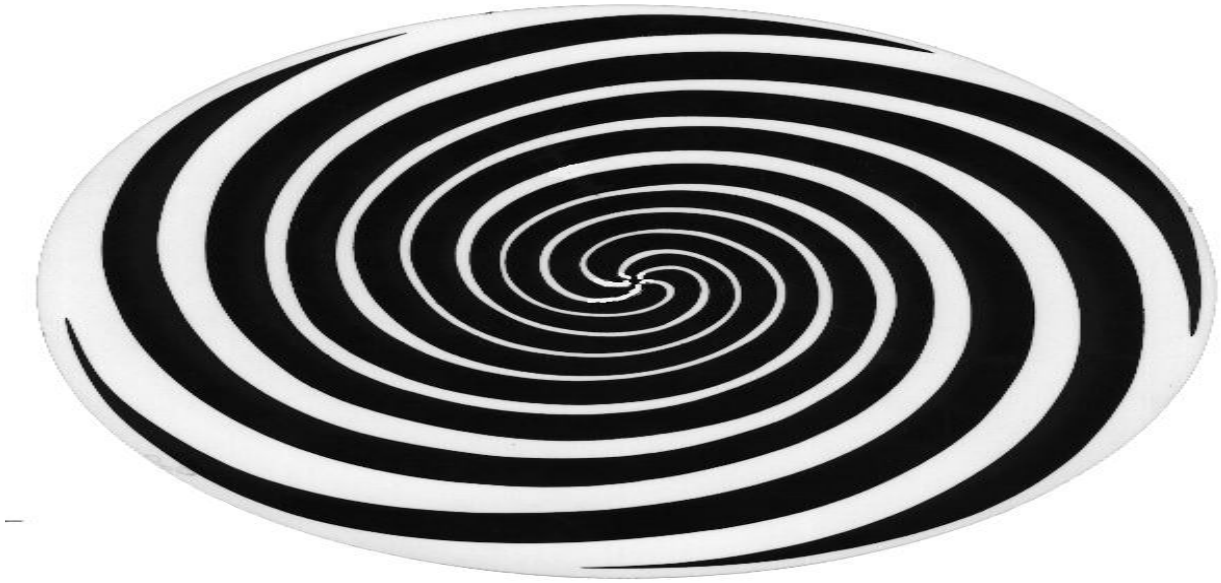
BÀI 8: TƯƠNG QUAN VÀ HỒI QUY



Nội dung

- Hệ số tương quan mẫu
- Đường hồi quy trung bình tuyến tính thực nghiệm

Thời lượng

- 4 tiết

Mục tiêu

- Giới thiệu hệ số tương quan mẫu và đường hồi quy bình phương trung bình tuyến tính thực nghiệm của hai biến ngẫu nhiên.
- Kiến thức nền quan trọng cho sinh viên tiếp thu kiến thức môn học Kinh tế lượng sau này.

Hướng dẫn học

Trong chương này các bạn cần nắm vững những kiến thức sau:

- Các khái niệm về hệ số tương quan giữa hai biến định lượng và ý nghĩa, các tính chất của hệ số tương quan, cách dùng hệ số tương quan để đánh giá mối quan hệ giữa hai đại lượng.
- Phương trình hồi quy tuyến tính đơn, đường hồi quy thực nghiệm, ý nghĩa của phương trình hồi quy và các hệ số hồi quy, tính chất của đường hồi quy thực nghiệm.
- Phương pháp sai số bình phương bé nhất, cách tính các hệ số hồi quy, cách dùng phương trình hồi quy để dự báo giá trị của biến phụ thuộc theo giá trị mới của biến giải thích.

Cần xem kỹ các ví dụ trong mỗi bài học và làm các bài tập của các phần tương ứng.

TÌNH HUỐNG KHỞI ĐỘNG BÀI

Tình huống

Siêu thị ABC muốn mở 01 siêu thị tại khu dân cư Vạn Phúc. Để xác định được quy mô siêu thị, doanh nghiệp cần biết được chi phí nhu yếu phẩm của người dân trong vùng. Biết chi phí nhu yếu phẩm của 01 cá nhân phụ thuộc chính vào mức thu nhập của cá nhân đó. Siêu thị tiến hành điều tra mức thu nhập (X) và chi tiêu (Y) cho những nhu cầu yếu phẩm của cá nhân. Kết quả cho bảng số liệu sau (đơn vị triệu đồng):

X\Y	0,5	0,8	1,0
1,5	4	3	0
2,0	6	2	1
2,5	2	5	2
3,0	1	1	4

Câu hỏi

1. Tính hệ số tương quan mẫu

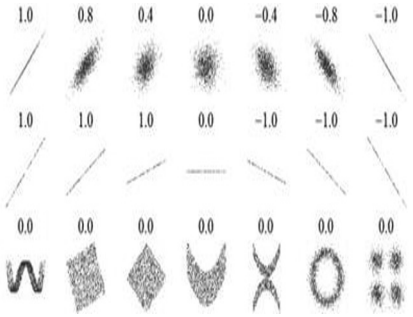
2. Viết phương trình hồi quy tuyến tính mẫu

3. Ước lượng sai số hồi quy

4. Dự báo giá trị của Y khi mức thu nhập X là 4,0 triệu đồng

8.1. Hệ số tương quan mẫu

Trong bài trước chúng ta đã đưa ra khái niệm hệ số tương quan mẫu và cách tính giá trị của hệ số tương quan mẫu ứng với mẫu cụ thể. Trong phần xác suất ta đã biết hệ số tương quan của hai biến ngẫu nhiên X, Y.



$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - (EX)E(Y)}{\sigma_X \sigma_Y}.$$

Ý nghĩa của hệ số tương quan là đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên X và Y. Trong thực tế nhiều khi ta chưa biết về phân phối của X và Y, do đó hệ số tương quan lý thuyết ρ giữa X và Y cũng chưa biết. Vậy ta cần dựa vào mẫu quan sát về véc tơ ngẫu nhiên (X, Y) để tìm cách ước lượng cho hệ số tương quan ρ.

Giả sử ta có mẫu ngẫu nhiên $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ rút ra từ véc tơ ngẫu nhiên (X, Y) với giá trị mẫu $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Hệ số tương quan mẫu được định nghĩa qua công thức:

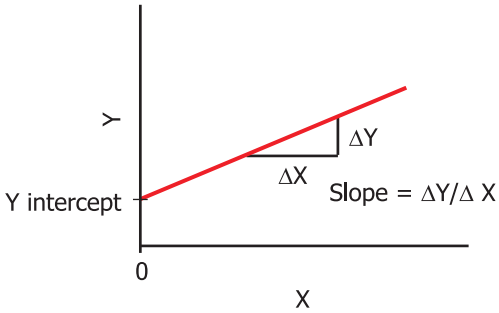
$$R = \frac{\overline{XY} - (\overline{X})(\overline{Y})}{S_X S_Y}.$$

trong đó thống kê $\overline{XY} = \frac{1}{n}\sum_{k=1}^{n} X_k Y_k$. Lúc đó R là ước lượng của hệ số tương quan lý thuyết ρ. Với mẫu cụ thể, giá trị của R là:

$$r = \frac{\overline{xy} - (\overline{x})(\overline{y})}{s_X s_Y}.$$

8.2. Đường hồi quy trung bình tuyến tính thực nghiệm

Ta có thể kiểm tra thấy rằng hệ số tương quan ρ và ước lượng r của nó đều là các đại lượng có giá trị tuyệt đối nhỏ hơn hoặc bằng 1. Khi $|\rho|$ càng gần 1 thì mức độ phụ thuộc tuyến tính giữa X và Y càng chặt chẽ, tức là ta có thể tính xấp xỉ Y theo *X* qua biểu thức dạng f(X) = aX + b . Thông thường khi $|\rho| > 0,8$ thì cách tính xấp xỉ đó được gọi là chặt chẽ.



Lúc đó ta có thể biểu diễn giá trị của Y qua giá trị của X bằng phương trình dạng

$$Y = aX + b + \varepsilon,$$

trong đó ε là sai số của phép lấy xấp xỉ.

Phương trình Y = aX + b được gọi là phương trình hồi quy tuyến tính của Y theo X, ε là sai số hồi quy. Hệ số *a* được gọi là độ dốc (slope), cho biết khi biến X tăng một

đơn vị thì giá trị của biến Y sẽ tăng hay giảm bao nhiêu đơn vị. Hệ số b được gọi là hằng số hồi quy (intercept), cho biết phương trình hồi quy có đi qua gốc tọa độ hay không và điểm xuất phát của Y khi X bằng 0 sẽ là bao nhiêu. Các hệ số a và b cũng được gọi là hệ số hồi quy.

Trong phương trình trên biến Y được gọi là biến được giải thích hay biến phụ thuộc, biến X được gọi là biến giải thích hoặc là biến độc lập, phương trình hồi quy được gọi là phương trình hồi quy tuyến tính đơn. Nếu biến Y được biểu diễn qua một phương trình dạng tuyến tính với nhiều hơn một biến giải thích thì phương trình được gọi là hồi quy tuyến tính bội.

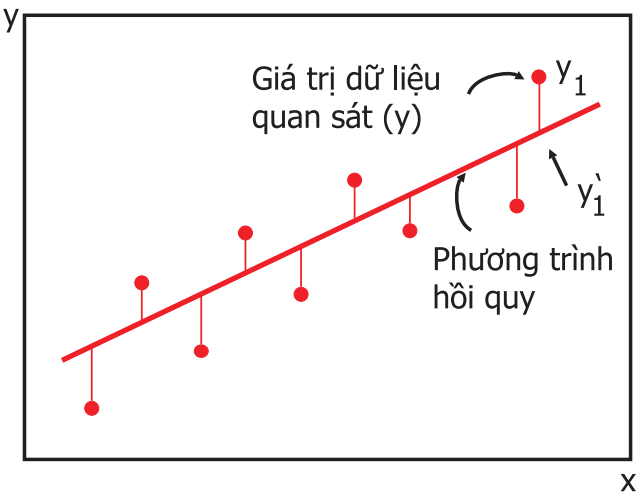
Với mẫu ngẫu nhiên (x_1, y_1), (x_2, y_2), …, (x_n, y_n) ta xây dựng hàm hồi quy mẫu (hàm hồi quy thực nghiệm) có dạng:

$$\hat{y} = \hat{a}x + \hat{b}.$$

Xét tại quan sát thứ i ta có

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$
$$y_i = \hat{y}_i + \varepsilon_i = \hat{a}.x_i + \hat{b} + \varepsilon_i$$

trong đó ε_i là giá trị của ε tại quan sát thứ i, $\hat{a}, \hat{b}$ là các ước lượng của a và b. Ta cần xác định các hệ số $\hat{a}, \hat{b}$ sao cho tổng sai số bình phương trung bình đạt giá trị nhỏ nhất.



$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{a}.x_i - \hat{b})^2 \to \min$$

Điều kiện cần để hàm L đạt min là:

$$\begin{cases} \dfrac{\partial L}{\partial \hat{b}} = -2 \sum\limits_{i=1}^{n} (y_i - \hat{a}.x_i - \hat{b}) = 0 \\ \dfrac{\partial L}{\partial \hat{a}} = -2 \sum\limits_{i=1}^{n} (y_i - \hat{a}.x_i - \hat{b})x_i = 0 \end{cases}$$

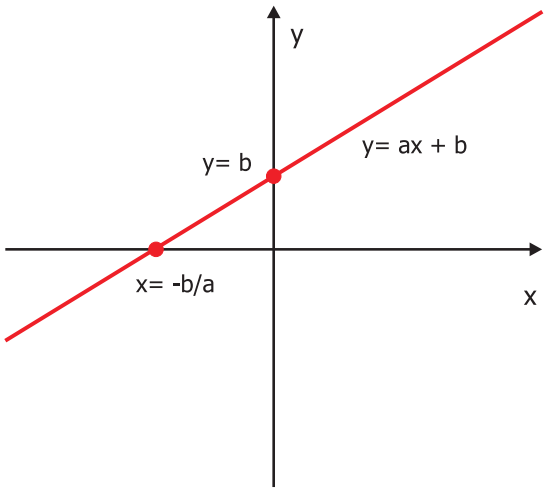
Giải hệ phương trình, ta có:

$$\begin{cases} n.\hat{b} + \hat{a}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{b}\sum_{i=1}^{n} x_i + \hat{a}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Từ đó ta có nghiệm

$$\hat{b} = \overline{y} - \hat{a}.\overline{x}$$

$$\hat{a} = \frac{\sum_{i=1}^{n} x_i y_i - (\overline{x})(\overline{y})}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2} = r\frac{S_y}{S_x}.$$



Do đó, ta có:

$$\hat{y}_i = r\frac{S_y}{S_x}x_i + \overline{y} - r\frac{S_y}{S_x}\overline{x}$$

$$\Rightarrow \hat{y}_i - \overline{y} = r\frac{S_y}{S_x}(x_i - \overline{x}).$$

Vậy phương trình hồi quy mẫu (đường hồi quy trung bình tuyến tính thực nghiệm) có dạng:

$$\hat{y} = r\frac{s_y}{s_x}x + \overline{y} - r\frac{s_y}{s_x}\overline{x} = \hat{a}x + \hat{b}.$$

Phương pháp ước lượng cho hệ số a, b như trên còn được gọi là phương pháp bình phương nhỏ nhất. Phương trình hồi quy xác định như trên có một số tính chất sau:

- Hàm hồi quy mẫu đi qua điểm $(\overline{x}, \overline{y})$.
- Các ước lượng $\hat{a}$, $\hat{b}$ được xác định duy nhất.
- Giá trị trung bình các sai số bằng 0:

$$\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i = 0.$$

- Giá trị trung bình của $\hat{y}_i$ bằng giá trị trung bình của các quan sát y_i:

$$\overline{\hat{y}} = \overline{y}.$$

Sai số $\varepsilon_i = y_i - \hat{y}_i$ cũng được gọi là phần dư (residual) nó biểu thị sự sai khác giữa quan sát y_i và giá trị $\hat{y}_i$ tính được từ phương trình hồi quy $\hat{y} = \hat{a}x + \hat{b}$.

Đặt:

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2.$$

RSS được gọi là tổng bình phương các phần dư và ta có

$$RSS = ns_y^2(1-r^2).$$

Ký hiệu $\varepsilon_{y/x}^2$ là ước lượng của sai số ε, ta có

$$\varepsilon_{y/x}^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - (\bar{\varepsilon}) = \frac{RSS}{n} = s_y^2(1-r^2).$$

Các đại lượng trên sẽ được dùng để đánh giá chất lượng của mô hình hồi quy.

Một trong những ứng dụng quan trọng của hồi quy là dự báo giá trị của biến Y khi biến X nhận giá trị mới. Nếu khi biến X nhận giá trị mới là x_0, khi đó ta có:

$$\hat{y}_0 = \hat{a}x_0 + \hat{b}$$

$\hat{y}_0$ là ước lượng điểm tương ứng cho giá trị y_0 của Y.

Ví dụ 1:

Theo dõi mức lãi suất (Y) và tỷ lệ lạm phát (X) ở một số nước ta có số liệu sau:

Y	17,5	15,6	9,8	5,3	7,9	10,0	19,2	13,1
X	14,2	11,7	6,4	2,1	4,8	8,1	15,4	9,8

Hãy:

- Tính hệ số tương quan mẫu.
- Xây dựng phương trình hồi quy mẫu.
- Ước lượng sai số hồi quy.
- Dự báo giá trị của mức lãi suất nếu tỷ lệ lạm phát là 22,5.

Giải:

- Với số liệu mẫu ta tính được

$$\bar{x} = 9,0625;\ \bar{y} = 12,3;\ \overline{xy} = 130,9813;$$
$$s_x^2 = 18,59;\ s_x = 4,312$$
$$s_y^2 = 20,76;\ s = 4,56.$$

Vậy hệ số tương quan mẫu sẽ là

$$r = \frac{130,9813 - 9,0625.12,3}{4,56.4,312} = 0,99.$$

Ta có:

$$\hat{a} = r\frac{s_y}{s_x} = 0,99.\frac{4,56}{4,312} = 1,045$$
$$\hat{b} = \bar{y} - \hat{a}.\bar{x} = 12,3 - 1,045.9,0625 = 2,83.$$

- Vậy ta có phương trình hồi quy mẫu

$$\hat{y} = 1,045.x + 2,83.$$

- Ước lượng của sai số hồi quy là:

$$\varepsilon^2_{y/x} = s^2_y(1-r^2) = 20,76(1-0,99^2) = 0,413.$$

- Nếu tỷ lệ lạm phát $x_0 = 22,5$ thì mức lãi suất ngân hàng sẽ là:

$$y_0 = 1,045.22,5 + 2,83 = 26,343.$$

Ví dụ 2:

Điều tra mức thu nhập (X) và chi tiêu (Y) cho những nhu cầu yếu phẩm của cá nhân ta có bảng số liệu sau (đơn vị triệu đồng)

X \ Y	0,5	0,8	1,0
1,5	4	3	0
2,0	6	2	1
2,5	2	5	2
3,0	1	1	4

Hãy

- Tính hệ số tương quan mẫu.
- Viết phương trình hồi quy tuyến tính mẫu.
- Ước lượng sai số hồi quy.
- Dự báo giá trị của Y khi mức thu nhập X là 4,0 triệu đồng.

Giải:

- Với số liệu đã cho tính toán như ví dụ trong bài 6 ta có:

$$\overline{x} = 2,23;\ \overline{y} = 0,72;\ \overline{xy} = 1,695;$$
$$s^2_x = 0,27;\ s_x = 0,52;$$
$$s^2_y = 0,04;\ s_y = 0,2.$$

Vậy, hệ số tương quan mẫu sẽ là:

$$r = \frac{1,695 - 2,23.0,72}{0,52.0,2} = 0,86.$$

Ta có:

$$\hat{a} = r\frac{s_y}{s_x} = 0,86.\frac{0,2}{0,52} = 0.33$$

$$\hat{b} = \overline{y} - \hat{a}.\overline{x} = 0,72 - 0,33.2,23 = -0,016.$$

- Phương trình hồi quy mẫu:

$$\hat{y} = 0,33.x - 0,016.$$

- Ước lượng sai số hồi quy:

$$\varepsilon^2_{y/x} = s^2_y(1-r^2) = 0,04(1-0,86^2) = 0,01.$$

- Dự báo giá trị của Y khi giá trị của X là $x_0 = 4,0$. Ta có $y_0 = 0,33.4 - 0,016 = 1,3.$

TÓM LƯỢC CUỐI BÀI

Bài này nghiên cứu phương pháp tương quan hồi quy, có tác dụng giải quyết rất nhiều bài toán trong khoa học, kỹ thuật, kinh tế đòi hỏi phai khai thác mối quan hệ giữa hai biến hay nhiều hơn hai biến. Bài này chỉ nghiên cứu hồi quy đơn, phần mở rộng hơn và những ứng dụng của hồi quy sẽ được đề cập tiếp trong phần kinh tế lượng.

BÀI TẬP TRẮC NGHIỆM

1. Từ mẫu thực nghiệm (x_1, x_2, …, x_{25}) rút ra từ biến ngẫu nhiên X ta tính được:

$$s' = 2,4 \text{ và } \sum_{k=1}^{25} x_k^2 = 165.$$

Trung bình mẫu $\overline{x}$ sẽ có giá trị bằng:

a. 1,036

b. 1,035

c. 1,034

d. 1,033

e. 1,045

2. Cho mẫu ngẫu nhiên (X_1, X_2, …,X_{25}) rút ra từ biến ngẫu nhiên X có phân phối chuẩn $N(\mu, \sigma^2)$. Với mức ý nghĩa 95% ta có khoảng ước lượng cho kỳ vọng là (16,75; 17,02). Khi đó giá trị trung bình mẫu $\overline{x}$ là:

a. 16,885

b. 17,885

c. 18,885

d. 19,885

e. không xác định được

3. Cho hai biến ngẫu nhiên X và Y có mẫu ngẫu nhiên là: (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) và giá trị trung bình mẫu là $\overline{x} = 14,12$; $\overline{y} = 15,5$, phương trình hồi quy tuyến tính mẫu là: y = a. x - 2,15. Khi đó giá trị của hệ số hồi quy a là:

a. 1,35

b. 1,25

c. 1,45

d. 1,27

e. không xác định được

4. Cho hai biến ngẫu nhiên X, Y có hệ số tương quan là 0,96 và các phương sai mẫu tương ứng là $s_x^2 = 2,7$; $s_y^2 = 3,01$, trung bình mẫu $\overline{y} = 17,5$; $\overline{x} = 14,25$. Phương trình hồi quy tuyến tính mẫu của Y theo X là

a. y = 1,014.x + 3,056

b. y = 1,015.x + 3,125

c. $y = 1,015.x + 3,045$

d. $y = 1,114.x + 3,125$

5. Cho hai biến ngẫu nhiên X và Y có phương trình hồi quy tuyến tính mẫu là: $y = a.x + b$. Biết rằng hai mẫu ngẫu nhiên tương ứng có trung bình mẫu là $\overline{x} = 9,75; \overline{y} = 12,45$ và khi dự báo giá trị của X là 4,5 thì giá trị của Y là 2,25. Khi đó cặp hệ số hồi quy a, b là:

a. 1,4125; -0,7675

b. 1,4125; 0,7675

c. -1,4125; -0,7675

d. 1,4125; 0,7675

e. 1,425; 0,975

BÀI TẬP

1. Điều tra giữa tổng giá trị hàng hoá xuất khẩu X và tiền trợ cấp hưu trí Y trong vòng 10 năm, thu được kết quả sau.

Y	5,1	5,3	5,2	4,9	4,8	4,7	4,5	5,0	4,6	4,4
X	32	30	35	39	37	36	40	39	42	45

a. Xác định hệ số tương quan giữa X và Y.

b. Tìm phương trình hồi quy tuyến tính mẫu của Y theo X.

c. Xác định sai số hồi quy.

d. Giả sử xuất khẩu X có giá trị là 53 tỷ/năm, hãy dự báo tiền trợ cấp hưu trí.

2. Điều tra về năng suất cây trồng Y và lượng đầu tư cải tạo đất X (triệu/ha) tại một tỉnh ta thu được số liệu sau.

X	20	21	21	23	25	25	26	28	30	30	31
Y	2	2	3	3	4	5	6	6	7	8	8

a. Tính hệ số tương quan mẫu.

b. Tìm hàm hồi quy trung bình tuyến tính mẫu của Y theo X.

c. Tính sai số hồi quy.

d. Nếu lượng đầu tư X bằng 35(triệu/ha) thì có thể dự báo năng suất Y cây trồng bằng bao nhiêu?

3. Điều tra mức tiêu dùng Y và thu nhập X của 14 người, thu được số liệu sau:

X	1	1	2	3	4	5	6	7	8	9	10	10	10	10
Y	2,3	2,5	2,6	3	3,2	3,5	4	4	4,5	5	5,2	5,5	5,5	6

a. Tính hệ số tương quan giữa X và Y.

b. Tìm hàm hồi quy trung bình tuyến tính mẫu của Y theo X.

c. Tính sai số hồi quy.

d. Nếu thu nhập X bằng 7 triệu thì dự báo mức chi tiêu Y là bao nhiêu?