CHƯƠNG 9. Tương quan và hồi quy tuyến tính đơn

7.1. Tương quan tuyến tính đơn
7.2. Hồi quy tuyến tính đơn
7.3. Một số mô hình phi tuyến có thể tuyến tính hoá

Bài 7.1. Tương quan tuyến tính đơn

1. Hệ số tương quan mẫu:

Giả sử X và Y là 2 BNN. Trong nhều trường hợp X và Y phụ thuộc lẫn nhau, ví dụ, GS X là chiều dài của bàn chân của 1 người và Y là chiều cao của người đó.

Để đo mức độ phụ thuộc tuyến tính giữa 2 BNN X và Y, người ta đưa ra khái niệm hệ số tương quan ρ:

$$\rho = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y}$$

Người ta đã chứng minh được $-1 \le \rho \le 1$.

Khi ρ=0 thì không có sự *tương quan tuyến tính* giữa X và Y. Đặc biệt khi (X, Y) có phân phối chuẩn đồng thời thì ρ=0 khi và chỉ khi X, Y độc lập. Ngược lại, khi |ρ| càng gần 1 thì sự phụ thuộc tuyến tính giữa X và Y càng mạnh.

Nếu |ρ|=1 thì Y là một hàm tuyến tính của X.

Ví dụ: Cho cặp BNN (X, Y) có hàm khối lượng xác suất đồng thời được cho bởi bảng sau:

Y X	-1	0	2	4	P_Y
1	0.08	0.12	0.10	0.05	0.35
2	0.06	0.10	0.14	0.10	0.40
3	0.05	0.09	0.06	0.05	0.25
P_X	0.19	0.31	0.3	0.2	1.00

μx	1.9	(бx)2	0.59
μy	1.21	(бy)2	3.1259

COV(X, Y)	0.071	ρ(X, Y)	0.052281

$$\hat{y}(x) = ax + b$$

$$b = \mu_y - a * \mu_x$$

$$a = \rho(X, Y) * \frac{\sigma_Y}{\sigma_X}$$

Muốn biết ρ chúng ta phải biết phân bố của tập chính bao gồm tất cả các giá trị của cặp (X, Y). Tuy nhiên, điều này là không thực tế.

Vì vậy, chúng ta có bài toán ước lượng và kiểm định hệ số tương quan ρ dựa vào mẫu ngẫu nhiên: (x_1, y_1), (x_2, y_2), …, (x_n, y_n) các giá trị của (X, Y).

Để ước lượng hệ số tương quan ρ, chúng ta sử dụng hệ số tương quan mẫu:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Chúng ta thường áp dụng công thức tính toán sau cho thuận lợi:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

Chú ý: $-1 \le r \le 1$

Ví dụ 1. Tính hệ số tương quan mẫu r dựa trên mẫu gồm 10 quan sát sau:

i	1	2	3	4	5	6	7	8	9	10
x_i	80	85	88	90	95	92	82	75	78	85
y_i	2.4	2.8	3.3	3.1	3.7	3	2.5	2.3	2.8	3.1

Giải

Cách 1. Tính trực tiếp

Đầu tiên tính các tổng

$$\sum x, \sum y, \sum xy, \sum x^2, \sum y^2$$

Và thay vào công thức tính r: $r = 0.858983$

Cách 2 : Dựa vào Excel

GS 10 giá trị của x_i được xếp vào các ô từ A1 đến J1, 10 giá trị của y_i được xếp vào các ô từ A2 đến J2. Khi đó, chỉ cần viết =CORREL(A1:J1,A2:J2), kết quả nhận được là 0.858983

Ví dụ 2

Giả sử ta có Danh sách điểm GT2 và Mạng MT của 10 SV như sau:

i	1	2	3	4	5	6	7	8	9	10
x_i	8.4	8	10	9	9	10	10	8	5.5	10
y_i	8.7	8.5	9	8.7	9.2	8	6.5	5.5	7	9.1

Câu hỏi: Liệu điểm GT 2 và MMT có không tương quan hay không?

Giải

r = 0.35

T= 1.071, α=5%

Ví dụ 3

GS ta có danh sách điểm – TT HCM và MMT:

i	1	2	3	4	5	6	7	8	9	10
x_i	6.6	6	7	7.2	7.2	7	7.5	6.3	5	8.2
y_i	8.7	8.5	9	8.7	9.2	8	6.5	5.5	7	9.1

Ví dụ 4

GS ta có danh sách điểm – MMT và NLHĐH:

i	1	2	3	4	5	6	7	8	9	10
x_i	8.7	8.5	9	8.7	9.2	8	6.5	5.5	7	9.1
y_i	8.5	8	8.5	8.2	8.6	8	8	6.5	7.5	6.9

Ví dụ 5

GS ta có danh sách điểm – GT2 và TRR:

i	1	2	3	4	5	6	7	8	9	10
x_i	8.4	8	10	9	9	10	10	8	5.5	10
y_i	5.8	7	9	9.6	8.8	8	6	5	5.5	8.4

Tiếp theo chúng ta đề cập đến bài toán kiểm định giả thiết về hệ số tương quan lý thuyết ρ.

Bài toán đầu tiên và quan trọng nhất là kiểm định xem X và Y có tương quan với nhau hay không.

2. Bài toán kiểm định giả thiết:

- ***Giả thiết H_0: $\rho=0$***
- ***Đối thiết H_1: $\rho\neq0$***

Tiêu chuẩn kiểm định được xây dựng dựa trên định lý sau:

Định lý: Nếu (X, Y) có phân bố chuẩn 2 chiều thì dưới giả thiết H_0, BNN

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Có phân bố Student với n-2 bậc tự do.

Với mức ý nghĩa α, ta sẽ bác bỏ H_0 nếu $|T|>t_{n-2}(\alpha/2)$.

Ví dụ: Trong một mẫu gồm 42 quan sát (x_i, y_i) rút ra từ tập hợp chính các giá trị của (X, Y), chúng ta tính được hệ số tương quan mẫu là r=0.22. Giả sử cặp BNN (X, Y) có phân phối chuẩn đồng thời. Với mức ý nghĩa α=5%, có thể kết luận rằng X và Y có tương quan hay không?

Giải

Ta có

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.22\sqrt{40}}{\sqrt{1-0.22^2}} = \frac{0.22}{0.154} = 1.43$$

Với bậc tự do 40, α=5% ta tra bảng

=TINV(0.05,40)=2.021075

So sánh, ta thấy |T|<2.021075, vì vậy chưa đủ cơ sở bác bỏ giả thiết H_0.

3. Với bài toán kiểm định giả thiết:

- ***Giả thiết H_0: $\rho=\rho_0$***
- ***Đối thiết H_1: $\rho\neq\rho_0$***

  ở đây ρ_0 là một giá trị khác 0 cho trước.

  Chúng ta sẽ xây dựng tiêu chuẩn thống kê

$$Z = \frac{u-m}{\sigma}.$$

  Trong đó:

$$u = \frac{1}{2}\ln\frac{1+r}{1-r}; m = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}; \sigma = \frac{1}{\sqrt{n-3}}$$

Người ta chứng minh được rằng nếu H_0 đúng, thì Z có phân bố xấp xỉ phân bố chuẩn tắc N(0,1). Do đó, H_0 sẽ bị bác bỏ ở mức ý nghĩa α nếu $|z|>z_{\alpha/2}$.

Ví dụ: Từ mẫu cỡ n=35 rút ra từ tập chính các giá trị của (X, Y), ta tính được hệ số tương quan là mẫu là r=0.8. Với mức ý nghĩa α=5%, kiểm định giả thiết:

- ***Giả thiết H_0: ρ= 0.9***
- ***Đối thiết H_1: $\rho\neq$ 0.9***

Giải

Ta có

$$u = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\ln\frac{1+0.8}{1-0.8} = 1.009;$$

$$m = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\ln\frac{1+0.9}{1-0.9} = 1.472;$$

$$\sigma = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{32}} = 0.177$$

Từ đó

$$z = \frac{u-m}{\sigma} = \frac{1.099-1.472}{0.177} = -2.11$$

Với α=5%, ta tìm được $z_{\alpha/2}$= 1.96.

Vì |T|=2.11> $u_{\alpha/2}$= 1.96, nên ta bác bỏ giả thiết H_0, chấp nhận đối thiết H_1, nghĩa là chấp nhận kết luận $\rho \neq 0.9$.

Ví dụ: Từ mẫu cỡ n=35 rút ra từ tập chính các giá trị của (X, Y), ta tính được hệ số tương quan là mẫu là r=0.8.

Với mức ý nghĩa α=5%, kiểm định giả thiết:

- ***Giả thiết H_0: $\rho = 0.85$***
- ***Đối thiết H_1: $\rho \neq 0.85$***

Giải

Ta có

$$u = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\ln\frac{1+0.8}{1-0.8} = 1.009;$$

$$m = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\ln\frac{1+0.85}{1-0.85} = 1.256;$$

$$\sigma = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{32}} = 0.177$$

Từ đó

$$z = \frac{u-m}{\sigma} = \frac{1.099-1.256}{0.177} = -0.887$$

Với α=5%, ta tìm được $z_{\alpha/2}$= 1.96.
Vì |T|=0.887< $z_{\alpha/2}$= 1.96, nên ta không đủ cơ sở bác bỏ giả thiết H_0, nghĩa là không chấp nhận kết luận $\rho \neq 0.85$.

Tiêu chuẩn thống kê $Z = \frac{u-m}{\sigma}$ cũng cho phép ta xác định được khoảng tin cậy cho hệ số tương quan lý thuyết ρ.
Ví dụ: Trong một mẫu có cỡ n=52 được rút ra từ tập hợp chính các giá trị của (X, Y), ta tính được hệ số tương quan mẫu là r=0.53. Căn cứ trên kết quả đó hãy xác định khoảng tin cậy 95% cho hệ số tương quan lý thuyết ρ giữa X và Y.
Giải
Ta có

$$u = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\ln\frac{1+0.53}{1-0.53} = 0.59;$$

$$\sigma = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{49}} = \frac{1}{7} = 0.143$$

Với α=5%, tra bảng ta có $u_{\alpha/2}$=1.96. Với xác suất 95% ta có:

$$-z_{\alpha/2}\sigma < u - m < z_{\alpha/2}\sigma$$

$$\Leftrightarrow u - z_{\alpha/2}\sigma < m < u + z_{\alpha/2}\sigma$$

Thay giá trị của

$u, u_{\alpha/2}, \sigma$ vào ta được

$$0.31 < m < 0.87$$

Hay

$$0.31 < \frac{1}{2}\ln\frac{1+\rho}{1-\rho} < 0.87$$

$$\Leftrightarrow 0.62 < \ln\frac{1+\rho}{1-\rho} < 1.74$$

$$\Leftrightarrow e^{0.62} < \frac{1+\rho}{1-\rho} < e^{1.74}$$

$$\Leftrightarrow 1.858 < \frac{1+\rho}{1-\rho} < 5.7$$

Giải bất đẳng thức trên ta tìm được:

$$0.3 < \rho < 0.7$$

Đây là khoảng tin cậy 95% cho ρ.

4. Kiểm tra tính độc lập

Giả sử ta có mẫu ngẫu nhiên cỡ n các quan sát đồng thời về hai biến ngẫu nhiên X và Y: $(x_1, y_1), (x_2,y_2), \ldots, (x_n, y_n)$.

Giả thiết H_0: X và Y độc lập với nhau

Đối thiết H_1: X và Y không độc lập.

- Ta ghép các giá trị mẫu $(x_1, x_2, \ldots, x_n)$ thành các khoảng, chẳng hạn r khoảng. Ghép các giá trị mẫu $(y_1, y_2, \ldots, y_n)$ thành s khoảng. Khi đó ta nhận được bảng

hai lối vào gồm rs ô chữ nhật con. Gọi (i, j) là ô ở hàng i cột j.

- Đếm số các quan sát từ mẫu đã cho rơi vào ô (i, j). Ký hiệu số đó là $n_{ij}, i = \overline{1,r}, j = \overline{1,s}$.

Nói cách khác n_{ij} là số các giá trị mẫu mà có giá trị mẫu theo X rơi vào khoảng thứ i và có giá trị mẫu theo Y rơi vào khoang thứ j.
Cần lưu ý rằng, các khoảng theo X và các khoảng theo Y không nhất thiết được phân chia theo định lượng, mà có thể theo định tính, chẳng hạn tốt, trung bình, xấu hoặc giỏi, khá, trung bình, kém hoặc màu xanh, đỏ, trắng, vàng, ...

- Tính

$$n_{i.} = \sum_{j=1}^{s} n_{ij} \text{ (lấy tổng theo hàng)}$$

$$n_{.j} = \sum_{i=1}^{r} n_{ij} \text{ (lấy tổng theo cột)}$$

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

- Đối với mỗi ô (i, j) ở trong bảng, ta tính $\frac{n_{i.} x n_{.j}}{n}$. Để tiện tính toán, ta đặt số này trong ô (i, j) cạnh số n_{ij}, nhưng ta đặt trong ngoặc.

- Tính

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right)$$

- Với α đã cho, tra bảng phân phối khi-bình phương (χ^2) với (r-1)(s-1) bậc tự do ta tìm được $\chi^2_{(r-1)(s-1)}(\alpha)$.

- Nếu $\chi^2 \geq \chi^2_{(r-1)(s-1)}(\alpha)$ ta bác bỏ tính độc lập của X và Y. (Thực chất tiêu chuẩn này là ứng dụng tiêu chuẩn phù hợp χ^2).

Trong thực hành ta hay sử dụng công thức :

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right)$$

Khi r=s=2 thì :

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right) = \frac{n\begin{vmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{vmatrix}}{n_{.1}n_{.2}n_{1.}n_{2.}}$$

Ví dụ : Ở các cây ngọc trâm lá có hai dạng, « lá phẳng » hoặc « lá nhăn », hoa có hai dạng, « hoa bình thường » hoặc « hoa hoàng hậu ».
Quan sát một mẫu gồm 560 cây ngọc trâm ta thu được kết quả sau :

Hoa Lá	Bình thường	Hoàng hậu	Tổng số
Phẳng	328	122	450
Nhăn	77	33	110
Tổng số	405	155	560

Có thể chấp nhận giả thiết hai đặc tính về hoa và lá nói trên là độc lập hay không ? Hay giữa chúng có sự tương quan ?

Giải
Ta có

$$\chi^2 = \frac{n\begin{vmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{vmatrix}}{n_{.1}n_{.2}n_{1.}n_{2.}} = \frac{560\begin{vmatrix} 328 & 122 \\ 77 & 33 \end{vmatrix}}{(450).(110).(405).(155)} = 0.368$$

Với mức ý nghĩa 5%, tra bảng phân phối χ^2 với 1 bậc tự do ta được $\chi_1^2(0.05) = 3.841$. Do $\chi^2 < \chi_1^2(0.05) = 3.841$, nên ta chấp nhận giả thiết H_0, chấp nhận giả thiết hai đặc tính về hoa và lá nói trên là độc lập.

Ví dụ : Giả sử X và Y tương ứng là số đo huyết áp và trọng lượng (tính bằng pound) (1pound=0.454 kg) của trẻ em 14 tuổi.

Để thuận tiện, số đo huyết áp X được chia thành các mức :

$B_1=\{X\leq 99\}$

$B_2=\{99<X\leq 110\}$

$B_3=\{110<X\leq 120\}$

$B_4=\{X>120\}$

Và Y chia làm 2 mức :

$A_1=\{Y\leq 102\}$

$A_2=\{Y>102\}$

Dựa vào mẫu ngẫu nhiên gồm 200 trẻ em được đo huyết áp và trọng lượng cho thấy số liệu sau :

Huyết áp Trọng lượng	B_1	B_2	B_3	**B_4**	Tổng số
A_1	10	20	11	5	46
A_2	6	48	50	50	154
Tổng số	16	68	61	55	200

Hãy kiểm định giả thiết về sự độc lập giữa trọng lượng và huyết áp của trẻ em.

Giải

Ta có :

$$\chi^2 = 200\left[\frac{10^2}{(16).(46)} + \frac{20^2}{(68).(46)} + ... + \frac{50^2}{(55).(154)} - 1\right] = 22.53$$

Với mức ý nghĩa α=1%, tra bảng phân phối χ^2 với bậc tự do là (2-1).(4-1)=3, ta tìm được $\chi_3^2(0.01) = 11.345$.

Vì $\chi^2 > \chi_3^2(0.01) = 11.345$ nên ta bác bỏ H_0 và kết luận :

Giữa huyết áp và trọng lượng trẻ 14 tuổi có sự phụ thuộc lẫn nhau.

Bài 7.2. Hồi quy tuyến tính đơn

Giả sử Y là đại lượng ngẫu nhiên phụ thuộc vào X (có thể là biến ngẫu nhiên hay không ngẫu nhiên). Nếu X=x thì Y sẽ có kỳ vọng là αx+β, với α, β là hằng số và phương sai là σ^2 (không phụ thuộc x). Khi đó ta nói Y có hồi quy tuyến tính theo X và đường thẳng y= αx+β được gọi là đường thẳng hồi quy lý thuyết của Y đối với X. Các hệ số α, β được gọi là hệ số hồi quy lý thuyết. X được gọi là biến độc lập. Y được gọi là biến phụ thuộc.

Bài toán đặt ra là ước lượng các hệ số hồi quy lý thuyết dựa trên mẫu quan sát (x_1, y_1), …, (x_n,y_n). Ước lượng α và β dựa trên phương pháp bình phương bé nhất. Các số a và b được dung làm ước lượng cho α và β nếu nó làm cực tiểu tổng

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

Ta tìm được

$$\frac{\partial Q}{\partial a} = -2\sum_{i=1}^{n} x_i(y_i - ax_i - b) = 0 \rightarrow a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \quad (1)$$

$$\frac{\partial Q}{\partial b} = -2\sum_{i=1}^{n} (y_i - ax_i - b) = 0 \quad \rightarrow a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i \quad (2)$$

Giải (1) và (2) ta được:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Hoặc $a = \hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$

$$b = \bar{y} - a\bar{x} = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Hoặc

$$b = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Trong đó a, b được gọi là hệ số hồi quy. Phương trình y=ax+b được gọi là đường hồi quy.

Ví dụ 11: Các số liệu về số trang của một cuốn sách (X) và giá bán của nó (Y) được cho bảng dưới đây :

Tên sách	X	Y(nghìn)
A	400	44
B	600	47
C	500	48
D	600	48
E	400	43
F	500	46

Hãy tìm đường thẳng hồi quy của Y theo X căn cứ trên số liệu nói trên.

Giải

Ta có :

$$\sum xy = 138800$$
$$\sum x = 3000$$
$$\sum y = 276$$
$$\sum x^2 = 1540000$$
$$\sum y^2 = 12718$$

Từ đó :

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{6(138800) - (3000)(276)}{6(1540000) - (3000)^2}$$

$$= \frac{4800}{240000} = 0.02$$

$$b = \bar{y} - a\bar{x} = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{276 - (0.02).(3000)}{6} = 36$$

Vậy đường hồi quy là : y=0.02x+36.

Ngoài việc ước lượng hệ số hồi quy a và b, ta còn ước lượng đại lượng đo sự phân tán của Y xung quanh đường

thẳng hồi quy, ký hiệu là $S^2_{Y,X}$ và được xác định theo công thức sau :

$$S^2_{Y,X} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ax_i - b_i)^2 = \frac{\sum y^2 - a\sum xy - b\sum y}{n-2}$$

Đại lượng $S_{Y,X}$ được gọi là sai số tiêu chuẩn của đường hồi quy, nó cho ta số đo sự phân tán của đám mây điểm (x_i, y_i) xung quanh đường thẳng hồi quy.

Ví dụ 12. Hãy tính sai số tiêu chuẩn của đường hồi quy $S_{Y,X}$ trong ví dụ 11 vừa nêu.

Giải

$$S^2_{Y,X} = \frac{\sum y^2 - a\sum xy - b\sum y}{n-2}$$

$$= \frac{12718 - (0.02).(1388000) - 36(276)}{6-2} = 1.5$$

Vậy $S_{Y,X} = \sqrt{1.5} = 1.22$

Dựa trên phương trình đường thẳng hồi quy tìm được, ta có thể dự báo được giá trị của Y nếu biết giá trị của X. Giá trị được dự báo của Y khi $X=x_0$ sẽ là :

$$\hat{y}_0 = ax_0 + b$$

Đây đồng thời cũng là giá trị được dự báo cho kỳ vọng của Y ứng với $X=x_0$ (ký hiệu là μ_{x_0}) : $\hat{\mu}_{x_0} = ax_0 + b$.

Sau đây, chúng ta xét bài toán tìm khoảng tin cậy cho giá trị dự báo của Y, cũng như khoảng tin cậy cho giá trị dự báo của μ_{x_0} .

+ Công thức tính khoảng tin cậy γ cho giá trị dự báo của Y khi X=x_0 :

$$\hat{y}_0 \pm t_{n-2}(\alpha/2) S_{Y,X} \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})}{\sum x^2-\frac{(\sum x)^2}{n}}}$$

+ Công thức tính khoảng tin cậy γ cho giá trị dự báo của μ_{x_0} là:

$$\hat{y}_0 \pm t_{n-2}(\alpha/2) S_{Y,X} \sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})}{\sum x^2-\frac{(\sum x)^2}{n}}}$$

Ví dụ 13. Với số liệu trong VD 11, hãy dự báo về giá bán của một cuốn sách với 450 trang.

Giải

Theo phương trình hồi quy : y=0.02x+36, giá cuốn sách đó được dự báo là : $\hat{y} = 0.02(450) + 36 = 45$ (nghìn).

Khoảng tin cậy 95% cho giá của một cuốn sách 450 trang là :

$$45 \pm t_{6-2}(0.025)(1.22)\sqrt{1+\frac{1}{6}+\frac{(450-500)}{154000-\frac{(3000)^2}{6}}}$$

$$= 45 \pm 3.77$$

Với $t_4(0.025) = 2.776$.

Vậy khoảng tin cậy cần tìm là : $41.23 < y_0 < 48.77$.

Vậy, với độ tin cậy 95%, cuốn sách với 450 trang sẽ được bán với giá trong khoảng từ 41230 đồng đến 48770 đồng.

Ví dụ 14. Với số liệu trong VD 13, chúng ta muốn dự báo giá bán trung bình của tất cả các cuốn sách 450 trang.

Giải

Giá trung bình của dự báo là : $\hat{\mu} = 0.02(450) + 36 = 45$ nghìn.

Khoảng tin cậy 95% cho giá trung bình của tất cả các cuốn sách 450 trang là :

$$45 \pm (2.776)(1.22)\sqrt{\frac{1}{6} + \frac{(450-500)}{154000 - \frac{(3000)^2}{6}}}$$

$$= 45 \pm 3.4\sqrt{0.23} = 45 \pm 1.63$$

Hay $43.37 < \mu < 46.63$.

Vậy với độ tin cậy 95% giá trung bình của tất cả các cuốn sách 450 trang sẽ nằm trong khoảng từ 43370 đồng đến 46630 đồng.

Một vấn đề quan trọng là phải kiểm tra xem hệ số hồi quy lý thuyết có bằng 0 hay không. Nếu hệ số hồi quy lý thuyết bằng 0 thì E(Y)=β- là một hằng số, không phụ thuộc vào X.

Người ta đã chứng minh được rằng hệ số hồi quy mẫu a có độ lệch tiêu chuẩn là :

$$S_a = \frac{S_{Y,X}}{S_X\sqrt{n-1}} = \frac{S_{Y,X}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

Thống kê

$$T = \frac{a}{S_a}$$

Sẽ có phân bố Student với n-2 bậc tự do, nếu giả thiết H_0: $\alpha=0$ là đúng. Vì vậy, giả thiết H_0 sẽ bị bác bỏ ở mức ý nghĩa α nếu $|T|>t_{n-2}(\alpha/2)$.

Ví dụ 15. Với mức ý nghĩa $\alpha=5\%$, hãy kiểm định giả thiết H_0: « Hệ số góc α của đường thẳng hồi quy lý thuyết của Y đối với X bằng 0 », ở đây X và Y là 2 biến xét trong VD 11.

Giải

Ta có

$$S_a = \frac{S_{Y,X}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}} = \frac{1.22}{\sqrt{1540000 - \frac{(3000)^2}{6}}} = \frac{1.225}{200} = 0.0061.$$

Vậy T=0.02/0.0061=3.33.

Với mức ý nghĩa $\alpha=5\%$, tra bảng phân phối Student ta tìm được $t_4(0.025)=2.776$.

Ta có $|T|> t_4(0.025)=2.776$, do đó ta bác bỏ H_0.

Vậy ta chấp nhận giả thiết hệ số góc α của đường thẳng hồi quy lý thuyết của Y đối với X là khác 0.

Bài 7.3. Phân tích tương quan phi tuyến

Như ta đã biết, hệ số tương quan dùng để đo mức độ phụ thuộc tuyến tính giữa hai BNN. Như vậy, chúng ta chưa có một chỉ tiêu để đo mức độ phụ thuộc nói chung. Vì khi hệ số tương quan giữa X và Y rất bé, hay thậm chí bằng 0 thì giữa X và Y vẫn có thể có 1 mối liên hệ phi tuyến rất chặt chẽ.
Để đo mức độ phụ thuộc nói chung của BNN Y vào BNN X, người ta đưa ra khái niệm tỷ số tương quan. Tỷ số tương quan lý thuyết của Y theo X được ký hiệu bởi $\eta^2_{Y/X}$ và được xác định theo công thức sau:

$$\eta^2_{Y/X} = 1 - \frac{E(Y - E(Y/X))^2}{V(Y)}$$

$$= \frac{V(Y) - E(Y - E(Y/X))^2}{V(Y)}$$

Trong đó E(Y/X) ký hiệu kỳ vọng của Y tính trong điều kiện X cố định một giá trị. Đại lượng E(Y/X) được gọi là kỳ vọng có điều kiện của Y với điều kiên X.
Người ta chứng minhđược rằng :

$$0 \le \eta^2_{Y/X} \le 1; \rho^2 \le \eta^2_{Y/X}$$

Hiệu số

$$\eta^2_{Y/X} - \rho^2$$

Đo mức độ phụ thuộc phi tuyến giữa Y và X.

Chúng ta xét vấn đề ước lượng và kiểm định giả thiết về tỷ số tương quan.
GS $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ là một mẫu gồm n quan sát độc lập rút ra từ tập chính các giá trị của (X, Y). Chúng ta cần giả thiết rằng trong dãy các giá trị của X : $x_1, x_2, \ldots, x_n$, mỗi giá trị x_i đều được lặp lại ít nhất 1 lần. Giả sử $x_{(1)}<x_{(2)}<\ldots< x_{(k)}$ là các giá trị khác nhau trong dãy (x_i). Ta sẽ trình bày dãy số liệu (x_i, y_i) thành bảng sau đây, gọi là bảng tương quan :

X Y	$x_{(1)}$	$x_{(2)}$	…	$x_{(k)}$	
	y_{11} y_{21} … $y_{n_1 1}$	y_{12} y_{22} … $y_{n_2 2}$	… … … …	y_{1k} y_{2k} … $y_{n_k k}$	
	n_1	n_2	…	n_k	$n = \sum n_i$
	T_1	T_2	…	T_k	$T = \sum T_i$

Ký hiệu :

$$T_i = \sum_{j=1}^{n_i} y_{ji}$$

$$T = \sum T_i$$

+ Tổng bình phương chung SST :

$$SST = \sum\sum y_{ji}^2 - \frac{T^2}{n}$$

+ Tổng bình phương do nhân tố SSF :

$$SSF = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{n}$$

Đại lượng sau được dùng để ước lượng cho tỷ số tương quan lý thuyết :

$$\hat{\eta}^2_{Y/X} = \frac{SSF}{SST}$$

Đại lượng

$$\hat{\eta}^2_{Y/X} = \frac{SSF}{SST}$$

Được gọi là tỷ số tương quan mẫu của Y đối với X. Để cho gọn ta sẽ viết $\hat{\eta}^2$ thay cho $\hat{\eta}^2_{Y/X}$.

Người ta đã chứng minh được rằng :

$$0 \le r^2 \le \hat{\eta}^2$$

Bình phương của hệ số tương quan r^2 được gọi là hệ số xác định.

Ví dụ 9: Cho mẫu quan sát sau đây của cặp BNN (X, Y) :

X	8	8	12	12	20	20	24	24	8	8
Y	82	78	65	50	60	47	52	41	87	58

X	8	12	12	12	20	20	20	24	24	24
Y	70	62	55	52	44	66	41	57	50	47

X	8	12	20	24
Y	65	49	57	65

Hãy tính hệ số tương quan, hệ số xác định và tỷ số tương quan mẫu của Y đối với X.

Giải

Trước hết, ta cần trình bày các số liệu trên dưới dạng bảng tương quan sau đây :

X Y	8	12	20	24	
	82 78 87	65 50 62	60 47 44	52 41 57	

	58 70 65	55 52 49	66 41 57	50 47 63	
n_i	6	6	6	6	n=24
T_i	440	333	315	310	T=1398

+ Tính hệ số tương quan :

Ta có :

$$\sum x = 6(8) + 6(12) + 6(20) + 6(24) = 384$$

$$\sum y = T = 1398$$

$$\sum x^2 = 6(8^2) + 6(12^2) + 6(20^2) + 6(24^2) = 7104$$

$$\sum y^2 = 82^2 + 78^2 + ... + 63^2 = 84908$$

$$\sum xy = 8.(440) + 12.(333) + 20.(315) + 24.(310) = 21256$$

Vậy

$$n\sum xy - (\sum x)(\sum y) = -26688$$

$$\sqrt{n\sum x^2 - (\sum x)^2} = \sqrt{24(7104) - 384^2} = 151.789$$

$$\sqrt{n\sum y^2 - (\sum y)^2} = \sqrt{24(84909) - 1398^2} = 288.77$$

Kết quả :

$$r = \frac{-26688}{(151.789).(288.77)} = 0.6089$$

Hệ số xác định

$$r^2 = 0.6089^2 = 0.37.$$

Tính tỷ số tương quan :

Ta có :

$$SST = \sum\sum y_{ji}^2 - \frac{T^2}{n} = 84908 - \frac{1398^2}{24} = 3474.5$$

$$SSF = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{n} = \frac{440^2 + ... + 310^2}{36} - \frac{1398^2}{24} = 1868.83$$

Từ đó :

$$\hat{\eta}^2_{Y/X} = \frac{SSF}{SST} = \frac{1868.83}{3474.5} = 0.5378$$

Hiệu số $\eta^2 - \rho^2$ giữa tỷ số tương quan lý thuyết và hệ số xác định lý thuyết cho ta hình ảnh về sự phụ thuộc phi tuyến của Y đối với X. Nếu hiệu số đó bằng 0 thì điều đó có nghĩa là chỉ có tương quan tuyến tính giữa Y và X.
Để giải BT kiểm định giả thiết :

- Giả thiết H_0 : $\eta^2 - \rho^2 = 0$ (Không có tương quan phi tuyến)
- Đối thiết H_1 : $\eta^2 - \rho^2 > 0$ (Có tương quan phi tuyến)
  Ta dùng tiêu chuẩn thống kê sau :

$$F = \frac{\frac{\hat{\eta}^2 - r^2}{k-2}}{\frac{1-\hat{\eta}^2}{n-k}} = \frac{(\hat{\eta}^2 - r^2)(n-k)}{(1-\hat{\eta}^2)(k-2)}$$

  Người ta chứng minh được rằng, nếu H_0 đúng thì F sẽ có phân bố Fisher với bậc tự do là (k-2, n-k). Khi đó, giả thiết H_0 bị bác bỏ với mức ý nghĩa α nếu F lớn hơn α phân vị của phân phối Fisher với bậc tự do là (k-2 ; n-k).

Ví dụ 10. Với số liệu trong VD 9, kiểm tra xem liệu có tương quan phi tuyến của Y đối với X hay không ?

Giải

Ta có

$$T = \frac{(0.5378 - 0.37)}{1 - 0.5378} \cdot \frac{(24 - 4)}{(4 - 2)} = 3.63$$

Tra bảng phân phối Fisher với α=5% phân vị và (2 ; 20) bậc tự do, ta được : 3.49.

Vì F>3.49, nên ta bác bỏ H_0. Vậy ta khẳng định, có mối tương quan phi tuyến của Y đối với X. Xác suất sai của khẳng định này là 5%.

Ví dụ

Hãy tính hệ số tương quan mẫu của bộ số liệu sau

Toán	9.4	9.2	9.1	8.5	9.4	8.6	8.1	8.8	8.5
Văn	8.2	5.7	7.9	7	6.5	7.5	5.5	6.5	7.5

Toán	7.9	9.2	7.5	9.4	9.2	9.1	8.8	9.2	9.0
Văn	78	6.9	5.4	7.4	7.2	7.5	7	7.1	7