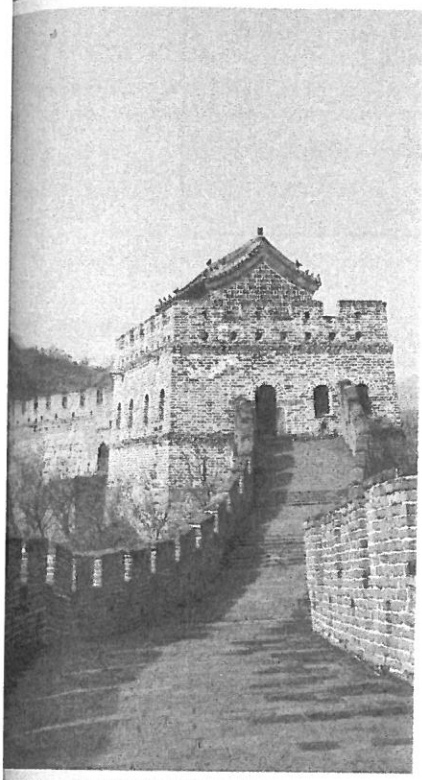
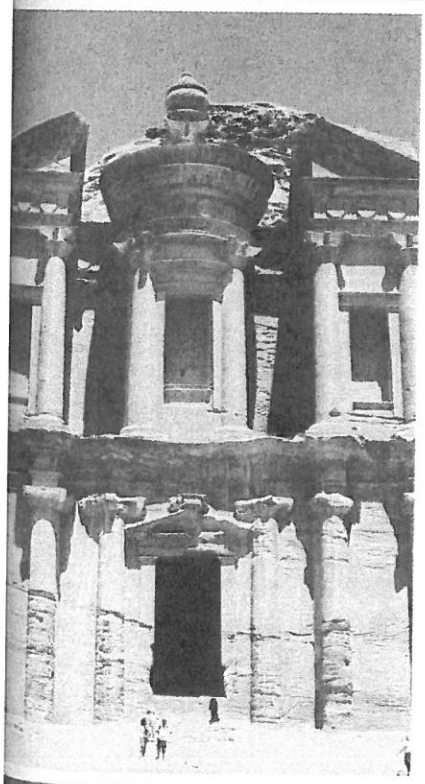
Chapter 12





Linear correlation

Contents:

- **A** Correlation
- **B** Measuring correlation
- **C** Line of best fit by eye
- **D** Least squares regression line

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 What happens to the distance thrown as the age of the athlete increases?

c How could you graph the data to more clearly see the relationship between the variables?

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

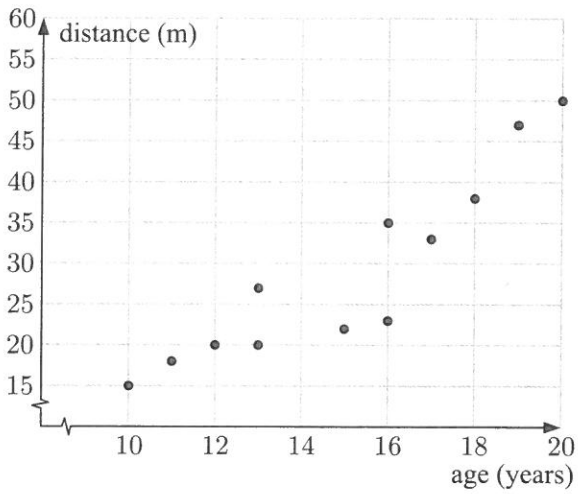
Statisticians are often interested in how two variables are **related**.

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

We can observe the relationship between the variables using a **scatter plot**. We place the independent variable *age* on the horizontal axis, and the dependent variable *distance* on the vertical axis.

We then graph each data value as a point on the scatter plot. 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*.

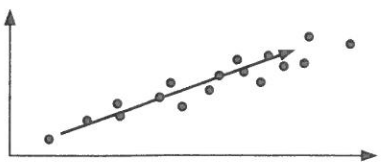


A CORRELATION

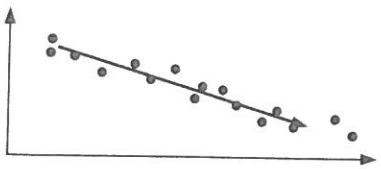
Correlation refers to the relationship or association between two variables.

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 generally results in an increase in the dependent variable.



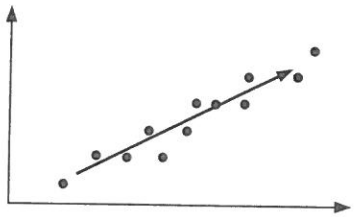
For a generally *downward* trend, we say that the correlation is **negative**. An increase in the independent variable generally results in a decrease in the dependent variable.

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

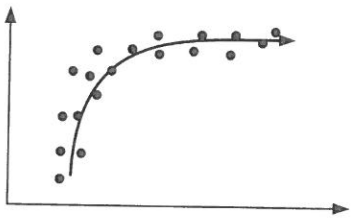
LINEARITY

When a trend exists, if the points approximately form a straight line, we say the trend is **linear**.

These points are roughly linear.



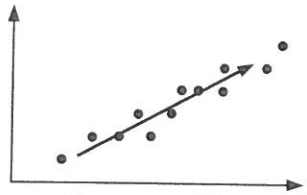
These points do not follow a linear trend.



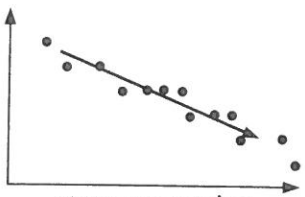
STRENGTH

To describe how closely the data follows a pattern or trend, we talk about the strength of correlation. It is usually described as either strong, moderate, or weak.

strong

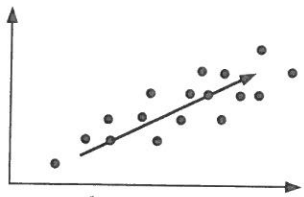


strong positive

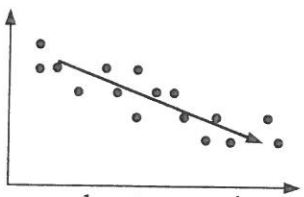


strong negative

moderate

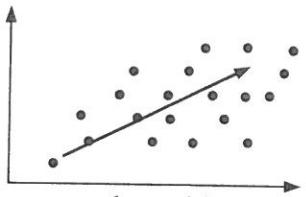


moderate positive

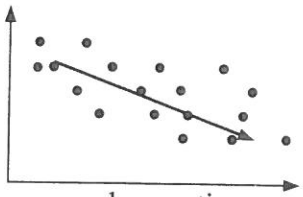


moderate negative

weak



weak positive

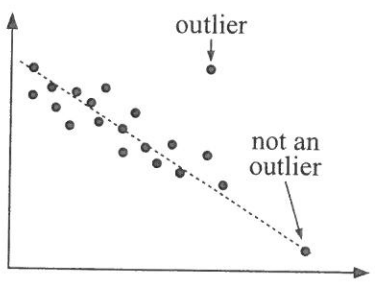


weak negative

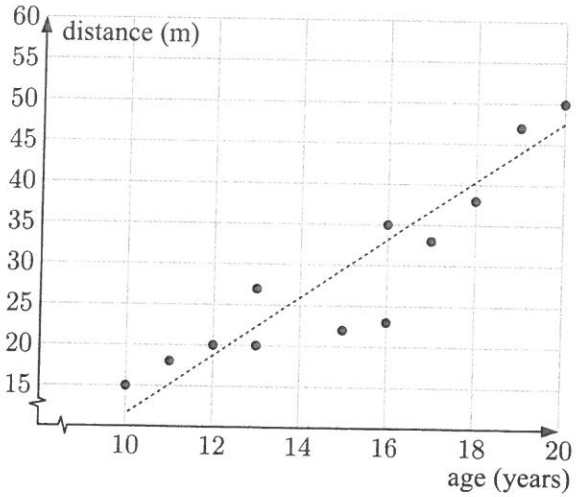
OUTLIERS

Outliers are 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 is a genuine piece of data, it should be kept.



For the scatter plot of 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.

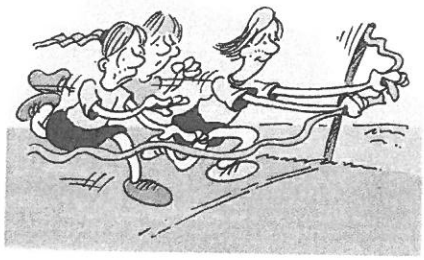


CAUSALITY

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

For example:

- The *arm length* and *running speed* of a sample of young children were measured, and a strong, positive correlation was found between the variables.
  This does *not* mean that short arms cause a reduction in running speed, or that a high running speed causes your arms to grow long.
  Rather, there is a strong, positive correlation between the variables because 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*.



- The number of television sets sold in Ballarat and the number of stray dogs collected in Bendigo were recorded over several years. 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. It is coincidental that the variables both increased over this period of time.

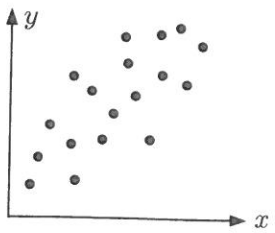
If a change in one variable *causes* a change in the other variable then we say that a **causal relationship** exists between them.

In cases where a causal relationship is not apparent, we cannot conclude that a causal relationship exists based on high correlation alone.

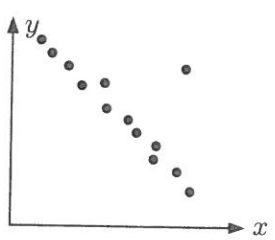
EXERCISE 12A

1 For each scatter plot, describe the relationship between the variables. Consider the direction, strength, and linearity of the relationship, as well as the presence of any outliers.

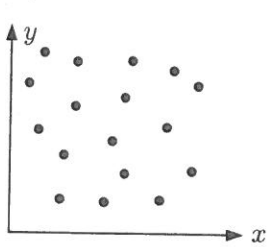
a



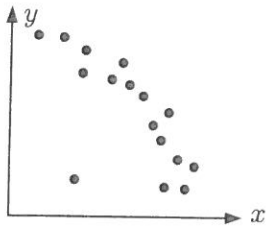
b



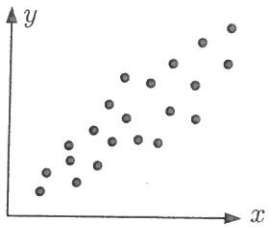
c



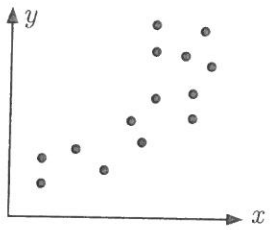
d



e



f



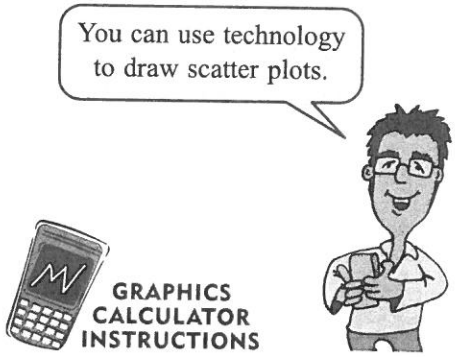
2 The scores awarded by two judges at an ice skating competition are shown in the table.

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 plot for the 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 plot:
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? Explain your answer.



3 Paul owns a company which installs industrial air conditioners. The table below shows the number of workers at the company's last 10 jobs, and the time it took to complete the job.

Job	A	B	C	D	E	F	G	H	I	J
Number of workers	5	3	8	2	5	6	1	4	2	7
Time (hours)	4	6	2.5	9	3	4	10	4	7.5	3

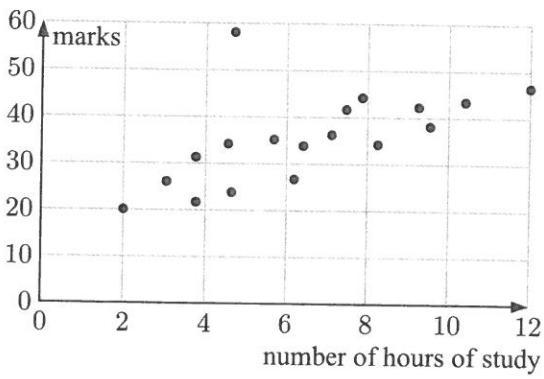
a Which job:

i took the longest **ii** involved the most workers?

b Draw a scatter plot to display the data.

c Describe the relationship between the variables *number of workers* and *time*.

4 The scatter plot 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.

c Do you think there is a causal relationship between the variables? Explain your answer.

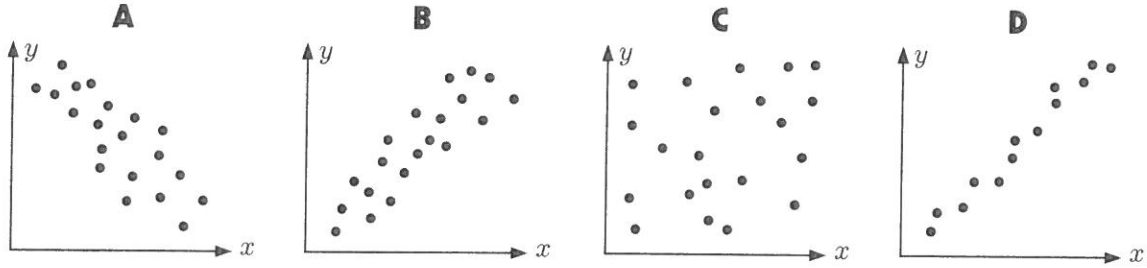
5 Choose the scatter plot which would best illustrate the relationship between the variables x and y.

a $x =$ the number of apples bought by customers, $y =$ the total cost of apples bought

b $x =$ the number of pushups a student can perform in one minute,
$y =$ the time taken for the student to run 100 metres

c $x =$ the height of a person, $y =$ the weight of the person

d $x =$ the distance a student travels to school, $y =$ the height of the student's uncle



6 When the following pairs of variables were measured, a strong positive correlation was found between each pair. Discuss whether a causal relationship exists between the variables. If not, suggest a third variable to which they may both be related.

a The lengths of one's left and right feet.

b The damage caused by a fire and the number of firemen who attend it.

c A company's expenditure on advertising, and the sales they make the following year.

d The heights of parents and the heights of their adult children.

e The numbers of hotels and numbers of service stations in rural towns.

B MEASURING CORRELATION

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 plot, and judged how clearly the points formed a linear relationship.

Since this method is *subjective* and relies on the observer's opinion, it is important to get a more precise measure of the strength of linear correlation between the variables. We achieve this using **Pearson's 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, but you should be able to calculate the value of r using technology.

GRAPHICS CALCULATOR INSTRUCTIONS

PROPERTIES OF PEARSON'S CORRELATION COEFFICIENT

- 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 variable results in an increase in the other.
  - A negative value for r indicates the variables are **negatively correlated**. An increase in one variable results in a decrease in the other.
  - If $r = 0$ then there is **no correlation** between the variables.
- 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.

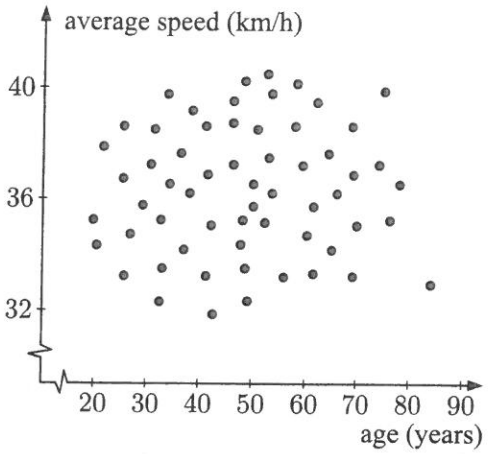
The following table is a guide for describing the strength of linear correlation using r.

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
$0.7 \leqslant r < 0.87$	moderate positive correlation	$-0.87 < r \leqslant -0.7$	moderate negative correlation
$0.5 \leqslant r < 0.7$	weak positive correlation	$-0.7 < r \leqslant -0.5$	weak negative correlation
$0 < r < 0.5$	very weak positive correlation	$-0.5 < r < 0$	very weak negative correlation

Example 1 **Self Tutor**

The Department of Road Safety wants to know if there is any association between *average speed* in the metropolitan area and the *age of drivers*. They commission a device to be fitted in the cars of drivers of different ages.

The results are shown in the scatter plot.
The r-value for this association is $+0.027$.
Describe the association.



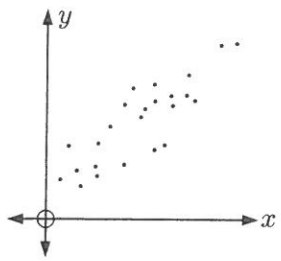
As r is close to zero, there is no correlation between the two variables.
We observe this in the graph as the points are randomly scattered.

EXERCISE 12B.1

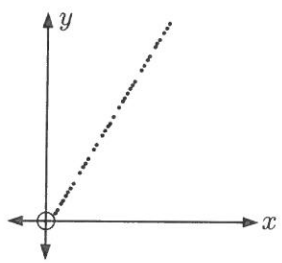
1 In a recent survey, the Department of International Commerce compared the *number of employees of a company* with its *export earnings*. A scatter plot of their data is shown alongside. The corresponding value of r is 0.556.
Describe the association between the variables.

2 Match each scatter plot with the correct value of r.

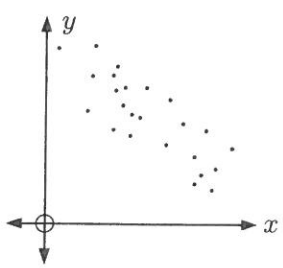
a



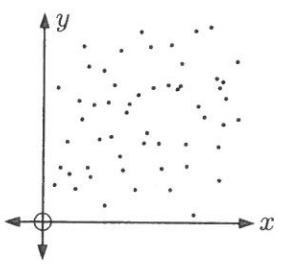
b



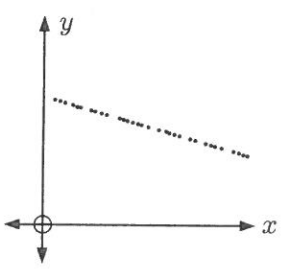
c



d



e



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

Example 2

Self Tutor

The botanical gardens have been trying 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 plot for the data.

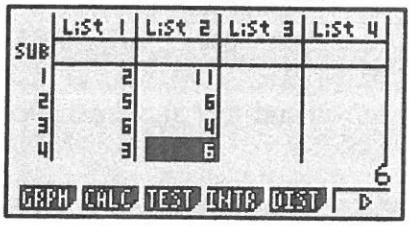
b Determine Pearson's correlation coefficient r.

c Describe the correlation between the *quantity of chemical* and the *number of surviving lawn beetles.*

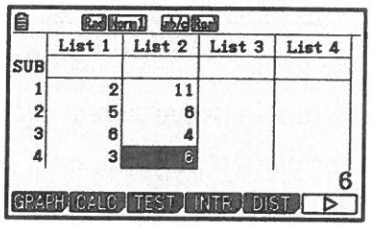
Sample	*Quantity of chemical* (g)	*Number of surviving beetles*
A	2	11
B	5	6
C	6	4
D	3	6
E	9	3

We first enter the data into separate lists:

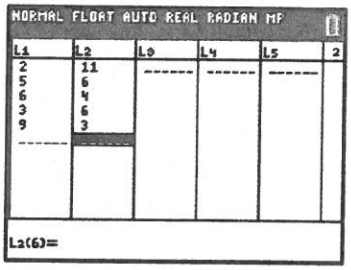
Casio fx-9860G PLUS



Casio fx-CG20 AU

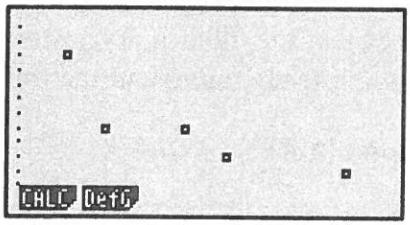


TI-84 Plus CE

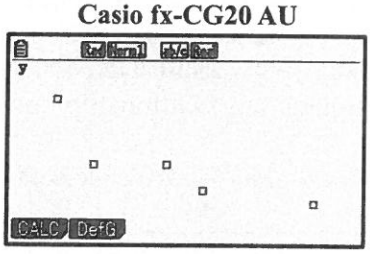


a

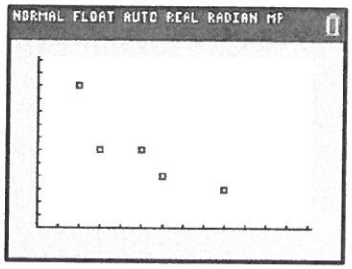
Casio fx-9860G PLUS



Casio fx-CG20 AU

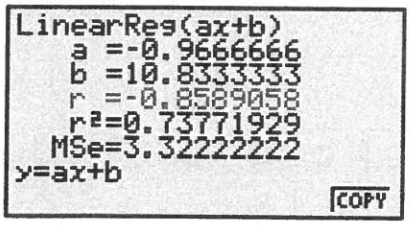


TI-84 Plus CE

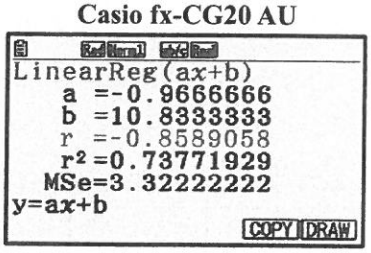


b

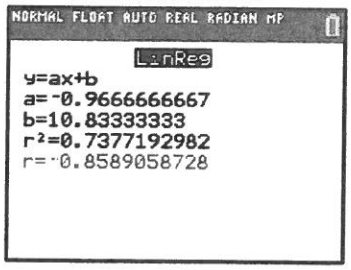
Casio fx-9860G PLUS



Casio fx-CG20 AU



TI-84 Plus CE



So, $r \approx -0.859$.

c There is a moderate negative correlation between the *quantity of chemical used* and the *number of surviving beetles.*

In general, the more chemical that is used, the fewer beetles that survive.

3 For each of the following data sets:

i Draw a scatter plot for the data.

ii Calculate Pearson's correlation coefficient r.

iii Describe the linear correlation between x and y.

a

x	1	2	3	4	5	6
y	3	2	5	5	9	6

b

x	3	8	5	14	19	10	16
y	17	12	15	6	1	10	4

c

x	3	6	11	7	5	6	8	10	4
y	2	8	8	4	7	9	11	1	5

4 A selection of students were asked how many phone calls and text messages they had received the previous day. The results are shown below.

Student	A	B	C	D	E	F	G	H
Phone calls received	4	7	1	0	3	2	2	4
Text messages received	6	9	2	2	5	8	4	7

a Draw a scatter plot for the data. **b** Calculate r.

c Describe the linear correlation between *phone calls received* and *text messages received.*

d Give a reason why this correlation may occur.

5 Consider the **Opening Problem** on page **282**.

a Calculate r for the data.

b Hence describe the association between the variables.

6 Jill hangs her clothes out to dry every Saturday. She notices that the clothes dry faster 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 plot for the data. **b** Calculate r.

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

7 This table shows the number of supermarkets in 10 towns, and the number of car accidents that have occurred in these towns in the last month.

Number of supermarkets	5	8	12	7	6	2	15	10	7	3
Number of car accidents	10	13	27	19	10	6	40	30	22	37

a Draw a scatter plot for the data. **b** Calculate r.

c Identify the outlier in the data.

d It was found that the outlier was due to an error in the data collection process.

i Recalculate r with the outlier removed.

ii Describe the relationship between the variables.

iii Discuss the effect of removing the outlier on the value of r.

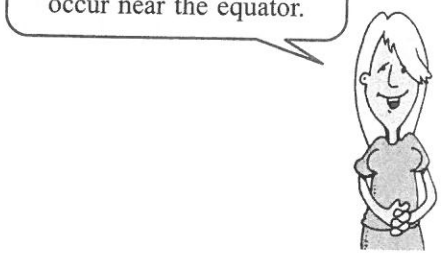
e Do you think there is a causal relationship between the variables? Explain your answer.

8 A health researcher notices that the incidence of Multiple Sclerosis (MS) is higher in some parts of the world than in others.
To investigate further, she records the *latitude* and *incidence of MS per 100 000 people* of 20 countries.

Latitude (degrees)	55	25	41	22	47	37	56	14	34	25
MS incidence per 100 000	165	95	75	20	180	140	230	15	45	65

Latitude (degrees)	27	65	10	24	4	56	46	8	50	40
MS incidence per 100 000	30	140	5	15	2	290	95	8	160	105

a Draw a scatter plot for the data.
b Calculate the value of r.
c Describe the relationship between the variables.
d Is the incidence of MS higher near the equator, or near the poles?



Activity 1 — Comparing height and foot length

In this Activity, you will explore the relationship between the *height* and *foot length* of the students in your class.

You will need: ruler, tape measure

What to do:

1 Predict whether there will be positive correlation, no correlation, or negative correlation between the *height* and *foot length* of the students in your class.

2 Measure the height and foot length of each student in your class. Record your measurements in a table like the one alongside.

Student	*Height* (cm)	*Foot length* (cm)

3 Use technology to draw a scatter plot for the data.

4 Calculate Pearson's correlation coefficient r for the data.

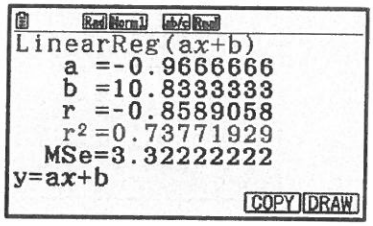
5 Describe the relationship between *height* and *foot length*. Was your prediction correct?

6 Do you think that a high value of r indicates a causal relationship in this case?

THE COEFFICIENT OF DETERMINATION r^2

To help describe the correlation between two variables, we can also calculate the **coefficient of determination** r^2. This is simply the square of Pearson's correlation coefficient r, so the direction of correlation is eliminated.

We can find r^2 using technology, or if r is already known, we can simply square this value.



INTERPRETATION OF THE COEFFICIENT OF DETERMINATION

> If there is a causal relationship, then r^2 indicates the degree to which change in the independent variable explains change in the dependent variable.

For example, an investigation into many different brands of muesli found that there is strong positive correlation between the variables *fat content* and *kilojoule content.* It was found that $r \approx 0.862$ and $r^2 \approx 0.743$.

An interpretation of this r^2 value is:

dependent variable ↙ ↘ independent variable

74.3% of the variation in *kilojoule content* of muesli can be explained by the variation in *fat content* in the muesli.

In this case, we assume that the other $100\% - 74.3\% = 25.7\%$ of the variation in *kilojoule content* in the muesli can be explained by other factors.

Example 3

Self Tutor

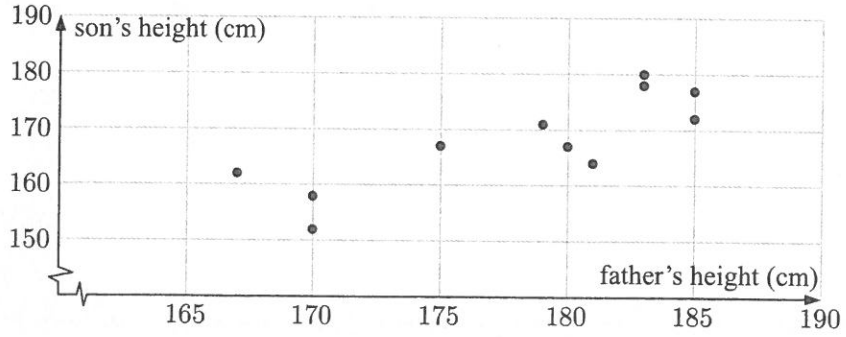
At a father-son camp, the heights of the fathers and their sons were measured.

Father's height (x cm)	175	183	170	167	179	180	183	185	170	181	185
Son's height (y cm)	167	178	158	162	171	167	180	177	152	164	172

a Draw a scatter plot for the data.

b Calculate r^2 for the data, and interpret its value.

a



b

Casio fx-9860G PLUS

```
LinearReg(ax+b)
   a =1.11190476
   b =-29.919047
   r =0.82658103
   r²=0.68323621
 MSe=26.7489417
y=ax+b
                COPY
```

Casio fx-CG20 AU

```
LinearReg(ax+b)
   a =1.11190476
   b =-29.919047
   r =0.82658103
   r² =0.68323621
 MSe=26.7489417
y=ax+b
                COPY
```

TI-84 Plus CE

```
NORMAL FLOAT AUTO REAL RADIAN MP
        LinReg
y=ax+b
a=1.111904762
b=-29.91904762
r²=0.6832362155
r=0.8265810399
```

So, $r^2 \approx 0.683$.

$\therefore$ 68.3% of the variation in the son's height can be explained by variation in the father's height.

EXERCISE 12B.2

1 From an investigation at an aquatic centre, the coefficient of determination between the variables *number of visitors* and *maximum temperature* is found to be 0.578. Complete the following interpretation of the coefficient of determination:

..... % of the variation in the ..... can be explained by the variation in .....

2 An investigation has found the association between the variables *time spent gambling* and *money lost* has correlation coefficient 0.7732. Find the coefficient of determination and interpret its meaning.

3 For a group of children, the correlation coefficient -0.365 is found between the variables *heart rate* and *age*. Find the coefficient of determination and interpret its meaning.

4 Joanne is a real estate agent. This table shows the *distance from the city* and the *selling price* of the last 12 houses she has sold.

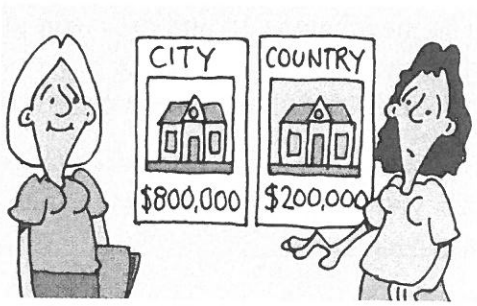
Distance from city (km)	10	4	23	16	3	35	8	7	12	24	14	12
Selling price ($ \times 1000$)	380	495	350	420	540	260	480	340	350	310	470	350

a Draw a scatter plot for the data.

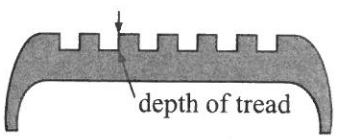
b Calculate the coefficient of determination r^2.

c What percentage of the variation in selling price can be explained by the variation in the distance from the city?

d What other factors could explain the variation in selling price?



5 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 thousand)	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 plot for the data.

b Calculate r^2 for the data, and interpret its meaning.

C LINE OF BEST FIT BY EYE

If there is a sufficiently strong linear correlation between two variables, we can draw a line of best fit to illustrate their relationship. In general, it is only worth drawing a line of best fit if the coefficient of determination $r^2 \geqslant 0.7$.

If we draw the line just by observing the points, we call it a **line of best fit by eye**. This line will vary from person to person.

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 plot.
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.

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

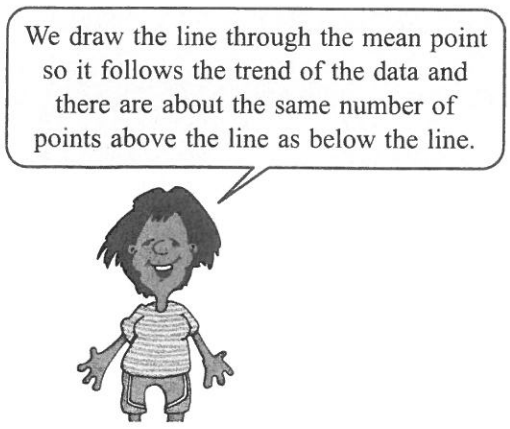
Athlete	A	B	C	D	E	F	G	H	I	J	K	L
Age (years)	12	16	16	18	13	19	11	10	20	17	15	13
Distance thrown (m)	20	35	23	38	27	47	18	15	50	33	22	20

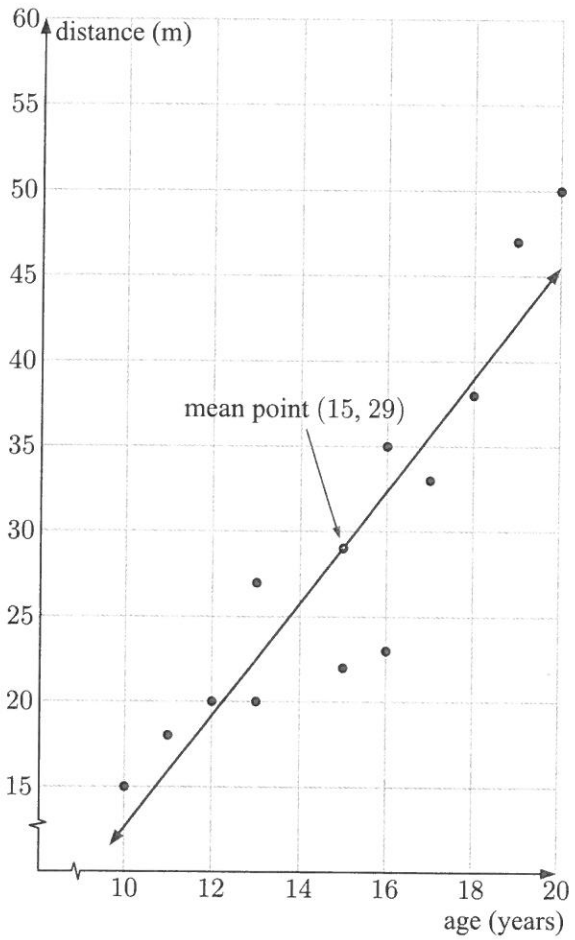
We have seen that there is a strong positive linear correlation between *age* and *distance thrown*.

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

The mean age is 15 and the mean distance thrown is 29. We therefore draw our line of best fit through the mean point $(15, 29)$.

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





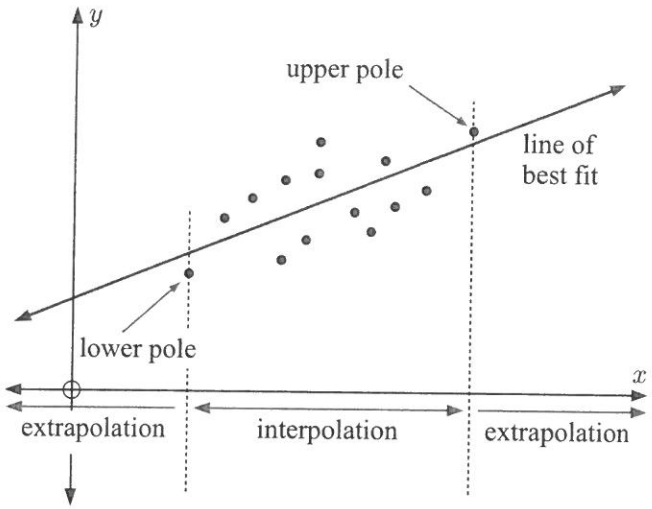
INTERPOLATION AND EXTRAPOLATION

Consider the data in the scatter plot alongside. The data with the highest and lowest values are called the **poles**.

A line of best fit has been drawn so we can predict the value of one variable for a given value of the other.

If we predict a y value for an x value **in between** the poles, we say we are **interpolating** in between the poles.

If we predict a y value for an x value **outside** the poles, we say we are **extrapolating** outside the poles.



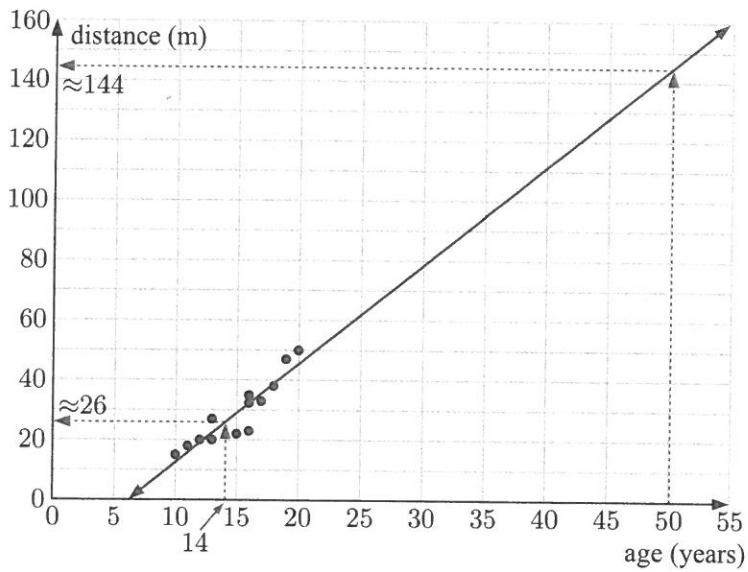
The accuracy of an interpolation depends on how linear the original data was. This can be gauged by the correlation coefficient and by ensuring that the data is randomly scattered around the line of best fit.

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.

For example, consider the line of best fit for the data in the **Opening Problem**.

The age 14 is within the range of ages already supplied, so it is reasonable to predict that a 14 year old will be able to throw the discus 26 m.

However, it is unlikely that the linear trend shown in the data will continue far beyond the poles. For example, according to the model, a 50 year old will throw the discus 144 m. This is almost twice the current world record of 76.8 m, and so it is an unreasonable prediction.



Example 4

Self Tutor

On a hot day, six cars were left in the sun in a car park. The length of time each car was left in the sun was recorded, as well as the temperature inside the car at the end of the period.

Car	A	B	C	D	E	F
Time (x min)	50	5	25	40	15	45
Temperature (y °C)	47	28	36	42	34	41

a Calculate $\overline{x}$ and $\overline{y}$. **b** Draw a scatter plot for the data.

c Locate the mean point $(\overline{x}, \overline{y})$ on the scatter plot, then draw a line of best fit through this point.

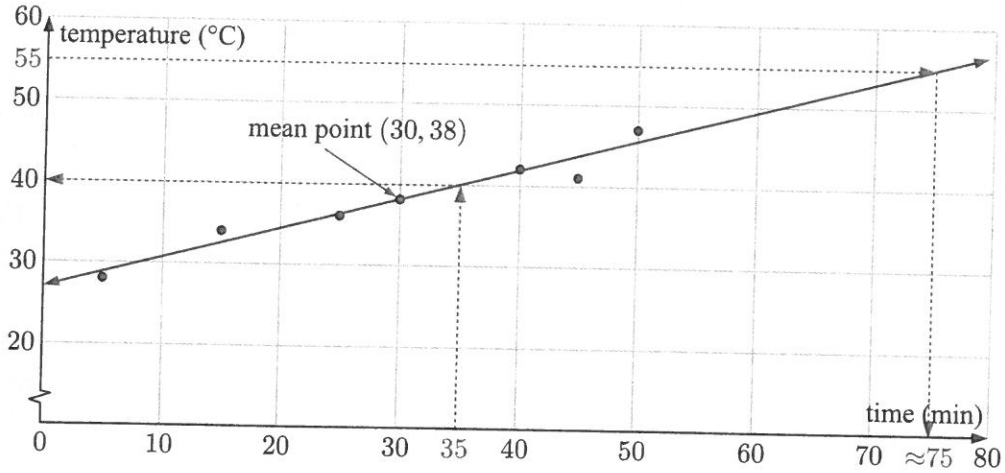
d Predict the temperature of a car which has been left in the sun for 35 minutes.

e Predict how long it would take for a car's temperature to reach 55°C.

f Comment on the reliability of your predictions in **d** and **e**.

a $\overline{x} = \dfrac{50+5+25+40+15+45}{6} = 30, \quad \overline{y} = \dfrac{47+28+36+42+34+41}{6} = 38$

b, c



d When $x = 35$, $y \approx 40$.
The temperature of a car left in the sun for 35 minutes will be approximately 40°C.

e When $y = 55$, $x \approx 75$.
It would take approximately 75 minutes for a car's temperature to reach 55°C.

f The prediction in **d** is reliable, as the data appears linear, and this is an interpolation.
The prediction in **e** may be unreliable, as it is an extrapolation and the linear trend displayed by the data may not continue beyond the 50 minute mark.

EXERCISE 12C

1 Consider the data set below.

x	5	12	20	17	10	8	25	15
y	28	19	4	18	22	20	7	10

a Draw a scatter plot for the data.

b Does the data appear to be positively or negatively correlated?

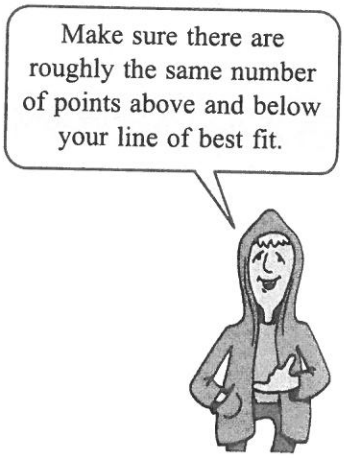
c Calculate r for the data.

d Describe the strength of the relationship between x and y.

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

f Locate the mean point, then draw a line of best fit through the mean point.

g Estimate the value of y when $x = 22$.



2 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 plot for the data.

b Calculate r.

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

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

e Locate the mean point on the scatter plot, then draw a line of best fit through the mean point.

f Estimate the pulse rate of a student who weighs 65 kg. Comment on the reliability of your estimate.

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 plot 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 Locate the mean point $(\overline{x}, \overline{y})$ on the scatter plot, then draw a line of best fit through the mean point.

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

f Estimate the number of car accidents in a city with 10 speed cameras.

4 The trunk widths and heights of the trees in a garden are given below:

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 plot for the data.

b Which of the points is an outlier?

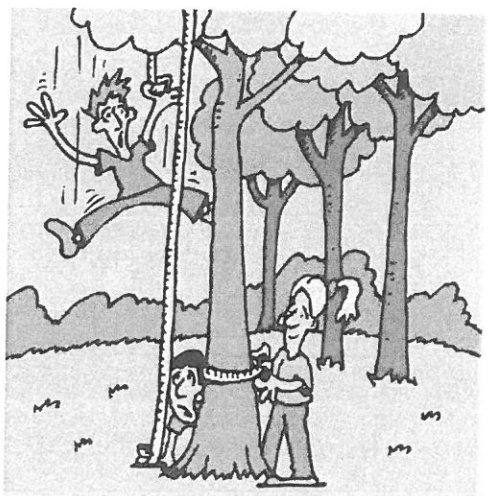
c How would you describe the tree represented by the outlier?

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

e Locate the mean point on the scatter plot, then draw a line of best fit through the mean point.

f Predict the height of a tree with trunk width 120 cm. Comment on the reliability of your prediction.

g Predict the trunk width of a tree with height 10 m. Comment on the reliability of your prediction.



D 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. For consistency, 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**".

Finding the equation of the least squares regression line by hand is quite complicated. Instead, we can use our **graphics calculator** or the **statistics package** to find the equation of the line.

Example 5

Self Tutor

The annual income and average weekly grocery bill for a selection of families is shown below:

Income (x thousand dollars)	55	36	25	47	60	64	42	50
Grocery bill (y dollars)	120	90	60	160	190	250	110	150

a Construct a scatter plot to illustrate the data.

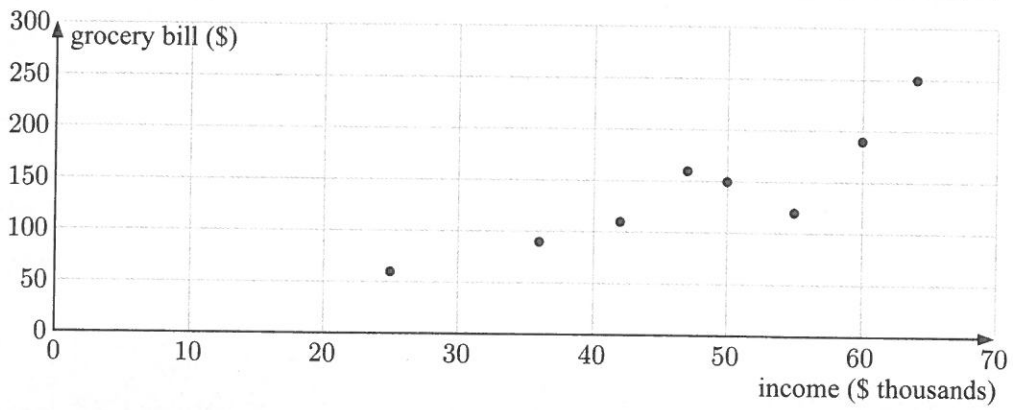
b Use technology to find the least squares regression line.

c Estimate the weekly grocery bill for a family with an annual income of \$95 000.

d Estimate the annual income of a family whose weekly grocery bill is \$100.

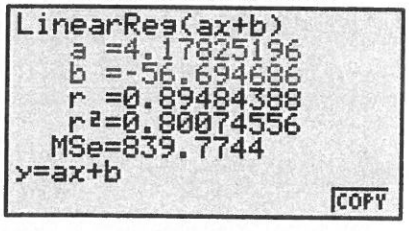
e Comment on whether the estimates in **c** and **d** are likely to be reliable.

a

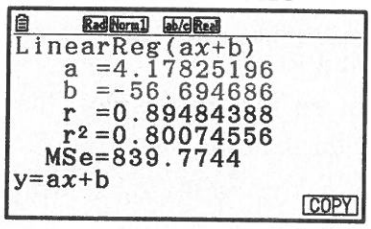


b

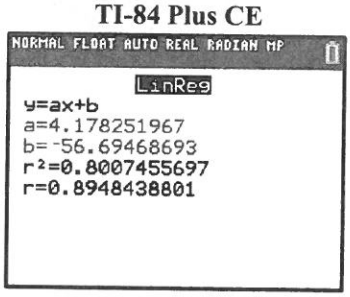
Casio fx-9860G PLUS



Casio fx-CG20 AU



TI-84 Plus CE



Using technology, the least squares regression line is $y \approx 4.18x - 56.7$

c When $x = 95$, $y \approx 4.18(95) - 56.7 \approx 340$

So, we expect a family with an income of \$95 000 to have a weekly grocery bill of approximately \$340.

d When $y = 100$, $100 \approx 4.18x - 56.7$

$\therefore \ 156.7 \approx 4.18x$ {adding 56.7 to both sides}

$\therefore \ x \approx 37.5$ {dividing both sides by 4.18}

So, we expect a family with a weekly grocery bill of \$100 to have an annual income of approximately \$37 500.

e The estimate in **c** is an extrapolation, so the estimate may not be reliable.

The estimate in **d** is an interpolation and there is strong linear correlation between the variables. We would therefore expect this estimate to be reliable.

EXERCISE 12D

1 Consider the data set below.

x	10	4	6	8	9	5	7	1	2	3
y	20	6	8	13	20	12	13	4	2	7

GRAPHICS CALCULATOR INSTRUCTIONS

a Draw a scatter plot for the data.

b Use technology to find the equation of the least squares regression line, and plot the line on your calculator.

c Use **b** to draw the least squares regression line on your scatter plot.

2 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 plot for 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.

3 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 the correlation coefficient 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.

d Estimate the number of customers per hour for a petrol station which sells petrol at 115.9 cents per litre.

e Estimate the petrol price at a petrol station which has 40 customers per hour.

f Comment on the reliability of your estimates in **d** and **e**.

4 The table below contains information about the *maximum speed* and *ceiling* (maximum altitude obtainable) for nineteen World War II fighter planes. The maximum speed is given in thousands of km/h, and the ceiling is given in km.

Maximum speed	*Ceiling*
0.46	8.84
0.42	10.06
0.53	10.97
0.53	9.906
0.49	9.448
0.53	10.36
0.68	11.73

Maximum speed	*Ceiling*
0.68	10.66
0.72	11.27
0.71	12.64
0.66	11.12
0.78	12.80
0.73	11.88

Maximum speed	*Ceiling*
0.67	12.49
0.57	10.66
0.44	10.51
0.67	11.58
0.70	11.73
0.52	10.36

a Draw a scatter plot for the data. **b** Calculate r.

c Describe the association between *maximum speed* (x) and *ceiling* (y).

d Use technology to find the least squares regression line, and draw the line on your scatter plot.

e Estimate the ceiling for a fighter plane with a maximum speed of 600 km/h.

f Estimate the maximum speed for a fighter plane with a ceiling of 11 km.

5 A group of children were asked the numbers of hours they spent exercising and watching television each week.

Exercise (x hours per week)	4	1	8	7	10	3	3	2
Television (y hours per week)	12	24	5	9	1	18	11	16

a Draw a scatter plot for the data. **b** Calculate r.

c Describe the correlation between *time exercising* and *time watching television.*

d Find the equation of the least squares regression line, and draw the line on your scatter plot.

e A child exercises for 5 hours each week. Estimate how much time this child spends watching television each week.

f Another child watches 30 hours of television each week. Estimate how much time this child spends exercising each week.

g Comment on the reliability of your estimates in **e** and **f**.

6 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 plot for 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? Explain your answer.

e Can you explain what may have caused the outlier? Do you think the outlier should be kept when analysing the data?

Activity 2 — Populations

Sydney and Melbourne are currently the most populous cities in Australia.

What to do:

1 Use the internet to find the populations of Sydney and Melbourne for at least 5 different years since 1991.

For each city, record your data in a table like the one below.

Year					
Sydney					
Melbourne					

2 Draw the scatter plot for each city on the same set of axes. Label the horizontal axis "years since 1991", so 1991 corresponds to 0, 1996 corresponds to 5, and so on. Does the data appear linear?

3 Find the equation of the least squares regression line for each city.

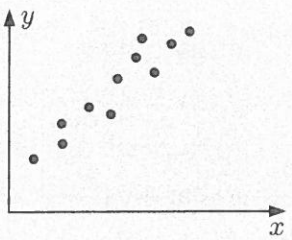
4 Which city's population is growing at a faster rate? How can you tell?

5 How can you use your regression lines to estimate when Melbourne will overtake Sydney as Australia's most populous city?

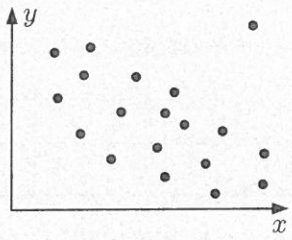
Review set 12A

1 For each scatter plot, describe the relationship between the variables. Consider the direction, strength, and linearity of the relationship, as well as the presence of any outliers.

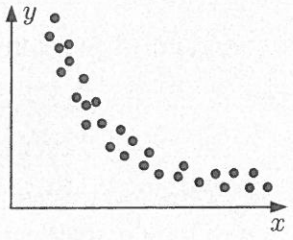
a



b



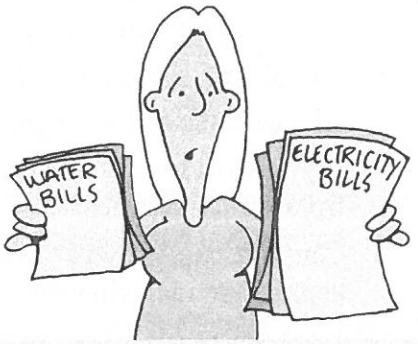
c



2 Kerry wants to investigate the relationship between the *water bill* and the *electricity bill* for the houses in her neighbourhood.

a Do you think the correlation between the variables is likely to be positive or negative? Explain your answer.

b Is there a causal relationship between the variables? Justify your answer.



3 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 plot for the data.

b Calculate Pearson's correlation coefficient r.

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

4 Jamie is a used car salesman. This table shows the age and selling price of the last 10 cars he has sold.

Age (years)	2	10	5	4	3	4	7	12	15	8
Selling price ($\$ \times 1000$)	25	10	12	7	17	15	4	3	2	7

a Draw a scatter plot for the data.

b Describe the correlation between *age* and *selling price*.

c Calculate the coefficient of determination r^2.

d What percentage of the variation in selling price can be explained by variation in age?

e What other factors could explain the variation in selling price?

5 A clothing store recorded the length of time customers were in the store and the amount of money 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 plot for the data. **b** Calculate the mean point.

c Locate the mean point on the scatter plot, then draw a line of best fit through the mean point.

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

e Estimate the amount of money spent by a person who is in the store for 15 minutes. Comment on the reliability of your estimate.

6 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. It involved one driver in the same car under the same test conditions.

Speed (x km/h)	10	20	30	40	50	60	70	80	90
Stopping time (y seconds)	1.23	1.54	1.88	2.20	2.52	2.83	3.15	3.45	3.83

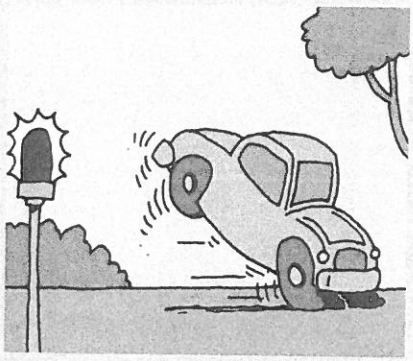
a Produce a scatter plot for the data.

b Find the least squares regression line which best fits the data, and draw the line on your scatter plot.

c Hence estimate the stopping time for a speed of:

i 55 km/h **ii** 110 km/h

d Comment on the reliability of your estimates in **c**.



7 Tomatoes are sprayed with a pesticide-fertiliser mix. The table below shows the *yield of tomatoes* per bush for various *spray concentrations.*

Spray concentration (x mL/L)	3	5	6	8	9	11	15
Yield of tomatoes per bush (y)	67	90	103	120	124	150	82

a Draw a scatter plot to display the data.

b Determine the value of r and interpret your answer.

c Is there an outlier present that is affecting the correlation?

d The outlier was found to be a recording error. Remove the outlier from the data set, and recalculate r. Is it reasonable to now draw a line of best fit?

e Determine the equation of the least squares regression line.

f Use your line to estimate:

i the yield if the spray concentration is 7 mL/L

ii the spray concentration if the yield is 200 tomatoes per bush.

g Comment on the reliability of your estimates in **f**.

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

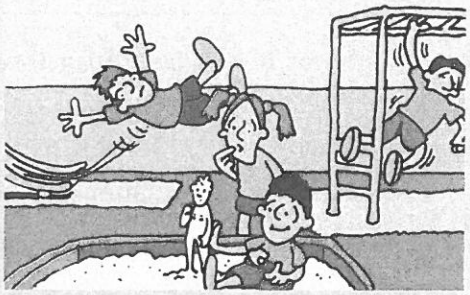
Age (x years)	3	9	7	4	4	12	8	6	5	10	13
Height (y cm)	94	132	123	102	109	150	127	110	115	145	157

a Draw a scatter plot for the data.

b Use technology to find the least squares regression line.

c Use the line to predict the height of a 5 year old child.

d Based on the given data, at what age would you expect a child to reach 140 cm in height?



Review set 12B

1 For each pair of variables, discuss whether the correlation between the variables is likely to be positive or negative, and whether a causal relationship exists between the variables:

a *price of tickets* and *number of tickets sold*

b *ice cream sales* and *number of shark attacks.*

2 A group of students is comparing their results for a Mathematics test and an Art project:

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

a Construct a scatter plot for the data.

b Describe the relationship between the Mathematics and Art marks.

3 For a group of foods, a correlation coefficient of 0.787 is found between the *fat content* and *energy*.

a Describe the relationship between the variables.

b Calculate the coefficient of determination, and interpret your answer.

4 A craft shop sells canvasses in a variety of sizes. The table alongside shows the area and price of each canvas type.

Area (x cm^2)	100	225	300	625	850	900
Price (\$$y$)	6	12	13	24	30	35

a Construct a scatter plot for the data.

b Calculate the correlation coefficient r.

c Describe the correlation between *area* and *price*.

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

e Locate the mean point on the scatter plot, then draw a line of best fit through the mean point.

f Estimate the price of a canvas with area 1200 cm^2. Discuss whether your estimate is likely to be reliable.

5 A drinks vendor varies the price of Supa-fizz on a daily basis. He records the number of sales of the drink as shown:

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

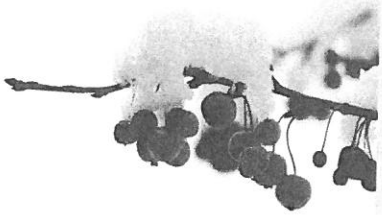
a Produce a scatter plot for the data.

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

c Calculate 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.

6 Winter frosts are important for producing good harvests of cherries and apples. The following data shows the *annual cherry yield* and *incidence of frosts* data for a cherry growing farm over a 7 year period.



Number of frosts (x)	27	23	7	37	32	14	16
Cherry yield (y tonnes)	5.6	4.8	3.1	7.2	6.1	3.7	3.8

a Draw a scatter plot for the data.

b Determine the r and r^2 values.

c Describe the association between *cherry yield* and the *number of frosts*.

d Determine the equation of the least squares regression line.

e Use the equation of the line to predict:

i the cherry yield in a year when 29 frosts were recorded

ii the number of frosts in a year when the cherry yield was 4 tonnes.

7 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 Draw a scatter plot for the data, and describe the correlation between the variables.

b Calculate the equation of the least squares regression line.

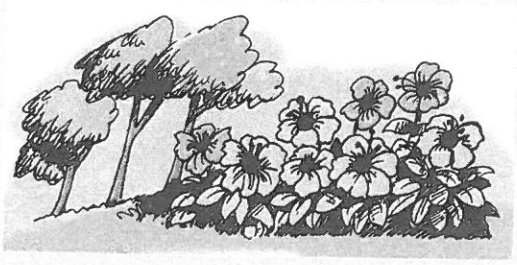
c Is it likely that a causal relationship exists between these two variables? Explain your answer.

d Plot the least squares regression line on the scatter plot.

e Violet has two beds of petunias. She waters one of the beds 5 times a fortnight and the other 10 times a week.

i How many flowers can she expect from each bed?

ii Discuss which of your estimates is likely to be more reliable.



8 Thomas rode his bicycle for an hour each day for eleven days. He recorded the number of kilometres he rode along with the temperature 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	22.8	23.5	32.6	28.7	29.4	24.2	23.2	29.7

a Using technology, construct a scatter plot for the data.

b Find and interpret Pearson's correlation coefficient.

c Find the equation of the least squares regression line.

d How far would you expect Thomas to ride on a 30°C day?