



TIN HỌC ỨNG DỤNG

(CH3- CÁC THỐNG KÊ CƠ BẢN, TƯƠNG QUAN VÀ HỒI QUY)

Phan Trọng Tiến

BM Công nghệ phần mềm

Khoa Công nghệ thông tin, VNUA

Email: phantien84@gmail.com

Website: <http://timoday.edu.vn>



Nội dung chính

1. Cài đặt chức năng phân tích dữ liệu trong Excel
2. Phân phối chuẩn
3. Thống kê mô tả
4. Biểu đồ tần suất
5. Tương quan
6. Hồi quy tuyến tính
7. Hồi quy phi tuyến

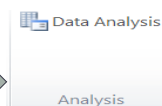


1. Cài đặt chức năng phân tích dữ liệu trong Excel

- Excel cung cấp công cụ phân tích dữ liệu bằng cách vào

□ **Data>Data Analysis**

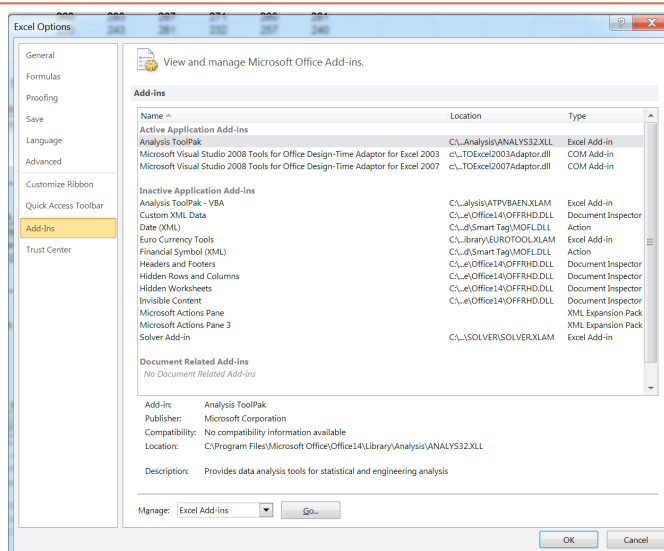
- Nếu không có chức năng này



□ **File>Options>Add-in>Analysis ToolPak** để cài đặt



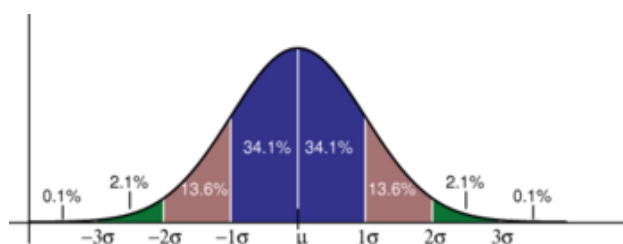
File>Options>Add-In>Analysis ToolPak





2. Phân phối chuẩn

- ❑ Còn gọi là phân bố Gauss
- ❑ Là phân bố cực kỳ quan trọng trong nhiều lĩnh vực
- ❑ Tâm phân phối chính là giá trị có tần suất lớn nhất và thường là giá trị kỳ vọng (hay gọi là giá trị trung bình của tập hợp)



Ch3 - Các thống kê cơ bản, tương quan và hồi quy

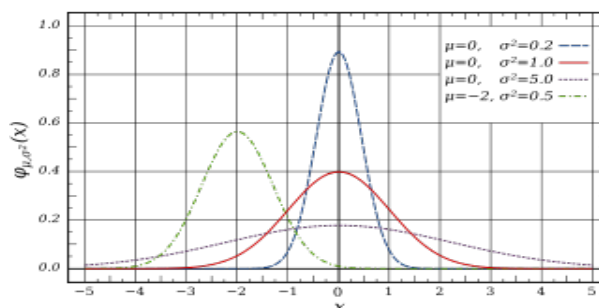
5



Phân phối chuẩn chuẩn hoá

standard normal distribution

- ❑ là phân bố có giá trị trung bình bằng 0 và phương sai bằng 1 (đường cong màu đỏ)
- ❑ Phân phối chuẩn còn được gọi là **đường cong chuông** (*bell curve*)



Ch3 - Các thống kê cơ bản, tương quan và hồi quy

6



Tính chất phân phối chuẩn

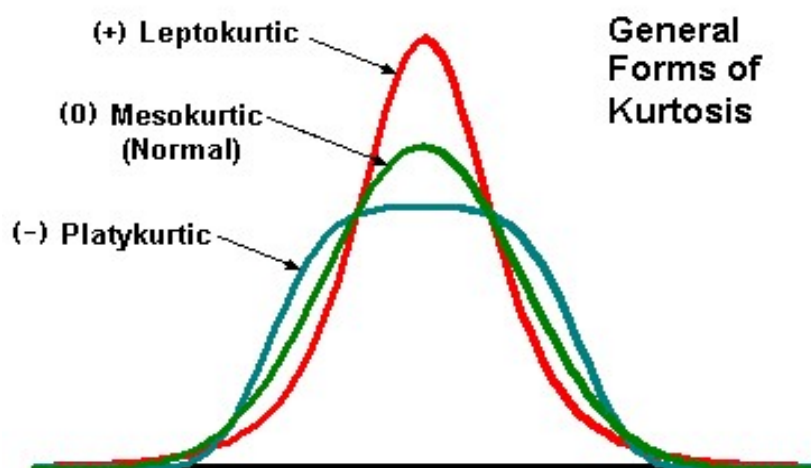
- ❑ Hàm mật độ là đối xứng qua giá trị trung bình.
- ❑ Trị trung bình cũng chính là mode và trung vị của nó.
- ❑ 68.26894921371% của diện tích dưới đường cong là nằm trong độ lệch chuẩn 1 tính từ trị trung bình.
- ❑ 95.44997361036% của diện tích dưới đường cong là nằm trong độ lệch chuẩn 2.
- ❑ 99.73002039367% của diện tích dưới đường cong là nằm trong độ lệch chuẩn 3.
- ❑ 99.99366575163% của diện tích dưới đường cong là nằm trong độ lệch chuẩn 4.
- ❑ 99.99994266969% của diện tích dưới đường cong là nằm trong độ lệch chuẩn 5.
- ❑ 99.9999980268% của diện tích dưới đường cong là nằm trong độ lệch chuẩn 6.
- ❑ 99.9999999974% của diện tích dưới đường cong là nằm trong độ lệch chuẩn 7.
- ❑ Điểm uốn của đường cong xảy ra tại độ lệch chuẩn 1 tính từ trị trung bình.

Ch3 - Các thống kê cơ bản, tương quan và hồi quy

7



Độ nhọn (Kurtosis)

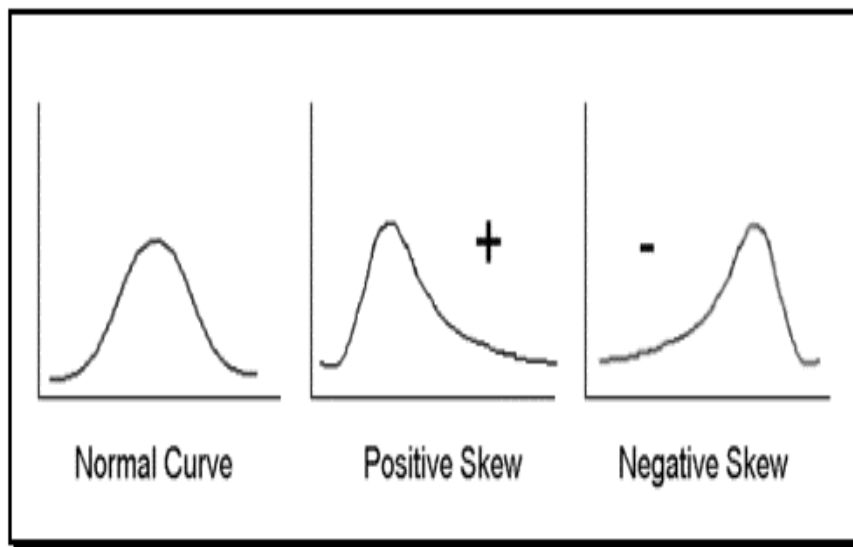


Ch3 - Các thống kê cơ bản, tương quan và hồi quy

8



Độ xiên (Skewness)



Ch3 - Các thống kê cơ bản, tương quan và hồi quy

9



Độ lệch chuẩn

- ❑ Trong thực nghiệm thường giả thiết dữ liệu lấy từ tổng thể có dạng phân phối xấp xỉ chuẩn
- ❑ Nếu giả thiết này được kiểm chứng thì
 - ❑ 68% số giá trị nằm trong khoảng 1
 - ❑ 95% nằm trong khoảng 2
 - ❑ 99.7% nằm trong khoảng 3 độ lệch chuẩn
- ❑ Đó gọi là "quy luật 68-95-99.7"

Ch3 - Các thống kê cơ bản, tương quan và hồi quy

10



3. Thống kê mô tả

- ❑ Cho phép tính các số đặc trưng mẫu/ các giá trị thống kê mẫu như trung bình, độ lệch chuẩn, sai số chuẩn, trung vị, mode ... Dữ liệu bố trí theo hàng hoặc theo cột.
- ❑ Các bước
 - ❑ **Data>Data Analysis**

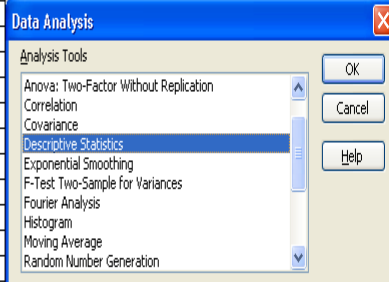
Ch3 - Các thống kê cơ bản, tương quan và hồi quy

11



Chọn *Descriptive Statistics*

	A	B	C	D	E	F	G	H
1	Bài 1 Nghiên cứu về lúa							
2	Dài bông	P1000	Số bông	Năng suất				
3	25	24	184	4				
4	26	25	186	4.5				
5	27	26	188	5				
6	27	26	166	3.8				
7	26	25	165	4				
8	25	27	164	4.3				
9	26	22	164	3.7				
10	27	23	165	4				
11	28	24	166	4.2				
12	26	25	163	3.9				
13	27	26	164	4.2				
14	28	27	165	4.5				
15	25	21	158	3.6				
16	26	22	159	3.8				
17	27	23	157	4				
18								



Ch3 - Các thống kê cơ bản, tương quan và hồi quy

12



Xuất hiện hộp thoại

Miền dữ liệu

Nhóm số liệu theo hàng hay theo cột

Nếu dữ liệu có cả nhãn đầu dòng thì tích

Nơi đặt kết quả

Hiện các thống kê cơ bản

Độ tin cậy

Số lớn thứ nhất

Số nhỏ nhất

Ch3 - Các thống kê cơ bản, tương quan và hồi quy

13



Kết quả

Dải bông			P1000			Số bông			Năng suất
Mean (trung bình)	26.4	Mean	24.4		Mean	167.6		Mean	4.1
Standard Error (sai số chuẩn)	0.25448	Standard Error	0.485994317035166		Standard Error	2.57052894315916		Standard Error	0.094112394811432
Median (trung vị)	26	Median	25		Median	165		Median	4
Mode	26	Mode	25		Mode	165		Mode	4
Standard Deviation (độ lệch chuẩn)	0.985611	Standard Deviation	1.88224789622865		Standard Deviation	9.95561578779957		Standard Deviation	0.364495737776375
Sample Variance (phương sai mẫu)	0.97142	Sample Variance	3.54285714285717		Sample Variance	99.1142857142841		Sample Variance	0.132857142857144
Kurtosis (độ nhọn)	-0.81046	Kurtosis	-0.974945969743058		Kurtosis	0.569278605644604		Kurtosis	1.3072120387417
Skewness (độ lệch)	0.06197	Skewness	-0.302552553959346		Skewness	1.34642659266614		Skewness	1.04158311363296
Range (phạm vi)	3	Range	6		Range	31		Range	1.4
Minimum (giá trị nhỏ nhất)	25	Minimum	21		Minimum	157		Minimum	3.6
Maximum (giá trị lớn nhất)	28	Maximum	27		Maximum	188		Maximum	5
Sum (tổng)	396	Sum	366		Sum	2514		Sum	61.5
Count (Số lượng)	15	Count	15		Count	15		Count	15
Largest(1) (Giá trị lớn nhất đầu tiên)	28	Largest(1)	27		Largest(1)	188		Largest(1)	5
Smallest(1) (Giá trị nhỏ nhất đầu tiên)	25	Smallest(1)	21		Smallest(1)	157		Smallest(1)	3.6
Confidence Level(95.0%) (nửa khoảng tin cậy)	0.54581	Confidence Level	1.04235413635581		Confidence Level	5.51323624113788		Confidence Level	0.201851010935121

Ch3 - Các thống kê cơ bản, tương quan và hồi quy

14



Phân tích kết quả

- ❑ **Mean**: cho ta trị trung bình của dãy số
- ❑ **Median**: cho giá trị điểm giữa của dãy số
 - ❑ Hai giá trị **Mean** và **Median** xấp xỉ nhau thì số liệu cân đối.
- ❑ Phương sai mẫu hay độ lệch chuẩn cho ta biết độ phân tán của số liệu quanh giá trị trung bình, nếu giá trị này càng nhỏ chứng tỏ số liệu càng tập chung.



Phân tích kết quả

- ❑ **Kurtosis** đánh giá đường mật độ phân phối của dãy số liệu có nhọn hơn hay tù hơn đường mật độ chuẩn tắc. Nếu trong $[-2,2]$ thì coi xấp xỉ chuẩn.
- ❑ **Skewness** đánh giá đường phân phối lệch trái hay lệch phải. Nếu trong $[-2,2]$ thì coi số liệu cân đối gần như số liệu trong phân phối chuẩn.
- ❑ **Confidence Level** là mức độ tin cậy. Ví dụ Confidence Level là m khoảng tin cậy trung bình tổng thể là: $(\text{Mean} - m, \text{Mean} + m)$



4. Biểu đồ tần xuất

- ❑ **Chức năng:** khi có nhiều số liệu cần chia lớp để thấy rõ các nét đặc trưng cơ bản của dãy số liệu, sau đó kiểm tra tính chuẩn của biến nghiên cứu.
- ❑ **Cách làm:**
 - ❑ Để số liệu trong một cột, một hàng hay một bảng chữ nhật
 - ❑ Tìm giá trị Min, Max của miền dl, tính $R = \text{Max} - \text{Min}$
 - ❑ Chọn khoảng k, thực tế k: 20-30, ít số liệu k: 6-10
 - ❑ Tính khoảng cách giữa các tổ: $h = R/k$ (làm tròn)
 - ❑ Có Max, Min, R, k, h ta tạo miền phân tổ bắt đầu từ Min, các giá trị tiếp theo cộng dồn với h, cho đến sát Max thì dừng



Tạo miền phân tổ

- ❑ Ví dụ:
 - ❑ $\text{Max} = 49, \text{Min} = 11, R = 38, k = 10, h = 38/10 = 3.8 \approx 4$
 - ❑ Xuất phát từ Min 11

$11 + 4 = 15$
 $15 + 4 = 19$

 - ❑ Cho đến sát Max (49) 47
- ❑ Tạo miền phân tổ tự động trong Excel (Ch2 – Slide 15)
- ❑ Vào **Data > Data Analysis**

Chọn *Histogram*

	C	D	E	F	G	H
con cá		Bin				
1	27	10				
7	34	15				
6	37	20				
1	35	25				
9	40	30				
2	26	35				
9	26	40				
10	34	45				
4	19	50				
9	20	55				

Data Analysis

Analysis Tools

- Fourier Analysis
- Histogram**
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances

Ch3 - Các thống kê cơ bản, tương quan và hồi quy **19**

Cửa sổ *Histogram* hiện ra

Miền dữ liệu → Input Range: \$A\$3:\$D\$12

Miền phân tổ → Bin Range: \$D\$3:\$D\$12

Nhấn ở đầu dòng nếu có → ☐ Labels

Nơi chứa kết quả → Output options

Tần số sắp xếp giảm dần → ☐ New Worksheet Ply:

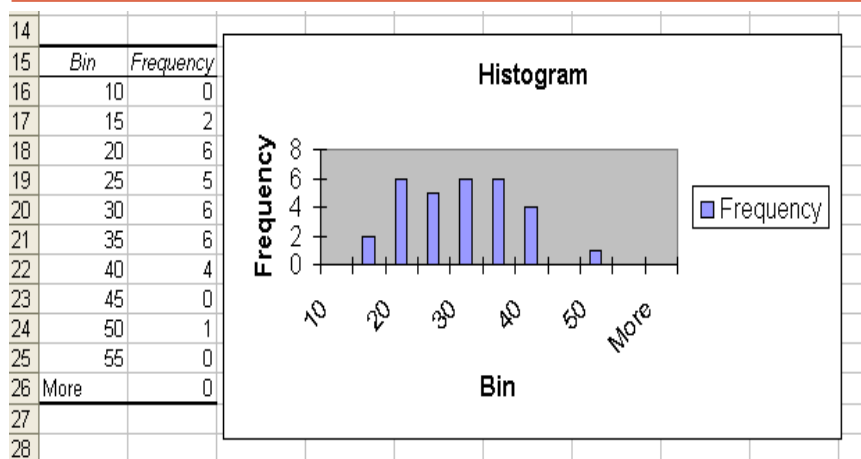
% Cộng dồn → ☐ Pareto (sorted histogram)

Biểu đồ → ☒ Chart Output

Ch3 - Các thống kê cơ bản, tương quan và hồi quy **20**



Kết quả biểu đồ



Ch3 - Các thống kê cơ bản, tương quan và hồi quy

21



Phân tích kết quả

- ❑ Tần số rơi vào từng khoảng được ghi ở cận trên của khoảng.
 - ❑ Ví dụ: (10,15] có 2 số liệu được ghi ứng với số 15 là cận trên
- ❑ Phải đưa ra được kết luận
 - ❑ Biểu đồ cho thấy khoảng nào số liệu xuất hiện nhiều nhất.
 - ❑ Biểu đồ có thể cho ta biết dãy số liệu khảo sát có tuân theo phân phối chuẩn hay không

Ch3 - Các thống kê cơ bản, tương quan và hồi quy

22



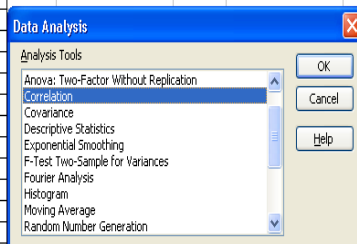
5. Tính hệ số tương quan

- ❑ Dùng hệ số tương quan để xác định mối quan hệ giữa hai đặc tính. Ví dụ nghiên cứu mối quan hệ nhiệt độ trung bình của một vị trí và việc dùng điều hòa.
- ❑ Tính hệ số tương quan giữa các biến sắp xếp thành một bảng gồm n hàng, m cột (mỗi cột là một biến)
- ❑ Vào **Data>Data Analysis**



Chọn Corelation

	A	B	C	D	E	F	G	H
1	Bài 1 Nghiên cứu về lúa							
2	Dải bông	P1000	Số bông	Năng suất				
3	25	24	184	4				
4	26	25	186	4.5				
5	27	26	188	5				
6	27	26	186	3.8				
7	26	25	185	4				
8	25	27	164	4.3				
9	26	22	164	3.7				
10	27	23	165	4				
11	28	24	166	4.2				
12	26	25	163	3.9				
13	27	26	164	4.2				
14	28	27	165	4.5				
15	25	21	158	3.6				
16	26	22	159	3.8				
17	27	23	157	4				
18								



Xuất hiện cửa sổ

Miền dữ liệu kể cả nhãn

Nhóm số liệu theo hàng hay cột

Chọn nhãn đầu dòng không

Chọn nơi để kết quả

Ch3 - Các thống kê cơ bản, tương quan và hồi quy **25**

Kết quả

	<i>Dài bông</i>	<i>P1000</i>	<i>Số bông</i>	<i>Năng suất</i>
Dài bông	1			
P1000	0.292619524226006	1		
Số bông	-0.0698826931062899	0.340772495753049	1	
Năng suất	0.377770181367487	0.666319715476545	0.661379393738142	1

Ch3 - Các thống kê cơ bản, tương quan và hồi quy **26**



Phân tích kết quả

- ❑ Hệ số tương quan của dòng và cột ghi ở ô giao giữa dòng và cột.
- ❑ Hệ số tương quan âm thể hiện mối tương quan nghịch biến.
- ❑ Các hệ số tương quan có giá trị tuyệt đối xấp xỉ 0.75 trở lên thể hiện mối tương quan tuyến tính mạnh.



6. Hồi quy tuyến tính

- ❑ Cho phép tìm phương trình hồi quy tuyến tính đơn $y = a \cdot x + b$ và hồi quy tuyến tính bội $y = a_1 \cdot x_1 + a_2 \cdot x_2 + \dots + a_n \cdot x_n + b$.
- ❑ Các biến độc lập chứa trong n cột, biến phụ thuộc y để trong một cột, các giá trị tương ứng giữa biến độc lập và biến phụ thuộc được xếp trên cùng một hàng.
- ❑ Ví dụ: Tìm đường hồi quy của năng suất lúa y phụ thuộc vào độ dài bông, trọng lượng 1000 hạt, và số bông.
- ❑ Vào **Data>Data Analysis**

Chọn Regression

	A	B	C	D	E	F	G	H
1	Bài 1 Nghiên cứu về lúa							
2	Dài bông	P1000	Số bông	Năng suất				
3	25	24	184	4				
4	26	25	186	4.5				
5	27	26	188	5				
6	27	26	166	3.8				
7	26	25	165	4				
8	25	27	164	4.3				
9	26	22	164	3.7				
10	27	23	165	4				
11	28	24	166	4.2				
12	26	25	163	3.9				
13	27	26	164	4.2				
14	28	27	165	4.5				
15	25	21	158	3.6				
16	26	22	159	3.8				
17	27	23	157	4				
18								

Data Analysis

Analysis Tools

- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression**
- Sampling
- t-Test: Paired Two Sample for Means

OK Cancel Help

Ch3 - Các thống kê cơ bản, tương quan và hồi quy 29

Xuất hiện hộp thoại

Miền dữ liệu Y

Miền dữ liệu X

Có để nhãn đầu dòng không?

Độ tin cậy

Hệ số tự do b = 0 khi tích vào

Hiện phần dư hay sai lệch giữa y thực nghiệm và y theo hồi quy

Hiện phần dư đã chuẩn hóa

Hiện đồ thị xác suất thông thường

Hiện đồ thị phần dư

Hiện đồ thị đường dự báo

Regression

Input

Input Y Range: \$D\$2:\$D\$17

Input X Range: \$A\$2:\$C\$17

☒ Labels ☐ Constant is Zero

☒ Confidence Level: 95 %

Output options

☒ Output Range: \$J\$11

☐ New Worksheet Ply:

☐ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

Ch3 - Các thống kê cơ bản, tương quan và hồi quy 30



Kết quả

SUMMARY OUTPUT						
Regression						
Multiple R	0.858900919592805					
R Square	0.737710789677366					
Adjusted R Square	0.666177368680284					
Standard Error	0.210596108744181					
Observations	15					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	1.3721420687999	0.457380689599967	10.3128129396671	0.001582207343336	
Residual	11	0.4878579312001	0.0443507210181909			
Total	14	1.86				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-4.06364336163536	1.8441848138248	-2.20349030703026	0.0497816657255703	-8.12266676680394	-0.0046199564667706
Dài bóng	0.111607803725866	0.0608118565619376	1.83529676671181	0.0936208158536078	-0.0222381900412795	0.245453797493012
P1000	0.0756835086172582	0.0337876990102569	2.23997226310921	0.0467035333349004	0.001317284548322	0.150049732686194
S? bóng	0.0201104996003066	0.006123394723325213	3.28420761831693	0.0072791914546221	0.0066329986932719	0.0335880005073412

Ch3 - Các thống kê cơ bản, tương quan và hồi quy 31



Phân tích kết quả

- ❑ Nếu hệ số tương quan bội (Multiple R) xấp xỉ ≥ 0.75 thì mô hình qui hoạch tuyến tính là thích hợp.
 - ❑ Ví dụ: Multiple R = 0.8589 -> mô hình tuyến tính coi là thích hợp.
- ❑ Hệ số tương quan (R Square) cho biết sự biến động y do x1, x2, x3 ... gây nên. Hệ số Adjusted R Square không sát R Square → không phải tất cả các biến đưa vào là cần thiết.
 - ❑ R Square = 0.7377 cho biết 73.77% sự biến động của y do x1, x2, x3 gây nên.
 - ❑ Adjusted R Square = 66.62% không sát R Square

Ch3 - Các thống kê cơ bản, tương quan và hồi quy 32



Phân tích kết quả

- ❑ F thực nghiệm = 10.31281 với xác suất $0.00158 < 0.05$ nên phương trình hồi quy tuyến tính được chấp nhận
- ❑ Dựa vào các hệ số ta viết được đường hồi quy dự báo

$$y = -4.06364 + 0.1116x_1 + 0.075684x_2 + 0.02011x_3$$

Hệ số x_1 không đáng tin cậy vì P-value = $0.093621 > 0.05$ (mức ý nghĩa đã chọn) -> cần tiến hành loại bỏ biến x_1 để đường hồi quy với các hệ số đều có ý nghĩa.



7. Hồi quy phi tuyến

- ❑ Các dạng hồi quy phi tuyến như hàm mũ, hàm logarit, hàm đa thức, hàm căn bậc hai ...
- ❑ Có hai cách:
 - ❑ Thông qua cách biến đổi đưa về dạng hồi quy tuyến tính bội.
 - ❑ Vẽ đồ thị và tìm đường ngoại suy.



6.1 Cách 1

- ❑ Thông qua biến đổi ta đưa về dạng hồi quy tuyến tính bội như hàm mũ, hàm logarit, hàm đa thức, hàm căn bậc hai ...

- ❑ Ví dụ

Nghiên cứu về dân số, x là năm, y là dân số. Tìm đường hồi quy phi tuyến dạng đa thức bậc hai: $Y = ax^2 + bx + c$. Khi đó ta thêm cột $X2 = X^2$ sau đó thực hiện tìm đường hồi quy bội tuyến tính với các biến độc lập X, X2 và hàm là Y.



Làm tương tự như mục 5

	A	B	C	D	E	F	G
1		Nghiên cứu về dân số 1850 đến năm 1950					
2	X	Dân số		X	X2	Dân số	
3	1	23.2		1	1	23.2	
4	2	31.4		2	4	31.4	
5	3	39.8		3	9	39.8	
6	4	50.2	→→→	4	16	50.2	
7	5	62.9		5	25	62.9	
8	6	76		6	36	76	
9	7	92		7	49	92	
10	8	105.7		8	64	105.7	
11	9	122.8		9	81	122.8	
12	10	131.7		10	100	131.7	
13	11	151.1		11	121	151.1	

Data Analysis

Analysis Tools

- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression**

OK Cancel Help



Phân tích kết quả

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.998917634					
R Square	0.99783644					
Adjusted R Square	0.997295551					
Standard Error	2.252508573					
Observations	11					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	18720.326	9360.163	1844.805	2.19117E-11	
Residual	8	40.59035897	5.073795			
Total	10	18760.91636				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	12.96242424	2.473722725	5.240047	0.000783	7.258009415	18.66683907
X	8.228951049	0.947455223	8.685319	2.41E-05	6.044115388	10.41378671
X ²	0.397435897	0.076899375	5.168259	0.000855	0.220105621	0.574766173

- Phân tích kết quả được phương trình:

$$Y = 0.397435 X^2 + 8.228951 X + 12.96242$$



6.2 Cách 2

- Vẽ đồ thị XY (Scatter) biểu diễn tương quan giữa y và x, đồ thị dạng điểm, sau đó tìm đường ngoại suy và hiện ra phương trình hồi quy.
- Sinh viên xem lại về phần đồ thị
- Sau khi vẽ xong đồ thị dạng XY bạn vào Menu **Layout>Trendline**

