Chapter 21

Linear modelling

Syllabus reference: 5.4

Contents:

OPENING PROBLEM

At a junior tournament, a group of young athletes throw a discus. The *age* and *distance thrown* are recorded for each athlete.

Athlete	A	B	C	D	E	F	G	H	I	J	K	L
Age (years)	12	16	16	18	13	19	11	10	20	17	15	13
Distance thrown (m)	20	35	23	38	27	47	18	15	50	33	22	20

Things to think about:

a Do you think the distance an athlete can throw is related to the person's age?

b How can you graph the data so we can clearly see the relationship between the variables?

c How can we *measure* the relationship between the variables?

d How can we use this data to predict the distance a 14 year old athlete can throw a discus?

Statisticians are often interested in how two variables are **related**.

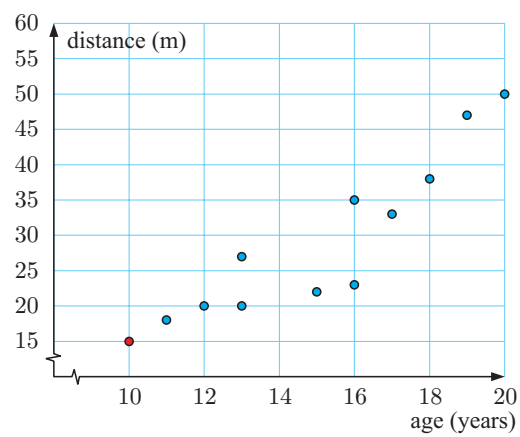
For example, in the **Opening Problem**, we want to know how a change in the *age* of the athlete will affect the *distance* the athlete can throw.

We can observe the relationship between the variables by plotting the data on a **scatter diagram**.

We place the **independent variable** *age* on the horizontal axis, and the **dependent variable** *distance* on the vertical axis.

We then plot each data value as a point on the scatter diagram. For example, the red point represents athlete H, who is 10 years old and threw the discus 15 metres.

From the general shape formed by the dots, we can see that as the *age* increases, so does the *distance thrown*.

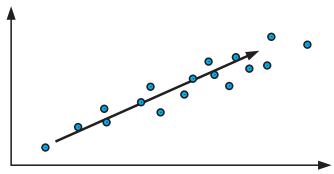


A CORRELATION

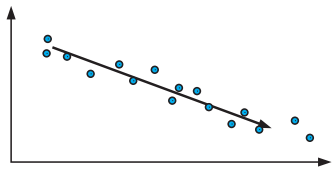
Correlation refers to the relationship or association between two variables.

There are several characteristics we consider when describing the correlation between two variables: direction, linearity, strength, outliers, and causation.

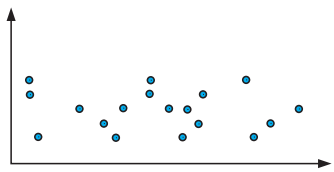
DIRECTION



For a generally *upward* trend, we say that the correlation is **positive**. An increase in the independent variable means that the dependent variable generally increases.



For a generally *downward* trend, we say that the correlation is **negative**. An increase in the independent variable means that the dependent variable generally decreases.

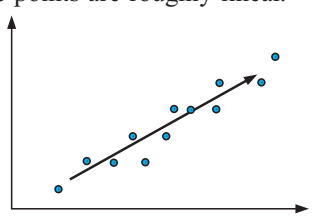


For *randomly scattered* points, with no upward or downward trend, we say there is **no correlation**.

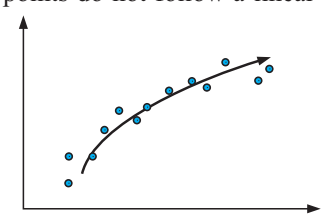
LINEARITY

We determine whether the points follow a **linear** trend, or in other words approximately form a straight line.

These points are roughly linear.



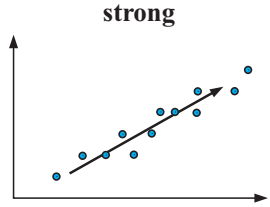
These points do not follow a linear trend.



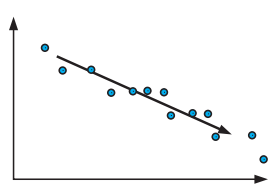
STRENGTH

We want to know how closely the data follows a pattern or trend. The strength of correlation is usually described as either strong, moderate, or weak.

strong

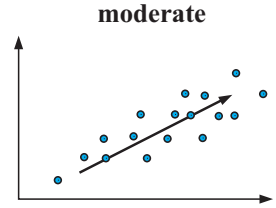


strong positive

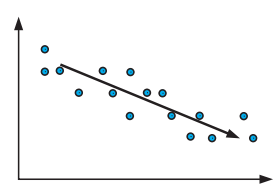


strong negative

moderate

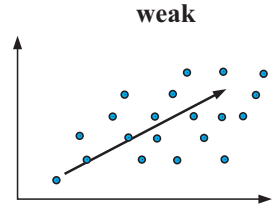


moderate positive

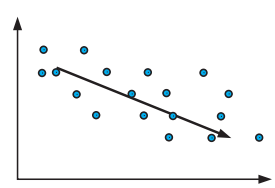


moderate negative

weak



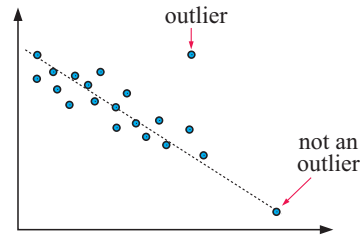
weak positive



weak negative

OUTLIERS

We observe and investigate any **outliers**, or isolated points which do not follow the trend formed by the main body of data.



If an outlier is the result of a recording or graphing error, it should be discarded. However, if the outlier proves to be a genuine piece of data, it should be kept.

For the scatter diagram for the data in the **Opening Problem**, we can say that there is a strong positive correlation between *age* and *distance thrown*. The relationship appears to be linear, with no outliers.

CAUSATION

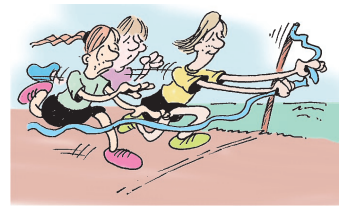
Correlation between two variables does not necessarily mean that one variable *causes* the other.

Consider the following:

1 The *arm length* and *running speed* of a sample of young children were measured, and a strong, positive correlation was found to exist between the variables.

Does this mean that short arms cause a reduction in running speed or that a high running speed causes your arms to grow long? This would clearly be nonsense.

Rather, the strong, positive correlation between the variables is attributed to the fact that both *arm length* and *running speed* are closely related to a third variable, *age*. Up to a certain age, both *arm length* and *running speed* increase with *age*.



2 The number of television sets sold in Ballarat and the number of stray dogs collected in Bendigo were recorded over several years and a strong positive correlation was found between the variables.

Obviously the number of television sets sold in Ballarat was not influencing the number of stray dogs collected in Bendigo. Both variables have simply been increasing over the period of time that their numbers were recorded.

If a change in one variable *causes* a change in the other variable then we say that a **causal relationship** exists between them.

For example, in the **Opening Problem** there is a causal relationship in which increasing the *age* of an athlete increases the *distance thrown*.

In cases where this is not apparent, there is no justification, based on high correlation alone, to conclude that changes in one variable cause the changes in the other.

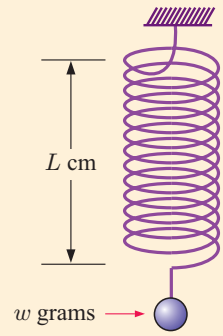
CASE STUDY — MASS ON A SPRING

Suppose we wish to examine the relationship between the *length* of a helical spring and the *mass* that is hung from the spring.

The force of gravity on the mass causes the spring to stretch.

The length of the spring depends on the force applied, so the dependent variable is the *length*.

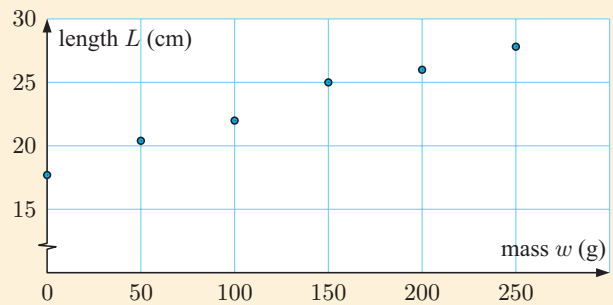
The following experimental results are obtained when objects of varying mass are hung from the spring:



Mass w (grams)	0	50	100	150	200	250
Length L (cm)	17.7	20.4	22.0	25.0	26.0	27.8

For each addition of 50 grams in mass, the consecutive increases in length are roughly constant.

There appears to be a strong positive correlation between the mass of the object hung from the spring, and the length of the spring. The relationship appears to be linear, with no obvious outliers.



EXERCISE 21A

1 Describe what is meant by:

a a scatter diagram **b** correlation **c** positive correlation

d negative correlation **e** an outlier.

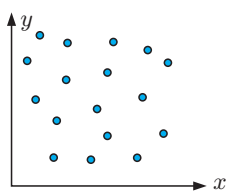
2 For the following scatter diagrams, comment on:

i the existence of any *pattern* (positive, negative or no correlation)

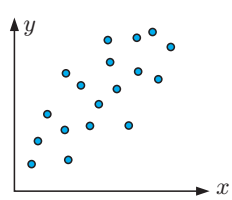
ii the relationship *strength* (zero, weak, moderate or strong)

iii whether the relationship is linear **iv** whether there are any outliers.

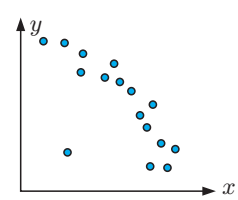
a



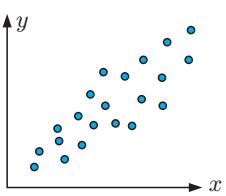
b



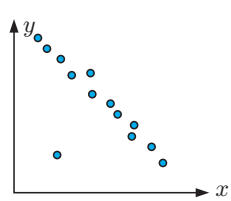
c



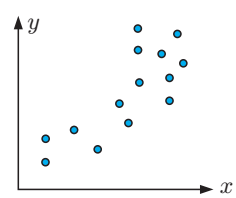
d



e



f



3 Ten students participated in a typing contest, where the students were given one minute to type as many words as possible. The table below shows how many words each student typed, and how many errors they made:

Student	A	B	C	D	E	F	G	H	I	J
Number of words x	40	53	20	65	35	60	85	49	35	76
Number of errors y	11	15	2	20	4	22	30	16	27	25

a Draw a scatter diagram for this data.

b Name the student who is best described as:

i slow but accurate

ii fast but inaccurate

iii an outlier.

c Describe the direction and strength of correlation between these variables.

d Is the data linear?

GRAPHICS CALCULATOR INSTRUCTIONS

4 The scores awarded by two judges at an ice skating competition are shown in the table.

Competitor	P	Q	R	S	T	U	V	W	X	Y
Judge A	5	6.5	8	9	4	2.5	7	5	6	3
Judge B	6	7	8.5	9	5	4	7.5	5	7	4.5

a Construct a scatter diagram for this data with Judge A's scores on the horizontal axis and Judge B's scores on the vertical axis.

b Copy and complete the following comments about the scatter diagram:
There appears to be ....., ... correlation between Judge A's scores and Judge B's scores. This means that as Judge A's scores increase, Judge B's scores .....

c Would it be reasonable to conclude that an increase in Judge A's scores *causes* an increase in Judge B's scores?

B PEARSON'S CORRELATION COEFFICIENT

In the previous section, we classified the strength of the correlation between two variables as either strong, moderate, or weak. We observed the points on a scatter diagram, and made a judgement as to how clearly the points formed a linear relationship.

However, this method can be quite inaccurate, so it is important to get a more precise measure of the strength of linear correlation between two variables. We achieve this using **Pearson's product-moment correlation coefficient** r.

For a set of n data given as ordered pairs (x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n),

Pearson's correlation coefficient is $$r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \sum(y-\overline{y})^2}}$$

where $\overline{x}$ and $\overline{y}$ are the means of the x and y data respectively, and $\sum$ means the sum over all the data values.

You are not required to learn this formula. Instead, we generally use technology to find the value of r.

The values of r range from -1 to $+1$.

The **sign** of r indicates the **direction** of the correlation.

- A positive value for r indicates the variables are **positively correlated**. An increase in one of the variables will result in an increase in the other.
- A negative value for r indicates the variables are **negatively correlated**. An increase in one of the variables will result in a decrease in the other.

The **size** of r indicates the **strength** of the correlation.

- A value of r close to $+1$ or -1 indicates strong correlation between the variables.
- A value of r close to zero indicates weak correlation between the variables.

The following table is a guide for describing the strength of linear correlation using r.

Positive correlation		**Negative correlation**	
$r = 1$	perfect positive correlation	$r = -1$	perfect negative correlation
$0.95 \leqslant r < 1$	very strong positive correlation	$-1 < r \leqslant -0.95$	very strong negative correlation
$0.87 \leqslant r < 0.95$	strong positive correlation	$-0.95 < r \leqslant -0.87$	strong negative correlation
$0.5 \leqslant r < 0.87$	moderate positive correlation	$-0.87 < r \leqslant -0.5$	moderate negative correlation
$0.1 \leqslant r < 0.5$	weak positive correlation	$-0.5 < r \leqslant -0.1$	weak negative correlation
$0 \leqslant r < 0.1$	no correlation	$-0.1 < r \leqslant 0$	no correlation

Example 1

Self Tutor

A chemical fertiliser company wishes to determine the extent of correlation between the *quantity of compound X* used and the *lawn growth* per day.

Find and interpret the correlation coefficient between the two variables.

Lawn	*Compound X* (g)	*Lawn growth* (mm)
A	1	3
B	2	3
C	4	6
D	5	8

	x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})(y-\overline{y})$	$(x-\overline{x})^2$	$(y-\overline{y})^2$
	1	3	-2	-2	4	4	4
	2	3	-1	-2	2	1	4
	4	6	1	1	1	1	1
	5	8	2	3	6	4	9
Totals:	12	20			13	10	18

$\therefore\ \overline{x} = \dfrac{\sum x}{n}$ $\qquad \overline{y} = \dfrac{\sum y}{n}$ $\qquad r = \dfrac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \sum(y-\overline{y})^2}}$

$= \dfrac{12}{4}$ $\qquad = \dfrac{20}{4}$ $\qquad = \dfrac{13}{\sqrt{10 \times 18}}$

$= 3$ $\qquad = 5$ $\qquad \approx 0.969$

There is a very strong positive correlation between the *quantity of compound X* used and *lawn growth*.

This suggests that the more of compound X used, the greater the lawn growth per day. However, care must be taken, as the small amount of data may provide a misleading result.

Example 2

Self Tutor

A group of adults is weighed, and their maximum speed when sprinting is measured:

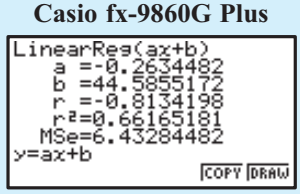
Weight x (kg)	85	60	78	100	83	67	79	62	88	68
Maximum speed y (km h^{-1})	26	29	24	17	22	30	25	24	19	27

a Use technology to find r for the data.

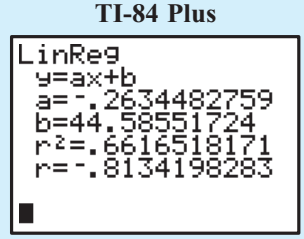
b Describe the correlation between *weight* and *maximum speed*.

a

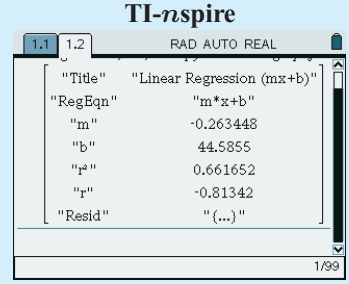
Casio fx-9860G Plus



TI-84 Plus



TI-*n*spire



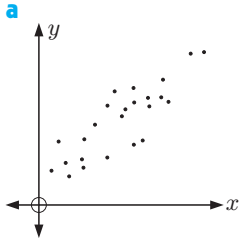
Using technology, $r \approx -0.813$.

b Since $-0.87 < r \leqslant -0.5$, there is a moderate negative correlation between *weight* and *maximum speed*.

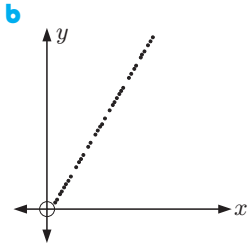
EXERCISE 21B

1 Match each scatter diagram with the correct value of r.

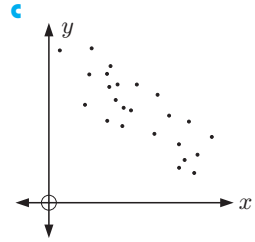
a



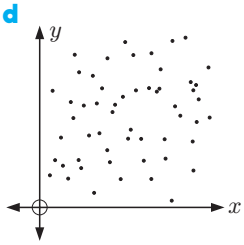
b



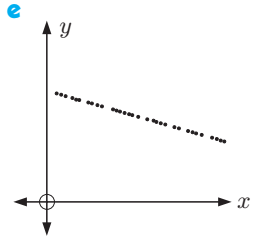
c



d



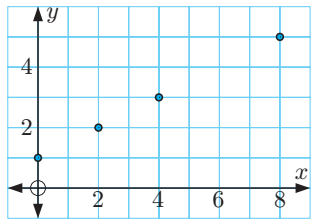
e



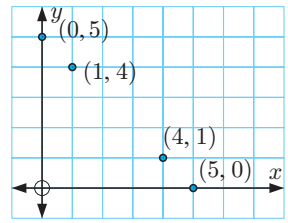
A $r = 1$ **B** $r = 0.6$ **C** $r = 0$ **D** $r = -0.7$ **E** $r = -1$

2 Use the formula $r = \dfrac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \sum(y-\overline{y})^2}}$ to determine the correlation coefficient r in the following:

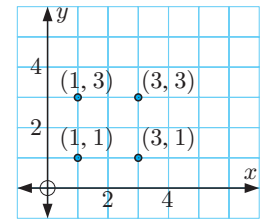
a



b



c



Check your answers using a calculator.

3 The table alongside shows the ages of five children, and the number of times they visited the doctor in the last year:

Age	2	5	7	5	8
Number of doctor visits	10	6	5	4	3

a Draw a scatter diagram of the data.

b Calculate the correlation coefficient by hand. *Check* your answer using technology.

c Describe the correlation between *age* and *number of doctor visits*.

4 Jill hangs her clothes out to dry every Saturday, and notices that the clothes dry more quickly some days than others. She investigates the relationship between the temperature and the time her clothes take to dry:

Temperature x (°C)	25	32	27	39	35	24	30	36	29	35
Drying time y (min)	100	70	95	25	38	105	70	35	75	40

a Draw a scatter diagram for this data. **b** Calculate r.

c Describe the correlation between *temperature* and *drying time*.

5 The table below shows the ticket and beverage sales for each day of a 12 day music festival:

Ticket sales ($\$x \times 1000$)	25	22	15	19	12	17	24	20	18	23	29	26
Beverage sales ($\$y \times 1000$)	9	7	4	8	3	4	8	10	7	7	9	8

a Draw a scatter diagram for this data. **b** Calculate r.

c Describe the correlation between *ticket sales* and *beverage sales*.

6 A local council collected data from a number of parks in the area, recording the size of the parks and the number of trees each contained:

Size (hectares)	2.8	6.9	7.4	4.3	8.5	2.3	9.4	5.2	8.0	4.9	6.2	3.3	4.5
Number of trees	18	31	33	24	13	17	40	32	37	30	32	25	28

a Draw a scatter diagram for this data.

b Would you expect r to be positive or negative? Explain your answer. **c** Calculate r.

d Are there any outliers? **e** Remove the outlier, and re-calculate r.

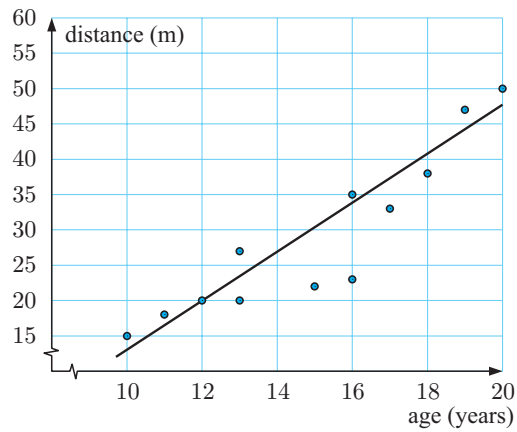
C LINE OF BEST FIT

Consider again the data from the **Opening Problem**:

Athlete	A	B	C	D	E	F	G	H	I	J	K	L
Age (years)	12	16	16	18	13	19	11	10	20	17	15	13
Distance thrown (m)	20	35	23	38	27	47	18	15	50	33	22	20

We have seen that there is a strong positive linear correlation between *age* and *distance thrown*.

We can therefore model the data using a **line of best fit**.



We draw a line of best fit connecting variables X and Y as follows:

Step 1: Calculate the mean of the X values $\overline{x}$, and the mean of the Y values $\overline{y}$.

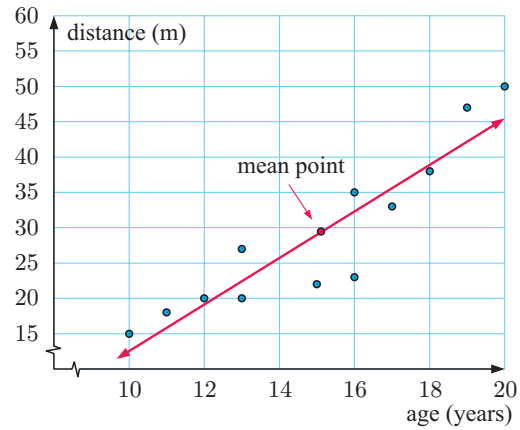
Step 2: Mark the **mean point** $(\overline{x}, \overline{y})$ on the scatter diagram.

Step 3: Draw a line through the mean point which fits the trend of the data, and so that about the same number of data points are above the line as below it.

The line formed is called a **line of best fit by eye**. This line will vary from person to person.

For the **Opening Problem**, the mean point is (15, 29). So, we draw our line of best fit through (15, 29).

We can use the line of best fit to estimate the value of y for any given value of x, and vice versa.



Example 3 **Self Tutor**

Consider the following data for a mass on a spring:

Mass w (grams)	0	50	100	150	200	250
Length L (cm)	17.7	20.4	22.0	25.0	26.0	27.8

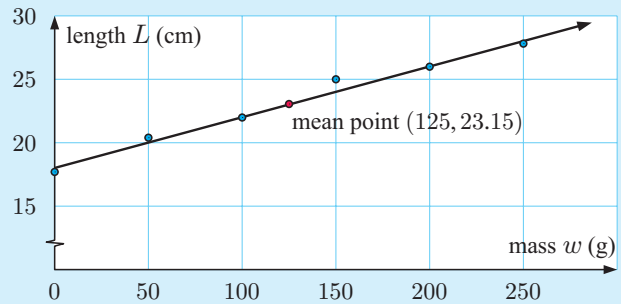
a Draw a scatter diagram for the data, and draw a line of best fit incorporating the mean point.

b Find the equation of the line you have drawn.

a The mean of the masses in the experiment is $\overline{w} = 125$ g.

The mean of the spring lengths is $\overline{L} = 23.15$ cm.

$\therefore$ the mean point is $(125, 23.15)$.



b The line of best fit above passes through $(125, 23.15)$ and $(200, 26)$.

The line has gradient $m = \dfrac{26 - 23.15}{200 - 125} \approx 0.04$.

Its equation is $\dfrac{y - 26}{x - 200} \approx 0.04$

$\therefore \quad y - 26 \approx 0.04x - 8$

$\therefore \quad y \approx 0.04x + 18$

or in this case $L \approx 0.04w + 18$

EXERCISE 21C

1 For each of the following data sets:

i draw the scatter diagram and draw a line of best fit incorporating the mean point

ii find the equation of the line you have drawn.

a

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

b

x	13	18	7	1	12	6	15	4	17	3
y	10	6	17	18	13	14	6	15	5	14

c

x	11	7	16	4	8	10	17	5	12	2	8	13	9	18	5	12
y	16	12	32	5	7	19	30	14	19	6	17	24	15	34	6	26

2 Over 10 days the *maximum temperature* and *number of car break-ins* was recorded for a city:

Maximum temperature x (°C)	22	17	14	18	24	29	33	32	26	22
Number of car break-ins y	30	18	9	20	31	38	47	40	29	25

a Draw a scatter diagram for the data.

b Describe the correlation between the *maximum temperature* and *number of car break-ins*.

c Draw a line of best fit through the data.

d Find the equation of the line of best fit.

e Use your equation to estimate the number of car break-ins you would expect to occur on a 25°C day.

3 To investigate whether speed cameras have an impact on road safety, data was collected from several cities. The number of speed cameras in operation was recorded for each city, as well as the number of accidents over a 7 day period.

Number of speed cameras x	7	15	20	3	16	17	28	17	24	25	20	5	16	25	15	19
Number of car accidents y	48	35	31	52	40	35	28	30	34	19	29	42	31	21	37	32

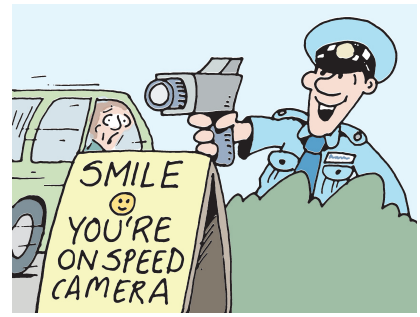
a Construct a scatter diagram to display the data.

b Calculate r for the data.

c Describe the relationship between the *number of speed cameras* and the *number of car accidents*.

d Plot the mean point $(\overline{x}, \overline{y})$ on the scatter diagram, and draw a line of best fit through the mean point.

e Where does your line cut the y-axis? Interpret what this answer means.



D THE LEAST SQUARES REGRESSION LINE

The problem with drawing a line of best fit by eye is that the line drawn will vary from one person to another.

Instead, we use a method known as **linear regression** to find the equation of the line which best fits the data. The most common method is the method of 'least squares'.

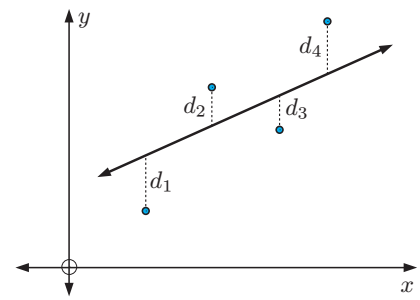
Consider the set of points alongside.

For any line we draw to model the linear relationship between the points, we can find the vertical distances d_1, d_2, d_3, ... between each point and the line.

We can then square each of these distances, and find their sum $d_1^2 + d_2^2 + d_3^2 + ...$

If the line is a good fit for the data, most of the distances will be small, and so will the sum of their squares.

The **least squares regression line** is the line which makes this sum as small as possible.



The demonstration alongside allows you to experiment with various data sets. Use trial and error to find the least squares regression line for each set.

DEMO

In practice, rather than finding the regression line by experimentation, we use a **calculator** or **statistics package**.

GRAPHICS CALCULATOR INSTRUCTIONS

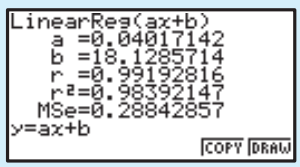
STATISTICS PACKAGE

Example 4 Self Tutor

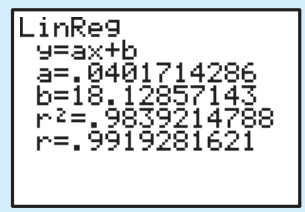
Use technology to find the least squares regression line for the mass on a spring data:

Mass w (grams)	0	50	100	150	200	250
Length L (cm)	17.7	20.4	22.0	25.0	26.0	27.8

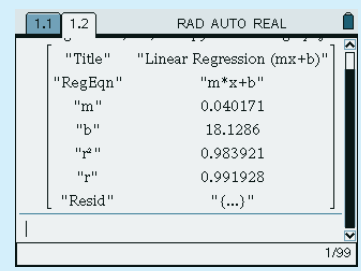
Casio fx-9860G Plus



TI-84 Plus



TI-*n*spire



The least squares regression line is

$$y \approx 0.0402x + 18.1,$$
or $L \approx 0.0402w + 18.1$.

Compare this equation with the one we obtained in **Example 3**.

EXERCISE 21D

1 Use technology to find the least squares regression line for each data set in **Exercise 21C** question **1**.

2 Find the least squares regression line for the *maximum temperature* vs *number of break-ins* data in **Exercise 21C** question **2**. Hence check your prediction made in part **e**.

3 Steve wanted to see whether there was any relationship between the temperature when he leaves for work in the morning, and the time it takes to get to work.
He collected data over a 14 day period:

Temperature x (°C)	25	19	23	27	32	35	29	27	21	18	16	17	28	34
Time y (min)	35	42	49	31	37	33	31	47	42	36	45	33	48	39

a Draw a scatter diagram of the data.

b Calculate r.

c Describe the relationship between the variables.

d Is it reasonable to try to find a line of best fit for this data? Explain your answer.

4 The table below shows the price of petrol and the number of customers per hour for sixteen petrol stations.

Petrol price x (cents per litre)	105.9	106.9	109.9	104.5	104.9	111.9	110.5	112.9
Number of customers y	45	42	25	48	43	15	19	10

Petrol price x (cents per litre)	107.5	108.0	104.9	102.9	110.9	106.9	105.5	109.5
Number of customers y	30	23	42	50	12	24	32	17

a Calculate r for the data.

b Describe the relationship between the *petrol price* and the *number of customers*.

c Use technology to find the least squares regression line.

E INTERPOLATION AND EXTRAPOLATION

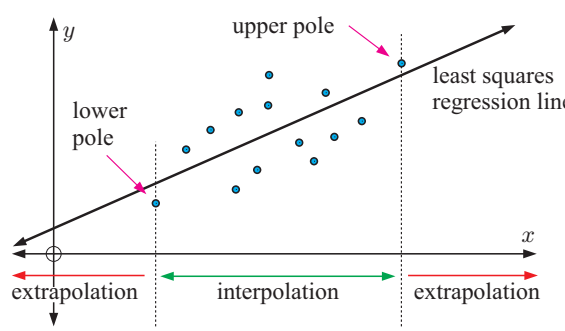
Suppose we have gathered data to investigate the association between two variables. We obtain the scatter diagram shown below. The data with the lowest and highest values of x are called the **poles**.

We use the least squares regression line to estimate values of one variable given a value for the other.

If we use values of x **in between** the poles, we say we are **interpolating** between the poles.

If we use values of x **outside** the poles, we say we are **extrapolating** outside the poles.

The accuracy of an interpolation depends on how linear the original data was. This can be gauged by determining the correlation coefficient and ensuring that the data is randomly scattered around the linear regression line.



The accuracy of an extrapolation depends not only on how linear the original data was, but also on the assumption that the linear trend will continue past the poles. The validity of this assumption depends greatly on the situation we are looking at.

As a general rule, it is reasonable to interpolate between the poles, but unreliable to extrapolate outside the poles.

Example 5 Self Tutor

The table below shows how far a group of students live from school, and how long it takes them to travel there each day.

Distance from school x (km)	7.2	4.5	13	1.3	9.9	12.2	19.6	6.1	23.1
Time to travel to school y (min)	17	13	29	2	25	27	41	15	53

a Draw a scatter diagram of the data.

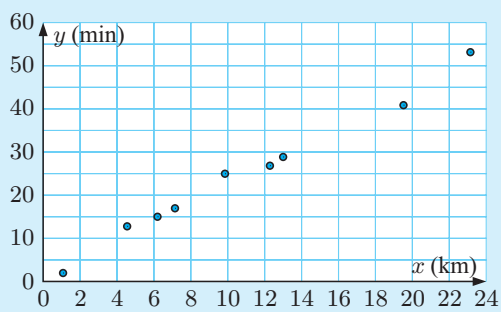
b Use technology to find: **i** r **ii** the equation of the least squares regression line.

c Pam lives 15 km from school.

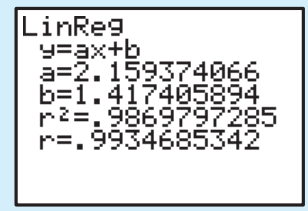
i Estimate how long it takes Pam to travel to school.

ii Comment on the reliability of your estimate.

a



b



i $r \approx 0.993$

ii The least squares regression line is $y \approx 2.16x + 1.42$.

c **i** When $x = 15$, $y \approx 2.16(15) + 1.42 \approx 33.8$
So, it will take Pam approximately 34 minutes to travel to school.

ii The estimate is an interpolation, and the correlation coefficient indicates a very strong correlation. This suggests that the estimate is reliable.

EXERCISE 21E

1 Consider the *temperature* vs *drying time* problem on page **553**.

a Use technology to find the equation of the least squares regression line.

b Estimate the time it will take for Jill's clothes to dry on a 28°C day.

c How reliable is your estimate in **b**?

2 Consider the *ticket sales* vs *beverage sales* problem on page **554**.

a Find the equation of the least squares regression line.

b The music festival is extended by one day, and \$35 000 worth of tickets are sold.

i Predict the beverage sales for this day.

ii Comment on the reliability of your prediction.

3 The table below shows the amount of time a collection of families spend preparing homemade meals each week, and the amount of money they spend each week on fast food.

Time on homemade meals x (hours)	3.3	6.0	4.0	8.5	7.2	2.5	9.1	6.9	3.8	7.7
Money on fast food y (\$)	85	0	60	0	27	100	15	40	59	29

a Draw a scatter diagram for the data.

b Calculate the value of r.

c Use technology to find the equation of the least squares regression line.

d Interpret the gradient and y-intercept of the least squares regression line.

e Another family spends 5 hours per week preparing homemade meals. Estimate how much money they spend on fast food each week. Comment on the reliability of your estimate.

4 The ages and heights of children at a playground are given below:

Age x (years)	3	9	7	4	4	12	8	6	5	10	13
Height y (cm)	94	132	123	102	109	150	127	110	115	145	157

a Draw a scatter diagram for the data.

b Use technology to find the least squares regression line.

c At what age would you expect children to reach a height of 140 cm?

d Interpret the gradient of the least squares regression line.

e Use the line to predict the height of a 20 year old. Do you think this prediction is reliable?

5 Once a balloon has been blown up, it slowly starts to deflate. A balloon's diameter was recorded at various times after it was blown up:

Time t (hours)	0	10	25	40	55	70	90	100	110
Diameter D (cm)	40.2	37.8	34.5	30.2	26.1	23.9	19.8	17.2	14.0

a Draw a scatter diagram of the data.

b Describe the correlation between D and t.

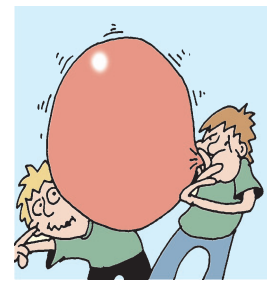
c Find the equation of the least squares regression line.

d Use this equation to predict:

i the diameter of the balloon after 80 hours

ii the time it took for the balloon to completely deflate.

e Which of your predictions in **d** is more likely to be reliable?



6 The mass of bacteria in a culture is measured each day for five days.

t days	1	2	3	4	5
M grams	3.6	5.7	9.1	14.6	23.3

a Add a row to the table for the values of $\ln M$.

b Use technology to graph M against t and $\ln M$ against t. Which of the scatter diagrams is linear?

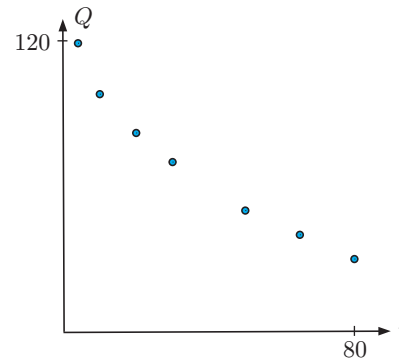
c Use technology to find the linear model $\quad Y = mX + c \quad$ where $\quad Y \equiv \ln M \quad$ and $\quad X \equiv t$.

d Hence show that $\quad M \approx 2.25 \times 1.60^t$.

e Estimate the original mass of bacteria.

7 The quantity Q of a chemical responsible for skin elasticity is measured at various ages (t years). The results are shown in the table and graph.

Age t (years)	2	10	20	30	50	65	80
Chemical Q (mg)	119	98	82	70	50	40	30



a What is the effect on Q as t increases?

b Is the graph linear?

c Construct a table of values comparing Q with $\sqrt{t}$.

d Explain why $Q = m\sqrt{t} + c$ is a likely model for the original data.

e Use your calculator to find m and c.

f Find the quantity of chemical in:

i a newly born baby **ii** a 25 year old.

g Are the answers in **f** likely to be reliable?

8 A bird bath is filled with water. Over time, the water evaporates as shown in the table below:

Time t (hours)	3	6	9	12	15	18	21	24
Water remaining V (litres)	6.7	3.6	2	1.1	0.6	0.32	0.18	0.10

a Draw a scatter diagram of V against t.

b Draw a scatter diagram of:

i $\ln V$ against t **ii** $\ln V$ against $\ln t$.

c Find the model connecting V and t.

d Estimate the amount of water remaining in the bird bath after 5 hours.

e Estimate the amount of water which has evaporated after 10 hours.

THEORY OF KNOWLEDGE

In the previous exercise we saw examples of data which was non-linear, but for which we could *transform* the variables so a linear model could be used.

In other situations we can use quadratic or trigonometric functions to model data.

1 Can all data be modelled by a known mathematical function?

2 How reliable is mathematics in predicting real-world phenomena?

Friedrich Wilhelm Bessel (1784 - 1846) was a German mathematician and astronomer who described the Bessel functions named after him. The Bessel functions are the solutions to a particular class of **differential equation**, which is an equation involving derivative functions. They are used in both classical and quantum physics to describe the dynamics of gravitational systems.

Friedrich Wilhelm Bessel

3 Are the Bessel functions defined by nature or by man?

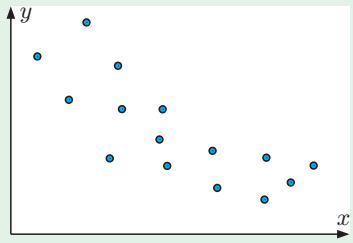
REVIEW SET 21A NON-CALCULATOR

1 For the following scatter diagrams, comment on:

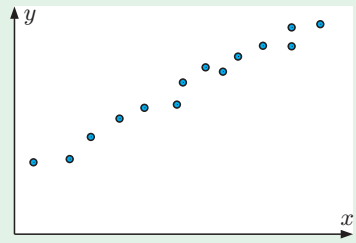
i the direction and strength of correlation between the two variables

ii whether the relationship is linear.

a



b



2 The results of a group of students for a Maths test and an Art essay are compared:

Student	A	B	C	D	E	F	G	H	I	J
Maths test	64	67	69	70	73	74	77	82	84	85
Art essay	85	82	80	82	72	71	70	71	62	66

a Construct a scatter diagram for the data. Make the scales on both axes from 60 to 90.

b Describe the relationship between the Mathematics and Art marks.

c Given the mean Maths score was 74.5 and the mean Art score was 74.1, draw a line of best fit on your graph.

3 The Botanical Gardens have been trying out a new chemical to control the number of beetles infesting their plants. The results of one of their tests are shown in the table.

a Draw a scatter diagram for the data.

b Given the correlation coefficient $r \approx -0.859$, describe the correlation between the *quantity of chemical* and the *number of surviving beetles*.

Sample	*Quantity of chemical* (g)	*Number of surviving beetles*
A	2	11
B	5	6
C	6	4
D	3	6
E	9	3

4 A clothing store recorded the length of time customers were in the store and the amount they spent.

Time (min)	8	18	5	10	17	11	2	13	18	4	11	20	23	22	17
Money (€)	40	78	0	46	72	86	0	59	33	0	0	122	90	137	93

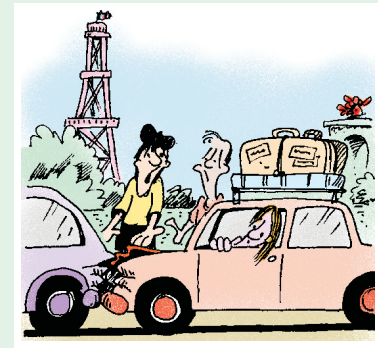
a Draw a scatter diagram of the data.

b Given the mean time ≈ 13.3 mins, and the mean amount $\approx$ €57.07, plot the mean point on your diagram and draw a line of best fit.

c Describe the relationship between *time in the store* and the *money spent*.

5 Safety authorities advise drivers to travel three seconds behind the car in front of them. This gives the driver a greater chance of avoiding a collision if the car in front has to brake quickly or is itself involved in an accident.

A test was carried out to find out how long it would take a driver to bring a car to rest from the time a red light was flashed. The following results are for one driver in the same car under the same test conditions.



Speed v (km h^{-1})	10	20	30	40	50	60	70	80	90
Stopping time t (s)	1.23	1.54	1.88	2.20	2.52	2.83	3.15	3.45	3.83

a Produce a scatter diagram of the data.

b The least squares regression line has the equation $t = 0.03v + 0.9$. Plot this line on your scatter diagram.

c Use the equation in **b** to estimate the stopping time for a speed of:

i 55 km h^{-1} **ii** 110 km h^{-1}

d Which of your estimates in **c** is more likely to be reliable?

REVIEW SET 21B CALCULATOR

1 Thomas rode for an hour each day for eleven days. He recorded the number of kilometres he rode and the temperature on that day.

Temperature T (°C)	32.9	33.9	35.2	37.1	38.9	30.3	32.5	31.7	35.7	36.3	34.7
Distance d (km)	26.5	26.7	24.4	19.8	18.5	32.6	28.7	29.4	23.8	21.2	29.7

a Construct a scatter diagram of the data.

b Find and interpret Pearson's correlation coefficient for the two variables.

c Calculate the equation of the least squares regression line.

d Using your answer to **c**, how hot must it get before Thomas does not ride at all?

2 A garden centre manager believes that during March, the number of customers is related to the temperature at noon. Over a period of a fortnight the number of customers and the noon temperature were recorded.

Temperature x (°C)	23	25	28	30	30	27	25	28	32	31	33	29	27
Number of customers y	57	64	62	75	69	58	61	78	80	35	84	73	76

a Draw a scatter diagram of the data.

b Calculate the correlation coefficient r.

c Are there any outliers?

d Remove the outlier, and re-calculate r.

e Using your answer to **d**, describe the association between the *number of customers* and the *noon temperature* at the garden centre.

3 Fifteen students were weighed, and their pulse rates were measured:

Weight x (kg)	61	52	47	72	62	79	57	45	67	71	80	58	51	43	55
Pulse rate y (beats per min)	65	59	54	74	69	87	61	59	70	69	75	60	56	53	58

a Draw a scatter diagram for the data.

b Describe the relationship between *weight* and *pulse rate*.

c Calculate the mean point $(\overline{x}, \overline{y})$.

d Draw a line of best fit through the data.

e Find the equation of your line of best fit.

f Hence estimate the pulse rate of a student who weighs 65 kg.

4 Eight identical flower beds contain petunias. The different beds were watered different numbers of times each week, and the number of flowers each bed produced was recorded in the table below:

Number of waterings n	0	1	2	3	4	5	6	7
Flowers produced f	18	52	86	123	158	191	228	250

a Which is the independent variable?

b Find the equation of the least squares regression line.

c Plot the least squares regression line on a scatter diagram of the data.

d Violet has two beds of petunias. One she waters five times a fortnight ($2\frac{1}{2}$ times a week), and the other ten times a week.

i How many flowers can she expect from each bed?

ii Which is the more reliable estimate?

5 The yield of pumpkins on a farm depends on the quantity of fertiliser used.

Fertiliser x (g m^{-2})	4	13	20	26	30	35	50
Yield y (kg)	1.8	2.9	3.8	4.2	4.7	5.7	4.4

a Draw a scatter diagram of the data and identify the outlier.

b Calculate the correlation coefficient:

i with the outlier included **ii** without the outlier.

c Calculate the equation of the least squares regression line:

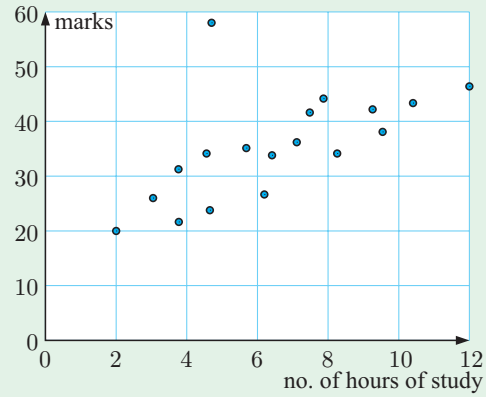
i with the outlier included **ii** without the outlier.

d If you wish to estimate the yield when 15 g m^{-2} of fertiliser is used, which regression line from **c** should be used?

e Attempt to explain what may have caused the outlier.

REVIEW SET 21C

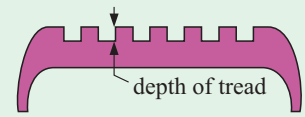
1 The scatter diagram alongside shows the marks obtained by students in a test out of 50 marks, plotted against the number of hours each student studied for the test.



a Describe the correlation between the variables.

b How should the outlier be treated? Explain your answer.

2 A sample of 8 tyres was taken to examine the association between the *tread depth* and the *number of kilometres travelled.*



tyre cross-section

Kilometres x ($\times 1000$)	14	17	24	34	35	37	38	39
Tread depth y (mm)	5.7	6.5	4.0	3.0	1.9	2.7	1.9	2.3

a Draw a scatter diagram of the data.

b Calculate the correlation coefficient r.

c Describe the correlation between the *tread depth* and the *number of kilometres travelled.*

3 The trunk widths and heights of the trees in a garden were recorded:

Trunk width x (cm)	35	47	72	40	15	87	20	66	57	24	32
Height y (m)	11	18	24	12	3	30	22	21	17	5	10

a Draw a scatter diagram of the data.

b Which of the points is an outlier?

c How would you describe the tree represented by the outlier?

d Calculate the mean point $(\overline{x}, \overline{y})$.

e Draw a line of best fit through the data.

f Find the equation of the line of best fit.

g Estimate the height of a tree with trunk width 120 cm. How reliable is this estimate?

4 A drinks vendor varies the price of Supa-fizz on a daily basis. He records the number of drinks sold in the following table:

Price p	\$2.50	\$1.90	\$1.60	\$2.10	\$2.20	\$1.40	\$1.70	\$1.85
Sales s	389	450	448	386	381	458	597	431

a Produce a scatter diagram for the data.

b Are there any outliers? If so, should they be included in the analysis?

c Find the equation of the least squares regression line.

d Do you think the least squares regression line would give an accurate prediction of sales if Supa-fizz was priced at 50 cents? Explain your answer.

5 The maximum speed of a canoe on a lake with different numbers of rowers is recorded in the table below:

Number of rowers r	4	6	10	14	18
Maximum speed S (km h^{-1})	8.7	10.3	12.6	14.2	15.9

a Draw a scatter diagram of S against r.

b Draw a scatter diagram of $\ln S$ against $\ln r$.

c Find the model connecting r and S.

d Predict the maximum speed of the canoe if there are 8 rowers.