

Bài 4 HỒI QUY TƯƠNG QUAN

Nếu có cặp biến ngẫu nhiên X, Y thì để đánh giá mối quan hệ tuyến tính giữa X và Y ta tính hệ số tương quan r_{XY} .

Hệ số tương quan lý thuyết được định nghĩa là

$$\rho_{XY} = \frac{\sqrt{E\{(X - EX)(Y - EY)\}}}{\sqrt{DX \cdot DY}} = \frac{\sqrt{E\{(X - EX)(Y - EY)\}}}{\sigma_X \sigma_Y}$$

với E là ký hiệu của kỳ vọng toán học, D là phương sai, σ là độ lệch chuẩn

Trong xử lý số liệu sau khi thu thập số liệu ta để n cặp số liệu thành 2 cột hay hai hàng (nên để 2 cột vì thuận tiện cho các thao tác sau này).

Nếu tính trực tiếp có thể dùng công thức sau:

$$\text{Phương sai của } X \quad s^2_X = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{(n-1)}$$

$$\text{Phương sai của } Y \quad s^2_Y = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{(n-1)} = \frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{(n-1)}$$

$$\text{Hiệp phương sai } XY \quad \text{covar}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{(n-1)}$$

Trong các công thức trên có thể tính tổng bằng hàm Sum(dãy số), tính tổng bình phương bằng hàm Sumsq (dãy số), tổng các tích số bằng hàm Sumproduct (dãy1, dãy 2)

Nếu dùng hàm định sẵn thì $s^2_X = \text{var}(\text{dãy}X)$ $s^2_Y = \text{var}(\text{dãy}Y)$

Hiệp phương sai = Covar(dãyX, dãyY)

Hệ số tương quan $r_{XY} = \text{Covar}(X, Y) / (s_X s_Y)$ hoặc hàm định sẵn Correl(dãyX, dãyY)

Hệ số tương quan dương kết luận tương quan thuận, âm kết luận tương quan nghịch

Để kiểm định giả thiết

$H_0: \rho_{XY} = 0$ đối thiết $H_1: \rho_{XY} \neq 0$ có thể tính $T_m = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ sau đó

so với $T_{lt} = \text{Tinv}(\alpha/2, n-2)$ tức là giá trị trong bảng Student ở mức ý nghĩa α với $n - 2$ bậc tự do.

Tiếp theo là tìm đường hồi quy tuyến tính Y theo X , đường thẳng này có phương trình

$$Y = a + bX$$

a là tung độ gốc, b là hệ số góc

b tính theo công thức $b = r s_Y / s_X$ $a = \bar{y} - a \bar{x}$

Dùng hàm định sẵn thì $a = \text{Intercept}(\text{dãy } Y, \text{dãy } X)$, $b = \text{Slope}(\text{dãy } Y, \text{dãy } X)$

Cũng có thể dùng hàm Linest (Dãy Y , dãy X , 0,1) để tìm đầy đủ a, b, r^2 , các sai số ...

Khi có một bảng nhiều biến thì nên vào Data Analysis để tính một bảng các hệ số tương quan giữa các biến bằng mục Correlation, tính một bảng các phương sai và hiệp phương sai bằng mục Covariance và tìm hồi quy tuyến tính hoặc bội tuyến tính bằng mục Regression, hai mục đầu không có gì đặc biệt, chỉ việc khai báo miền vào (Input range) bao gồm tất cả các cột số liệu, sau đó chọn Label (nếu có tên cột ở dòng đầu), miền ra (Output range)

Phần Regression nên chọn Option Residuals và Line fit plots sẽ được một bảng tóm tắt các thống kê, một bảng phân tích phương sai, một bảng các giá trị của các hệ số hồi quy cùng các kiểm định Student, một bảng các phần dư (Residu) và các đồ thị. Nếu làm hồi quy bội tuyến tính thì chú ý miền X bao gồm các cột chứa các biến, đặt sát nhau, bao gồm cả hàng đầu chứa tên biến, nếu có k biến thì sẽ được k đồ thị hồi quy đơn tuyến tính của Y theo biến 1, Y theo biến 2, . . . , Y theo biến k.

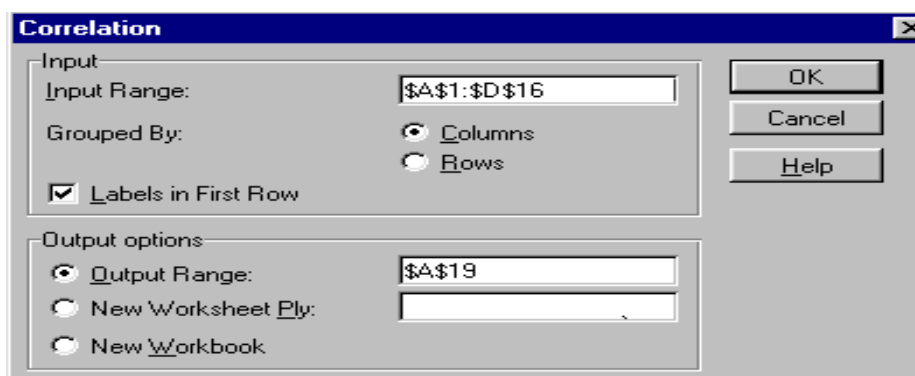
X1	X2	X3	Y
52	40	81	5.5
33	37	90	2.1
72	95	66	20.5
15	58	40	9.6
40	20	75	1.7
32	41	80	3.8
76	54	83	10.3
10	85	70	11.7
68	70	65	15.2
57	109	45	24.4
24	62	64	9.3
46	75	71	13
35	55	82	6.5
54	68	63	13.8
66	82	68	16.6

Bảng hệ số t-ong quan				
	X1	X2	X3	Y
X1	1			
X2	0.27507	1		
X3	0.09303	-0.58	1	
Y	0.49827	0.94999	-0.6297	1

Bảng hiệp ph-ong sai				
	X1	X2	X3	Y
X1	424.095			
X2	134.643	564.971		
X3	26.1667	-188.3	186.55	
Y	67.4238	148.371	-56.512	43.2

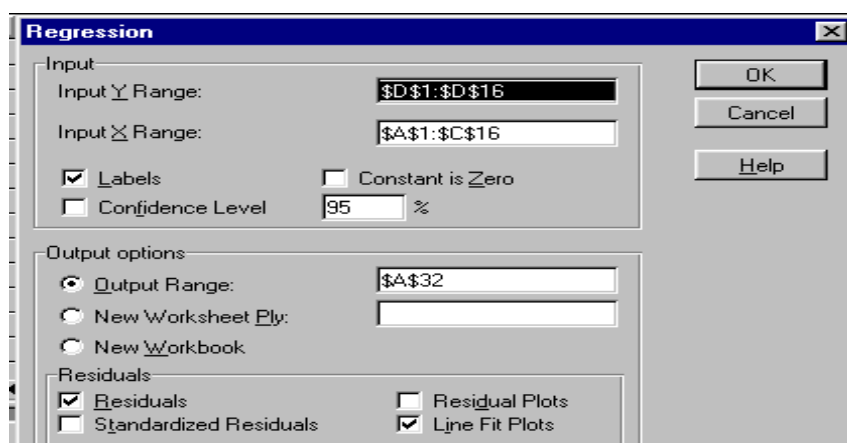
Trên đ-òng chéo là ph-ong sai của các biến X1, X2, X3, Y

Input Range trong mục Correlation bao gồm toàn bộ các cột chứa X1, X2, X3, Y Vì hàng đầu có tên các biến nên khai LabelsV, kết quả ta được một bảng, tại các chỗ giao nhau giữa hàng và cột ta được hệ số tương quan r_{X1X2} , r_{X1X3} , . . .



Tương tự như vậy vào mục Covariance và khai báo như mục Correlation, kết quả được một bảng, trên đường chéo là các phương sai còn tại chỗ giao nhau giữa hàng và cột ta được hiệp phương sai, đem hiệp phương sai của hai biến chia cho căn của tích số hai phương sai ta được hệ số tương quan.

Nếu chọn Regression thì Input Y Range là cột chứa Y, Input X Range gồm các cột chứa X1, X2, X3. Chọn Labels vì hàng đầu có tên các biến. Bỏ qua Constant is Zero và Confidence level (tức là để nguyên 95%). Phần Options chọn Residuals và Line fit plots còn bỏ qua Standardized residuals, Residual plots, Normal Probability Plot.



Kết quả nh- sau:

Summary statistics

<i>Regression Statistics</i>	
Multiple R	0.99891
R Square	0.99783
adjusted R	0.99723
Square	
Standard error	0.34561
observations	15

Anova

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	603.139	201.05	1683	6.38E-15	Mức ý nghĩa
Residual	11	1.31388	0.1194	F thực nghiệm		
Total	14	604.453	se ²			

Các hệ số trong hồi quy bội tuyến tính

	<i>Coefficients</i>	<i>Standard error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	1.69692	0.80419	2.1101	0.06	-0.0731	3.466928
X1	0.10283	0.00493	20.863	0	0.091978	0.113673
X2	0.19943	0.00522	38.212	0	0.187947	0.210922
X3	-0.116	0.00877	-13.232	0	-0.13535	-0.09674

Hệ số Sai số Giá trị T Mức ý nghĩa Cận dưới Cận trên
Phần dư (sai số)

Residuals

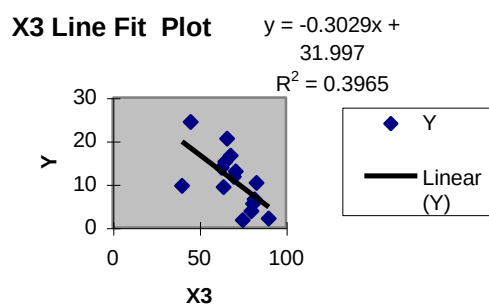
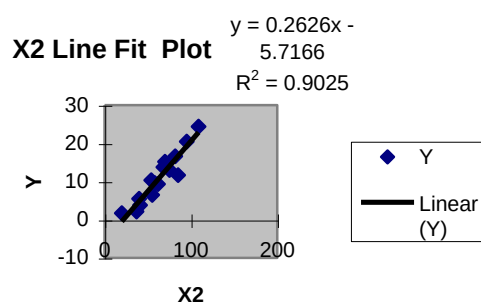
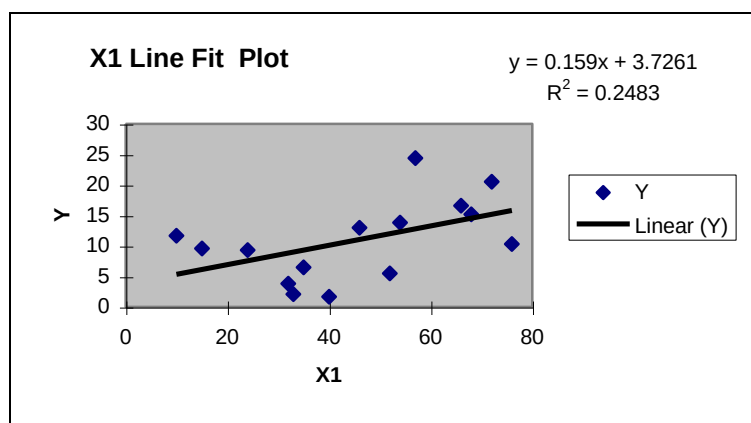
Quan sát	Dự báo theo hồi quy	Phần dư
<i>observation</i>	<i>Predicted</i>	<i>Residual</i>
<i>Y</i>	<i>Yhq</i>	<i>e</i>
1	5.62139	-0.1214
2	2.02496	0.07504
3	20.3875	0.11247
4	10.1646	-0.5646
5	1.09507	0.60493
6	3.88035	-0.0804
7	10.6492	-0.3492
8	11.5538	0.14621
9	15.1064	0.0936
10	24.0742	0.32578
11	9.10264	0.19736
12	13.1451	-0.1451
13	6.74882	-0.2488
14	13.5001	0.29993
15	16.9458	-0.3458

Chú thích



$$e = Y - Y_{hq}$$

Ngoài ra còn đ-ợc 3 đồ thị hồi quy tuyến tính đơn Y theo X1, Y theo X2, Y theo X3



Ba đồ thị này khi vẽ ra có 2 dãy điểm, dãy quan sát Y, dãy dự báo theo hồi quy mẫu Y_{hq} , ta có thể vào chế độ đồ thị xóa dãy Y_{hq} , sau đó nháy sáng dãy Y, Insert Trend line để vẽ đường hồi quy và thêm phương trình, thêm hệ số R^2 .

Vì Excel muốn giải quyết nhiều vấn đề nên ở phần thống kê chỉ tính toán những điểm chủ yếu, chứ không đi sâu phân tích như ở các bộ chương trình chuyên dụng để tính thống kê SPSS, Statistica, SAS, MstatC, Statgraphics, . . . Ngoài các phần trình bày ở trên còn có phần tạo số ngẫu nhiên, chọn mẫu, tìm đường trung bình trượt, làm trơn số liệu, biến đổi Fourier. . . mà do khuôn khổ của bài giảng chúng ta không đề cập tới.