng ngẫu nhiên có phân bố χ^2 với $r - k - 1$ bậc tự do).

13 Kiểm định về tính độc lập

Người ta có thể kiểm định về tính độc lập của các biến cố ngẫu nhiên, các đại lượng ngẫu nhiên. Chúng ta trình bày vấn đề dưới dạng sau đây:

Cho hai hệ đầy đủ các biến cố

$$A_1, A_2, ..., A_r; \quad B_1, B_2, ..., B_s.$$

Hãy kiểm định giả thiết hai hệ đó độc lập:

(H): $P(A_iB_j) = P(A_i)P(B_j)$ với mọi $i = 1, 2, ..., r; j = 1, 2, ..., s$.

Xét một mẫu ngẫu nhiên cỡ n (mẫu gồm n phần tử mẫu). Ta đưa vào các kí hiệu sau:

n_{ij} là số lần xảy ra biến cố tích A_iB_j trong tập hợp các phần tử mẫu.

$n_{i.} = \sum_{j=1}^{s} n_{ij}$ là số lần xảy ra biến cố A_i.

$n_{.j} = \sum_{i=1}^{r} n_{ij}$ là số lần xảy ra biến cố B_j.

Hiển nhiên

$$\sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.j} = n$$

và

$$\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} = n.$$

Các số n_{ij} được xếp vào bảng sau đây:

j / i	1	2	. . .	s	Tổng
1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2.}$
.			$\cdots$	.	
.			$\cdots$	.	
.			$\cdots$	.	
r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r.}$
Tổng	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.s}$	n

Ta tóm tắt quy tắc kiểm định trong bảng sau

Kiểm định về tính độc lập.

Cho hai hệ đầy đủ các biến cố

$$A_1, A_2, ..., A_r; \quad B_1, B_2, ..., B_s.$$

Hãy kiểm định giả thiết hai hệ đó độc lập, với mức ý nghĩa bằng α:

$$(H): \quad P(A_iB_j) = P(A_i)P(B_j) \text{ với mọi } i = 1, 2, ..., r; j = 1, 2, ..., s.$$

Quy tắc: Bác bỏ (H) nếu $\displaystyle\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} > \chi^2_\alpha,$

trong đó χ^2_α được xác định từ hệ thức $P(\chi^2 > \chi^2_\alpha) = \alpha,$

(χ^2là đại lượng ngẫu nhiên có phân bố χ^2 với $(r-1)(s-1)$ bậc tự do).

Chú ý rằng xấp xỉ tương đối tốt nếu $\frac{n_{i.}n_{.j}}{n^2} \geq 5$ *với mọi* i, j.

14 Hệ số tương quan mẫu

Trong lí thuyết xác suất, chúng ta biết rằng để đo mối quan hệ giữa hai hoặc nhiều đại lượng ngẫu nhiên, người ta thường tính các hệ số tương quan giữa chúng.

$$\varrho(X,Y)=\frac{cov(X,Y)}{\sigma_x\sigma_y}=\frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{D(X)}\sqrt{D(X)}}.$$

Nếu X và Y là hai đại lượng ngẫu nhiên độc lập khi đó hệ số tương quan $\varrho(X,Y)=0$. Trường hợp $|\varrho(X,Y)|=1$, giữa X và Y có mối quan hệ phụ thuộc tuyến tính $Y=aX+b$. Trong thống kê, thay vì hai đại lượng ngẫu nhiên X,Y ta xét mẫu ngẫu nhiên

$$(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$$

Có thể coi chúng như các điểm ngẫu nhiên trên mặt phẳng toạ độ. *Hệ số tương quan mẫu* được định nghĩa

$$r=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y})}{S_xS_Y}=\frac{\frac{1}{n}\sum_{i=1}^{n}x_iY_i-\overline{x}\cdot\overline{Y}}{S_xS_Y}.$$

S_X^2,S_Y^2 là phương sai mẫu của X,Y tương ứng

$$S_X^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2,S_Y^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\overline{Y}^2.$$

Dễ dàng chứng minh được

$$r=\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y})}{S_x^*S_Y^*}=\frac{\sum_{i=1}^{n}x_iY_i-n\overline{x}\cdot\overline{Y}}{\sqrt{\left(\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right)\left(\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2\right)}}.$$

Chẳng hạn ta xét bài toán dự báo đỉnh lũ hàng năm trên sông Hồng tại Hà nội, người ta thu thập các số liệu hàng năm về lượng mưa trong tháng Sáu trên thượng nguồn sông Hồng (X_i) và đỉnh lũ tương ứng với năm đó tại Hà nội (Y_i). Các số liệu giả định nhằm giúp độc giả nghiên cứu cách sử dụng hồi quy trong công việc dự báo được cho trong bảng dưới đây

STT	Năm	Lượng mưa (X)	Đỉnh lũ (Y)	STT	Năm	Lượng mưa (X)	Đỉnh lũ (Y)
1	1969	720	1405	13	1981	690	1337
2	1970	720	1405	14	1982	500	960
3	1971	730	1439	15	1983	460	879
4	1972	590	1133	16	1984	610	1176
5	1973	660	1272	17	1985	710	1382
6	1974	780	1519	18	1986	620	1178
7	1975	770	1524	19	1987	660	1271
8	1976	710	1364	20	1988	620	1194
9	1977	640	1253	21	1989	590	1161
10	1978	670	1324	22	1990	740	1449
11	1979	520	1002	23	1991	640	1225
12	1980	660	1303	24	1992	805	1377

Nếu ta minh hoạ các cặp số liệu $(x_i,y_i),i=1,2,...,24$ trong bảng trên bằng các điểm trên mặt phẳng, chúng ta cảm nhận thấy một mối liên hệ giữa lượng mưa (X) hàng năm và đỉnh lũ tại Hà nội (Y), lượng mưa càng lớn thì lũ do mưa gây nên càng cao. Hệ số tương quan mẫu sẽ giải thích mối quan hệ giữa hai đại lượng: lượng mưa hàng năm và đỉnh lũ tại Hà nội. Để tính hệ số tương quan mẫu giữa chúng, ta tính các đặc trưng kì vọng mẫu và phương sai mẫu của X và Y

$\overline{x}$	$\overline{y}$	S_x^2	S_y^2
$\frac{1}{n}\sum_{i=1}^{n}x_i$	$\frac{1}{n}\sum_{i=1}^{n}y_i$	$\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2$
658,95833	1272,16667	$85,02425^2$	$163,5071^2$

Hệ số tương quan mẫu do vậy bằng

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{S_x S_y} = 0,97045.$$

Dựa vào hệ số tương quan mẫu, sau này người ta giải thích được mức độ liên hệ giữa hai đại lượng ngẫu nhiên X và Y khi biểu diễn chúng thông qua mối quan hệ tuyến tính.

15 Hồi quy bình phương trung bình tuyến tính

Giả sử

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$$

là mẫu ngẫu nhiên tương ứng với hai đại lượng ngẫu nhiên X và Y. Chẳng hạn khi xét bài toán dự báo đỉnh lũ hàng năm trên sông Hồng tại Hà nội đã nói trong mục trước. Chúng ta cảm nhận được mối liên hệ giữa lượng mưa (X) hàng năm và đỉnh lũ tại Hà nội (Y), tuy nhiên không có thông tin nào hơn về mối liên hệ thực giữa X và Y, khi đó ta giả thiết giữa chúng có mối quan hệ *tuyến tính* (bậc nhất). Mặt khác do chúng ta xem lượng mưa và đỉnh lũ là các đại lượng ngẫu nhiên, vì vậy khi dự báo lượng mưa Y với điều kiện lượng mưa X bằng một giá trị x nào đó, ta chỉ có thể khảo sát hàm phân bố có điều kiện của Y. (X còn gọi là biến độc lập và Y được gọi là biến phụ thuộc). Đặc trưng quan trọng của phân bố có điều kiện là *kì vọng có điều kiện* $E(Y/X = x)$. Vì vậy trong chương này chúng ta hạn chế chỉ xét trường hợp *kì vọng có điều kiện* $E(Y/X = x)$ *là hàm tuyến tính đối với* X

$$\boxed{E(Y/X = x) = \alpha x + \beta.}$$

Chú ý rằng khi X tăng 1 đơn vị, kì vọng có điều kiện của Y sẽ tăng α

$$E(Y/X = x + 1) = \alpha(x + 1) + \beta = \alpha x + \beta + \alpha = E(Y/X = x) + \alpha.$$

Để chỉ ra được sự phụ thuộc hàm đó, với thông tin duy nhất là các cặp số liệu $(x_i, y_i), i = 1, 2, ..., n$, trong bài toán hồi quy người ta coi x_i là các biểu hiện cụ thể của biến ngẫu nhiên X, y_i là các biểu hiện cụ thể của biến ngẫu nhiên phụ thuộc Y_i tương ứng. Do đẳng thức trên, kì vọng có điều kiện của Y_i thoả mãn

$$\boxed{E(Y_i/X = x_i) = \alpha x_i + \beta \qquad i = 1, 2, ..., n.}$$

Như vậy sai số giữa Y_i và kì vọng có điều kiện $E(Y_i/X = x_i)$, kí hiệu

$$\varepsilon_i = Y_i - E(Y_i/X = x_i) = Y_i - (\alpha x_i + \beta)$$

là đại lượng ngẫu nhiên có kì vọng bằng 0

$$E(\varepsilon_i) = E(Y_i) - E(E(Y_i/X = x_i)) = E(Y_i) - E(Y_i) = 0.$$

Vậy **mẫu hồi quy tuyến tính** của Y đối với X được tóm tắt như sau:

Đại lượng ngẫu nhiên độc lập X nhận các giá trị x_i, khi đó

$$\boxed{Y_i = \alpha x_i + \beta + \varepsilon_i \qquad i = 1, 2, ..., n.} \tag{3}$$

trong đó α, β là các hệ số cần ước lượng, $y = \alpha x + \beta$ được gọi là đường thẳng hồi quy, ε_i là đại lượng ngẫu nhiên có kì vọng $E(\varepsilon_i) = 0$.

Ta gọi a, b là các ước lượng bất kì của các hệ số α, β tương ứng. Khi đó đường thẳng hồi quy được ước lượng là đường thẳng

$$y = ax + b.$$

Độ lệch (hay tạm gọi là sai số) giữa y_i với đường thẳng trên tại điểm x_i, kí hiệu e_i bằng

$$e_i = y_i - (ax_i + b).$$

Độ lệch này có thể dương hoặc âm tuỳ theo giá trị mẫu (x_i, y_i) là điểm nằm trên hoặc nằm dưới đường thẳng ước lượng $y = ax + b$. Một trong các phương pháp ước lượng có nhiều ưu điểm là tìm các ước lượng a, b của α, β sao cho tổng bình phương các độ lệch e_i đạt giá trị nhỏ nhất. Người ta gọi phương pháp ước lượng như vậy là *phương pháp bình phương bé nhất*. Đường thẳng hồi quy nhận được từ phương pháp bình phương bé nhất còn được gọi là *hồi quy bình phương trung bình tuyến tính*.

Các ước lượng a, b của α và β dựa trên phương pháp bình phương bé nhất, tức là làm cực tiểu hàm

$$u(a,b) = \sum_{i=1}^{n}(Y_i - ax_i - b)^2.$$

Bài toán trên có thể giải một cách dễ dàng bằng cách tìm điểm dừng của hàm $u(a,b)$:

$$\begin{cases} \frac{\partial u}{\partial a} = -2\sum_{i=1}^{n}(Y_i - ax_i - b)x_i = 0 \\ \frac{\partial u}{\partial b} = -2\sum_{i=1}^{n}(Y_i - ax_i - b) = 0 \end{cases}$$

Từ phương trình thứ hai suy ra

$$b = \overline{Y} - a\overline{x}. \tag{4}$$

Thay b vào phương trình thứ nhất, khi đó

$$\sum_{i=1}^{n}[(Y_i - \overline{Y}) - a(x_i - \overline{x})]x_i = \sum_{i=1}^{n}[(Y_i - \overline{Y}) - a(x_i - \overline{x})](x_i - \overline{x}) = 0.$$

Suy ra

$$a = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}x_iY_i - n\overline{x}\overline{Y}}{\sum_{i=1}^{n}x_i^2 - n\overline{x}^2} = r\frac{S_Y}{S_x}, \tag{5}$$

trong đó r là hệ số tương quan mẫu

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{S_xS_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_iY_i - \overline{x}\cdot\overline{Y}}{S_xS_Y}. \tag{6}$$

S_X^2, S_Y^2 là phương sai mẫu của X, Y tương ứng

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2,$$

$$S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \overline{Y}^2. \tag{7}$$

Vậy hàm hồi quy bình phương trung bình tuyến tính có dạng

$$\boxed{y = ax + b = \overline{y} + r\frac{S_y}{S_x}(x - \overline{x}).}$$

Trở lại ví dụ về dự báo lũ, ta đã tính

$$\overline{x} = 658,95833, \quad \overline{y} = 1272,16667, \quad S_x = 85,02425, \quad S_y = 163,5071$$

Hệ số tương quan mẫu $r = 0,97045$. Áp dụng công thức để tính các hệ số a và b của đường thẳng hồi quy $y = ax + b$

$$a = r\frac{S_y}{S_x} = 1,86623$$

$$b = \overline{y} - r\overline{x}\frac{S_y}{S_x} = 42,39808.$$

Vậy đường thẳng hồi quy của Y đối với X

$$y = 1,86623x + 42,39808.$$

Ta phát biểu định lí sau

Định lí 10 *[Định lí Gauss-Markov]*

Giả thiết rằng theo (3) mẫu hồi quy tuyến tính của Y đối với X :

$$Y_i = \alpha x_i + \beta + \varepsilon_i \qquad i = 1, 2, ..., n$$

thoả mãn

$$E(\varepsilon_i) = 0, \quad D(\varepsilon_i) = \sigma^2, \quad E(\varepsilon_i \varepsilon_j) = 0, \quad \text{với mọi} \quad i \neq j, i, j = 1...n$$

Khi đó các ước lượng a, b của α và β theo phương pháp bình phương bé nhất là các ước lượng không chệch có phương sai nhỏ nhất. Hơn nữa với mọi số thực u và v, $ua + vb$ cũng là ước lượng có phương sai nhỏ nhất trong số tất cả các ước lượng tuyến tính $\sum p_i Y_i = P'Y$ không chệch của $u\alpha + v\beta$.

Theo (4) và (5) a và b là các hàm tuyến tính của Y_i

$$a = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}, \quad b = \overline{Y} - a\overline{x}.$$

Vậy

$$E(a) = E\left(\frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) = \frac{\sum_{i=1}^n (x_i - \overline{x})(E(Y_i) - E(\overline{Y}))}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})\alpha(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \alpha$$

$$E(b) = E(\overline{Y} - a\overline{x}) = a\overline{x} + \beta - a\overline{x} = \beta.$$

Hay a, b là các ước lượng không chệch của α và β.

$$E(b) = \beta, \quad E(a) = \alpha.$$

Nhận xét rằng

$$E(b) = \beta, \quad E(a) = \alpha, \quad D(b) = \frac{\sigma^2}{n}, \quad D(a) = \frac{\sigma^2}{nS_x^2}.$$

Định lí 11 *Với các điều kiện của định lí Gauss-Markov, kì vọng của tổng bình phương sai số*

$$E(SSE) = (n-2)\sigma^2 \quad (SSE = \sum_{i=1}^n [y_i - (ax_i + b)]^2.)$$

Nói cách khác nếu kí hiệu

$$\sigma^{*2} = \frac{SSE}{n-2} \quad \left(= \frac{nS_Y^2(1-r^2)}{n-2}\right),$$

khi đó σ^{*2} là ước lượng không chệch của σ^2, σ^* còn được gọi là sai số chuẩn (Standard Error).

Ước lượng cho phương sai của α được tính như sau:

$$a = r\frac{S_Y}{S_x} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{S_x^2} = \sum_{i=1}^n \frac{x_i - \overline{x}}{nS_x^2} Y_i.$$

Đặt $C_i = \frac{x_i - \overline{x}}{nS_x^2}$, với mỗi giá trị cố định của x_i, phương sai của hệ số a bằng

$$D(a) = D\left(\sum_{i=1}^n \frac{x_i - \overline{x}}{nS_x^2} Y_i\right) = D(\sum_{i=1}^n C_i Y_i) = \sigma^2 \sum_{i=1}^n C_i^2 = \frac{\sigma^2}{nS_x^2}.$$

Sử dụng định lí trên, kí hiệu

$$s_a^2 = \frac{\sigma^{*2}}{nS_x^2} = \frac{SSE}{n(n-2)S_x^2}$$

ta có s_a^2 là ước lượng không chệch của $D(a)$, do vậy s_a được coi là sai số trung bình của hệ số góc α của phương trình đường thẳng hồi quy.

Chú ý rằng nếu cùng với các điều kiện của định lí Gauss-Markov, ta giả thiết thêm ε_i (sai số trong mẫu hồi quy) có *phân bố chuẩn*, khi đó thống kê

$$t = \frac{a - \alpha}{s_a}$$

có phân bố Student với $n-2$ bậc tự do. Do vậy khoảng tin cậy của α còn có thể viết dưới dạng

$$a - t_\epsilon s_a < \alpha < a + t_\epsilon s_a. \tag{8}$$

Cũng dựa trên cơ sở t có phân bố Student với $n-2$ bậc tự do, ta có thể kiểm định các giả thiết

$$H_0: \quad \alpha = \alpha_0 \qquad \text{hoặc} \qquad H_0: \quad \alpha \leqslant \alpha_0$$

với đối thiết

$$H_1: \qquad \alpha > \alpha_0,$$

$$\text{theo quy tắc bác bỏ} \quad H_0 \quad \text{nếu} \qquad t_{qs} = \frac{a - \alpha_0}{s_a} > t_\epsilon.$$

(Các kiểm định một phía khác hoặc kiểm định 2 phía cũng theo quy tắc tương tự đã biết).

Đặc biệt nếu giả thiết $\alpha = 0, \quad Y_i = \alpha + \varepsilon_i$ khi đó $E(Y_i) = \alpha$ không bị ảnh hưởng bởi biến độc lập X. Nói cách khác sự biến thiên của biến phụ thuộc Y hoàn toàn không một phần nào có thể giải thích bằng mối quan hệ tuyến tính với X.

Nhận xét rằng khi $\alpha = 0, t_{qs} = \frac{a}{s_a}$ là giá trị quan sát (t Stat) ứng với hệ số góc α trong bảng ANOVA phân tích hồi quy.

Tương tự xét hệ số tự do của hồi quy trung bình tuyến tính thực nghiệm

$$b = \overline{Y} - r\overline{x}\frac{S_Y}{S_x} = \overline{Y} - \overline{x}\frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{S_x^2} = \frac{1}{n}\sum_{i=1}^n Y_i - \sum_{i=1}^n \frac{x_i - \overline{x}}{nS_x^2} Y_i \cdot \overline{x}.$$

Đặt $C_i = \frac{x_i - \overline{x}}{nS_x^2}$, khi đó

$$b = \sum_{i=1}^n \left(\frac{1}{n} - C_i\overline{x}\right) Y_i.$$

Suy ra với mỗi giá trị cố định của x_i, phương sai của hệ số b bằng

$$D(b) = \sigma^2 \sum_{i=1}^n \left(\frac{1}{n} - C_i\overline{x}\right)^2 = \sigma^2 \sum_{i=1}^n \left(\frac{1}{n^2} - 2\frac{C_i}{n}\overline{x} + C_i^2\overline{x}^2\right) = \sigma^2 \sum_{i=1}^n \left(\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}\right).$$

Kí hiệu

$$s_b^2 = \sigma^{*2} \sum_{i=1}^n \left(\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}\right) = \frac{(1-r^2)S_Y^2(S_x^2 + \overline{x}^2)}{(n-2)S_x^2} = \frac{\sigma^{*2}(\sum_{i=1}^n x_i^2)}{n^2 S_x^2},$$

ta có s_b^2 là ước lượng không chệch của $D(b)$, s_b được coi là sai số trung bình của hệ số tự do β của phương trình đường thẳng hồi quy.

Cũng như hệ số góc của đường thẳng hồi quy, người ta chứng minh được rằng nếu ε_i có *phân bố chuẩn*, khi đó thống kê

$$t = \frac{b - \beta}{s_b}$$

có phân bố Student với $n-2$ bậc tự do. Do vậy áp dụng phương pháp ước lượng khoảng tin cậy cho giá trị trung bình, ta nhận được khoảng tin cậy của β

$$b - t_\epsilon s_b < \beta < b + t_\epsilon s_b. \tag{9}$$

Khi $\beta = 0, t_{qs} = \frac{b}{s_b}$ là giá trị quan sát (t Stat) ứng với hệ số tự do β trong bảng ANOVA phân tích hồi quy.

Ví dụ 1

Trong ví dụ ở mục trước, đường thẳng hồi quy của Y đối với X

$$y = 1,86623x + 42,39808.$$

Sai số trung bình

$$\sigma^* = \left(\frac{\sqrt{n}}{\sqrt{n-2}} S_Y\sqrt{1-r^2} = \right)\sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{37363,89302}{22}} = 41,21115.$$

1. *Sai số khi ước lượng các hệ số a và b của đường hồi quy*

Ta biết rằng

$$s_a^2 = \frac{S_Y^2(1-r^2)}{(n-2)S_X^2}$$

$$s_b^2 = \frac{(1-r^2)S_Y^2(S_X^2+\overline{X}^2)}{(n-2)S_X^2}.$$

Thay vào tính ta sẽ được các sai số khi ước lượng a và b. Sai số trung bình của a

$$s_a = 0,098939$$

Sai số của b

$$s_b = 65,73696$$

2. *Kiểm định quan hệ tuyến tính của hàm hồi quy*

Như đã trình bày ở trên, kiểm định về mối liên quan tuyến tính tương đương với kiểm định giả thuyết (H): $\alpha = 0$ với đối thiết (K): $\alpha \neq 0$

Khi giả thiết (H): $\alpha = 0$ đúng, giá trị quan sát của thống kê

$$t_{qs} = \frac{a-\alpha_0}{s_a} = \frac{1,86623}{0,098939} = 18,86$$

tra bảng phân vị phân bố Student với $n-2 = 22$ bậc tự do, mức ý nghĩa $\epsilon = 0,05$ ta có phân vị $t_{0,05} = 2,405468$. Giá trị quan sát lớn hơn nhiều so với phân vị $t_{0,05} = 2,405468$. Ta bác bỏ giả thiết $\alpha = 0$, mối quan hệ giữa Y và X là quan hệ tuyến tính.

Nhận xét rằng tương đương với kiểm định trên, ta có thể sử dụng thống kê F.

$$F_{qs} = \frac{(24-2)r^2}{1-r^2} = 355,7938$$

Với mức ý nghĩa $\epsilon = 0,05$ tra bảng phân vị phân bố F với 1 và $n-2 = 22$ bậc tự do, ta xác định

$$F_2 = 5,78632$$

Giá trị quan sát $F_{qs} = 355,7938$ lớn hơn rất nhiều so với $F_2 = 5,78632$, ta bác bỏ giả thiết (H): $\alpha = 0$, tức là mối quan hệ tuyến tính giữa Y và X khá chặt.

3. *Khoảng tin cậy cho hệ số góc α của đường hồi quy*

Thống kê

$$t = \frac{a-\alpha}{s_a}$$

có phân bố Student với 22 bậc tự do. Áp dụng công thức (8) tìm khoảng tin cậy với độ tin cậy 95% cho hệ số góc α: $a - t_\epsilon s_a < \alpha < a + t_\epsilon s_a$ (phân vị $t_{0,05} = 2,405468$) ta được khoảng tin cậy cho hệ số góc α là

$$(1,628237 \quad ; \quad 2,104225)$$

Ví dụ 2

Hãy phân tích hiệu quả của việc đầu tư quảng cáo (X) và doanh thu của một công ty (Y) trong khoảng thời gian một năm. Các số liệu được cho trong bảng dưới đây:

X	7	5	2	4	9	4
Y	14,99	12,08	5,55	9,79	16,38	9,68

X	9	6	3	4	7	5
Y	18,61	14,25	5,52	12,49	15,94	12,54

Sử dụng lệnh $\{= LINEST(Y, X, 1, 1)\}$ trong EXCEL (nhấn đồng thời các phím $CTRL + SHIFT + ENTER$) ta thu được bảng sau

1.72676783	2.965007587
0.199411812	1.161334855
0.882330203	1.47775679
74.98357456	10
163.7465154	21.83765129

Hàng thứ nhất là các hệ số hồi quy $a = 1.72676783, b = 2.965007587 \Rightarrow y = 1.72676783x + 2.965007587$
Sai số trung bình của các hệ số hồi quy a và b trong hàng thứ hai.

$$\sqrt{D(\alpha)} = 0.199411812 \quad \sqrt{D(\beta)} = 1.161334855.$$

Hàng thứ ba là hệ số số tương quan $r^2 = 0.882330203$ và sai số chuẩn (standard error) bằng

$$\sigma^* = 1.47775679.$$

Hàng thứ tư cho giá trị quan sát $F_{qs} = 74.98357456$ của phân bố F với 10 bậc tự do.
Hàng thứ năm là các tổng bình phương toàn phần theo Y (còn kí hiệu là SST) $nS_Y^2 = 163.7465154$ và phần dư $R_0^2 = 21.83765129$ (kí hiệu là SSR) trong bảng phân tích phương sai

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.939324333
R Square	0.882330203
Adjusted R Square	0.870563223
Standard Error	1.47775679
Observations	12

ANOVA

	df	SS	MS	F	Significance F
Regression	1	163.7465154	163.7465154	74.98357456	5.84643E-06
Residual	10	21.83765129	2.183765129		
Total	11	185.5841667			

	Coefficients	Stand Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.965007587	1.161335	2.5531	0.028710768	0.377392	5.552623
X Variable 1	1.72676783	0.199412	8.6593	5.84643E-06	1.282451	2.171085

Áp dụng công thức (8) ta được 2 cận trên, cận dưới $(1.282451; 2.171085)$ của hệ số góc của đường thẳng hồi quy với độ tin cậy 95%. Các nhận xét sau công thức (8) và (9):

$$t_{qs} = \frac{a}{\sqrt{D(a)}}, t_{qs} = \frac{b}{\sqrt{D(b)}}$$

cho ta các giá trị quan sát *t Stat* 8.6593 và 2.5531. Công thức (9) để tính khoảng tin cậy cho hệ số tự do b của đường thẳng hồi quy với độ tin cậy 95%

$$(0.377392; 5.552623).$$

16 Hồi quy nhiều chiều

Bài toán hồi quy nhiều chiều là bài toán xét tác động của nhiều biến ngẫu nhiên $(X_1, X_2, ...)$ tới một biến ngẫu nhiên khác (Y). Chẳng hạn khi muốn tìm hiểu lãi suất hàng năm của các công ty tài chính, người ta thấy lãi suất đó tỉ lệ thuận với tổng thu (từ thuế của nhà nước, đơn vị của tổng thu này tính theo % và kí hiệu là X_1), đồng thời cũng tỉ lệ nghịch với số văn phòng giao dịch (X_2). (Do sự cạnh tranh giữa các công ty, số văn phòng giao dịch được mở ngày một tăng). Gọi Y là tỉ lệ lãi suất hàng năm của công ty (đơn vị %).

Bảng sau cho ta số liệu quan sát được về các đại lượng này trong vòng 25 năm.

STT	X_1	X_2	Y	STT	X_1	X_2	Y
1	3.92	7298	0.75	14	3.78	6672	0.84
2	3.61	6855	0.71	15	3.82	6890	0.79
3	3.32	6636	0.66	16	3.97	7115	0.7
4	3.07	6506	0.61	17	4.07	7327	0.68
5	3.06	6450	0.7	18	4.25	7546	0.72
6	3.11	6402	0.72	19	4.41	7931	0.55
7	3.21	6368	0.77	20	4.49	8097	0.63
8	3.26	6340	0.74	21	4.7	8468	0.56
9	3.42	6349	0.9	22	4.58	8717	0.41
10	3.42	6352	0.82	23	4.69	8991	0.51
11	3.45	6361	0.75	24	4.71	9179	0.47
12	3.58	6369	0.77	25	4.78	9318	0.32
13	3.66	6546	0.78				

Mẫu hồi quy nhiều chiều

$$E(Y_i/X_1 = x_{1i}, X_2 = x_{2i}, ..., X_k = x_{ki}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}, \quad i = 1, 2, ..., n.$$

hay

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i,$$

trong đó β_i là các hằng số cần ước lượng và ε_i là biến ngẫu nhiên có kì vọng bằng 0. Các mẫu ngẫu nhiên là các điểm quan sát

$$(x_{1i}, x_{2i}, ..., x_{ki}, y_i), \quad i = 1, 2, ..., n.$$

Do mẫu hồi quy nhiều chiều

$$E(Y_i/X_1 = x_{1i}, X_2 = x_{2i}, ..., X_k = x_{ki}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}, \quad i = 1, 2, ..., n.$$

Suy ra

$$E(Y_i/X_1 = x_{1i} + 1, X_2 = x_{2i}, ..., X_k = x_{ki}) - E(Y_i/X_1 = x_{1i}, X_2 = x_{2i}, ..., X_k = x_{ki}) = \beta_1$$

(Nghĩa là trong ví dụ trên nếu tổng thu tăng thêm 1%, với số văn phòng giao dịch X_2 không đổi, khi đó tỉ lệ lãi suất hàng năm tăng thêm β_1.)

Gọi $a, b_1, b_2, ..., b_k$ là các ước lượng tương ứng, khi đó mẫu dự báo của biến ngẫu nhiên Y là

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + ... + b_k x_k.$$

Theo đó các sai số

$$e_i = y_i - (a + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki}), \quad i = 1, 2, ..., n.$$

Đối với mẫu hồi quy tuyến tính nhiều chiều, các ước lượng $a, b_1, b_2, ..., b_k$ cần xác định theo phương pháp bình phương bé nhất, tức là tổng bình phương các độ lệch

$$SS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - ... - b_k x_{ki})^2$$

đạt giá trị nhỏ nhất.

Phương trình

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_k x_k$$

được gọi là mặt phẳng hồi quy của Y đối với $X_1, X_2, ..., X_k$.

Trước hết ta phát biểu định lí sau

Định lí 12

Giả thiết rằng mẫu hồi quy tuyến tính của Y đối với $X_1, X_2, ..., X_k$:

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i,$$

trong đó

1. *$x_{1i}, x_{2i}, ..., x_{ki}$ là các thể hiện của $X_{1i}, X_{2i}, ..., X_{ki}$. Các biến ngẫu nhiên đó độc lập với ε_i.*
2. *$E(\varepsilon_i) = 0, \quad D(\varepsilon_i) = \sigma^2, \quad E(\varepsilon_i \varepsilon_j) = 0, \quad$ với mọi $\quad i \neq j, i, j = 1...n$*
3. *Hạng của ma trận (x_{ij}) bằng k.*

Khi đó các ước lượng $a, b_1, b_2, ..., b_k$ xác định theo phương pháp bình phương bé nhất của α và $\beta_1, \beta_2, ..., \beta_k$ là các ước lượng không chệch có phương sai nhỏ nhất. Hơn nữa với mọi số thực $d_0, d_1, d_2, ..., d_k$, ước lượng $d_0 + d_1 b_1 + d_2 b_2 + \cdots + d_k b_k$ cũng là ước lượng có phương sai nhỏ nhất trong số tất cả các ước lượng tuyến tính không chệch của

$$d_0 + d_1 \beta_1 + d_2 \beta_2 + \cdots + d_k \beta_k.$$

Từ hệ thức

$$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + e_i,$$

bình phương cả hai vế đẳng thức trên và cộng chúng lại theo i ta được

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} e_i^2.$$

Đẳng thức có ý nghĩa như sau: vế trái là tổng bình phương các độ lệch giữa các phần tử mẫu của Y với giá trị trung bình mẫu $\overline{y}$, kí hiệu SST (total sum of squares) được phân tích thành tổng của hai phần: phần thứ nhất là tổng bình phương các độ lệch giữa hồi quy $\hat{y}_i$ với trung bình mẫu $\overline{y}$ và phần thứ hai là phần dư: tổng bình phương các sai số. Kí hiệu

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2 = nS_y^2 \quad \text{(Tổng bình phương chung)}$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \quad \text{(Tổng bình phương hồi quy)}$$

$$SSE = \sum_{i=1}^{n} e_i^2 \quad \text{(Tổng bình phương sai số).}$$

Theo đẳng thức: $SST = SSR + SSE$, khi đó tỉ số

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

được gọi là *hệ số xác định* biểu diễn lực của hồi quy. $0 \leqslant R^2 \leqslant 1$ và khi R^2 càng gần với 1, phần dư SSE (tổng bình phương các sai số) càng nhỏ so với tổng bình phương các độ lệch chung của Y.

Chú ý: hệ số xác định điều chỉnh

$$\overline{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}.$$

Người ta chứng minh được rằng với các điều kiện của định lí trên

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1} = \frac{SSE}{n-k-1}$$

là ước lượng không chệch của σ^2. Ta gọi $s_e = \sqrt{s_e^2}$ là *sai số chuẩn*.

Việc tính *sai số chuẩn* của các hệ số hồi quy $b_k, b_{k-1}, ..., b_2, b_1, a$ phức tạp hơn (xem phần hồi quy đơn giản, một chiều). Các chương trình phần mềm thống kê sẽ tính giúp ta các sai số đó.

Thực hành trên EXCEL

Xét ví dụ về lãi suất hàng năm của các công ty tài chính, sử dụng lệnh $\{= LINEST(Y, X, 1, 1)\}$, ta được bảng sau

-0.000249079	0.237197475	1.564496771
3.20485E-05	0.055559366	0.079395981
0.865296068	0.053302217	
70.66057082	22	
0.40151122	0.06250478	

Hàng thứ nhất là các hệ số hồi quy viết theo đúng thứ tự

$$y = b_k x_k + b_{k-1}x_{k-1} + ... + b_2 x_2 + b_1 x_1 + a$$

Hay $y = -0.00025x_2 + 0.2372x_1 + 1.5645$.

Sai số trung bình (căn bậc hai của phương sai) của các hệ số hồi quy $b_k, b_{k-1}, ..., b_2, b_1, a$ cho trong hàng thứ hai.

$$\sqrt{D(b_2)} = 3.20485E-05, \quad \sqrt{D(b_1)} = 0.055559, \quad \sqrt{D(a)} = 0.079396.$$

Hàng thứ ba là hệ số xác định giải thích lực của hồi quy $R^2 = 0.865296068$ và sai số chuẩn (standard error) $s_e = 0.053302217$.

Hàng thứ tư cho giá trị quan sát $F_{qs} = 70.66057082$ của phân bố F với $(k, 22)$ bậc tự do. (Trong ví dụ này $k = 2$).

Hàng thứ năm là các tổng bình phương $SSR = 0.40151122$ và phần dư $SSE = 0.06250478$.

Chú ý rằng hồi quy tuyến tính nhiều chiều thường xuyên được sử dụng hơn hồi quy đơn giản (một chiều), nếu còn các biến độc lập tác động đáng kể tới biến phụ thuộc. Chẳng hạn trong ví dụ trên , biến phụ thuộc (lãi suất y) tỉ lệ thuận với tổng thu (x_1). Trong khi nếu ta chỉ quan tâm tới lãi suất và tổng thu, hồi quy đơn giản cho ta kết quả

$$y = 1.326 - 0.169x_1$$

lãi suất giảm khi x_1 tăng(!)

Tương quan bội và tương quan riêng

Ta nhấn mạnh rằng tương ứng với mẫu quan sát $y_i, \quad i = 1, 2, ..., n$ là mẫu dự báo

$$\hat{y_i} = a + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki}, \quad i = 1, 2, ..., n.$$

Hệ số tương quan giữa chúng được gọi là *hệ số tương quan bội*, nó đo mức độ tác dụng tuyến tính của $X = (X_1, \cdots, X_k)$ lên Y. (Dễ dàng chứng minh được: $Y - \hat{Y}$ không tương quan (trực giao) với $X_1, \cdots, X_k$. *Thực chất của phương pháp bình phương nhỏ nhất là sau khi tịnh tiến hệ trục toạ độ tới điểm $(EY, EX_1, ..., EX_k) \in R^{k+1}$, $\hat{Y}$ là phép chiếu vuông góc Y xuống $L_2(X_1, ..., X_k)$*). Suy ra, như đã biết trong lí thuyết về không gian Hilbert *hệ số tương quan chẳng qua là côsin của góc giữa hai véc tơ*, hệ số tương quan bội bằng căn bậc hai của hệ số xác định

$$R = \sqrt{R^2}.$$

Trong ví dụ của chúng ta $R = \sqrt{0.8652} = 0.93$.

Khi khảo sát mối tương quan ta tính hệ số tương quan giữa các đại lượng ngẫu nhiên, chẳng hạn $\varrho_{ij} = \varrho_{ij}(X_i, X_j)$. Đó là độ đo toàn phần mối tương quan giữa chúng (có kể đến mối quan hệ thông qua các biến ngẫu nhiên khác: $X_1, ..., X_k$). Như trên ta biết rằng có thể phân tích một đại lượng ngẫu nhiên thành tổng của hai đại lượng ngẫu nhiên không tương quan (chiếu vuông góc xuống $L_2(X_2, ..., X_k)$)

$$Y = \hat{Y}_{Y'2...k} + (Y - \hat{Y}_{Y'2...k}) = \hat{Y}_{Y'2...k} + \eta_{Y'2...k}, \quad X_1 = \hat{X_1} + (X_1 - \hat{X_1}) = \hat{X_1} + \eta_{1'2...k}$$

Có thể coi $\eta_{Y'2...k} = Y - \hat{Y}_{Y'2...k}$ là phần còn lại của Y sau khi đã loại đi các tác động tuyến tính của $X_2, ..., X_k$ vào Y. Tương tự $\eta_{1'2...k} = X_1 - \hat{X_1}$ là phần còn lại của X_1 sau khi đã loại đi các tác động tuyến tính của $X_2, ..., X_k$ vào X_1. Khi đó hệ số tương quan giữa hai phần dư $\eta_{Y'2...k} = Y - \hat{Y}_{Y'2...k}$ và $\eta_{1'2...k} = X_1 - \hat{X_1}$ được gọi là *hệ số tương quan riêng* (mối quan hệ nội tại, không phụ thuộc vào các đại lượng ngẫu nhiên khác: $X_2, ..., X_k$) giữa Y và X_1. Kí hiệu $\varrho_{Y.1} = \varrho(\eta_{Y'2...k}, \eta_{1'2...k})$.

Quay trở lại ví dụ trên, ta tính hệ hệ số tương quan riêng giữa lãi suất (Y) và số văn phòng giao dịch được mở ra (X_2). Ta lập bảng sau mà các cột dữ liệu là hồi quy của Y theo X_1 và hồi quy của Y theo X_2.

STT	$\eta_{Y'2\ldots k}$	$\eta_{1'2\ldots k} = X_1 - \hat{X}_1$	STT	$\eta_{Y'2\ldots k}$	$\eta_{1'2\ldots k} = X_1 - \hat{X}_1$
1	0.086830251	-53.63957787	14	0.153152011	-451.2549257
2	-0.005600136	9.06929472	15	0.109917223	-298.5076835
3	-0.104647917	263.1517884	16	0.045286765	-318.2055251
4	-0.196930487	540.9815244	17	0.042199793	-269.3374194
5	-0.10862179	501.2947138	18	0.112643243	-343.9748293
6	-0.080165276	371.7287666	19	-0.030295912	-219.9858604
7	-0.013252248	174.5968723	20	0.06323451	-184.4913759
8	-0.034795734	65.03092506	21	0.028751869	-156.0683541
9	0.152265111	-186.980106	22	-0.141543765	288.6899192
10	0.072265111	-183.980106	23	-0.022939434	383.2448354
11	0.007339019	-223.9196743	24	-0.059556828	538.6184565
12	0.049325955	-427.991137	25	-0.197717709	563.4261304
13	0.072856378	-381.4966525			

Hệ số tương quan riêng giữa lãi suất (Y) và số văn phòng giao dịch được mở ra (X_2) khi đó bằng $\varrho_{Y.1} = -0.85617$. (Sử dụng lệnh $CORREL$).

Bình phương hệ số tương quan riêng $(-0.85617)^2 = 0.73$, vậy 73% phần biến động của lãi suất (Y) được giải thích bởi sự phụ thuộc tuyến tính (tỉ lệ nghịch) vào số lượng văn phòng giao dịch được mở.

Tương tự hệ số tương quan riêng giữa lãi suất (Y) và (X_1) bằng $\varrho_{Y.2} = 0.6731$. (Tỉ lệ thuận).

*Ta cũng có thể tính tương quan riêng giữa lãi suất (Y) và (X_1) bằng cách sử dụng các công thức (**??**-**??**)*

$$\varrho_{01.(23\ldots n)} = \frac{-C_{10}}{\sqrt{C_{00}C_{11}}} = \frac{5.929936871}{\sqrt{3.10432981 \times 25}} = 0.673126.$$

Khoảng tin cậy và kiểm định giả thiết cho các tham số của hồi quy.

Các vấn đề về khoảng tin cậy và kiểm định giả thiết cho các tham số của hồi quy dựa trên định lí sau

Định lí 13

Với các giả thiết như trong định lí 12, đồng thời giả thiết thêm rằng rằng các số hạng sai số ε_i có phân bố chuẩn. Kí hiệu $s_{b_k}, s_{b_{k-1}}, \ldots, s_{b_2}, s_{b_1} s_a$ là các sai số chuẩn của các hệ số hồi quy $b_k, b_{k-1}, \ldots, b_2, b_1, a$, khi đó

$$t_a = \frac{a-\alpha}{s_a}, \quad t_{b_i} = \frac{b_i - \beta_i}{s_{b_i}}, \quad i = 1, 2, \ldots, k$$

là các đại lượng ngẫu nhiên có phân bố Student với $n-k-1$ bậc tự do.

Chẳng hạn trong ví dụ lãi suất của các công ty tài chính, với độ tin cậy 99%

$$0.081 < \beta_2 < 0.394, \quad -0.000339 < \beta_1 < -0.000159.$$

$$(s_{b_1} t_\epsilon - b_1 \leqslant \beta_1 \leqslant s_{b_1} t_\epsilon + b_1, \quad t_\epsilon = t_{0.01} = 2.81876, \quad s_{b_1} = 3.2 \times 10^{-5}, \quad b_1 = -0.000249)$$

Do mẫu hồi quy nhiều chiều

$$E(Y_i/X_1 = x_{1i}, X_2 = x_{2i}, \ldots, X_k = x_{ki}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}, \quad i = 1, 2, \ldots, n.$$

Suy ra

$$E(Y_i/X_1 = x_{1i} + 1, X_2 = x_{2i}, \ldots, X_k = x_{ki}) - E(Y_i/X_1 = x_{1i}, X_2 = x_{2i}, \ldots, X_k = x_{ki}) = \beta_1$$

Nghĩa là trong ví dụ đã nêu nếu số văn phòng giao dịch tăng thêm 1000, (với tổng thu X_1 không đổi), khi đó tỉ lệ lãi suất hàng năm giảm từ 0.159 tới 0.339.

Kiểm định giả thiết cho mỗi tham số của hồi quy.

Cũng dựa trên cơ sở t_{b_i} có phân bố Student với $n-k-1$ bậc tự do, ta có thể kiểm định các giả thiết

$$H_0: \quad \beta_i = \beta_{i,0} \qquad \text{hoặc} \qquad H_0: \quad \beta_i \leqslant \beta_{i,0}$$

với đối thiết

$$H_1: \qquad \beta_i > \beta_{i,0},$$

$$\text{theo quy tắc bác bỏ} \quad H_0 \quad \text{nếu} \qquad t_{qs} = \frac{\beta_i - \beta_{i,0}}{s_{b_i}} > t_\epsilon.$$

(Các kiểm định một phía khác hoặc kiểm định 2 phía cũng theo quy tắc tương tự đã biết).

Đặc biệt nếu giá trị thực của $\beta_1 = 0$

$$Y_i = \alpha + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$$

không bị ảnh hưởng bởi biến độc lập X_1 khi các biến $X_2, ..., X_k$ nhận các giá trị cố định cho trước. Nói cách khác X_1 không góp phần vào giải thích mối quan hệ tuyến tính giữa biến phụ thuộc với các biến độc lập.

Trong ví dụ trên kiểm định $H_0: \quad \beta_1 = 0$ với đối thiết $H_1: \quad \beta_1 > 0$

$$t_{qs} = \frac{b_i - \beta_{i,0}}{s_{b_i}} = \frac{0.237}{0.0555} = 4.27$$

Nhận xét rằng khi $\beta_i = 0, t_{qs} = \frac{b_i}{s_{b_i}}$ *là giá trị quan sát (t Stat) ứng với hệ số góc* β_i *trong bảng ANOVA phân tích hồi quy.*

Nếu mức ý nghĩa rất bé 0.5%, tra bảng 22 bậc tự do (1 phía) $t_\epsilon = 2.81876$, ta vẫn bác bỏ $H_0: \quad \beta_1 = 0$.

Tương tự xét bài toán kiểm định $H_0: \quad \beta_2 = 0$ với đối thiết $H_1: \quad \beta_2 < 0$

$$t_{qs} = \frac{b_2 - 0}{s_{b_2}} = \frac{-0.000249}{0.0000320} = -7.78 < -t_\epsilon = -2.81876,$$

ta bác bỏ $H_0: \quad \beta_2 = 0$ ở mức 0.5%.

Ta có thể kiểm định Bài toán (2): giả thiết

$$H_0: \quad \beta_i = \beta_{i,0} \qquad \textit{hoặc} \qquad H_0: \quad \beta_i \geq \beta_{i,0}$$

với đối thiết

$$H_1: \qquad \beta_i < \beta_{i,0},$$

$$\textit{theo quy tắc bác bỏ} \quad H_0 \quad \textit{nếu} \qquad t_{qs} = \frac{\beta_i - \beta_{i,0}}{s_{b_i}} < -t_\epsilon.$$

Bài toán (3):

$$H_0: \quad \beta_i = \beta_{i,0}$$

với đối thiết

$$H_1: \qquad \beta_i \neq \beta_{i,0},$$

$$\textit{theo quy tắc bác bỏ} \quad H_0 \quad \textit{nếu} \qquad |t_{qs}| = \left|\frac{\beta_i - \beta_{i,0}}{s_{b_i}}\right| > t_{\epsilon/2}.$$

Kiểm định giả thiết đồng thời cho các tham số của hồi quy.

$$H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

với đối thiết

$$H_1: \qquad \text{Tồn tại ít nhất một } i: \beta_i \neq 0.$$

Nếu giả thiết H_0 đúng, $Y_i = \alpha + \varepsilon_i$, nên $E(Y_i/X) = \alpha$ là hằng số. Các biến độc lập X_i không có ảnh hưởng (tuyến tính) tới Y. Kiểm định giả thiết H_0 thực chất nhằm bác bỏ tính phụ thuộc tuyến tính giữa các biến. Ta biết rằng $SST = SSR + SSE$, trong đó SSR nhằm giải thích sự biến động của hồi quy (sự phụ thuộc tuyến tính của biến phụ thuộc vào các biến độc lập), còn SSE là phần biến động ngoài hồi quy. Do vậy nếu giữa các biến ngẫu nhiên không tồn tại quan hệ tuyến tính khi đó SSR tương đối nhỏ so với SSE, nói cách khác tỉ số giữa SSR và SSE càng lớn, khả năng bác bỏ giả thiết không (quan hệ tuyến tính) càng cao. Vì thế để tạo ra một thống kê như vậy người ta sử dụng kết quả sau:

Nếu giả thiết $H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_k = 0$ đúng và ε_i có phân bố chuẩn, khi đó

$$F = \frac{SSR/k}{SSE/(n-k-1)}$$

có phân bố F với $(k, n-k-1)$ bậc tự do. Vậy ta có quy tắc ở mức α

$$\text{Bác bỏ} \quad H_0 \quad \text{nếu} \qquad F_{qs} = \frac{SSR/k}{SSE/(n-k-1)} > F_{k,n-k-1,\alpha},$$

trong đó

$$P(F_{k,n-k-1} > F_{k,n-k-1,\alpha}) = \alpha.$$

Nhận xét rằng do $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$, *suy ra*

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2}.$$

Trở lại ví dụ lãi suất tiết kiệm và cho vay

$$F_{qs} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{0.40151122/2}{0.06250478/22} = 70.66057082$$

Với mức ý nghĩa 1%, $F_{k,n-k-1,\alpha} = 5.719$, nhỏ hơn rất nhiều so với $F_{qs} = 70.66057082$, ta bác bỏ giả thiết H_0.

Kiểm định giả thiết đồng thời cho một tập con các tham số của hồi quy.

Giả thiết rằng ta cần kiểm định k_1 tham số đầu tiên của hồi quy bằng 0.

$$H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_{k_1} = 0$$

$$\text{(Với đối thiết} \quad H_1: \quad \text{Tồn tại ít nhất một } i, 1 \leqslant i \leqslant k_1 : \beta_i \neq 0.)$$

Nếu giả thiết H_0 đúng, các biến $X_1, X_2, ..., X_{k_1}$ không có ảnh hưởng gì tới Y, do vậy ta tiến hành ước lượng hồi quy của Y chỉ thông qua các biến $X_{k_1+1}, X_{k_1+2}, ..., X_k$

$$Y_i = \alpha^* + \beta^*_{k_1+1} x_{k_1+1,i} + \cdots + \beta^*_k x_{ki} + \varepsilon^*_i$$

Khi đó ta hy vọng SSE của mẫu hồi quy cũ khác nhiều so với SSE^* của mẫu hồi quy mới.

Thống kê

$$F = \frac{(SSR^* - SSE)/k_1}{SSE/(n-k-1)}$$

có phân bố F với $(k_1, n-k-1)$ bậc tự do. Vậy ta có quy tắc ở mức α

$$\text{Bác bỏ} \quad H_0 \quad \text{nếu} \qquad F_{qs} = \frac{(SSE^* - SSE)/k_1}{SSE/(n-k-1)} > F_{k_1,n-k-1,\alpha}.$$

Dự báo.

Với mẫu hồi quy như đã nói ở trên, kí hiệu $a, b_1, b_2, ..., b_k$ là các ước lượng theo phương pháp bình phương bé nhất các hệ số hồi quy, khi đó với mẫu thứ $n+1$ của các biến độc lập:

$$(x_{1,n+1}, x_{2,n+1}, ..., x_{k,n+1})$$

dự báo của biến phụ thuộc $(Y_{n+1} = \alpha + \beta_1 x_{1,n+1} + \cdots + \beta_k x_{k,n+1} + \varepsilon_{n+1})$

$$\hat{Y}_{n+1} = a + b_1 x_{1,n+1} + b_2 x_{2,n+1} + \cdots + b_k x_{k,n+1}$$

là ước lượng tuyến tính không chệch tốt nhất của Y_{n+1}.

Trở lại ví dụ quen thuộc nếu $x_{1,n+1} = 4.50$ và số lượng các văn phòng $x_{2,n+1} = 9000$ khi đó

$$\hat{Y}_{n+1} = a + b_1 x_{1,n+1} + b_2 x_{2,n+1} = 0,39.$$

Ngoài ra nếu giả thiết ε_i *có phân bố chuẩn khi đó chúng ta có thể tính các khoảng tin cậy cho các dự báo* $\hat{Y}_{n+1}$.