CORRELATION AND REGRESSION

Text

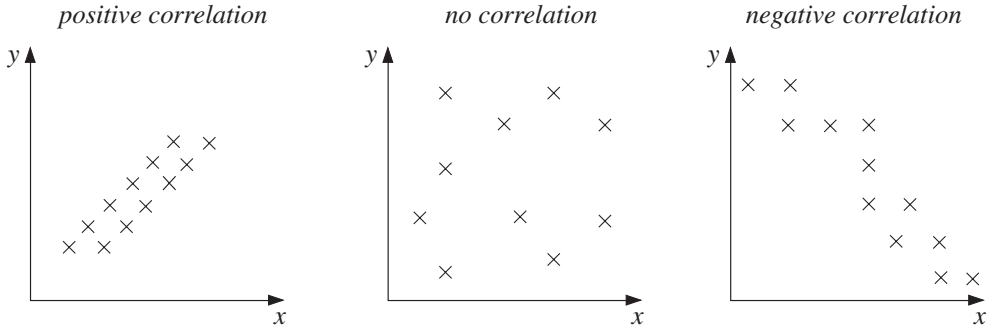
Contents

© University of Plymouth

Correlation and Regression

1 Correlation

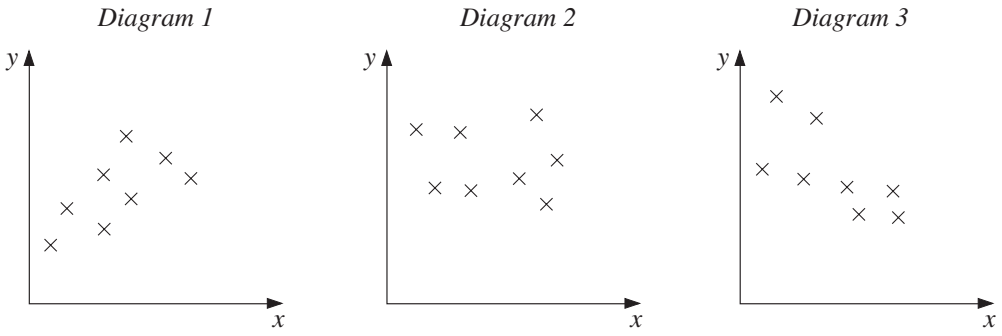
You have probably already met the concept of *correlation* between two sets of data and will have used scatter diagrams for plotting a series of data points (x_i, y_i), and then defined:



These are examples of *bivariate* data, where two variables are related.

The next two Worked Examples illustrate what you need to know.

Worked Example 1



(a) Write down the number of the diagram that is most likely to represent

(i) the heights of a group of boys on the x axis and the sizes of shirts worn by them on the y axis,

(ii) the mean temperature during the day on the x axis and the amount of gas used to heat a house on that day on the y axis,

(iii) the shoe sizes of a group of adults on the x axis and their ages on the y axis.

© University of Plymouth

(b) Which diagram shows two variables which have

(i) positive correlation,

(ii) negative correlation,

(iii) no correlation?

Solution

(a) (i) Diagram 1 (size increase with height)

(ii) Diagram 3 (the colder the day, the more gas is used)

(iii) Diagram 2 (no particular relationship between shoe size and age)

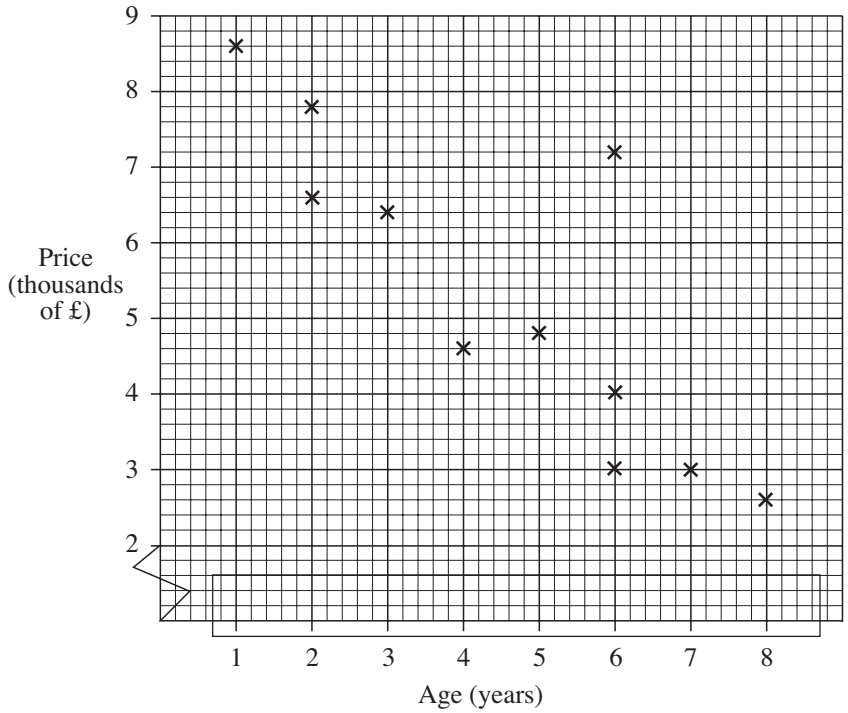
(b) (i) Diagram 1

(ii) Diagram 3

(iii) Diagram 2

Worked Example 2

The scatter diagram shows the ages, in years, and the selling prices, in thousands of pounds, of second-hand cars of the same model. The cars have been advertised for sale in a local paper.



The price of one of these cars has been advertised wrongly.

(a) Give the age and price of the car that you think is incorrectly advertised.

(b) How many cars are being advertised for sale?

© University of Plymouth

(c) Another car is to be included in the advertisement next week.

The car is four years old.

Do you think the price will be more than £6500 or less?

Give a reason for your answer.

(d) What type of correlation does the diagram show?

Solution

(a) 6 years and £7200

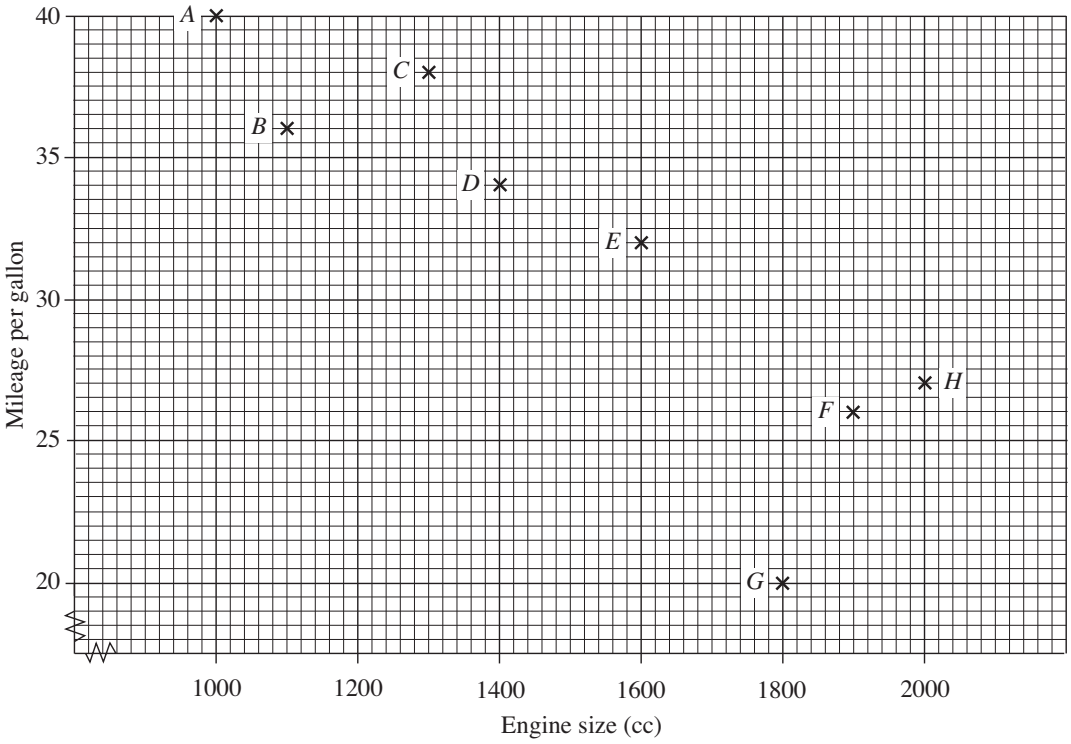
(b) 11

(c) Less than £6500 if it follows the above trend.

(d) Negative correlation

Exercises

1. A guide to used cars shows the engine size in cc and the mileage per gallon.



(a) Complete a copy of the table below.

Car	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Engine size (cc)	1000	1100	1300	1400		1900	1800	
Mileage per gallon	40	36	38	34	32		20	

(b) Another car, with engine size 1600 cc, was tested and its mileage was 30 per gallon. Plot this on the diagram, labelling it *I*.

(c) One car does not appear to follow the trend. Which one is it? Give a reason for your answer.

© University of Plymouth

2. Mary was ill. Her temperature was taken and recorded every hour between 12 noon and 6 pm.

Time	12 noon	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm
Temp (°C)	37.0	39.5	40.0	39.0	38.2	37.7	37.2

(a) Draw a suitable diagram to represent these data.

(b) Use your diagram to estimate:

(i) the time at which Mary's temperature first rose above 38.5 °C.

(ii) Mary's temperature at 3.30 pm.

A normal temperature is 37 °C.

(c) Her temperature continues to fall at the same rate until it reaches 37 °C.

Use your diagram to estimate the time at which Mary's temperature returns to 37 °C.

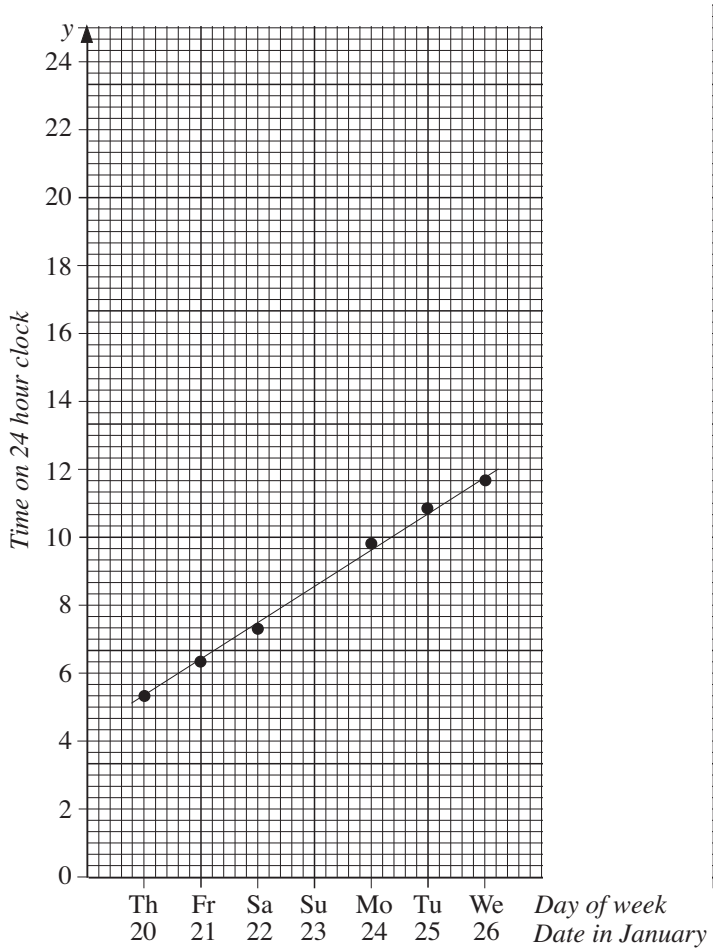
3. Southend has two high tides every 24 hours.

The table shows some high sides at Southend for a period in January.

		Time	
Day and Date		*1st high tide*	*2nd high tide*
Thursday	20	05.27	18.05
Friday	21	06.22	19.03
Saturday	22	07.29	20.10
Sunday	23	*	*
Monday	24	09.54	22.25
Tuesday	25	10.52	23.17
Wednesday	26	11.40	*

The time for the 1st high tide of each day has been plotted on the following axes.

© University of Plymouth



(a) On the same axes plot the time for the 2nd high tide of each day.

(b) Use your graph to estimate the times of high tide on Sunday 23 January.

High tide at Tilbury is always 30 minutes later than at Southend.

(c) Calculate the next two high tide times at Tilbury *after* Tuesday 25 January.

4. The following data relate to the age and weight of ten randomly chosen children in Bedway Primary School.

Age (years)	7.8	8.1	6.4	5.2	7.0	9.9	8.4	6.0	7.2	10.0
Weight (kg)	29	28	26	20	24	35	30	22	25	36

(a) Draw a scatter diagram to show this information.

The mean age of this group of children is 7.6 years.

(b) Calculate the mean weight of this group.

(c) On the graph, draw the line of best fit.

(d) Use your graph to find the equation of this line of best fit, in the form $y = mx + c$.

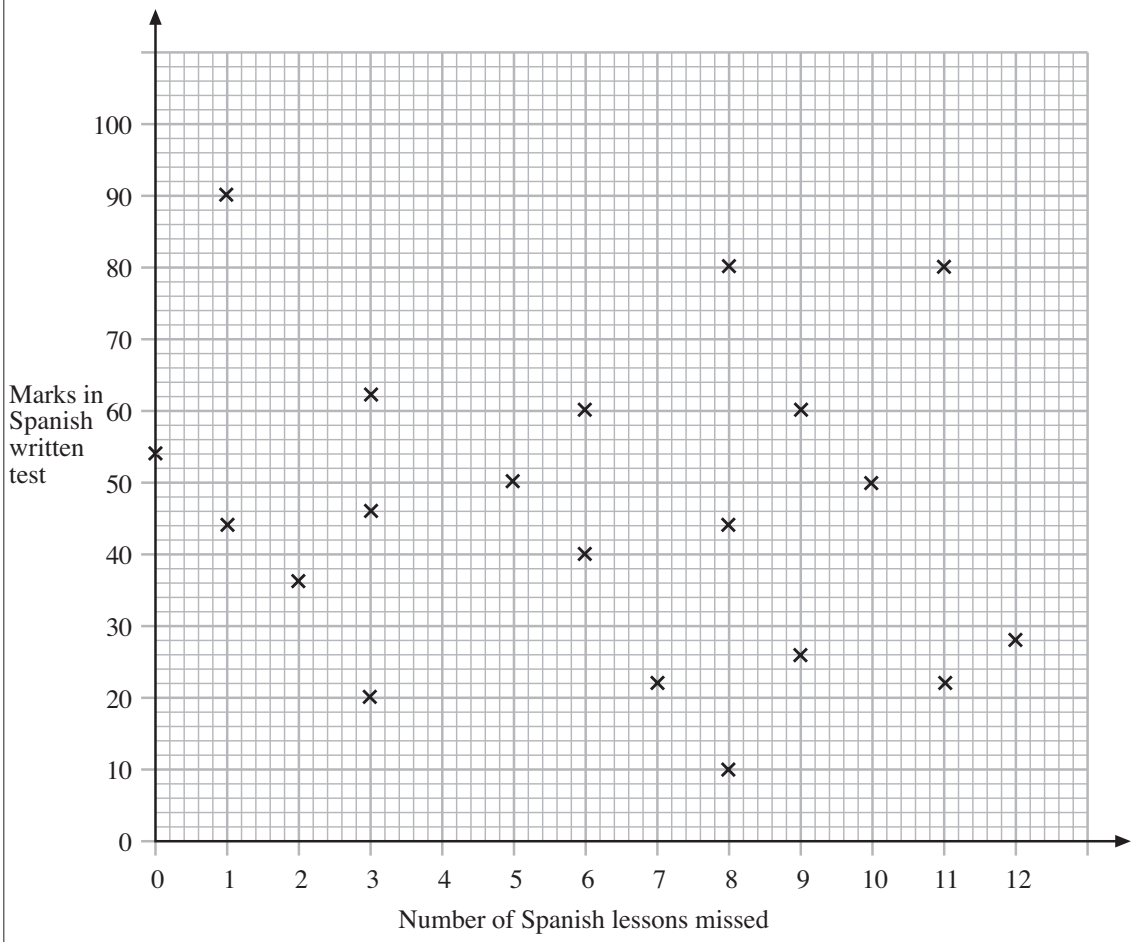
Jane is a pupil at Bedway Primary School and her age is 8.0 years.

© University of Plymouth

(e) Use your answer to (d) to estimate Jane's weight.

(f) Give *one* reason why a prediction of the weight of a twelve year old from your graph might not be reliable.

5. 20 students take a Spanish written test
The scatter diagram shows their marks and the number of Spanish lessons they had missed during the year.



(a) Write down the mark of the student who missed most lessons.

(b) Write down the number of lessons missed by the student having a mark of 36.

(c) One student missed many lessons but still had a high mark in the test.
Write down the mark and number of lessons missed by this student.

(d) The teacher looks at the scatter diagram and concludes:

"*The more Spanish lessons a student attends, the higher their mark in the written test*."

Does the information in the scatter diagram support this conclusion?
Give a reason for your answer.

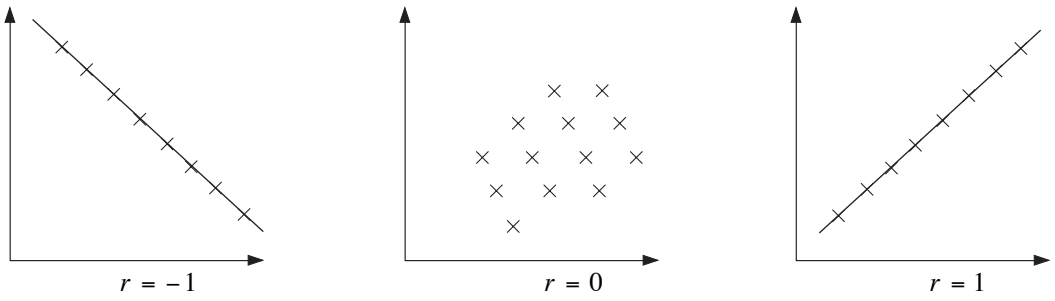
© University of Plymouth

2 Spearman's Rank Coefficient of Correlation

This is a method used to assign a meaning to the correlation between pairs of data points. Such a coefficient, call it r, is designed so that

$$-1 \le r \le 1$$

and $r = -1$ corresponds to *perfect negative correlation*, $r = 0$ to *no correlation* and $r = 1$ to *perfect positive correlation* (as illustrated below).



Spearman's rank correlation coefficient is based on the squares of the differences between data points when they have been ranked – that is, put in numerical order and then given the values 1, 2, 3, ..., etc. The formula is

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

Here n is the number of data points and d the difference between values

You will see a justification for this in the final worked example, but first we will see how to use the formula.

Worked Example 1

At the Deepdale 'Best of British Pie' competition two judges award marks for nine different pies as follows:

Pie	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Judge 1	18	24	23	13	27	19	30	10	20
Judge 2	7	18	9	4	17	8	29	5	10

(a) What do the scores tell you about the two judges?

(b) (i) Calculate Spearman's coefficient of rank correlation between the two judges.

(ii) What does your result tell you about the judges' decision?

Solution

(a) The first judge appears to be using much higher scores than the second judge.

(b) (i) We first find the 'ranks' and then the differences, and square them (note that squaring a negative number results in a positive value).

© University of Plymouth

Pie	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Judge 1	18	24	23	13	27	19	30	10	20
Judge 2	7	18	9	4	17	8	29	5	10
Rank 1	3	7	6	2	8	4	9	1	5
Rank 2	3	8	5	1	7	4	9	2	6
d	0	–1	1	1	1	0	0	–1	–1
d^2	0	1	1	1	1	0	0	1	1

Summing the d^2 gives

$$\sum d^2 = 0 + 1 + 1 + 1 + 1 + 0 + 0 + 1 + 1 = 6$$

and, using the formula,

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} \qquad \text{with } \sum d^2 = 6 \text{ and } n = 9 \text{ gives}$$

$$r = 1 - \frac{6 \times 6}{9 \times 80} = 1 - 0.05 = 0.95$$

(ii) The value of r is very close to 1, showing that there is highly positive correlation between the two judges' rankings (but not in their actual scores).

Worked Example 2

An investigation was conducted by a company on the value of various assessment methods for recruiting employees. The data are shown in this table.

Employee	Educational Test Score	Assessment Score by Personnel Officer
A	9	12
B	10	14
C	15	16
D	14	15
E	16	17
F	11	10
G	12	11
H	17	18

This is based on 8 employees, giving their educational test scores, together with an assessment score by the Personnel Officer of their ability one year after joining the company. Possible test scores in each case can range from a low of 1 to a high of 20.

© University of Plymouth

(a) Rank each employee in terms of Educational Test score and Assessment score by the Personnel Officer.

(b) Hence for these scores calculate, to 2 decimal places, the Spearman rank correlation coefficient.

(c) The recruits also took an aptitude test and the comparable value for the rank coefficient based on aptitude test score and assessment score by the Personnel Officer was – 0.21.

With reference to this result and your answer in (b) comment on the effectiveness of the tests in providing the Personnel Officer with an indication of the suitability of applicants for employment.

Solution

(a)

Employee	Educational Test Score	Assessment Score by Personnel Officer	R_E	R_A	d	d^2
A	9	12	1	3	–2	4
B	10	14	2	4	–2	4
C	15	16	6	6	0	0
D	14	15	5	5	0	0
E	16	17	7	7	0	0
F	11	10	3	1	2	4
G	12	11	4	2	2	4
H	17	18	8	8	0	0

(b) $n = 8$ and $\sum d^2 = 16$, so

$$r = 1 - \frac{6 \times 16}{8 \times 63} = 1 - 0.19 = 0.81$$

(c) The educational test seems to work well (fairly positive correlation) but the aptitude test does not work well (slightly negative correlation).

Worked Example 3

In a music festival, each competitor is judged on his performance on two different musical instruments. The judge awards marks out of 100 for each instrument, as follows.

Competitor	*A*	*B*	*C*	*D*	*E*	*F*
1st Instrument	90	75	62	70	75	56
2nd Instrument	95	76	64	76	86	60

(a) Complete a table of ranks.

The rank correlation coefficient for these data was found to be 0.96.

© University of Plymouth

It was later discovered that the marks from one of the judges, for one competitor, had been misread. This competitor should have had 10 more marks on his second instrument.

The mark was changed and on recalculation it was found that the correlation coefficient remained the same at 0.96.

(b) (i) Which competitor's mark was originally incorrect?

(ii) Give a reason for your answer.

Solution

(a)

Competitor	*A*	*B*	*C*	*D*	*E*	*F*
1st Instrument	90	75	62	70	75	56
2nd Instrument	95	76	64	76	86	60
Rank 1	1	2.5	5	4	2.5	6
Rank 2	1	3.5	5	3.5	2	6

(Note that we have marked the highest rather than the lowest as 1; this is not a problem provided that both sets of rankings are done in the same way.

Also, if there are two tied ranks, we use the average of the two values, here 2 and 3, so we use 2.5 for each.)

(b) (i) Competitor C

(ii) Only the '64' entry (i.e. competitor C, 2nd instrument) will not have their rank affected by an increase of 10 marks. Hence competitor C must have the incorrect mark.

Worked Example 4

For sets of paired data, find the value of $\sum d^2$ for

(i) perfect positive correlation,

(ii) perfect negative correlation when $n = 2, 3, 4, \ldots, 8$.

Hence deduce Spearman's rank correlation coefficient formula, assuming it is of the form

$$r = 1 - k\sum d^2$$

Solution

Perfect positive correlation

n	A	B	d	d^2	
2	1	1	0	0	
	2	2	0	0	$\sum d^2 = 0$
3	1	1	0	0	
	2	2	0	0	$\sum d^2 = 0$
	3	3	0	0	

Perfect negative correlation

n	A	B	d	d^2	
	1	2	–1	1	
2	2	1	1	1	$\sum d^2 = 2$
	1	3	–2	4	
3	2	2	0	0	
	3	1	2	4	$\sum d^2 = 8$

© University of Plymouth

Perfect positive correlation

$\sum d^2 = 0$ for perfect positive correlation for all values of n.

Perfect negative correlation

4	1	4	–3	9	
	2	3	–1	1	
	3	2	1	1	$\sum d^2 = 20$
	4	1	3	9	
5	1	5	–4	16	
	2	4	–2	4	
	3	3	0	0	
	4	2	2	4	$\sum d^2 = 40$
	5	1	4	16	
6	1	6	–5	25	
	2	5	–3	9	
	3	4	–1	1	
	4	3	1	1	
	5	2	3	9	$\sum d^2 = 70$
	6	1	5	25	
7	1	7	–6	36	
	2	6	–4	16	
	3	5	–2	4	
	4	4	0	0	
	5	3	2	4	
	6	2	4	16	$\sum d^2 = 112$
	7	1	6	36	
8	1	8	–7	49	
	2	7	–5	25	
	3	6	–3	9	
	4	5	–1	1	
	5	4	1	1	
	6	3	3	9	
	7	2	5	25	$\sum d^2 = 168$
	8	1	7	49	

Assume that Spearman's correlation coefficient takes the form

$$r = 1 - k\sum d^2$$

(since this gives $r = 1$ when $\sum d^2 = 0$, i.e. perfect positive correlation).

Now the constant k will depend on n (the number of data points) and must be chosen so that when there is perfect negative correlation, then $r = -1$; i.e.

$$-1 = 1 - k\sum d^2 \Rightarrow k = \frac{2}{\sum d^2}$$

© University of Plymouth

Tabulating the values obtained gives

n	$\sum d^2$
2	2
3	8
4	20
5	40
6	70
7	112
8	168

from which you can see that $\sum d^2 = \frac{n}{3}\left(n^2 - 1\right)$ fits these values. (You could in fact first deduce that it is a cubic expression since the third differences are constant, and then fit a general cubic to the data.)

Hence, using the formula for $\sum d^2$,

$$k = \frac{6}{n\left(n^2 - 1\right)}$$

and

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

(which is Spearman's formula).

Exercises

1. There was a vacancy for a typist at Betterprint.

   Six people applied for the job.

   The manager gave each applicant a test, which consisted of typing a page of writing. Marks were awarded for the speed and for the accuracy of the typing.

   The following table shows the results of the test.

Typist	*A*	*B*	*C*	*D*	*E*	*F*
Time to complete (seconds)	56	44	60	50	80	30
Number of errors	3	4	2	4	1	8

   (a) Copy and complete the following table of ranks.

© University of Plymouth

	A	*B*	*C*	*D*	*E*	*F*
Rank Time						
Rank errors						
d						
d^2						

(b) Given that Spearman's rank correlation coefficient is

$$1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

use the table to calculate this coefficient for these data.

The manager decided to offer the typist vacancy to applicant A.

(c) Give a reason why you think this was a sensible decision.

2. Mrs Maden is a wine expert. At a wine tasting evening she is asked to taste the wines of a producer taken from each of ten different years and place them in order of quality. She regards 1983 as the best drink and ranks it 1.

Year	*1990*	*1989*	*1988*	*1987*	*1986*	*1985*	*1984*	*1983*	*1982*	*1981*
Rank age of wine	10	9	8	7	6	5	4	3	2	1
Rank quality	8	3	7	6	2	9	5	1	4	10

(a) Calculate Spearman's coefficient of rank correlation between the age and the quality of the wine. The formula for calculating Spearman's rank correlation coefficient is

$$p = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

(b) On the basis of your answer to (a) comment on the statement 'wine improves with age'.

(c) A similar tasting between the age and quality of beer resulted in a correlation coefficient of –1. What does this suggest about the age and quality of beer?

3. The following table shows the positions in a Sunday league of 8 cricket clubs at the end of a season together with the average attendances (in hundreds) at their home matches during the season.

© University of Plymouth

Club	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Position in league	1	3	6	2	7	8	5	4
Average attendance	34	12	18	32	15	25	27	19
Rank of attendance	1	8						
Difference in ranks (d)	0	5						
d^2	0	25						

(a) Copy and complete the table.

(b) Using the formula $1 - \dfrac{6\sum d^2}{n\left(n^2 - 1\right)}$

calculate Spearman's rank correlation coefficient for these data.

(c) Explain what the value you have calculated in (b) shows.

(d) If the value of the rank correlation coefficient had been $+0.95$ describe what this would have implied in relation to position in the league and average league attendance for each club.

3 Product Moment Correlation Coefficient

This is also known as Pearson's correlation coefficient and, like Spearman's, it satisfies

$$-1 \le r \le 1$$

but it is not based on ranking the data but uses the actual data values. We will first introduce the idea of 'covariance'.

Worked Example 1

The data below gives the marks obtained by 10 pupils taking Maths and Physics tests.

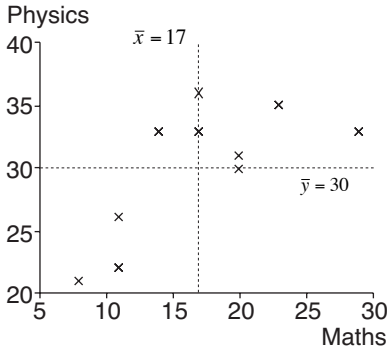
Pupil	A	B	C	D	E	F	G	H	I	J
Maths mark (out of 30) x	20	23	8	29	14	11	11	20	17	17
Physics mark (out of 40) y	30	35	21	33	33	26	22	31	33	36

Is there a connection between the marks gained by ten pupils, A, B, C ..., J in Maths and Physics tests?

© University of Plymouth

Solution

As a starting point, plot the marks as a scatter diagram.



The areas in the bottom right and top left of the graph are largely vacant so there is a tendency for the points to run from bottom left to top right.

Now calculate the means,

$$\bar{x} = \frac{170}{10} = 17$$

and $$\bar{y} = \frac{300}{10} = 30$$

Using the means to divide the graph into four shows this tendency clearly.

The problem though is to find a way to measure how strong the tendency is.
We can attempt to quantify the tendency to go from bottom left to top right by evaluating the expression

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

which is known as the **covariance** and denoted by $\text{cov}(X,Y)$ or s_{xy}. For shorthand it is normally written as

$$\frac{1}{n}\sum (x - \bar{x})(y - \bar{y})$$

where the summation over *i* is assumed.

The points in the top right have *x* and *y* values greater than $\bar{x}$ and $\bar{y}$ respectively, so $x - \bar{x}$ and $y - \bar{y}$ are both positive and so is the product $(x - \bar{x})(y - \bar{y})$.

Those in the bottom left have values less than $\bar{x}$ and $\bar{y}$, so $x - \bar{x}$ and $y - \bar{y}$ are both negative and again the product $(x - \bar{x})(y - \bar{y})$ is positive.

Points in the other two areas have one of $x - \bar{x}$ and $y - \bar{y}$ positive and the other negative, so $x - \bar{x}$ and $y - \bar{y}$ is negative.

The $\frac{1}{n}$ factor accounts for the fact that the number of points will affect the value of the covariance.

In the example above, most of the points give positive values of $(x - \bar{x})(y - \bar{y})$.

There is another form of the expression for covariance which is easier to use in calculations.

$$\frac{1}{n}\sum(x - \bar{x})(y - \bar{y}) = \frac{1}{n}\sum(xy - \bar{x}y - x\bar{y} + \bar{x}\bar{y})$$

$$= \frac{1}{n}\left(\sum xy - \sum \bar{x}y - \sum x\bar{y} + \sum \bar{x}\bar{y}\right)$$

© University of Plymouth

$$= \frac{1}{n}\left(\sum xy - \bar{x}\sum y - \bar{y}\sum x + n\bar{x}\bar{y}\right)$$

$$= \frac{1}{n}\left(\sum xy - \bar{x}n\bar{y} - \bar{y}n\bar{x} + n\bar{x}\bar{y}\right) \qquad \text{since } \bar{y} = \frac{\sum y}{n}, \quad \bar{x} = \frac{\sum x}{n}$$

$$= \frac{1}{n}\left(\sum xy - n\bar{x}\bar{y}\right)$$

Thus $$\boxed{\frac{1}{n}\sum (x-\bar{x})(y-\bar{y}) = \frac{1}{n}\sum xy - \bar{x}\bar{y}}$$

The right hand side is quicker to evaluate. For the example on page 15, this form of the expression is usually used when calculating covariance.

$$s_{xy} = \frac{1}{10}\Sigma xy - 17 \times 30$$

$$= \frac{1}{10} \times 5313 - 510$$

$$= 21.3$$

(Σxy is a function available on calculators with LR mode.)

The fact that $s_{xy} > 0$ indicates that the points follow a trend with a positive slope. The size of the number, however, conveys little as it can easily be altered by a change of scale.

The following examples show this.

Worked Example 2

Find the covariance for the following data.

(a)

Height (m) x	1.60	1.64	1.71
Weight (kg) y	53	57	60

(b)

Height (cm) x	160	164	171
Weight (kg) y	53	57	60

Solution

(a) $$s_{xy} = \frac{1}{3} \times 280.88 - \frac{170}{3} \times \frac{4.95}{3}$$

$$= 0.12\dot{6}$$

(b) $$s_{xy} = \frac{1}{3} \times 28088 - \frac{170}{3} \times \frac{495}{3}$$

$$= 12.\dot{6}$$

You can, of course, get quite different values by measuring in pounds and inches or kg and feet, etc. They will all be positive but their sizes will not convey useful information.

© University of Plymouth

–

$$\frac{x-\bar{x}}{s_x}$$

will give the same answer regardless of the units or scale involved. The quantity

$$\frac{1}{n}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

can therefore be relied on to produce a value with more meaning than the covariance.

Since $$\frac{1}{n}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{s_x s_y}$$

and the latter is easier to evaluate, **Pearson's product moment correlation coefficient** is often given as

$$\boxed{r = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{s_x s_y}}$$

where $s_x = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2}$ and $s_y = \sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2}$.

(Note that *r* is a function given on calculators with LR mode.)

Returning to the Worked Example above:

Pupil	A	B	C	D	E	F	G	H	I	J
Maths mark (out of 30) x	20	23	8	29	14	11	11	20	17	17
Physics mark (out of 40) y	30	35	21	33	33	26	22	31	33	36

$$r = \frac{\frac{1}{10}\times 5313 - 17\times 30}{s_x \times s_y}$$

$$s_x = \sqrt{\frac{1}{10}\times 3250 - 17^2} = \sqrt{36} = 6$$

$$s_y = \sqrt{\frac{1}{10}\times 9250 - 30^2} = \sqrt{25} = 5$$

$$\Rightarrow \quad r = \frac{531.3 - 510}{6\times 5}$$

$$= 0.71$$

© University of Plymouth

Worked Example 3

A group of twelve children participated in a psychological study designed to assess the relationship, if any, between age, x years, and average total sleep time (ATST), y minutes. To obtain a measure for ATST, recordings were taken on each child on five consecutive nights and then averaged. The results obtained are shown in the table.

Child	Age (***x* years**)	ATST (***y* minutes**)
A	4.4	586
B	6.7	565
C	10.5	515
D	9.6	532
E	12.4	478
F	5.5	560
G	11.1	493
H	8.6	533
I	14.0	575
J	10.1	490
K	7.2	530
L	7.9	515

$\sum x = 108 \ \sum y = 6372 \ \sum x^2 = 1060.1 \ \sum y^2 = 3396942 \ \sum xy = 56825.4$

Calculate the value of the product moment correlation coefficient between x and y. Assess the statistical significance of your value and interpret your results.

Solution

Use the formula

$$s_{xy} = \frac{1}{n}\sum xy - \bar{x}\bar{y}$$

when $\bar{x} = \frac{108}{12} = 9$ and $\bar{y} = \frac{6372}{12} = 531$

Thus $s_{xy} = \frac{1}{12}(56825.4) - 9 \times 531 = -43.55$

Also $$s_x = \sqrt{\frac{1}{12} \times 1060.1 - 9^2} \approx 2.7096$$

$$s_y = \sqrt{\frac{1}{12} \times 3396942 - 531^2} \approx 33.4290$$

Hence $$r = \frac{-43.55}{2.7096 \times 33.4290} \approx -0.481$$

© University of Plymouth

Exercises

1. For each of the following sets of data,

   (a) draw a scatter diagram

   (b) calculate the product moment correlation coefficient.

   (i)

x	1	3	6	10	12
y	5	13	25	41	49

   (ii)

x	1	3	5	7	9
y	44	34	24	14	4

   (iii)

x	1	1	3	5	5
y	5	1	3	1	5

   (iv)

x	1	3	6	9	11
y	12	28	37	28	12

2. (a) Calculate the value of r for the random variables X and Y using the following values

x	11	17	26
y	23	18	19

   (b) The random variable Z is converted to Y by the equation $Z = \frac{Y}{10} + 3$.

x	11	17	26
z			

   Complete the table above and evaluate r for X and Z.

   (c) State the value of r for Y and Z.

3. A metal rod was gradually heated and its length, L, was measured at various temperatures, T.

Temperature (°C)	15	20	25	30	35	40
Length (cm)	100	103.8	106.1	112	116.1	119.9

   (a) Draw a scatter diagram to show the data and evaluate r. (Plot L against T.)

   (b) Do you suspect a major inaccuracy in any of the recorded values? If so, discard any you consider untrustworthy and find the new value of r.

© University of Plymouth

4. The diameter of the longest lichens growing on gravestones were measured.

Age of gravestone x (years)	**Diameter of lichen y (mm)**
9	2
18	3
20	4
31	20
44	22
52	41
53	35
61	22
63	28
63	32
64	35
64	41
114	51
141	52

(a) Draw a scatter diagram to show the data.

(b) Calculate the values of $\bar{x}$ and $\bar{y}$ and show these on the diagram.

(c) Find the values of s_x, s_y and r.

5. In a biology experiment a number of cultures were grown in the laboratory. The numbers of bacteria, in millions, and their ages, in days, are given below.

Age (x)	1	2	3	4	5	6	7	8
No. of bacteria (y)	34	106	135	181	192	231	268	300

(a) (i) Plot these on a scatter diagram with the x-axis having a scale up to 15 days and the y-axis up to 410 millions.

(ii) Calculate the value of r and comment on your results.

(b) Some late readings were taken and are given below.

x	13	14	15
y	400	403	405

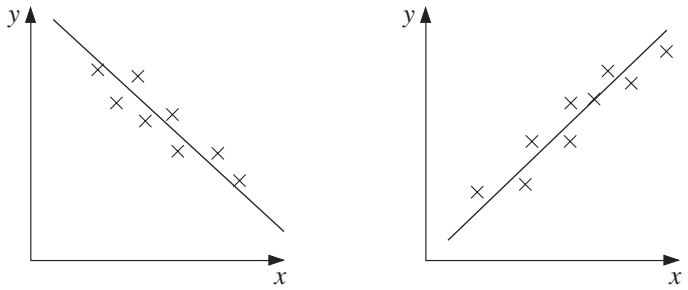
(i) Add these points to your graph.

(ii) Describe what these points show.

© University of Plymouth

4 Regression Lines

You will have met the idea of trend lines – these are essentially 'lines of best fit'. Note that there is little value in attempting to draw lines of best fit unless there is either strong positive or strong negative correlation between the points plotted, as shown in the following diagrams.



Also note that the lines of best fit should always pass through the point representing the mean values, $\bar{x}$ and $\bar{y}$, of the data points.

Worked Example 1

A sample of 8 U.S. Companies showed the following Sales and Profit levels for the year ending April 1994.

Sales Turnover	($m) ($x$)	22	36	26	14	25	34	6	18
Profit	($m) (y)	1.8	4.9	0.8	0.9	3.2	3.7	0.5	2.1

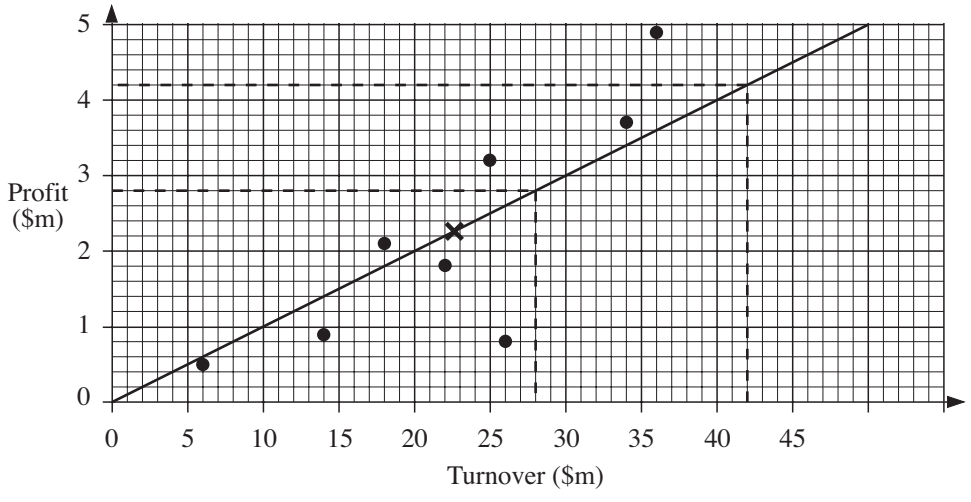
(a) Draw a scatter diagram of this information.

(b) After making suitable calculations draw in a line of best fit and use this to estimate Profit levels for *two* companies with annual turnovers respectively of $28m and $42m.

(c) State briefly which of the estimates in (b) is likely to be more accurate. Justify your choice.

Solution

(a)



(b) The mean values are calculated as $\bar{x} = 22.6$, $\bar{y} = 2.24$, and shown on the scatter diagram. (The line of best fit will pass through this point.)

For $x = \$28\text{m}$ the estimate of the profit is \$2.8m, and for $x = \$42\text{m}$, the estimate is \$4.2m.

(c) The estimate for a turnover of \$28m is likely to be more accurate than for \$42m, as the latter is outside the range of data on which the line of best fit is based.

Worked Example 2

Brunel plc is keen to set up a forecasting system which will enable them to estimate maintenance for delivery vehicles of various ages.

The following table summarises the age in months (x) and maintenance cost (y) for a sample of ten such vehicles.

Vehicle	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Age, months (x)	63	13	34	80	51	14	45	74	24	82
Maintenance Cost £, (y)	141	14	43	170	95	21	72	152	31	171

(a) Draw a scatter diagram of the data on graph paper.

(b) Find the mean value of the ages (x) and maintenance cost (y).

(c) Use your results from (b) and the fact that the line of best fit for the data passes through the point (20, 24.5) to draw this line on the graph.

(d) Estimate from your line the maintenance cost for a vehicle aged

(i) 85 months

(ii) 5 months

(iii) 60 months.

(e) Order these forecasts in terms of their reliability, listing the most reliable first. Justify your choice.

Solution

(a) See the scatter diagram on the following page.

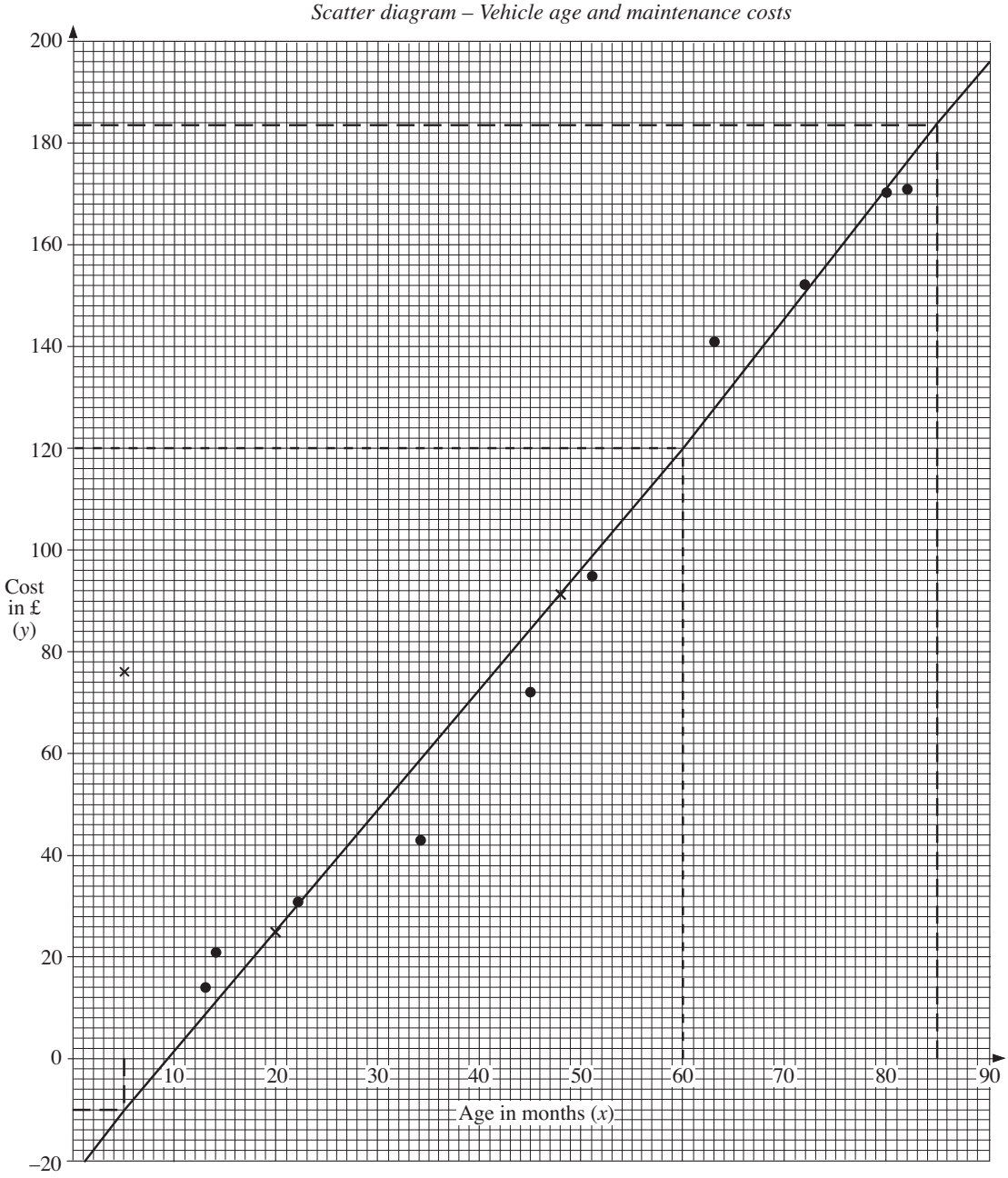
(b) $\bar{x} = 48$, $\bar{y} = 91$

(c) See diagram on the following page.

(d) (i) £184 (ii) −£10 (iii) £120

(e) £120 (in middle of data range), £185 (just outside data range), −£10 (makes no sense!)

Scatter diagram – Vehicle age and maintenance costs



Worked Example 3

An electric heater was switched on in a cold room and the temperature of the room was taken at 5 minute intervals. The results were recorded and plotted on the following graph.

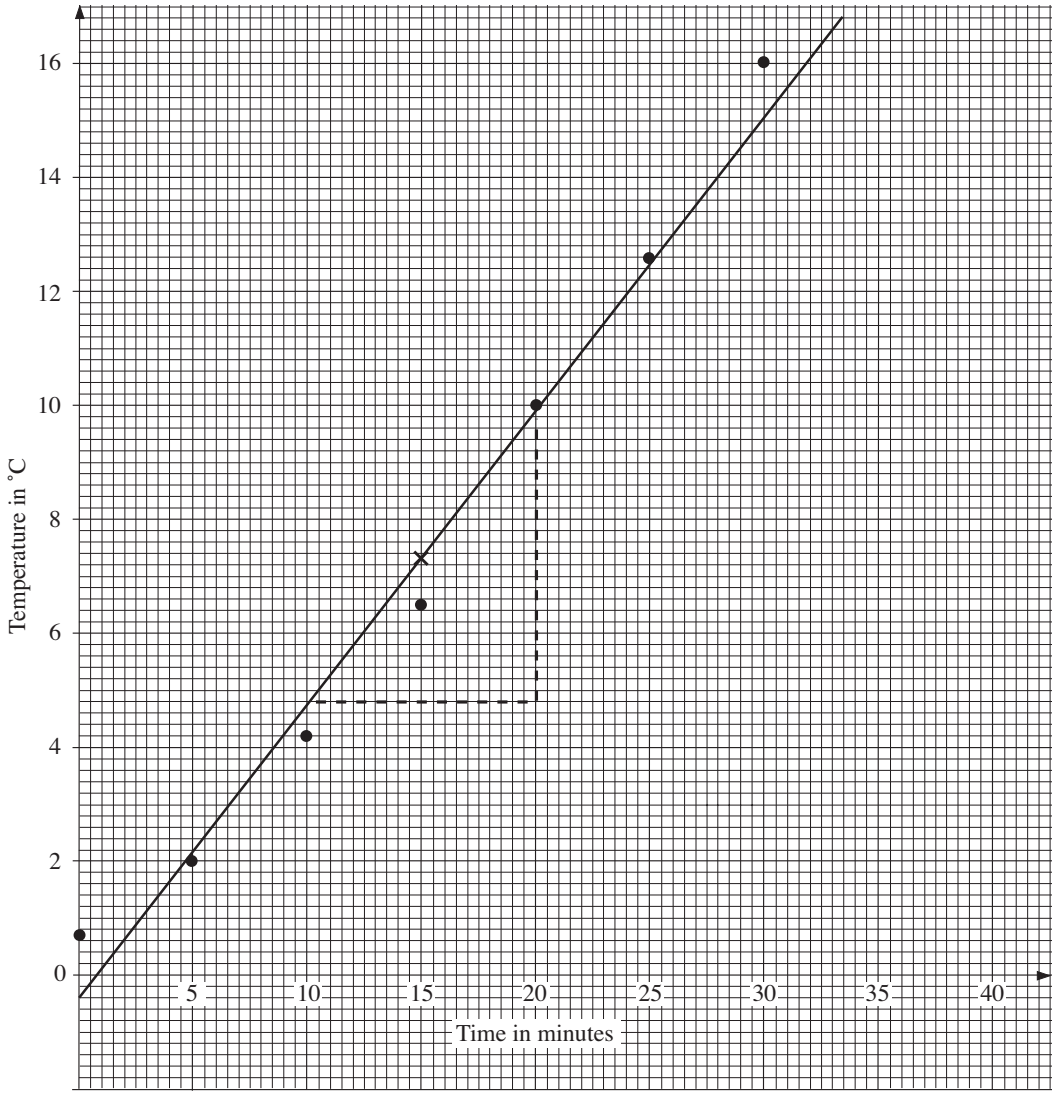
(a) Given that $\bar{x} = 15$ and $\bar{y} = 7.4$, draw a line of best fit for these data.

(b) Obtain the equation of this line of best fit in the form $y = mx + c$, stating clearly your values of m and c.

(c) Use your equation to predict the temperature of the room 40 minutes after switching on the fire,

(d) Give *two* reasons why this result may not be reliable.

© University of Plymouth



Solution

(a) See diagram above.

(b) Intercept $c = -0.4$ and slope, $m = \dfrac{10 - 4.8}{10}$, (see triangle drawn on graph)

$$\text{i.e. } m = 0.52,$$

so $$\boxed{y = 0.52x - 0.4}$$

An alternative approach would be to note the intercept $c = -0.4$ from the diagram, so that

$$y = mx - 0.4$$

To pass through the point $\bar{x} = 15,\ \bar{y} = 7.4$ means that

$$7.4 = 15m - 0.4$$

$$15m = 7.4 + 0.4$$

$$m = \frac{7.8}{15} \approx 0.52$$

© University of Plymouth

giving the equation

$$y = 0.52x - 0.4$$

(c) Predicted temperature $= 0.52 \times 40 - 0.4$

$= 20.4\ ^\circ\text{C}$

(d) The value of 40 minutes is outside the range of values on which the line of regression is based; the heater may not continue to increase if it has a thermostat on it.

Worked Example 4

Mr Bean often travels by taxi and has to keep details of the journeys in order to complete his claim form at the end of the week. Details for journeys made during a week are:

Distance travelled (miles)	2	7	$8\frac{1}{2}$	11	6	3	$4\frac{1}{2}$
Cost (£)	3.00	5.40	6.10	7.40	5.00	3.20	4.20

(a) On graph paper, plot the above points.

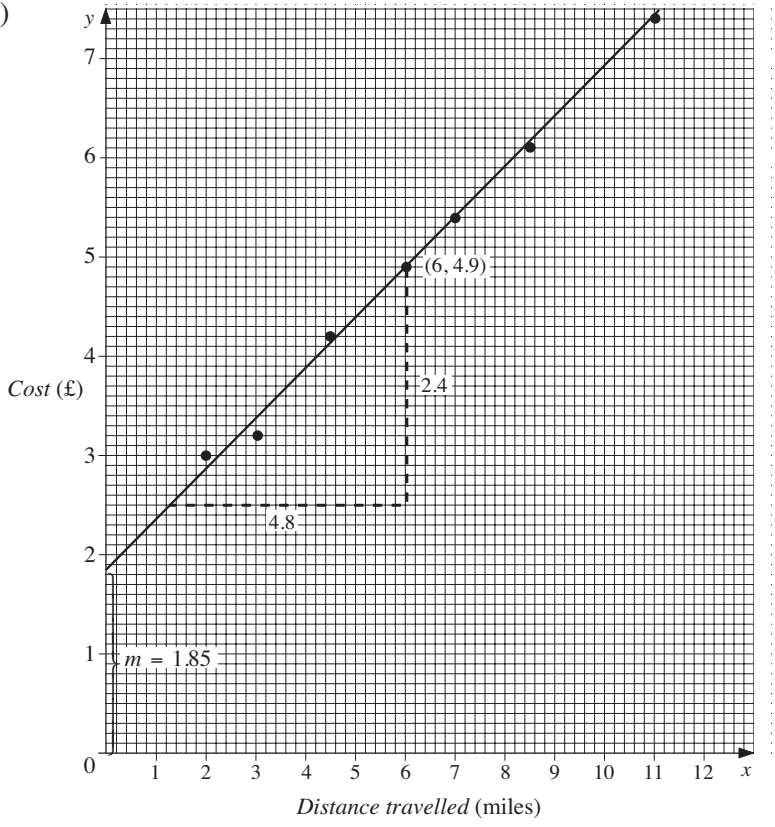
(b) Calculate the mean point of these data and use this line to draw the line of best fit on your graph.

(c) Obtain the equation of your line of best fit in the form $y = mx + c$.

(d) Give an interpretation for the value of c in your calculation.

Solution

(a)



(b) distance, $\bar{x} = 6$; cost, $\bar{y} = 4.9$

(c) From the intercept, $c = 1.85$, and from the construction, the gradient $= \dfrac{2.4}{4.8} = 0.5$

Thus

$$y = 0.50\,x + 1.85$$

(d) The intercept is the value of y when $x = 0$, i.e. the cost involved when the taxi is not being used.

Exercises

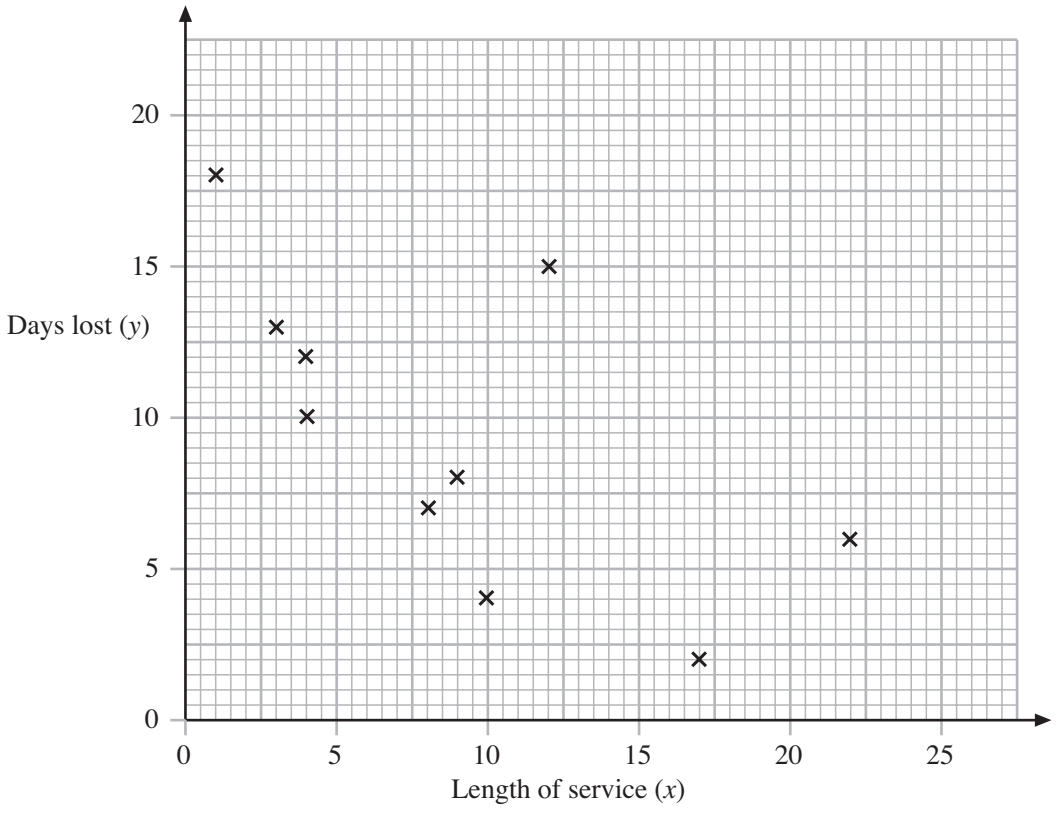
1. Management at a building company are worried about the number of working days lost, by employees, due to illness.

   At the start 2011 they select a random sample of ten employees.

   For each employee they record the length of service (in years) with the company and the number of working days lost due to illness during 2100.

   The results are summarised in the table and shown on the scatter diagram.

Employee	A	B	C	D	E	F	G	H	I	J
Length of service (x)	9	3	4	10	8	12	17	1	22	4
Days lost (y)	8	13	10	4	7	15	2	18	6	12



(a) The line of best fit for the data

- has a gradient of -0.5
- passes through the point $(9, 9.5)$

Draw this line on the scatter diagram.

(b) (i) Write down the equation for the line of best fit in the form $y = mx + c$.

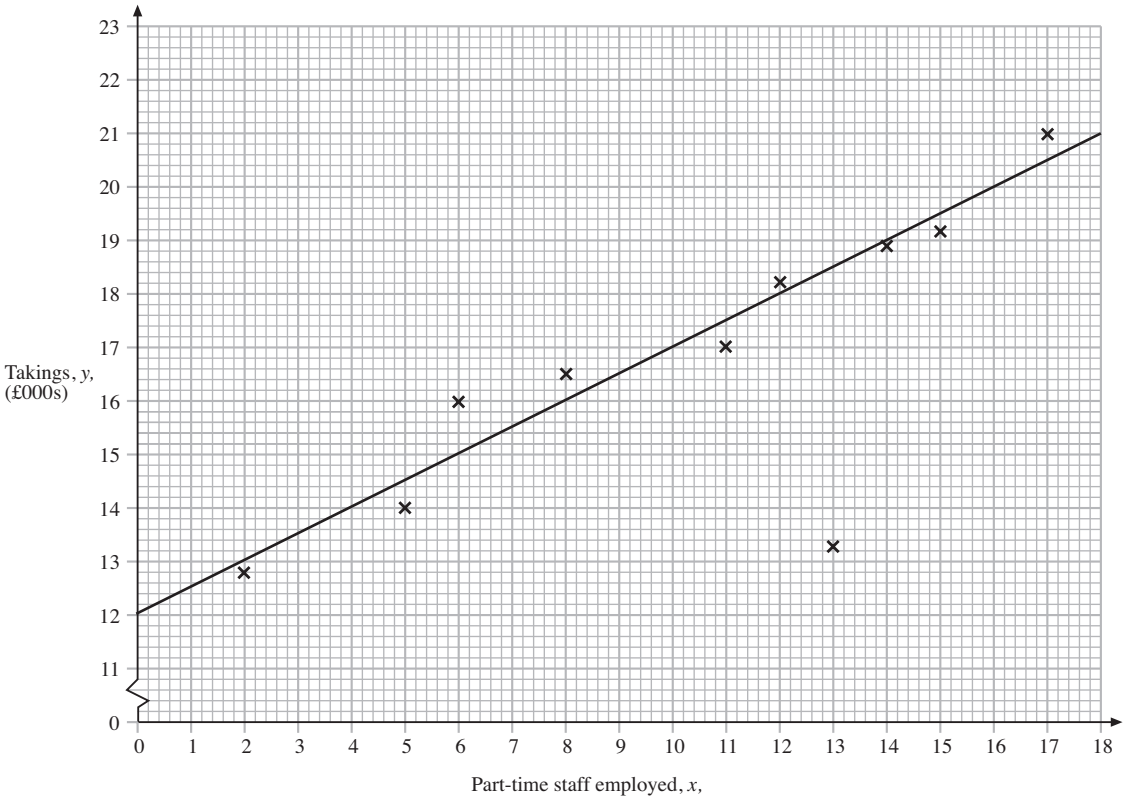
(ii) Explain why no sensible interpretation can be given for your value of c in the equation.

(c) Estimate the number of days lost due to illness, during 2011, for an employee having 20 years' service with the company.

(d) Give **two** reasons why the equation found in part (b) can **not** be used to estimate the number of days lost due to illness, for an employee having 30 years' service with the company.

2. A large department store employs additional part-time sales staff on Saturdays. The manager wants to know whether the number of part-time staff employed increases daily takings.

The scatter diagram below shows the number of part-time staff employed and takings over successive Saturdays.



© University of Plymouth

(a) On one of the Saturdays the store's car park was closed to customers.

(i) How many part-time staff do you think were employed on that Saturday?

(ii) Justify your choice.

(b) A line of best fit has been drawn on the scatter diagram.

Work out the equation of this line in the form $y = mx + c$,

(c) The manager comments that

"*Increasing the number of part-time staff will result in an increase in Saturday takings*."

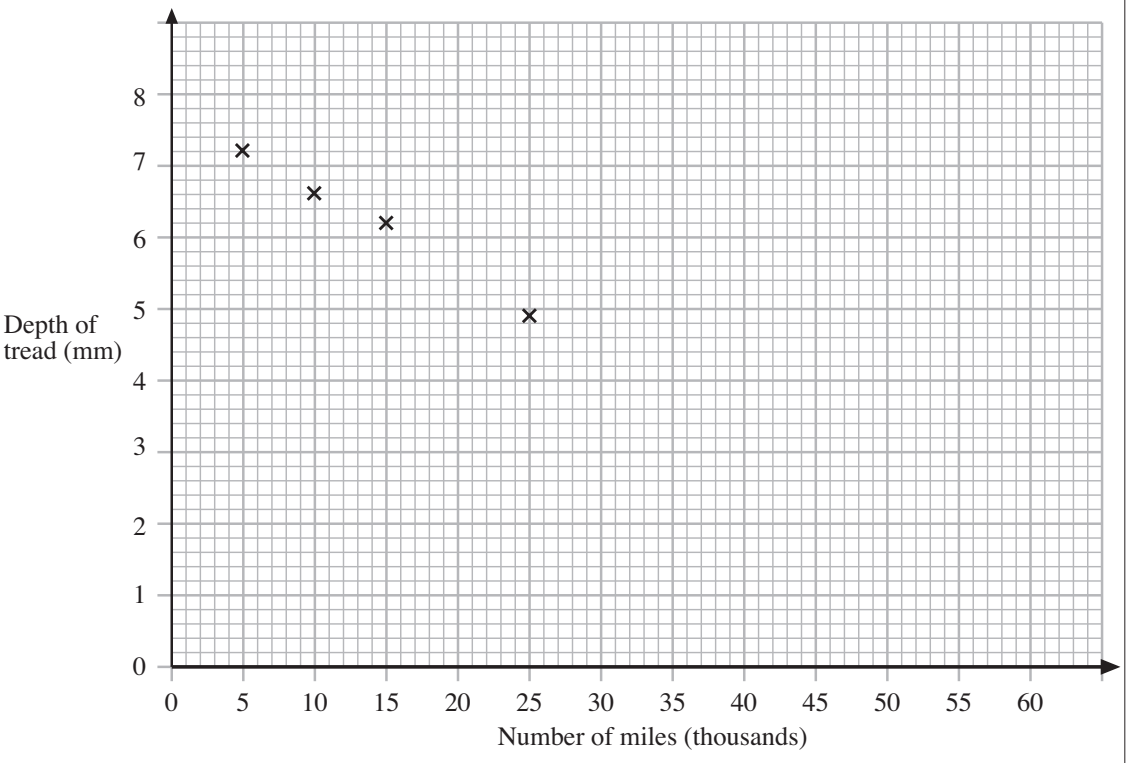
State **two** reasons why this may **not** be a valid conclusion to draw based on the data provided.

3. The table shows the number of miles travelled (thousands) and depth of tread (mm) on eight tyres of the same type.

Number of miles (thousands)	5	10	15	25	31	36	40	46
Depth of tread (mm)	7.2	6.6	6.2	4.9	4.8	3.8	3.3	2.4

(a) Complete the scatter diagram for the data.

The first four points have been plotted for you.



(b) For the data, the mean number of miles is 26 thousand.

(i) Work out the mean depth of tread.

(ii) Use these mean values to help you draw a line of best fit on the scatter diagram.

(c) Use your line of best fit to estimate the depth of tread for a tyre that had travelled 20 thousand miles.

(d) It is illegal to have less than 1.6 mm of tread on a tyre.

Use your line to estimate the number of miles travelled before a tyre becomes illegal.

(e) Which of your answers, 3(c) or 3(d), do you think is **more** reliable?

Give a reason for your choice.

(f) Is there likely to be a **causal** relationship between the number of miles travelled and the depth of tread?

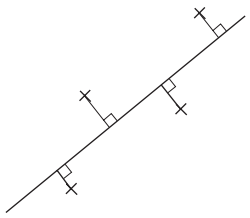
Give a reason for your answer.

5 Line of Regression Equation

In the previous section, we used lines of regression by drawing these as accurately, by eye, as possible.

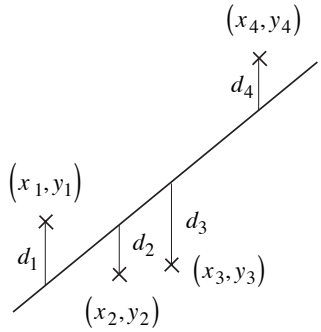
Here we will extend this and show how to use a formula to find the accurate line of regression.

If you have a set of paired data points (x_i, x_i) for which there is strong correlation (positive or negative), the problem is to find the line which best fits the data.



It may seem natural to try to find the line so that the points' distances from it have as small a total as possible, However, since the line will need to produce values of *y* for given values of *x* (or vice versa) it is more sensible to seek to produce a line so that any distances in the *y* direction, and therefore any errors in preparing *y* given *x,* should be a minimum.

If the line is to be used to predict values of *y* based on known values of *x* it is called the '*y* on *x*' line and its equation is determined by making $d_1^2 + d_2^2 + ... = \sum d^2$ as small as possible. The equation of this line can be shown to be



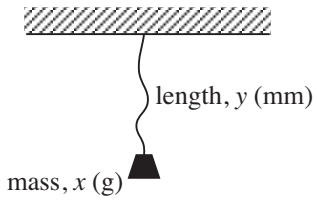
$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

and for this line it can also be shown that $\sum d^2 = n s_y^2 (1 - r^2)$. You will notice that when $r = \pm 1$, (i.e. the points lie exactly on a straight line), then $\sum d^2 = 0$ as would be expected. The procedure used to obtain the equation is called the **method of least squares** and the '*d*'s are often referred to as **residuals**. The gradient is called the **regression coefficient**.

© University of Plymouth

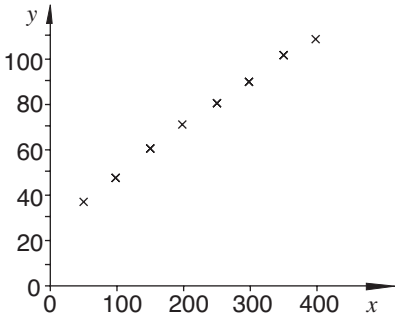
Worked Example 1

Data obtained from an experiment in which different weights were attached to an elastic band and the length of the band measured, are given in the following table.



Mass g (x)	50	100	150	200	250	300	350	400
Length mm (y)	37	48	60	71	80	90	102	109

In the diagram opposite, the points lie very close to a straight line and the value of r is 0.999.



Solution

For this data,

$$\bar{x} = \frac{1800}{8} = 225, \quad \bar{y} = \frac{597}{8} = 74.625$$

$$s_{xy} = \frac{156150}{8} - 225 \times 74.625 = 2728.125$$

$$s_x^2 = \frac{510000}{8} - 225^2 = 13125$$

$$\Rightarrow \quad y - 74.625 = \frac{2728.125}{13125}(x - 225)$$

$$\Rightarrow \quad \boxed{y = 0.208x + 27.857}$$

The values of 0.208 and 27.857 represent the gradient of the line and its intercept on the y-axis and are available directly from a calculator with LR mode. The gradient has units mm/g and tells us how much extension would be caused by the addition of 1 extra gram to the suspended mass. This line can now be used to find values of y given values of x.

Worked Example 2

Using the equation found in Worked Example 1, what length would you expect the elastic band to be if a weight of

(a) 375 g (b) 1 kg

was attached to it?

Solution

$y = 0.208x + 27.857$

(a) When $x = 375$ g,

$$y = 0.208 \times 375 + 27.857$$

$$\approx 106 \text{ mm}$$

© University of Plymouth

(b) When $x = 1$ kg $= 1000$ g,

$$y = 0.208 \times 1000 + 27.857$$

$$\approx 236 \text{ mm}$$

Exercises

1. A student counted the number of words in an essay she had written, recording the total every 10 lines.

No. of lines (x)	10	20	30	40	50	60	70	80
No. of words (y)	75	136	210	291	368	441	519	588

Find the formula to convert lines to words. How many words (approximately) has she written if she writes

(a) 65 lines (b) 100 lines (c) 1000 lines?

Are you happy with all these estimates?

2. Eight test areas were given different concentrations of a new fertiliser and the resulting crop was weighed.

Concentration g/L (x)	1	2	3	4	5	6	7	8
Weight of crop kg(y)	7	11.1	14	16.2	20	23.9	27	29

(a) (i) Draw a scatter diagram to show the data.

(ii) Calculate the equation of the regression line y on x and show it on your diagram.

(b) What increase in weight of crop might be expected from raising the concentration of fertiliser by 1 g/L?

3. In an investigation into prediction using the stars and planets a celebrated astrologist, Horace Cope, predicted the ages at which thirteen young people would first marry. The complete data, of predicted and actual ages at first marriage, are now available and are summarised in the table.

Person	A	B	C	D	E	F	G	H	I	J	K	L	M
Predicted age (x years)	24	30	28	36	20	22	31	28	21	29	40	25	27
Actual age y (years)	23	31	28	35	20	25	45	30	22	27	40	27	26

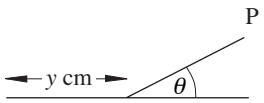
(a) Draw a scatter diagram of these data.

(b) Calculate the equation of the regression line of y on x and draw this line on the scatter diagram.

(c) Comment upon the results obtained, particularly in view of the data for person G. What further action would you suggest?

© University of Plymouth

4. The experimental data below were obtained by measuring the horizontal distance y cm, rolled by an object released from the point P on a plane inclined at $\theta°$ to the horizontal, as shown in the diagram.



Distance y	Angle $\theta°$
44	8.0
132	25.0
152	31.5
87	17.5
104	20.0
91	10.5
142	28.5
76	14.5
$\Sigma y = 828$,	$\Sigma y\theta = 18147$
$\Sigma \theta = 155.5$,	$\Sigma \theta^2 = 3520.25$

(a) Illustrate the data by a scatter diagram.

(b) Calculate the equation of the regression line of distance on angle and draw this line on the scatter diagram.

(c) It later emerged that one of the points was obtained using a different object. Suggest which point this was.

(d) Estimate the distance the original object would roll if released at an angle of (i) $12°$, (ii) $40°$.

Discuss the uncertainty of each of these estimates.

5. The variables H and T are known to be linearly related. Fifty pairs of experimental observations of the two variables gave the following results:

$\Sigma H = 83.4$, $\Sigma T = 402.0$,

$\Sigma HT = 680.2$, $\Sigma H^2 = 384.6$,

$\Sigma T^2 = 3238.2$

Obtain the regression equation from which one can estimate H when T has the value 7.8 and give, to 1 decimal place, the value of this estimate.

© University of Plymouth