

NỘI DUNG 13: HỒI QUY

1 LÝ THUYẾT

1.1 Viết phương trình đường hồi quy và tìm hệ số tương quan

Phương trình hồi quy tuyến tính mẫu của Y theo X: $y = a + bx$, với:

$$\begin{cases} a = \bar{y} - b.\bar{x} \\ b = \frac{\overline{xy} - \bar{x}.\bar{y}}{\widehat{s}_x^2} \end{cases}$$

Phương trình hồi quy tuyến tính mẫu của X theo Y: $x = c + dy$, với:

$$\begin{cases} c = \bar{x} - d.\bar{y} \\ d = \frac{\overline{xy} - \bar{x}.\bar{y}}{\widehat{s}_y^2} \end{cases}$$

Hệ số tương quan mẫu:

$$r_{XY} = \frac{\overline{xy} - \bar{x}.\bar{y}}{\widehat{s}_x.\widehat{s}_y}$$

Hệ số xác định: r^2

Tham khảo cách trình bày khác:

Dạng 1: Viết phương trình đường hồi quy và tìm hệ số tương quan:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \\ S_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \\ S_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i) \cdot (\sum_{i=1}^n y_i)}{n} \end{aligned}$$

Phương trình hồi quy Y theo X có dạng: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ với:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i) \cdot (\sum_{i=1}^n y_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

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

Phương trình hồi quy X theo Y có dạng: $\hat{x} = \hat{\beta}_0 + \hat{\beta}_1 y$ với:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i) \cdot (\sum_{i=1}^n y_i)}{n}}{\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}} = \frac{S_{xy}}{S_{yy}}$$

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

Xác định hệ số tương quan:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Hệ số xác định: r^2

1.2 Ước lượng sai số chuẩn

$$SSE = SST - SSR = n.\widehat{s}_y^2 - \frac{n.(\overline{xy} - \bar{x}.\bar{y})^2}{\widehat{s}_x^2}$$

$$\text{Sai số chuẩn của ước lượng: } \sigma = \sqrt{\frac{SSE}{n-2}}$$

1.3 Khoảng tin cậy cho các hệ số hồi quy

Khoảng tin cậy cho hệ số tự do A: $(a - \varepsilon_a; a + \varepsilon_a)$. Trong đó:

$$\varepsilon_a = t_{\alpha/2}^{n-2} \cdot \frac{\sqrt{SSE \cdot \overline{x^2}}}{\hat{s}_x \cdot \sqrt{n(n-2)}}$$

Khoảng tin cậy cho hệ số góc B: $(b - \varepsilon_b; b + \varepsilon_b)$. Trong đó:

$$\varepsilon_b = t_{\alpha/2}^{n-2} \cdot \frac{\sqrt{SSE}}{\hat{s}_x \cdot \sqrt{n(n-2)}}$$

1.4 Kiểm định cho các hệ số hồi quy

1.4.1 Kiểm định hệ số tự do A

a) Phát biểu giả thiết

Giả thiết	Miền bác bỏ (1)	Miền bác bỏ (2)
$H_0 : A = A_0; H_1 : A \neq A_0$	$W_\alpha = (-\infty; -t_{\alpha/2}^{n-2}) \cup (t_{\alpha/2}^{n-2}; +\infty)$	$ T > t_{\alpha/2}^{n-2}$
$H_0 : A = A_0; H_1 : A < A_0$	$W_\alpha = (-\infty; -t_{\alpha}^{n-2})$	$T < -t_{\alpha}^{n-2}$
$H_0 : A = A_0; H_1 : A > A_0$	$W_\alpha = (t_{\alpha}^{n-2}; +\infty)$	$T > t_{\alpha}^{n-2}$

Thông thường thì $A_0 = 0$.

b) Tính tiêu chuẩn kiểm định:

$$T = \frac{a - A_0}{\frac{\sqrt{SSE \cdot \overline{x^2}}}{\hat{s}_x \cdot \sqrt{n(n-2)}}}$$

1.4.2 Kiểm định hệ số góc B

a) Phát biểu giả thiết

Giả thiết	Miền bác bỏ (1)	Miền bác bỏ (2)
$H_0 : B = B_0; H_1 : B \neq B_0$	$W_\alpha = (-\infty; -t_{\alpha/2}^{n-2}) \cup (t_{\alpha/2}^{n-2}; +\infty)$	$ T > t_{\alpha/2}^{n-2}$
$H_0 : B = B_0; H_1 : B < B_0$	$W_\alpha = (-\infty; -t_{\alpha}^{n-2})$	$T < -t_{\alpha}^{n-2}$
$H_0 : B = B_0; H_1 : B > B_0$	$W_\alpha = (t_{\alpha}^{n-2}; +\infty)$	$T > t_{\alpha}^{n-2}$

Thông thường thì $B_0 = 0$.

b) Tính tiêu chuẩn kiểm định:

$$T = \frac{b - B_0}{\frac{\sqrt{SSE}}{\hat{s}_x \cdot \sqrt{n(n-2)}}}$$

1.5 Kiểm định sự phù hợp của hàm hồi quy tuyến tính đơn

a) Phát biểu giả thiết:

$H_0: R^2 = 0$ hoặc $(B = 0)$: Phương trình đường hồi quy không thích hợp

$H_1: R^2 \neq 0$ hoặc $(B \neq 0)$: Phương trình đường hồi quy thích hợp

b) Miền bác bỏ: $W_\alpha = (F_{\alpha}^{1;n-2}; +\infty)$ hoặc $F > F_{\alpha}^{1;n-2}$

c) Tính tiêu chuẩn kiểm định:

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

1.6 Kiểm định mối tương quan tuyến tính X, Y

a) Phát biểu giả thiết:

$H_0: R = 0$: X, Y không có tương quan tuyến tính

$H_1: R \neq 0$: X, Y có tương quan tuyến tính

b) Miền bác bỏ: $W_\alpha = (-\infty; -t_{\alpha/2}^{n-2}) \cup (t_{\alpha/2}^{n-2}; +\infty)$ hoặc $|T| > t_{\alpha/2}^{n-2}$

c) Tiêu chuẩn kiểm định:

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

1.7 Khoảng tin cậy cho giá trị dự đoán

$$a + bx \pm t_{\alpha/2}^{n-2} \cdot \sigma \cdot \sqrt{\left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{n \cdot \widehat{s_x^2}} \right]}$$

2 BÀI TẬP

Bài 1: Khi khảo sát mối liên hệ giữa lượng phân bón X (kg/ha) và năng suất tương ứng Y (tạ/ha) của một loại cây trồng trong vùng, người ta lấy số liệu mẫu trên 100 thửa ruộng có cùng diện tích. Kết quả được biểu diễn trong bảng sau:

	X			
Y	300	340	360	380
32	10	5		
34		10	7	
36		4	20	14
38			10	20

Hãy ước lượng hệ số tương quan giữa X, Y; phương trình đường hồi quy tuyến tính của Y theo X; và dự kiến năng suất cây trồng nếu sử dụng lượng phân bón là 390 kg/ha.

Bài 2: Người ta khảo sát một loại cây được liệu trưởng thành về chỉ số chiều cao X (cm) và chỉ số trọng lượng Y (100 gram). Dưới đây là số liệu của mẫu thu được:

	Y				
X	4	5	6	7	8
100	5	5			
110	4	6	7		
120		5	9	8	
130			4	6	9
140				5	7

Giả thiết rằng chiều cao và trọng lượng của cây tuân theo phân phối chuẩn.

Hãy tìm hệ số tương quan mẫu (X, Y), viết phương trình đường hồi quy tuyến tính mẫu của Y theo X và dự đoán trọng lượng của cây loại này có chiều cao 145 cm.

Bài 3: Xét một mẫu ngẫu nhiên 2 chiều (X, Y). Trong đó: X là số công nhân trong 1 phân xưởng và Y là số thành phẩm trong 1 ngày của phân xưởng đó. Giả sử các phân xưởng cùng sản xuất 1 loại sản phẩm.

X (số công nhân)	68	58	67	78	72	59	82	89	65	78
Y (số sản phẩm)	1695	1475	1650	2000	1820	1490	2040	2220	1630	2015

- Hãy dự đoán số sản phẩm được sản xuất bởi một phân xưởng có 70 công nhân.
- Hãy dự đoán số công nhân cần có để một phân xưởng có thể đạt 2100 sản phẩm một ngày.

Bài 4: Một nghiên cứu về hàm lượng đường (tính theo phần trăm) được chuyển hoá trong một quá trình xác định khi tăng nhiệt độ ở các mức khác nhau cho dữ liệu sau:

Mức tăng nhiệt độ ($^{\circ}\text{C}$)	1.1	1.2	1.3	1.4	1.5
Hàm lượng đường	8.5	8.6	9.2	9.1	10.1

- Tìm hệ số tương quan và tìm phương trình đường hồi quy tuyến tính Y theo X.
- Hãy xác định sai số chuẩn của đường hồi quy, và tìm khoảng tin cậy cho hệ tự do và hệ số góc của đường hồi quy tuyến tính với độ tin cậy 95%.
- Hãy tìm khoảng tin cậy 95% cho hàm lượng đường được chuyển hoá khi mức tăng nhiệt độ đạt 1.7°C .

Bài 5: Một nhà thực vật học khảo sát mối liên hệ giữa tổng diện tích bề mặt (đơn vị: cm^2) của các lá cây đậu nành và trọng lượng khô (đơn vị: gam) của các cây này. Nhà thực vật học trồng 13 cây trong nhà kính và đo tổng diện tích lá và trọng lượng của các cây này sau 16 ngày trồng, kết quả cho bởi bảng sau:

X	411	550	471	393	427	431	492	371	470	419	407	489	439
Y	2.00	2.46	2.11	1.89	2.05	2.30	2.46	2.06	2.25	2.07	2.17	2.32	2.12

- Xác định hệ số tương quan và đường thẳng hồi quy biểu diễn mối liên hệ giữa trọng lượng cây Y theo diện tích lá X?
- Nêu ý nghĩa của hệ số tương quan và cho nhận xét về hệ số tương quan tìm được ở bài toán.
- Xác định sai số chuẩn của ước lượng.
- Đường hồi quy và các hệ số có ý nghĩa hay không?
- Cho kết luận về mối tương quan tuyến tính giữa X và Y?
- Thiết lập khoảng tin cậy 95% cho các hệ số hồi quy.
- Thiết lập khoảng dự đoán 95% cho trọng lượng khô của cây có diện tích lá bằng 469 cm^2 .

TÍNH CÁC ĐẶC TRƯNG CỦA BẢNG PHÂN BỐ TẦN SỐ 2 CHIỀU

Máy tính CASIO 570VN, ES, VINACAL

Bước 1: Mở cột tần số

Nhấn **SHIFT** \Rightarrow **MODE** \Rightarrow $\nabla \Rightarrow$ **4 (STAT)**.

Bước 2: Mở chế độ nhập bảng:

Nhấn **MODE** \Rightarrow **3 (STAT)** \Rightarrow **2 (A + BX)**.

Bước 3: Sau khi xuất hiện 3 cột, ta nhập số liệu vào bảng:

Nhập các giá trị của X vào cột **X**, Nhập các giá trị của Y vào cột **Y**, nhập tần số tương ứng vào cột **FREQ**. Nhập xong nhấn **AC**.

Lưu ý: Nhập giá trị X, Y không giống như nhập bảng 1 chiều.

Bước 4:

Tìm các đặc trưng bằng cách:

Nhấn **SHIFT** \Rightarrow **1** \Rightarrow **4 (VAR)** với:

1 (n) : kích thước mẫu

2 (\bar{x}) : trung bình mẫu \bar{x}

3 (σ_x hay $x\sigma n$) : độ lệch mẫu $\hat{\sigma}_x$

\Rightarrow Nếu muốn tìm phương sai mẫu $\hat{\sigma}_x^2$ thì bình phương độ lệch mẫu $\hat{\sigma}_x$

4 (s_x hay $x\sigma n - 1$) : độ lệch mẫu hiệu chỉnh s_x

\Rightarrow Nếu muốn tìm phương sai mẫu hiệu chỉnh s_x^2 thì bình phương độ lệch mẫu hiệu chỉnh s_x

5 (\bar{y}) : trung bình mẫu \bar{y}

6 (σ_y hay $y\sigma n$) : độ lệch mẫu $\hat{\sigma}_y$

\Rightarrow Nếu muốn tìm phương sai mẫu $\hat{\sigma}_y^2$ thì bình phương độ lệch mẫu $\hat{\sigma}_y$

7 (s_y hay $y\sigma n - 1$) : độ lệch mẫu hiệu chỉnh s_y

\Rightarrow Nếu muốn tìm phương sai mẫu hiệu chỉnh s_y^2 thì bình phương độ lệch mẫu hiệu chỉnh s_y

hoặc tìm hệ số tương quan r_{XY} bằng cách:

Nhấn **SHIFT** \Rightarrow **1** \Rightarrow **5** \Rightarrow **3 (r)**

hoặc tìm hệ số tự do a bằng cách:

Nhấn **SHIFT** \Rightarrow **1** \Rightarrow **5** \Rightarrow **1 (A)**

hoặc tìm hệ số góc b bằng cách:

Nhấn **SHIFT** \Rightarrow **1** \Rightarrow **5** \Rightarrow **2 (B)**

hoặc tìm $\overline{xy} = \frac{\sum xy}{n}$ bằng cách:

Nhấn **SHIFT** \Rightarrow **1** \Rightarrow **3** \Rightarrow **5 ($\sum xy$)** \Rightarrow \div \Rightarrow **SHIFT** \Rightarrow **1** \Rightarrow **4** \Rightarrow **1 (n)**

TÍNH CÁC ĐẶC TRƯNG CỦA BẢNG PHÂN BỐ TẦN SỐ 2 CHIỀU

Máy tính 580VNX

Bước 1: Mở cột tần sốNhấn **SHIFT** \Rightarrow **MODE** \Rightarrow $\nabla \Rightarrow$ **4 (STAT)**.**Bước 2:** Mở chế độ nhập bảng:Nhấn **MODE** \Rightarrow **6 (THỐNG KÊ)** \Rightarrow **2 (a + bx)**.**Bước 3:** Sau khi xuất hiện 3 cột, ta nhập số liệu vào bảng:Nhập các giá trị của X vào cột **X**, Nhập các giá trị của Y vào cột **Y**, nhập tần số tương ứng vào cột **FREQ**. Nhập xong nhấn **AC**.Lưu ý: Nhập giá trị X, Y không giống như nhập bảng 1 chiều.**Bước 4:**

Tìm các đặc trưng bằng cách:

Nhấn **OPTION** \Rightarrow $\nabla \Rightarrow$ **2 (BIẾN THỐNG KÊ)** với:**1** (\bar{x}) : trung bình mẫu \bar{x} **2** (σ_x^2) : phương sai mẫu \hat{s}_x^2 **3** (σ_x) : độ lệch mẫu \hat{s}_x **4** (s_x^2) : phương sai mẫu hiệu chỉnh s_x^2 **5** (s_x) : độ lệch mẫu hiệu chỉnh s_x **6** (n) : kích thước mẫu**7** (\bar{y}) : trung bình mẫu \bar{y} **8** (σ_y^2) : phương sai mẫu \hat{s}_y^2 $\nabla \Rightarrow$ **1** (σ_y) : độ lệch mẫu \hat{s}_y $\nabla \Rightarrow$ **2** (s_y^2) : phương sai mẫu hiệu chỉnh s_y^2 $\nabla \Rightarrow$ **3** (s_y) : độ lệch mẫu hiệu chỉnh s_y hoặc tìm hệ số tương quan r_{XY} bằng cách:Nhấn **OPTION** \Rightarrow $\nabla \Rightarrow$ **4 (HỒI QUY)** \Rightarrow **3 (r)**

hoặc tìm hệ số tự do a bằng cách:

Nhấn **OPTION** \Rightarrow $\nabla \Rightarrow$ **4 (HỒI QUY)** \Rightarrow **1 (a)**

hoặc tìm hệ số góc b bằng cách:

Nhấn **OPTION** \Rightarrow $\nabla \Rightarrow$ **4 (HỒI QUY)** \Rightarrow **2 (b)**hoặc tìm $\overline{xy} = \frac{\sum xy}{n}$ bằng cách:Nhấn **OPTION** \Rightarrow $\nabla \Rightarrow$ **1 (PHÉP TÍNH TỔNG)** \Rightarrow **5** ($\sum xy$) \Rightarrow $\div \Rightarrow$ **OPTION** \Rightarrow $\nabla \Rightarrow$ **2 (BIẾN THỐNG KÊ)** \Rightarrow **6 (n)**