Chửơng 6

LÝ THUYẾT TƯƠNG QUAN VÀ HÀM HỒI QUI

1. MỐI QUAN HỆ GIỮA HAI ĐẠI LƯỢNG NGẪU NHIÊN

Khi khảo sát hai đại lượng ngẫu nhiên X, Y ta thấy giữa chúng có thể có một số quan hệ sau:

i) X và Y độc lập với nhau, tức là việc nhận giá trị của đại lượng ngẫu nhiên này không ảnh hưởng đến việc nhận giá trị của đại lượng ngẫu nhiên kia.

ii) X và Y có mối phụ thuộc hàm số $Y = \varphi(X)$.

iii) X và Y có sự phụ thuộc tương quan và phụ thuộc không tương quan.

2. HỆ SỐ TƯƠNG QUAN

2.1 Moment tửơng quan (Covarian)

□ **Định nghĩa 1**

** Moment tương quan (hiệp phương sai) của hai đại lượng ngẫu nhiên X và Y, kí hiệu $cov(X,Y)$ hay μ_{XY}, là số được xác định như sau*

$$cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

** Nếu $cov(X,Y) = 0$ thì ta nói hai đại lượng ngẫu nhiên X và Y không tương quan.*

⊙ **Chú ý**

$$\boxed{cov(X,Y) = E(XY) - E(X).E(Y)}$$

Thật vậy, ta có

$$\begin{aligned} cov(XY) &= E\{X.Y - X.E(Y) - Y.E(X) + E(X).E(Y) \\ &= E(XY) - E(X).E(Y) - E(X).E(Y) + E(X).E(Y) \\ &= E(XY) - E(X).E(Y) \end{aligned}$$

⊕ **Nhận xét 1**

* Nếu (X, Y) rời rạc thì

$$cov(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(x_i, y_j) - E(X)E(Y)$$

* Nếu (X, Y) liên tục thì

$$cov(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy - E(X)E(Y)$$

⊕ **Nhận xét**

i) Nếu X và Y là hai đại lượng ngẫu nhiên độc lập thì chúng không tương quan.

ii) Cov(X,X)=Var(X).

2.2 Hệ số tương quan

□ **Định nghĩa 2** *Hệ số tương quan của hai đại lượng ngẫu nhiên X và Y, kí hiệu r_{XY}, là số được xác định như sau*

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

với S_x, S_Y là độ lệch tiêu chuẩn của X, Y.

• **Ý nghĩa của hệ số tương quan**

Hệ số tương quan đo mức độ phụ thuộc tuyến tính giữa X và Y. Khi $|r_{XY}|$ càng gần 1 thì mối quan hệ tuyến tính càng chặt, khi $|r_{XY}|$ càng gần 0 thì quan hệ tuyến tính càng "lỏng lẻo".

2.3 Ước lượng hệ số tương quan

Lập mẫu ngẫu nhiên $W_{XY} = [(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)]$.

Để ước lượng hệ số tương quan $r_{XY} = \dfrac{E(XY) - E(X).E(Y)}{S_X.S_Y}$ ta dùng thống kê

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

trong đó

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \overline{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

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

Với mẫu cụ thể, ta tính được giá trị của R là

$$\boxed{r_{XY} = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x.s_y}}$$

trong đó

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2, \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\overline{y})^2$$

Ta có

$$\boxed{r_{XY} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}.\sqrt{n(\sum y^2) - (\sum y)^2}}}$$

2.4 Tính chất của hệ số tưởng quan

Hệ số tương quan $r = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{s_x.s_y}$ được dùng để đánh giá mức độ chặt chẻ của sự phụ thuộc tương quan tuyến tính giữa hai đại lượng ngẫu nhiên X và Y, nó có các tính chất sau đây:

i) $|r| \leq 1$.

ii) Nếu $|r| = 1$ thì X và Y có quan hệ tuyến tính.

iii) Nếr $|r|$ càng lớn thì sự phụ thuộc tương quan tuyến tính giữa X và Y càng chặt chẻ.

iv) Nếu $|r| = 0$ thì giữa X và Y không có phụ thuộc tuyến tính tương quan.

v) Nếu $r > 0$ thì X và Y có tương quan thuận (X tăng thì Y tăng). Nếu $r < 0$ thì X và Y có tương quan nghịch (X giảm thì Y giảm).

• **Ví dụ 1** *Từ số liệu được cho bởi bảng sau, hãy xác định hệ số tương quan của Y và X*

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Giải

Ta lập bảng sau

x_i	y_i	x_i^2	x_iy_i	y_i^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
$\sum x = 56$	$\sum y = 40$	$\sum x^2 = 524$	$\sum xy = 364$	$\sum y^2 = 256$

Hệ số tương quan của X và Y là

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

$$= \frac{8.364 - (56).(40)}{\sqrt{8.524 - (56)^2}.\sqrt{8.256 - (40)^2}} = \frac{672}{687,81} = 0,977$$

2.5 Tỷ số tưởng quan

Để đánh giá mức độ chặt chẻ của sự phụ thuộc tương quan phi tuyến, người ta dùng *tỷ số tương quan*:

$$\boxed{\eta_{Y/X} = \frac{s_{\overline{y}}}{s_y}}$$

trong đó

$$s_{\overline{y}} = \sqrt{\frac{1}{n}\sum n_i.(\overline{y_{x_i}} - \overline{y})^2}; \qquad s_y = \sqrt{\frac{1}{n}\sum m_j.(y_j - \overline{y})^2}$$

Tỷ số tương quan có các tính chất sau:

i) $0 \leq \eta_{Y/X} \leq 1$.

ii) $\eta_{Y/X} = 0$ khi và chỉ khi Y và X không có phụ thuộc tương quan.

iii) $\eta_{Y/X} = 1$ khi và chỉ khi Y và X phụ thuộc hàm số.

iv) $\eta_{Y/X} \geq |r|$.

Nếu $\eta_{Y/X} = |r|$ thì sự phụ thuộc tương quan của Y và X có dạng tuyến tính.

2.6 Hệ số xác định mẫu

Trong thống kê, để đánh giá chất lượng của mô hình tuyến tính người ta còn xét *hệ số xác định mẫu* $\beta = r^2$ với r là hệ số tương quan. Ta có $0 \leq \beta \leq 1$.

3. HỒI QUI

3.1 Kỳ vọng có điều kiện

i) Đại lượng ngẫu nhiên rời rạc

* Kỳ vọng có điều kiện của đại lượng ngẫu nhiên rời rạc Y với điều kiện $X = x$ là

$$E(Y/x) = \sum_{j=1}^{m} y_j P(X = x, Y = y_j)$$

* Tương tự, kỳ vọng có điều kiện của đại lượng ngẫu nhiên rời rạc X với điều kiện $Y = y$ là

$$E(X/y) = \sum_{i=1}^{n} x_i P(X = x_i, Y = y)$$

ii) Đại lượng ngẫu nhiên liên tục

$$E(Y/x) = \int_{-\infty}^{+\infty} y f(y/x) dy$$

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

trong đó

$f(y/x) = f(x, y)$ với x không đổi

$f(x/y) = f(x, y)$ với y không đổi

3.2 Hàm hồi qui

* Hàm hồi qui của Y đối với X là $f(x) = E(Y/x)$.

* Hàm hồi qui của X đối với Y là $f(y) = E(X/y)$.

Trong thực tế ta thường gặp hai đại lượng ngẫu nhiên X, Y có mối liên hệ với nhau, trong đó việc khảo sát X thì dễ còn khảo sát Y thì khó hơn thậm chí không thể khảo sát được. Người ta muốn tìm mối liên hệ $\varphi(X)$ nào đó giữa X và Y để biết X ta có thể dự đoán được Y.

Giả sử biết X, nếu dự đoán Y bằng $\varphi(X)$ thì sai số phạm phải là $E[Y - \varphi(X)]^2$. Vấn đề được đặt ra là tìm $\varphi(X)$ như thế nào để $E[Y - \varphi(X)]^2$ là nhỏ nhất.

Ta sẽ chứng minh khi chọn $\varphi(X) = E(Y/X)$ (với $\varphi(x) = E(Y/x)$) thì $E[Y - \varphi(X)]^2$ sẽ nhỏ nhất.

Thật vậy, ta có

$$\begin{aligned} E[Y - \varphi(X)]^2 &= E\{([Y - E(Y/X)] + [E(Y/X) - \varphi(X)])^2\} \\ &= E\{[Y - E(Y/X)]^2\} + E\{[E(Y/X) - \varphi(X)]^2\} \\ &\quad +2E\{[Y - E(Y/X)][E(Y/X) - \varphi(X)]\} \end{aligned}$$

Ta thấy $E(Y/X)$ chỉ phụ thuộc vào X nên có thể đặt $T(X) = E(Y/X) - \varphi(X)$.

Vì $E[E(Y/X)T(X)] = E[YT(X)]$ nên

$$\begin{aligned} 2E[Y - E(Y/X)][E(Y/X) - \varphi(X)] &= 2E\{[Y - E(Y/X)]T(X)\} \\ &= 2E[YT(X)] - 2E[E(Y/X)T(X)] = 0 \end{aligned}$$

Do đó

$$E\{[Y - \varphi(X)]^2\} = E\{[Y - E(Y/X)]^2\} + E\{E(Y/X) - \varphi(X)]^2$$

nhỏ nhất khi

$$E\{[(Y/X) - \varphi(X)]^2 = 0$$

Ta chỉ cần chọn

$$\varphi(X) = E(Y/X) \tag{6.1}$$

Phương trình (6.1) được gọi là *phương trình tương quan* hay *phương trình hồi qui.*

3.3 Xác định hàm hồi qui

a) Trường hợp ít số liệu (tương quan cặp)

Giả sử giữa hai đại lượng ngẫu nhiên X và Y có tương quan tuyến tính, tức là $E(Y/X) = AX + B$.

Dựa vào n cặp giá trị $(x_1, x_2), (x_2, y_2), \ldots, (x_n, y_n)$ của (X, Y) ta tìm hàm

$$\overline{y_x} = y = ax + b \quad (*)$$

để ước lượng hàm $Y = AX + B$.

(*) được gọi là *hồi qui tuyến tính mẫu.*

Vì các cặp giá trị trên là trị xấp xỉ của x và y nên thỏa (*) một cách xấp xỉ.

Do đó $y_i = ax_i + b + \varepsilon_i$ hay $\varepsilon_i = y_i - ax_i - b$.

Ta tìm a, b sao cho các sai số ε_i $(i = \overline{1, n})$ có trị tuyệt đối nhỏ nhất hay hàm

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

đạt cực tiểu. Phương pháp tìm này được gọi là *phương pháp bình phương bé nhất.*

Ta thấy S sẽ đạt giá trị nhỏ nhất tại điểm dừng thỏa mãn

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

$$0 = \frac{\partial S}{\partial b} = -2 \sum_{i=1}^{n} (y_i - ax_i - b)$$

hay

$$\boxed{\begin{aligned}\left(\sum_{i=1}^{n} x_i^2\right).a + \left(\sum_{i=1}^{n} x_i\right).b &= \sum_{i=1}^{n} x_i y_i \\ \left(\sum_{i=1}^{n} x_i\right).a + nb &= \sum_{i=1}^{n} y_i\end{aligned}} \quad (6.2)$$

Hệ trên có định thức

$$D = \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix} = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2$$

Vì các x_i khác nhau nên theo bất đẳng thức Bunhiakovsky ta có $\left(\sum_{i=1}^{n} x_i\right)^2 < n\sum_{i=1}^{n} x_i^2$. Do đó $D > 0$. Suy ra hệ trên có nghiệm duy nhất

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

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

Nếu đặt

$$\overline{x} = \frac{1}{n}.\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}.\sum_{i=1}^{n} y_i, \quad \overline{xy} = \frac{1}{n}.\sum_{i=1}^{n} x_i y_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

thì nghiệm của hệ có thể viết lại dưới dạng

$$a = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x^2}; \qquad b = \frac{\overline{x^2}.\overline{y} - \overline{x}.\overline{xy}}{\overline{x^2} - (\overline{x})^2} = \frac{\overline{x^2}.\overline{y} - \overline{x}.\overline{xy}}{s_x^2}$$

Tóm lại, ta có thể tìm hàm $\overline{y_x} = ax + b$ từ các công thức

$$\boxed{\begin{gathered} a = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x^2} = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \\ b = \overline{y} - a.\overline{x} \end{gathered}}$$

⊙ **Chú ý**

-bb-error =

Đường gấp khúc nối các điểm (x_1, y_1), (x_2, y_2) , . . . , (x_n, y_n) được gọi là *đường hồi qui thực nghiệm.*

Đường thẳng $y = ax + b$ nhận được bởi công thức bình phương bé nhất không đi qua được tất cả các điểm nhưng là đường thẳng "gần" các điểm đó nhất được gọi là *đường thẳng hồi qui* và thủ tục làm thích hợp đường thẳng thông qua các điểm dữ liệu cho trước được gọi là *hồi qui tuyến tính.*

Theo trên ta có $b = \overline{y} - a.\overline{x}$, do đó điểm $(\overline{x}, \overline{y})$ luôn nằm trên đường thẳng hồi qui.

• **Ví dụ 2** *Ước lượng hàm hồi qui tuyến tính mẫu xủa Y theo X trên cơ sở bảng tương quan cặp sau*

X	15	38	23	16	16	13	20	24
Y	145	228	150	130	160	114	142	265

Giải

Ta lập bảng sau

x_i	y_i	x_i^2	x_iy_i
15	145	225	3175
38	228	1444	8664
23	150	529	3450
16	130	256	2080
16	160	256	2560
13	114	169	1482
20	142	400	2840
24	265	576	6360
$\sum x = 165$	$\sum y = 1334$	$\sum x^2 = 3855$	$\sum xy = 29611$

Ta có

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

$$= \frac{8(19611) - (165)(1334)}{8(3855)(165)^2} = \frac{16778}{3615} = 4,64$$

$$b = \overline{y} - a\overline{x} = \frac{1334}{8} - \left(\frac{16778}{3615}\right)\left(\frac{165}{8}\right) = 71$$

Vậy hàm hồi qui tuyến tính mẫu là $\overline{y_x} = 4,64x + 71$.

• **Ví dụ 3** *Độ ẩm của không khí ảnh hưởng đến sự bay hơi của nước trong sơn khi phun ra. Người ta tiến hành nghiên cứu mối liên hệ giữa độ ẩm của không khí X và độ bay hơi Y. Sự hiểu biết về mối quan hệ này sẽ giúp ta tiết kiệm được lượng sơn bằng cách chỉnh súng phun sơn một cách thích hợp. Tiến hành 25 quan sát ta được các số liệu sau:*

Quan sát	Độ ẩm (%)	Độ bay hơi (%)	Quan sát	Độ ẩm (%)	Độ bay hơi (%)
1	35,3	11,0	14	39,1	9,6
2	29,7	11,1	15	46,8	10,9
3	30,8	12,5	16	48,5	9,6
4	58,8	8,4	17	59,3	10,1
5	61,4	9,3	18	70,0	8,1
6	71,3	8,7	19	70,0	6,8
7	74,4	6,4	20	74,4	8,9
8	76,7	8,5	21	72,1	7,7
9	70,7	7,8	22	58,1	8,5
10	57,5	9,1	23	44,6	8,9
11	46,4	8,2	24	33,4	10,4
12	28,9	12,2	25	28,6	11,1
13	28,1	11,9			

Hãy tìm hàm hồi qui tuyến tính mẫu $\overline{y_x} = ax + b$.

Giải

Ta có

$$n = 25 \qquad \sum x = 1314,9 \qquad \sum y = 235,7$$

$$\sum x^2 = 76308,53 \qquad \sum y^2 = 2286,07$$

$$\sum xy = 11824,44$$

Do đó

$$a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{25 \times 11824,44 - (1314,9 \times 235,7)}{25 \times 76308,53 - (1314,9)^2} = -0,08$$

$$b = \overline{y} - a\overline{x} = 9,43 - (-0,08) \times 52,6 = 13,64$$

Vậy hàm hồi qui tuyến tính mẫu là $\overline{y_x} = -0,08x + 13,64$

b) Trường hợp nhiều số liệu (tương quan bảng)

Giả sử

X nhận các giá trị x_i với tần suất n_i $i = \overline{1,k}$,

Y nhận các giá trị y_j với tần suất m_j $j = \overline{1,h}$,

XY nhận các giá trị x_iy_j với tần suất n_{ij} $i = \overline{1,k}$, $j = \overline{1,h}$,

Ta tìm hồi qui tuyến tính mẫu $\overline{y_x} = ax + b$ trong trường hợp có nhiều số liệu. Theo (6.2) ta có

$$\begin{aligned}\left(\sum_{i=1}^{k} n_i x_i^2\right).a + \left(\sum_{i=1}^{k} n_i x_i\right).b &= \sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} x_i y_j \\ \left(\sum_{i=1}^{k} n_i x_i\right).a + nb &= \sum_{j=1}^{h} m_j y_j\end{aligned} \tag{6.3}$$

Thay $\sum_{i=1}^{k} n_i x_i = n\overline{x}$, $\sum_{j=1}^{h} m_j y_j = n\overline{y}$, $\sum_{i=1}^{k} n_i x_i^2 = n\overline{x^2}$, $\sum_{j=1}^{h} m_j y_j^2 = n\overline{y^2}$,

$\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} x_i y_j = n\overline{xy}$ vào (6.3) ta được

$$\begin{aligned}\overline{x^2}.a + \overline{x}.b &= \overline{xy} \quad (i) \\ \overline{x}.a + nb &= \overline{y} \quad (ii)\end{aligned}$$

Từ (ii) ta có $b = \overline{y} - a.\overline{x}$

Thay b vào $\overline{y_x} = ax + b$ ta suy ra

$$\overline{y_x} - \overline{y} = a(x - \overline{x}) \tag{6.4}$$

Ta tìm a bởi

$$a = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij}x_i y_j - (\sum_{i=1}^{k} n_i x_i)(\sum_{j=1}^{h} m_j y_j)}{n\sum_{i=1}^{k} n_i x_i^2 - (\sum_{i=1}^{k} n_i x_i)^2} = \frac{n^2\overline{xy} - n\overline{x}.n\overline{y}}{n.n\overline{x^2} - (n\overline{x})^2}$$

$$= \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x^2}$$

Tóm lại, ta tìm hồi qui tuyến tính mẫu $\overline{y_x} = ax + b$ với $\boxed{a = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{s_x^2}, \quad b = \overline{y} - a\overline{x}}$.

⊙ **Chú ý**

i) Ta biết hệ số tương quan $r_{XY} = \dfrac{\overline{xy} - \overline{x}\overline{y}}{s_x.s_y}$ nên $a = r_{XY}\dfrac{s_y}{s_x}$

Thay a vào (6.4) ta có

$$\overline{y_x} - \overline{y} = r_{XY}\frac{s_y}{s_x}(x - \overline{x})$$

hay

$$\frac{\overline{y_x} - \overline{y}}{s_y} = r_{XY}\frac{(x - \overline{x})}{s_x}$$

Từ phương trình này ta có thể suy ra phương trình hồi qui tuyến tính mẫu $\overline{y_x} = ax+b$ một cách thuận lợi hơn vì thông qua việc tìm r_{XY} ta đã tính s_x, s_y.

ii) Khi các giá trị của X, Y khá lớn, ta có thể dùng phép đổi biến

$$u_i = \frac{x_i - x_0}{h_x} \quad (\forall i = \overline{1,k}); \qquad v_j = \frac{y_j - y_0}{h_y} \quad (\forall j = \overline{1,h})$$

trong đó

* x_0, y_0 là những giá trị tùy ý (thường chọn x_0, y_0 là giá trị của X, Y ứng với tần số n_{ij} lớn nhất trong bảng tương quan thực nghiệm),

* h_x, h_y là các giá trị tùy ý (thường chọn h_x, h_y là khoảng cách các giá trị kế tiếp nhau của X, Y).

Lập bảng tương quan đối với các biến mới U, V và tính toán các giá trị cần thiết ta tìm được hàm hồi qui tuyến tính mẫu

$$\overline{v_u} = a_0.u + b_0$$

trong đó

$$a_0 = \frac{\overline{uv} - \overline{u}.\overline{v}}{s_u^2}, \quad b_0 = \overline{v} - a_0.\overline{u}$$

Khi đó ta suy ra hàm $\overline{y_x} = ax + b$ với a, b được tìm bởi công thức

$$a = a_0\frac{h_y}{h_x}, \qquad b = y_0 + b_0.h_y - a_0.\frac{h_y}{h_x}.x_0$$

• **Ví dụ 4** *Xác định hệ số tương quan và hàm hồi qui tuyến tính mẫu $\overline{y_x} = ax + b$ của các đại lương ngẫu nhiên X và Y cho bởi bảng tương quan thực nghiệm sau:*

Y \ *X*	*1*	*2*	*3*
10	*20*		
20		*30*	*1*
30		*1*	*48*

Giải

Ta lập bảng sau

Y \ X	1	2	3	m_j	m_jy_j	$m_jy_j^2$
10	200 $\lvert\overline{20}$			20	200	2000
20		1200 $\lvert\overline{30}$	60 $\lvert\overline{1}$	31	620	12400
30		60 $\lvert\overline{1}$	4320 $\lvert\overline{48}$	49	1470	44100
n_i	20	31	49	n=100	$\sum y = 2290$	$\sum y^2 = 58500$
n_ix_i	20	62	147	$\sum x = 229$		
$n_ix_i^2$	20	124	441	$\sum x^2 = 585$		$\sum xy = 5840$

$$\sum xy = 200 + 1200 + 60 + 60 + 4320 = 5840$$

Phần trên góc trái của ô ghi các tích $n_{ij}x_iy_j$. Ta có

$$\overline{x} = \frac{229}{100} = 2,29; \qquad \overline{y} = \frac{2290}{100} = 22,9;$$

$$\overline{x^2} = \frac{585}{100} = 5,58; \qquad \overline{y^2} = \frac{58500}{100} = 585 \qquad \overline{xy} = \frac{5840}{100} = 58,4;$$

$$s_x^2 = \overline{x^2} - (\overline{x})^2 = 5,85 - (2,29)^2 \approx 0,6059 \quad \Longrightarrow \quad s_x \approx 0,78$$

$$s_y = \sqrt{\overline{y^2} - (\overline{y})^2} = \sqrt{585 - (22,9)^2} \approx 7,78$$

Do đó

$$a = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x^2} = \frac{58,4 - 2,29 \times 22,9}{0,6059} = 9,835$$

$$b = \overline{y} - a.\overline{x} = 22,9 - 9,835 \times 2,29 = 0,378$$

Hàm hồi qui tuyến tính mẫu là $\overline{y_x} = 9,835x + 0,378$

Hệ số tương quan là

$$r_{xy} = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x.s_y} = \frac{58,4 - 2,29 \times 22,9}{0,78 \times 7,78} \approx 0,982$$

4. BÀI TẬP

1. Cho các giá trị quan sát của hai đại lượng ngẫu nhiên X và Y ở bảng sau:

X	5	10	10	10	15	15	15	20	20	20
Y	20	20	30	30	30	40	50	50	60	60

   Giả sử X và Y có sự phụ thuộc tương quan tuyến tính. Tìm hàm hồi qui tuyến tính mẫu: $\overline{y}_x = ax + b$.

2. Người ta đo chiều dài vật đúc và khuôn thì thấy chúng lệch khỏi qui định nhưsau:

X	0.90	1,22	1,32	0,77	1,30	1,20	1,32	0,95	0,45	1,30	1,20
Y	-0,30	0,10	0,70	-0,28	0,25	0,02	0,37	-0,70	0,55	0,35	0,32

   Trong đó X, Y là các độ lệch.

   Xác định hệ số tương quan.

3. Số liệu thống kê nhằm nghiên cứu quan hệ giữa tổng sản phẩm nông nghiệp Y với tổng giá trị tài sản cố định X của 10 nông trại (tính trên 100 ha) như sau:

X	11,3	12,9	13,6	16,8	18,8	20,0	22,2	23,7	26,6	27,5
Y	13,2	15,6	17,2	18,8	20,2	23,9	22,4	23,0	24,4	24,6

Xác định đường hồi qui tuyến tính mẫu $\overline{y}_x = ax + b$. Sau đó tìm phương sai sai số thực nghiệm và khoảng tin cậy 95% cho hệ số góc của đường hồi qui trên.

4. Đo chiều cao X (cm) và trọng lượng Y (kg) của 100 học sinh, ta được kết quả sau:

X / Y	145 − 150	150 − 155	155 − 160	160 − 165	165 − 170
35 − 40	3				
40 − 45	5	10			
45 − 50		14	20	6	
50 − 55			15	12	5
55 − 60				6	4

Giả thuyết X và Y có mố phụ thuộc tương quan tuyến tính. Tìm các hàm hồi qui

a) $\overline{y}_x = ax + b$;

b) $\overline{x}_y = cy + d$

5. Theo dõi lượng phân bón và năng suất lúa của 100 hecta lúa ở một vùng, ta thu được bảng số liệu sau:

X / Y	120	140	160	180	200
2,2	2				
2,6	5	3			
3,0		11	8	4	
3,4			15	17	
3,8			10	6	7
4,2					12

Trong đó X là phân bón (kg/ha) và Y là năng suất lúa (tấn/ha).

a) Hãy ước lượng hệ số tương quan tuyến tính r.
b) Tìm phương trình tương quan tuyến tính: $\overline{y}_x = ax + b$.

6. Đo chiều cao và đường kính của một loại cây, ta được kết quả cho bở bảng sau:

X / Y	6	8	10	12	14
30	2	17	9	3	
35		10	17	9	
40		3	24	16	13
45			6	24	12
50			2	11	22

Trong đó X là đường kính (cm) và Y là chiều cao (m).

a) Xác định hệ số tương quan tuyến tính mẫu r.
b) Tìm các phương trình hồi qui tuyến tính mẫu.
c) Các phương trình trên sẽ thay đổi như thế nào nếu X được tính theo đơn vị là mét (m)?

◙ TRẢ LỜI BÀI TẬP

1. $\overline{x} = 14$, $\overline{y} = 39$, $\overline{y}_x = \frac{8}{3}x + \frac{5}{3}$.

2. $r = -0,3096$.

3. $\overline{y}_x = 0,67x + 7,18$, $\sigma^2 = 1,126$, $(0,6280\,;\,0,7176)$.

4. a) $\overline{y}_x = 0,7018x - 61,5537$, b) $\overline{x}_y = 0,91y + 112,96$.

5. $r = 0,8165$; $\overline{y}_x = 0,017x + 0,5622$.

6. a) $r = 0,69$, b) $\overline{y}_x = 0,218x + 2,434$, $\overline{x}_y = 2,18y + 15,87$.

c) $\overline{y}_{x'} = 21,8x' + 2,434$, $\overline{x}_y = 0,0218y' + 0,1587$.