

Chương 5: TƯƠNG QUAN VÀ HỒI QUY

5.1. Tương quan

Cho hai biến ngẫu nhiên X, Y .

Tương quan (Correlation) giữa hai biến X, Y là khái niệm thể hiện mức độ của mối quan hệ giữa X và Y .

Hệ số tương quan (Correlation coefficient) là đại lượng đo lường cường độ của mối quan hệ giữa hai biến X và Y (không phân biệt biến nào là phụ thuộc, biến nào là độc lập).

Ký hiệu ρ : hệ số tương quan tổng thể. ($-1 \leq \rho \leq 1$).

Biểu đồ phân tán (scatter diagram): đồ thị thể hiện các cặp giá trị (X_i, Y_i)

5.2. Hệ số tương quan Pearson

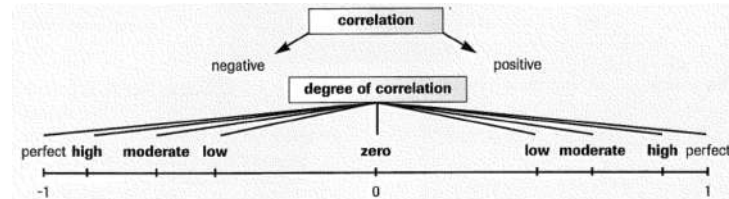
Cho hai biến ngẫu nhiên X, Y .

X	x_1	x_2	\dots	x_n
Y	y_1	y_2	\dots	y_n

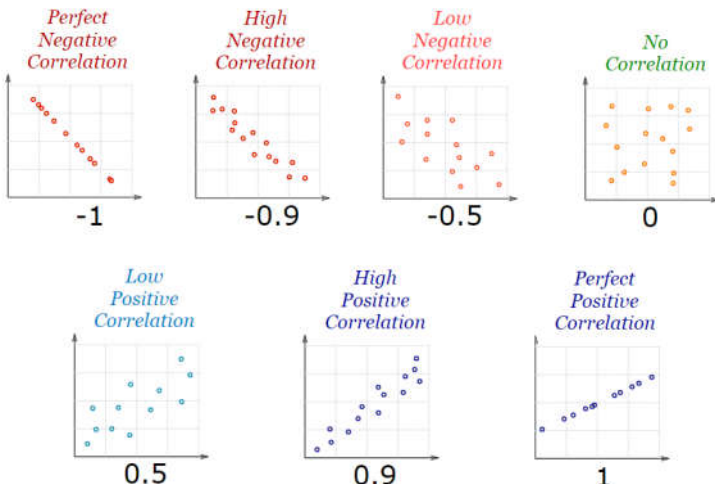
Hệ số tương quan mẫu (Pearson) của X và Y , ký hiệu r_{XY}

$$r_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2 \cdot \sum_{i=1}^n (y_i - \bar{Y})^2}}$$

$$= \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{X} \cdot \bar{Y}}{\sqrt{\left[\sum_{i=1}^n x_i^2 - n(\bar{X})^2 \right] \cdot \left[\sum_{i=1}^n y_i^2 - n(\bar{Y})^2 \right]}}$$



r_{XY}	Strength of Correlation
$ r_{XY} = 1$	Perfect
$0,75 < r_{XY} < 1$	High (Strong)
$0,5 < r_{XY} < 0,75$	Moderate
$0,2 < r_{XY} < 0,5$	Low (Weak)
$0 < r_{XY} < 0,2$	No Correlation



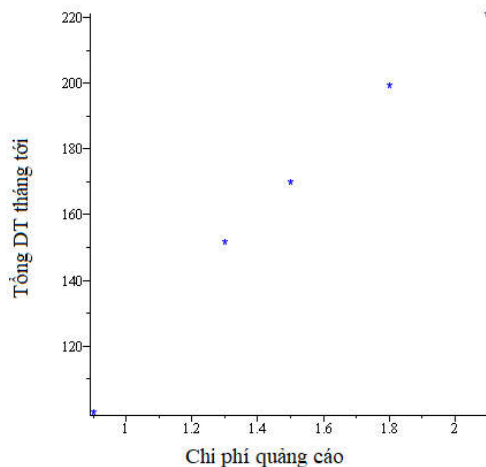
Ví dụ 78

Một công ty muốn nghiên cứu ảnh hưởng của quảng cáo (QC) tới doanh số (DS) bán hàng. Dữ liệu về quảng cáo và doanh thu một số tháng thu thập được như sau.

Chi phí QC (Tr.đ)	0,9	1,3	1,5	1,8	2,1
Tổng DS tháng tới (Tr.đ)	100,1	151,6	170	199,3	221,2

Vẽ biểu đồ phân tán và tính hệ số tương quan Pearson.

Biểu đồ phân tán



X	Y	X^2	Y^2	$X.Y$
0,9	100,1			
1,3	151,6			
1,5	170			
1,8	199,3			
2,1	221,2			
$\sum X =$	$\sum Y$	$\sum X^2$	$\sum Y^2$	$\sum X.Y$

$n =$ $\bar{X} =$ $\bar{Y} =$
 Hệ số tương quan:

$$r_{XY} = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sqrt{\left[\sum_{i=1}^n X_i^2 - n(\bar{X})^2 \right] \cdot \left[\sum_{i=1}^n Y_i^2 - n(\bar{Y})^2 \right]}} =$$

Nhận xét: . . .

Ví dụ 79

Có số liệu về thời gian quảng cáo (TGQC/tuần, đv: phút) trên truyền hình và lượng sản phẩm tiêu thụ (LTT/tuần, đv: 1000 sp) ở một công ty sản xuất đồ chơi trẻ em như sau

TGQC	28	37	44	36	47	35	26	29	33	32	31	28
LTT	41	32	49	42	38	33	27	24	35	30	34	25

Tính hệ số tương quan Pearson.

5.3. Vấn đề xếp hạng

Cho biến ngẫu nhiên $X = (X_1, \dots, X_n)$. Ký hiệu:

- R_{X_i} là hạng của X_i . - $R_X = (R_{X_1}, \dots, R_{X_n})$

Cách làm:

- Quy ước thứ tự xếp hạng trước khi làm.
- Đánh thứ tự.
- Nếu có các giá trị X_i bằng nhau \Rightarrow trung bình vị trí.

Ví dụ 80

Điểm trung bình năm 2018-2019 của 10 học sinh lớp 12A1 như sau

HS	A	B	C	D	E	F	G	H	I	J
Điểm TB(X)	9,8	7	9	8	9,8	10	6	9,8	7	5
Hạng(R_X)	3	7,5	5	6	3	1	9	3	7,5	10

5.4. Hệ số tương quan Spearman

Trường hợp X, Y được thể hiện dưới hình thức xếp hạng.

Bước 1:

-Xếp hạng các giá trị của X_i của biến X: $R_X = (R_{X_1}, \dots, R_{X_n})$.

-Xếp hạng các giá trị của Y_i của biến Y: $R_Y = (R_{Y_1}, \dots, R_{Y_n})$.

Bước 2: Tính hệ số tương quan hạng

$$r_{XY} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

trong đó: $d_i = R_{X_i} - R_{Y_i}$.

Ví dụ 81

Khi tuyển dụng, một công ty đánh giá các ứng viên thông qua phỏng vấn và bài kiểm tra. Khi phỏng vấn, các ứng viên được đánh giá từ A (xuất sắc) đến E (không phù hợp) và bài kiểm tra được tính theo thang điểm 100. Kết quả của 5 ứng viên như sau:

Ứng viên	28	37	44	36	47
Điểm phỏng vấn	A	B	A	E	D
Điểm bài thi	60	72	49	88	75

Tính hệ số tương quan hạng.

Giải

Ú.V	Điểm PV X	Hạng PV R_X	Điểm BT Y	Hạng BT R_Y	d	d^2
1	A		60			
2	B		72			
3	A		49			
4	E		88			
5	D		75			

Hệ số tương quan hạng

$$r_{XY} = \dots\dots\dots = -0,925$$

Ví dụ 82

Có số liệu về thời gian quảng cáo (TGQC/tuần, đv: phút) trên truyền hình và lượng sản phẩm tiêu thụ (LTT/tuần, đv: 1000 sp) ở một công ty sản xuất đồ chơi trẻ em như sau

TGQC	28	37	44	36	47	35	26	29	33	32	31	28
LTT	41	32	49	42	38	33	27	24	35	30	34	25

Tính hệ số tương quan Spearman.

5.5. Hồi quy

Hồi quy là phương pháp phân tích được dùng để xem xét mối liên hệ giữa hai biến X, và Y, trong đó

- X: biến độc lập (biến giải thích, ảnh hưởng đến biến Y).
- Y: biến phụ thuộc (biến được giải thích, chịu ảnh hưởng bởi X)

Ví dụ 83

- Trong phân tích ảnh hưởng của chi phí quảng cáo đến doanh số bán hàng, ta có
X = chi phí quảng cáo.
Y = doanh số bán hàng.
- Trong một cuộc khảo sát về việc đầu tư cho giáo dục và lợi ích thu được sau khi ra trường, ta có
X = đầu tư cho giáo dục.
Y = lợi ích thu được.

5.6. Hồi quy tuyến tính đơn giản (Simple Linear Regression)

Giả sử rằng mối quan hệ thực sự giữa x và Y là đường thẳng và quan sát Y tại mỗi giá trị x là một biến ngẫu nhiên thỏa

$$E(Y|x) = \beta_0 + \beta_1 x$$

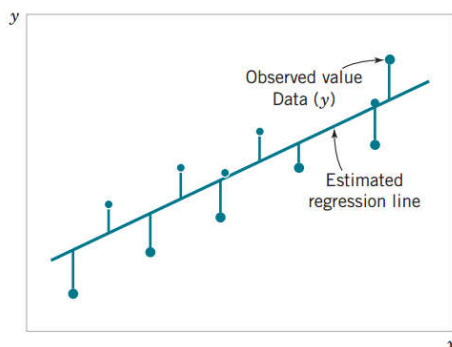
trong đó: tung độ gốc β_0 và hệ số góc β_1 là những hệ số hồi quy chưa biết.

Giả sử những giá trị quan sát y được biểu diễn thông qua mô hình

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad (12)$$

trong đó ε là sai số ngẫu nhiên có trung bình bằng 0 và phương sai σ^2 (không biết). Sai số của mỗi quan sát Y khác nhau được giả sử là những biến ngẫu nhiên không tương quan.

Giả sử có n cặp quan sát $(x_1, y_1); (x_2, y_2); \dots; (x_n, y_n)$ và một đường thẳng ứng viên cho đường hồi quy ước lượng như hình



Câu hỏi: đường nào (β_0, β_1) là tốt nhất cho bộ dữ liệu quan sát?

Karl Gauss (1777–1855) đề xuất ước tính β_0, β_1 bằng phương pháp bình phương tối thiểu.

Sử dụng phương trình 12, ta có biểu diễn của n cặp quan sát như sau

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon \quad (13)$$

và tổng bình phương sai số giữa các quan sát và đường hồi quy đúng là

$$L = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (14)$$

Giải bài toán cực tiểu phương trình 14, ta được

Định lý 0.1

Ước lượng bình phương cực tiểu của β_0 và β_1 là

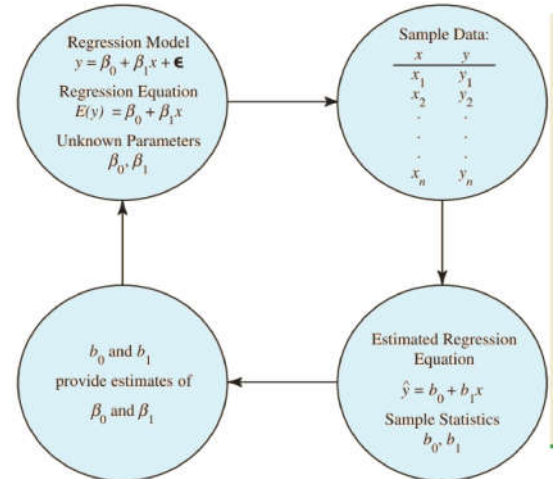
$$b_0 = \bar{y} - b_1 \bar{x} \quad (15)$$

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2} \quad (16)$$

Vậy, ước lượng cho đường thẳng hồi quy có dạng

$$\hat{y} = b_0 + b_1 x \quad (17)$$

Quá trình ước lượng trong hồi quy tuyến tính đơn giản



Ví dụ 84

Cho dữ liệu về số lượng sinh viên và doanh số bán hàng theo quý của 10 nhà hàng trong chuỗi cửa hàng ARMAND'S PIZZA PARLORS

Nhà hàng	Số lượng SV (1000 người)	Doanh số bán theo quý (1000 USD)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Giải.

Cách 1: Lập bảng tính.

NH I	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma (x_i - \bar{x})(y_i - \bar{y})$	$\Sigma (x_i - \bar{x})^2$

Từ bảng trên, ta tính được

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

$$b_0 = \bar{y} - b_1 \bar{x} = 130 - 5.14 = 60.$$

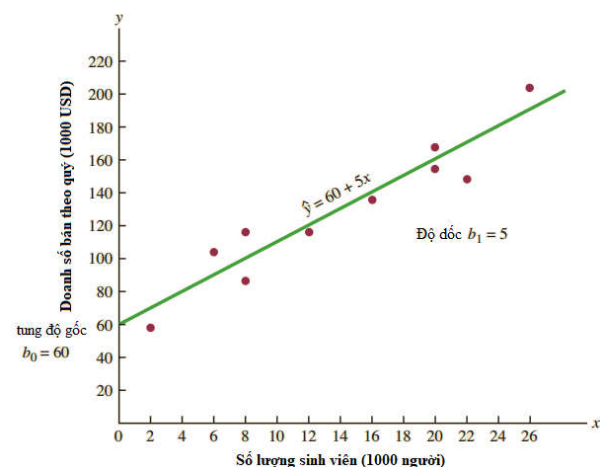
Phương trình hồi quy ước lượng là:

$$\hat{y} = 60 + 5x$$

Cách 2: dùng máy tính.

$$\begin{aligned} b_1 &= \\ b_0 &= \\ \Rightarrow \hat{y} &= \end{aligned}$$

Đồ thị PT hồi quy ước lượng ARMAND'S PIZZA PARLORS



Dùng máy tính tìm hệ số tương quan mẫu - phương trình hồi quy tuyến tính

Bước chuẩn bị:

Xóa bộ nhớ: $\text{Shift} \Rightarrow 9 \Rightarrow 2 \Rightarrow =$

Khai báo cột tần số:

$\text{Shift} \Rightarrow \text{Mode} \Rightarrow \downarrow \Rightarrow 4$ (chọn STAT) $\Rightarrow 1$ (chọn ON)

Nhập dữ liệu

$\text{Mode} \Rightarrow 3$ (chọn STAT) $\Rightarrow 2$ (chọn A+BX)

nhập dữ liệu xong nhấn AC .

Xuất kết quả

$\text{Shift} \Rightarrow 1 \Rightarrow 5$ (chọn Reg) \Rightarrow chọn kết quả cần xuất
($b_0 = 1$, $b_1 = 2$, $r = 3$).

Ví dụ 85

Đối với sở thuế (Internal Revenue Service - IRS), tính hợp lý của tổng số chi được khấu trừ thuế phụ thuộc vào tổng thu nhập đã được điều chỉnh của người nộp thuế. Các khoản chi được khấu trừ lớn, trong đó bao gồm chi đóng góp từ thiện và các đóng góp y tế, là các khoản hợp lý hơn cho người nộp thuế có tổng thu nhập điều chỉnh lớn. Ở một mức thu nhập nào đó, nếu người nộp thuế kê khai số chi cần phải khấu trừ thuế lớn hơn số chi được khấu trừ thuế trung bình, thì khả năng phải thực hiện kiểm toán IRS tăng lên. Dữ liệu (ngàn USD) về tổng thu nhập điều chỉnh và số chi khấu trừ trung bình hoặc hợp lý như sau

Tổng thu nhập điều chỉnh (1000 USD)	Số tiền chi khấu trừ thuế hợp lý (1000 USD)
22	9.6
27	9.6
32	10.1
48	11.1
65	13.5
85	17.7
120	25.5

- Hãy vẽ biểu đồ phân tán với tổng thu nhập điều chỉnh là biến độc lập.
- Dùng phương pháp bình phương nhỏ nhất để xây dựng pt hồi quy.
- Hãy ước tính mức hợp lý của tổng số chi được khấu trừ cho một người nộp thuế có tổng thu nhập điều chỉnh là 52 500 USD. Nếu người nộp thuế này khai số thuế chi được khấu trừ là 20 400 USD, thì nhân viên IRS có yêu cầu kiểm toán hay không?

5.7. Hệ số xác định

A. Tổng bình phương do sai số

(SUM OF SQUARES DUE TO ERROR - SSE)

Thể hiện phần biến thiên của y do các nhân tố không được nghiên cứu đến.

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

B. Tổng bình phương do hồi quy

(SUM OF SQUARES DUE TO REGRESSION - SSR)

Thể hiện phần biến thiên của Y được giải thích bởi biến X .

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (b_1^2) \cdot \left[\sum_{i=1}^n x_i^2 - n(\bar{x})^2 \right]$$

C. Tổng bình phương toàn bộ (TOTAL SUM OF SQUARES - SST)

Thể hiện toàn bộ biến thiên của Y .

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n(\bar{y})^2$$

Quan hệ giữa: SST, SSR, SSE:

$$SST = SSR + SSE$$

Hệ số xác định

$$R^2 = \frac{SSR}{SST}$$

Ý nghĩa:

Biến X giải thích được $(R^2 \cdot 100)\%$ sự thay đổi của biến phụ thuộc Y . Phần còn lại do các yếu tố ngẫu nhiên.

Tính chất:

- $0 \leq R^2 \leq 1$
- R^2 càng gần 1 càng tốt.

Ví dụ 86

Tính SSR, SSE, SST, R^2 trong ví dụ 84

Cách 1:

i	x_i	D.S y_i	DS dự đoán $\hat{y} = 60 + 5x$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	2	58	49	42	38	33	27
2	6	105					
3	8	88					
4	8	118					
5	12	117					
6	16	137					
7	20	157					
8	20	169					
9	22	149					
10	26	202					
					1530		15730

Ta có:

$$SST = 15730; \quad SSE = 1530$$

$$\Rightarrow SSR = SST - SSE = 15730 - 1530 = 14200$$

Suy ra:

$$R^2 =$$

Ta nói 90,27% sự thay đổi trong doanh số bán hàng có thể giải thích bởi mối quan hệ tuyến tính giữa số lượng SV và doanh số bán hàng.

Cách 2:

x_i	y_i	x_i^2	y_i^2
2	58	4	3364
6	105	36	11025
8	88	64	7744
8	118	64	13924
12	117	144	13689
16	137	169	118769
20	157	400	24649
20	169	400	28561
22	149	484	22201
26	202	676	40804
$\bar{X} = 14$	$\bar{Y} = 130$	$\sum = 2528$	$\sum = 184730$

$$SSR = (b_1^2) \cdot \left[\sum_{i=1}^n x_i^2 - n(\bar{x})^2 \right] = (5^2) [2528 - 10.14^2] = 14200.$$

$$SST = \sum_{i=1}^n y_i^2 - n(\bar{y})^2 = 184730 - 10.130^2 = 15730$$

$$R^2 = \dots$$

Cách 3: dùng máy tính.

Xem cuối chương.

Dữ liệu X là dãy số theo thời gian

Cho hai biến ngẫu nhiên X, Y.

X	x_1	x_2	\dots	x_n
Y	y_1	y_2	\dots	y_n

trong đó X là biến theo thời gian.

Cách giải

Cách 1: giải bình thường.

Cách 2: đánh số 1, 2, 3, ...

Cách 3: đánh số lại biến X: $x_i \Rightarrow t_i$ với điều kiện: $\sum t_i = 0$.

- Nếu: n lẻ, $t_i \in \{\dots, -4, -3, -2, -1, 0, 1, 2, 3, 4, \dots\}$.

- Nếu n chẵn, $t_i \in \{\dots - 5, -3, -1, 1, 3, 5, \dots\}$.

Ví dụ 87

Có tài liệu thống kê về sản lượng sản xuất của một công ty qua các năm như sau

Năm	2010	2011	2012	2013	2014	2015
Sản lượng (Tấn)	20	23	25	29	35	40

- Viết phương trình hàm hồi quy tuyến tính sản lượng sản xuất của công ty theo thời gian (theo 2 cách).
- Dự báo giá trị sản lượng sản xuất của công ty trong 2 năm tiếp theo
- Nếu năm 2016, sản lượng sản xuất của công ty là 44 tấn. Viết lại hàm hồi quy.

5.3. Hồi quy bội

Mô hình hồi quy bội

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon \quad (18)$$

trong đó $\beta_0, \beta_1, \dots, \beta_p$ là các tham số của mô hình.

ε sai số ngẫu nhiên.

Phương trình hồi quy bội

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (19)$$

Phương trình hồi quy bội ước lượng

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p \quad (20)$$

trong đó b_0, b_1, \dots, b_p là các ước lượng của $\beta_0, \beta_1, \dots, \beta_p$.

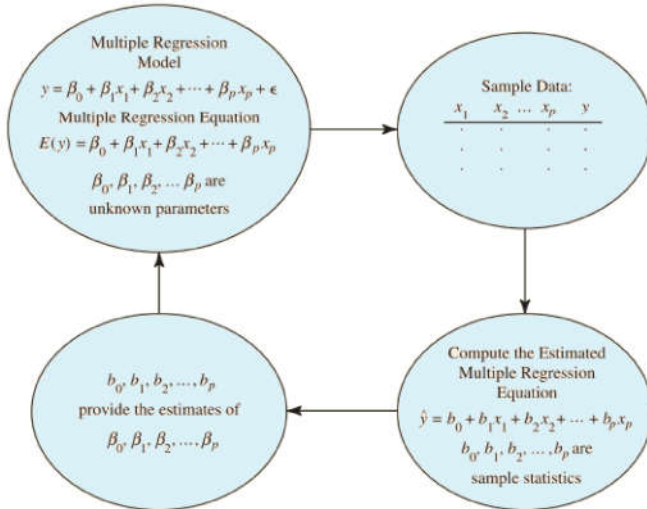
\hat{y} : giá trị ước lượng của biến phụ thuộc.

Hồi quy 3 biến

Cho biến ngẫu nhiên X_1, X_2 , và Y .

X_1	x_{11}	x_{12}	\cdots	x_{1n}
X_2	x_{21}	x_{22}	\cdots	x_{2n}
Y	y_1	y_2	\cdots	y_n

Mục tiêu: viết PT hồi quy: $\hat{y} = b_0 + b_1x_1 + b_2x_2$.



Phương pháp

Bước 1: Từ bảng dữ liệu, lập ma trận

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix}, \quad X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Đặt $B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$

Bước 2: Tính

$$- X^T \cdot X = \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} \cdot x_{2i} \\ \sum x_{2i} & \sum x_{1i} \cdot x_{2i} & \sum x_{2i}^2 \end{bmatrix}_{3 \times 3}$$

$$- (X^T \cdot X)^{-1}$$

$$- X^T \cdot Y = \begin{bmatrix} \sum y_i \\ \sum x_{1i} \cdot y_i \\ \sum x_{2i} \cdot y_i \end{bmatrix}_{3 \times 1}$$

Bước 3: Tính ma trận B. $B = (X^T X)^{-1} \cdot (X^T \cdot Y) = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$

Bước 4: viết PT hồi quy: $\hat{y} = b_0 + b_1x_1 + b_2x_2$.

Ví dụ 88

Cho dãy dữ liệu sau với x_1, x_2 là biến độc lập, y là biến phụ thuộc

X_1	30	47	25	51
X_2	12	10	17	16
Y	94	108	112	178

Viết PT hồi quy.

Giải

Cách 1:

$$X = \begin{bmatrix} 1 & 30 & 12 \\ 1 & 47 & 10 \\ 1 & 25 & 17 \\ 1 & 51 & 16 \end{bmatrix}, \quad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 30 & 47 & 25 & 51 \\ 12 & 10 & 17 & 16 \end{bmatrix}$$

$$Y = \begin{bmatrix} 94 \\ 108 \\ 112 \\ 178 \end{bmatrix}$$

$$\begin{aligned}
 X^T X &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 30 & 47 & 25 & 51 \\ 12 & 10 & 17 & 16 \end{bmatrix} \begin{bmatrix} 1 & 30 & 12 \\ 1 & 47 & 10 \\ 1 & 25 & 17 \\ 1 & 51 & 16 \end{bmatrix} \\
 &= \begin{bmatrix} 4 & 153 & 55 \\ 153 & 6335 & 2071 \\ 55 & 2071 & 789 \end{bmatrix} \\
 (X^T X)^{-1} &= \begin{bmatrix} \frac{354637}{29475} & \frac{-26}{450} & \frac{-15781}{58950} \\ \frac{29475}{-26} & \frac{225}{450} & \frac{29475}{1931} \\ \frac{225}{-15781} & \frac{450}{1} & \frac{450}{1931} \end{bmatrix} \\
 X^T Y &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 30 & 47 & 25 & 51 \\ 12 & 10 & 17 & 16 \end{bmatrix} \begin{bmatrix} 94 \\ 108 \\ 112 \\ 178 \end{bmatrix} = \begin{bmatrix} 492 \\ 19774 \\ 6960 \end{bmatrix}
 \end{aligned}$$

$$\begin{aligned}
 B &= (X^T X)^{-1} \cdot (X^T Y) \\
 &= \begin{bmatrix} \frac{354637}{29475} & \frac{-26}{450} & \frac{-15781}{58950} \\ \frac{29475}{-26} & \frac{225}{450} & \frac{29475}{1931} \\ \frac{225}{-15781} & \frac{450}{1} & \frac{450}{1931} \end{bmatrix} \cdot \begin{bmatrix} 492 \\ 19774 \\ 6960 \end{bmatrix} \\
 &= \begin{bmatrix} \frac{354637}{29475} & \frac{-26}{450} & \frac{-15781}{58950} \\ \frac{29475}{-26} & \frac{225}{450} & \frac{29475}{1931} \\ \frac{225}{-15781} & \frac{450}{1} & \frac{450}{1931} \end{bmatrix} \cdot \begin{bmatrix} 492 \\ 19774 \\ 6960 \end{bmatrix} \\
 \text{PT hồi quy: } \hat{y} &= \frac{-108184}{1179} + \frac{23}{9}x_1 + \frac{10033}{1179}x_2.
 \end{aligned}$$

Hệ số xác định

A. Tổng bình phương do sai số

(SUM OF SQUARES DUE TO ERROR - SSE)

Thể hiện phần biến thiên của Y do các nhân tố không được nghiên cứu đến.

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = Y^T Y - B^T X^T Y$$

B. Tổng bình phương do hồi quy

(SUM OF SQUARES DUE TO REGRESSION - SSR)

Thể hiện phần biến Y được giải thích bởi biến X_1, X_2 .

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = B^T \cdot (X^T Y) - n(\bar{y})^2$$

C. Tổng bình phương toàn bộ

(TOTAL SUM OF SQUARES - SST)

Thể hiện toàn bộ biến thiên của Y .

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n(\bar{y})^2$$

Quan hệ giữa: SST, SSR, SSE:

$$SST = SSR + SSE$$

Hệ số xác định

$$R^2 = \frac{SSR}{SST}$$

Ví dụ 89

Tìm hệ số xác định trong ví dụ 88.

Ma trận hệ số tương quan

$$\text{Ma trận hệ số tương quan: } R = \begin{bmatrix} 1 & r_{YX_1} & r_{YX_2} \\ r_{X_1Y} & 1 & r_{X_1X_2} \\ r_{X_2Y} & r_{X_2X_1} & 1 \end{bmatrix}.$$

Trong đó

$$r_{YX_1} = r_{X_1Y} = \frac{\sum_{i=1}^n x_{1i}y_i - n \cdot \bar{x}_1 \cdot \bar{y}}{\sqrt{\left[\sum_{i=1}^n x_{1i}^2 - n(\bar{x}_1)^2 \right] \cdot \left[\sum_{i=1}^n y_i^2 - n(\bar{y})^2 \right]}}$$

$$r_{YX_2} = r_{X_2Y} = \frac{\sum_{i=1}^n x_{2i}y_i - n \cdot \bar{x}_2 \cdot \bar{y}}{\sqrt{\left[\sum_{i=1}^n x_{2i}^2 - n(\bar{x}_2)^2 \right] \cdot \left[\sum_{i=1}^n y_i^2 - n(\bar{y})^2 \right]}}$$

$$r_{X_1X_2} = r_{X_2X_1} = \frac{\sum_{i=1}^n x_{1i}x_{2i} - n \cdot \bar{x}_1 \cdot \bar{x}_2}{\sqrt{\left[\sum_{i=1}^n x_{1i}^2 - n(\bar{x}_1)^2 \right] \cdot \left[\sum_{i=1}^n x_{2i}^2 - n(\bar{x}_2)^2 \right]}}$$

Ví dụ 90

Cho bảng dữ liệu của công ty vận tải Butler như sau

Nhiệm vụ vận chuyển	x1 = số dặm di chuyển	x2 = số lần vận chuyển	y = thời gian vận chuyển (giờ)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Viết phương trình hồi quy ước lượng. Tính R^2 và ma trận tương quan.

Hướng dẫn dùng máy tính casio

▷ Hồi quy đơn

Cho hai biến ngẫu nhiên X, Y .

X	x_1	x_2	\cdots	x_n
Y	y_1	y_2	\cdots	y_n

- Viết phương trình hồi quy ước lượng $\hat{y} = b_0 + b_1.x$, Tìm hệ số tương quan mẫu r_{XY} .
- Tìm SST, SSR, SSE, R^2 .

Cách giải:

1.

Bước chuẩn bị:

Xóa bộ nhớ: $\boxed{Shift} \Rightarrow \boxed{9} \Rightarrow \boxed{2} \Rightarrow \boxed{=}$

Khai báo cột tần số: $\boxed{Shift} \Rightarrow \boxed{Mode} \Rightarrow \boxed{\downarrow} \Rightarrow \boxed{4}$ (chọn STAT) $\Rightarrow \boxed{1}$ (chọn ON)

Nhập dữ liệu

$\boxed{Mode} \Rightarrow \boxed{3}$ (chọn STAT) $\Rightarrow \boxed{2}$ (chọn A+BX)

nhập dữ liệu xong nhấn \boxed{AC} .

Xuất kết quả

$\boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{5}$ (chọn Reg) \Rightarrow chọn kết quả cần xuất ($b_0 = \boxed{1}, b_1 = \boxed{2}, r_{XY} = \boxed{3}$).

2.

Tính $SSR = (b_1^2) \cdot \left[\sum_{i=1}^n x_i^2 - n(\bar{x})^2 \right]$

Tính $\sum_{i=1}^n x_i^2$: $\boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3}$ (chọn Sum) $\Rightarrow \boxed{1}$

Tính \bar{x} : $\boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{4}$ (chọn Var) $\Rightarrow \boxed{2}$

$\Rightarrow SSR$.

Tính $SST = \sum_{i=1}^n y_i^2 - n(\bar{y})^2$

Tính $\sum_{i=1}^n y_i^2$: $\boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3}$ (chọn Sum) $\Rightarrow \boxed{1}$

Tính \bar{y} : $\boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{4}$ (chọn Var) $\Rightarrow \boxed{5}$

$\Rightarrow SST$.

▷ Hồi quy bội 3

Cho biến ngẫu nhiên X_1, X_2 , và Y .

X_1	X_{11}	X_{12}	\cdots	X_{1n}
X_2	X_{21}	X_{22}	\cdots	X_{2n}
Y	Y_1	Y_2	\cdots	Y_n

- Viết phương trình hồi quy ước lượng $\hat{y} = b_0 + b_1x_1 + b_2x_2$.
- Tìm SST, SSR, SSE, R^2 .
- Tìm ma trận hệ số tương quan.

Cách giải:

1.

Bước chuẩn bị:

Xóa bộ nhớ: $\boxed{Shift} \Rightarrow \boxed{9} \Rightarrow \boxed{2} \Rightarrow \boxed{=}$

Khai báo cột tần số: $\boxed{Shift} \Rightarrow \boxed{Mode} \Rightarrow \boxed{\downarrow} \Rightarrow \boxed{4}$ (chọn STAT) $\Rightarrow \boxed{1}$ (chọn ON)

Nhập X_1, X_2

$\boxed{Mode} \Rightarrow \boxed{3}$ (chọn STAT) $\Rightarrow \boxed{2}$ (chọn A+BX)

X_1 nhập vào cột X, X_2 nhập vào cột Y.

Tính $X^T X$

$$X^T \cdot X = \begin{bmatrix} \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{2i}^2 & \sum x_{1i}x_{2i} \end{bmatrix}_{3 \times 3}$$

Trong đó

$$\sum x_{1i} = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{2}$$

$$\sum x_{2i} = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{4}$$

$$\sum x_{1i}^2 = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{1}$$

$$\sum x_{2i}^2 = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{3}$$

$$\sum x_{1i}x_{2i} = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{5}$$

Nhập $(X^T X)$ **vô máy dưới dạng ma trận** $A_{\text{máy}}$

Đặt $A_{\text{máy}} = X^T X$

Nhập ma trận A: $\boxed{Mode} \Rightarrow \boxed{6} \Rightarrow \boxed{1} \Rightarrow \boxed{1}$

Tính $X^T Y$

$$X^T \cdot Y = \begin{bmatrix} \sum y_i \\ \sum x_{1i} \cdot y_i \\ \sum x_{2i} \cdot y_i \end{bmatrix}_{3 \times 1}$$

Tính tương tự $X^T \cdot X$.

Thay X_2 trong máy tính bằng Y để tính $\sum y_i$ và $\sum x_{1i} \cdot y_i$.

Thay X_1 trong máy tính để tính tiếp $\sum x_{2i} \cdot y_i$.

Nhập ma trận $X^T Y$ vào máy dưới dạng **ma trận** $C_{\text{máy}}$: $C_{\text{máy}} = X^T Y$

Nhập ma trận C: $\boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{2} \Rightarrow \boxed{3} \Rightarrow \boxed{3}$

Tính ma trận B

$$B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{3} \Rightarrow \boxed{x^{-1}} \Rightarrow \boxed{\times} \Rightarrow \boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{5}$$

$$\Rightarrow B^T = \begin{bmatrix} b_0 & b_1 & b_2 \end{bmatrix}$$

Lưu ma trận B^T vào máy tính dưới dạng **ma trận** $B_{\text{máy}}$: $B_{\text{máy}} = B^T$

Nhập ma trận B^T : $\boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{2} \Rightarrow \boxed{2} \Rightarrow \boxed{\nabla} \Rightarrow \boxed{1}$

2. Tính SSR - SST

Tính \bar{Y}

$$\text{SSR: } SSR = \underbrace{B^T}_{B_{\text{máy}}} \cdot \underbrace{(X^T Y)}_{C_{\text{máy}}} - n(\bar{y})^2$$

Tính $B^T \cdot X^T Y$: $\boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{4} \Rightarrow \boxed{\times} \Rightarrow \boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{5}$

Tính \bar{y} .

$$\Rightarrow SSR$$

$$\text{SST: } SST = \sum_{i=1}^n Y_i^2 - n(\bar{Y})^2.$$

Ví dụ 1. Cho dữ liệu về số lượng sinh viên và doanh số bán hàng theo quý của 10 nhà hàng trong chuỗi cửa hàng ARMAND'S PIZZA PARLORS

Nhà hàng	Số lượng SV (1000 người)	Doanh số bán theo quý (1000 USD)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

- Viết phương trình hồi quy ước lượng.
- Tính R^2 .

Giải. Dùng máy tính

Máy

Bước chuẩn bị:

$\boxed{\text{Shift}} \Rightarrow \boxed{9} \Rightarrow \boxed{2} \Rightarrow \boxed{=}$

$\boxed{\text{Shift}} \Rightarrow \boxed{\text{Mode}} \Rightarrow \boxed{\downarrow} \Rightarrow \boxed{4} \Rightarrow \boxed{1}$

Nhập dữ liệu

$\boxed{\text{Mode}} \Rightarrow \boxed{3} \Rightarrow \boxed{2}$

Xuất kết quả

$b_0 : \boxed{\text{Shift}} \Rightarrow \boxed{1} \Rightarrow \boxed{5} \Rightarrow \boxed{1}$

$b_1 : \boxed{\text{Shift}} \Rightarrow \boxed{1} \Rightarrow \boxed{5} \Rightarrow \boxed{2}$

$r : \boxed{\text{Shift}} \Rightarrow \boxed{1} \Rightarrow \boxed{5} \Rightarrow \boxed{3}$

Tính SSR

$\boxed{\text{Shift}} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{1}$

$\boxed{\text{Shift}} \Rightarrow \boxed{1} \Rightarrow \boxed{4} \Rightarrow \boxed{2}$

Tính SST

$\boxed{\text{Shift}} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{1}$

$\boxed{\text{Shift}} \Rightarrow \boxed{1} \Rightarrow \boxed{4} \Rightarrow \boxed{5}$

.

Ghi bài

.

.

.

.

.

$$b_0 = 60$$

$$b_1 = 5$$

$$r = 0,9501$$

$$\Rightarrow \hat{y} = 60 + 5x$$

$$\sum X^2 = 2528$$

$$\bar{X} = 14$$

$$\Rightarrow SSR = (5^2) [2528 - 10.14^2] = 14200$$

$$\sum Y^2 = 184730$$

$$\bar{Y} = 130$$

$$\Rightarrow SSR = 184730 - 10.130^2 = 15730$$

$$R^2 \approx 0,9027$$

Ví dụ 2. Cho bảng dữ liệu của công ty vận tải Butler như sau

Nhiệm vụ vận chuyển	x1 = số dặm đi chuyển	x2 = số lần vận chuyển	y = thời gian vận chuyển (giờ)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Viết phương trình hồi quy ước lượng. Tính R^2 và ma trận tương quan.

Nhận xét: Với các bài có nhiều dữ liệu, ta không thể tính toán bằng cách nhân ma trận
 \Rightarrow máy tính hỗ trợ

Giải.

Bấm máy	Ghi bài
$\boxed{Shift} \Rightarrow \boxed{9} \Rightarrow \boxed{2} \Rightarrow \boxed{=}$ $\boxed{Shift} \Rightarrow \boxed{Mode} \Rightarrow \boxed{\downarrow} \Rightarrow \boxed{4} \Rightarrow \boxed{1}$ Nhập X_1, X_2 vào cột $X_{máy}, Y_{máy}$ $\boxed{Mode} \Rightarrow \boxed{3} \Rightarrow \boxed{2}$ $\sum x_{1i} = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{2}$ $\sum x_{2i} = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{4}$ $\sum x_{1i}^2 = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{1}$ $\sum x_{2i}^2 = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{3}$ $\sum x_{1i} \cdot x_{2i} = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{5}$ Nhập Y thay cho X_2 $\sum y_i = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{4}$ $\sum x_{1i} \cdot y_i = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{5}$	Tính $X^T \cdot X$ $X^T \cdot X = \begin{bmatrix} \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i} \cdot x_{2i} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{2i}^2 & \sum_{i=1}^n x_{1i} \cdot x_{2i} \end{bmatrix}_{3 \times 3}$ $X^T \cdot X = \begin{bmatrix} 10 & 800 & 29 \\ 800 & 67450 & 2345 \\ 29 & 2345 & 91 \end{bmatrix}_{3 \times 3}$ Tính $X^T \cdot Y$ $X^T \cdot Y = \begin{bmatrix} \sum y_i \\ \sum x_{1i} \cdot y_i \\ \sum x_{2i} \cdot y_i \end{bmatrix}_{3 \times 1}$ $X^T \cdot Y = \begin{bmatrix} 67 \\ 5594 \\ \sum x_{2i} \cdot y_i \end{bmatrix}_{3 \times 1}$

Bấm máy	Ghi bài
Nhập X_2 vào máy thay cho X_1	
$\sum x_{2i} \cdot y_i = \boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{5}$	$X^T \cdot Y = \begin{bmatrix} 67 \\ 5594 \\ 202,2 \end{bmatrix}_{3 \times 1}$
Nhập ma trận $X^T X$ vào ma trận $A_{\text{máy}}$	
$\boxed{Mode} \Rightarrow \boxed{6} \Rightarrow \boxed{1} \Rightarrow \boxed{1} : \text{nhập}$	
Nhập ma trận $X^T Y$ vào ma trận $C_{\text{máy}}$	
$\boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{2} \Rightarrow \boxed{3} \Rightarrow \boxed{3} : \text{nhập}$	
Tính B : $\boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{3} \Rightarrow \boxed{x^{-1}}$	Tính $B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -0,8687 \\ 0,0611 \\ 0,9234 \end{bmatrix}$
$\Rightarrow \boxed{\times} \Rightarrow \boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{5}$	
Nhập ma trận B^T vào ma trận $B_{\text{máy}}$	PT hồi quy: $\hat{y} = -0,8687 + 0,0611 \cdot x_1 + 0,9234 \cdot x_2$
$\boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{2} \Rightarrow \boxed{2} \Rightarrow \boxed{\nabla} \Rightarrow \boxed{1}$	$\Rightarrow B^T = \begin{bmatrix} -0,8687 & 0,0611 & 0,9234 \end{bmatrix}$
Tính $B^T \cdot (X^T Y)$	Tính $SSR = \underbrace{B^T}_{B_{\text{máy}}} \cdot \underbrace{(X^T Y)}_{C_{\text{máy}}} - n(\bar{y})^2$
$\boxed{Shift} \Rightarrow \boxed{4} \Rightarrow \boxed{4} \Rightarrow \boxed{\times} \Rightarrow \boxed{Shift}$	$B^T \cdot (X^T Y) = 470,302$
$\Rightarrow \boxed{4} \Rightarrow \boxed{5}$	$\bar{y} = 6,7$
Tính \bar{y}	$SSR = 470,302 - 10 \times 6,7^2 = 21,402$
$\boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{4} \Rightarrow \boxed{5}$	Tính $SST = \sum_{i=1}^n Y_i^2 - n(\bar{y})^2$
Tính $\sum_{i=1}^n Y_i^2$	$\sum_{i=1}^n Y_i^2 = 472,8$
$\boxed{Shift} \Rightarrow \boxed{1} \Rightarrow \boxed{3} \Rightarrow \boxed{2}$	$SST = 472,8 - 10 \times 6,7^2 = 23,9$
	$R^2 = \frac{21,402}{23,9} = 0,9055$