Ch.8

i	x_i	y_i	$r(x_i)$	$r(y_i)$	$d_i = r(x_i) - r(y_i)$	d_i^2
1						
⋮						
n						

n , $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i^2$, $\sum_{i=1}^{n} x_i y_i$, $r(x_i)$: رتب x ، $r(y_i)$: رتب y

$$\Rightarrow S_{xy} = S_{yx} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

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

① Simple linear regression (regression eq.) معادلة الإنحدار

$\hat{y} = a + bx$ where ;

$\hat{y}$: Predicted value القيمة المتنبأة

$b = \frac{S_{xy}}{S_{xx}}$: *regression coefficient معامل الإنحدار

* For every unit increase in x we expect that y to increase de " by |b| حسب اشارة b

$a = \bar{y} - b\bar{x}$: الحد الثابت

② Coefficient of determination (proportion of variation) معامل التحديد (نسبة التغير)

$r^2 = b \frac{S_{xy}}{S_{yy}}$: This mean that $r^2(100)\%$ of the variation in y can be explained by the variation in x

$= 1 - \frac{\quad}{\quad} = \frac{\quad}{\quad}$ ($r^2(100)\%$ من التغير في y يرجع بسبب علاقته الخطية بالمتغير x)

③ Correlation الإرتباط

a) Pearson correlation coefficient معامل الإرتباط لبيرسون

$r_p = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = ($ b اشارة $) \sqrt{r^2} \Rightarrow r^2 = (r_p)^2$: إذا كان x و y متغيرات كمية

b) Spearman rank correlation معامل الإرتباط لسبيرمان للرتب

$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$: إذا كان x و y متغيرات كمية أو وصفية يمكن ترتيبها

c) $-1 \le r_p, r_s \le 1$ ← اتجاه العلاقة بين x و y

-ve (indirect) : العلاقة بين x و y عكسية

+ve (direct) : العلاقة بين x و y طردية

قوة العلاقة بين x و y ← حسب قيمة $|r_p|$ أو $|r_s|$:

strong: قوي

moderate: متوسط

weak: ضعيف

0 — .5 — .8 — 1

استخدام الآلة الحاسبة في الإحصاء:

Mood → stat → A+BX → X | Y → Ac

shift + stat (2)

→ sum

1: Σx^2 2: Σx

3: Σy^2 4: Σy

5: Σxy

→ var

1: n 2: $\bar{x}$

5: $\bar{y}$

→ Reg.

1: A 2: B

3: r

5: $\hat{y}$

هذا هو جذر معامل التحديد وبأخذ [x²] = يكون لدينا r^2

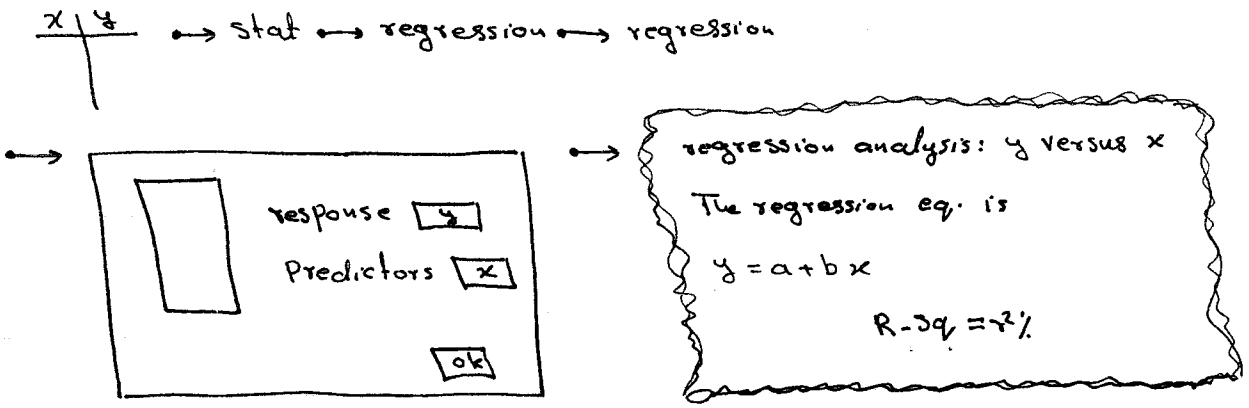
هذه تساعد على إيجاد قيمة y عندما يعطى قيمة لـ x وتكون استخدامها كما يلي:

أتأكد من أن الشاشة لا توجد عليها أي رقم وذلك بالضغط على Ac

ثم أدخل القيمة لـ x = ثم أختار $\hat{y}$ =

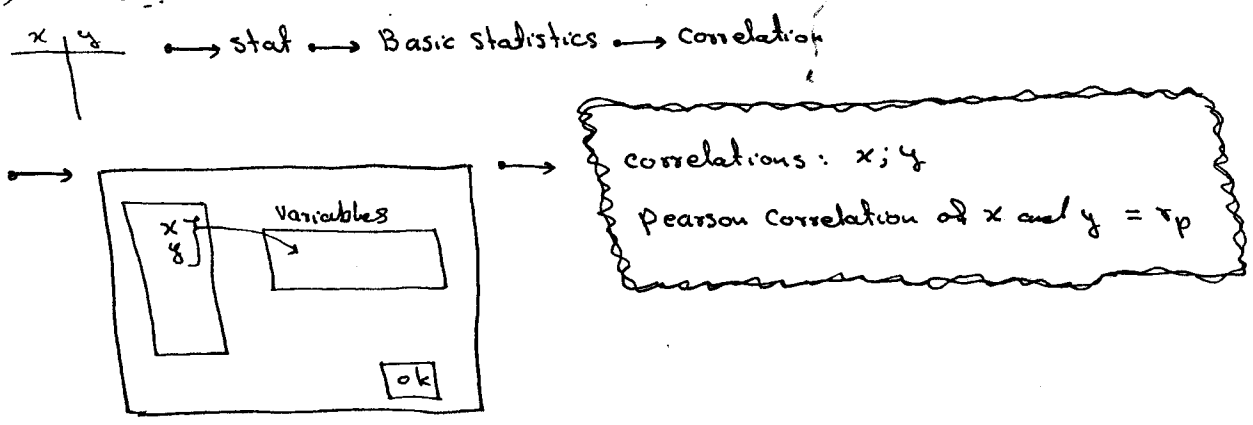
1) استخدام Minitab في الانحدار:

إيجاد معادلة الانحدار ومعامل التحديد r^2

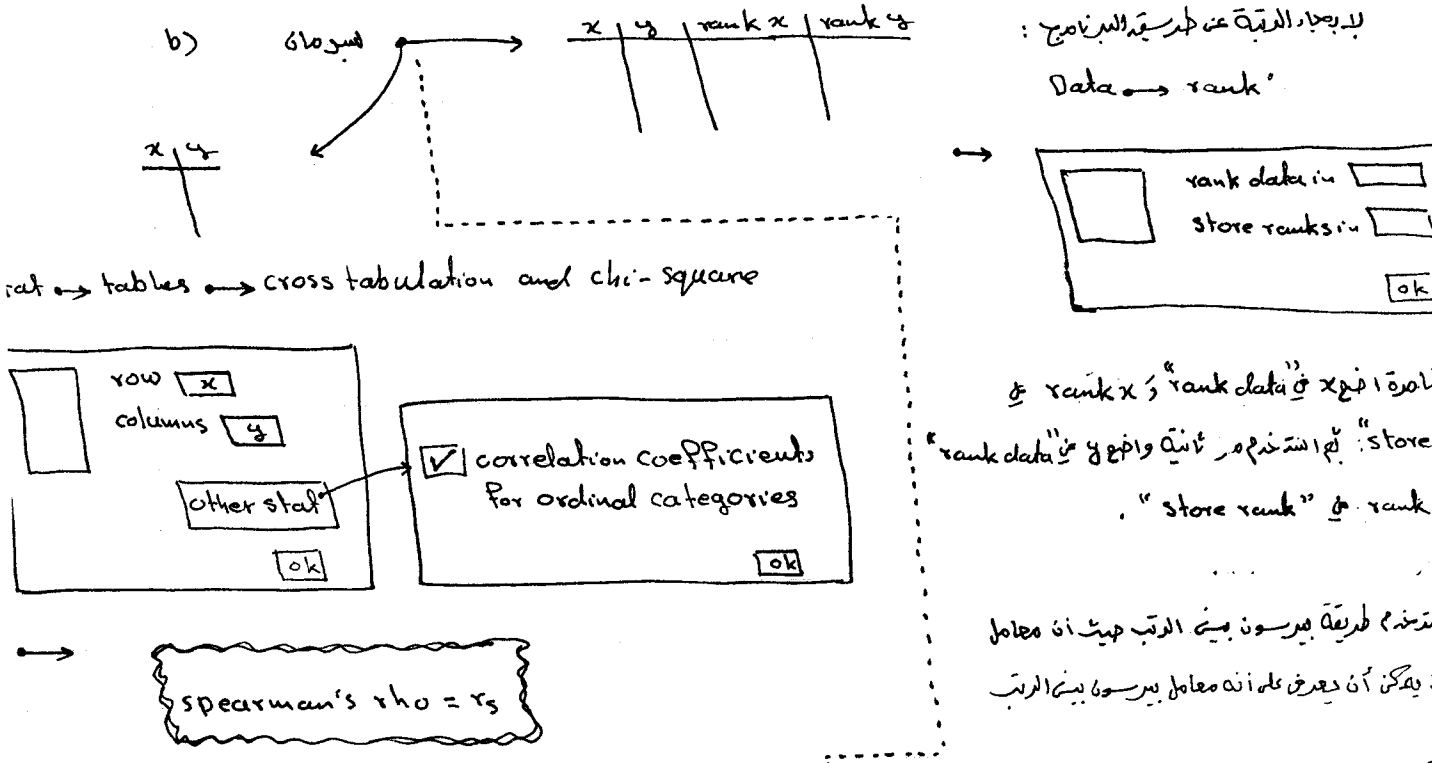


2) إيجاد معامل الارتباط

a) بيرسون



b) سبيرمان



If the data of the variables X and Y is given as follows:

$\sum x = 30$, $\sum y = 17$, $\sum xy = 91$, $\sum x^2 = 150$, $\sum y^2 = 59$, $n = 8$, then:

1) $\bar{x} = \Sigma x_i/n = 3.75$

a) 2.125 b) 3.75 c)5 d) none of these

2) $\bar{y} = \Sigma y_i/n = 2.125$

a) 2.125 b) 3.75 c)5 d) none of these

3) $S_{xx} = \Sigma x_i^2 - n(\bar{x})^2 = 37.5$

a)37.5 b) 22.875 c) 27.25 d) none of these

4) $S_{yy} = \Sigma y^2 - n(\bar{y})^2 = 22.875$

a)37.5 b) 22.875 c) 27.25 d) none of these

5) $S_{xy} = \Sigma xy - n\bar{x}\bar{y} = 27.25$

a)37.5 b) 22.875 c) 27.25 d) none of these

6) $b = S_{xy}/S_{xx} = .726666$

a) 0.727 b) - 0.6 c)1.19 d) none of these

7) $a = \bar{y} - b\bar{x} = -.6$

a) 0.727 b) - 0.6 c)1.19 d) none of these

8) The regression equation of y on x is:

a) y=- 0.6+0.727x b) y=1.19+0.727x c) y=1.19-0.6x d) none of these

$a + bx$

9) If X=10 then the predictive value is: $\hat{y} = -.6 + (.727)(10) = 6.67$

a) $\hat{y} = 8.46$ b) $\hat{y} = 6.67$ c) $\hat{y} = -4.81$ d) none of these

10) The determination coefficient r^2 is:

a) 0.9304 b) 0.8656 c) 0.9646 d) none of these

$= b\frac{S_{xy}}{S_{yy}} = .727\ \frac{27.25}{22.875} = .8660437$

Minitab ادخل بالاستخدام ✓

	C2	C3	C4
x	y	rank x	rank y
1	10.1	6.2	1	1
2	22.2	14.9	10	12
3	21.6	6.4	8	2
4	27.4	8.4	12	6
5	29.4	10.2	13	7
6	30.8	13.3	14	10
7	26.4	16.3	11	13
8	22.0	8.3	9	5
9	18.8	16.4	7	14
10	14.8	12.1	5	9
11	12.0	7.0	4	3
12	11.7	13.8	3	11
13	10.5	11.3	2	8
14	17.3	7.2	6	4

11/26/2013 2:18:03 PM

Welcome to Minitab, press F1 for help.

Regression Analysis: y versus x

The regression equation is
y = 8.53 + 0.118 x

a b ⇒ at $x = 25 \Rightarrow y = 8.53 + (.118)(25) = 11.48$

Predictor	Coef	SE Coef	T	P
Constant	8.531	3.004	2.84	0.015
x	0.1177	0.1443	0.82	0.431

S = 3.72641 R-Sq = 5.3% R-Sq(adj) = 0.0%

r^2 (100)%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	9.24	9.24	0.67	0.431
Residual Error	12	166.63	13.89		
Total	13	175.87			

Correlations: x, y

Pearson correlation of x and y = 0.229 = r_p

Tabulated statistics: x, y

Rows: x Columns: y

3	6.2	6.4	7.0	7.2	8.3	8.4	10.2	11.3	12.1	13.3	13.8	14.9	16.
10.10	1	0	0	0	0	0	0	0	0	0	0	0	
10.50	0	0	0	0	0	0	0	1	0	0	0	0	
11.70	0	0	0	0	0	0	0	0	0	0	1	0	
12.00	0	0	1	0	0	0	0	0	0	0	0	0	
14.80	0	0	0	0	0	0	0	0	1	0	0	0	
17.30	0	0	0	1	0	0	0	0	0	0	0	0	
18.80	0	0	0	0	0	0	0	0	0	0	0	0	
21.60	0	1	0	0	0	0	0	0	0	0	0	0	
22.00	0	0	0	0	1	0	0	0	0	0	0	0	
22.20	0	0	0	0	0	0	0	0	0	0	0	1	
26.41	0	0	0	0	0	0	0	0	0	0	0	0	
27.40	0	0	0	0	0	1	0	0	0	0	0	0	
29.40	0	0	0	0	0	0	1	0	0	0	0	0	
30.80	0	0	0	0	0	0	0	0	0	1	0	0	
All 1	1	1	1	1	1	1	1	1	1	1	1	1	

16.4 All

10.1	0	1
10.5	0	1
11.7	0	1
12.0	0	1
14.8	0	1
17.3	0	1
18.8	1	1
21.6	0	1
22.0	0	1
22.2	0	1
26.4	0	1
27.4	0	1
29.4	0	1
30.8	0	1
All	1	14

Cell Contents: Count

Pearson's r 0.301099
Spearman's rho 0.301099

r_s

Correlations: rank x, rank y

Pearson correlation of rank x and rank y = 0.301

$$\hat{y} = 7.0928686 + 0.5301812\ (25) = 20.347399$$

So we predict that a camel will consume 20.3 liters if the temperature is 25°C.

The proportion of the variation in the water consumed by a camel (Y) that is explained by the temperature (X) is given by the coefficient of determination where

$$R^2 = \frac{SSR}{SST} = \frac{bS_{xy}}{S_{yy}}$$

$$= \frac{0.5301812(353.76571)}{1361.2293} = \frac{187.55993}{1361.2293} = 0.1377871$$

Therefore, only about 14% of the differences in water consumption of camels is explained by temperature.

Note that when we use the regression line to predict, we must be careful. The scatter diagram shows us that the relationship between Y and X is linear only for the range in which we have taken data on X. In Example 5.1, this range of X is from 10.1 to 30.8. We call this *range of X data from the smallest value of X to the largest value of X* the scope of the model. When we predict for Y, we should not predict using X values outside the scope (either larger or smaller) unless we can be sure that the model does indeed hold outside.

EXERCISES

For Correlation and regress

Suppose that as in Example 5.1, we are interested in using temperature to predict water consumption but now for donkeys. Consider the data:

Temperature	Water consumption	Temperature	Water consumption
10.1	6.2	22.0	8.3
22.2	14.9	18.8	16.4
21.6	6.4	14.8	12.1
27.4	8.4	12.0	7.0
29.4	10.2	11.7	13.8
30.8	13.3	10.5	11.3
26.4	16.3	17.3	7.2

and complete the following:

a) Construct a scatter diagram.

b) Find the equation of the least squares regression lin

c) Plot the line on the scatter diagram.

d) Predict the water consumption of a donkey if t temperature is 25°C.

e) What is the scope of the model for prediction?

f) Find the proportion of variation in a donkey's wa consumption that is explained by temperature.

In a study on the saw-toothed grain beetle [Ali et (1980)], newly-hatched larvae were placed in flour at fi densities (that is, with a certain number of larvae per g of flour) and the percent mortality of the larvae recorded:

Density (X)	% Mortality (Y)
0.8	0.0
1.6	3.2
2.4	8.8
3.2	14.3
4.0	19.7

scatter diagram

least squares regression line.

5.3. Given below are the daily average percent relative humidity and the daily average temperature (in °C) for a sample of days taken in the Qassim region [Moustafa et al. (1978)]:

Relative Humidity	Temperature	Relative Humidity	Temperature
54	10.9	19	32.9
45	12.5	19	34.0
39	19.8	21	31.5
49	22.5	24	26.9
31	27.7	59	17.9
20	32.0	60	10.2

If we want to predict the relative humidity from the temperature,

a) Construct a scatter diagram.
b) Find the equation of the least squares regression line.
c) Plot the line on the scatter diagram.
d) Predict the relative humidity if the temperature is 25°C in Qassim.
e) What is the scope of the model?
f) What proportion of the variation in relative humidity is explained by temperature?

5.4. Suppose we want to use the dry weight of a soybean shoot (in grams) to predict the iron content of the shoot (in mg/g). A sample of 12 soybean shoots harvested thirty-five days after planting gave [as estimated from a graph in Sliman et al. (1987)]:

Dry Weight	Iron Content	Dry Weight	Iron Content
2.35	60	3.00	107
2.75	80	3.10	85
2.85	65	3.15	98
2.95	76	3.45	106
2.95	114	3.60	119
3.00	96	3.95	124

For this data,

a) Construct a scatter diagram.
b) Find the equation of the least squares regression line.
c) Plot the line on the scatter diagram.
d) Predict the iron content of a 35-day-old shoot if its dry weight is 2.9 g.
e) What is the scope of the model?
f) What proportion [illegible]

In a study, the interest was in predicting the long-run average cost (in S.R.) for tomatoes using the output or level of production (in tons). Given below are data for selected output levels [Al-Ibrahim and Sherif (1993)]:

Output	Average Cost
6	1526
10	1510
20	1472
30	1434
40	1398
50	1363
100	1206
150	1079
200	982
250	915
300	878

a) Construct a scatter diagram.
b) Find the equation of the least squares regression line.
c) Plot the line on the scatter diagram.
d) What is the scope of the model?
e) Additional data were given in the study as:

Output	Average Cost
350	871
400	894
450	947
500	1030
550	1143

Plot this additional data on the scatter diagram using a different symbol. Do you think that a linear model is appropriate outside the scope of the model used in finding the equation in b)? Why or why not?