S12

MATHEMATICS SUPPORT CENTRE

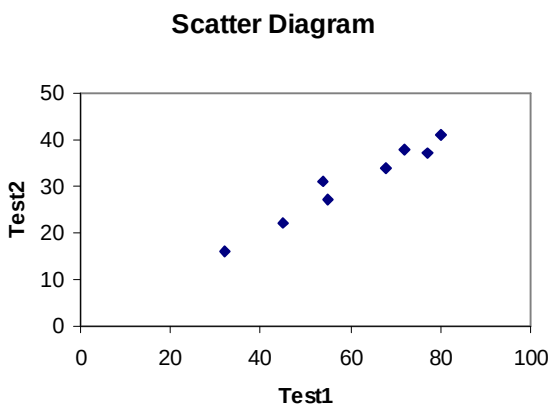
Title: Regression and Correlation

Target: On completion of this worksheet you should be able to calculate the equation of a regression line, the correlation coefficient and the coefficient of determination.

Eight students took two mathematics tests. We would like to know if we could predict the result of test 2 from test 1. The percentage results are given below:

Test1	54	72	32	68	55	80	45	77
Test2	31	38	16	34	27	41	22	37

These results can be plotted on a scatter diagram.



We can see that those students with the higher marks in test 1 get higher marks in test 2 and the points almost lie on a straight line. If we can find this line then we can use it to predict test 2 marks. We want to make sure that the line is as close to all the points as possible. This line is called the regression line and is found by the Method of Least Squares. We will use formulae to find the equation of the regression line.

Suppose the equation of the regression line is $y = a + bx$ (if you are not sure about this see graph sheet G6). If we call the test 1 results 'x' and the test 2 results 'y' then a and b are found from the formulae:

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \qquad a = \frac{\Sigma y - b \times \Sigma x}{n}$$

(n is the number of pairs of values)

Note: We must find b first then use it to find a.

In this example the number of pairs of values is 8 ie $n = 8$. We will put the calculations in a table:

	Test 1	Test 2			
	x	y	x^2	y^2	xy
	54	31	2916	961	1674
	72	38	5184	1444	2736
	32	16	1024	256	512
	68	34	4624	1156	2312
	55	27	3025	729	1485
	80	41	6400	1681	3280
	45	22	2025	484	990
	77	37	5929	1369	2849
Total	483	246	31127	8080	15838
	Σx	Σy	Σx^2	Σy^2	Σxy

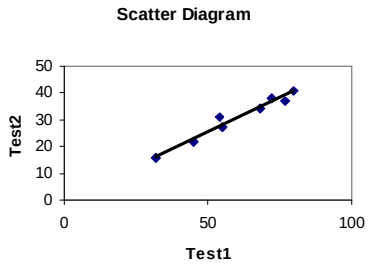
$$b = \frac{8 \times 15838 - 483 \times 246}{8 \times 31127 - 483^2} = 0 \cdot 5014$$

$b = 0 \cdot 50$ correct to 2 decimal places

$$a = \frac{246 - 0 \cdot 5014 \times 483}{8} = 0 \cdot 48 \text{ to 2 d.p.}$$

so $y = 0 \cdot 48 + 0 \cdot 50x$

This is the equation of the regression line and can be drawn on the scatter diagram:



If another student achieved a score of 40 in test 1 then we can predict that this student will get $(0{\cdot}48 + 0{\cdot}50 \times 40) = 20$ in test 2.

The value of b (the gradient of the line) is called the regression coefficient and shows that for each extra mark in test 1 the mark in test 2 goes up by 0·50.

Exercise

Plot a scatter diagram and find the values of *a* and *b* and the equation of the line of best fit for the following sets of data:

1.

x	0	1	2
y	4	7	10

2.

x	1	2	3	4
y	70	70	80	100

3. Five fruit buns were weighed and the number of sultanas in each was noted:

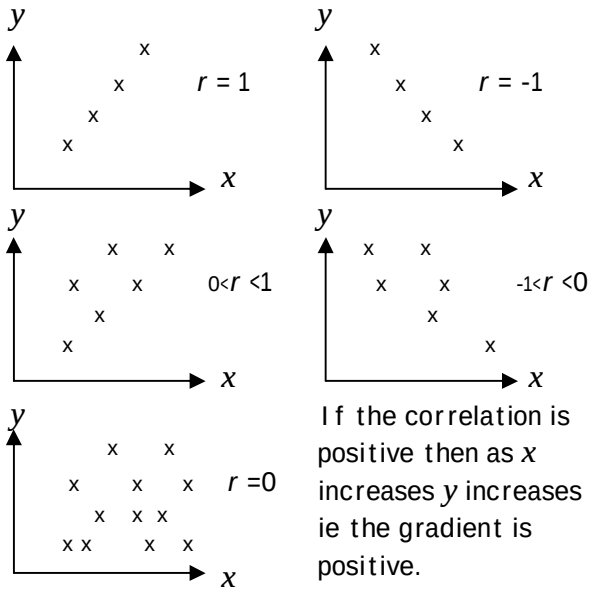
Weight (g)	22	38	47	50	53
No. sultanas	7	18	20	24	26

Give an interpretation of *b* and predict how many sultanas there would be in a bun weighing 35g.

(Answers: $a = 4$, $b = 3$, $y = 4 + 3x$
$a = 59$, $b = 7$, $y = 59 + 7x$
$a = -5{\cdot}56$, $b = 0{\cdot}58$, $y = -5{\cdot}56 + 0{\cdot}58x$
For each 1g increase in weight there will be 0·58 of a sultana. 15 sultanas)

Correlation

In all the examples above the scatter diagrams show that there seems to be a linear relationship connecting the variables so we are justified in finding the regression line. We can also calculate the correlation coefficient, *r,* to give a measure of how good this relationship is. If the points lie exactly on a straight line then we have perfect correlation and $r = 1$ or –1. Generally *r* lies between these values.



If the correlation is positive then as x increases y increases ie the gradient is positive.

If y decreases as x increases then the correlation is negative (negative slope).

Example

Using the data from the previous example calculate the correlation coefficient, *r*.

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{(n\Sigma x^2 - (\Sigma x)^2)(n\Sigma y^2 - (\Sigma y)^2)}}$$

We have already calculated these totals:

$\Sigma x = 483 \qquad \Sigma x^2 = 31127 \qquad \Sigma xy = 15838$

$\Sigma y = 246 \qquad \Sigma y^2 = 8080 \qquad n = 8$

$$r = \frac{8 \times 15838 - 483 \times 246}{\sqrt{(8 \times 31127 - 483^2)(8 \times 8080 - 246^2)}}$$

$r = 0{\cdot}98$

This is close to 1 which is to be expected as we can see from the scatter diagram that the points lie very close to the regression line.

Exercise

Find the correlation coefficients for the data in the previous exercise. (Answers: 1, 0·91, 0·99)

The coefficient of determination is r^2. This tells us how much of the variation in y is explained by the variation in x. In the above example $r^2 = 0{\cdot}98^2$ $r^2 = 0{\cdot}96$ so we can say that 96% of the variation in y is explained by the variation in x.

Exercise

1–3. Find the coefficients of determination for the data in the previous exercise and interpret it.

4. The following table gives the price of a particular item and the number bought.

Price (£)	100	120	140	160	180
Quantity	15	10	9	8	5

Plot a scatter diagram and find the equation of the regression line if appropriate. Interpret the coefficient of regression. Calculate the correlation coefficient and coefficient of determination. Estimate how many items are bought if the price is a) £130 b) £210. Which is the more reliable estimate and why?

(Answers: 1. 1, 100% of the variation in y is explained by the variation in x.
2. 0·83, 83% of the variation in y is explained by the variation in x.
3. 0·97, 97% of the variation in the number of sultanas is explained by the variation in the weight.
4. q=24·8-0·11p, for each £ increase buy 0·11 less items, -0·95, 0·91, 11, 2. a) is more reliable as within the given range of the data.)