Chương 4

ÁP DỤNG MS-EXCEL TRONG PHÂN TÍCH TƯƠNG QUAN VÀ HỒI QUY

Phân tích tương quan

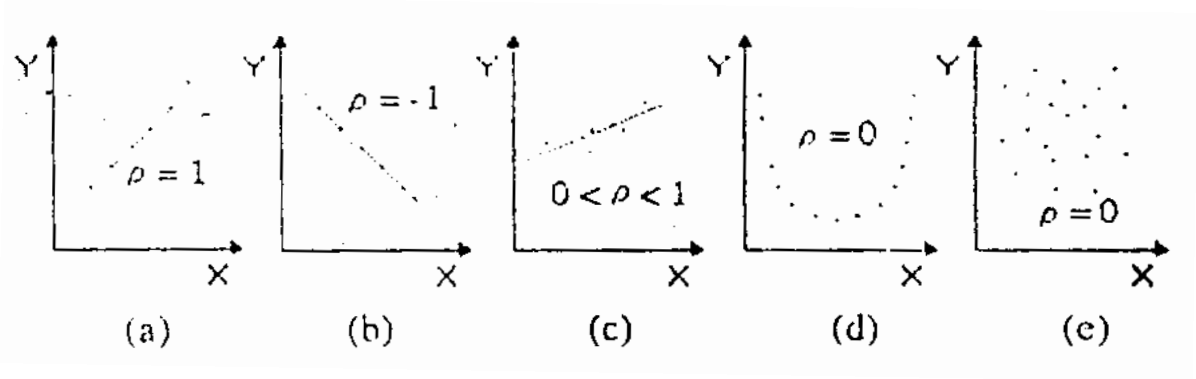
Phân tích hồi quy

- **Đơn giản**
- **Đa tham số**

A- PHÂN TÍCH TƯƠNG QUAN

6.1 Khái niệm thống kê

Hai biến số ngẫu nhiên Y và X có thể: liên quan tuyến tính (a và b), có khuynh hướng tuyến tính (c) hoặc không có liên quan (d và c).



Hệ số tương quan Pearson:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_X};\ \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)^2;\ \sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu_Y)^2$$

Sự phân tích **tương quan** *(correlation)* khảo sát *khuynh hướng* và **mức độ** của sự liên quan, trong sự phân tích **hồi quy**(regrestion) xác định sự liên quan định lượng giữa hai biến số ngẫu nhiên Y và X. Hệ số tương quan có thể được ước tính bởi biểu thức:

$$\hat{\rho} = R = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Hệ số tương quan được dùng trong việc đánh giá mức độ liên quan:

Giá trị |R|	Mức độ
<0,70	Nghèo nàn
0,70-0,80	Khá
0,80-0,90	Tốt
<0,90	Xuất sắc

6.2 Áp dụng MS-EXCEL

Thí dụ 16: Người ta tiến hành song song hai thí nghiệm lão hóa cấp tốc một dạng thuốc với hai điều kiện: độ ẩm 90% và nhiệt độ 60^{o}C.Tỷ lệ phân hủy *(%)* của hoạt chất theo thời gian *(phút)* như sau:

Thời gian	5	10	15	20	25
Độ ẩm	3,5	5,1	5,8	6,7	7,1
Nhiệt độ	2,7	3,2	4,7	6,1	6,2

Giữ độ ẩm, nhiệt độ và thời gian có liên quan như thế nào?

6.2.1 Nhập dữ liệu vào bảng tính

	A	B	C
1	Thời gian	Độ ẩm	Nhiệt độ
2	5	3,5	2,7
3	10	5,1	3,2
4	15	5,8	4,7
5	20	6,7	6,1
6	25	7,1	6,2

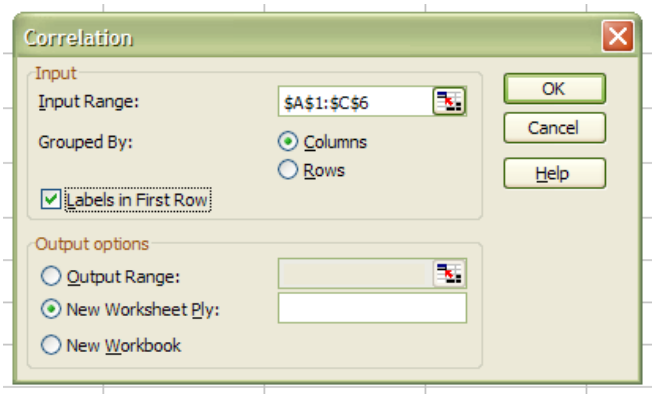
6.2.3 Áp dụng "Correlation"

a- Nhấp lần lượt đơn lệnh Tools và lệnh Data Analysic

b- Chọn phương trình Correlation trong hợp thoại Data Analysic rồi nhấp nút OK.

c- Trong hợp Correlation, lần lượt ấn định các chi tiết:

- Phạm vi đầu vào *(Input Range),*
- Cách xắp xếp theo hàng hay cột *(Group By),*
- Nhãn dữ liệu *(Labels Fisrt Row/Column),*
- Phạm vi đầu ra *(Output Range*)



Hộp thoại Correlation

	Thời gian	*Độ âm*	*Nhiệt độ*
Thời gian	1		
Độ ẩm	0.974654263	1	
Nhiệt độ	0.971335416	0.952366944	1

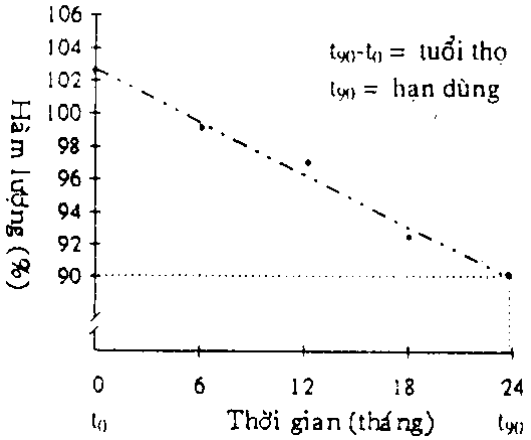
Kết quả

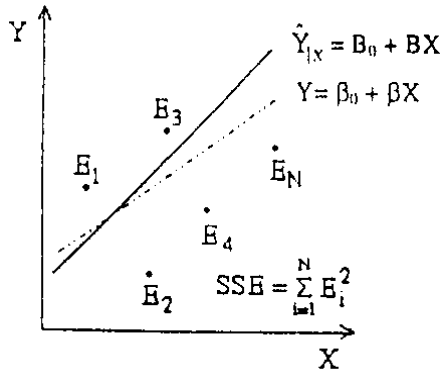
Các hệ số tương quan: R *(ẩm/thời gian)* = 0,97; R(nhiệt/thời gian) = 0,97 và R *(ẩm / nhiệt)* = 0,95

B- PHÂN TÍCH HỒI QUY

6.4 Khái nhiệm thống kê

Phép phâm tích hồi quy tuyến tính (liner regression) hay được áp dụng trong khoa học. Thí dụ, đường hồi quy (regression line / line of best fit) thường dùng để dự đoán về tuổi thọ hay hạn dùng của thuốc





(Lý thuyết)

(Ước tính)

Phương trình hồi quy có thể được ước tính bằng phương pháp bình phương cực tiểu *(least-squares estimation).*

C- HỒI QUY TUYẾN TÍNH ĐƠN GIẢN

6.5 Phương trình tổng quát

$\hat{Y}_{|X} = B_0 + BX$

$B_0 = \bar{Y} - B\bar{X}$

$B = \dfrac{\sum X_i Y_i - \sum X_i Y_i / N}{X_i - \bar{X}}$

Y: biến số phụ thuộc

(dependent / reponse variable)

X: là biến số độc lập

(independent / predictor variable)

B_0 và B là các hệ số hồi quy

(regression coefficients)

Bảng ANOVA

Nguồn sai số	Bậc tự do	Tổng số bình phương	Bình phương trung bình	Giá trị thống kê
Hồi quy	1	$SSR = \sum (Y_i^{'} - \bar{Y}^{'})^2$	MSR = SSR	$F = \dfrac{MSR}{MSE}$
Sai số	N − 2	$SSE = \sum (Y_i - Y_i^{'})^2$	MSE = SSE/(N-2)	
Tổng cộng	N −1	$SST = \sum (Y_i - \bar{Y})^2$ = SSR + SSE		

Giá trị thống kê

Giá trị R bình phương (R square):

$R = \dfrac{SSR}{SST}$ *(100R^2: % của biến đổi trên Y được giải thích bởi X)*

Độ lệch chuẩn (Standard Error):

$S = \sqrt{\dfrac{1}{N-2} \sum (Y_i - Y_i^{'})^2}$

(Sự phân tán của dữ liệu càng ít thì giá trị của S càng gần zero)

Trắc nghiệm thống kê

Đối với một phương trình hồi quy, $\hat{Y}_{|X} = B_0 + BX$, ý nghĩa thống kê của các hệ số B_i *(B_0 hay B)* được đánh giá bằng trắc nghiệm t *(phân phối Student)* trong khi tính chất thích hợp của phương trình $\hat{Y}_{|X} = f(X)$ được đánh giá bằng trắc nghiệm F *(phân bố Fischer)*

Trắc nghiệm t

- Giả thuyết:

H_0: $\beta_i = 0$ "Hệ số hồi quy không có ý nghĩa"

H_0: $\beta_i \neq 0$ "Hệ số hồi quy có ý nghĩa"

- Giá trị thống kê:

$$t = \frac{|B_i - \beta_i|}{\sqrt{S_n^2}};\ S_n^2 = \frac{S^2}{\sum (X_i - \bar{X})^2}$$

$$= \frac{B}{\sqrt{S_n^2}}$$

Phân bố Student $\gamma = N-2$

- Biện luận:

Nếu $t < t_\alpha (N-2) \Rightarrow$ Chấp nhận giả thuyết H_0 .

Trắc nghiệm F

- Giả thuyết:

H_0: $\beta_i = 0$ "Phương trình hồi quy không thích hợp"

H_0: $\beta_i \neq 0$ "Phương trình hồi quy thích hợp"

- Giá trị thống kê: $F = \frac{MSR}{MSE}$

Phân bố Fischer $v_1 = 1$, $v_2 = N-2$

- Kết luận:

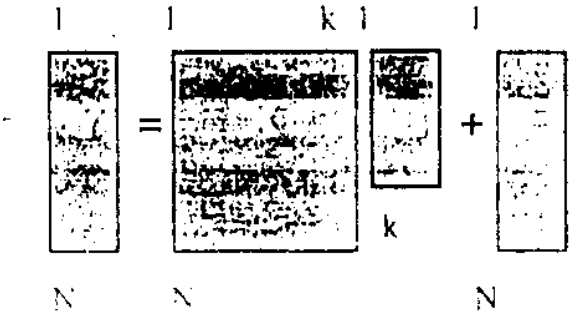
Nếu $F < F_\alpha (1,N-2) \Rightarrow$ Chấp nhận giả thuyết H_0 .

D- HỒI QUY TUYẾN TÍNH ĐA THAM SỐ

Trong phương trình hồi quy tuyến tính đa tham số biến số phụ thuộc Y có liên quan đến k biến số độc lập X_i (i = 1,2,…k) thay vì chỉ có một như trong hồi quy tuyến tính đơn giản.

Phương trình tổng quát : $\widehat{Y}_{|X_0,X_1,...,X_k} = B_0 + B_1X_1 + B_2X_2 + ... + B_kX_k$

Phương trình hồi quy đa tham số có thể được trình bày dưới dạng ma trận:



Bảng ANOVA

Nguồn sai số	**Bậc tự do**	**Tổng số bình phương**	**Bình phương trung bình**	**Giá trị thống kê**
Hồi quy	k	SSR	MSR = SSR/k	$F = \dfrac{MSR}{MSE}$
Sai số	N – k – 1	SSE	MSE = SSE/(N-k-1)	
Tổng cộng	N – 1	SST= SSR + SSE		

Giá trị thống kê:

Giá trị R bình phương:

Giá trị R^2 được hiệu chỉnh *(Adjusted R Square)*

$$R^2 = \frac{SSR}{SST} = \frac{kF}{(N-k-1)+kF}$$

($R^2 \geq 0,81$ là khá tốt)

Giá trị R^2 được hiệu chỉnh *(Adjusted R square):*

$$R_{ii}^2 = \frac{(N-1)R^2 - k}{N-k-1} = R^2 - \frac{k(1-R^2)}{(N-k-1)}$$

(R_{ii}^2 sẽ trở nên âm hay không xác định nếu R^2 hay N nhỏ)

Độ lệch chuẩn:

$$S = \sqrt{\frac{SSE}{(N-k-1)}} \quad (S \le 0{,}30 \text{ là khá tốt})$$

Trắc nghiệm thống kê

Tương tự hồi quy đơn giản, song bạn cần chú ý:

- Trong trắc nghiệm t

H_0: $\beta_i = 0$ "Các hệ số hồi quy không có ý nghĩa"

H_0: $\beta_i \neq 0$ "Có ít nhất vài hệ số hồi quy có ý nghĩa"

Bậc tự do của giá trị t: $\gamma = N - k - 1$.

$$t = \frac{|B_i - \beta_i|}{\sqrt{S_n^2}};\ S_n^2 = \frac{S^2}{\sum (X_i - \bar{X})^2}$$

- Trong trắc nghiệm F:

H_0: $\beta_i = 0$ "Phương trình hồi quy không thích hợp"

H_0: $\beta_i \neq 0$ "Phương trình hồi quy thích hợp" với ít nhất vài B_i .

Bậc tự do của giá trị F: $v_1 = 1$; $v_2 = N-k-1$.

Áp dụng MS-EXCEL

Thí dụ 17: Người ta đã dùng ba mức nhiệt độ gồm 105, 120 và 135^{0}C kết hợp với ba khoảng thời gian là 15, 30 và 60 phút để thực hiện một phản ứng tổng hợp. Các hiệu suất của phản ứng (%) được trình bày trong bảng sau đây:

Thời gian (phút) X_1	Nhiệt độ (^{0}C) X_2	Hiệu suất (%) Y
15	105	1.87
30	105	2.02
60	105	3.28
15	120	3.05
30	120	4.07
60	120	5.54
15	135	5.03
30	135	6.45
60	135	7.26

Hãy cho biết yếu tố nhiệt độ và/ hoặc yếu tố thời gian có liên quan tuyến tính với hiệu suất của phản ứng tổng hợp ? Nếu có thì điều kiện nhiệt độ 115^{0}C trong vòng 50 phút thì hiệu suất phản ứng sẽ là bao nhiêu?

Nhập dữ liệu vào bảng tính

Dữ liệu nhất thiết phải được nhất thiết được nhập theo cột:

	A	B	C
1	X_1	X_2	Y
2	15	105	1.87
3	30	105	2.02
4	60	105	3.28
5	15	120	3.05
6	30	120	4.07
7	60	120	5.54
8	15	135	5.03
9	30	135	6.45
10	60	135	7.26

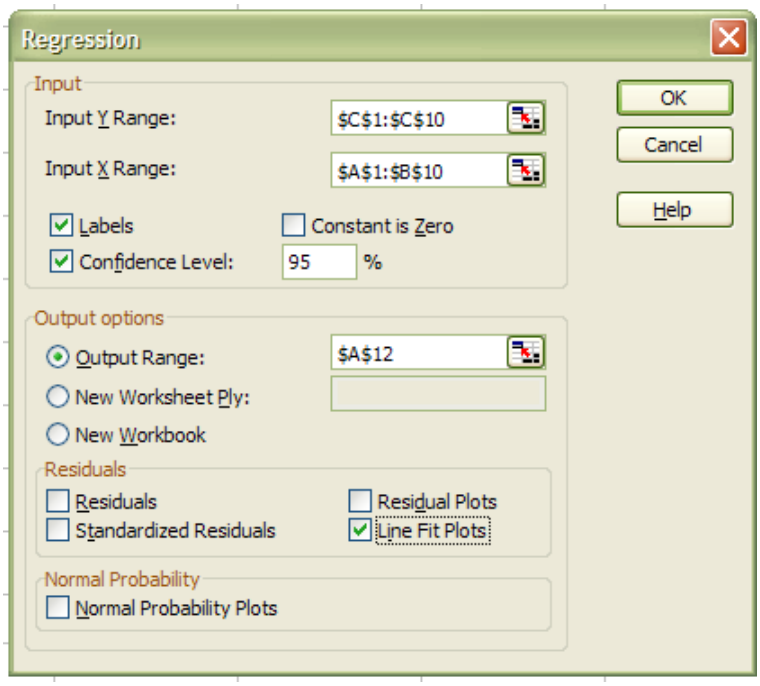
Sử dụng "Regression"

Nhấn lần lượt đơn lệnh Tools và lệnh Data Analysis.

Chọn chương trình Regression trong hộp thoại Data Analysis rồi nhấp OK.

Trong hộp thoại Regression, lần lượt ấn định các chi tiết:

- Phạm vi của biến số Y (Input Y Range).
- Phạm vi của biến số X (Input Y Range)
- Nhãn dữ liệu (Labels)
- Mức tin cậy (Confidence Level)
- Tọa độ đầu ra (Output Range)
- Và một số tùy chọn khác như đường hồi quy (Line Fit Plots), biểu thức sai số (Residuals Plots)…



Hộp thoại Regression

Phương trình hồi quy $\widehat{Y}_{|X_1} = f(X_1)$

$\widehat{Y}_{|X_1} = 2,73 + 0,04X_1$

(R^2 = 0,21; S=1,81)

Regression Statistics					
Multiple R	0.462512069				
R Square	0.213917414				
Adjusted R Square	0.101619901				
Standard Error	1.811191587				
Observations	9				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6.24891746	6.24891746	1.904917	0.209994918
Residual	7	22.96290476	3.280414966		
Total	8	29.21182222			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	2.726666667	1.280705853	2.129034282	0.070771	-0.301721453
X1	0.044539683	0.032270754	1.38018722	0.209995	-0.031768525

$t_0 = 2,19 < t_{0,05} = 2,365$ (Hay $P_V^2 = 0,071 > \alpha = 0,05$)

$\Rightarrow$ Chấp nhận giả thuyết H_0.

$t_1 = 1,38 < t_{0,05} = 2,365$ (Hay $P_V = 0,209 > \alpha = 0,05$)

$\Rightarrow$ Chấp nhận giả thuyết H_0.

$F = 1,905 < F_{0,05}^3 = 5,590$ (Hay $F_S^4 = 0,209 > \alpha = 0,05$)

$\Rightarrow$ Chấp nhận giả thuyết H_0.

Vậy cả hai hệ số 2,37(B_0) và 0,04(B_1) của phương trình hồi quy $\widehat{Y}_{|X_i} = 2,73 + 0,04X_i$ đều không có ý nghĩa thống kê. Nói một cách khác, phương trình hồi quy này không thích hợp.

Kết luận: Yếu tố thời gian không có liên quan tuyến tính với hiệu suất của phản ứng tổng hợp.

Phương trình hồi quy $\widehat{Y}_{X_2} = f(X_2)$

$\widehat{Y}_{|X_2} = 2,73 + 0,04X_2$

(R^2 = 0,76; S=0,99)

Regression Statistics					
Multiple R	0.873933544				
R Square	0.76375984				
Adjusted R Square	0.730011246				
Standard Error	0.99290379				
Observations	9				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	22.31081667	22.31082	22.63086	0.002066188
Residual	7	6.901005556	0.985858		
Total	8	29.21182222			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-11.14111111	3.25965608	-3.41788	0.011168	-18.84897293
X2	0.128555556	0.027023418	4.757191	0.002066	0.064655325

$t_0 = 3,418 < t_{0,05} = 2,365$ (Hay $P_V = 0,011 > \alpha = 0,05$)

$\Rightarrow$ Bác bỏ giả thuyết H_0.

$t_2 = 4,757 < t_{0,05} = 2,365$ (Hay $P_V = 0,00206 < \alpha = 0,05$)

$\Rightarrow$ Bác bỏ giả thuyết H_0.

$F = 22,631 < F_{0,05} = 5,590$ (Hay $F_S = 0,00206 < \alpha = 0,05$)

$\Rightarrow$ Bác bỏ giả thuyết H_0.

Vậy cả hai hệ số -11,14(B_0) và 0,13(B_2) của phương trình hồi quy $\widehat{Y}_{|X_2} = -11,14 + 0,13X_2$ đều có ý nghĩa thống kê. Nói một cách khác, phương trình hồi quy này thích hợp.

Kết luận: Yếu tố nhiệt độ có liên quan tuyến tính với hiệu suất của phản ứng tổng hợp.

Phương trình hồi quy $\widehat{Y}_{|X_1,X_2} = f(X_1, X_2)$

$$\widehat{Y}_{|X_1,X_2} = -12,70 + 0,04X_1 + 0,13X_2$$

($R^2 = 0,97$; S=0,33)

Regression Statistics					
Multiple R	0.988775634				
R Square	0.977677254				
Adjusted R Square	0.970236338				
Standard Error	0.329668544				
Observations	9				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	28.55973413	14.27987	131.3921	1.11235E-05
Residual	6	0.652088095	0.108681		
Total	8	29.21182222			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-12.7	1.101638961	-11.5283	2.56E-05	-15.39561342
X1	0.044539683	0.005873842	7.582718	0.000274	0.03016691
X2	0.128555556	0.008972441	14.32782	7.23E-06	0.106600783

$t_0 = 11,528 > t_{0,05} = 2,365$ (Hay $P_V = 2,260.10^{-5} > \alpha = 0,05$)

$\Rightarrow$ Bác bỏ giả thuyết H_0.

$t_1 = 7,583 > t_{0,05} = 2,365$ (Hay $P_V = 0,00027 < \alpha = 0,05$)

$\Rightarrow$ Bác bỏ giả thuyết H_0.

$t_2 = 14,328 > t_{0,05} = 2,365$ (Hay $P_V = 7,233.10^{-6} < \alpha = 0,05$)

$\Rightarrow$ Bác bỏ giả thuyết H_0.

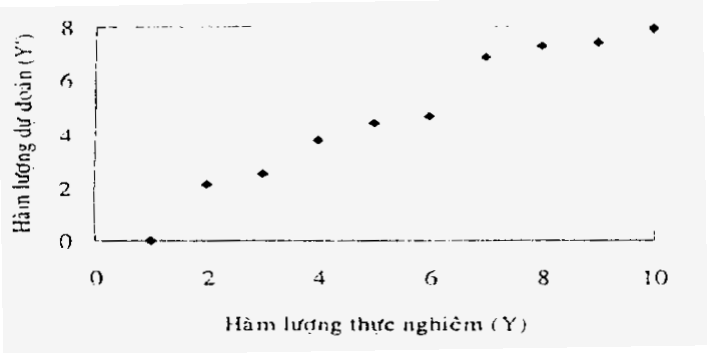
$F = 131,392 > F_{0,05} = 5,140$ (Hay $F_S = 1,112.10^{-5} < \alpha = 0,05$)

$\Rightarrow$ Bác bỏ giả thuyết H_0.

Vậy cả hai hệ số -12,70(B_0), 0,04(B_1) và 0,13(B_2) của phương trình hồi quy $\widehat{Y}_{|X_1,X_2} = -12,70 + 0,04X_1 + 0,13X_2$ đều có ý nghĩa thống kê. Nói một cách khác, phương trình hồi quy này thích hợp.

Kết luận: Hiệu suất của phản ứng tổng hợp có liên quan tuyến tính với cả hai yếu tố là thời gian và nhiệt độ.

Sự tuyến tính của phương trình $\widehat{Y}_{|X_1,X_2} = -12,70 + 0,04X_1 + 0,13X_2$ có thể được trình bày trên biểu đồ phân tán (scatterplots):



Muốn dự đoán hiệu suất của phản ứng bằng phương trình hồi quy $\widehat{Y}_{|X_1,X_2} = -12,70 + 0,04X_1 + 0,13X_2$, bạn chỉ cần chọn một ô, thí dụ B21, sau đó nhập hàm và được kết quả như sau:

	B21	⇓	**= B17 + B18 * 50 + B19 * 115**	
	A	**B**	**C**	**D**
7	Interrcept	-12,7	1,101638961	-11,52827782
8	X1	0,044539683	0,005873842	7.582717626
9	X2	0,128555556	0,008972441	14,32782351
0				
1	Dự đoán	4,310873016		

Ghi chú: B17 tọa độ của B_0, B_{18} tọa độ của B_1, B_{19} tọa độ của B_2, 50 là giá trị của X_1(thời gian) và 115 là giá trị của X_2(nhiệt độ).

PHỤ LỤC:

Bảng giá trị tới hạn dùng trong trắc nghiệm loại giá trị bất thường:

Giá trị thống kê G_1	**Số trường hợp khảo sát N**	**Trị số tới hạn G_P (P=0,01)**
N=3÷7	3	0,976
$G_1 = \frac{Y_2 - Y_1}{Y_N - Y_1}$	4	0,846
	5	0,729
	6	0,644
	7	0,586
N=8÷13	8	0,780
$G_2 = \frac{Y_3 - Y_1}{Y_{N-1} - Y_1}$	9	0,725
	10	0,678
	11	0,638
	12	0,605
	13	0,578
N=14÷24	14	0,602
$G_3 = \frac{Y_3 - Y_1}{Y_{N-2} - Y_1}$	15	0,579
	16	0,559
	17	0,542
	18	0,527
	19	0,514
	20	0,502
	21	0,491
	22	0,481
	23	0,472
	24	0,464