12 CORRELATION AND REGRESSION

Objectives

After studying this chapter you should

- be able to investigate the strength and direction of a relationship between two variables by collecting measurements and using suitable statistical analysis;
- be able to evaluate and interpret the product moment correlation coefficient and Spearman's correlation coefficient;
- be able to find the equations of regression lines and use them where appropriate.

12.0 Introduction

Is a child's height at two years old related to her later adult height? Is it true that people aged over twenty have slower reaction times than those under twenty? Does a connection exist between a person's weight and the size of his feet?

In this chapter you will see how to quantify answers to questions of the type above, based on observed data.

12.1 Ideas for data collection

Undertake at least one of the three activities below. You will need your data for further analysis later in this chapter.

Activity 1

Collect a random sample of twenty stones. For each stone measure its

(i) maximum dimension
(ii) minimum dimension
(iii) weight.

Does there appear to be a connection between (i) and (ii), (i) and (iii), or (ii) and (iii)?

Activity 2

Measure the heights and weights of a random sample of 15 students of the same sex. Is there any apparent relationship between the two variables?

Would you expect the same relationship (if any) to exist between the heights and weights of the opposite sex?

Activity 3

Collect a dozen volunteers and time them running a forty metre straight sprint. Ask them to do two long jumps each and record the better one. (Measure the jump from the point of take-off rather than any board.)

Is there a connection between the times and distances recorded?

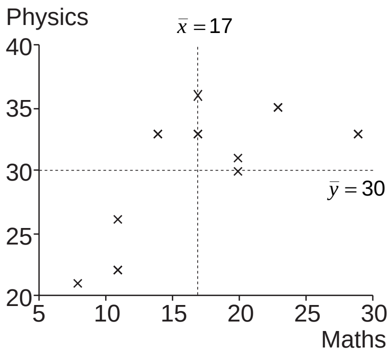
12.2 Studying results

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?

A starting point would be to 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.

Calculating the means,

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

and

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

and using them to divide the graph into four shows this clearly.

The problem is to find a way to measure how strong this tendency is.

Covariance

An attempt to quantify the tendency to go from bottom left to top right is to evaluate the expression

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

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

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

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 $(x - \bar{x})(y - \bar{y})$ 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.

$$\begin{aligned}\frac{1}{n}\Sigma(x - \bar{x})(y - \bar{y}) &= \frac{1}{n}\Sigma(xy - \bar{x}y - x\bar{y} + \overline{xy}) \\ &= \frac{1}{n}(\Sigma xy - \Sigma \bar{x}y - \Sigma\, x\bar{y} + \Sigma \overline{xy})\end{aligned}$$

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

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

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

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 216, this form of the expression is usually used when calculating covariance.

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

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

$$= 21.3$$

($\sum xy$ 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.

Example

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.

Activity 4

Find the covariance for the data you collected in any of the first three activities.

12.3 Pearson's product moment correlation coefficient

Dividing $(x - \bar{x})$ by the standard deviation s_x gives the distance of each x value above or below the mean as so many standard deviations. For the example on height and weight above, the standard deviations in m and cm are related, with the second being one hundred times the first, so

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

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

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

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 example in Section 12.2:

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

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

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

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

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

$$=0.71$$

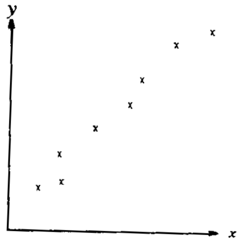
The value of r gives a measure of how close the points are to lying on a straight line. It is always true that

$$\boxed{-1\le r\le 1}$$

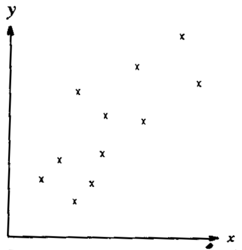
and $r=1$ indicates that all the points lie exactly on a straight line with positive gradient, while $r=-1$ gives the same information with a line having negative gradient, and $r=0$ tells us that there is no connection at all between the two sets of data.

The sketches opposite indicate these and in between cases.

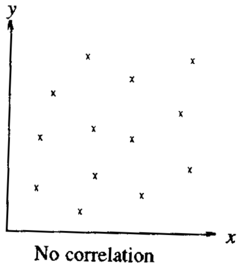
(Note that s_{xy} is not a calculator key, but its value may be checked by $r \times s_x \times s_y$ which are all available.)



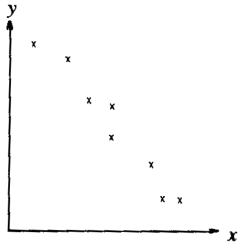
Strong positive correlation



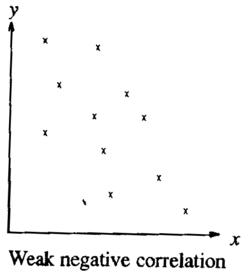
Weak positive correlation



No correlation



Strong negative correlation



Weak negative correlation

The significance of r

With only **two** pairs of values it is unlikely that they will lie on the same horizontal or vertical line, giving a correlation coefficient of zero but any other arrangement will produce a value of r equal to plus or minus one, depending on whether the line through them has a positive or negative gradient. With **six** points, however, the fact that they lie on, or close to, a straight line becomes much more significant.

The following table, showing critical values at 5% significance level, gives some indication of how likely some values of the correlation coefficient are. For example, for $n = 5$, $r = 0.878$ means that there is only a 5% chance of getting a result of 0.878 or greater if there is **no** correlation between the variables. Such a value, therefore, indicates the likely existence of a relationship between the variables.

(no.of pairs) n	r
3	0.997
4	0.950
5	0.878
6	0.811
7	0.755
8	0.707
9	0.666
10	0.632

More detailed tables of critical values are available for a range of significant levels and values of n. Their calculation relies on the data being drawn from joint normal distributions, so using them in other circumstances cannot provide an accurate assessment of significance.

Example

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

(a) Use the formula

$$s_{xy} = \frac{1}{n}\sum xy - \overline{xy}$$

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

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

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

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

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

This indicates weak negative correlation. But to apply a significance test, the null and alternative hypotheses need to be defined:

$$H_0 : r = 0$$
$$H_1 : r \neq 0$$

significance level : 5% (two tailed).

Using the table of critical values in the Appendix, for $n = 12$,

$$r_{crit} = \pm\, 0.576$$

That is, the critical region where H_0 is rejected is $r < -\,0.576$ and $r > 0.576$.

Since $r = -0.481$, there is insufficient evidence to reject the null hypothesis.

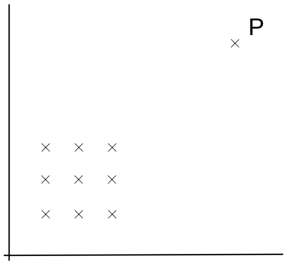
Limitations of correlation

You should note that

(1) r is a measure of **linear** relationship only. There may be an exact connection between the two variables but if it is not a straight line r is no help. It is well worth studying the scatter diagram carefully to see if a non-linear relationship may exist. Perhaps studying x and $\ln y$ may provide an answer but this is only one possibility.

(2) Correlation does not imply **causality**. A survey of pupils in a primary school may well show that there is a strong correlation between those with the biggest left feet and those who are best at mental arithmetic. However it is unlikely that a policy of 'left foot stretching' will lead to improved scores. It is possible that the oldest children have the biggest left feet and are also best at mental arithmetic.

(3) An unusual or freak result may have a strong effect on the value of r. What value of r would you expect if point P were omitted in the scatter diagram opposite?



Exercise 12A

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

(iii)

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

(ii)

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

(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. 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

Draw a scatter diagram to show the data.

Calculate the values of $\bar{x}$ and $\bar{y}$ and show these as vertical and horizontal lines. Which three points are the odd ones out?

Find the values of s_x, s_y and r.

4. 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

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

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

x	13	14	15
y	400	403	405

Add these points to your graph and describe what they show.

5. 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

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

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.

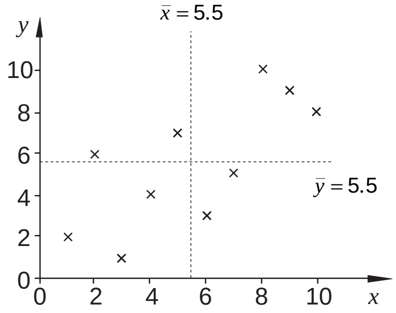
12.4 Spearman's rank correlation coefficient

Two judges at a fete placed the ten entries for the 'best fruit cakes' competition in order as follows (1 denotes first, etc.)

Entry	A	B	C	D	E	F	G	H	I	J
Judge 1 (x)	2	9	1	3	10	4	6	8	5	7
Judge 2 (y)	6	9	2	1	8	4	3	10	7	5

No actual marks like 73/100 have been awarded in this case where only ranks exist.

Is there a linear relationship between the rankings produced by the two judges?



Spearman's rank correlation coefficient answers this question by simply using the ranks as data and in the product moment coerrelation coefficient, *r*, and denoting it r_s. Again a scatter diagram may be drawn and the presence of the points plotted in, or very near, the top right and bottom left areas indicates a positive correlation.

Spearman's rank correlation coefficient,

$$r_s = \frac{\frac{1}{10}\Sigma xy - \bar{x}\bar{y}}{s_x\, s_y}$$

where $\bar{x} = \bar{y} = \dfrac{55}{10} = 5.5$

$$s_x = s_y = \sqrt{\frac{385}{10} - 5.5^2} = \sqrt{8.25}$$

and $\Sigma xy = 2\times 6 + 9\times 9 + \ \dots\ + 7\times 5 \ = 362$

$$\Rightarrow \qquad r_s = \frac{\frac{1}{10}\times 362 - 5.5^2}{\sqrt{8.25}\sqrt{8.25}}$$

$$= \frac{36.2 - 30.25}{8.25}$$

$$= 0.721$$

(The significance tables for *r* should certainly not be used here as the ranks definitely do not come from normal distributions.)

It can be shown that, when there are no tied ranks,

$$\frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{s_x s_y} = 1 - \frac{6\Sigma d^2}{n\left(n^2 - 1\right)}$$

and so $$\boxed{r_s = 1 - \frac{6\Sigma d^2}{n\left(n^2 - 1\right)}}$$

where $d = x - y$, is the difference in ranking.

For the example just considered

Entry	A	B	C	D	E	F	G	H	I	J
Judge 1	2	9	1	3	10	4	6	8	5	7
Judge 2	6	9	2	1	8	4	3	10	7	5
$\lvert d \rvert$	4	0	1	2	2	0	3	2	2	2
d^2	16	0	1	4	4	0	9	4	4	4

So $\quad \Sigma d^2 = 16 + 0 + 1 + \ \ldots \ + 4 = 46$

$$\Rightarrow \quad r_s = 1 - \frac{6 \times 46}{10(100 - 1)}$$

$$= 1 - \frac{6 \times 46}{10 \times 99}$$

$$= \frac{119}{165}$$

≈ 0.721 to 3 decimal places.

As with the product moment correlation coefficient, Spearman's correlation coefficient also obeys

$$-1 \le r_s \le 1$$

where $r = 1$ corresponds to perfect positive correlation and $r = -1$ to perfect negative correlation.

The definition of the formula from the product moment correlation coefficient will not be given here but you will see in the following Activity how it can be deduced.

Activity 5

You can verify Spearman's formula by first assuming that

$$r_s = 1 - K \sum d^2$$

where K is a constant for each value of n .

(a) Show that $r_s = 1$ for perfect positive correlation.

(b) Use the fact that $r_s = -1$ for perfect negative correlation to complete the table below.

n	2	3	4	5	6	7	8
K							

(For example, for $n = 4$, perfect negative correlation corresponds to

1	2	3	4
4	3	2	1)

Check that these values agree with the Spearman's formula, that is

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

Significance of r_s

If the tables of significance for r cannot be used here, you can still assess the importance of the value by noting that the formula

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

contains the term Σd^2. Tables giving the critical values of r_s for various values of n are available.

So at 5% significance level, the hypotheses are defined by

$$H_0 : r_s = 0$$

$$H_1 : r_s \neq 0 \qquad \text{(two tailed)}$$

and, with $n = 10$, the tables show that

$$p\left(\left|r_s\right| > 0.6485\right) = 0.05$$

Note: $\left|r_s\right| > 0.6485$ means $r_s < -0.6485$ or $r > 0.6485$.

So for a two tailed test, you should reject H_0 since in the example on page 226, $r_s = 0.721$, and accept H_1, the alternative hypothesis, which says that there is significant correlation.

You can test for positive correlation, by using the hypothesis

$$H_0 : r_s = 0$$

$$H_1 : r_s > 0 \qquad \text{(one tailed)}$$

At 5% level, and with $n = 10$ as before,

$$p\left(r_s > 0.5636\right) = 0.05$$

and since 0.721 > 0.5636, again H_0 is rejected. You accept the alternative hypothesis that there is significant positive correlation.

Tied ranks

The formula $r_s = 1 - \frac{6\Sigma d^2}{n\left(n^2 - 1\right)}$ does not give the correct value for r_s when there are **tied** ranks, but as long as you do not have too many ties, the inaccuracies are negligible and the use of this equation allows the table of significance for Σd^2 to be employed.

Example

Find the value of n for the following data

Ranks x	1	2 =	2 =	5	4	6	7	8
Ranks y	1	3	4	2	5	6 =	6 =	6 =

Solution

Those tied in the x rankings are given a value of $\frac{2+3}{2} = 2\frac{1}{2}$ and those tied in y are allocated $\frac{6+7+8}{3} = 7$. (In general, each tie is given the mean of the places that would have been occupied if a strict order had been produced.) The table, therefore, becomes

Ranks x	1	$2\frac{1}{2}$	$2\frac{1}{2}$	5	4	6	7	8
Ranks y	1	3	4	2	5	7	7	7
$\lvert d \rvert$	0	$\frac{1}{2}$	$1\frac{1}{2}$	3	1	1	0	1
d^2	0	$\frac{1}{4}$	$2\frac{1}{4}$	9	1	1	0	1

Hence $\Sigma d^2 = 14.5$

and
$$r_s = 1 - \frac{6 \times 14.5}{8(64-1)}$$
$$= 0.827$$

and again this is a significant result. That is, you would conclude that there is positive correlation.

Exercise 12B

1.

Item	A	B	C	D	E	F
Ranks (x)	1 =	1 =	1 =	4 =	4 =	4 =
Ranks (y)	1	2	3 =	3 =	5	6

Use both

(i) $r = 1 - \dfrac{6\Sigma d^2}{n\left(n^2-1\right)}$ and (ii) $r = \dfrac{\frac{1}{n}\Sigma xy - \overline{x}\overline{y}}{s_x s_y}$

to evaluate rank correlation coefficients for the two sets of rankings given.

Comment on your results.

2. The performances of the six fastest male sprinters in a school were noted in their winter cross-country race. The details are shown in the table.

	A	B	C	D	E	F
Sprint ranking	1	2	3	4	5	6
Position in cross-country	70	31	4	32	12	17

Give each athlete a rank for cross-country and evaluate r_s. Comment on the significance of your result.

3. At an agricultural show 10 Shetland sheep were ranked by a qualified judge and by a trainee judge. Their rankings are shown in the table.

Qualified Judge	1	2	3	4	5	6	7	8	9	10
Trainee Judge	1	2	5	6	7	8	10	4	3	9

Calculate a rank correlation coefficient for these data. Is this result significant at the 5% level?

4. Five sacks of coal, A, B, C, D and E have different weights, with A being heavier than B, B being heavier than C, and so on. A weight lifter ranks the sacks (heaviest first) in the order A, D, B, E, C. Calculate a coefficient of rank correlation between the weight lifter's ranking and the true ranking of the weights of the sacks.

5. A company is to replace its fleet of cars. Eight possible models are considered and the transport manager is asked to rank them, from 1 to 8, in order of preference. A saleswoman is asked to use each type of car for a week and grade them according to their suitability for the job (A - very suitable to E - unsuitable). The price is also recorded.

Model	Transport manager's ranking	Saleswoman's grade	Price (£10's)
S	5	B	611
T	1	B+	811
U	7	D-	591
V	2	C	792
W	8	B+	520
X	6	D	573
Y	4	C+	683
Z	3	A-	716

(a) Calculate Spearman's rank correlation coefficient between

(i) price and transport manager's rankings,

(ii) price and saleswoman's grades.

(b) Based on the results in (a) state, giving a reason, whether it would be necessary to use all three different methods of assessing the cars.

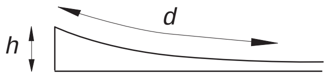
Ideas for data collection

Activity 6

Use a strong elastic band or spring as a simple weighing machine. Carefully hang weights from it and record its length for each one.

Activity 7

Mount two metres of toy railway track on flexible board. Raise one end and record the distance travelled by a railway truck for each different height.



Activity 8

Run water into a container on a set of scales. The water should flow in at as steady a rate as possible from just above the level of the container. Record the time taken for the scales to show different masses. e.g.

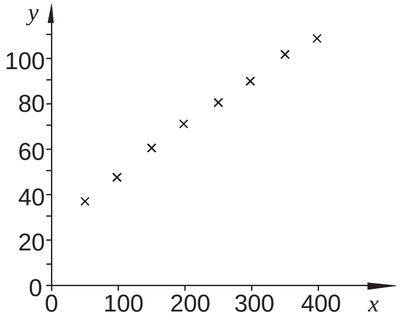
Mass (g) x	200	250	300	350	400
Time (secs) y					

12.5 Linear regression

In linear regression you start by looking at a set of points to see if there is a relationship between them and if there is you proceed to establish it in such a way that further points may be deduced from it with the minimum possible error. That is, start with points, proceed to a line and regress to points again.

Here are some results for the elastic band experiment suggested in Activity 6.

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.

Activity 9

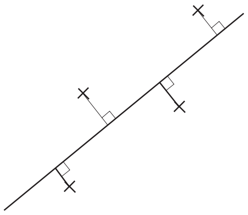
Find the value of

(a) r, the product moment correlation coefficient.

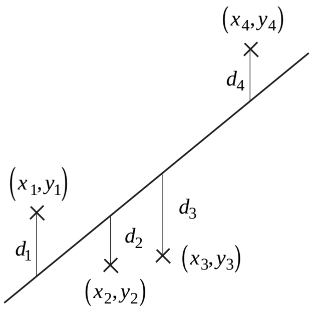
(b) r_s, Spearman's rank correlation coefficient.

Comment on their values.

Having decided that the points follow a straight line, with some small variations due to errors in measurement, changes in the environment etc, 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 predicting 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 $d_1^{\,2}+d_2^{\,2}+\ \dots\ =\Sigma d^2$ as small as possible. The equation of this line can be shown to be

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

and for this line $\Sigma d^2 = n s_y^2\left(1-r^2\right)$. You will notice that when $r=\pm 1$ (i.e. the points lie exactly on a straight line) then $\Sigma d^2=0$ 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 the **residuals**. The gradient is called the **regression coefficient**.

For the elastic band example,

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

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

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

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

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

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.

Example

What length would you expect the elastic band to be if a weight of

(a) 375 g (b) 1 kg

was suspended by it?

Solution

(a) $\hat{y} = 0.208 \times 375 + 27.857$

$= 105.9 \text{ mm}$

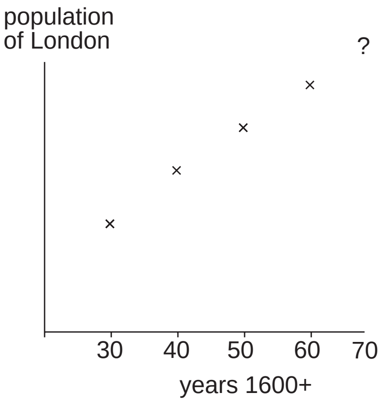
(The ^ above the y indicates that this is an estimate, however accurate. Calculators with LR mode usually have a $\hat{y}$ function giving the answer directly.)

(b) $\hat{y} = 0.208 \times 1000 + 27.857$

$= 235.9$

The first of these answers is an example of **interpolating**, (that is 'putting between' known values) and is quite trustworthy. The latter, though, is a case of **extrapolating** (that is 'putting beyond' known values) and may be wildly inaccurate. The elastic may well break under the action of the 1 kg mass!

The mass x is known as the **independent** or **exploratory variable** and is controlled by the experimenter. The length y is called the **dependent** or **response variable** and is less accurate. For any fixed value of x used repeatedly the resulting readings for y will form a normal distribution.



It may be tempting to extrapolate in the example illustrated opposite, and modern day planners have to do just that, but the Plague of 1665 and the Great Fire of 1666 would be guaranteed to sabotage any attempt in this case.

Any estimates outside the range of the data are dangerous and the further away they are the less trust can be placed in them.

Estimates of x based on given values of y may be obtained from the line but since it was constructed to minimise errors in the y direction it was not designed for this use, so answers are bound to be unreliable.

Drawing the line

Looking at the equation

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

we can see that $x = \bar{x}$, $y = \bar{y}$ satisfies it so $(\bar{x}, \bar{y})$ will always be a point on the line. To find a couple more points to enable you to draw the line use the $\hat{y}$ values with the two x values at the ends of the given set of values.

So, for the elastic band example,

$$x = 50 \quad \Rightarrow \quad \hat{y} = 38.3$$

$$x = 400 \quad \Rightarrow \quad \hat{y} = 111.$$

Other forms of the equation

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

$$\Rightarrow \quad \frac{y - \bar{y}}{s_y} = \frac{s_{xy}}{s_x s_y} \left(\frac{x - \bar{x}}{s_x}\right)$$

$$\Rightarrow \quad \boxed{\frac{y - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right)}$$

Also $$\frac{s_{xy}}{{s_x}^2} = \frac{\frac{1}{n}\Sigma(x - \bar{x})(y - \bar{y})}{\frac{1}{n}\Sigma(x - \bar{x})^2}$$

$$= \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

so $$y - \bar{y} = \hat{\beta}(x - \bar{x})$$

where $$\hat{\beta} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

Exercise 12C

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

Draw a scatter diagram to show the data. Calculate the equation of the regression line y on x and show it on your diagram.

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

3. An experiment was carried out to investigate variation of solubility of chemical X in water. The quantities in kg that dissolved in 1 litre at various temperatures are show in the table.

Temp.°C (y)	15	20	25	30	35	50	70
Mass of X (x)	2.1	2.6	2.9	3.3	4.0	5.1	7.0

Draw a scatter diagram to show the data. Calculate the equation of the regression line of y on x. Draw the line and plot the point $(\bar{x},\bar{y})$ on your diagram. What quantity might be expected to dissolve at 42° C? Find the quantity that your equation indicates would dissolve at $-10°$ C and comment on your answer.

Calculate the sum of the squares of the residuals and comment on your result.

12.6 Bivariate distributions

In many situations it may not be possible to control either variable.

Example

In a decathlon held over two days the following performances were recorded in the high jump and long jump. All distances are in metres.

Competitor	A	B	C	D	E	F	G
High jump x	1.90	1.85	1.96	1.88	1.88	Abs	1.92
Long jump y	6.22	6.24	6.50	6.36	6.32	6.44	Abs

What performances might have been expected from F in the high jump and G in the long jump if they had competed?

Solution

To estimate G's performance in the long jump we use the y on x line.

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

and using competitors A to E,

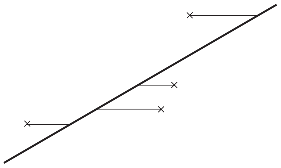
$$\bar{y}=\frac{31.64}{5}=6.328\,,\qquad \bar{x}=\frac{9.47}{5}=1.894$$

Also $$s_{xy}=\frac{1}{5}\times 59.9404-6.328\times 1.894=0.002848$$

$$s_x{}^2=\frac{1}{5}\times 17.9429-1.894^2=0.001344$$

$$\Rightarrow\quad y-6.328=\frac{0.002848}{0.001344}\quad(x-1.894)$$

$$\Rightarrow\quad y=2.119x+2.315$$



Thus $x = 1.92$ gives $\hat{y} = 2.119 \times 1.92 + 2.315 = 6.38\text{m}$

Now to estimate F's high jump accurately we need a line for which the sum of the horizontal distances from it is a minimum.

This is the *x* on *y* line and its equation is

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

$$s_y{}^2 = \frac{1}{5} \times 200.268 - 6.328^2 = 0.010016$$

$$\Rightarrow \qquad x - 1.894 = \frac{0.002848}{0.010016}\ (y - 6.328)$$

$$\Rightarrow \qquad x = 0.284y + 0.095\,.$$

Now $y = 6.44 \quad \Rightarrow \quad \hat{x} = 0.284 \times 6.44 + 0.095 = 1.92\ \text{m}$

(To use all the functions available in LR mode the coordinates can be typed in with the pairs reversed)

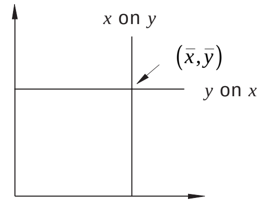
Notice that

$$x = 1.92 \qquad \Rightarrow \qquad \hat{y} = 6.38$$
$$y = 6.44 \qquad \Rightarrow \qquad \hat{x} = 1.92$$

Might we have expected $x = 1.92 \Rightarrow \hat{y} = 6.44$**?**

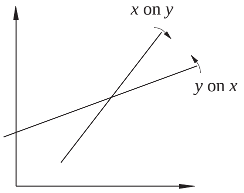
Not really as the two predictions are made from different lines.

y on *x* and *x* on *y* lines



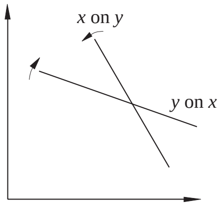
When $r = 0$ the *y* on *x* line is horizontal as can be seen from the formula

$$\frac{y - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right).$$



Similarly the *x* on *y* line is vertical as it has the form

$$\frac{x - \bar{x}}{s_x} = r\left(\frac{y - \bar{y}}{s_y}\right).$$



As *r* increases from zero the lines rotate about their point of intersection until they coincide when $r = 1$ as a line with positive gradient.

As *r* decreases from zero they turn about $(\bar{x},\bar{y})$ until they meet as a single line with negative gradient when $r = -1$.

Exercise 12D

1. 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

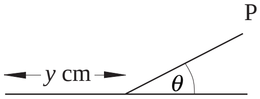
(a) Draw a scatter diagram of these data.

(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?

(AEB)

2. 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

$$\Sigma y = 828, \quad \Sigma y\theta = 18147$$

$$\Sigma \theta = 155.5, \quad \Sigma \theta^2 = 3520.25$$

(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.

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

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

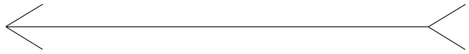
$$\Sigma HT = 680.2, \quad \Sigma H^2 = 384.6$$

$$\Sigma T^2 = 3238.2.$$

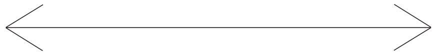
Obtain the regression equation from which one can estimate H when T has the value 7.8 and give, to 1 decimal place, the value of this estimate.

4. Students were asked to estimate the centres of the two 10 cm lines shown below.

(i)



(ii)



Their errors are shown in the following table with '–' indicating an error to the left of the centre (all in mm).

Error on (i) x	1	4	7	6	2	0	1	4
Error on (ii) y	0	1	2	2	–1	0	–1	3

Draw a scatter diagram to show the data. Calculate the equations of the regression lines y on x and x on y.

Draw both lines and plot $(\bar{x}, \bar{y})$ on your diagram.

Estimate

(a) y when $x = 5$

(b) x when $y = 1$

(AEB)

12.7 Miscellaneous Exercises

1. The yield of a batch process in the chemical industry is known to be approximately linearly related to the temperature, at least over a limited range of temératures. Two measurements of the yield are made at each of eight temperatures, within this range, with the following results:

Temperature (°C) x	180	190	200	210	220	230	240	250
Yield (tonnes) y	136.2	147.5	153.0	161.7	176.6	194.2	194.3	196.5
	136.9	145.1	155.9	167.8	164.4	183.0	175.5	219.3

$$\sum x = 1720 \qquad \sum x^2 = 374000$$

(a) Plot the data on a scatter diagram.

(b) For each temerature, calculate the mean of the two yields. Calculate the equation of the regression line of this mean yield on temperature. Draw the regression line on your scatter diagram.

(c) Predict, from the regression line, the yield of a batch at each of the following temperatures:

(i) 175 (ii) 185 (iii) 300

Discuss the amount of uncertainty in each of your three predictions.

(d) In order to improve predictions of the mean yield at various temperatures in the range 180 to 250 it is decided to take a further eight measurements of yield. Recommend, giving a reason, the temperatures at which these measurements could be carried out. (AEB)

2. Some children were asked to eat a variety of sweets and classify each one on the following scale:

strongly dislike/dislike/neutral/like/like very much.

This was then converted to a numerical scale 0, 1, 2, 3, 4 with 0 representing 'strongly dislike'. A similar method produced a score on the scale 0, 1, 2, 3 for the sweetness of each sweet assessed by each child (the sweeter the sweet the higher the score). The following frequency distribution resulted

		Liking				
		0	1	2	3	4
Sweetness	0	5	2	0	0	0
	1	3	14	16	9	0
	2	8	22	42	29	37
	3	3	4	36	58	64

(a) Calculate the product moment correlation coefficient for these data. Comment briefly on the data and on the correlation coefficient.

(b) A child was asked to rank 7 sweets according to preference and sweetness with the following results:

	Ranks						
Sweet	A	B	C	D	E	F	G
Preference	3	4	1	2	6	5	7
Sweetness	2	3	4	1	5	6	7

Calculate Spearman's rank correlation coefficient for these data.

(c) It is suggested that the product moment correlation coefficient should be calculated for (b) and Spearman's rank correlation coefficient for (a). Comment on this suggestion. (AEB)

3. A lecturer gave a group of students an assignment consisting of two questions. The following table summarises the number of numerical errors made on each question by the group of students.

	Errors on Question 1 (x)				
Errors on Question 2 (y)	0	1	2	3	4
0				4	3
1				4	5
2		5	7	5	2
3	1	4	3	4	

(a) Find the product moment correlation coefficient between x and y.

(b) Give a written interpretation of your answer.

The scores on each question for a random sample of 8 of the group are as shown below.

Student	1	2	3	4	5	6	7	8
Question 1	42	68	32	84	71	55	55	70
Question 2	39	75	43	79	83	65	62	68

(c) Calculate the Spearman rank correlation coefficient between the scores on the two questions.

(d) Give an interpretation of your result. (AEB)

4. Sets of china are individually packed to customers' requirements. The packaging manager introduces a new procedure in which each packer is responsible for all stages of an order from its initial receipt to final despatch. In order to be able to estimate the time to pack particular orders, he recorded the time taken by a particular packer to complete his first 11 sets packed by the new system. The data are in order of packing.

No. of items in set x	40	21	62	49	21	30
Time in min. to complete packaging, y	545	370	525	440	315	285

No. of items in set x	10	57	48	20	38
Time in min. to complete packaging, y	220	410	360	285	320

(a) Draw a scatter diagram of the data. Label the points from 1 to 11 according to the order of packing.

(b) Calculate the regression line of 'time' on 'number of items' and draw it on your scatter diagram. Comment on the pattern revealed and suggest why it has occurred.

(c) The regression line for the last 6 points only is $y = 188 + 3.70x$. Draw this line on your scatter diagram.

(d) The packaging manager estimated that the next order, which consisted of 44 items, would take 406.31 minutes. Comment on this estimate and the method by which you think it was made.

Make your own estimate of the packaging time for this order and explain why you think it is better than the packaging manager's. (AEB)

5. A headteacher wished to investigate the relationship between coursework marks for GCSE and marks for internal school examinations. She asked the Head of English and the Head of Science to provide some data. The Head of English reported that the marks for his four best students were as follows:

Exam mark, x	84	79	89	92
Coursework mark, y	86	85	81	91

(a) Calculate the product moment correlation coefficient for these marks.

(b) The Head of Science reported that he had asked every teacher in the school to supply him with full details of all marks. Not everyone had cooperated and some subjects used letter grades instead of marks. However, he had converted all information received into a score from 1 to 3 (the better the grade the higher the score). He produced the following frequency distribution:

		Examination score x		
		1	2	3
Course work score y	1	940	570	310
	2	630	1030	720
	3	290	480	1910

Calculate the product moment correlation coefficient for these marks.

(c) Comment on the two sets of data provided and their appropriateness to the investigation. What advice would you give the headteacher if she were to carry out a similar exercise next year? (AEB)

6. In an attempt to increase the yield (kg/h) of an industrial process a technician varies the percentage of a certain additive used, while keeping all other conditions as constant as posible. The results are shown below.

Yield y	127.6	130.2	132.7	133.6	133.9	133.8	133.3	131.9
% additive x	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0

You may assume that $\sum x = 34$ $\sum y = 1057$

$\sum xy = 4504.55$ $\sum x^2 = 155$.

(a) Draw a scatter diagram of the data.

(b) Calculate the equation of the regression line of yield on percentage additive and draw it on the scatter diagram.

The technician now varies the temperature (°C) while keeping other conditions as constant as possible and obtains the following results:

Yield y	127.6	128.7	130.4	131.2	133.6
Temperature t	70	75	80	85	90

He calculates (correctly) that the regression line is $y = 107.1 + 0.29t$.

(c) Draw a scatter diagram of these data together with the regression line.

(d) The technician reports as follows, 'The regression coefficient of yield on percentage additive is larger than that of yield on temperature, hence the most effective way of increasing the yield is to make the percentage additive as large as possible, within reason'.

Criticise the report and make your own recommendations on how to achieve the maximum yield. (AEB)

7. An instrument panel is being designed to control a complex industrial process. It will be necessary to use both hands independently to operate the panel. To help with the design it was decided to time a number of operators, each carrying out the same task once with the left hand and once with the right hand.

The times, in seconds, were as follows:

Operator	A	B	C	D	E	F	G	H	I	J	K
l.h.,x	49	58	63	42	27	55	39	33	72	66	50
r.h.,y	34	37	49	27	49	40	66	21	64	42	37

You may assume that

$\sum x = 554 \quad \sum x^2 = 29902 \quad \sum y = 466$

$\sum y^2 = 21682 \quad \sum xy = 24053$

(a) Plot a scatter diagram of the data.

(b) Calculate the product moment correlation coefficient between the two variables and comment on this value.

(c) Further investigation revealed that two of the operators were left handed. State, giving a reason, which you think these were. Omitting their two results, calculate Spearman's rank correlation coefficient and comment on this value.

(d) What can you say about the relationship between the times to carry out the task with left and right hands? (AEB)

8. An electric fire was switched on in a cold room and the temperature of the room was noted at intervals.

Time in minutes, from switching on the fire, x								
0	5	10	15	20	25	30	35	40
Temperature, °C, y								
0.4	1.5	3.4	5.5	7.7	9.7	11.7	13.5	15.4

You may assume that

$\sum x = 180 \quad \sum y = 68.8 \quad \sum xy = 1960$

$\sum x^2 = 5100$

(a) Plot the data on a scatter diagram.

(b) Calculate the regression line $y = a + bx$ and draw it on your scatter diagram.

(c) Predict the temperature 60 minutes from switching on the fire. Why should this prediction be treated with caution?

(d) Starting from the equation of the regression line $y = a + bx$, derive the equation of the regression line of

(i) y on t where y is temperature in °C (as above) and t is time in hours.

(ii) z on x where z is temperature in °K and x is time in minutes (as above).

(A temperature in °C is converted to °K by adding 273, e.g. $10°\text{C} \rightarrow 283\ °\text{K}$)

(e) Explain why, in (b), the line $y = a + bx$ was calculated rather than $x = a' + b'x$. If, instead of the temperature being measured at 5 minute intervals, the time for the room to reach predetermined temperatures (e.g. 1, 4, 7, 10, 13°C) had been observed what would the appropriate calculation have been? Explain your answer. (AEB)

9. The data in the following table show the length and breadth (in mm) of a group of skulls discovered during an excavation.

Length (x)	165	170	172	176	178	179	182	184	186	190
Breadth (y)	139	141	147	147	149	149	159	145	155	152

(You may assume that $\sum x^2 = 318086$, $\sum xy = 264582$ and $\sum y^2 = 220257$.)

(a) Calculate the regression lines of length on breadth and breadth on length.

(b) Plot these data on a scatter diagram and draw both your regression lines on your diagram.

(c) State, in symbolic form, the point of intersection of your two lines.

(d) Using in each case the appropriate regression line predict the breadth of a skull of length 185 mm and the length of a skull of breadth 155 mm.

(e) Under what circumstances would your two lines be coincident? (AEB)

10. A small firm negotiates an annual pay rise with each of its twelve employees. In an attempt to simplify the process it is proposed that each employee should be given a score, *x*, based on his/her level of responsibility. The annual salary will be £$(a+bx)$ and the annual negotiations will only involve the values of *a* and *b*. The following table gives last year's salaries (which were generally accepted as fair) and the proposed scores.

Employee	*x*	Annual salary (£), *y*
A	10	5750
B	55	17300
C	46	14750
D	27	8200
E	17	6350
F	12	6150
G	85	18800
H	64	14850
I	36	9900
J	40	11000
K	30	9150
L	37	10400

(You may assume that $\sum x = 459$, $\sum x^2 = 22889$, $\sum y = 132600$ and $\sum xy = 6094750$)

(a) Plot the data on a scatter diagram.

(b) Estimate values that could have been used for *a* and *b* last year by fitting the regression line $y = a + bx$ to the data. Draw the line on the scatter diagram.

(c) Comment on whether the suggested method is likely to prove reasonably satisfactory in practice.

(d) Without recalculating the regression line find the appropriate values of *a* and *b* if every employee were to receive a rise of

(i) £500 per year

(ii) 8%

(iii) 4% plus £300 per year.

(e) Two employees, B and C, had to work away from home for a large part of the year. In the light of this additional information, suggest an improvement to the model. (AEB)

11. The following data show the IQ and the score in an English test of a sample of 10 pupils taken from a mixed ability class.

The English test was marked out of 50 and the range of IQ values for the class was 80 to 140.

Pupil	A	B	C	D	E	F	G	H	I	J
IQ (*x*)	110	107	127	100	132	130	98	109	114	124
English Score (*y*)	26	31	37	20	35	34	23	38	31	36

(a) Estimate the product moment correlation coefficient for the class.

(b) What does this coefficient measure?

(c) Outline briefly how other information given in the data of the question might have affected your coefficient.

For two other groups within the class, the teacher assessed each individual in terms of scholastic aptitude and perseverance. A rating scale of 0–100 was used for each assessment and the following table summarises the ratings for one of the groups.

Scholastic aptitude	42	68	32	84	71	55	58	70
Perseverance	39	75	43	79	83	65	62	68

(d) Show that the Spearman rank correlation coefficient between the two sets of ratings for the group is 0.905.

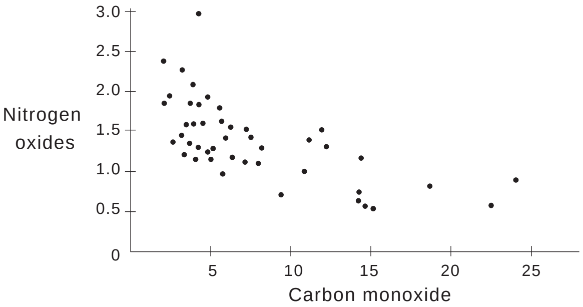
(e) The value of the Spearman rank correlation coefficient between the sets of ratings for the other group is -0.886. Interpret briefly the sign of each of these coefficients.

(f) When these two groups are combined, the value of the Spearman rank correlation coefficient is 0.66. Interpret and explain the effect of this combining on the correlation between scholastic aptitude and perseverance. (AEB)

12. (a) The product moment correlation coefficient between the random variables *W* and *X* is 0.71 and between the random variables *Y* and *Z* is -0.05.

For each of these pairs of variables, sketch a scatter diagram which might represent the results which gave the correlation coefficients.

(b) The scatter diagram on the next page shows the amounts of the pollutants, nitrogen oxides and carbon monoxide, emitted by the exhausts of 46 vehicles. Both variables are measured in grams of the pollutant per mile driven.



Write down three noticeable features of this scatter diagram.

It has been suggested that,

'If an engine is out of tune, it emits more of all the important pollutants. You can find out how badly a vehicle is polluting the air by measuring any one pollutant. If that value is acceptable, the other emissions will also be acceptable.'

State, giving your reason, whether or not this scatter diagram supports the above suggestion.

(c) When investigating the amount of heat evolved during the hardening of cement, a scientist monitored the amount of heat evolved, Y, in calories/g of cement, and four explanatory variables, X_1, X_2, X_3 and X_4. Based on thirteen observations, the scientist produced the following correlation matrix.

	Y	X_1	X_2	X_3	X_4
Y	1	0.731	0.816	−0.535	−0.821
X_1		1	0.229	r	−0.245
X_2			1	−0.139	−0.973
X_3				1	0.030
X_4					1

The values of X_1 and X_3 are as follows.

x_1	7	1	11	11	7	11	3	1	2	21	1	11	10
x_3	6	15	8	8	6	9	17	22	18	4	23	9	8

Assuming $\sum x_1^2 = 1139$ and $\sum x_3^2 = 2293$, find r, the product moment correlation coefficient between X_1 and X_3.

Write down two noticeable features of the correlation matrix. (AEB)