Mathematics

Edexcel IAL

S1

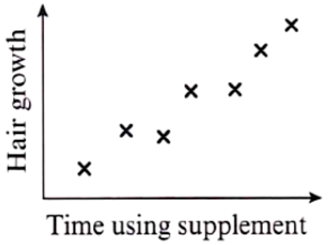
Worksheets

Correlation and Regression

Eng. Nagy Elraheb

Correlation and Regression:

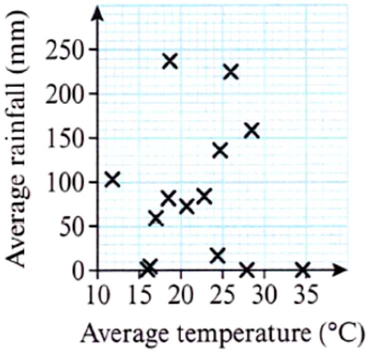
Exercise 1:



1 A lab technician researched the effectiveness of a herbal hair growth supplement. Seven people recorded their hair growth and this was compared with the length of time they had been using the supplement. A scatter diagram was drawn to represent the data.

a Describe the type of correlation shown by the scatter diagram.

b Interpret the correlation in context.



2 The average temperature and rainfall were collected for a number of cities around the world.
The scatter diagram shows this information.

a Describe the correlation between average temperature and average rainfall.

b Comment on the claim that hotter cities have less rainfall.

3 Eight students were asked to estimate the mass of a bag of rice in grams. First they were asked to estimate the mass without touching the bag and then they were told to pick up the bag and estimate the mass again. The results are shown in the table below.

Student	A	B	C	D	E	F	G	H
Estimate of mass before holding the bag (g)	25	18	32	27	21	35	28	30
Estimate of mass after holding the bag (g)	16	11	20	17	15	26	22	20

a Draw a scatter diagram to represent the data.

b Describe and interpret the correlation between the two variables.

4 Donal was interested to see if there was a relationship between the value of a house and the speed of its internet connection, as measured by the time taken to download a 100 Mb file. The table shows his results.

Time taken (s)	5.2	5.5	5.8	6.0	6.8	8.3	9.3	13	13.6	16.0
House value (€1000s)	300	310	270	200	230	205	208	235	175	180

a Draw a scatter diagram to represent the data.

b Describe the type of correlation shown.

Donal says that his data shows that a slow internet connection reduces the value of a house.

c Give one reason why Donal's conclusion may not be valid.

5 The table shows the daily total rainfall, r mm, and daily total hours of sunshine, s, in Edinburgh, for a random sample of 11 days in August.

r	0	6.8	0.9	4.9	0.1	22.3	1.8	4.5	0.1	2	0.2
s	8.4	4.9	10.5	4.1	3.3	4.2	5.8	1.8	10.2	1	4.6

The median and quartiles for the rainfall data are: $Q_1 = 0.1$ $Q_2 = 1.8$ $Q_3 = 4.9$

An outlier is defined as a value which lies either 1.5 × the interquartile range above the upper quartile or 1.5 × the interquartile range below the lower quartile.

a Show that $r = 22.3$ is an outlier. **(1 mark)**

b Give a reason why you might:
i include this day's readings **ii** exclude this day's readings. **(2 marks)**

c Exclude this day's readings and draw a scatter diagram to represent the data for the remaining ten days. **(3 marks)**

d Describe the correlation between rainfall and hours of sunshine. **(1 mark)**

e Do you think there is a causal relationship between the amount of rain and the hours of sunshine on a particular day? Explain your reasoning. **(1 mark)**

Exercise 2:

1 An accountant monitors the number of items produced per month by a company, together with the total production costs. The table shows these data.

Number of items, n (1000s)	21	39	48	24	72	75	15	35	62	81	12	56
Production costs, p (€1000s)	40	58	67	45	89	96	37	53	83	102	35	75

a Draw a scatter diagram to represent these data.

The equation of the regression line of p on n is $p = 21.0 + 0.98n$

b Draw the regression line on your scatter diagram.

c Interpret the meaning of the numbers 21.0 and 0.98

The company expects to produce 74 000 items in June, and 95 000 items in July.

d Comment on the suitability of this regression line equation to predict the production costs in each of these months.

2 The relationship between the number of coats of paint applied to a boat and the resulting weather resistance was tested in a laboratory. The data collected are shown in the table.

Coats of paint, x	**Protection, y (years)**
1	4.4
2	5.9
3	7.1
4	8.8
5	10.2

a Draw a scatter diagram to represent the data.

The equation of the regression line is $y = 2.93 + 1.45x$

Joti says that a gradient of 1.45 means that if 10 coats of paint are applied then the protection will last 14.5 years.

b Comment on Joti's statement.

3 The table shows the ages of some chickens and the number of eggs that they laid in a month.

Age of chicken, a (months)	18	32	44	60	71	79	99	109	118	140
Number of eggs laid in a month, n	16	18	13	7	12	7	11	13	6	9

a Draw a scatter diagram to show this information.

Mehmet calculates the regression line of n on a as $n = 16.1 + 0.063a$

b Without further calculation, explain why Mehmet's regression equation is incorrect.

4 Aisha collected data on the number of bedrooms, x, and the value, y (€1000s), of the houses in her village. She calculated the regression equation of y on x to be $y = 190 + 50x$

She states that the value of the constant in her regression equation means that a house with no bedrooms in her village would be worth €190 000. Explain why this is not a reasonable statement.

5 The table below shows data on the number of visitors to Ireland in a month, V (in '000s), and the amount of money they collectively spend, M (€, millions), for each of eight months.

Number of visitors, V (in '000s)	2450	2480	2540	2420	2350	2290	2400	2460
Amount of money spent, M (€, millions)	1370	1350	1400	1330	1270	1210	1330	1350

The equation on the regression line of M on $V = -467 + 0.740V$ (3 s.f.)

a Give an interpretation of the gradient of the regression line. **(2 marks)**

b Use the regression line to estimate the amount of money spent when the number of visitors to Ireland in a month is 2 200 000. **(2 marks)**

c Comment on the reliability of your estimate in part **b**.
Give a reason for your answer. **(2 marks)**

Exercise 3:

1 The equation of a regression line in the form $y = a + bx$ is to be found. Given that $S_{xx} = 15$, $S_{xy} = 90$, $\overline{x} = 3$ and $\overline{y} = 15$, work out the values of a and b.

2 Given that $S_{xx} = 30$, $S_{xy} = 165$, $\overline{x} = 4$ and $\overline{y} = 8$, find the equation of the regression line of y on x.

3 The equation of a regression line is to be found. The following summary data are given:

$S_{xx} = 40$ $\quad S_{xy} = 80$ $\quad \overline{x} = 6$ $\quad \overline{y} = 12$

Find the equation of the regression line in the form $y = a + bx$

4 Data are collected and summarised as follows:

$\sum x = 10$ $\quad \sum x^2 = 30$ $\quad \sum y = 48$ $\quad \sum xy = 140$ $\quad n = 4$

a Work out $\overline{x}$, $\overline{y}$, S_{xx} and S_{xy}

b Find the equation of the regression line of y on x in the form $y = a + bx$

5 For the data in the table,

x	2	4	5	8	10
y	3	7	8	13	17

Hint You can check your answer using the statistical functions on your calculator.

a calculate S_{xx} and S_{xy}

b find the equation of the regression line of y on x in the form $y = a + bx$

6 Research was done to see if there is a relationship between finger dexterity and the ability to do work on a production line. The data are shown in the table.

Dexterity score, x	2.5	3	3.5	4	5	5	5.5	6.5	7	8
Productivity, y	80	130	100	220	190	210	270	290	350	400

The equation of the regression line for these data is $y = -59 + 57x$

a Use the equation to estimate the productivity of someone with a dexterity of 6.

b Give an interpretation of the value of 57 in the equation of the regression line.

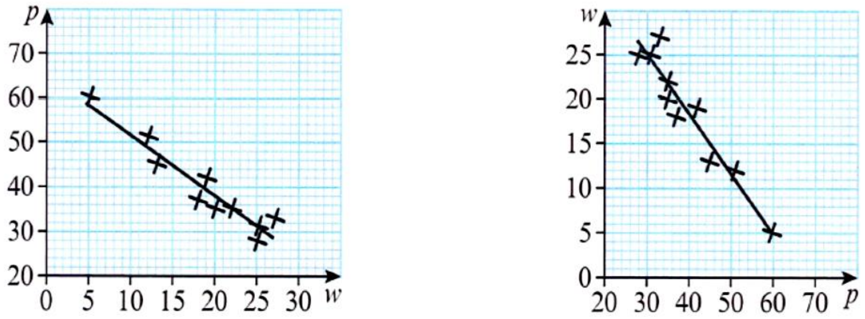
c State, giving a reason in each case, whether or not it would be reasonable to use this equation to work out the productivity of someone with a dexterity score of:

i 2 **ii** 14

7 A field was divided into 12 plots of equal area. Each plot was fertilised with a different amount of fertiliser (h). The yield of grain (g) was measured for each plot. Find the equation of the regression line of g on h in the form $g = a + bh$, given the following summary data.

$$\sum h = 22.09 \quad \sum g = 49.7 \quad \sum h^2 = 45.04 \quad \sum g^2 = 244.83 \quad \sum hg = 97.778 \quad n = 12$$

8 Research was done to see if there was a relationship between the number of hours in the working week (w) and productivity (p). The data are shown in the two scatter diagrams below.



(You may use $\sum p = 397 \quad \sum p^2 = 16\,643 \quad \sum w = 186 \quad \sum w^2 = 3886 \quad \sum pw = 6797$)

a Calculate the equation of the regression line of p on w.
Give your answer in the form $p = a - bw$

b Rearrange this equation into the form $w = c + dp$

The equation of the regression line of w on p is $w = 45.0 - 0.666p$

c Comment on the fact that your answer to part **b** is different to this equation.

d Decide which equation you should use to predict:

i the productivity for a 23-hour working week

ii the number of hours in a working week that achieves a productivity score of 40.

9 In an experiment, the mass of chemical produced, y, and the temperature, x, are recorded.

x (°C)	100	110	120	130	140	150	160	170	180	190	200
y (mg)	34	39	41	45	48	47	41	35	26	15	3

Maya thinks that the data can be modelled using a linear regression line.

a Calculate the equation of the regression line of y on x.
Give your answer in the form $y = a + bx$

b Draw a scatter diagram for these data.

c Comment on the validity of Maya's model.

10 An accountant monitors the number of items produced per month by a company (n) together with the total production costs (p). The table shows these data.

Number of items, n (1000s)	21	39	48	24	72	75	15	35	62	81	12	56
Production cost, p (€1000s)	40	58	67	45	89	96	37	53	83	102	35	75

(You may use $\sum n = 540$ $\quad \sum n^2 = 30\,786$
$\sum p = 780$ $\quad \sum p^2 = 56\,936$
$\sum np = 41\,444$)

Watch out The number of items are given in 1000s. Be careful to choose the correct value to substitute into your regression equation.

a Calculate S_{nn} and S_{np} **(2 marks)**

b Find the equation of the regression line of p on n in the form $p = a + bn$ **(3 marks)**

c Use your equation to estimate the production costs of 40 000 items. **(2 marks)**

d Comment on the reliability of your estimate. **(1 mark)**

11 A printing company produces leaflets for different advertisers. The number of leaflets, n, measured in 100s, and printing costs, \$$p$, are recorded for a random sample of 10 advertisers. The table shows these data.

n (100s)	1	3	4	6	8	12	15	18	20	25
p (\$)	22.5	27.5	30	35	40	50	57.5	65	70	82.5

(You may use $\sum n = 112$ $\quad \sum n^2 = 1844$ $\quad \sum p = 480$ $\quad \sum p^2 = 26\,725$ $\quad \sum np = 6850$)

a Calculate S_{nn} and S_{np} **(2 marks)**

b Find the equation of the regression line of p on n in the form $p = a + bn$ **(3 marks)**

c Give an interpretation of the value of b. **(1 mark)**

An advertiser is planning to print t hundred leaflets. A rival printing company charges 5 cents per leaflet.

d Find the range of values of t for which the first printing company is cheaper than the rival. **(2 marks)**

12 The relationship between the number of coats of paint applied to a boat and the resulting weather resistance was tested in a laboratory. The data collected are shown in the table.

Coats of paint, x	1	2	3	4	5
Protection, y (years)	1.4	2.9	4.1	5.8	7.2

a Find an equation of the regression line of y on x as a model for these results, giving your answer in the form $y = a + bx$ **(2 marks)**

b Interpret the value b in your model. **(1 mark)**

c Explain why this model would not be suitable for predicting the number of coats of paint that had been applied to a boat that had remained weather resistant for 7 years. **(1 mark)**

d Use your answer to part **a** to predict the number of years of protection when 7 coats of paint are applied. **(2 marks)**

In order to improve the reliability of its results, the laboratory made two further observations:

Coats of paint, x	6	8
Protection, y (years)	8.2	9.9

e Using all 7 data points:

i produce a refined model

ii use your new model to predict the number of years of protection when 7 coats of paint are applied

iii give two reasons why your new prediction might be more accurate than your original prediction. **(5 marks)**

Exercise 4:

1 Given that the coding $p = x + 2$ and $q = y - 3$ has been used to get the regression equation $p + q = 5$, find the equation of the regression line of y on x in the form $y = a + bx$

2 Given the coding $x = p - 10$ and $y = s - 100$ and the regression equation $x = y + 2$, work out the equation of the regression line of s on p.

3 Given that the coding $g = \frac{x}{3}$ and $h = \frac{y}{4} - 2$ has been used to get the regression equation $h = 6 - 4g$, find the equation of the regression line of y on x.

4 The regression line of t on s is found by using the coding $x = s - 5$ and $y = t - 10$. The regression equation of y on x is $y = 14 + 3x$. Work out the regression line of t on s.

5 A regression line of c on d is worked out using the coding $x = \frac{c}{2}$ and $y = \frac{d}{10}$

a Given that $S_{xy} = 120$, $S_{xx} = 240$,, $\bar{x} = 5$, and $\bar{y} = 6$, calculate the regression line of y on x.

b Find the regression line of d on c.

6 Some data on the coverage area, $a\,\text{m}^2$, and cost, $\$c$, of five boxes of flooring were collected.

The results were coded such that $x = \frac{a-8}{2}$ and $y = \frac{c}{5}$

The coded results are shown in the table.

x	1	5	10	16	17
y	9	12	16	21	23

a Calculate S_{xy} and S_{xx} and use them to find the equation of the regression line of y on x. **(4 marks)**

b Find the equation of the regression line of c on a. **(2 marks)**

c Estimate the cost of a box of flooring which covers an area of $32\,\text{m}^2$. **(2 marks)**

7 A farmer collected data on the annual rainfall, x cm, and the annual yield of potatoes, p tonnes per acre.

The data for annual rainfall were coded using $v = \frac{x-4}{8}$ and the following statistics were found:

$S_{vv} = 10.21$ $\quad S_{pv} = 15.26$ $\quad S_{pp} = 23.39$ $\quad \overline{p} = 9.88$ $\quad \overline{v} = 4.58$

a Find the equation of the regression line of p on v in the form $p = a + bv$ **(3 marks)**

b Using your regression line, estimate the annual yield of potatoes per acre when the annual rainfall is 42 cm. **(2 marks)**

Exercise 5:

1 Given that $S_{xx} = 92$, $S_{yy} = 112$ and $S_{xy} = 100$, find the value of the product moment correlation coefficient between x and y.

2 Given the following summary data,

$\sum x = 367 \quad \sum y = 270 \quad \sum x^2 = 33\,845 \quad \sum y^2 = 12\,976 \quad \sum xy = 17\,135 \quad n = 6$

calculate the product moment correlation coefficient, r, using the formula

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

3 The ages, a years, and heights, h cm, of seven members of a team were recorded. The data were summarised as follows:

$\sum a = 115 \quad \sum a^2 = 1899 \quad S_{hh} = 571.4 \quad S_{ah} = 72.1$

a Find S_{aa} **(1 mark)**

b Find the value of the product moment correlation coefficient between a and h. **(1 mark)**

c Describe and interpret the correlation between the age and height of these seven people based on these data. **(2 marks)**

4 In research on the quality of lamb produced by different breeds of sheep, data were obtained about the leanness, L, and taste, T, of the lamb. The data are shown in the table.

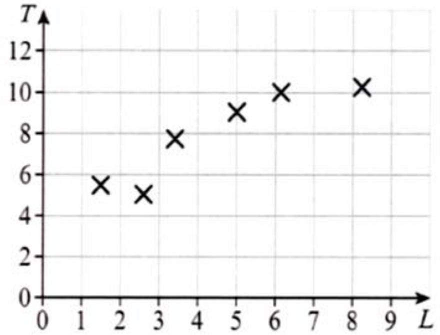
Leanness, L	1.5	2.6	3.4	5.0	6.1	8.2
Taste, T	5.5	5.0	7.7	9.0	10.0	10.2

a Find S_{LL}, S_{TT} and S_{LT} **(3 marks)**

b Calculate the product moment correlation coefficient between L and T using the values found in part **a**. **(2 marks)**

A scatter diagram is drawn for the data.

c With reference to your answer to part **b** and the scatter diagram, comment on the suitability of a linear regression model for these data. **(2 marks)**



5 Eight children had their IQ measured and then took a general knowledge test. Their IQ, x, and their marks, y, for the test were summarised as follows:

$\sum x = 973$ $\sum x^2 = 120\,123$ $\sum y = 490$ $\sum y^2 = 33\,000$ $\sum xy = 61\,595$

a Calculate the product moment correlation coefficient. **(3 marks)**

b Describe and interpret the correlation coefficient between IQ and general knowledge. **(2 marks)**

6 Two variables, x and y, were coded using $A = x - 7$ and $B = y - 100$
The product moment correlation coefficient between A and B is found to be 0.973
Find the product moment correlation coefficient between x and y.

7 The following data are to be coded using the coding $p = x$ and $q = y - 100$

x	0	5	3	2	1
y	100	117	112	110	106

a Complete a table showing the values of p and q.

b Use your values of p and q to find the product moment correlation coefficient between p and q.

c Hence write down the product moment correlation coefficient between x and y.

8 The PMCC is to be worked out for the following data set using coding.

x	50	40	55	45	60
y	4	3	5	4	6

a Using the coding $p = \frac{x}{5}$ and $t = y$, find the values of S_{pp}, S_{tt} and S_{pt}

b Calculate the product moment correlation coefficient between p and t.

c Write down the product moment correlation coefficient between x and y.

9 A shopkeeper thinks that the more newspapers he sells in a week the more sweets he sells. He records the amount of money (m dinars) that he takes in newspaper sales and also the amount of money he takes in sweet sales (s dinars) each week for seven weeks. The data are shown in the following table.

Newspaper sales, m (dinars)	380	402	370	365	410	392	385
Sweet sales, s (dinars)	560	543	564	573	550	544	530

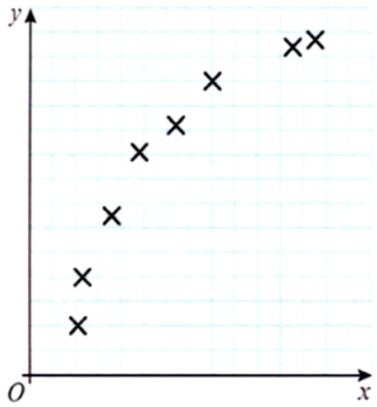
a Use the coding $x = m - 365$ and $y = s - 530$ to find S_{xx}, S_{yy} and S_{xy} **(4 marks)**

b Calculate the product moment correlation coefficient for m and s. **(1 mark)**

c State, with a reason, whether or not what the shopkeeper thinks is correct. **(1 mark)**

10 A student vet collected 8 blood samples from a horse with an infection. For each sample, the vet recorded the amount of drug, f, given to the horse and the amount of antibodies present in the blood, g. She coded the data using $f = 10x$ and $g = 5(y + 10)$ and drew a scatter diagram of x against y.

Unfortunately, she forgot to label the axes on her scatter diagram and left the summary data calculations incomplete.



$\sum g^2 = 74\,458.75$

$S_{fg} = 5667.5$

$\sum y = 70.9$

$S_{xx} = 111.48$

A second student was asked to complete the analysis of the data.

a Show that $S_{ff} = 11\,148$ **(3 marks)**

Problem-solving

Substitute the code into the formula for S_{xx}

b Find the value of the product moment correlation coefficient between f and g. **(4 marks)**

c With reference to the scatter diagram, comment on the result in part **b**. **(1 mark)**

11 Ji-yoo, a market gardener, measures the amount of fertiliser, x litres, that she adds to the compost for a random sample of 7 chilli plant beds. She also measures the yield of chillies, y kg. The data are shown in the table below:

x, litres	1.1	1.3	1.4	1.7	1.9	2.1	2.5
y, kg	6.2	10.5	12	15	17	18	19

$\left(\sum x = 12 \quad \sum x^2 = 22.02 \quad \sum y = 97.7 \quad \sum y^2 = 1491.69 \quad \sum xy = 180.37\right)$

a Show that the product moment correlation coefficient for these data is 0.946, correct to 3 significant figures. **(4 marks)**

The equation of the regression line of y on x is given as $y = -1.2905 + 8.8945x$

b Calculate the residuals. **(3 marks)**

Ji-yoo thinks that because the PMCC is close to 1, a linear relationship is a good model for these data.

c With reference to the residuals, evaluate Ji-yoo's conclusion. **(2 marks)**