

Chapter 26

Bivariate statistics

Contents:

- A** Association between numerical variables
- B** Pearson's product-moment correlation coefficient
- C** Line of best fit by eye
- D** The least squares regression line
- E** The regression line of x against y



OPENING PROBLEM

At a junior tournament, some young athletes each 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:

- Do you think the distance an athlete can throw is related to the person's age?
- What happens to the distance thrown as the age of the athlete increases?
- How could you graph the data to more clearly see the relationship between the variables?
- How can we *measure* the relationship between the variables?

In the **Opening Problem**, each athlete has had *two* variables (*age* and *distance thrown*) recorded about them. This type of data is called **bivariate data**. We study it to understand the **relationship** between the two variables.

For example, we expect the *distance thrown* will *depend* on the athlete's *age*, so *age* is the **independent variable** and *distance thrown* is the **dependent variable**.

The **independent** and **dependent** variables are sometimes called the **explanatory** and **response** variables respectively.



In this Chapter we **describe** and **model** relationships between pairs of numerical variables.

A

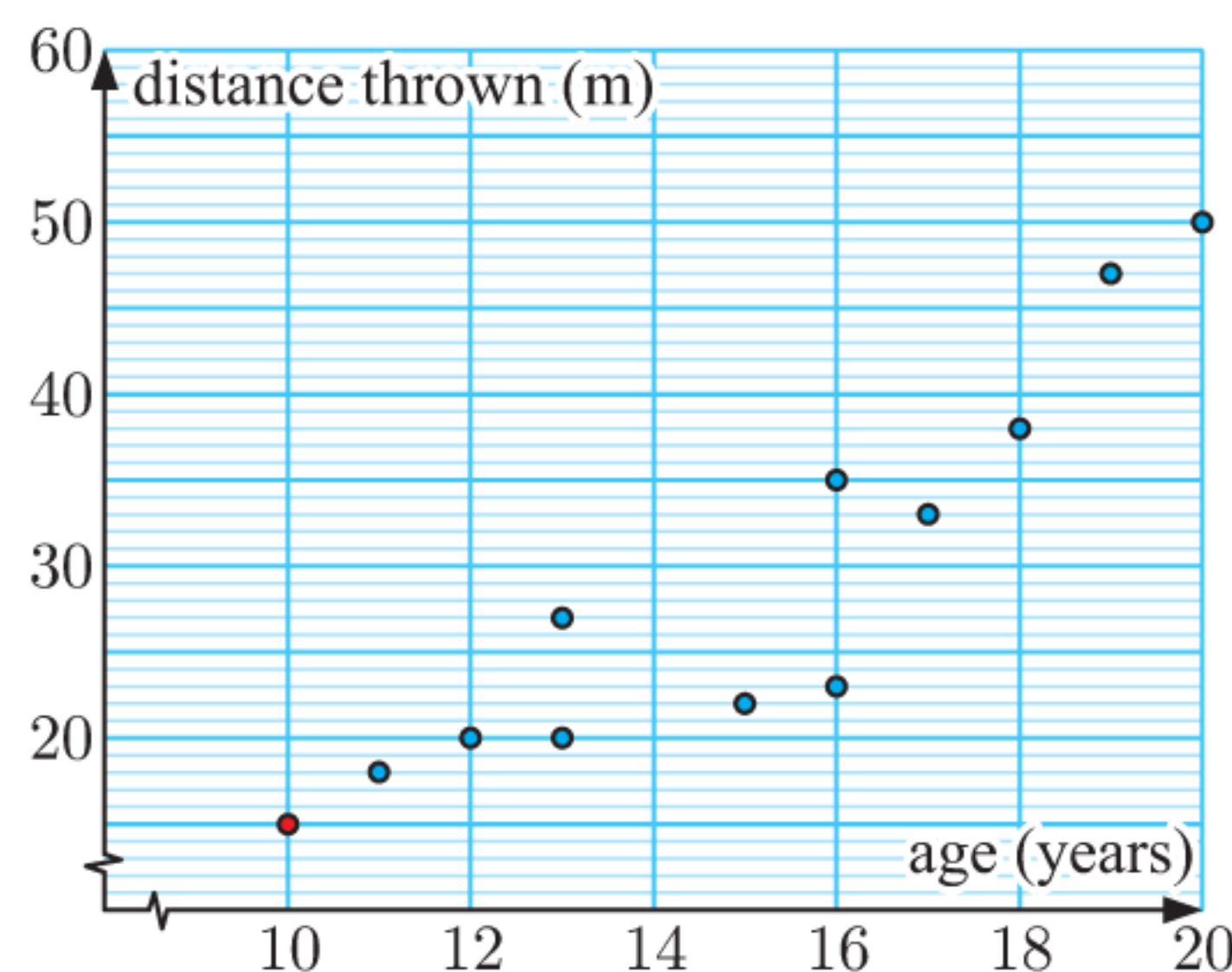
ASSOCIATION BETWEEN NUMERICAL VARIABLES

We can observe the relationship between two numerical variables using a **scatter diagram**. We usually place the independent variable on the horizontal axis, and the dependent variable on the vertical axis.

In the **Opening Problem**, the independent variable *age* is placed on the horizontal axis, and the dependent variable *distance thrown* is placed on the vertical axis.

We then graph 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*.

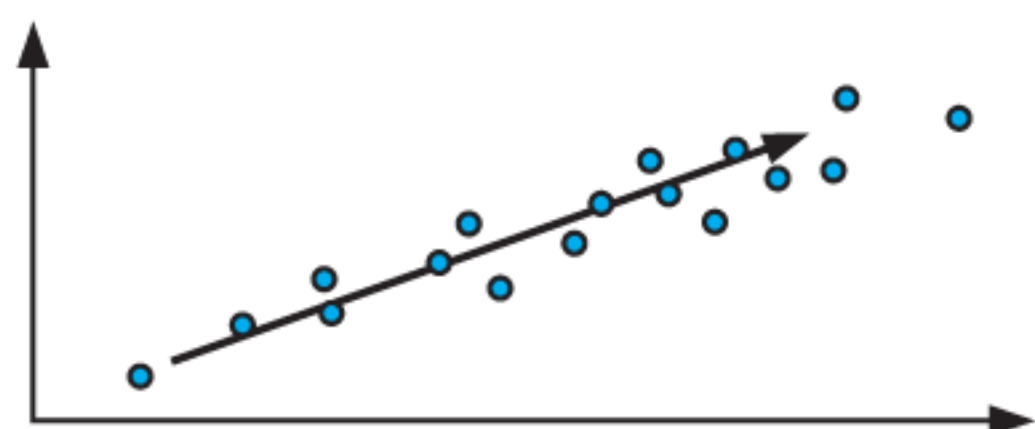


CORRELATION

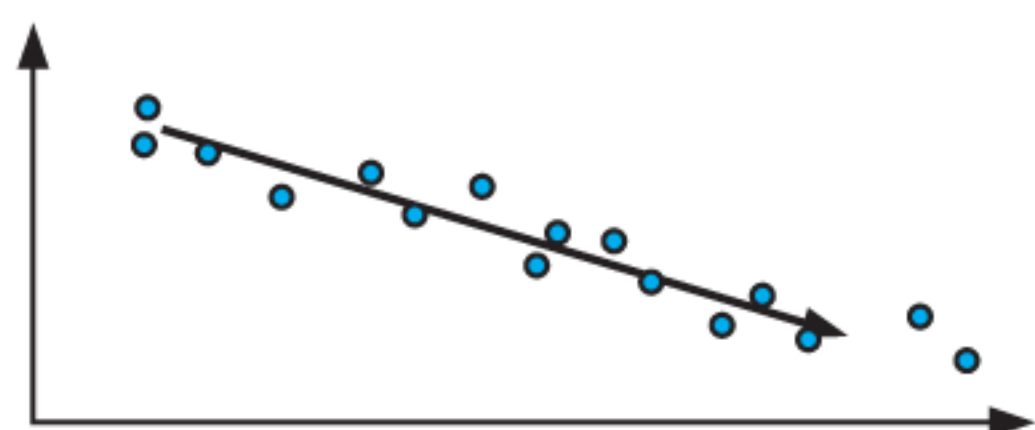
Correlation refers to the relationship or association between two numerical variables.

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

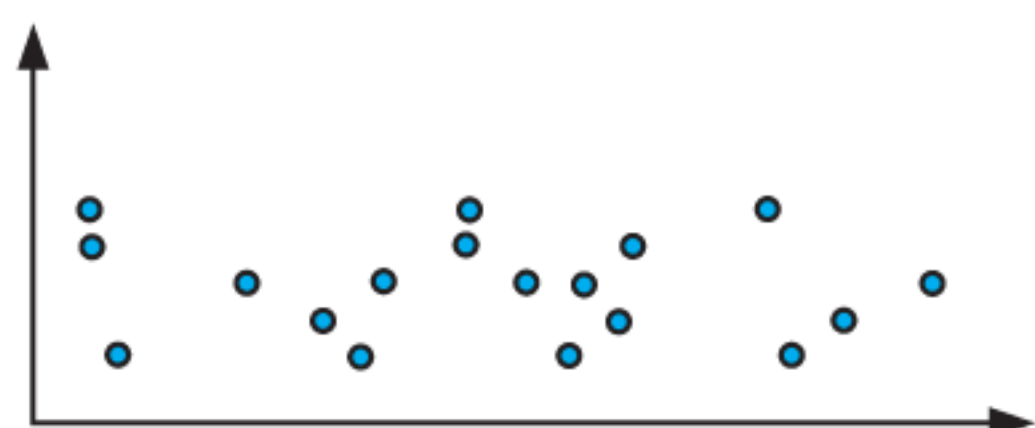
DIRECTION



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.

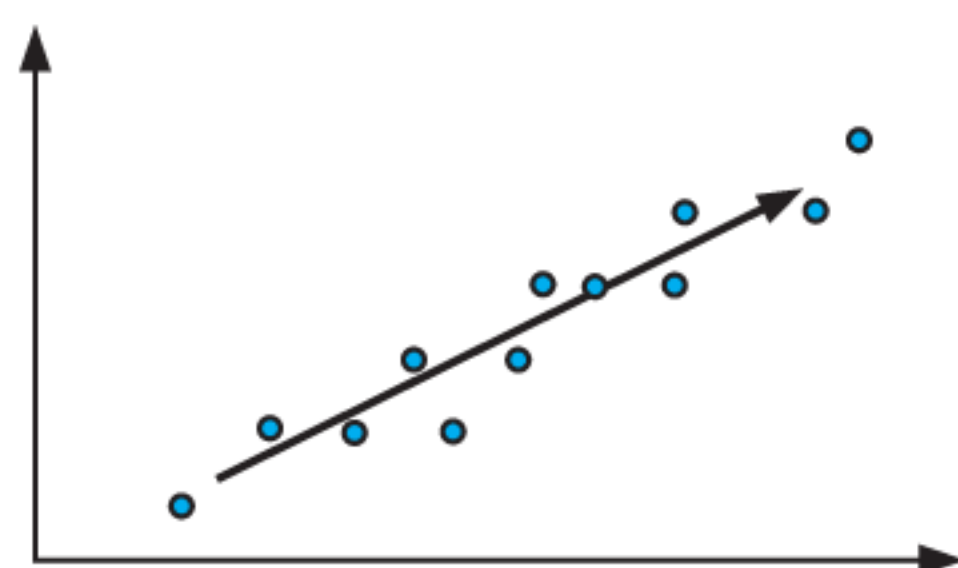


For *randomly scattered* points, with no upward or downward trend, we say there is **no correlation**.

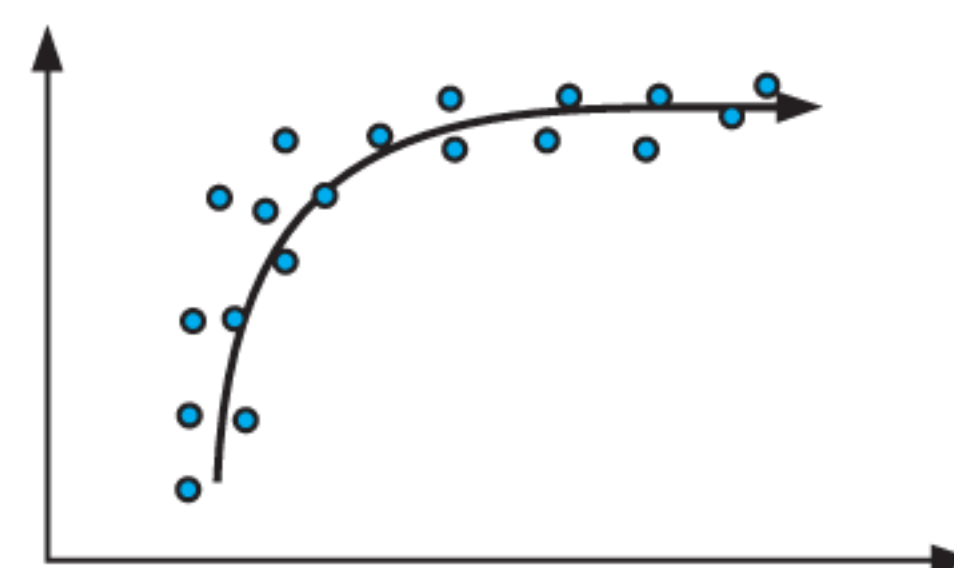
LINEARITY

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

These points are roughly linear.

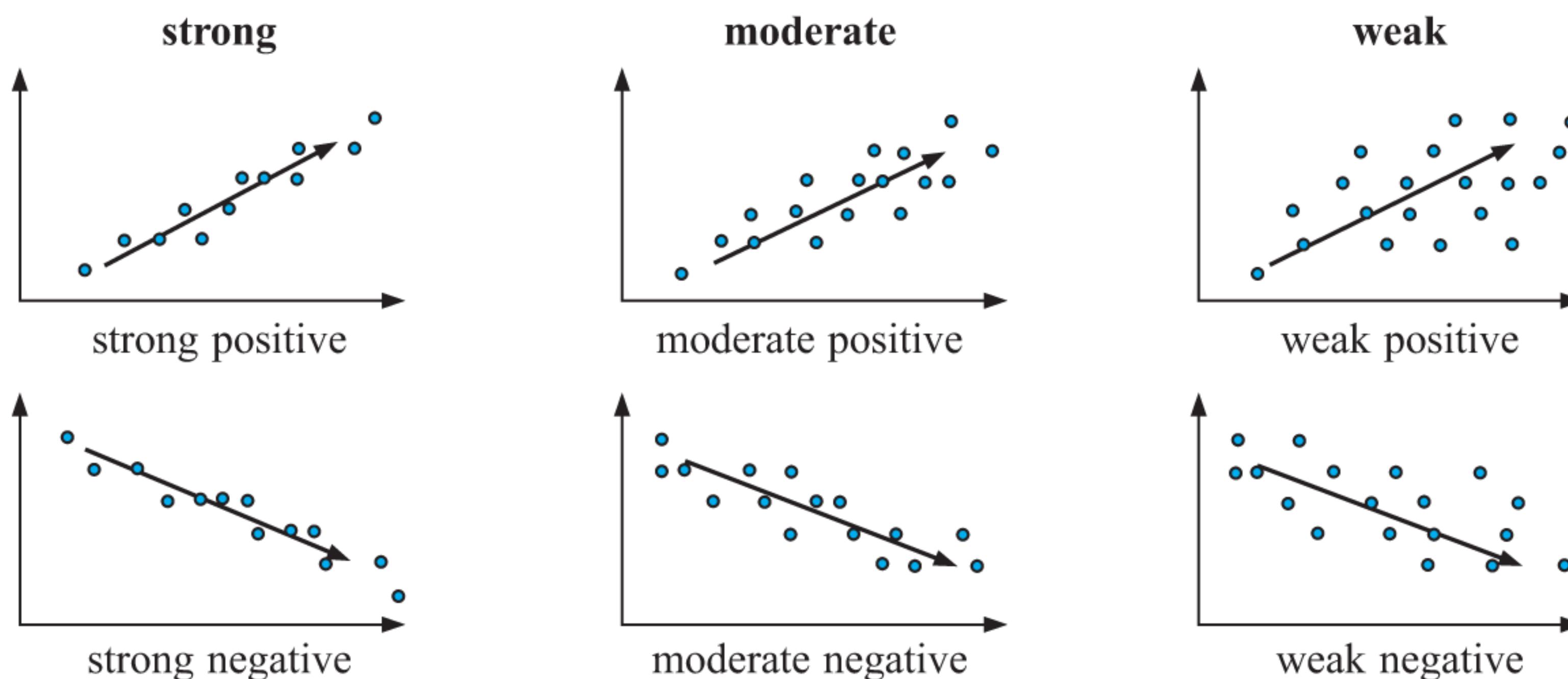


These points do not follow a linear trend.



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**.

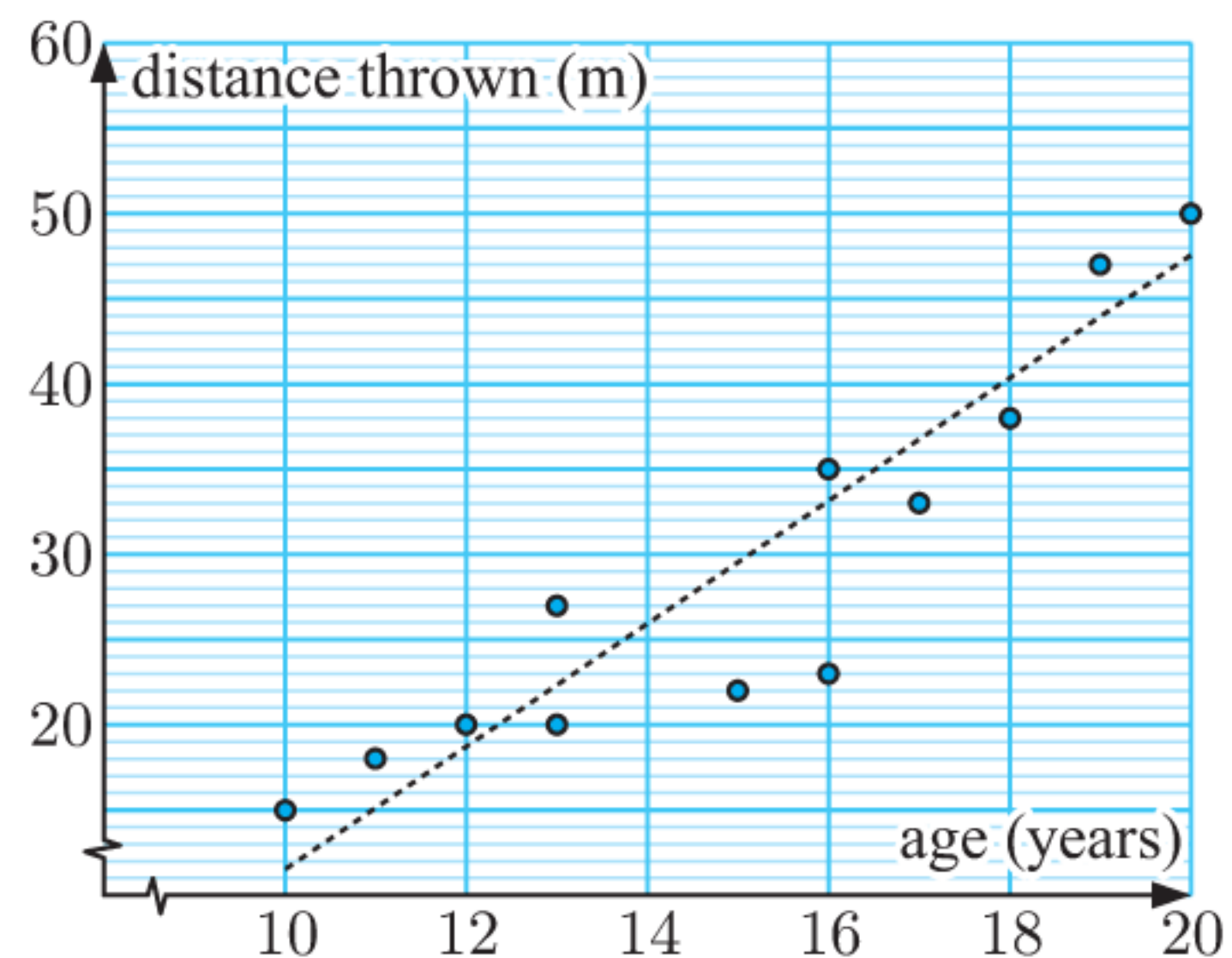
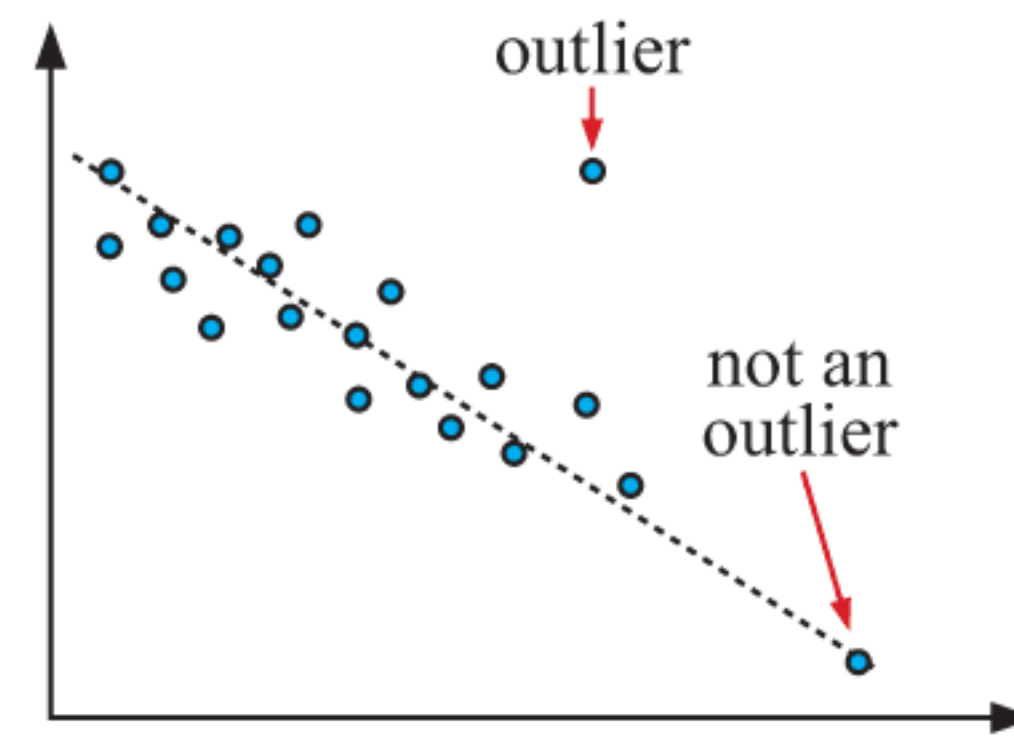


OUTLIERS

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 diagram 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

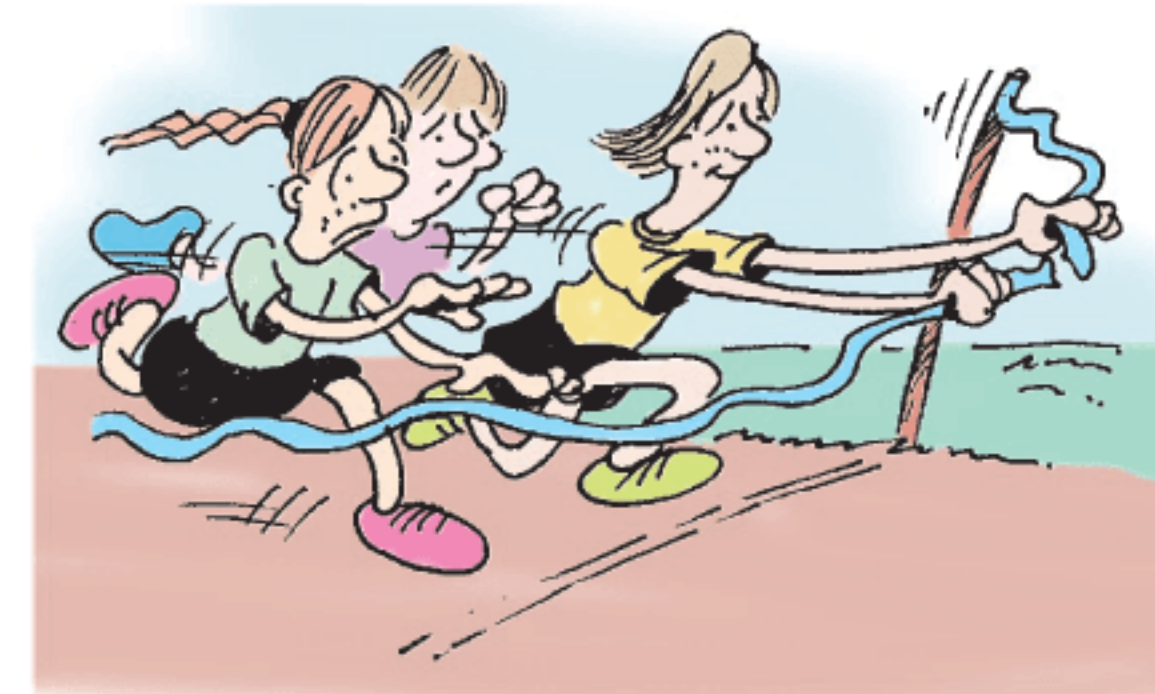
Correlation between two variables does not necessarily mean that one variable *causes* the other.

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.

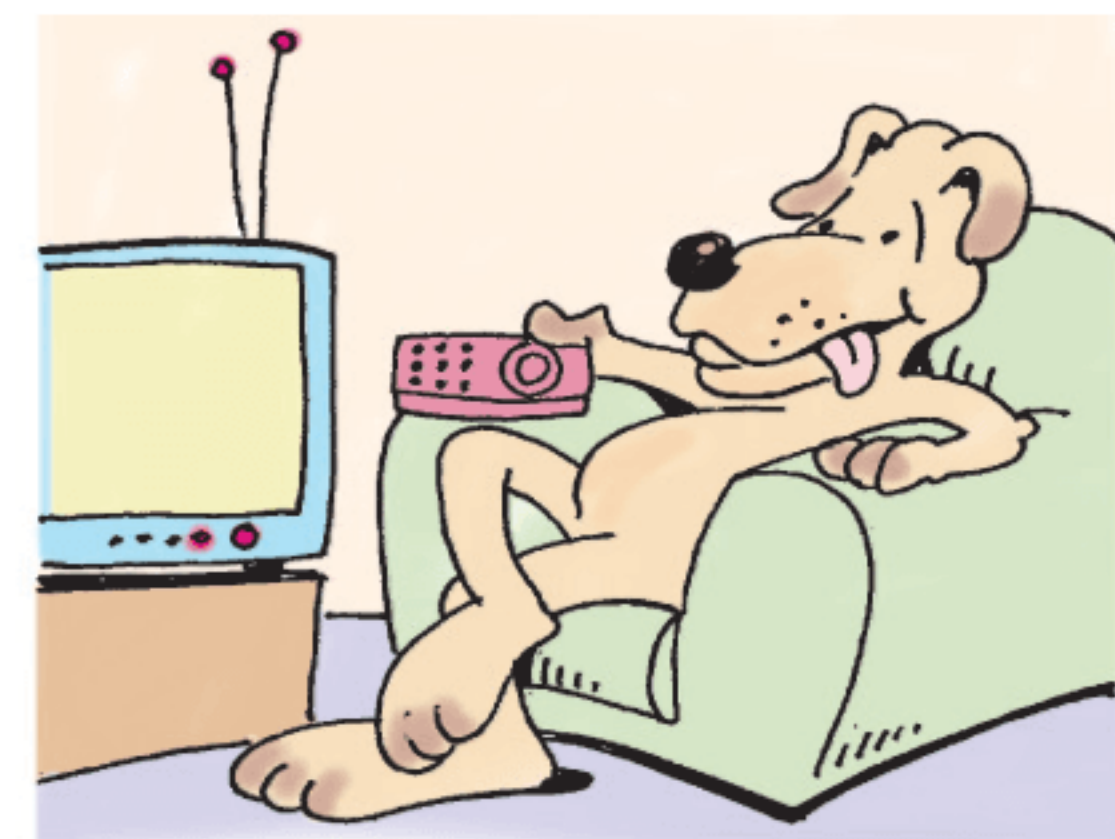
This does *not* mean that short arms cause a reduction in running speed, or that a high running speed causes your arms to grow long.

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 London and the number of stray dogs collected in Boston were recorded over several years. A strong, positive correlation was found between the variables.

Obviously the number of television sets sold in London was not influencing the number of stray dogs collected in Boston. It is coincidental that the variables both increased over this period of time.

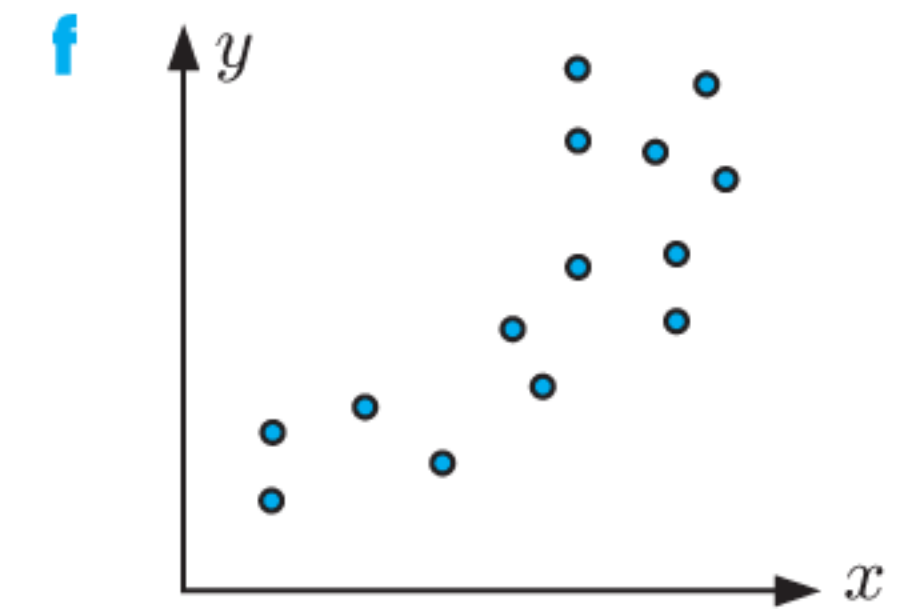
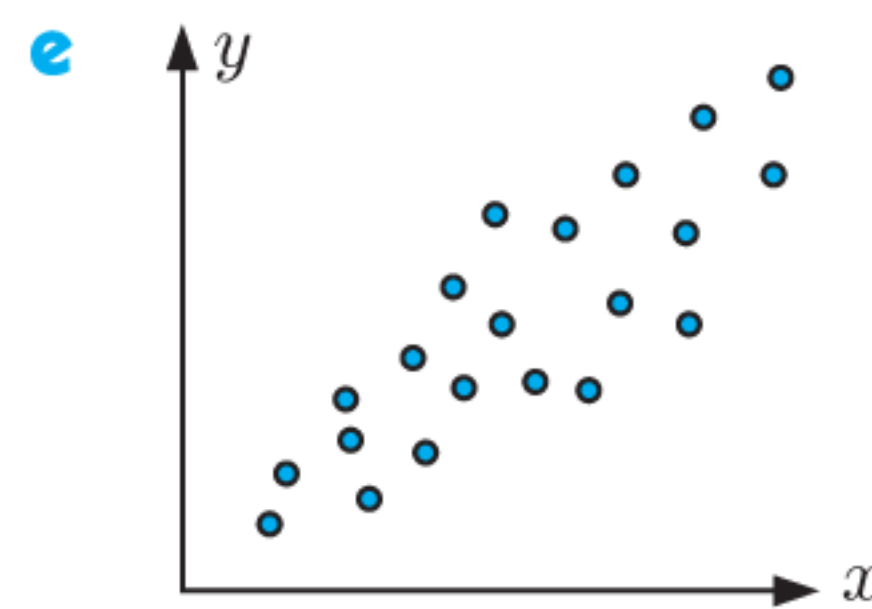
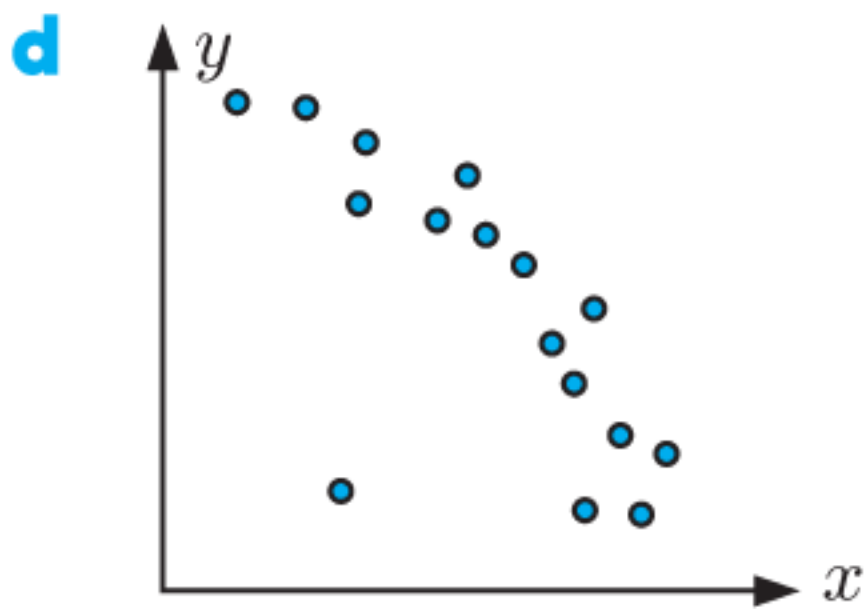
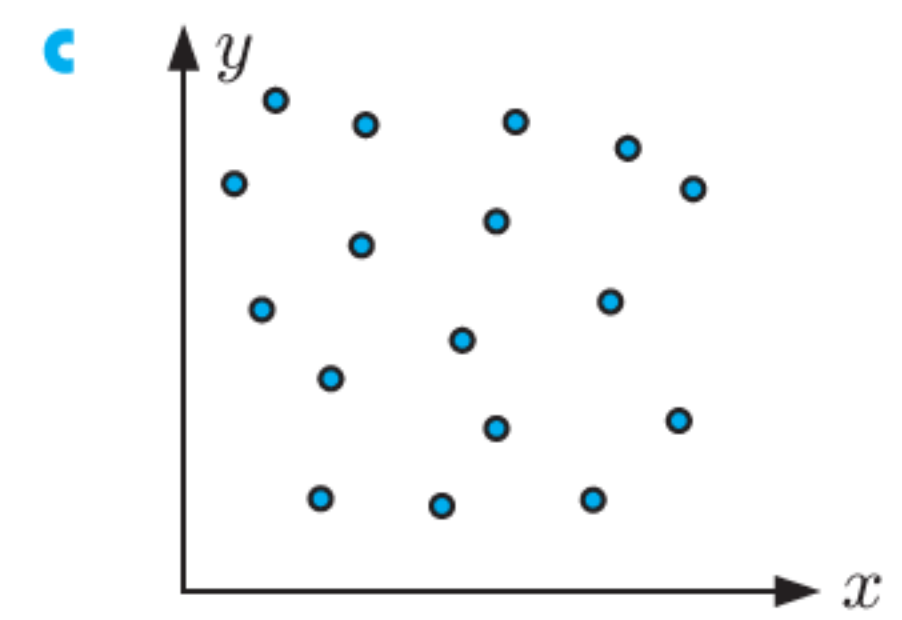
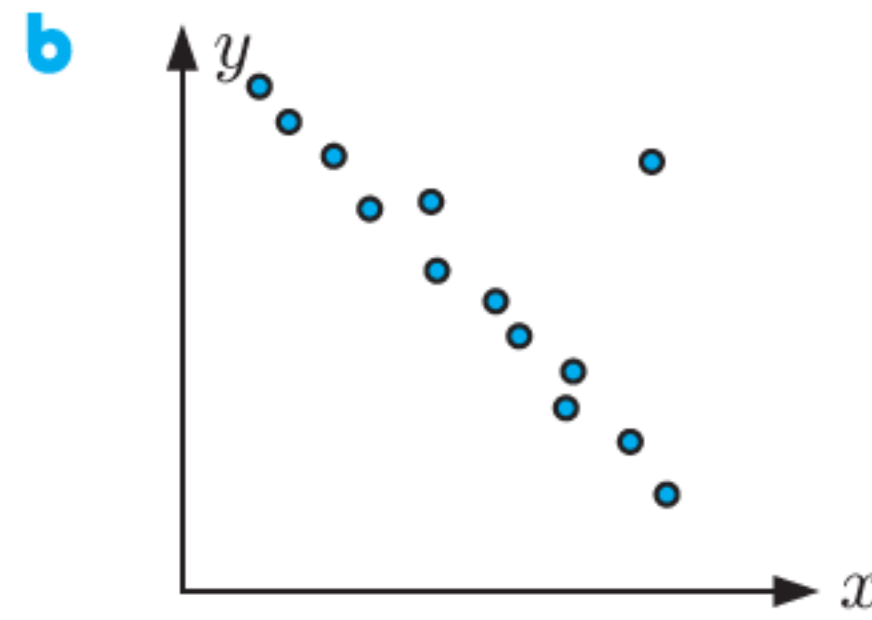
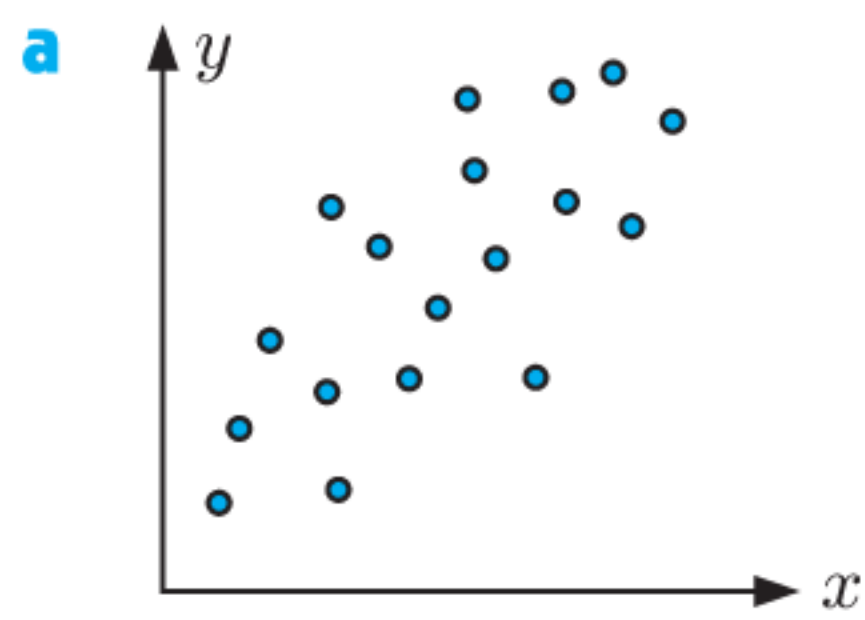


If a change in one variable *causes* a change in the other variable then we say that a **causal relationship** exists between them. In these cases, we can say that the independent variable *explains* the dependent variable. It may be more natural to use the terminology **explanatory variable** and **response variable**.

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

EXERCISE 26A

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



2 Tiffany is a hairdresser. The table below shows the number of hours she worked each day last week, and the number of customers she had.

Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Hours worked	8	4	5	10	8	3	6
Number of customers	9	6	5	12	7	4	5

- a Which is the explanatory variable, and which is the response variable?
- b Draw a scatter diagram of the data.
- c On which two days did Tiffany:
 - i work the same number of hours
 - ii have the same number of customers?
- d Explain why you would expect a positive correlation between the variables.

You can use technology to help draw scatter diagrams.



GRAPHICS CALCULATOR INSTRUCTIONS



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

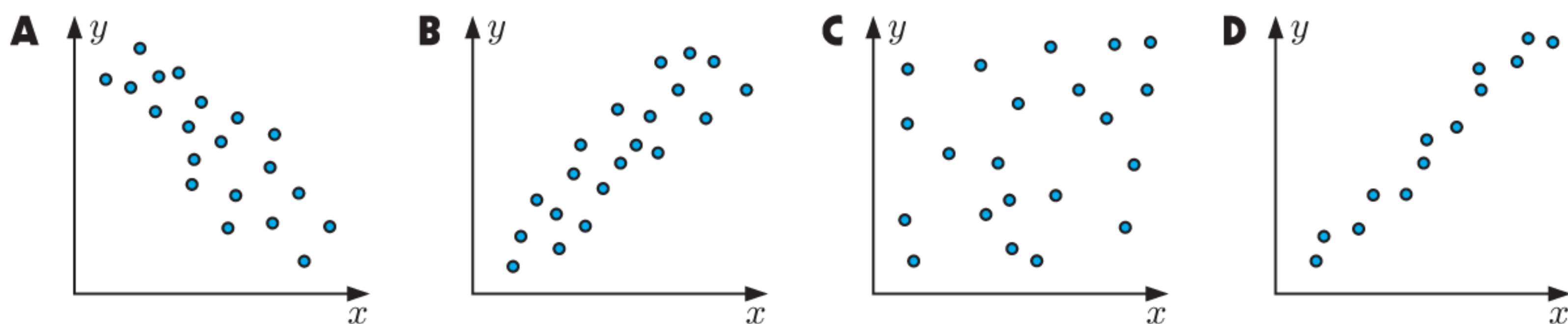
Competitor	P	Q	R	S	T	U	V	W	X	Y
Judge A	5	6.5	8	9	4	2.5	7	5	6	3
Judge B	6	7	8.5	9	5	4	7.5	5	7	4.5

- a Construct a scatter diagram for 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 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? Explain your answer.

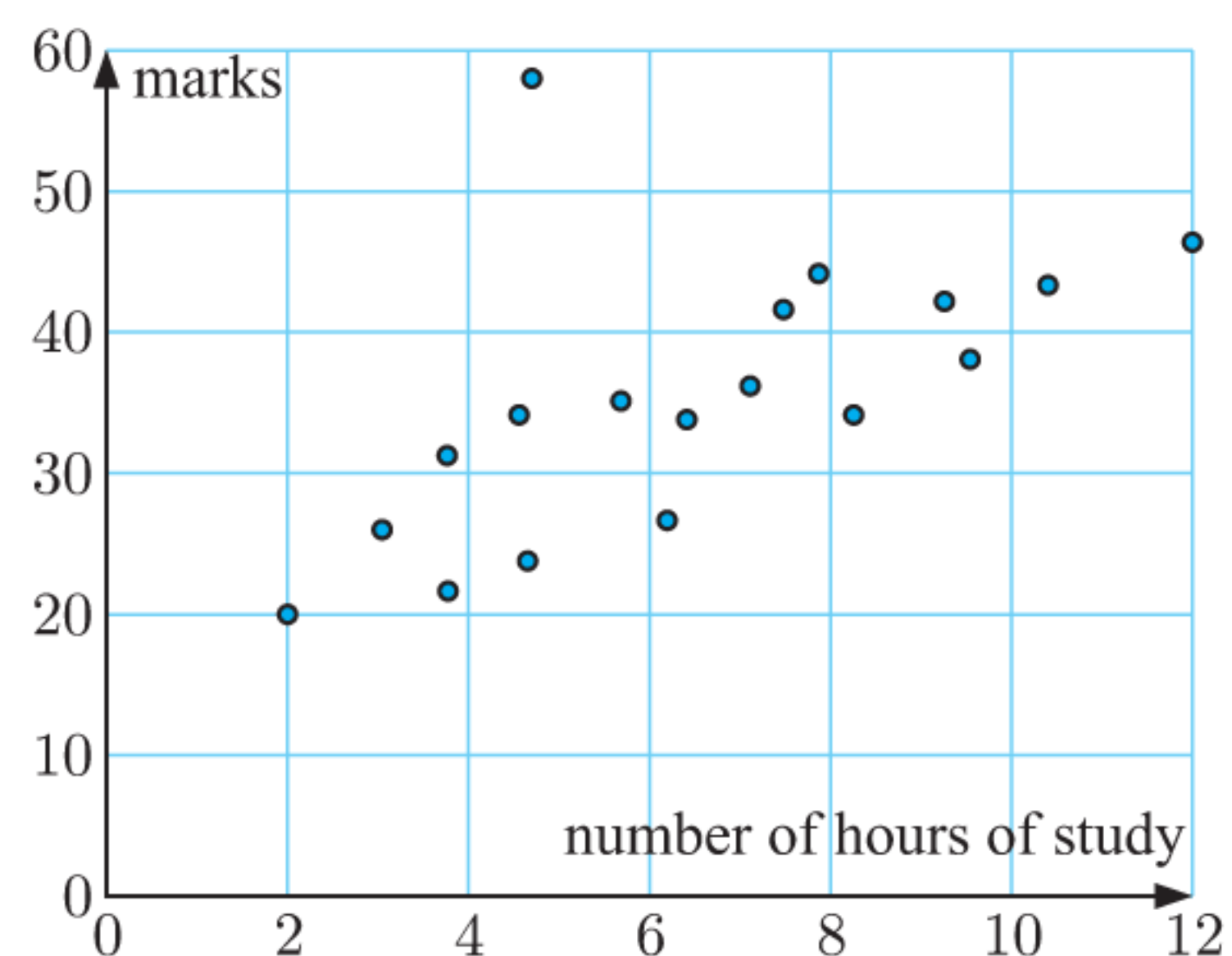
- 4 Paul owns a company which installs industrial air conditioners. The table below shows the number of workers at the company's last 10 jobs, and the time it took to complete the job.

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 diagram to display the data.
- c Describe the relationship between the variables *number of workers* and *time*.
- 5 Choose the scatter diagram 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 The scatter diagram 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.



- 7 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 firefighters 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

PEARSON'S PRODUCT-MOMENT 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 judged how clearly the points formed a linear relationship.

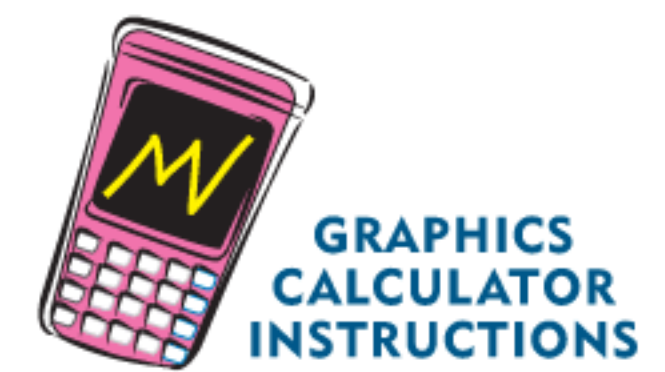
Since this method is *subjective* and relies on the observer's opinion, it is important to get a more precise measure of the strength of linear correlation between the variables. We achieve this using **Pearson's product-moment correlation coefficient** r .

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

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

where \bar{x} and \bar{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.



HISTORICAL NOTE

Karl Pearson (1857 - 1936) was an English statistician who developed the product-moment correlation coefficient together with his academic advisor **Sir Francis Galton**.

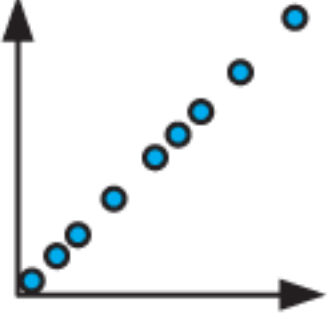
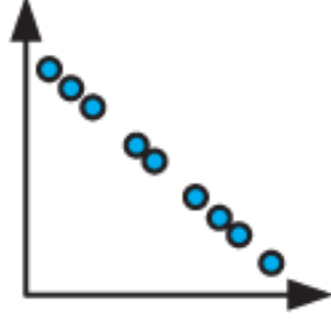
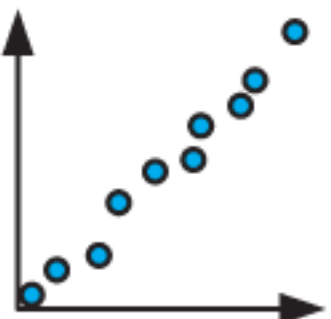
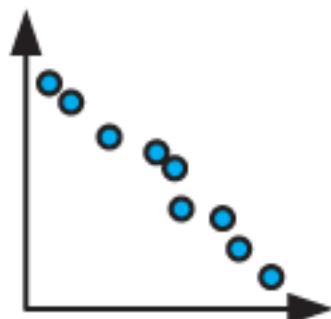
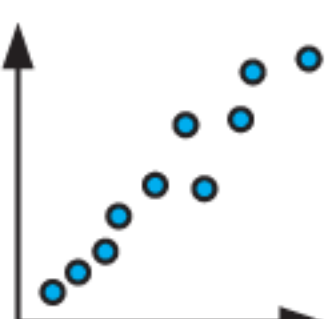
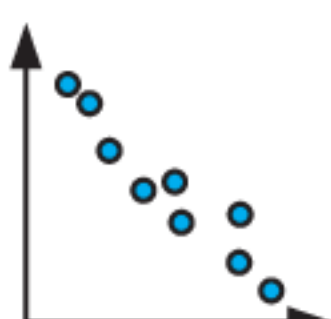
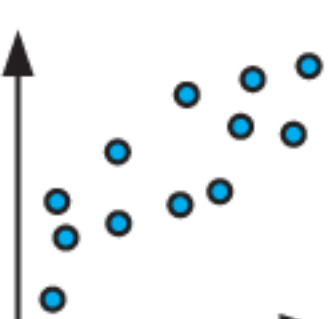
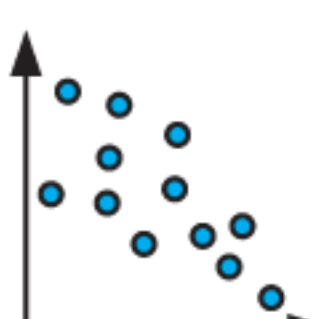
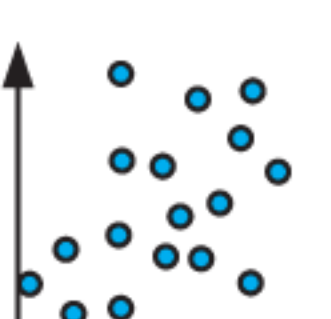
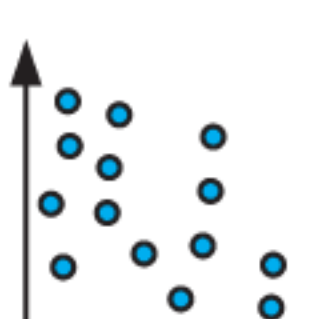
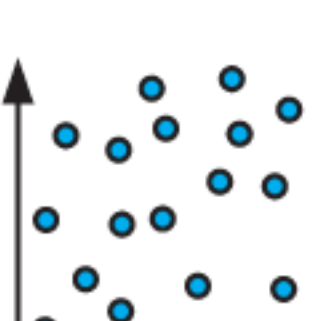
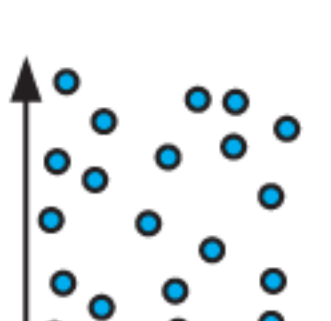
Pearson made many other contributions to statistics including the use of histograms in exploratory data analysis, parameter estimation, and hypothesis testing.

He is considered a key figure in the development of mathematical statistics.

PROPERTIES OF PEARSON'S PRODUCT-MOMENT 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 .

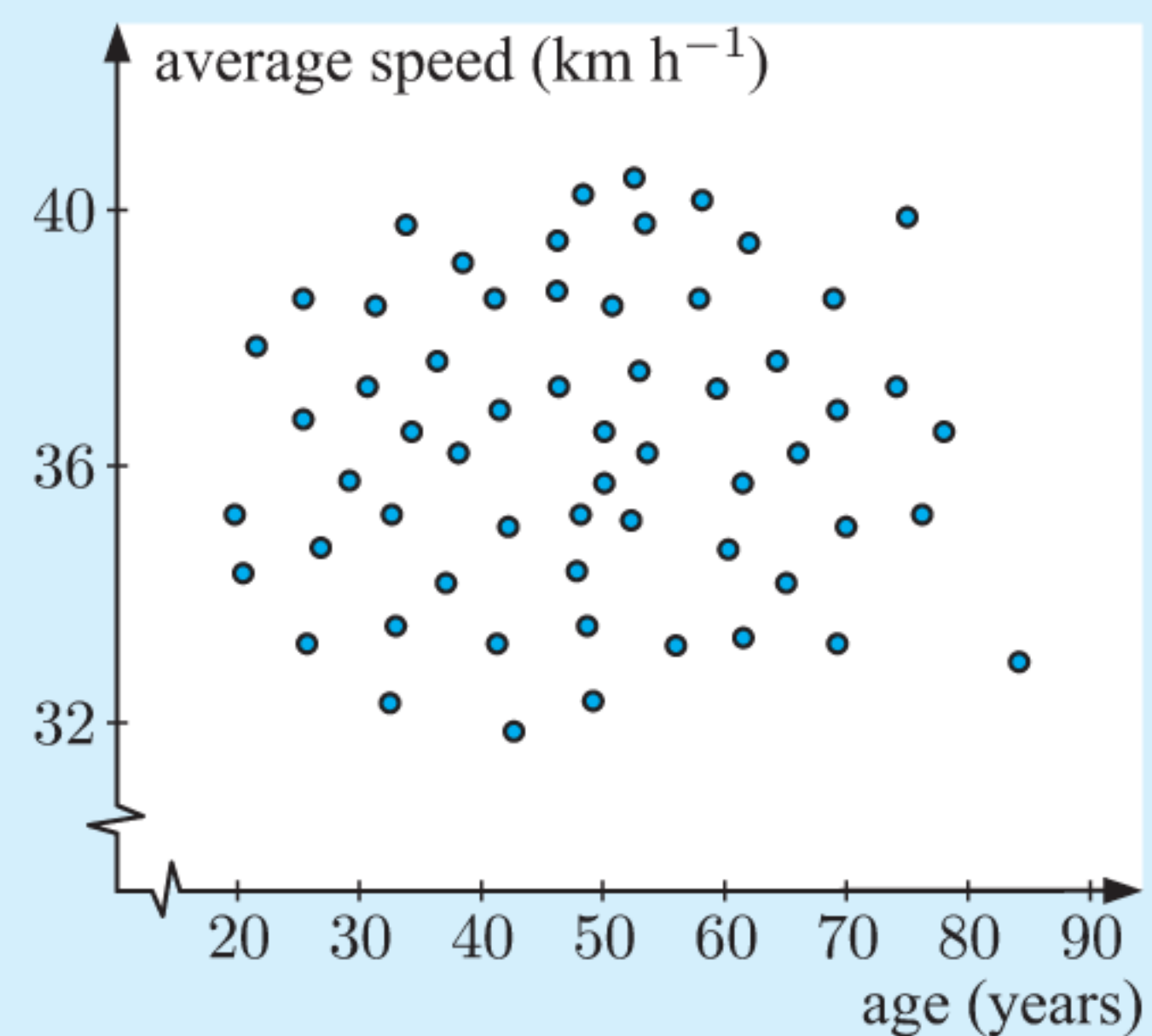
Positive correlation			Negative correlation		
$r = 1$	perfect positive correlation		$r = -1$	perfect negative correlation	
$0.95 \leq r < 1$	very strong positive correlation		$-1 < r \leq -0.95$	very strong negative correlation	
$0.87 \leq r < 0.95$	strong positive correlation		$-0.95 < r \leq -0.87$	strong negative correlation	
$0.7 \leq r < 0.87$	moderate positive correlation		$-0.87 < r \leq -0.7$	moderate negative correlation	
$0.5 \leq r < 0.7$	weak positive correlation		$-0.7 < r \leq -0.5$	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 diagram. The r -value for this association is $+0.027$. Describe the association.



Since $0 < r < 0.5$, there is a very weak positive correlation between the two variables. We observe this in the graph as the points are randomly scattered.

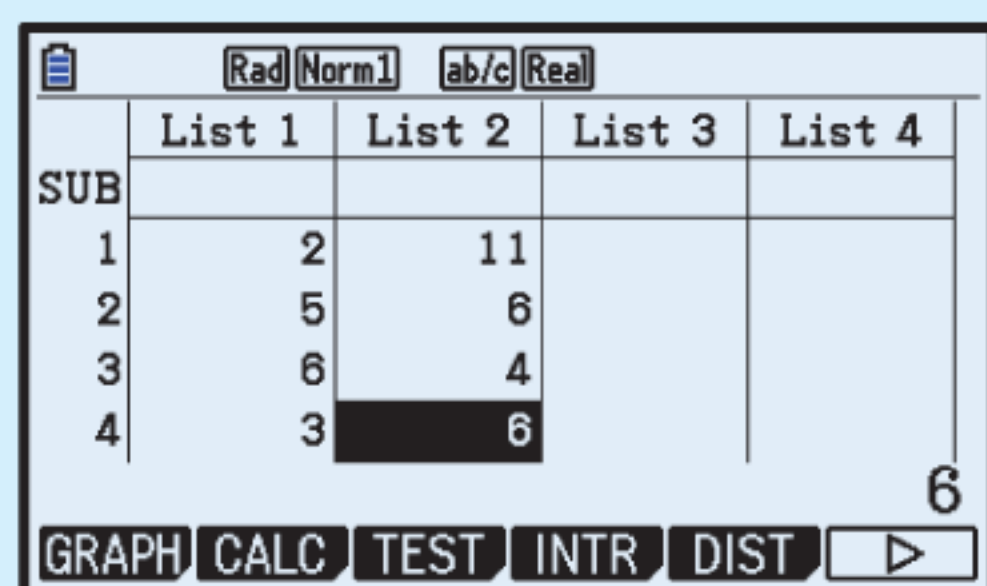
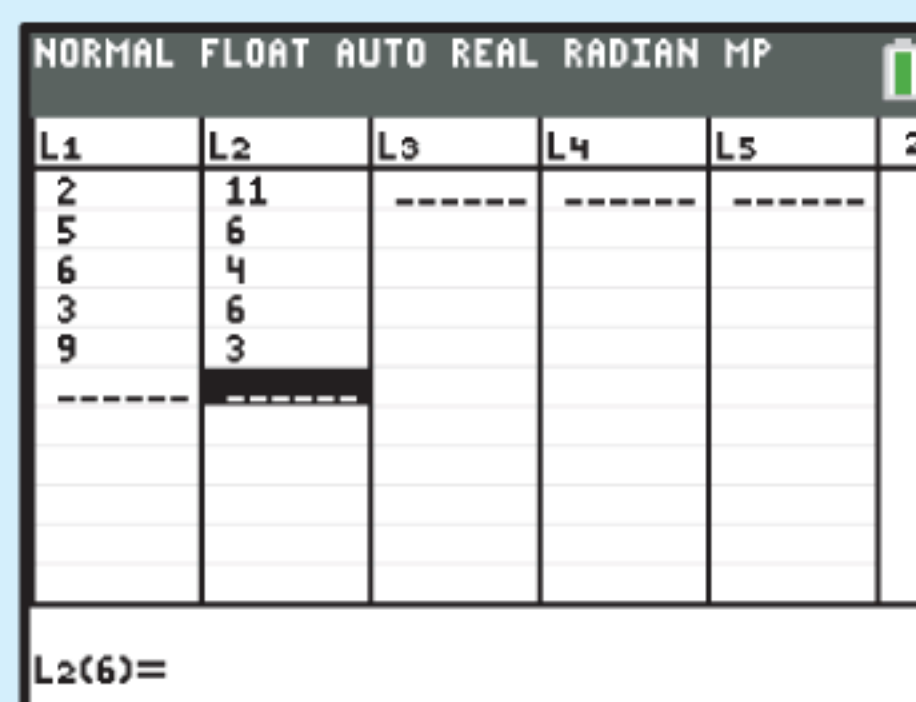
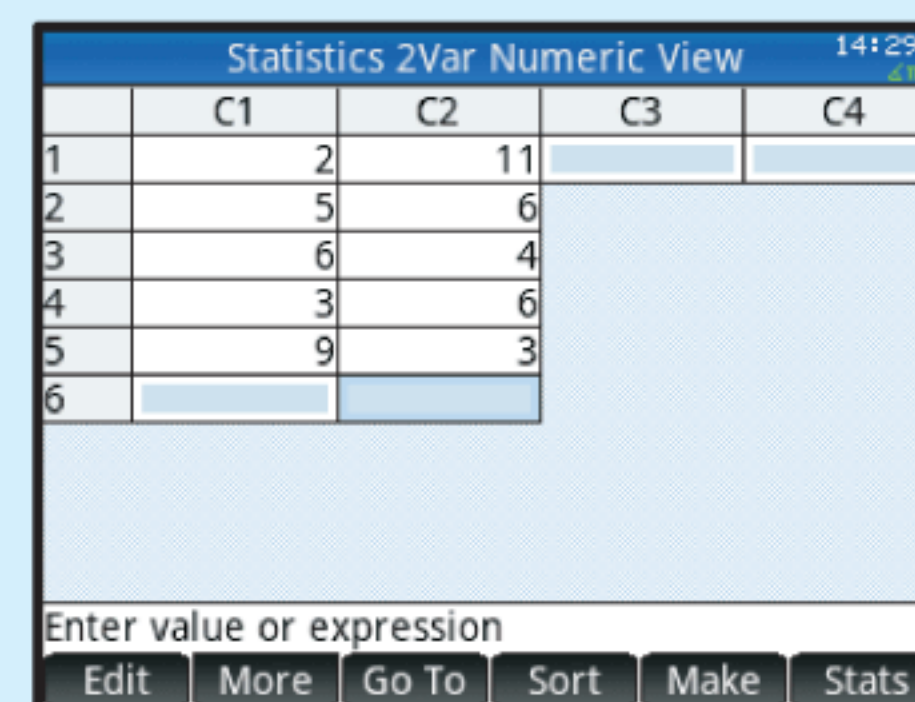
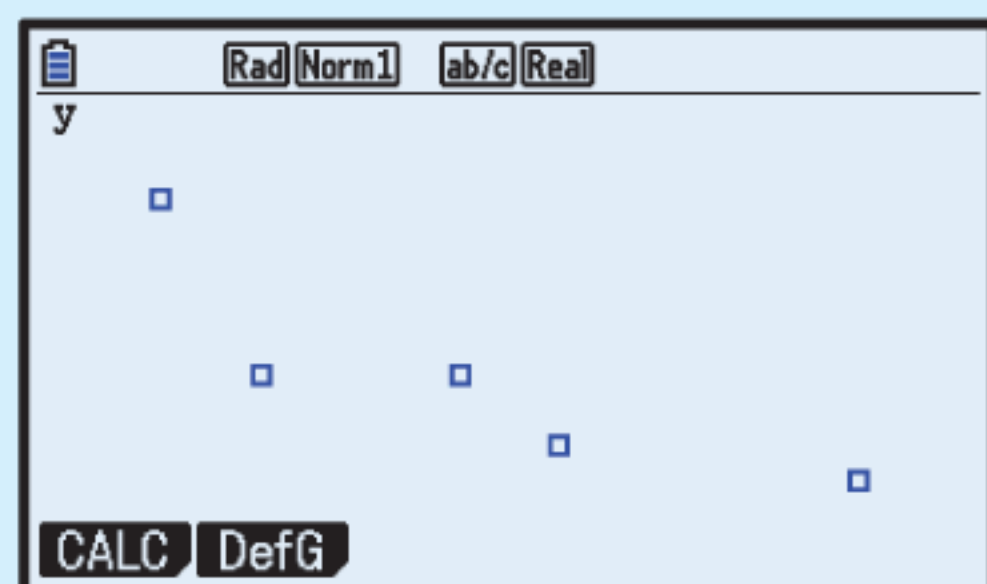
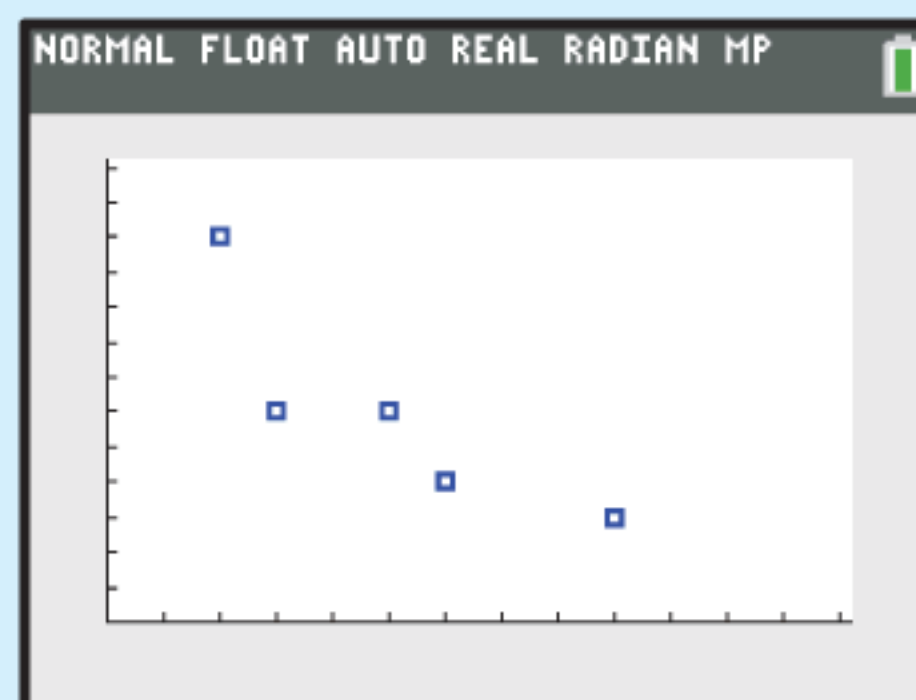
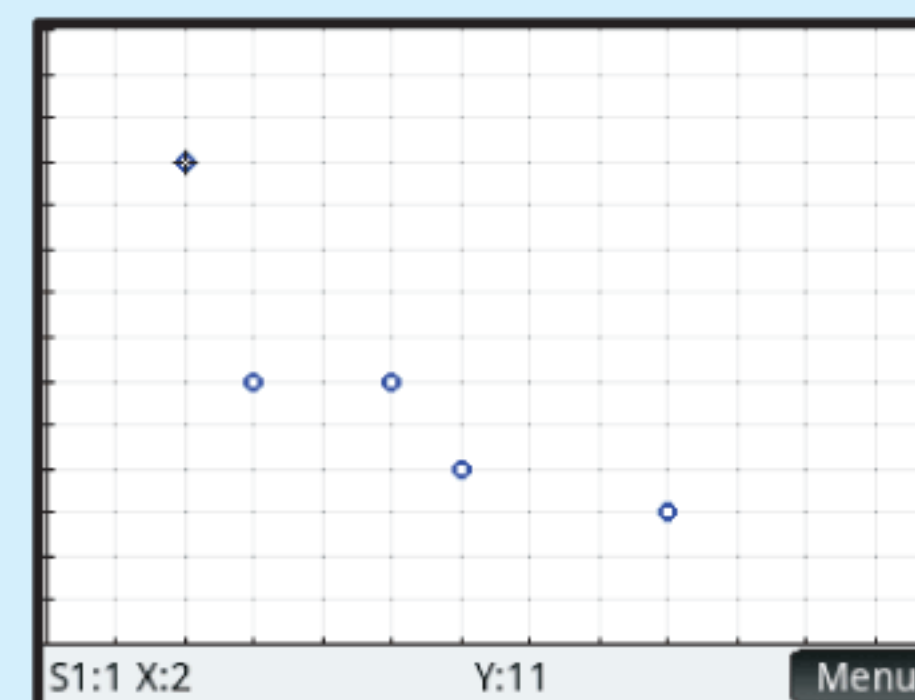
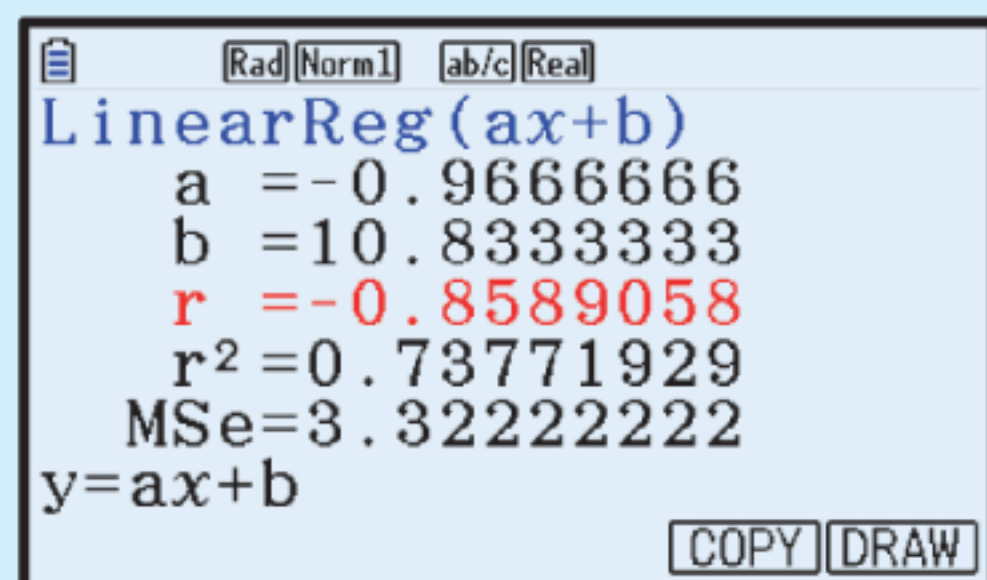
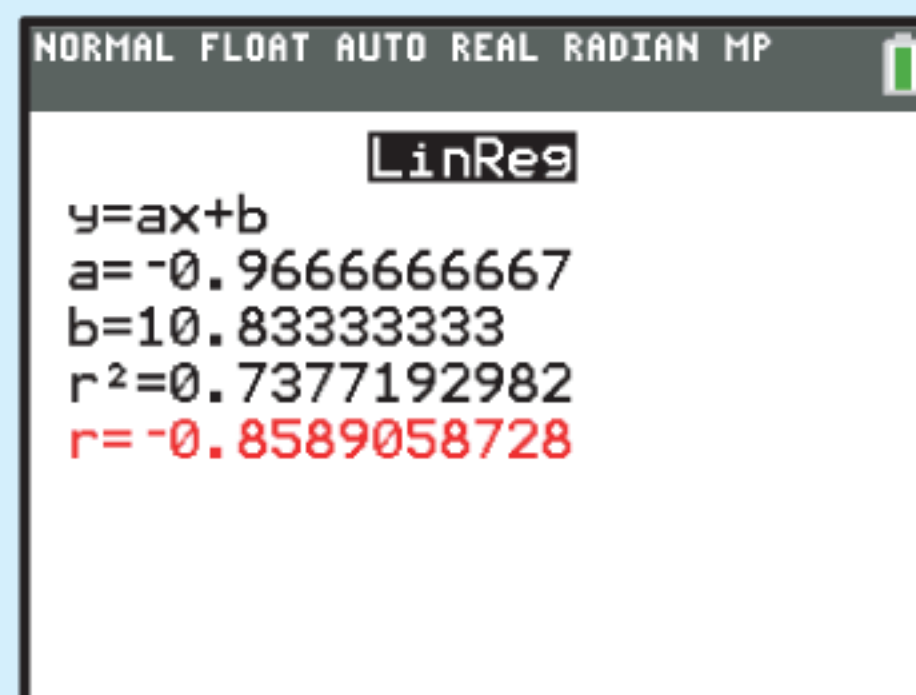
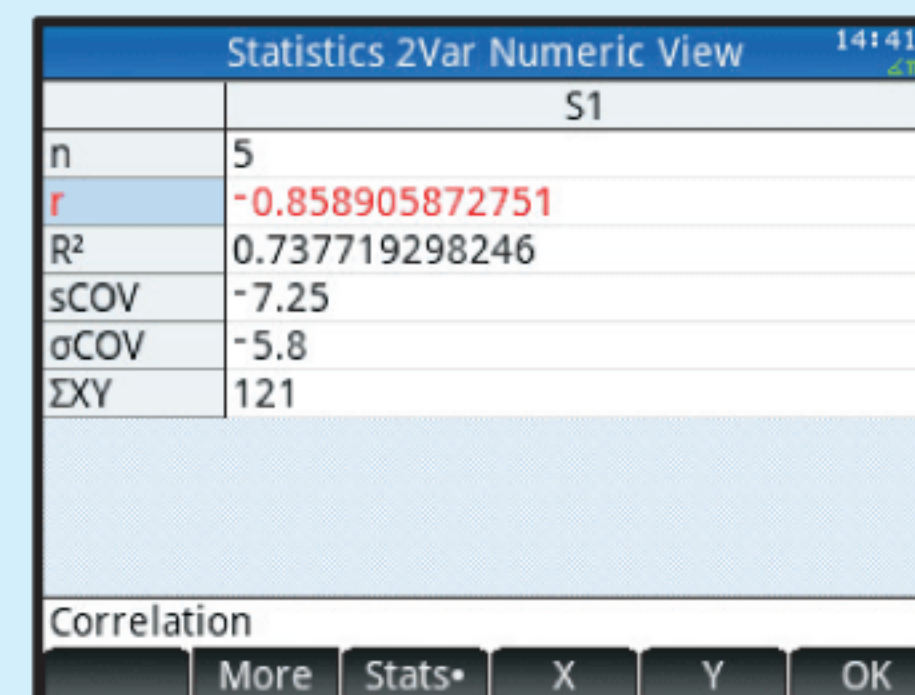
Example 2


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.

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

- Draw a scatter diagram for the data.
- Determine the correlation coefficient r .
- Describe the correlation between the *quantity of chemical* and the *number of surviving beetles*.

We first enter the data into separate lists:

Casio fx-CG50

TI-84 Plus CE

HP Prime

a
Casio fx-CG50

TI-84 Plus CE

HP Prime

b
Casio fx-CG50

TI-84 Plus CE

HP Prime


So, $r \approx -0.859$.

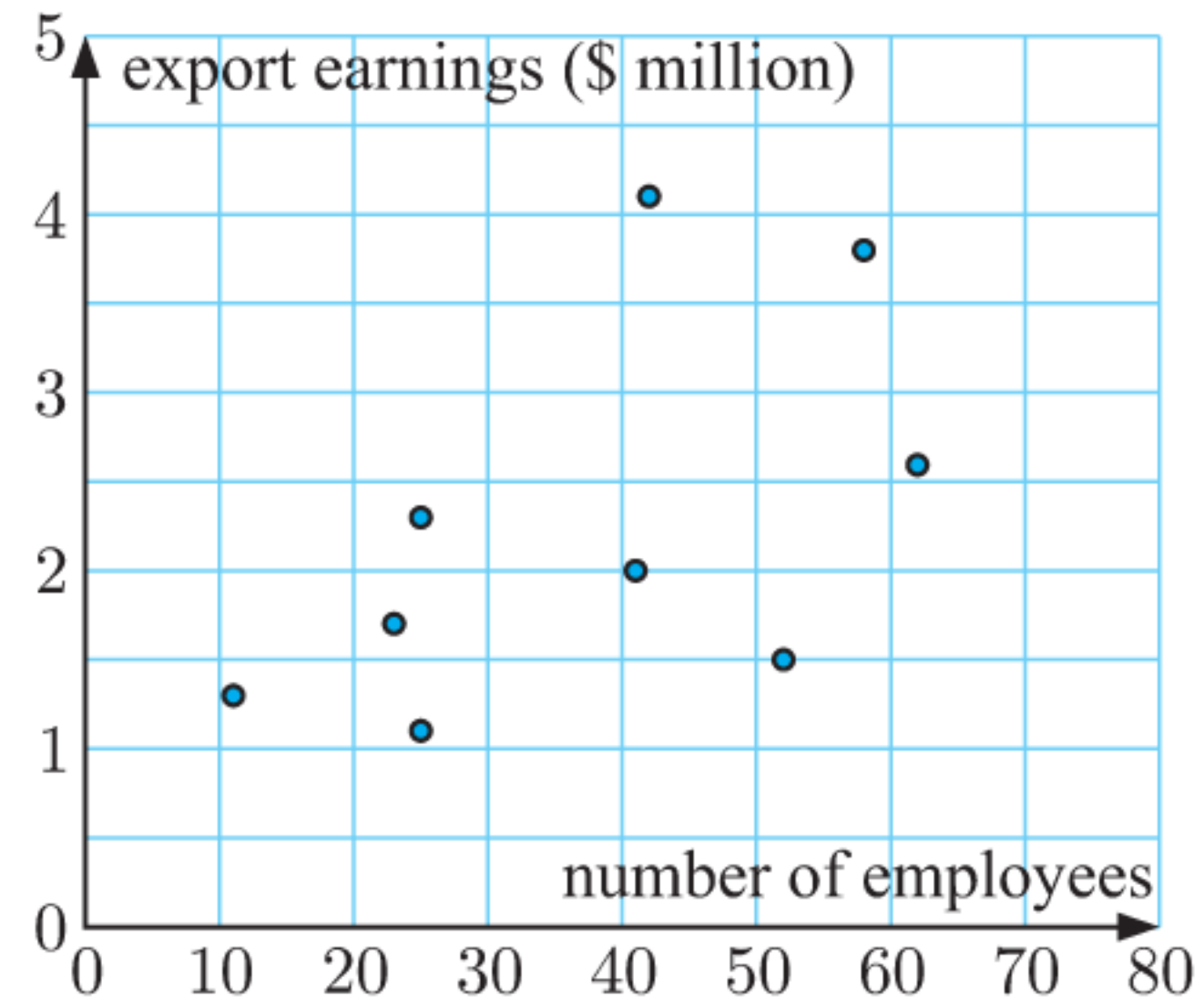
- There is a moderate negative correlation between the *quantity of chemical used* and the *number of surviving beetles*.

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

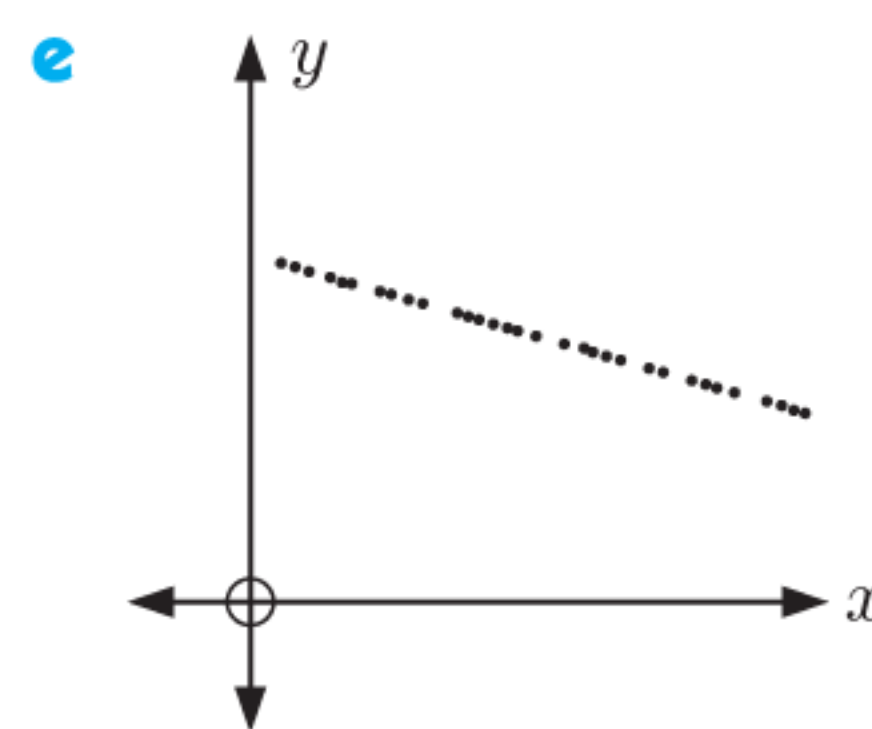
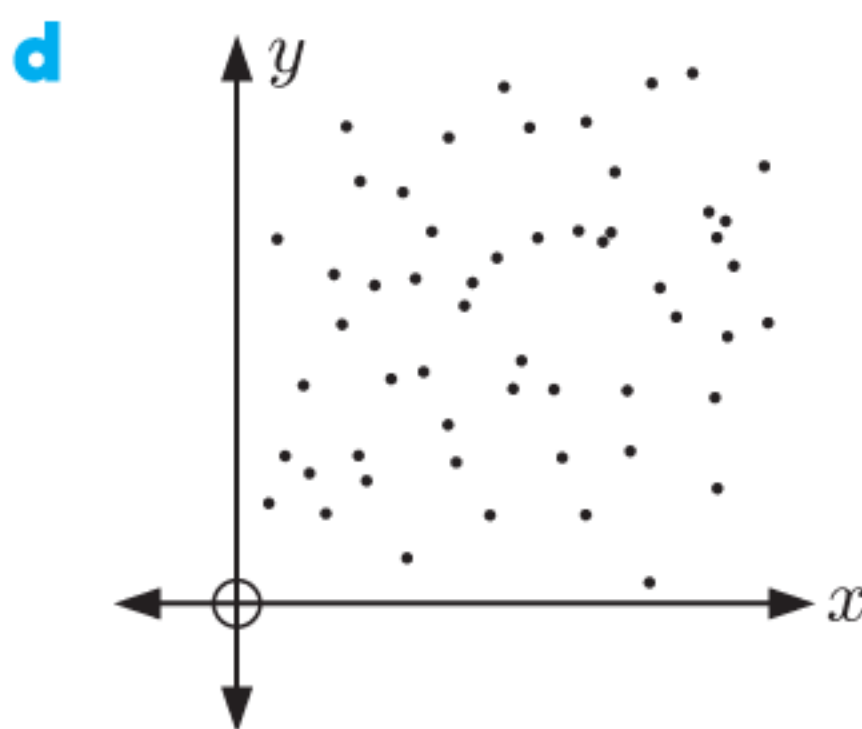
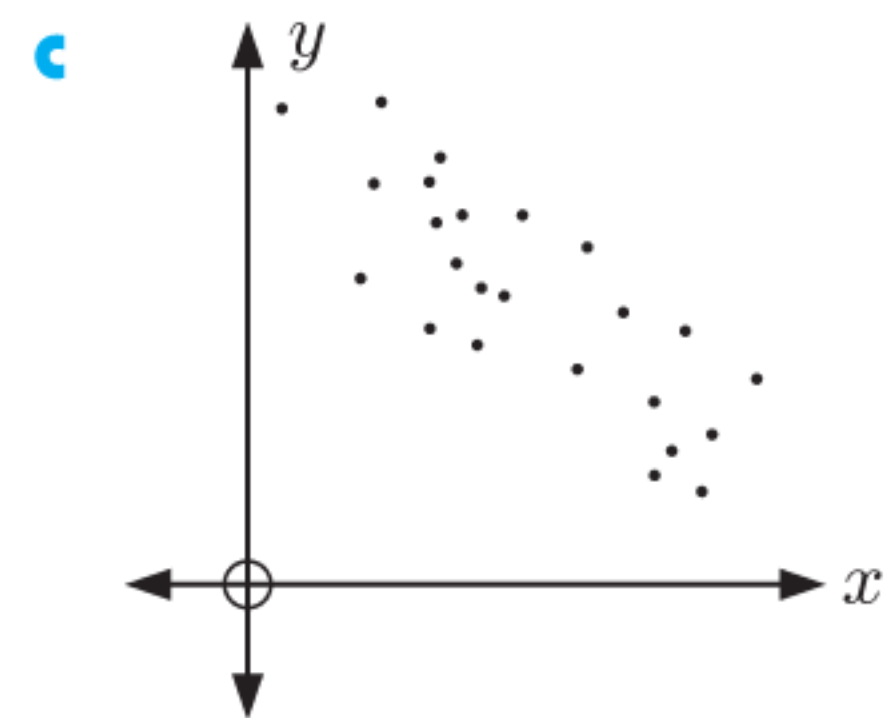
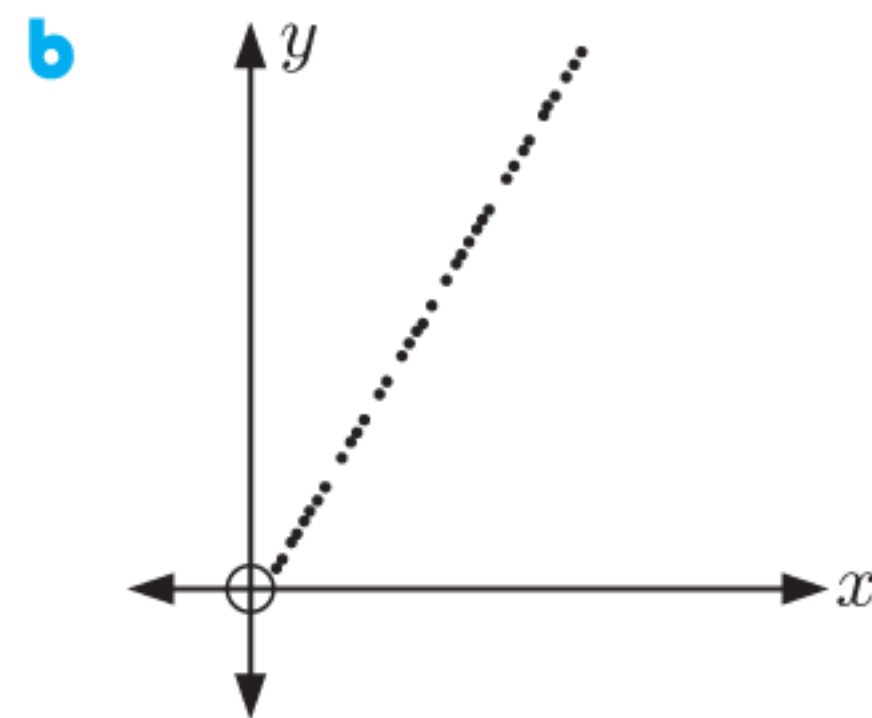
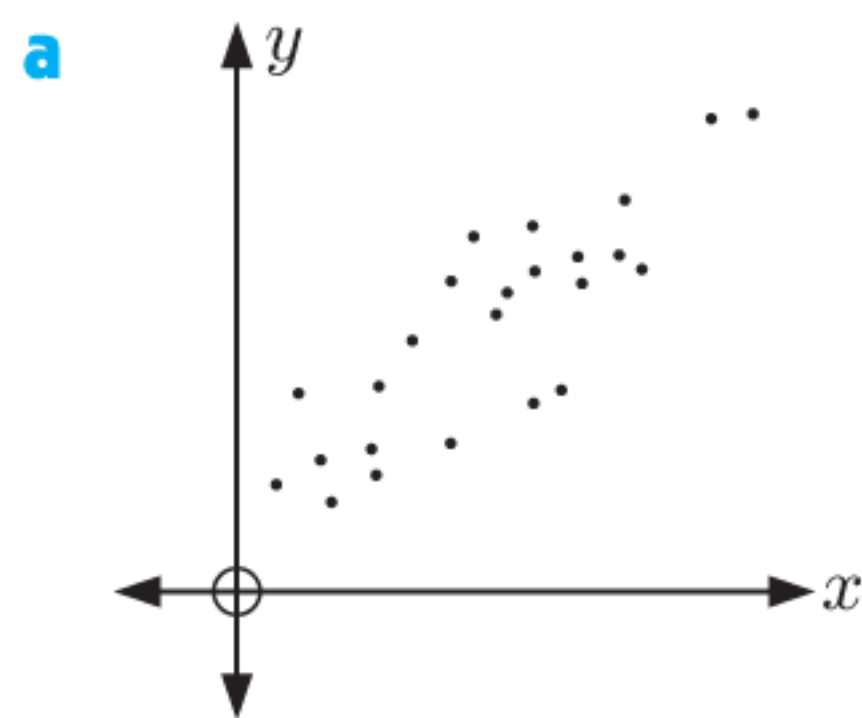
EXERCISE 26B

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

Describe the association between the variables.



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



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

3 For each of the following data sets:

- i** Draw a scatter diagram for the data.
- ii** Calculate Pearson's product-moment 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 was asked how many phone calls and text messages they received the previous day. The results are shown alongside.

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

- a** Draw a scatter diagram 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 708.
- Calculate r for the data.
 - Hence describe the association between the variables.
- 6 Jill does her washing every Saturday and hangs her clothes out to dry. 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:

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

- Draw a scatter diagram for the data.
 - Calculate r .
 - 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.

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

- Draw a scatter diagram for the data.
 - Calculate r .
 - Identify the outlier in the data.
 - It was found that the outlier was due to an error in the data collection process.
 - Recalculate r with the outlier removed.
 - Describe the relationship between the variables.
 - Discuss the effect of removing the outlier on the value of r .
 - 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.

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

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

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

Higher latitudes occur near the poles. Lower latitudes occur near the equator.



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 below:

Student	Height (cm)	Foot length (cm)



- 3 Use technology to draw a scatter diagram for the data.
- 4 Calculate Pearson's product-moment 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?

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 correlation between the variables is strong. There is no fixed rule, but we suggest that a line of best fit is not appropriate if $|r| < 0.85$.

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 \bar{x} , and the mean of the y values \bar{y} .

Step 2: Mark the **mean point** (\bar{x}, \bar{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.

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

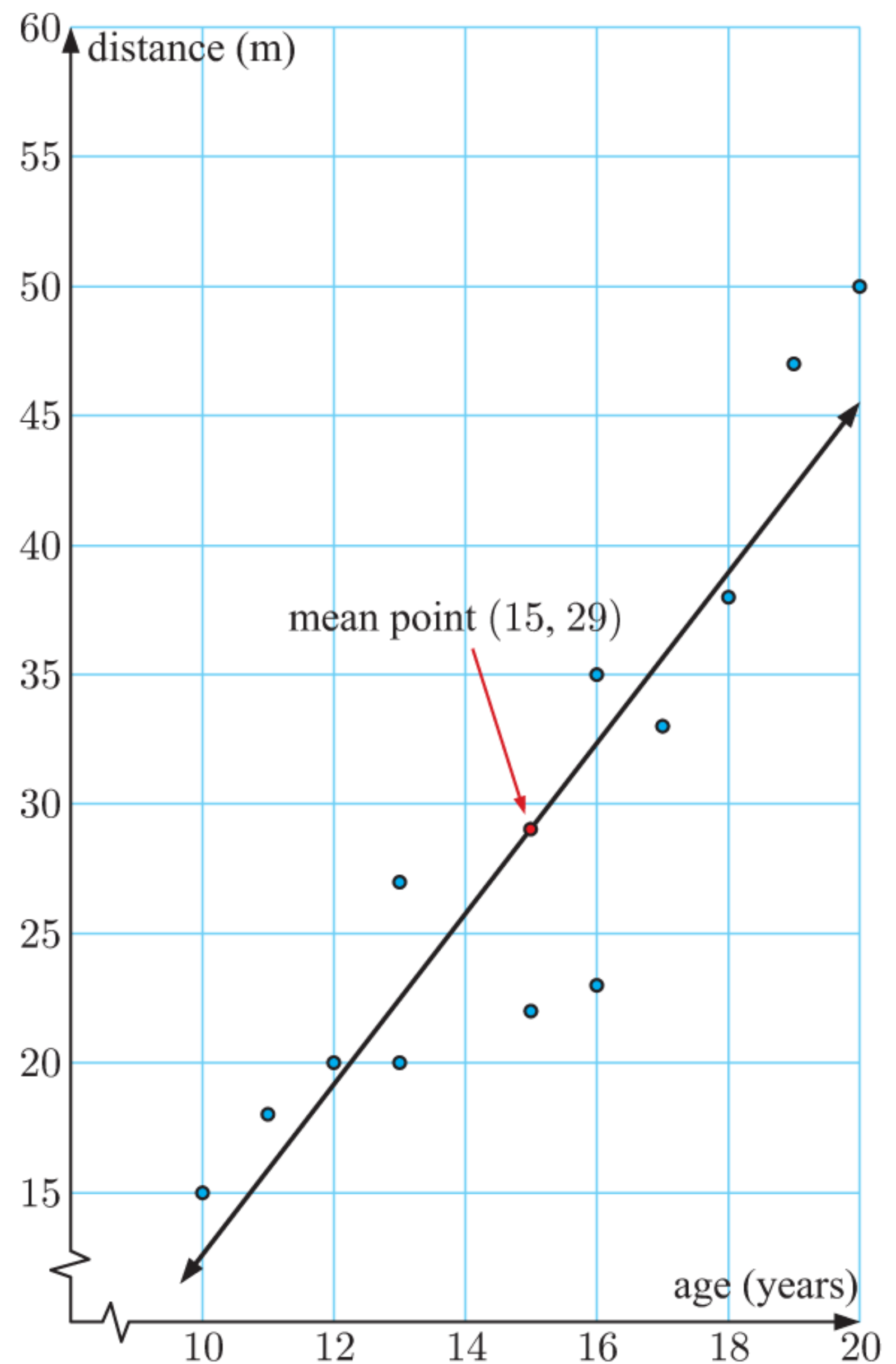
We have seen that there is a strong positive linear correlation between *age* and *distance thrown*.

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

The mean age is 15 years and the mean distance thrown is 29 m. 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.

We draw the line through the mean point so it follows the trend of the data and there are about the same number of points above the line as below the line.

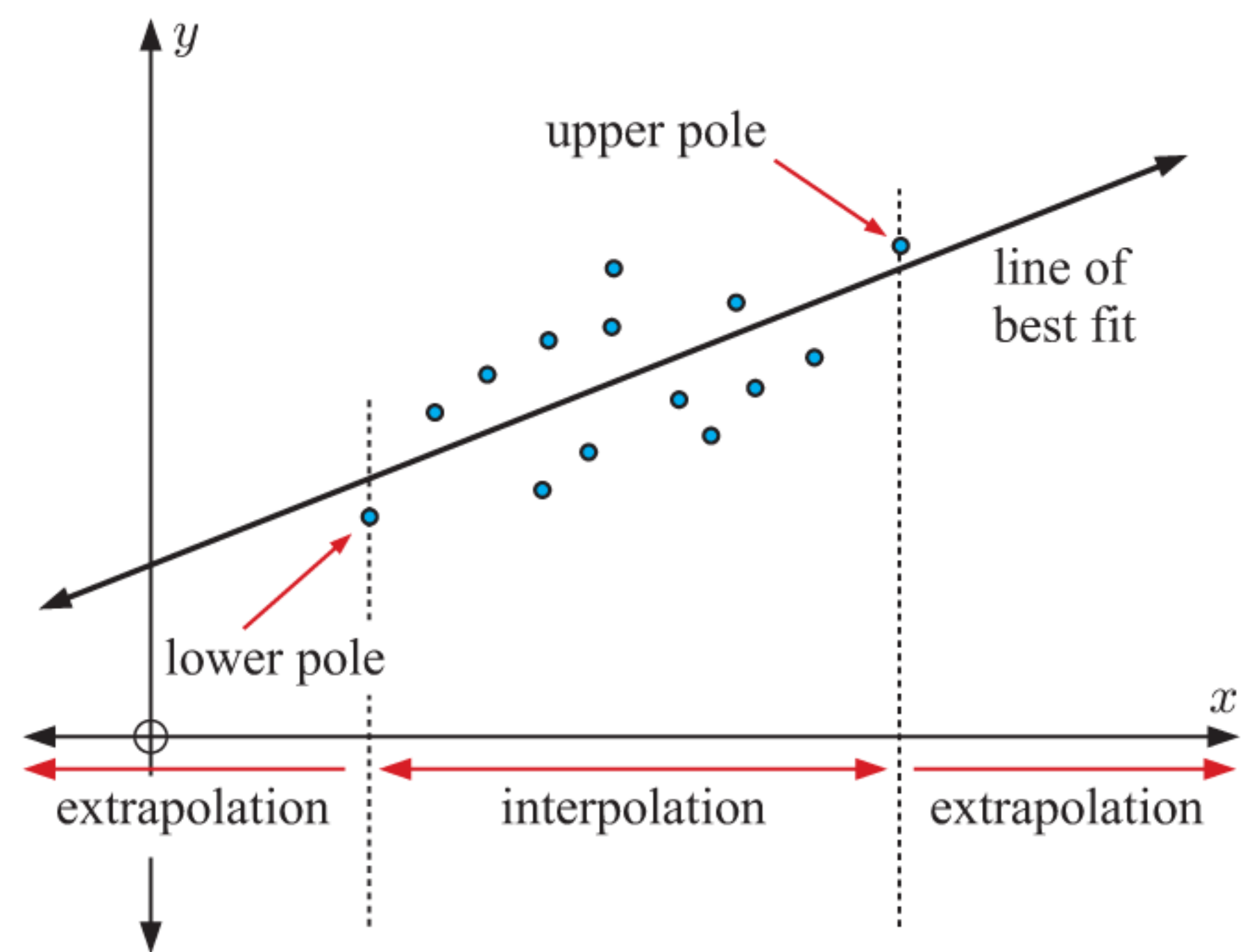


INTERPOLATION AND EXTRAPOLATION

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

The line of best fit for the data is also drawn on the scatter diagram. We can use this line to 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.



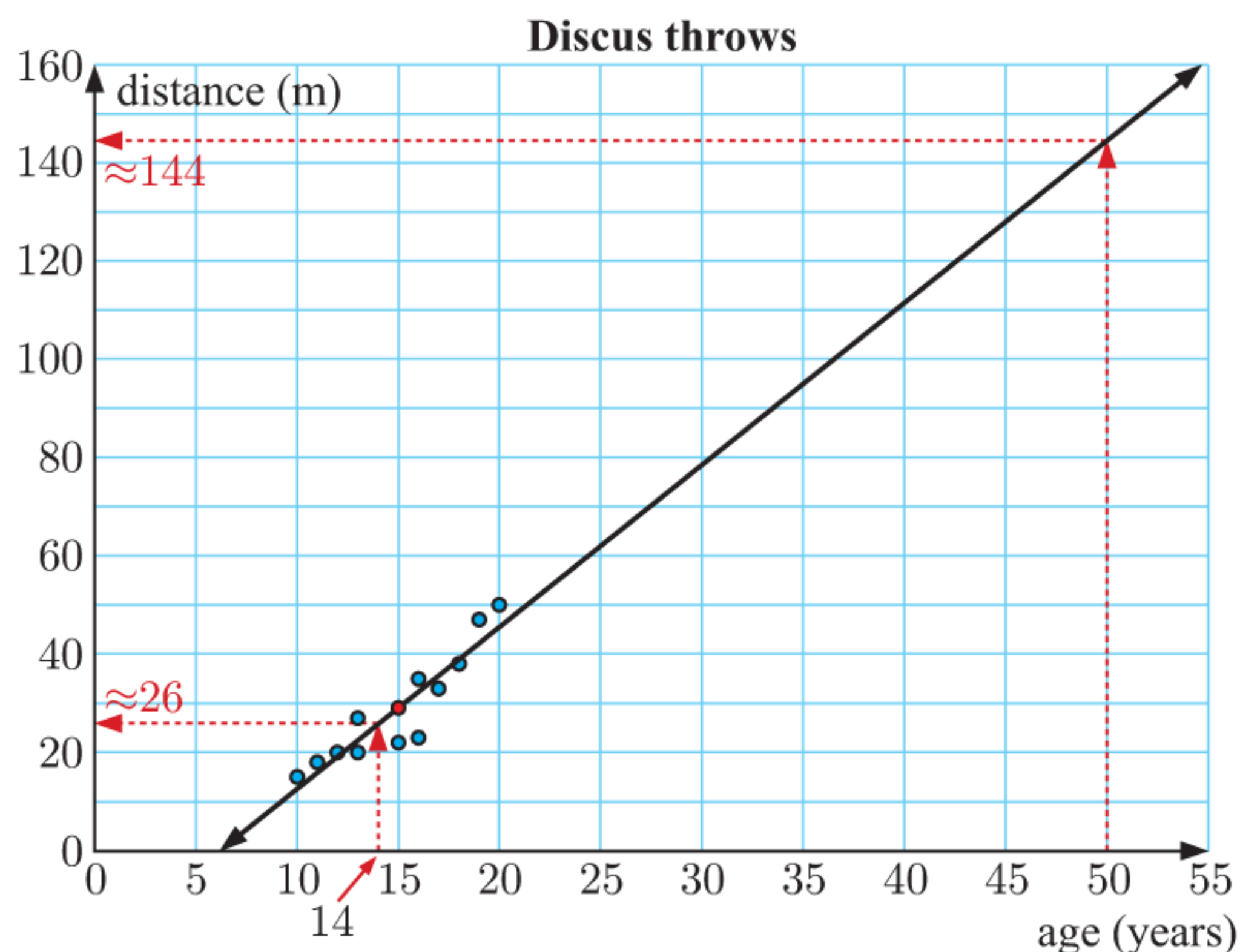
The accuracy of an interpolation depends on how well the linear model fits the data. This can be gauged by the correlation coefficient and by ensuring that the data is randomly scattered around the line of best fit.

The accuracy of an extrapolation depends not only on how well the model fits, 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**. It can be used to predict the distance a discus will be thrown by an athlete of a particular age.

The age 14 is within the range of ages in the original data, 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 might throw the discus 144 m. This is almost twice the current world record of 76.8 m, so it would clearly be an unreasonable prediction.

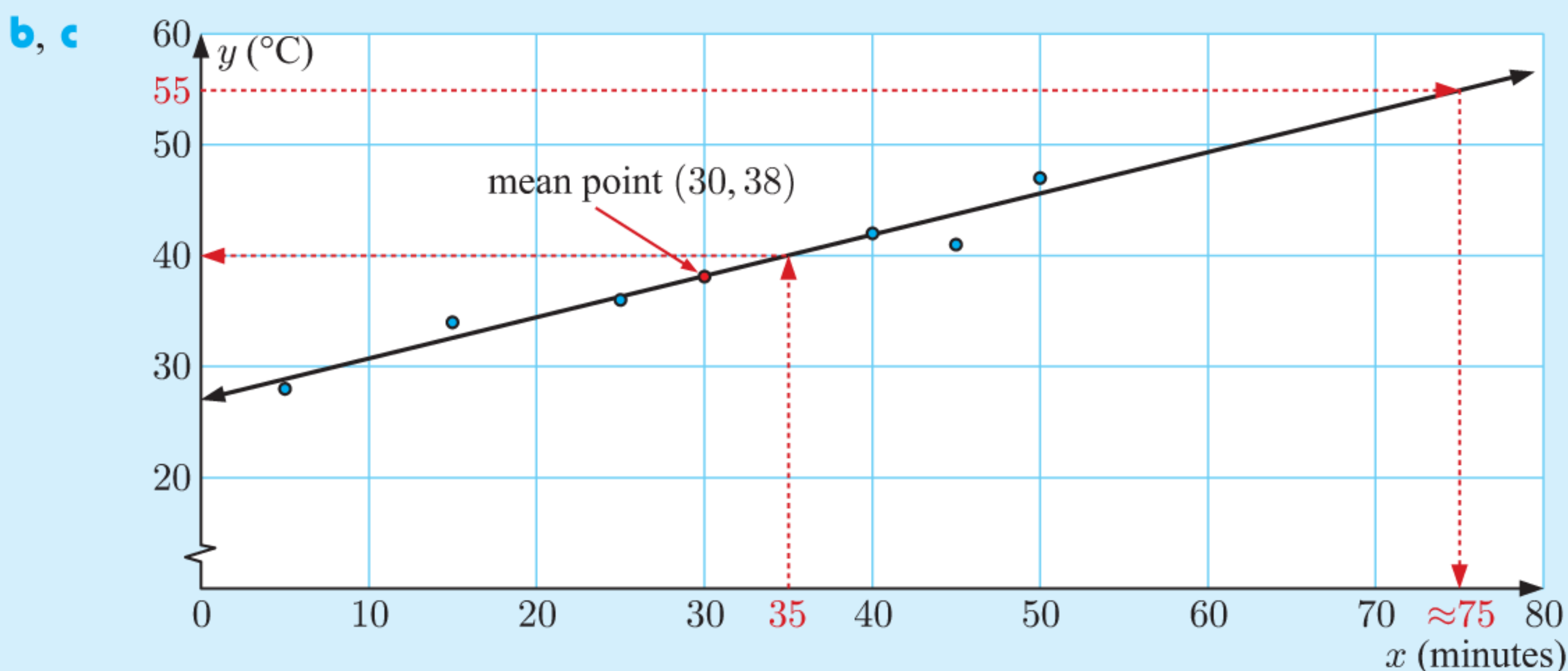
**Example 3****Self Tutor**

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 minutes)	50	5	25	40	15	45
Temperature (y °C)	47	28	36	42	34	41

- Calculate \bar{x} and \bar{y} .
- Draw a scatter diagram for the data.
- Locate the mean point (\bar{x}, \bar{y}) on the scatter diagram, then draw a line of best fit through this point.
- Predict the temperature of a car which has been left in the sun for 35 minutes.
- Predict how long it would take for a car's temperature to reach 55°C.
- Comment on the reliability of your predictions in **d** and **e**.

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



- 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 26C

- 1** Consider the data set:

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

- a** Draw a scatter diagram 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 (\bar{x}, \bar{y}) .
 - f** Locate the mean point, then use it in drawing a line of best fit.
 - g** Estimate the value of y when $x = 22$.
- 2** Fifteen students were weighed and their pulse rates were measured:

<i>Weight</i> (x kg)	46	37	32	57	47	64	42	30	52	56	65	43	36	28	40
<i>Pulse rate</i> (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** Calculate r .
 - c** Describe the relationship between *weight* and *pulse rate*.
 - d** Calculate the mean point (\bar{x}, \bar{y}) .
 - e** Locate the mean point on the scatter diagram, then use it in drawing a line of best fit.
 - f** Estimate the pulse rate of a 50 kg student. Comment on the reliability of your estimate.
- 3** The trunk widths and heights of the trees in a garden are given below:

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

- a** Draw a scatter diagram 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 (\bar{x}, \bar{y}) .
- e** Locate the mean point on the scatter diagram, 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.



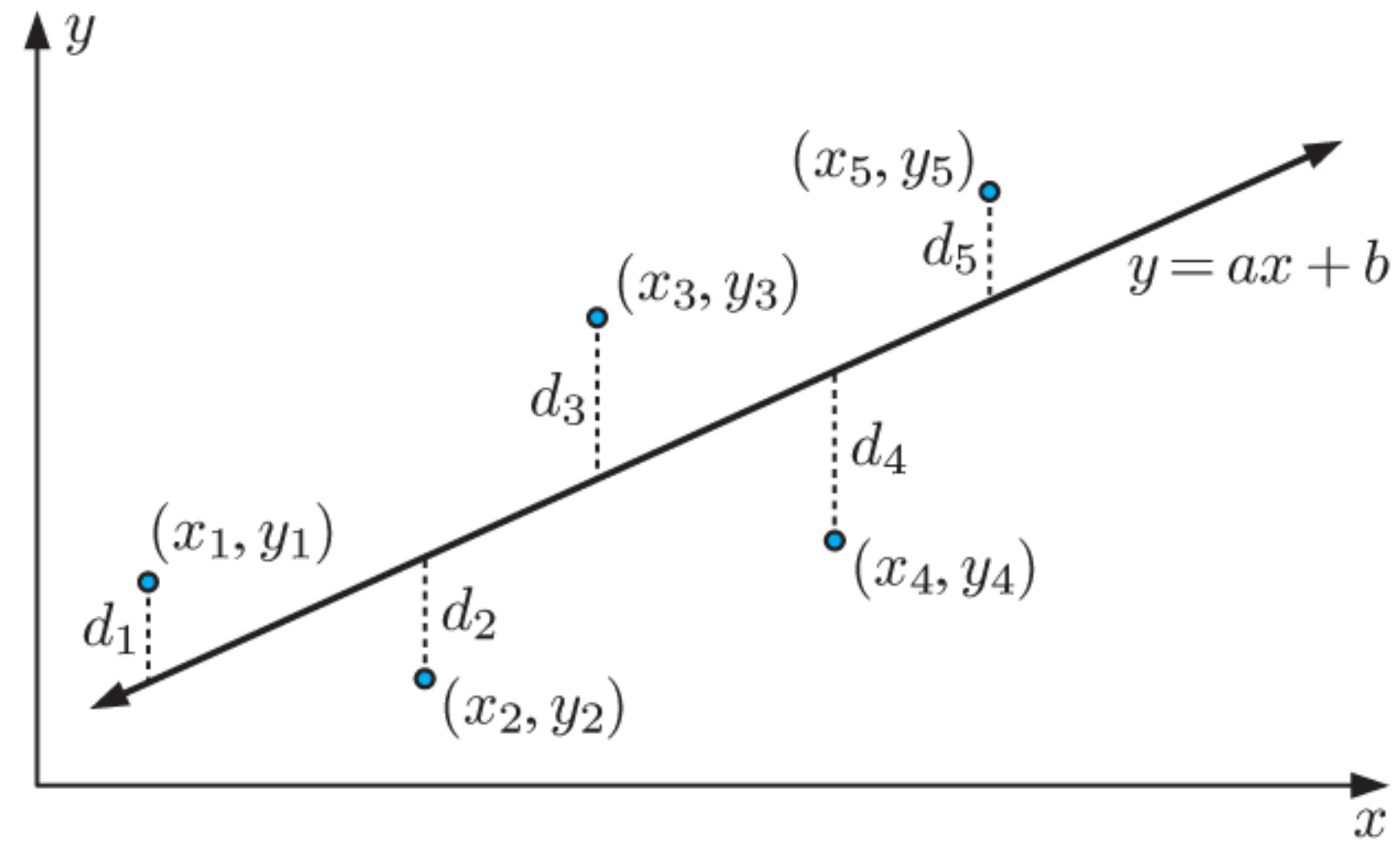
D THE LEAST SQUARES REGRESSION LINE

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

In least squares linear regression, we minimise the sum of the squares of the *vertical* distances between each data point and the **regression line**.

In other words, we need to find the straight line $y = ax + b$ where a and b are chosen to minimise

$$D = \sum_{i=1}^n d_i^2.$$



The required values of a and b are $a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ and $b = \bar{y} - a\bar{x}$.

Click on the icon to see how these formulae are derived.

LEAST SQUARES REGRESSION



In this course you will not be required to find the equation of the least squares regression line by hand.

Instead, you can use your **graphics calculator** or the **statistics package**.

STATISTICS PACKAGE



GRAPHICS CALCULATOR INSTRUCTIONS

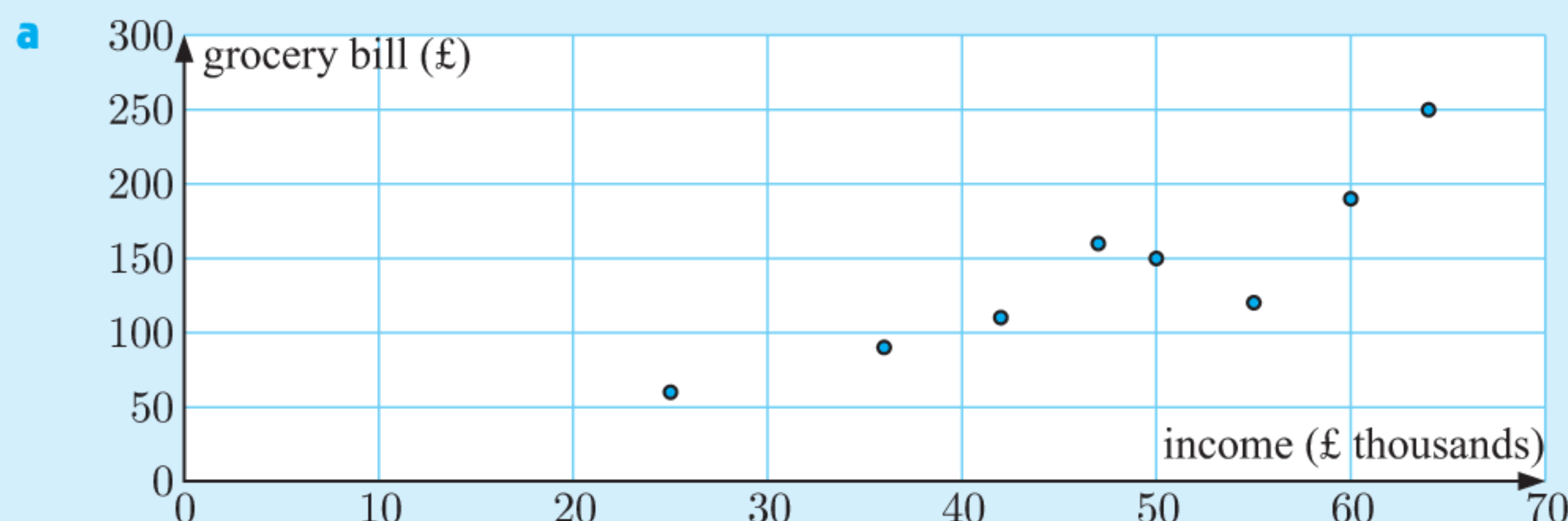
Example 4

Self Tutor

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

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

- Construct a scatter diagram to illustrate the data.
- Find the equation of the regression line. State and interpret its gradient.
- Estimate the weekly grocery bill for a family with an annual income of £95 000.
- Estimate the annual income of a family whose weekly grocery bill is £100.
- Comment on whether the estimates in **c** and **d** are likely to be reliable.



Casio fx-CG50

```

LinearReg(ax+b)
a =4.17825196
b =-56.694686
r =0.89484388
r2=0.80074556
MSe=839.7744
y=ax+b
                    
```

TI-84 Plus CE

```

NORMAL FLOAT AUTO REAL RADIAN MP
LinReg
y=ax+b
a=4.178251967
b=-56.69468693
r2=0.8007455697
r=0.8948438801
                    
```

TI-nspire

```

LinRegMx income,bill,1: CopyVar stat.RegEq
["Title" "Linear Regression (mx+b)"]
["RegEqn" "m*x+b"]
["m" 4.17825]
["b" -56.6947]
["r2" 0.800746]
["r" 0.894844]
["Resid" "{...}"]
                    
```

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

The gradient of the regression line ≈ 4.18 . This means that for every additional £1000 of income, a family's weekly grocery bill will increase by an average of £4.18.

- 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}

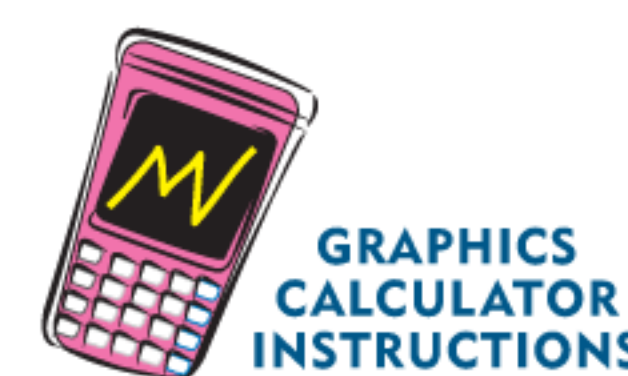
So, we expect a family with a weekly grocery bill of £100 to have an annual income of approximately £37 500.

- 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 therefore expect this estimate to be reliable.

EXERCISE 26D

- 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



- a** Draw a scatter diagram for the data.
- b** Use technology to find the equation of the regression line, and plot the line on your calculator.
- c** Use **b** to draw the regression line on your scatter diagram.

- 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 for him to get to work.

He collected data over a 14 day period:

<i>Temperature</i> ($x^\circ\text{C}$)	25	19	23	27	32	35	29	27	21	18	16	17	28	34
<i>Time</i> (y minutes)	35	42	49	31	37	33	31	47	42	36	45	33	48	39

- a** Draw a scatter diagram for the data.
- b** Calculate r .
- c** Describe the relationship between the variables.
- d** Is it reasonable to fit a linear model to this data? Explain your answer.

- 3 The prices of petrol and the number of customers per hour for sixteen petrol stations are:

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

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

- Calculate Pearson's product-moment correlation coefficient for the data.
 - Describe the relationship between the *petrol price* and the *number of customers*.
 - Use technology to find the equation of the regression line.
 - State and interpret the gradient of the regression line.
 - Estimate the number of customers per hour for a petrol station which sells petrol at 115.9 cents per litre.
 - Estimate the petrol price at a petrol station which has 40 customers per hour.
 - Comment on the reliability of your estimates in e and f.
- 4 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.

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

- Draw a scatter diagram for the data.
 - Calculate r .
 - Describe the relationship between the *number of speed cameras* and the *number of car accidents*.
 - Find the equation of the regression line.
 - State and interpret the gradient and y -intercept of the regression line.
 - Estimate the number of car accidents in a city with 10 speed cameras.
- 5 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 km h^{-1} , and the ceiling is given in km.



<i>Maximum speed</i>	<i>Ceiling</i>
460	8.84
420	10.06
530	10.97
530	9.906
490	9.448
530	10.36
680	11.73

<i>Maximum speed</i>	<i>Ceiling</i>
680	10.66
720	11.27
710	12.64
660	11.12
780	12.80
730	11.88

<i>Maximum speed</i>	<i>Ceiling</i>
670	12.49
570	10.66
440	10.51
670	11.58
700	11.73
520	10.36

- a** Draw a scatter diagram for the data.
- b** Calculate r .
- c** Describe the association between *maximum speed* (x) and *ceiling* (y).
- d** Use technology to find the regression line, and draw the line on your scatter diagram.
- e** State and interpret the gradient of the regression line.
- f** Estimate the ceiling for a fighter plane with a maximum speed of 600 km h^{-1} .
- g** Estimate the maximum speed for a fighter plane with a ceiling of 11 km .

- 6** A group of children was asked the numbers of hours they spent exercising and watching television each week.

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

- a** Draw a scatter diagram for the data.
- b** Calculate r .
- c** Describe the correlation between *time exercising* and *time watching television*.
- d** Find the equation of the regression line, and draw the line on your scatter diagram.
- e** State and interpret the gradient and y -intercept of the regression line.
- f**
 - i** One of the children in the group exercised for 7 hours each week. How much television does this child watch weekly?
 - ii** Use the regression line to predict the amount of television watched each week by a child who exercises for 7 hours each week.
 - iii** Compare your answers to **i** and **ii**.

- 7** The yield of pumpkins on a farm depends on the quantity of fertiliser used.

<i>Fertiliser</i> (x g per m^2)	4	13	20	26	30	35	50
<i>Yield</i> (y kg)	1.8	2.9	3.8	4.2	4.7	5.7	4.4

- a** Draw a scatter diagram for the data, and identify the outlier.
- b** What effect do you think the outlier has on:
 - i** the strength of correlation of the data
 - ii** the gradient of the regression line?
- c** Calculate the correlation coefficient:
 - i** with the outlier included
 - ii** without the outlier.
- d** Calculate the equation of the regression line:
 - i** with the outlier included
 - ii** without the outlier.
- e** If you wish to estimate the yield when 15 g per m^2 of fertiliser is used, which regression line from **d** should be used? Explain your answer.
- f** Can you explain what may have caused the outlier? Do you think the outlier should be kept when analysing the data?

- 8** Consider fitting a line through the origin to the bivariate data set $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. The model has the form $y = \beta x$.

- a** Show that the sum of the squared vertical distances from each point to the line is

$$D = \beta^2 \sum_{i=1}^n x_i^2 - 2\beta \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2.$$

- b** Hence find, in terms of x_i and y_i , the value of β that minimises D .

ACTIVITY 2

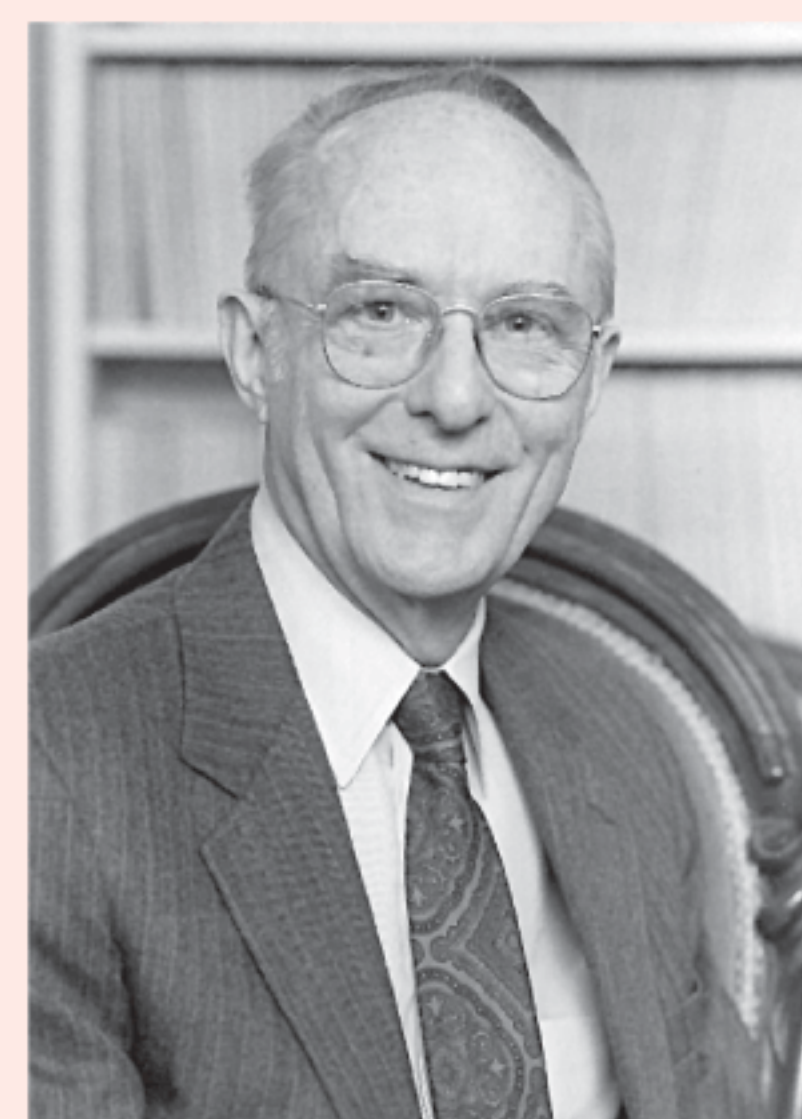
ANSCOMBE'S QUARTET

Anscombe's quartet is a collection of four bivariate data sets which have interesting statistical properties.

It was first described in 1973 by the English statistician **Francis Anscombe** (1918 - 2001). At the time, computers were becoming increasingly popular in statistics, as they allowed for more large scale and complex computations to be done within a reasonable amount of time. However, many common statistical packages primarily performed numerical calculations rather than produce graphs. Such output was often limited to those with advanced programming skills.

In his 1973 article, Anscombe stressed that:

“A computer should make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.”



Francis Anscombe
Photo courtesy of
Yale University.

The data values for Anscombe's quartet are given in the tables below:

Data set A:

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Data set B:

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.1	6.13	3.1	9.13	7.26	4.74

Data set C:

x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Data set D:

x	8	8	8	8	8	8	8	19	8	8	8
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.5	5.56	7.91	6.89

Enter the data into your **graphics calculator** or click on the icon to access the data in the **statistics package**.

STATISTICS
PACKAGE

**What to do:**

- For each data set, use technology to calculate:
 - the mean of each variable
 - the population variance of each variable.
 Comment on your answers.
- Find the regression line for each data set. What do you notice?
- Construct a scatter diagram for each data set, and plot the corresponding regression line on the same set of axes.
- How do your calculations in **1** and **2** compare to your graphs in **3**? Is a linear model necessarily appropriate for each data set?
- Why is it important to consider both graphs *and* descriptive statistics when analysing data?

ACTIVITY 3

RESIDUAL PLOTS

In addition to the *correlation coefficient* and the *linearity* of a scatter diagram, we can use a **residual plot** to decide whether a linear model is appropriate. Click on the icon to explore these graphs.



THEORY OF KNOWLEDGE

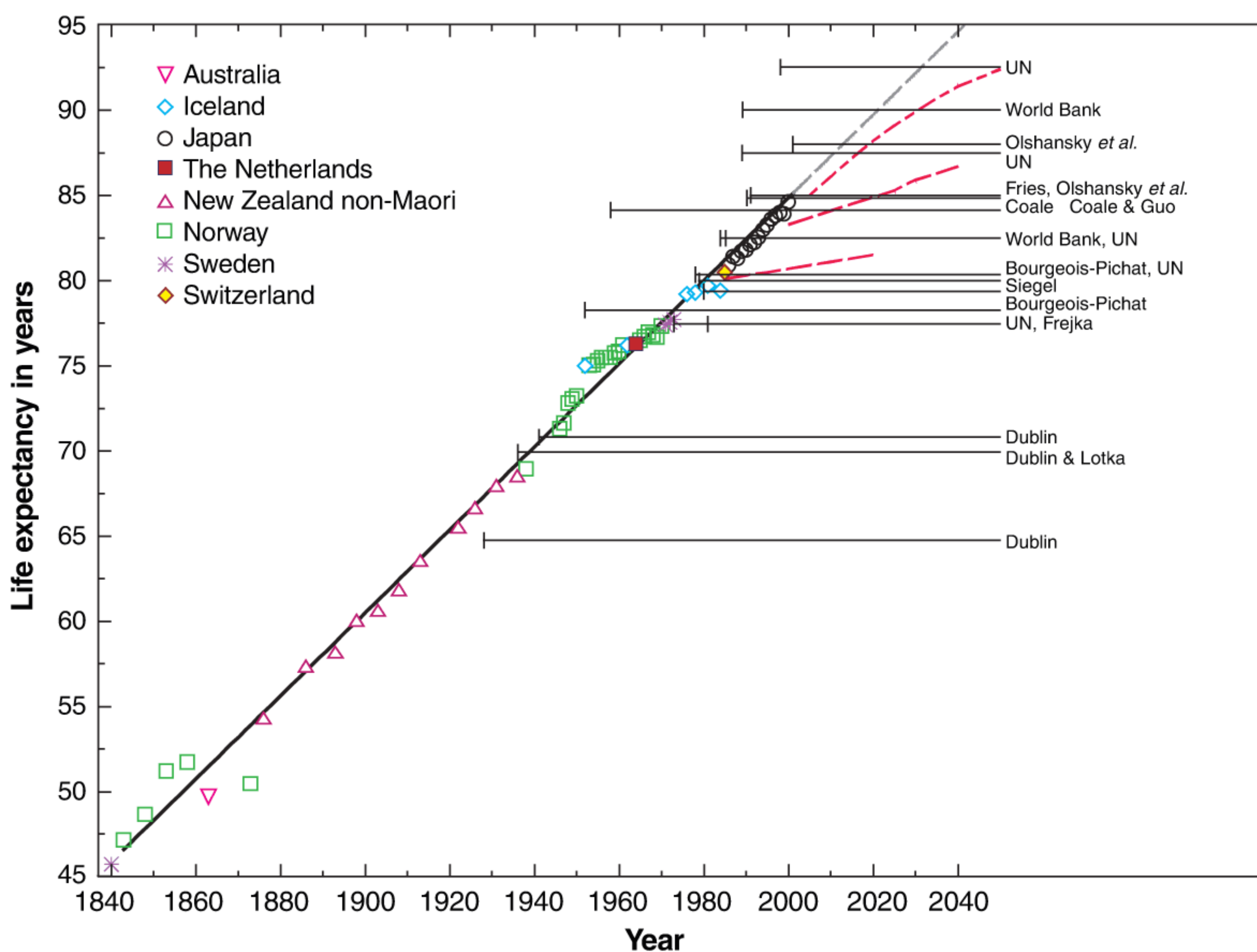
The use of extrapolation for predicting the future leads to debate on many global issues. Even when data shows a strong linear correlation, we need to consider whether it is reasonable for the trend to continue in the long term.

For example, the graph below is based on the article by Oeppen and Vaupel (2002)^[1]. It shows female life expectancy from 1840 to the early 2000s, and the country with the highest female life expectancy at each point in time.

Notice that:

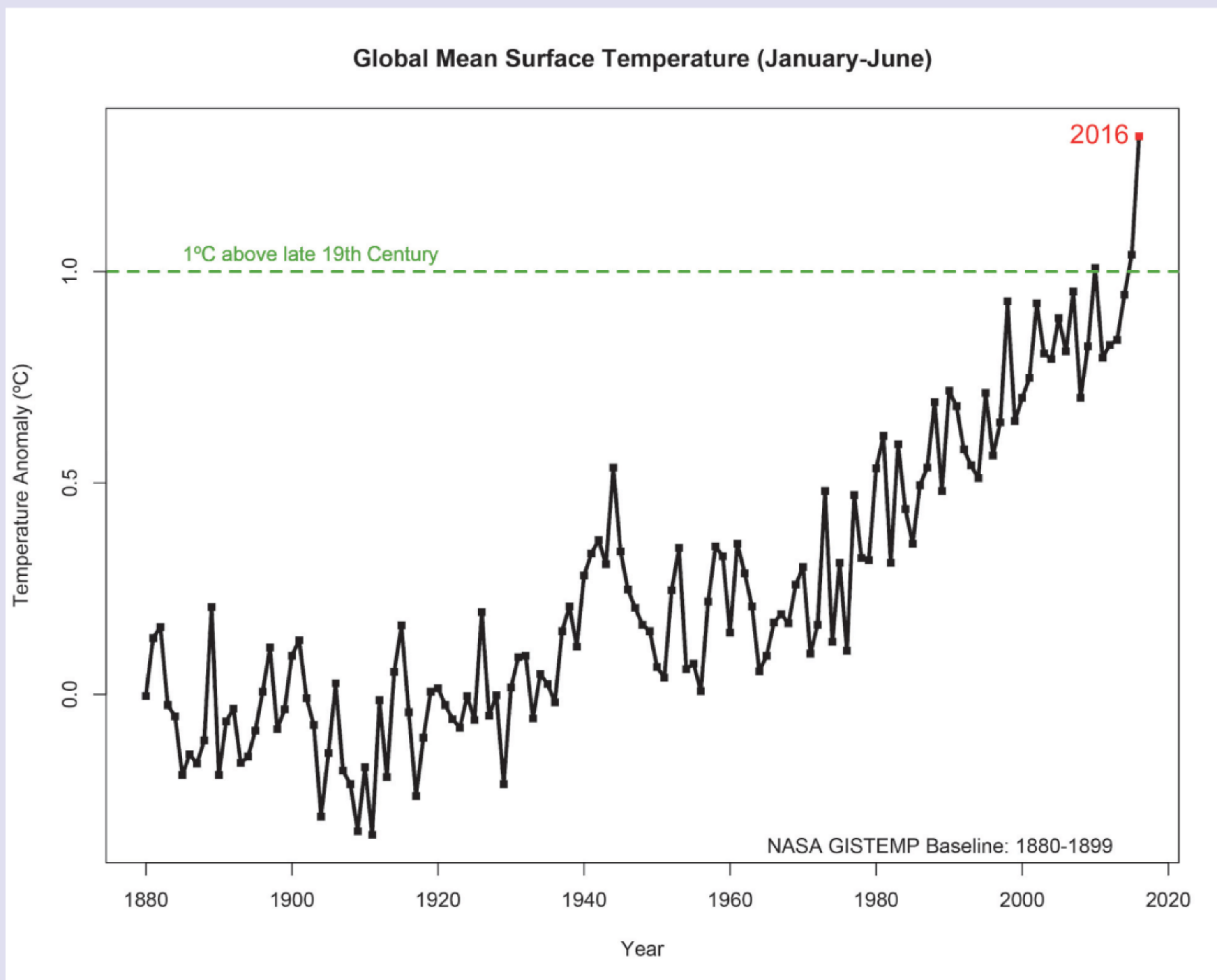
- The linear regression trend line is drawn in black, and extrapolated in grey.
- The horizontal black lines show asserted “ceilings” on life expectancy. The vertical line at the left end shows the year of publication.
- The dashed red lines denote projections of female life expectancy in Japan published by the United Nations (UN) in 1986, 1999, and 2001.

Record female life expectancy from 1840 to the present - Oeppen and Vaupel (2002)



- 1 Discuss the relationship between the variables.
- 2 Use the regression line to predict female life expectancy in the year 2100. Do you think this is realistic?
- 3 Discuss the “ceilings” suggested by publishers over time. Is there evidence to suggest that human life expectancy will approach a limiting “ceiling”?
- 4 Discuss the accuracy of the UN projections for females in Japan from 1986 to 1999. Is there reason to expect the latest projection will be more reliable?

The graph below shows data from the NASA Goddard Institute for Space Studies^[2]. The data for each point is for the first six months of the corresponding year.



- 5 Discuss the relationship between the variables. Is it reasonable to use a linear model to describe the mean surface temperature of the Earth over time? Is it reasonable to even conclude that the mean surface temperature of the Earth is increasing?
- 6 How can we predict the mean surface temperature of the Earth in the future?
- 7 Is mathematical extrapolation valid evidence for dictating environmental policy?

References:

- [1] Oeppen and Vaupel, *Broken limits to life expectancy*, *Science*, **296**, 5570, 1029-1031, 2002.
- [2] www.nasa.gov/feature/goddard/2016/climate-trends-continue-to-break-records

E THE REGRESSION LINE OF x AGAINST y

In the previous Section, we saw how linear regression can be used to find a linear model for the response variable y in terms of the explanatory variable x .

In these cases, we generally rely on the x -values being more precise than the y -values. This means that either there is less error involved with their measurement, or that there is naturally less variation associated with the x variable. The distance of each data point from the line is measured in the y -direction, so we are associating all of the “error” with the response variable y . However, in some cases the y -values may be more precisely measured.

The response variable y is always on the vertical axis.



For example:

- When a student studies for a test, their time spent studying x explains their test score y . However, the test score will be more precisely measured than the amount of time spent studying.
- At a breath testing station, police use a breathalyser to estimate the blood alcohol concentration (BAC) of drivers. If the result x is sufficiently high, the driver is required to take a blood test to establish their actual BAC. The blood test result y is a much more precise measurement.

In these scenarios, we consider the regression line of x against y . This means that we minimise the *horizontal* distances of points from the line, so all of the “error” is associated with the explanatory variable x .

We consider a line of the form $x = my + c$, and choose the constants m and c to minimise

$$H = \sum_{i=1}^n h_i^2.$$

It can be shown that $m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2}$ and $c = \bar{x} - m\bar{y}$.

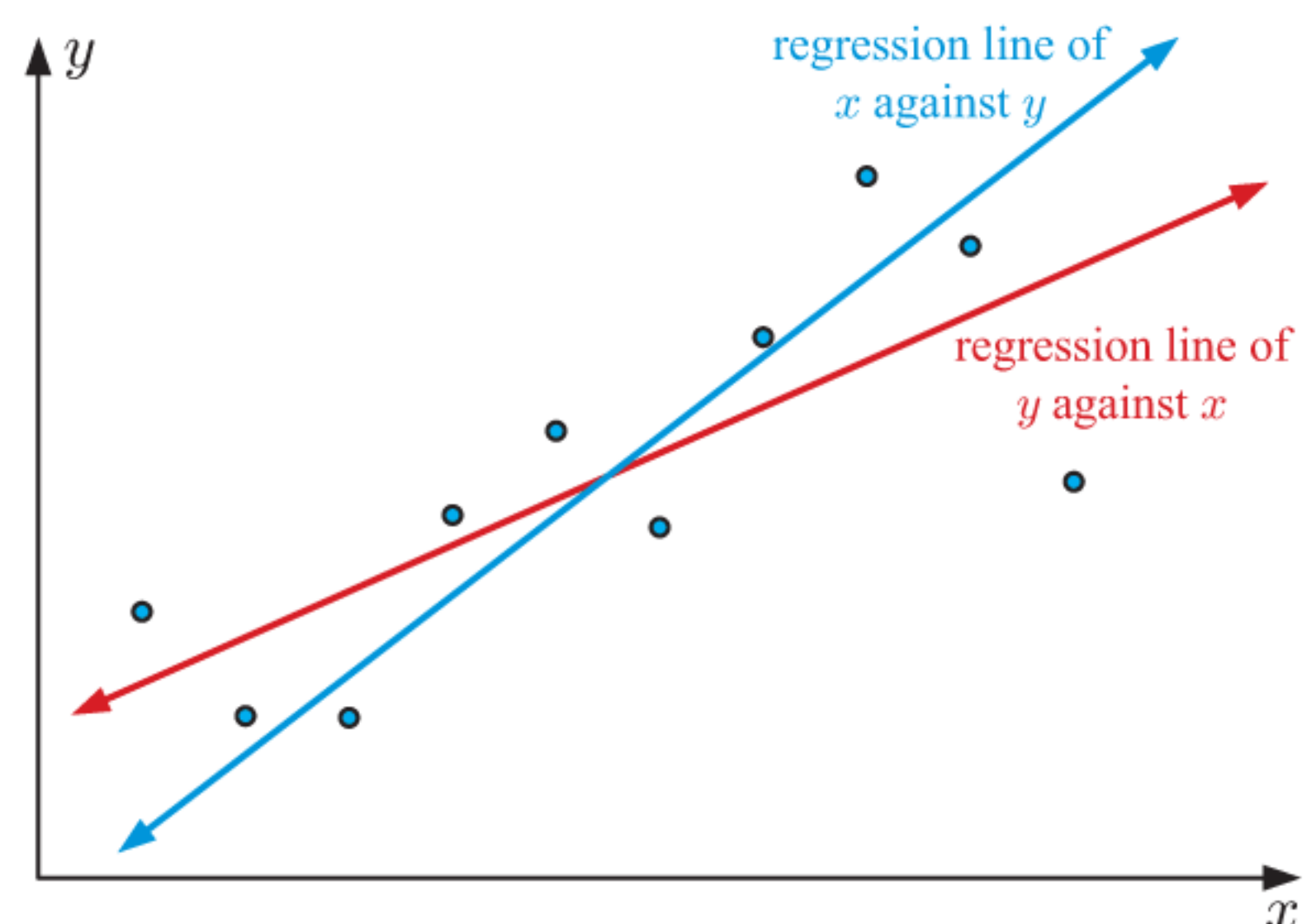
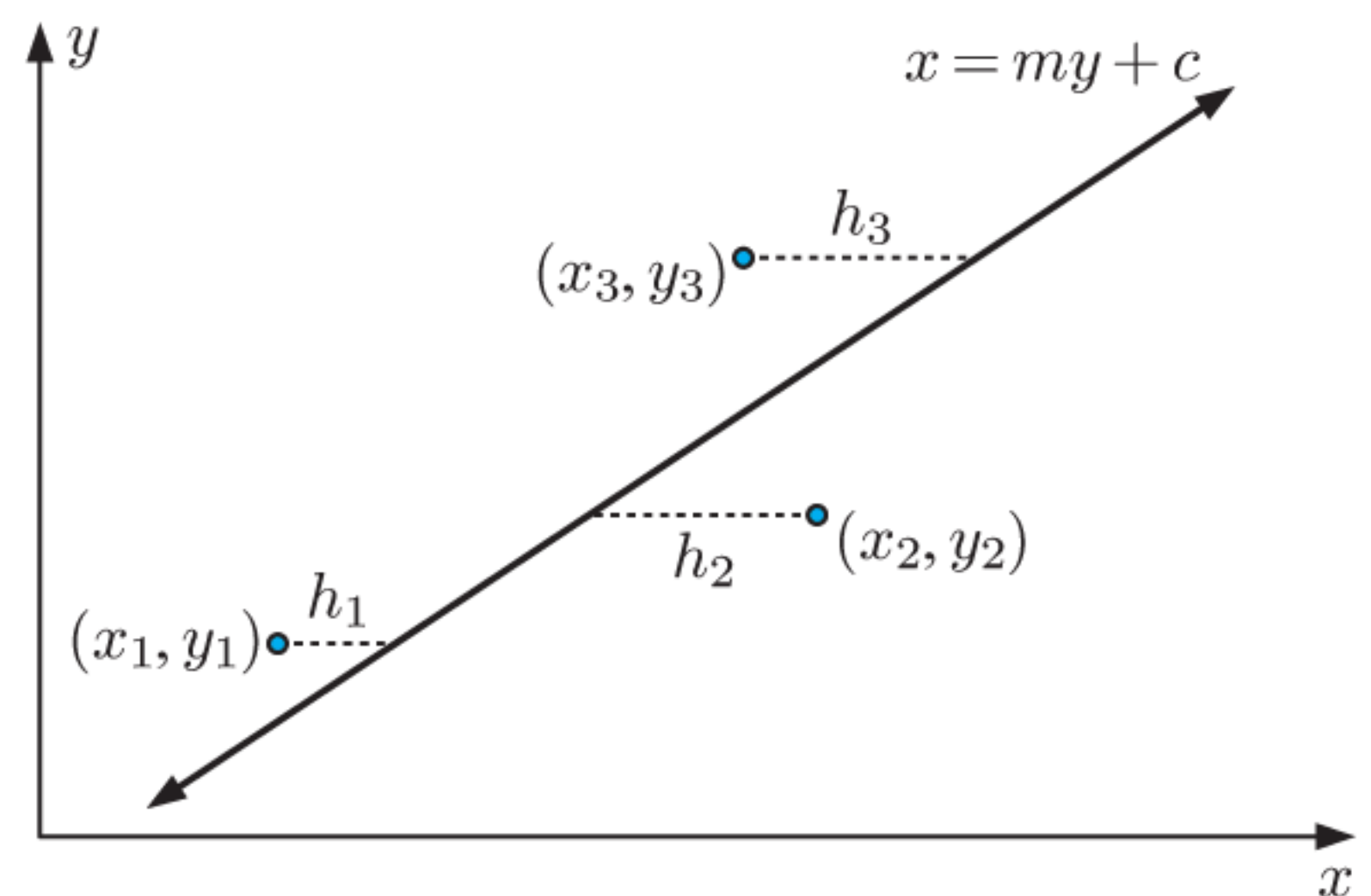
Rearranging the regression line $x = my + c$ into gradient-intercept form gives $y = \frac{1}{m}x - \frac{c}{m}$.

The gradient of the rearranged line is $\frac{1}{m} = \frac{\sum(y_i - \bar{y})^2}{\sum(x_i - \bar{x})(y_i - \bar{y})} \neq a$

and the y -intercept is

$$\begin{aligned} -\frac{c}{m} &= \frac{-(\bar{x} - m\bar{y})}{m} \\ &= \frac{m\bar{y} - \bar{x}}{m} \\ &= \bar{y} - \frac{1}{m}\bar{x} \\ &\neq b \quad \left\{ \text{since } \frac{1}{m} \neq a \right\} \end{aligned}$$

So, in general, the regression line of x against y is **not** the same as the regression line of y against x .



Example 5**Self Tutor**

The data below shows the quantity of feed eaten and the number of eggs laid in a fortnight for 12 Leghorn chickens:

<i>Feed (x kg)</i>	0.57	0.86	0.49	1.37	0.91	0.50
<i>Number of eggs (y)</i>	4	6	3	9	6	3
<i>Feed (x kg)</i>	0.97	1.06	1.00	0.68	1.34	0.94
<i>Number of eggs (y)</i>	6	7	7	4	9	6

- Explain why it would be appropriate to use the regression line of x against y in this case.
- Find the regression line of x against y .
- Use the regression line to estimate:
 - the quantity of feed required for a Leghorn chicken to lay a dozen eggs
 - the number of eggs laid in a fortnight by a hen that ate 1.2 kg of feed.

- The number of eggs laid can be counted exactly, whereas the quantity of feed usually cannot be measured exactly.

Since the response variable y is more precisely measured than the explanatory variable x , it would be appropriate to use the regression line of x against y in this case.

b

	List 1	List 2	List 3	List 4
SUB				
1	0.57	4		
2	0.86	6		
3	0.49	3		
4	1.37	9		

	List 1	List 2	List 3	List 4
SUB				
1	0.57	4		
2	0.86	6		
3	0.49	3		
4	1.37	9		

The regression line of x against y is $x \approx 0.142y + 0.0603$ kg.

- When $y = 12$, $x \approx 0.142(12) + 0.0603$
 ≈ 1.76
 We expect a Leghorn chicken to need about 1.76 kg of feed to lay a dozen eggs.
 - When $x = 1.2$, $1.2 \approx 0.142y + 0.0603$
 $\therefore 1.1397 \approx 0.142y$
 $\therefore y \approx 8$
 We expect a Leghorn chicken eating 1.2 kg of feed to lay about 8 eggs.

EXERCISE 26E

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

<i>Time on homemade meals (x hours)</i>	3.5	6.0	4.0	8.5	7.0	2.5	9.0	7.0	4.0	7.5
<i>Money on fast food (\$y)</i>	85	0	60	0	27	100	15	40	59	29

- Explain why it would be appropriate to use the regression line of x against y in this case.

- b** Find the regression line of x against y .
- c** Use the regression line to estimate:
- i** the time spent preparing homemade meals by a family that spends \$45 on fast food
 - ii** the amount of money spent on fast food by a family that spends 5 hours preparing homemade meals.
- 2** The table below shows how far a group of students live from school, and how long it takes them to travel there each day.

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

- a** Draw a scatter diagram of the data.
 - b** Which regression line should be used to model the relationship between the variables? Explain your answer.
 - c** Use an appropriate regression line to estimate the travel time of a student who lives 15 km from school.
 - d** Comment on the reliability of your estimate.
- 3** Eight students swim 200 m breaststroke. Their times y in seconds, and arm lengths x in cm, are shown in the table below:

<i>Length of arm</i> (x cm)	78	73	71	68	76	72	63	69
<i>Breaststroke</i> (y seconds)	123.1	123.7	127.3	132.0	120.8	125.0	140.9	129.0

- a** Draw a scatter diagram for the data.
 - b** Find the equation of the regression line of:
 - i** y against x
 - ii** x against y .
 - c** Plot both regression lines on your scatter diagram. What do you notice?
- 4** Consider the bivariate data set $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. The equation of the regression line of y against x is $y = ax + b$ and the equation of the regression line of x against y is $x = my + c$.
- a** Show that $ma = r^2$, where r is Pearson's product-moment correlation coefficient.
 - b** Hence find the condition(s) under which the two regression lines will be the same.

DISCUSSION

Suppose the variables x and y are measured with equally poor precision.

- 1** Is it necessarily appropriate to choose one regression line over the other in this case?
- 2** How could we take the poor precision of *both* variables into account when formulating the linear regression model? Consider the "distance" that we minimise when we perform the linear regression.
- 3** In some situations, the variables depend equally on each other. In these cases we say that the variables are **co-dependent**, and the variables can be placed on either axis of the scatter diagram. Is it sensible to use a regression line to describe the relationship between co-dependent variables?

THEORY OF KNOWLEDGE

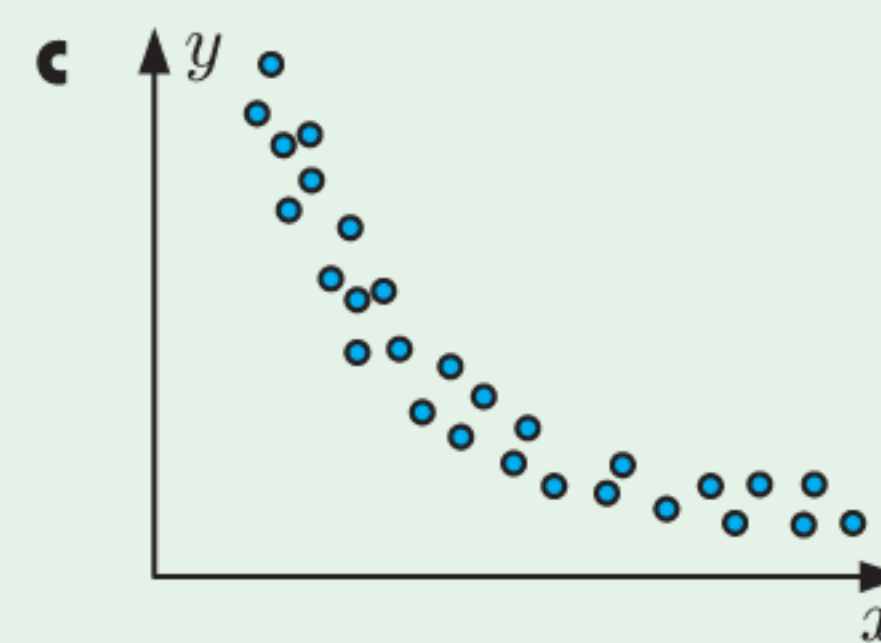
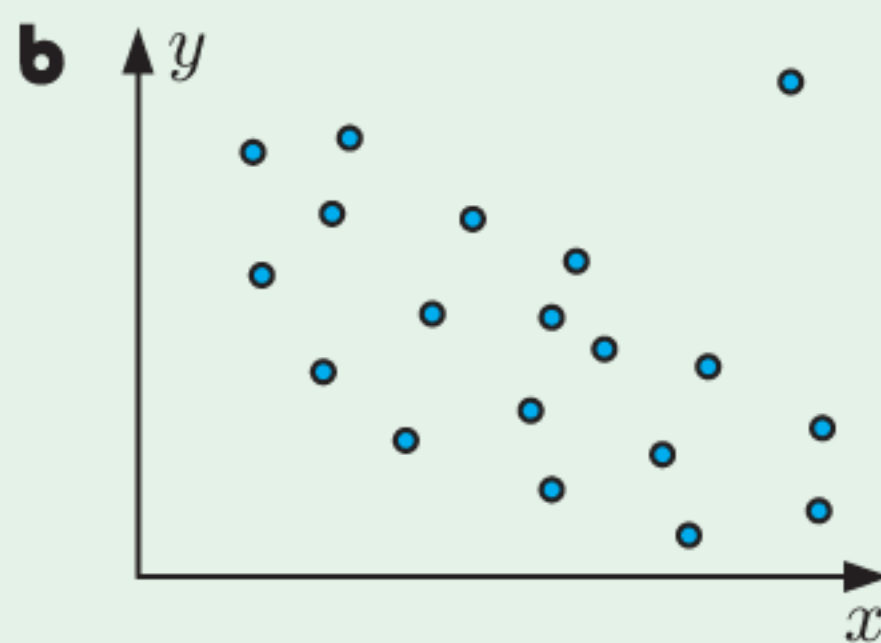
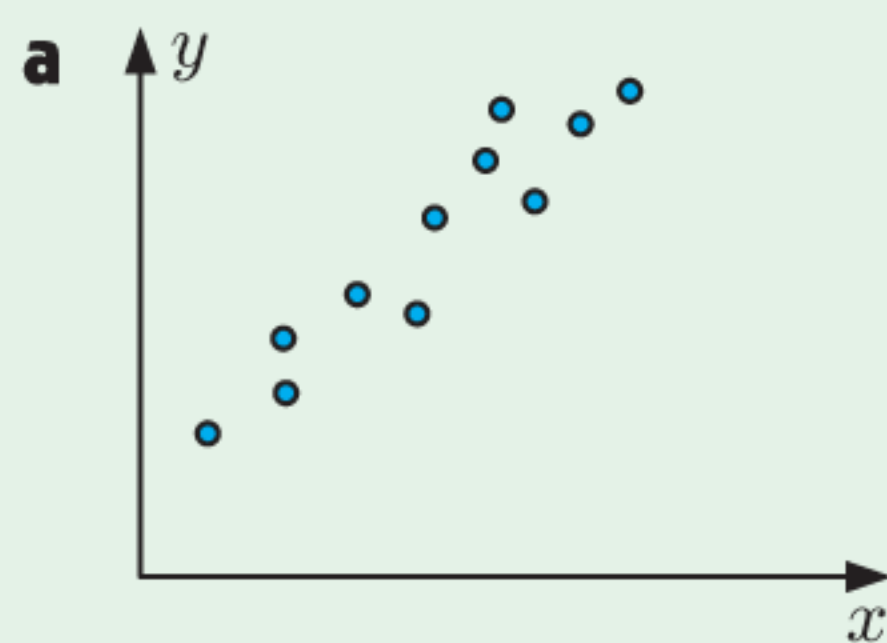
Since the 1970s and 1980s, wage discrimination between men and women has been a topic of debate. During that time, Conway and Roberts^[1] published a study which used linear regression to show that on average, women with the same qualifications as men were paid less. This would seem to imply that given the same salary, women would be more qualified. However, when the regression was applied the other way, the opposite conclusion was observed.

- 1 Can you explain why this occurred?
- 2 Are these two questions necessarily equivalent?
 - “Given the same qualifications, do men and women earn the same wage?”
 - “Given the same wage, do men and women have the same qualifications?”
- 3 Is it necessary to consider both regression lines in order to conclude whether discrimination has occurred?
- 4 Should people be paid according to their qualifications, the job they do, or their capability in doing that job?

[1] Conway, Delores A. and Harry V. Roberts (1983). “Reverse Regression, Fairness, and Employment Discrimination”. In: *Journal of Business & Economic Statistics* 1.1, pp. 75-85.

REVIEW SET 26A

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



- 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 Consider the data set alongside.

x	2	5	7	10	12	15
y	18	10	13	7	7	5

 - a Draw a scatter diagram for the data.
 - b Does the correlation between the variables appear to be positive or negative?
 - c Calculate Pearson's product-moment correlation coefficient r .

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

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

- a Draw a scatter diagram for the data.

- b** Calculate Pearson's product-moment correlation coefficient r .
- c** Describe the correlation between *ticket sales* and *beverage sales*.

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

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

- a** Find the mean for each variable.
- b** Draw a scatter diagram for the data. Plot the mean point, and draw a line of best fit by eye.
- c** Describe the relationship between *time in the store* and the *money spent*.

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

<i>Age (x years)</i>	3	9	7	4	4	12	8	6	5	10	13
<i>Height (y cm)</i>	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 regression line of y against x .
- c** State and interpret the gradient of the regression line.
- d** Use the regression line to predict the height of a 5 year old child.
- e** Based on the given data, at what age would you expect a child to reach 140 cm in height?



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

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

- a** Draw a scatter diagram 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 fit a linear model?
- e** Determine the equation of the regression line of y against x .
- f** State and interpret the gradient and y -intercept of the regression line.
- g** Use your line to estimate:
 - i** the yield if the spray concentration is 7 mL per L
 - ii** the spray concentration if the yield is 200 tomatoes per bush.
- h** Comment on the reliability of your estimates in **g**.

- 8 Thomas rode his bicycle for an hour each day for eleven days. He recorded the number of kilometres he rode, along with his estimate of the temperature that day:

<i>Temperature</i> (T °C)	23	24	25	27	28	20	22	21	25	26	24
<i>Distance</i> (d km)	26.5	26.7	24.4	22.8	23.5	32.6	28.7	29.4	24.2	23.2	29.7

- Draw a scatter diagram for the data.
- Explain why it would be appropriate to use the regression line of T against d in this case.
- Find the equation of the regression line of T against d .
- How far would you expect Thomas to ride on a 30°C day?

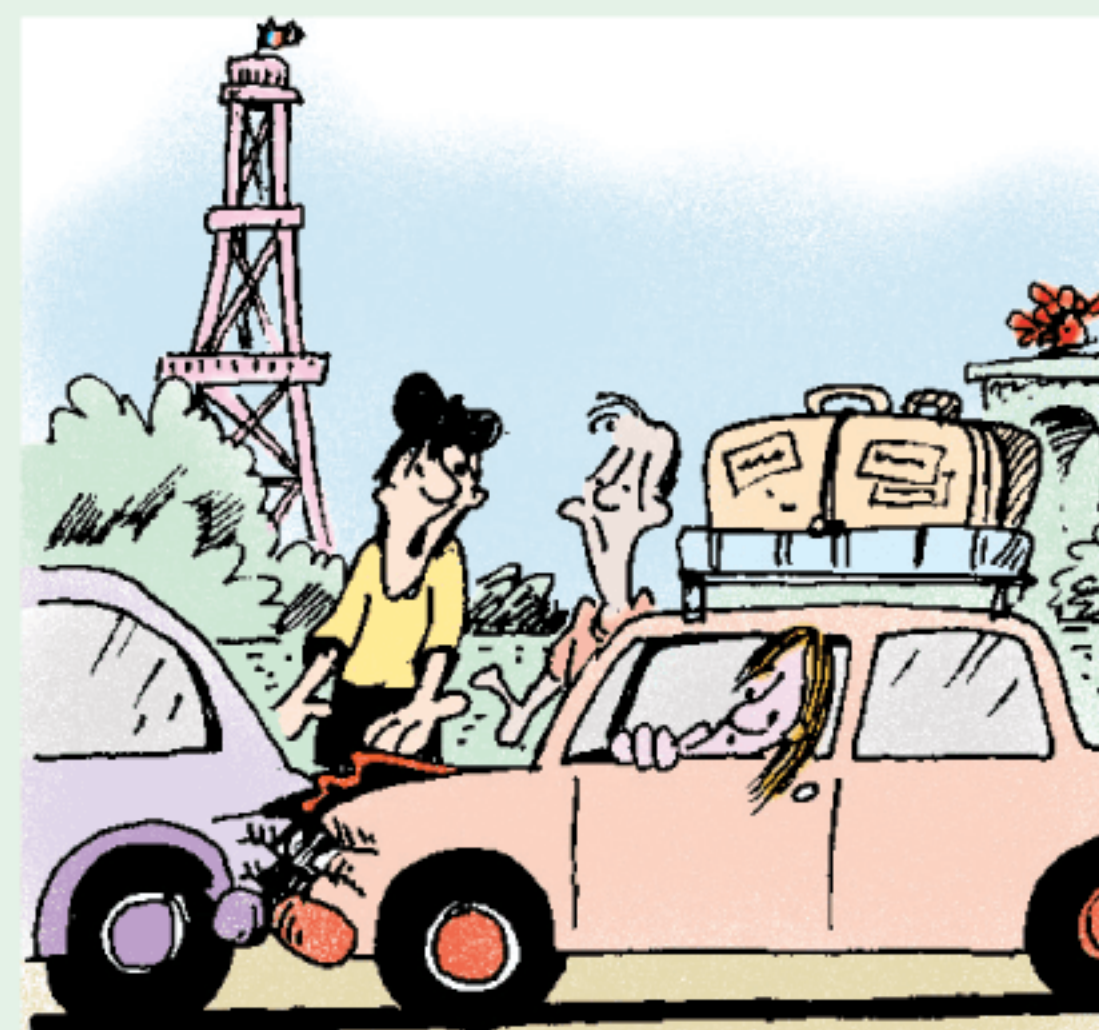
REVIEW SET 26B

- 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:
 - price of tickets* and *number of tickets sold*
 - ice cream sales* and *number of shark attacks*.
- A group of students is comparing their results for a Mathematics test and an Art project:

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

- Construct a scatter diagram for the data.
 - Describe the relationship between the Mathematics and Art marks.
 - Calculate the correlation coefficient r between the variables.
- 3 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 were recorded for a particular driver in the same car under the same test conditions.



<i>Speed</i> (v km h ⁻¹)	10	20	30	40	50	60	70	80	90
<i>Stopping time</i> (t s)	1.23	1.54	1.88	2.20	2.52	2.83	3.15	3.45	3.83

- Find the mean point (\bar{v}, \bar{t}) .
- Draw a scatter diagram of the data. Add the mean point and draw a line of best fit by eye.
- Hence estimate the stopping time for a speed of:
 - 55 km h⁻¹
 - 110 km h⁻¹
- Which of your estimates in **c** is more likely to be reliable?

4 Consider the data set alongside.

x	2	3	6	8	13	16
y	12	17	32	41	50	61

- Calculate the correlation coefficient r .
- Find the regression line of y against x .
- Estimate the value of y when $x = 10$.

5 A craft shop sells canvasses in a variety of sizes. The table below 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

- Construct a scatter diagram for the data.
 - Calculate the correlation coefficient r .
 - Describe the correlation between *area* and *price*.
 - Find the regression line of y against x , then draw the line on the scatter diagram.
 - Estimate the price of a canvas with area 1200 cm². Discuss whether your estimate is likely to be reliable.
- 6 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

- Produce a scatter diagram for the data.
 - Are there any outliers? If so, should they be included in the analysis?
 - Calculate the equation of the regression line of s against p .
 - State and interpret the gradient of the regression line.
 - Do you think the regression line would give a reliable prediction of sales if Supa-fizz was priced at 50 cents? Explain your answer.
- 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

- Draw a scatter diagram for the data, and describe the correlation between the variables.
- Find the equation of the regression line of f against n .
- Is it likely that a causal relationship exists between these two variables? Explain your answer.
- Plot the regression line on the scatter diagram.
- Violet has two beds of petunias. She waters one of the beds 5 times a fortnight and the other 10 times a week.
 - How many flowers can she expect from each bed?
 - Discuss which of your estimates is likely to be more reliable.



- 8** An archer shoots 10 arrows at a target from each of 12 different positions. The table below shows the distance of each position from the target, and how many shots were successful.

<i>Distance from target (x m)</i>	20	25	15	35	40	55	30	45	60	80	65	70
<i>Hits (y)</i>	9	8	8	8	7	6	9	7	4	2	3	3

- Draw a scatter diagram for the data.
- Explain why it would be appropriate to use the regression line of x against y in this case.
- Find the equation of the regression line of x against y .
- Predict the number of hits out of 10 shots fired at a distance of 100 m. Discuss the reliability of your estimate.

13 $y^2 = 2x^2(\ln|x| + c)$

14 a $y = 2x\sqrt{x} - 4x$

b $y = -\frac{1}{2} \cot x \cos x$

15 a $y = x^2 - \frac{1}{2}x^4 + \frac{1}{3}x^6 + \dots$

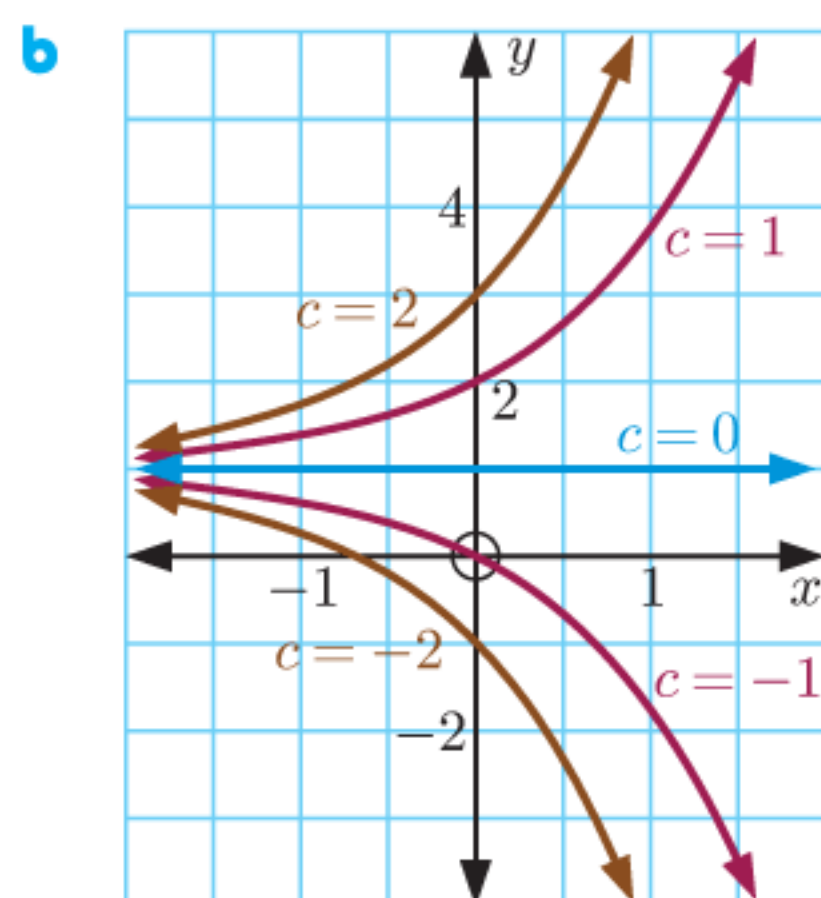
b $y = \ln(x^2 + 1)$

16 a $x(t) = \sum_{k=0}^{\infty} \frac{(-1)^k (mt)^{2k}}{(2k)!}$

b $x(t) = \cos mt$

REVIEW SET 25B

2 a $\frac{dy}{dx} = ce^x = (ce^x + 1) - 1 = y - 1$ ✓



c $y = 3e^x + 1$

d $y = 3x + 4$

3 $y(0.5) \approx 0.8555$

4 a $y = \ln|e^x - 2| + c$

b $y = -\frac{1}{4} \sin\left(\frac{\pi}{3} - 2x\right) - \frac{\sqrt{3}}{8}$

6 a $y = \frac{1}{\sqrt[3]{c - 6x}}$

b $P = A\sqrt{t^2 + 1}$

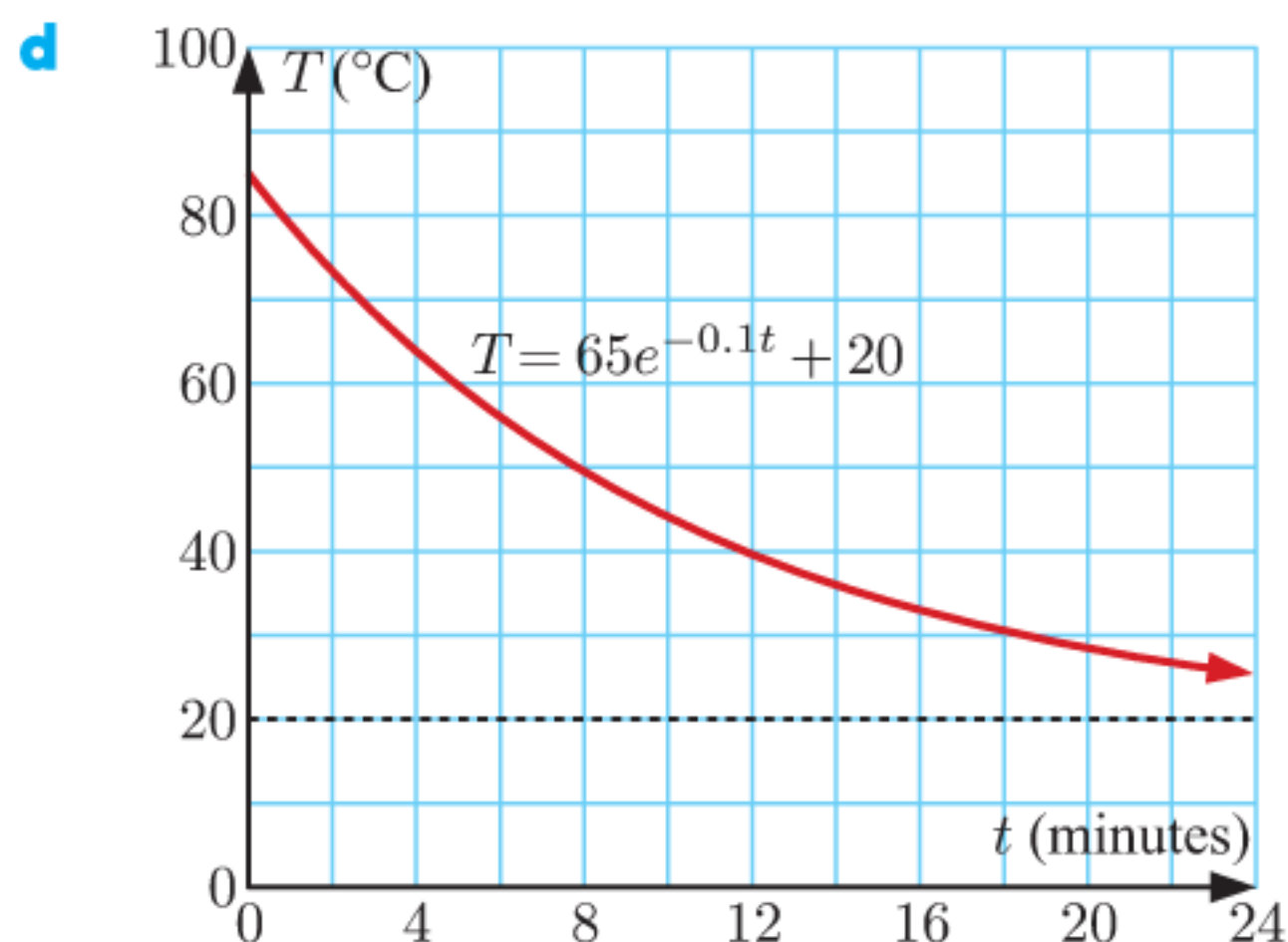
7 a $y = \left(\frac{1}{2}x \pm 2\right)^2$

b $y = e^{\sin x - 3}$

8 $y = \frac{(x + 2)^2}{25(x - 2)}$

9 a $T = 65e^{-0.1t} + 20$

b $\approx 63.6^\circ\text{C}$

c as $t \rightarrow \infty$, $T \rightarrow 20$ 

e i ≈ 3.68 minutes

ii ≈ 8.11 minutes

10 $y = \frac{5}{\sqrt{x}}$

11 a $\frac{dV}{dt} = -k\sqrt{h}$

b $V = 2 \times 2 \times h$, $\frac{dh}{dt} = -\frac{k}{4}\sqrt{h}$

c 20 minutes

12 a $P(t) \approx \frac{200\,000}{1 + 7e^{-0.129t}}$

b $\approx 42\,800$ ostriches

c after ≈ 15.1 years

13 $y^2 = 3x^2 - 2x - 1$

14 $y = x^5 - \frac{1}{x^3}$

15 a $y = 1 + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \dots$

b $y = (x + 1)^{x+1}e^{-x}$

16 d $y^2 = 2cx + c^2$

e The mirror is a parabola (x is a quadratic in y).

EXERCISE 26A

1 a weak, positive, linear correlation, with no outliers

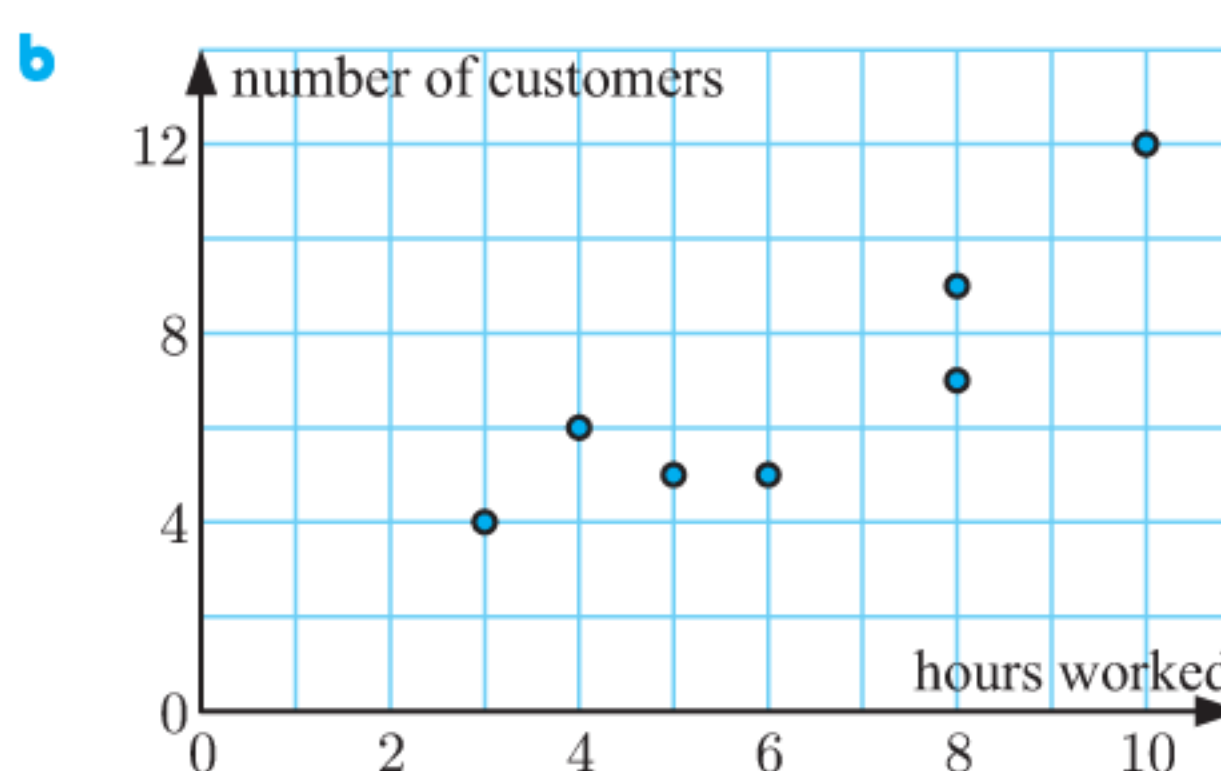
b strong, negative, linear correlation, with one outlier

c no correlation

d strong, negative, non-linear correlation, with one outlier

e moderate, positive, linear correlation, with no outliers

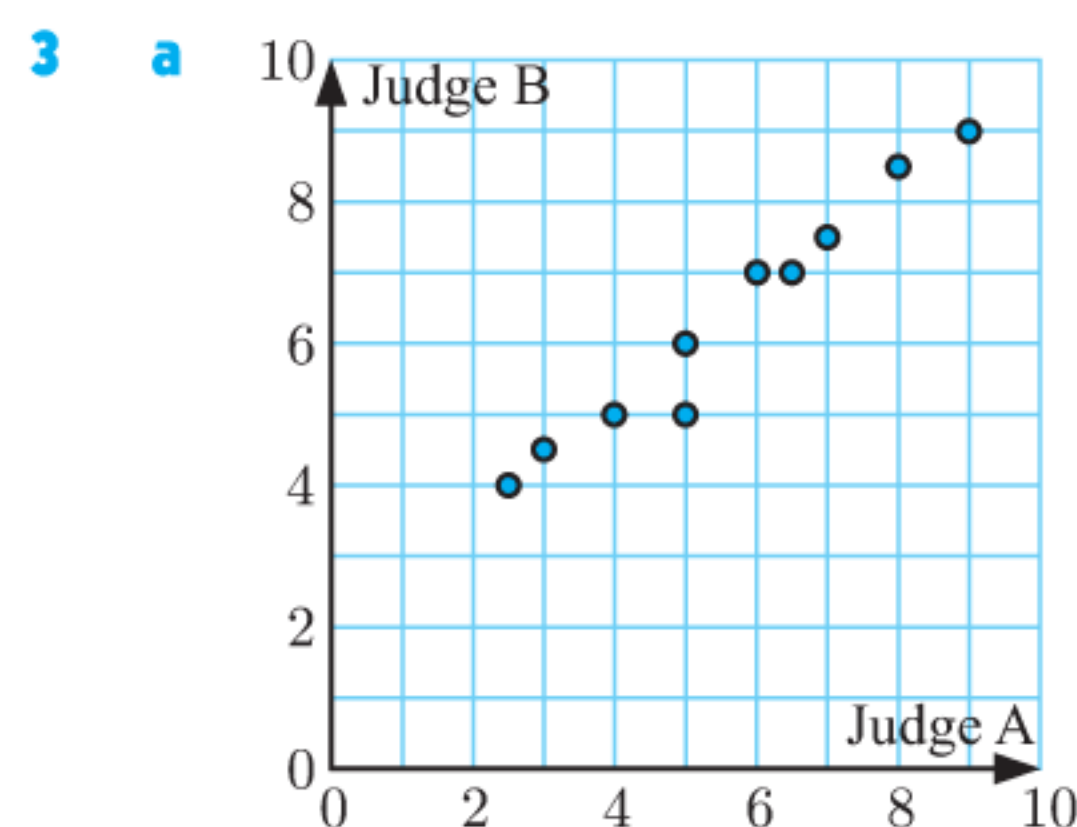
f weak, positive, non-linear correlation, with no outliers

2 a *Hours worked* is the explanatory variable.*Number of customers* is the response variable.

c i Monday and Friday

ii Wednesday and Sunday

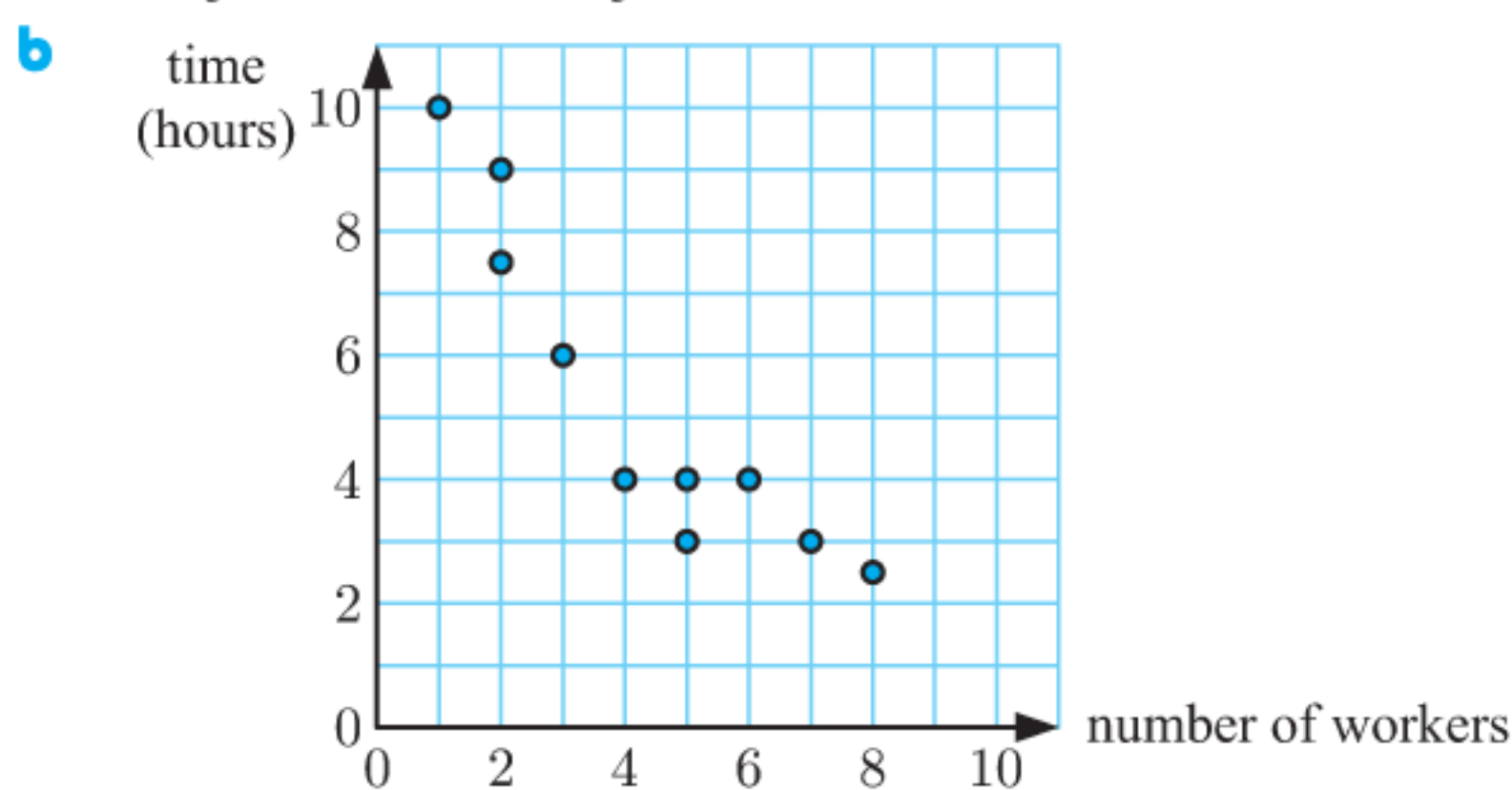
d The more hours that Tiffany works, the more customers she is likely to have.

b There appears to be **strong, positive, linear** correlation between Judge A's scores and Judge B's scores. This means that as Judge A's scores increase, Judge B's scores **increase**.

c No, the scores are related to the quality of the ice skaters' performances.

4 a i job G

ii job C

c There is a strong, negative, non-linear correlation between *number of workers* and *time*.

5 a D b A c B d C

6 a There is a moderate, positive, linear correlation between *hours of study* and *marks obtained*.b The test is out of 50 marks, so the outlier (> 50) appears to be an error. It should be discarded.

c Yes, this is a causal relationship as spending more time studying for the test is likely to cause a higher mark.

7 a Not causal, dependent on genetics and/or age.

b Not causal, dependent on the size of the fire.

c Causal, an increase in advertising is likely to cause an increase in sales.

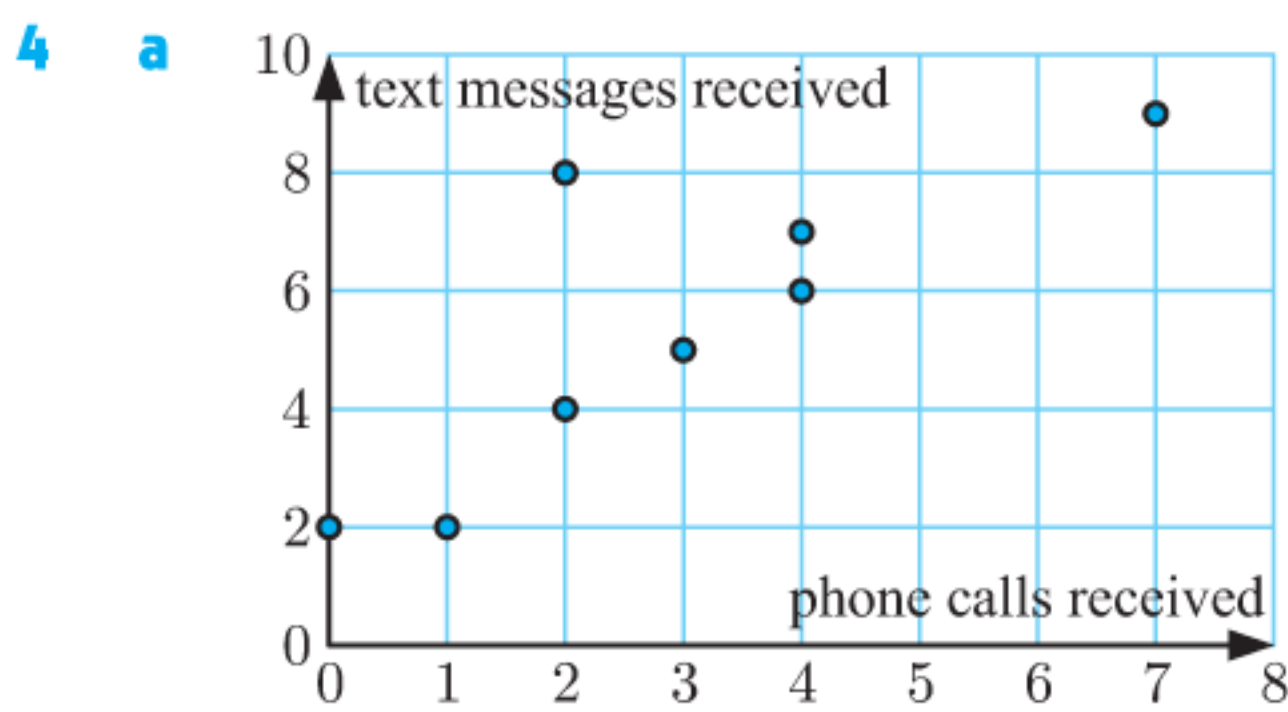
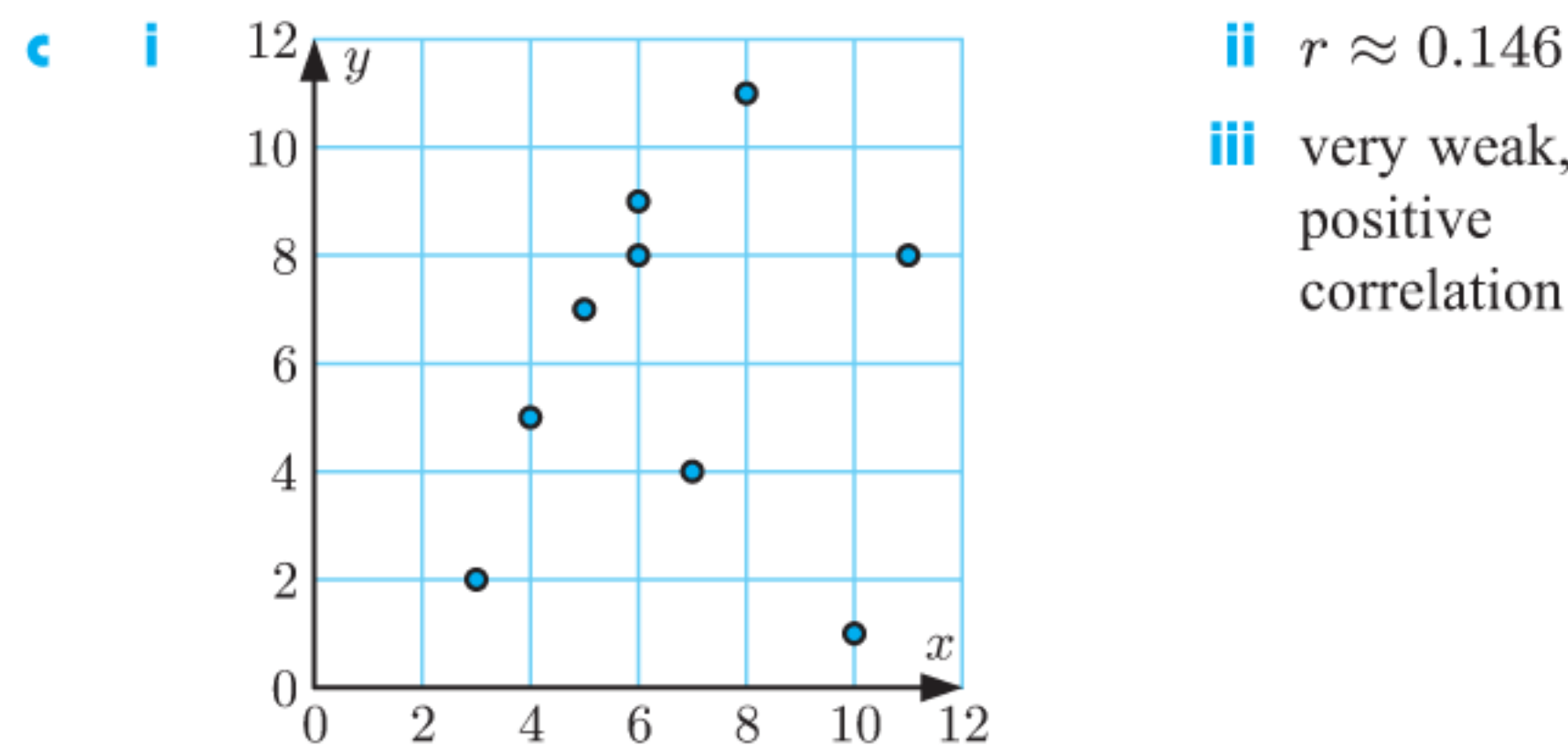
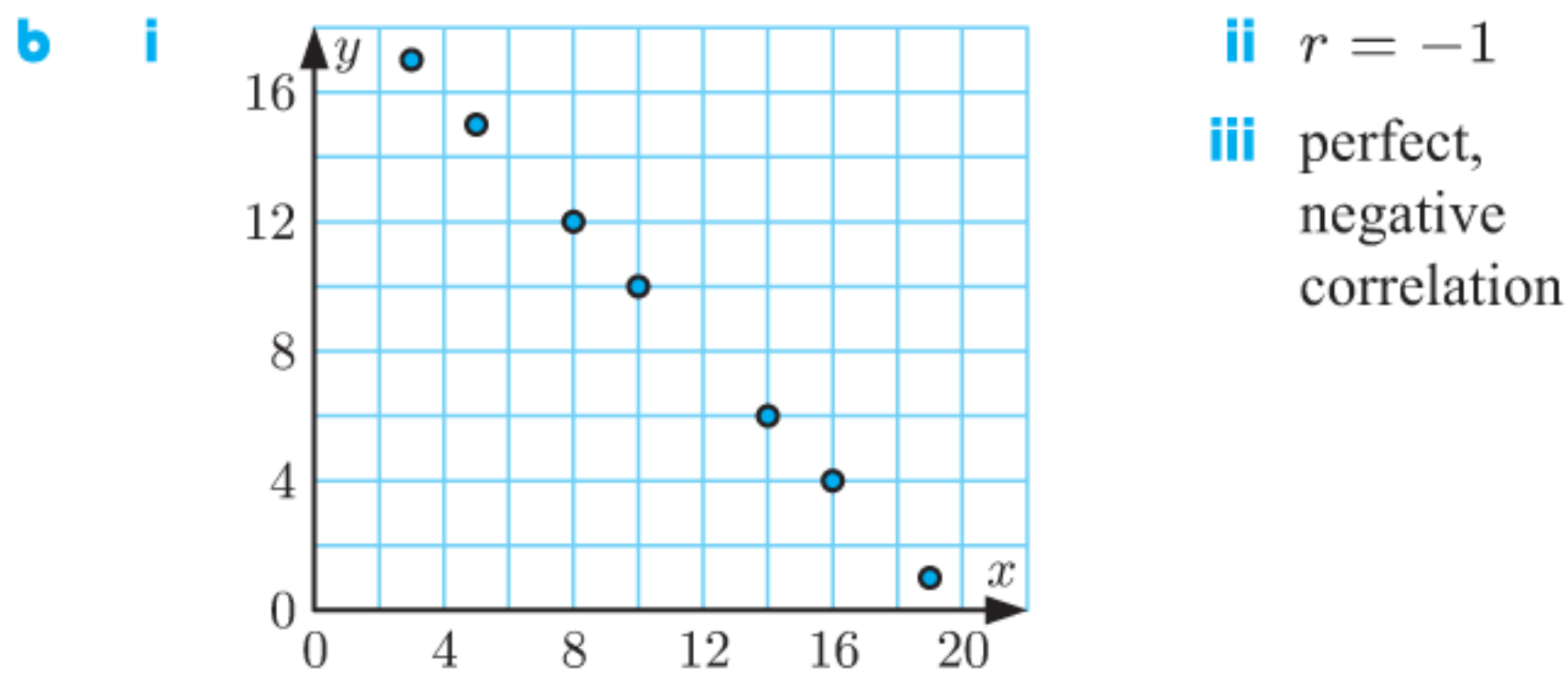
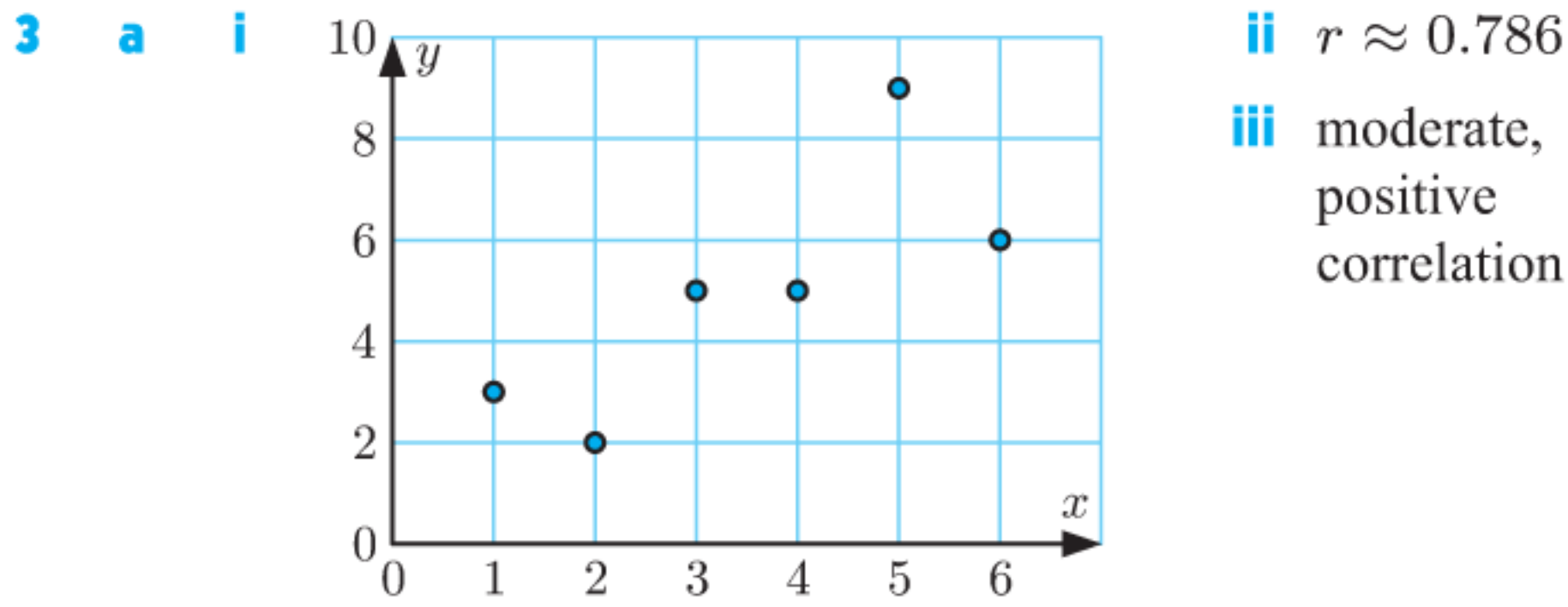
d Causal, the childrens' adult height is determined by the genetics they receive from their parents to a great extent.

e Not causal, dependent on population of town.

EXERCISE 26B

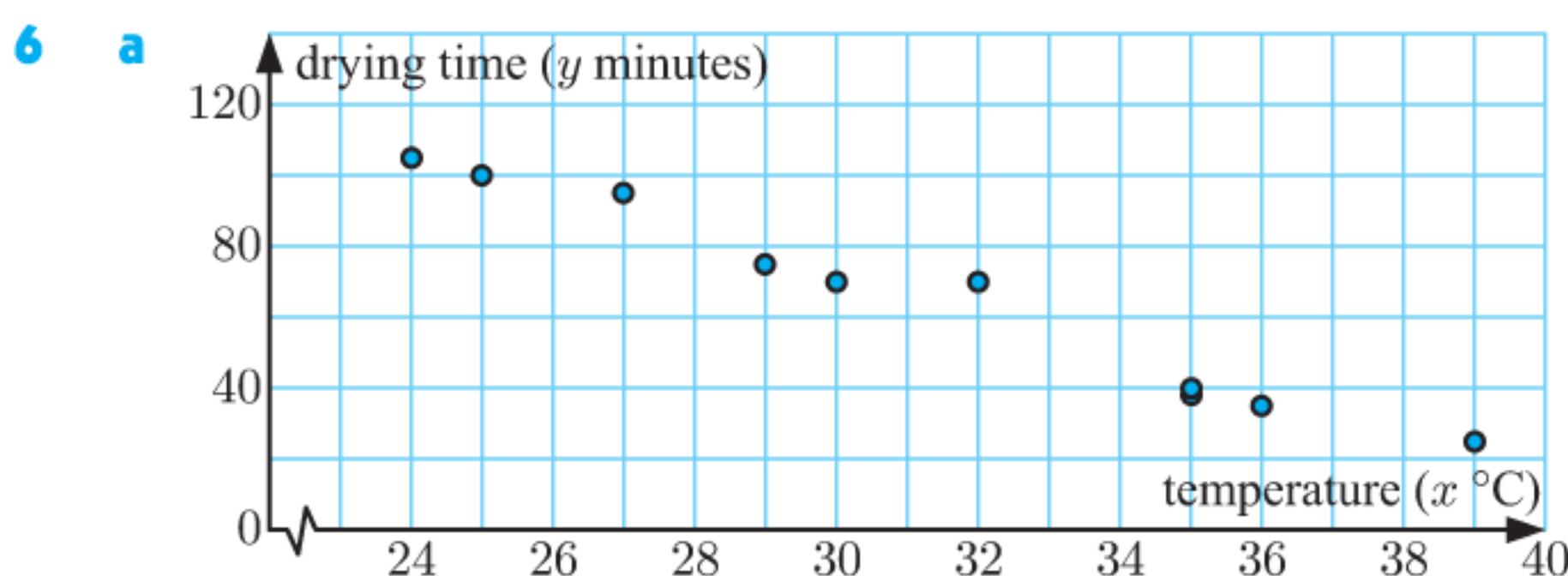
1 weak, positive correlation

2 a B b A c D d C e E

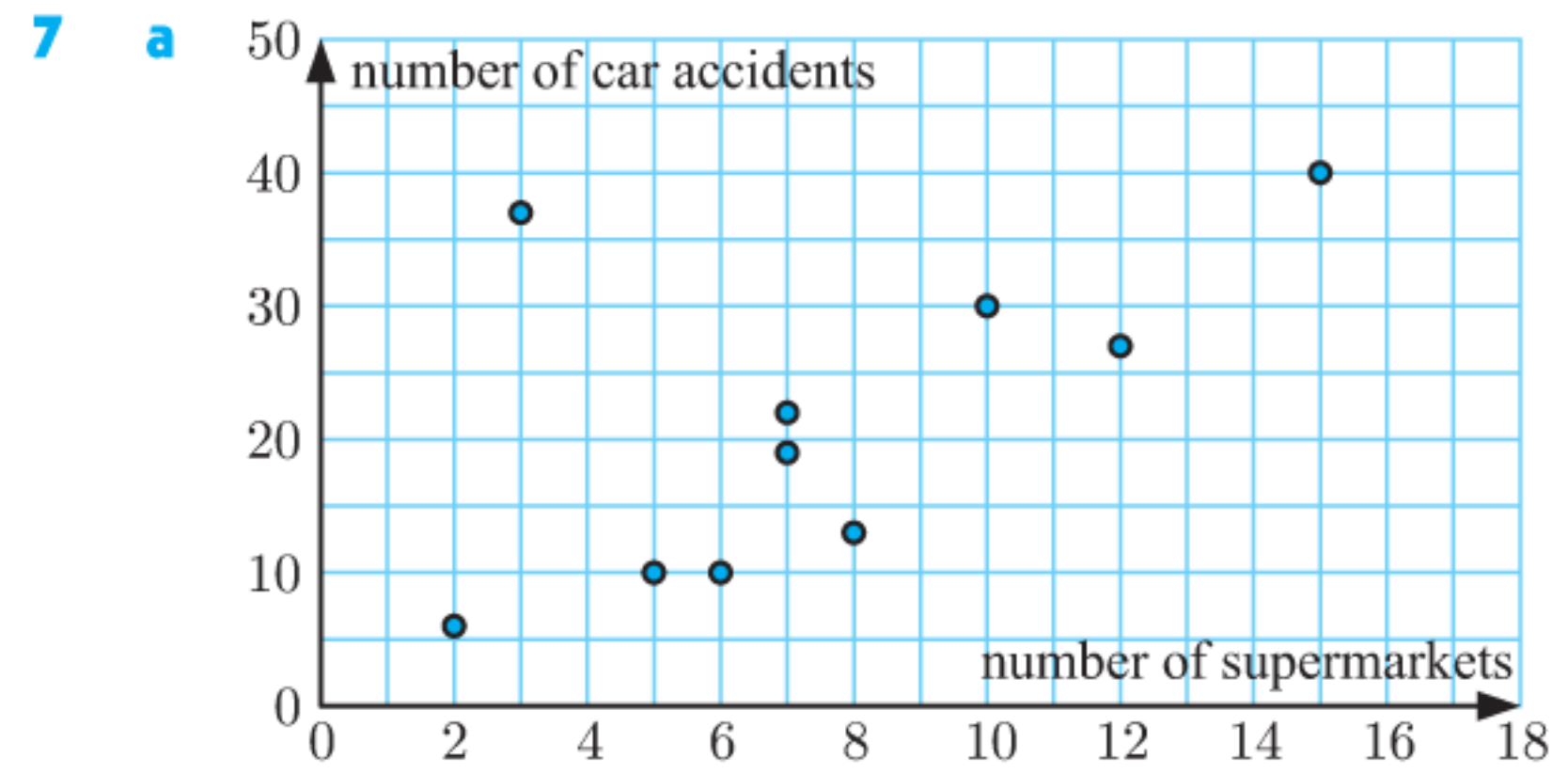


b $r \approx 0.816$
c moderate, positive correlation
d Those students who receive several phone calls are also likely to receive several text messages and vice versa.

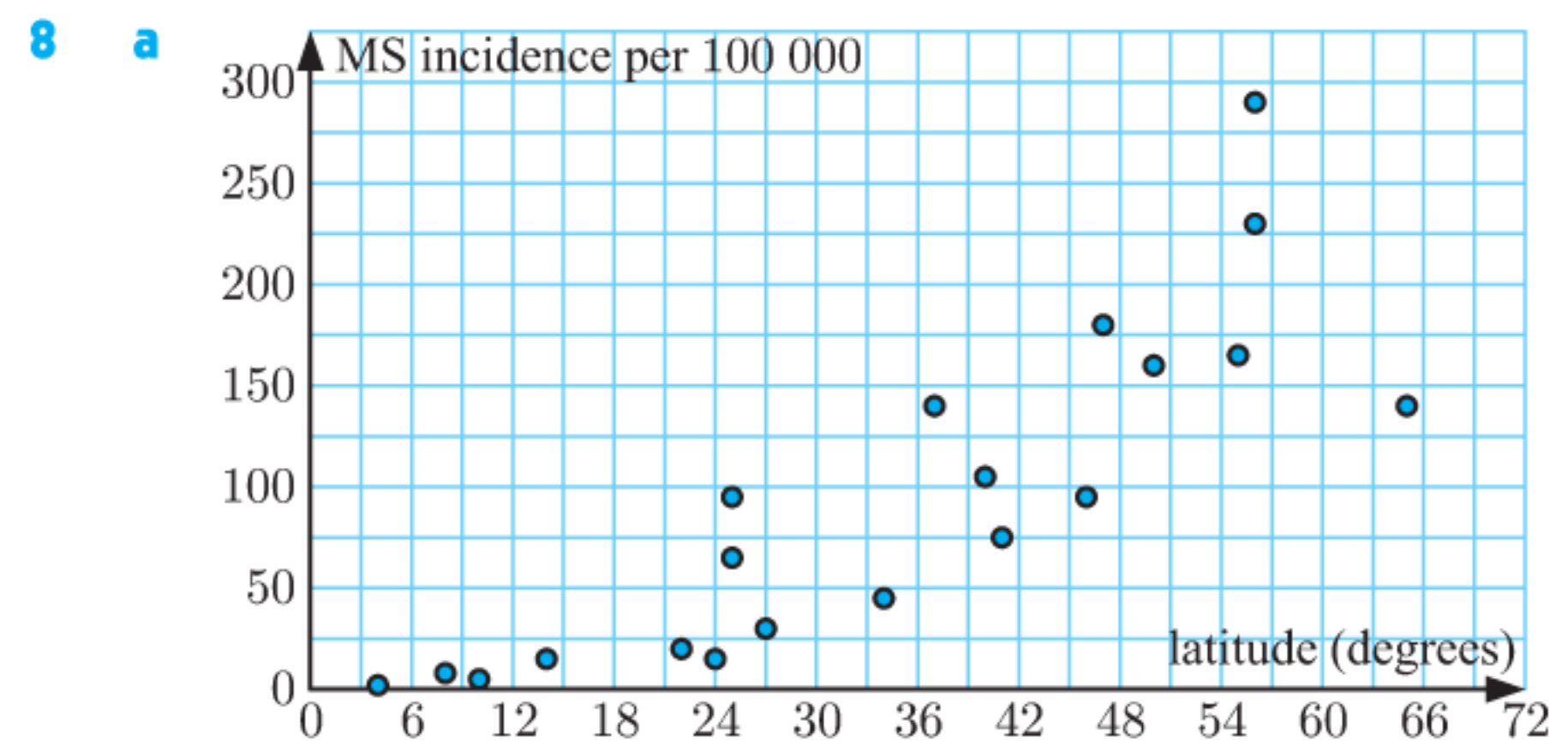
5 a $r \approx 0.917$
b strong, positive correlation
In general, the higher the young athlete's age, the further they can throw a discus.



b $r \approx -0.987$ c very strong, negative correlation

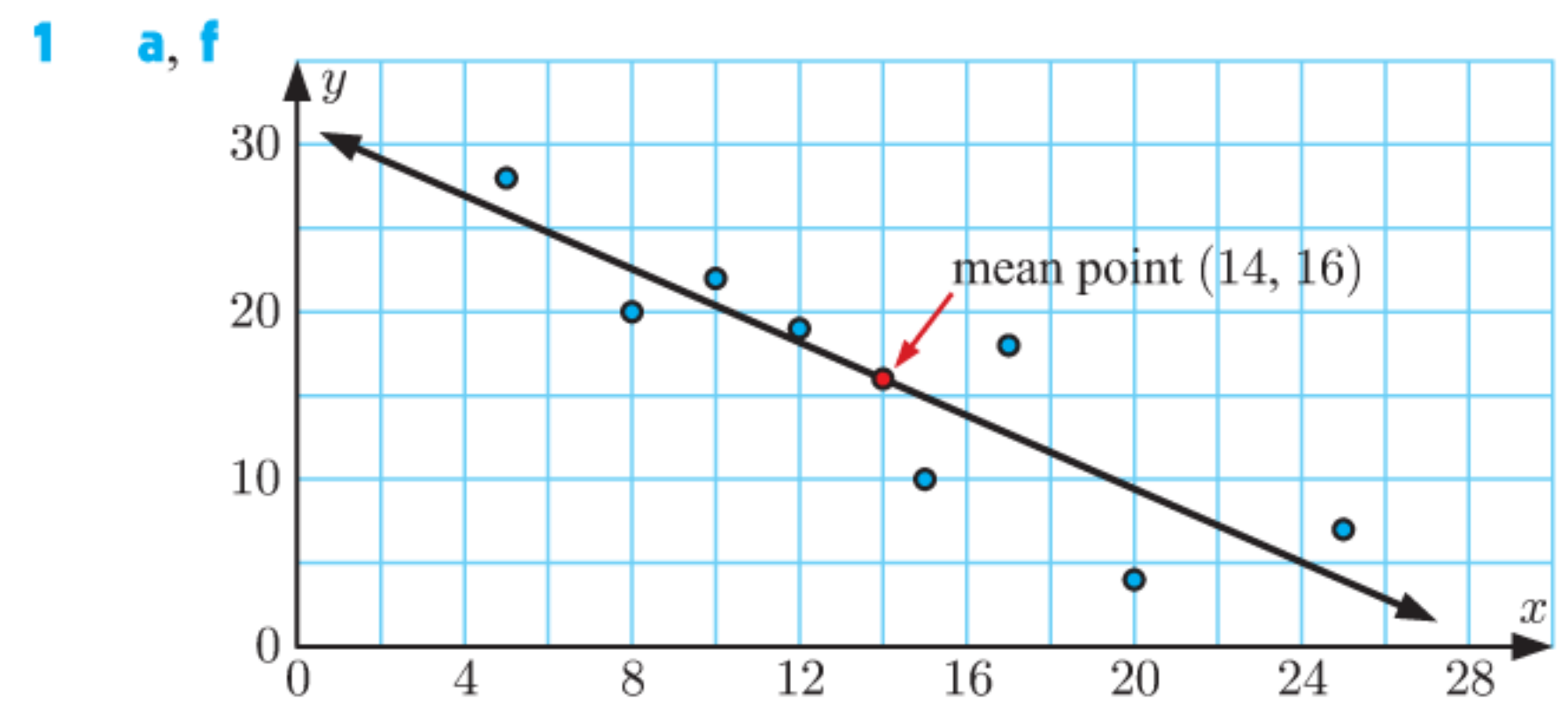


b $r \approx 0.572$
c The point (3, 37), which represents 37 car accidents in a town with 3 supermarkets, is an outlier.
d i $r \approx 0.928$
ii strong, positive correlation
iii Removing the outlier had a very significant effect on the value of r .
e No, it is not a causal relationship. Both variables depend on the number of people in each town, not on each other.

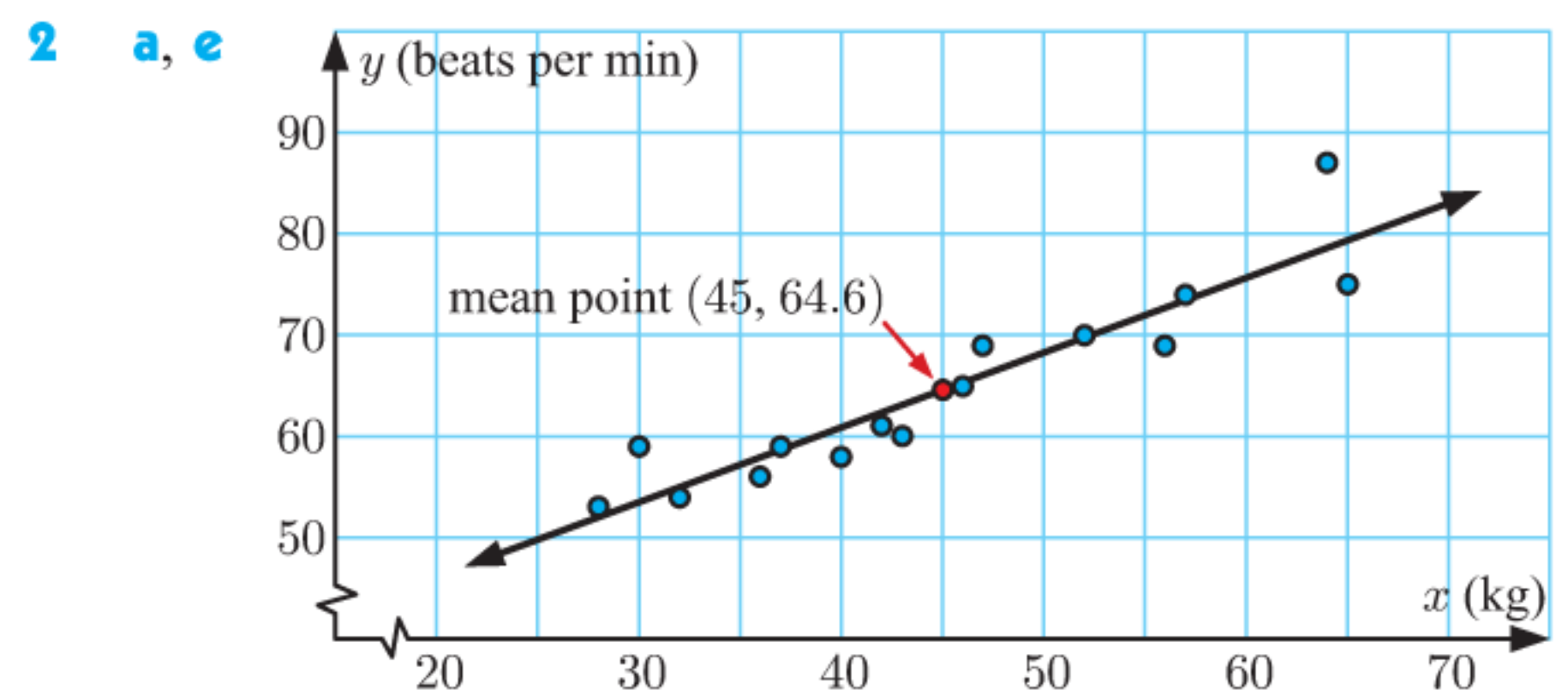


b $r \approx 0.849$ c moderate, positive correlation
d The incidence of MS is higher near the poles.

EXERCISE 26C

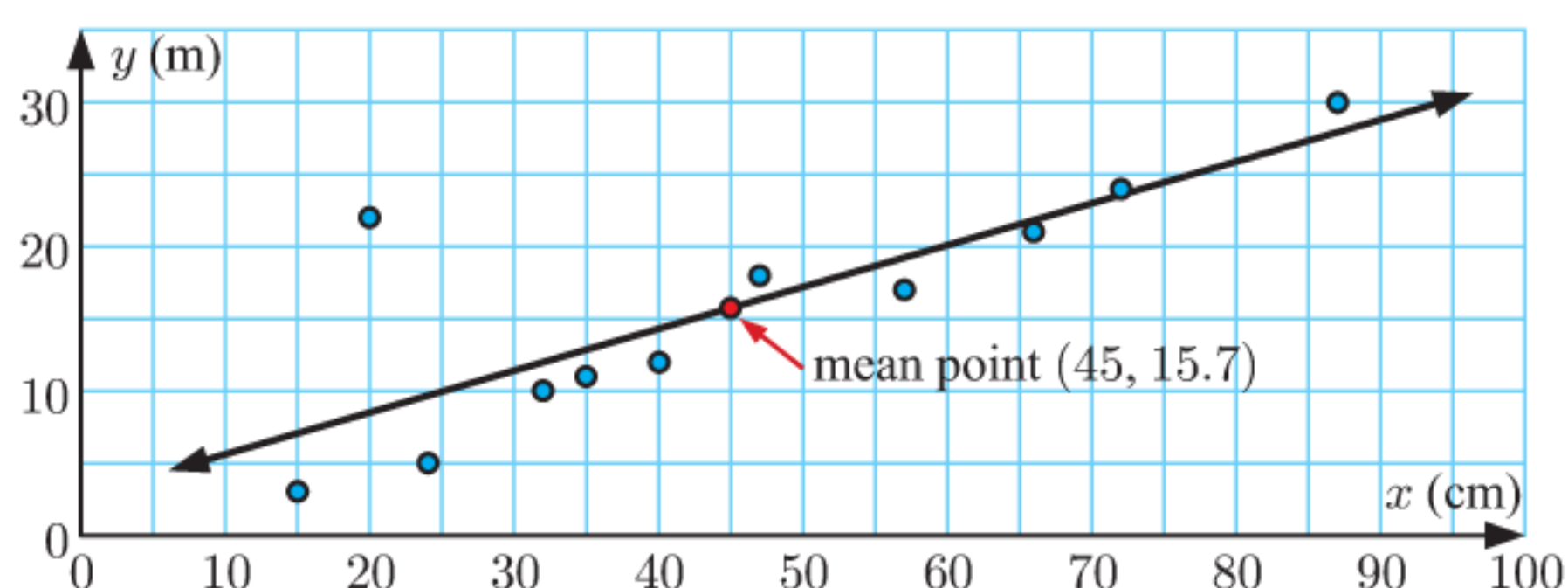


b negatively correlated c $r \approx -0.881$
d strong, negative correlation e (14, 16) g $y \approx 7$



b $r \approx 0.929$
c There is a strong, positive correlation between *weight* and *pulse rate*.
d (45, 64.6)
f 68 beats per minute. This is an interpolation, so the estimate is reliable.

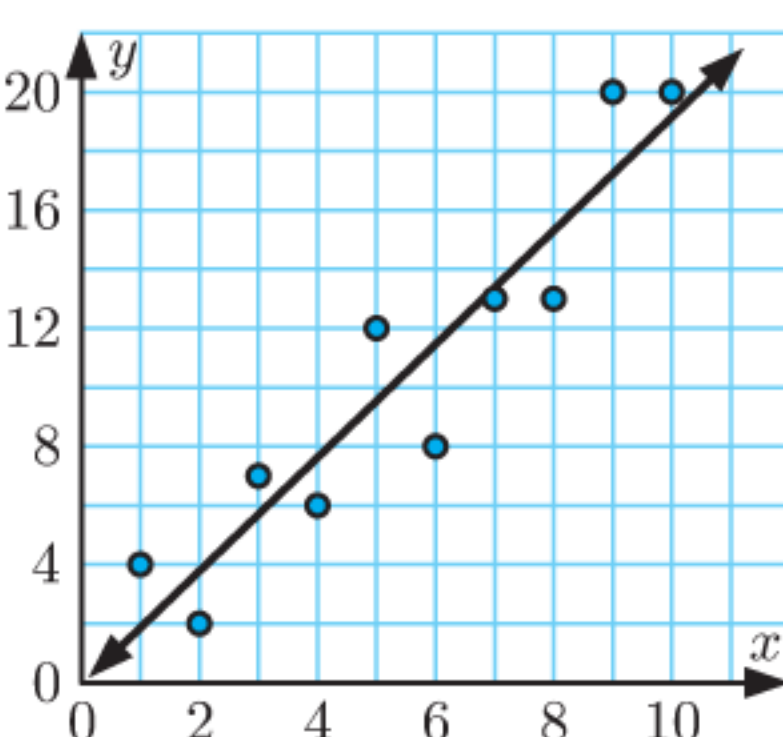
3 a, e



- b (20, 22) c very tall and thin d (45, 15.7)
 f ≈ 37 m. This is an extrapolation, so the prediction may not be reliable.
 g ≈ 25 cm. This is an interpolation, so the estimate is reliable.

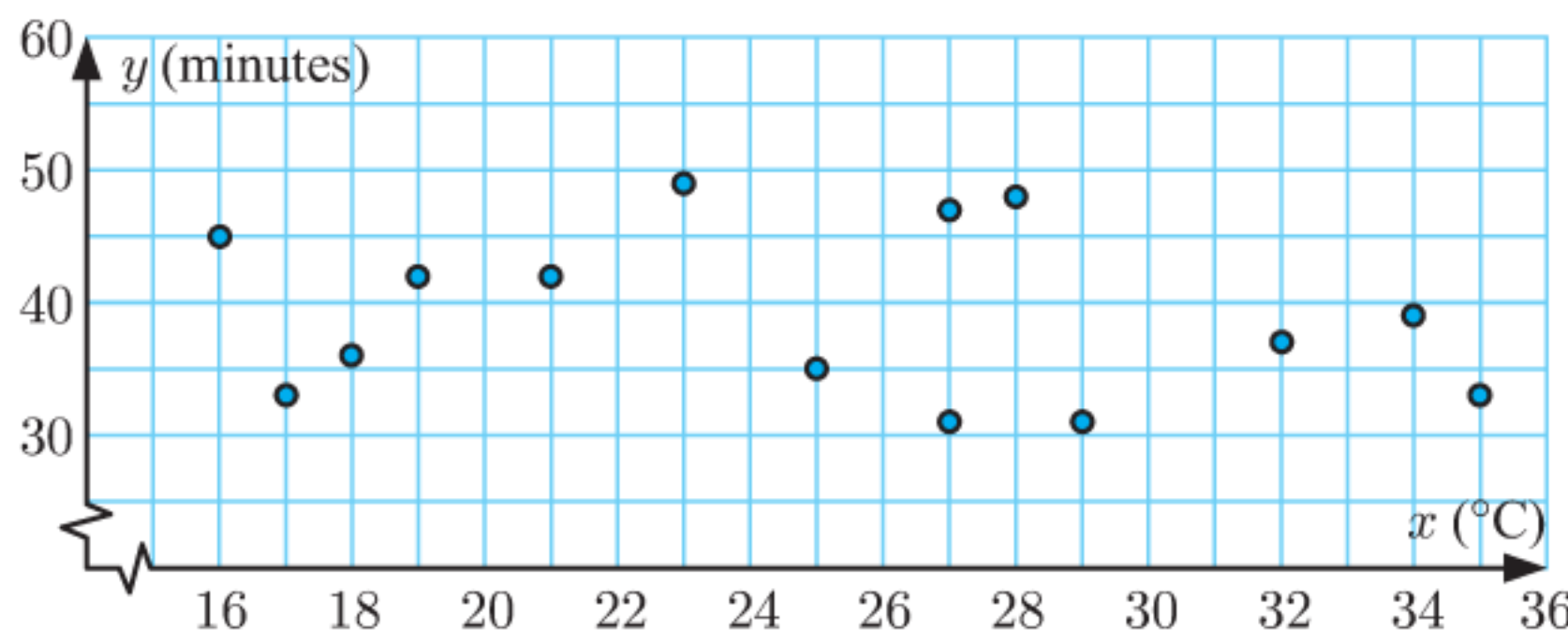
EXERCISE 26D

1 a, c



b $y \approx 1.92x - 0.0667$

2 a



- b $r \approx -0.219$
 c There is a very weak, negative correlation between temperature and time.
 d No, as there is almost no correlation.

3 a $r \approx -0.924$

b There is a strong, negative, linear correlation between the petrol price and the number of customers.

c $y \approx -4.27x + 489$

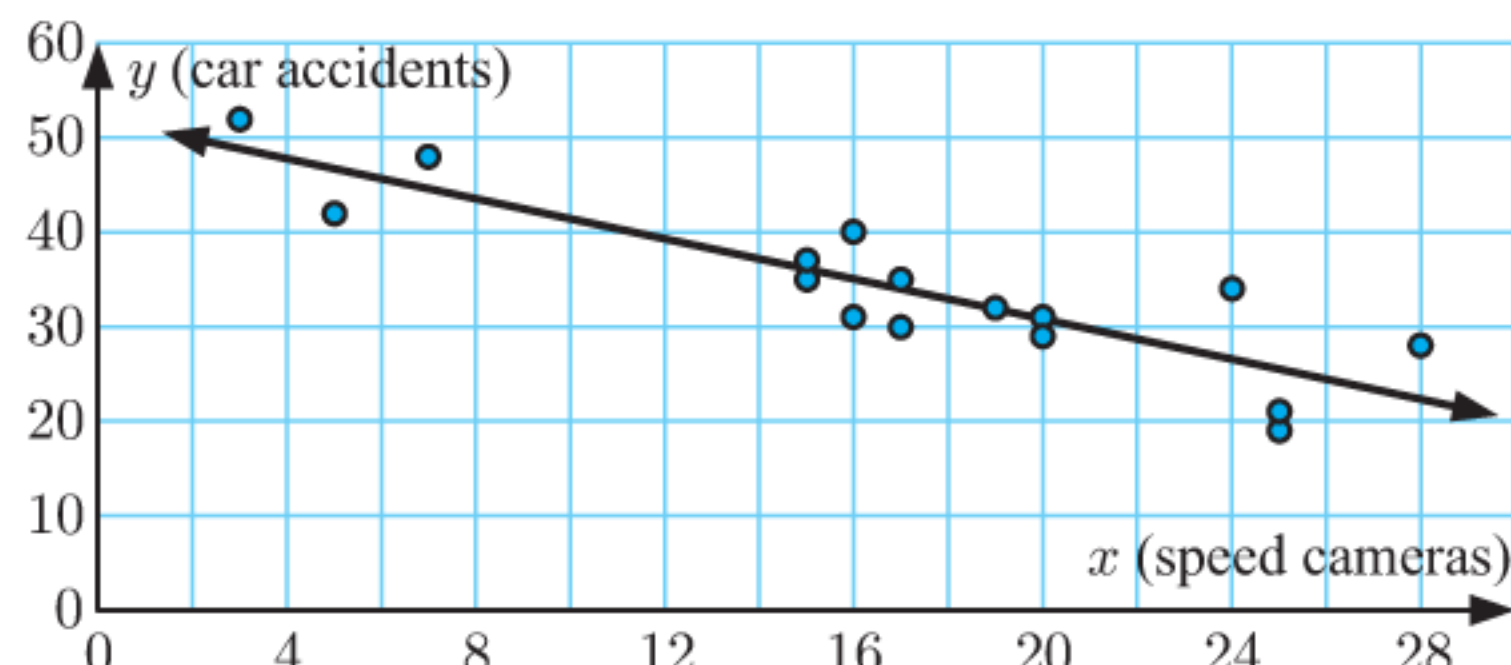
d ≈ -4.27 ; this indicates that for every cent per litre the petrol price increases by, the number of customers will decrease by approximately 4.27.

e ≈ -5.10 customers f ≈ 105.3 cents per litre

g In e, it is impossible to have a negative number of customers. This extrapolation is not valid.

In f, this is an interpolation, so this estimate is likely to be reliable.

4 a



b $r \approx -0.878$

c There is a strong, negative correlation between number of speed cameras and number of car accidents.

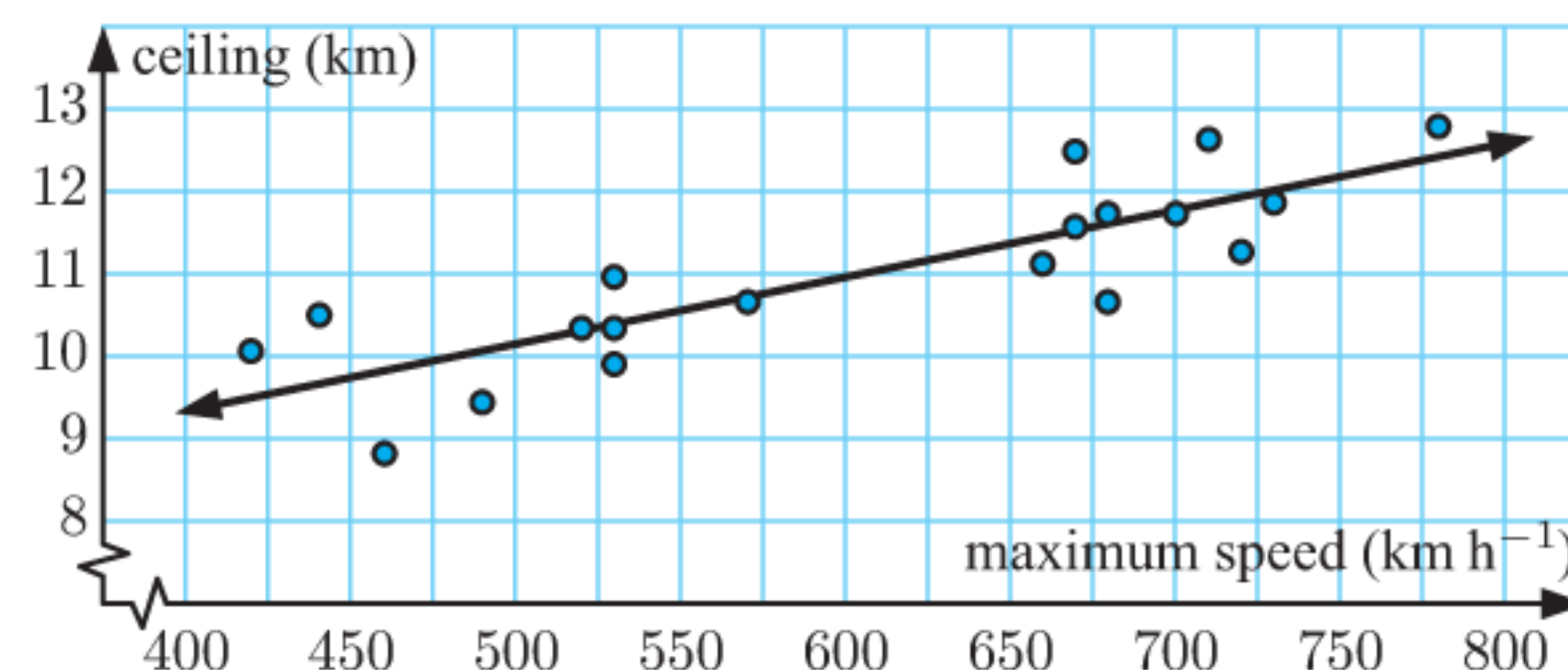
d $y \approx -1.06x + 52.0$

e gradient: ≈ -1.06 ; this indicates that for every additional speed camera, the number of car accidents per week decreases by an average of 1.06.

y-intercept: ≈ 52.0 ; this indicates that if there were no speed cameras in a city, an average of 52.0 car accidents would occur each week.

f ≈ 41.4 car accidents

5 a, d



b $r \approx 0.840$

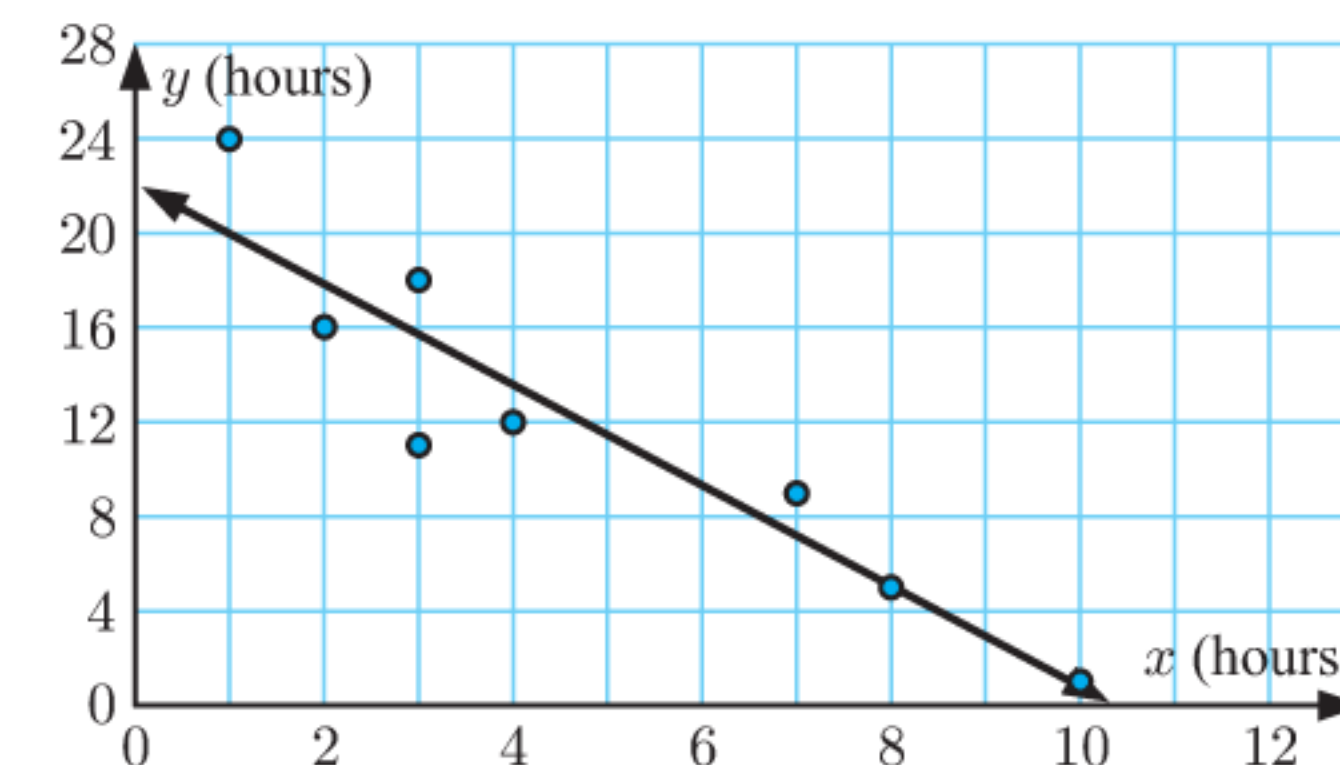
c moderate, positive, linear correlation

d $y \approx 0.00812x + 6.09$

e ≈ 0.00812 ; this indicates that for each additional km h^{-1} , the ceiling increases by an average of 0.00812 km or 8.12 m.

f ≈ 11.0 km g $\approx 605 \text{ km h}^{-1}$

6 a, d



b $r \approx -0.927$

c There is a strong, negative, linear correlation between time exercising and time watching television.

d $y \approx -2.13x + 22.1$

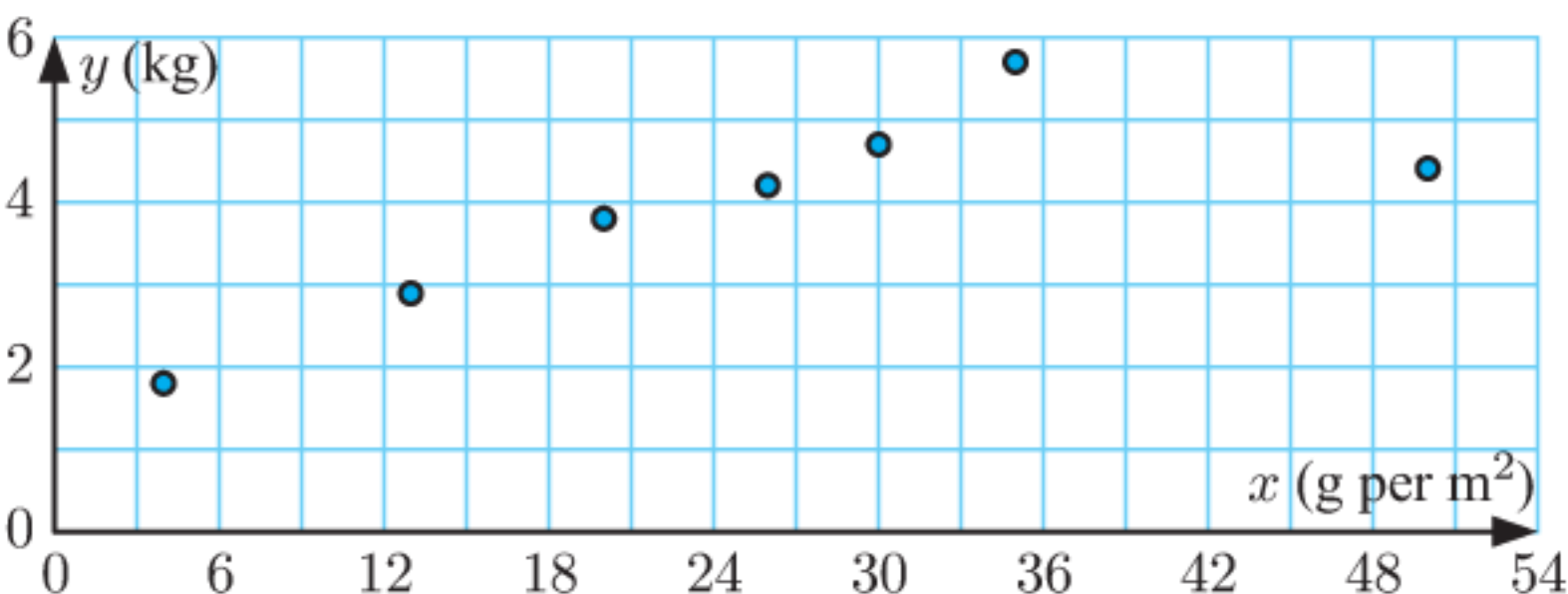
e gradient: ≈ -2.13 ; this indicates that for each additional hour a child exercises each week, the number of hours they spend watching television each week decreases by 2.13.

y-intercept: ≈ 22.1 ; this indicates that for children who do not spend time exercising, they would watch television for an average of 22.1 hours per week.

f i 9 hours per week ii ≈ 7.22 hours per week

iii This particular child spent more time watching television than predicted.

7 a



(50, 4.4) is the outlier.

b i reduces the strength of the correlation

ii decreases the gradient of the regression line

c i $r \approx 0.798$

ii $r \approx 0.993$

d i $y \approx 0.0672x + 2.22$

ii $y \approx 0.119x + 1.32$

e The one which excludes the outlier, as this will be more accurate for an interpolation.

f Too much fertiliser often kills the plants. In this case, the outlier should be kept when analysing the data as it is a valid data value. If the outlier is a recording error caused by bad measurement or recording skills, it should be removed before analysing data.

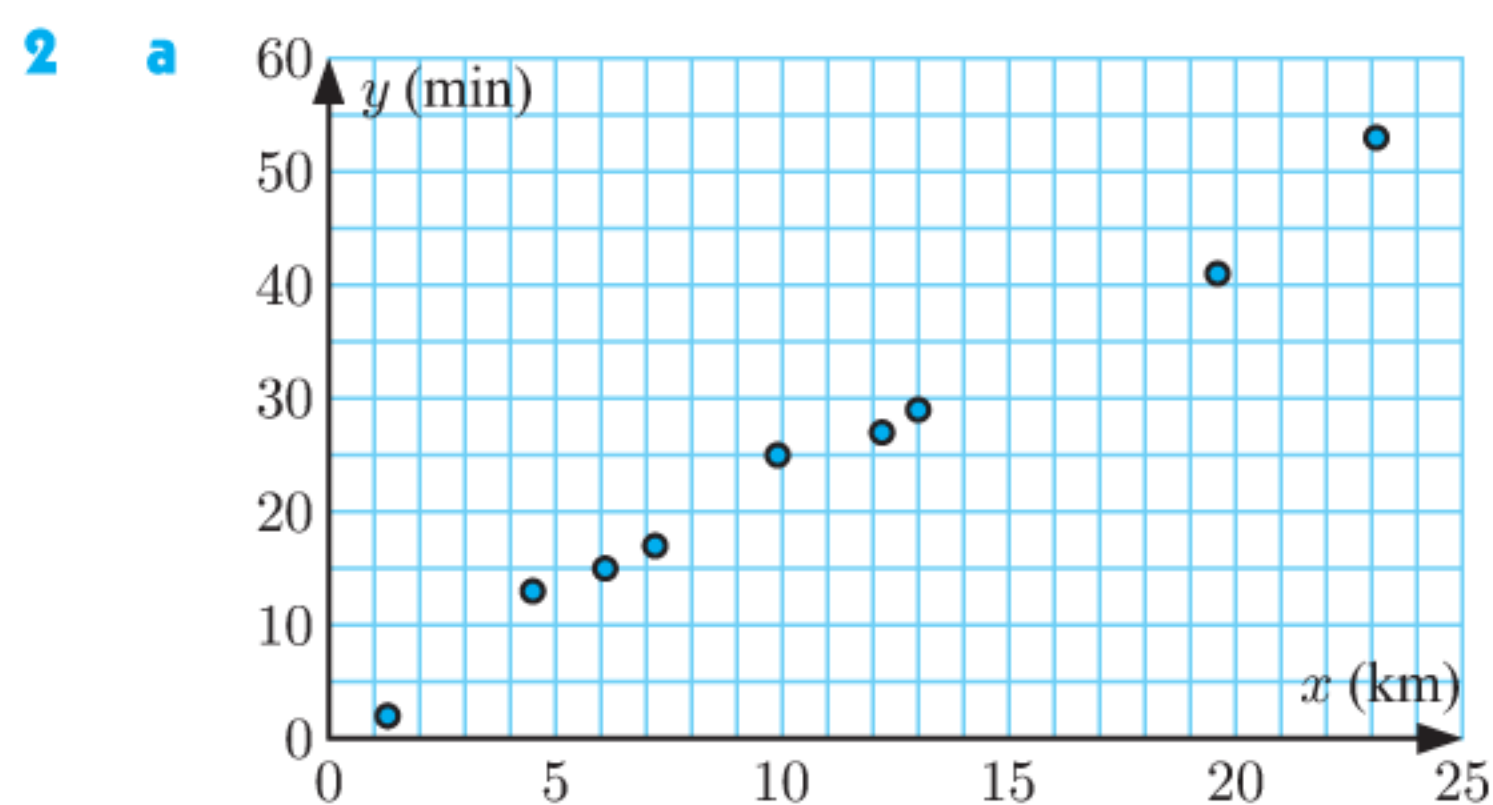
$$8 \quad b \quad \beta = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

EXERCISE 26E

1 a The y variable, money spent on fast food, can be measured exactly. The x variable, time spent on homemade meals, will not be measured exactly.

b $x \approx -0.0576y + 8.29$

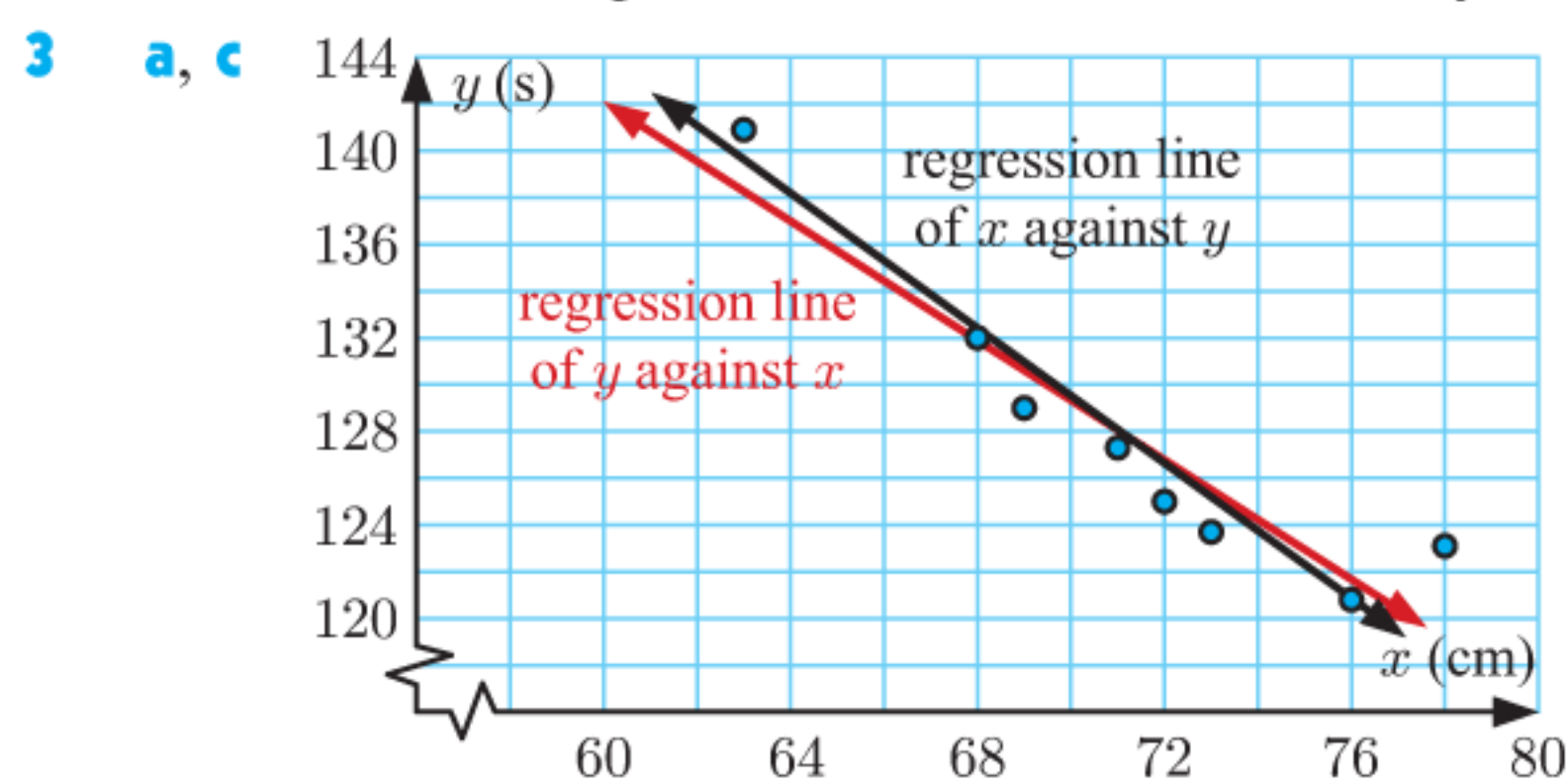
c i ≈ 5.70 hours ii $\approx \$57.13$



b x against y , since a student's time taken to travel to school can be more precisely measured than their distance from school.

c ≈ 33.9 min

d This is an interpolation, so this estimate is likely to be reliable.



b i $y \approx -1.28x + 219$

ii $x \approx -0.693y + 160$

c The two regression lines are very similar. The regression line of x against y is slightly steeper.

4 b The regression lines are the same if $r^2 = 1$.

REVIEW SET 26A

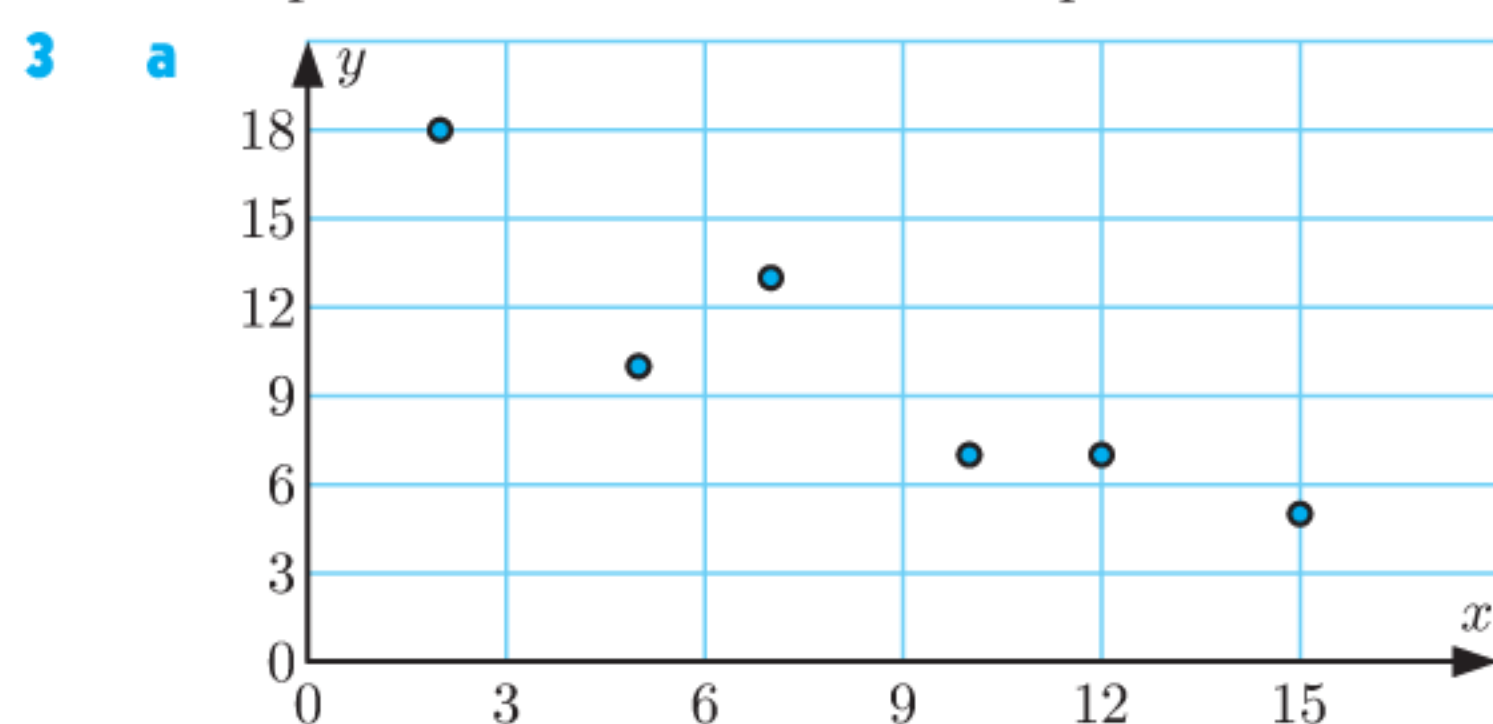
1 a strong, positive, linear correlation, with no outliers

b weak, negative, linear correlation, with one outlier

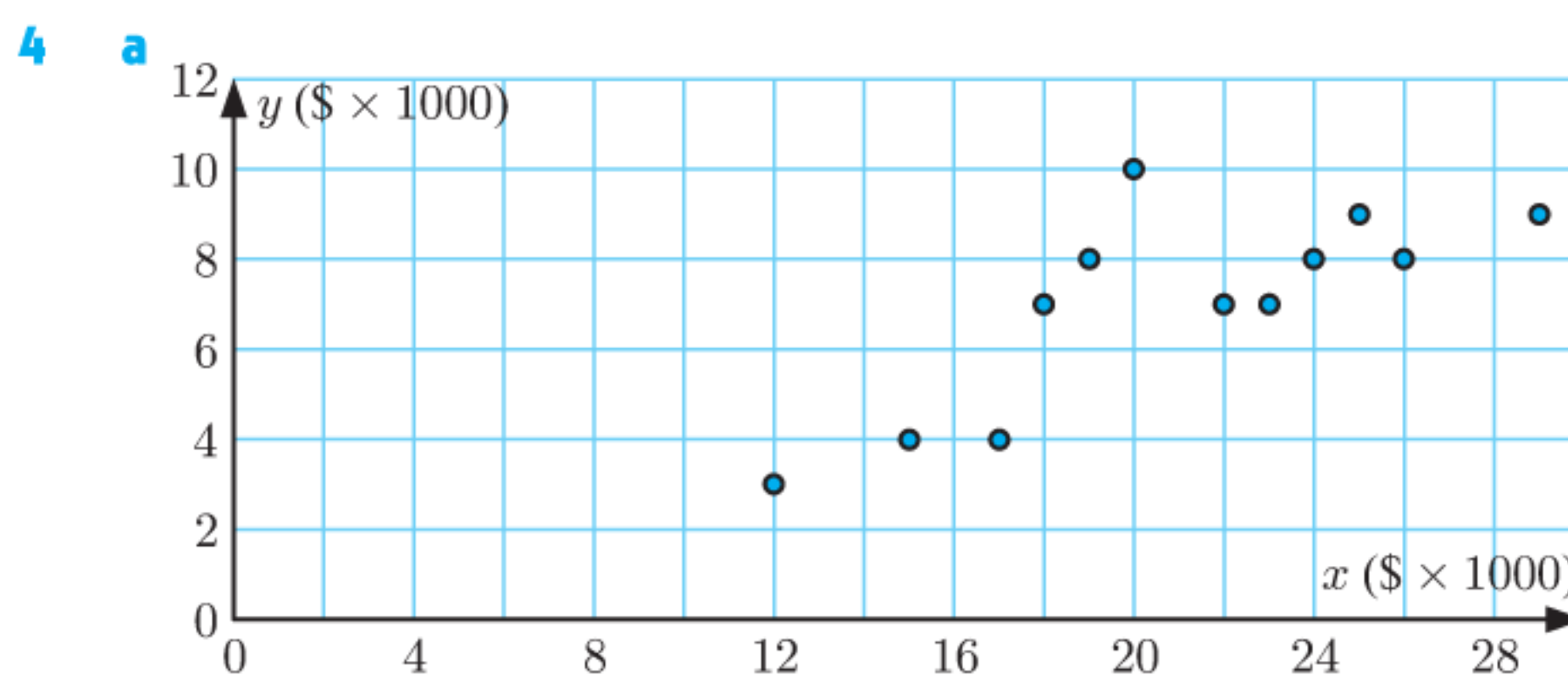
c strong, negative, non-linear correlation, with no outliers

2 a The correlation between water bills and electricity bills is likely to be positive, as a household with a high water bill is also likely to have a high electricity bill, and vice versa.

b No, there is not a causal relationship. Both variables mainly depend on the number of occupants in each house.

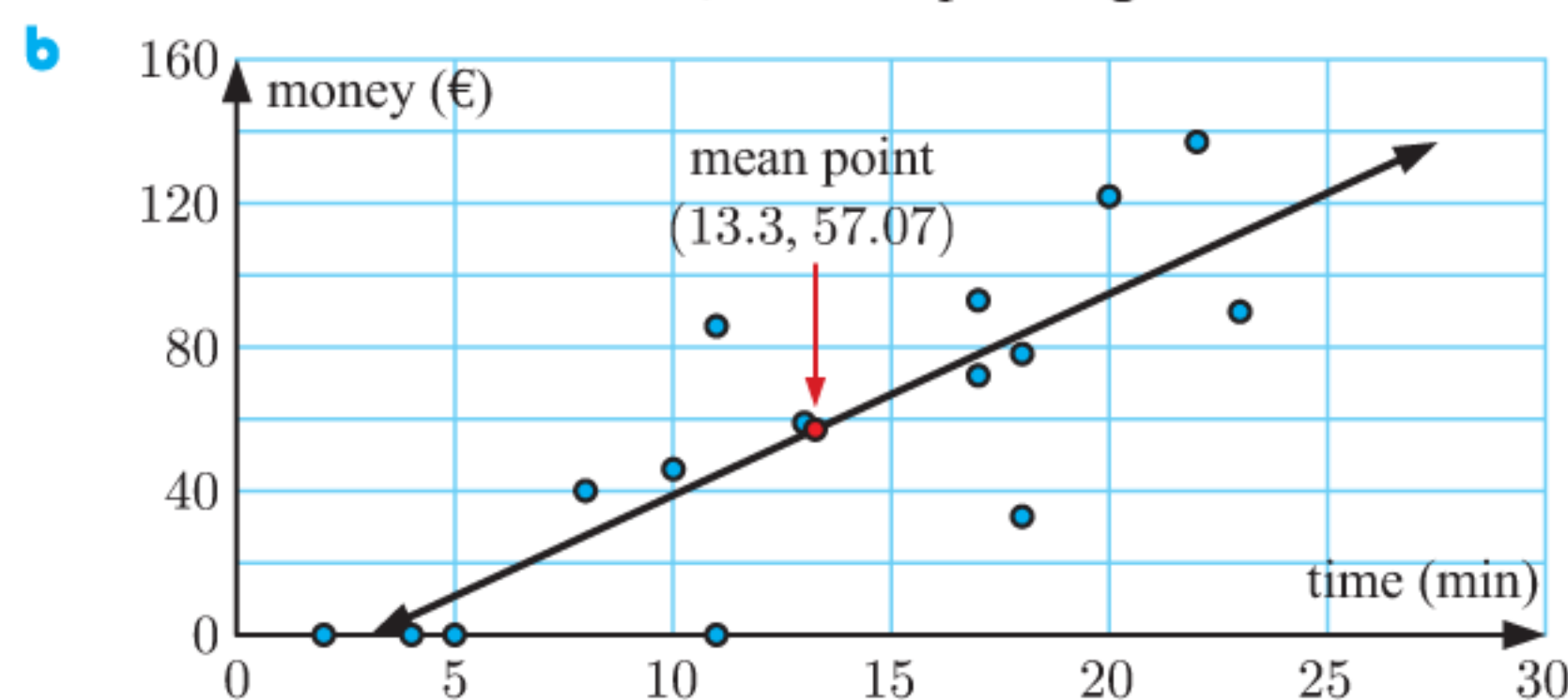


b negative c $r \approx -0.906$

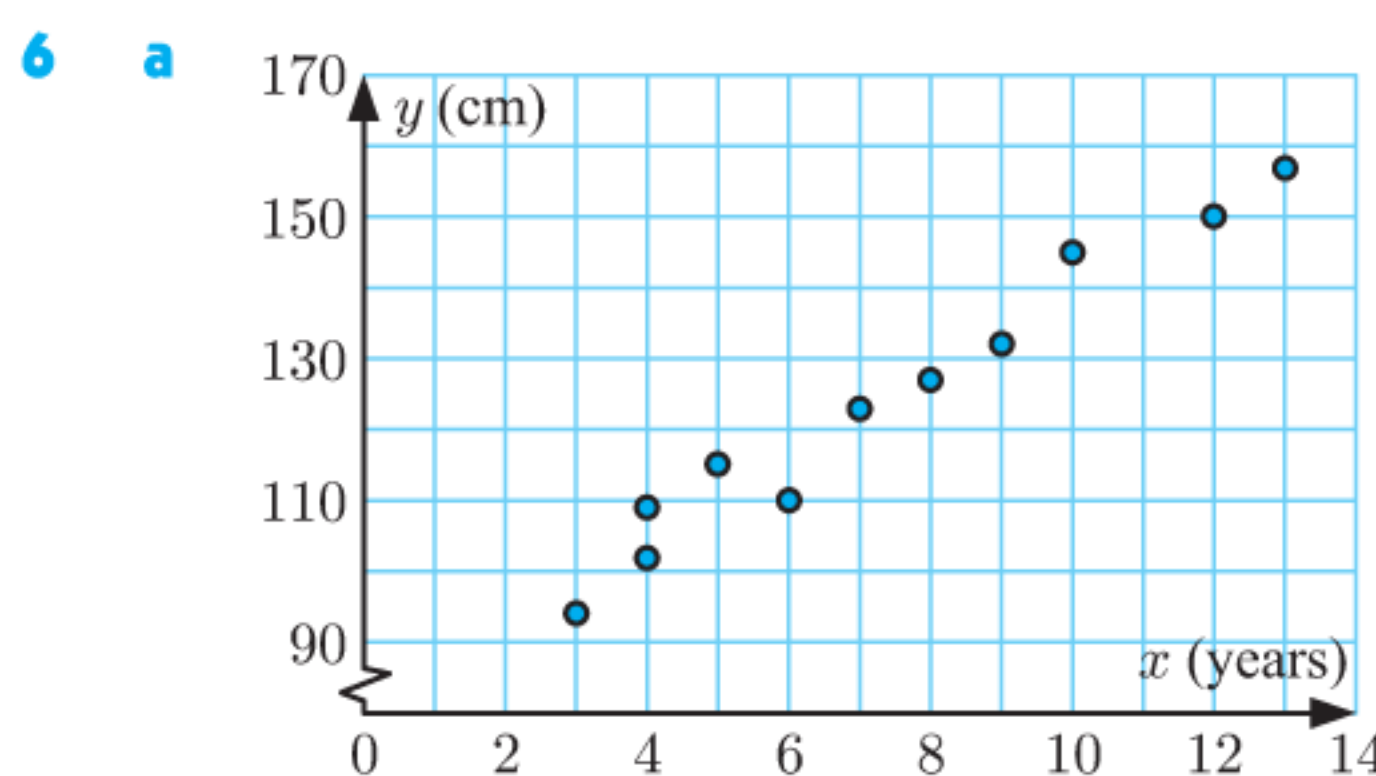


b $r \approx 0.776$ c moderate, positive correlation

5 a mean time ≈ 13.3 min, mean spending $\approx €57.07$



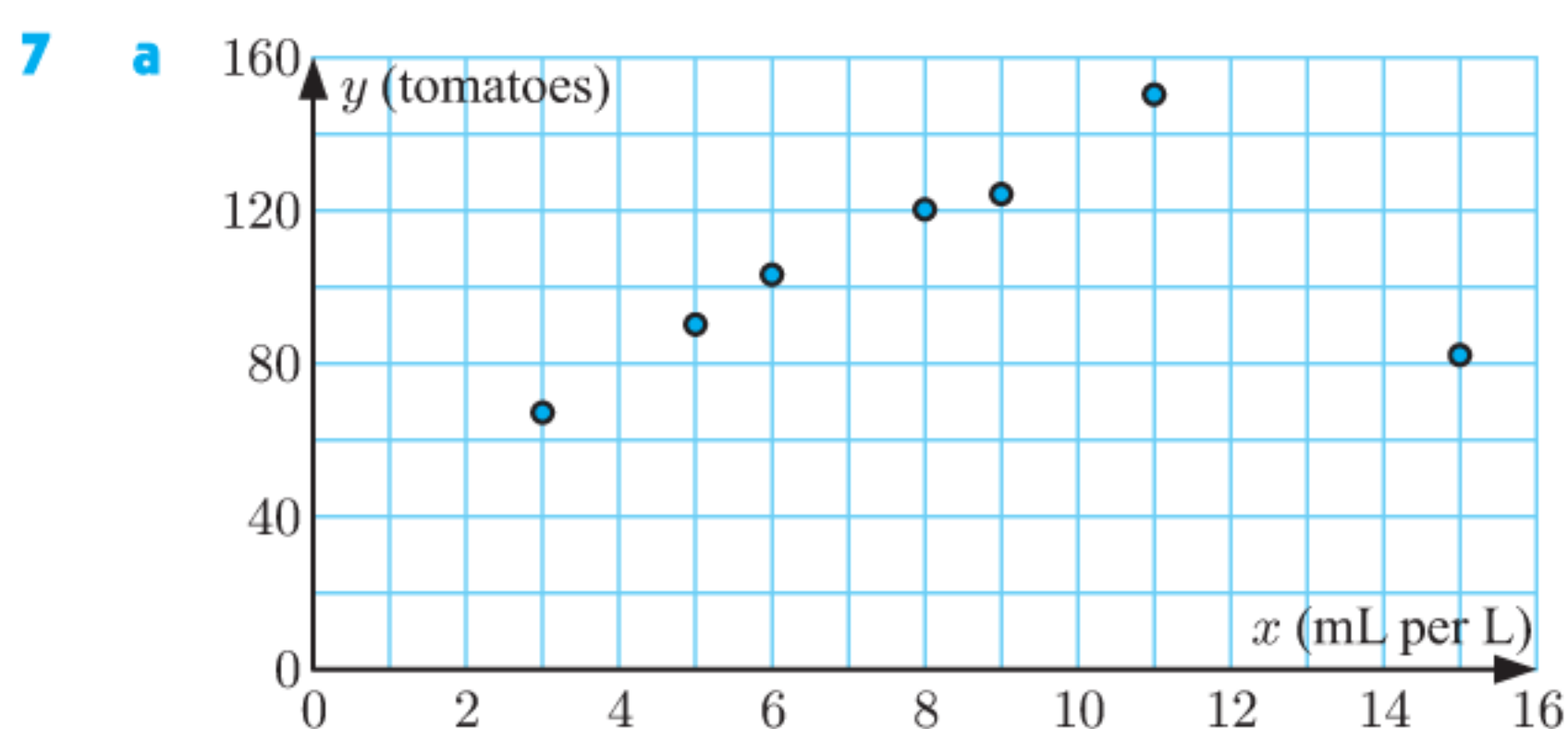
c There is a moderate positive linear correlation between *time in the store* and *money spent*.



b $y \approx 5.98x + 80.0$

c ≈ 5.98 ; this indicates that each year, a child grows taller by an average of 5.98 cm.

d ≈ 110 cm e 10 years old



b $r \approx 0.340$. There is a very weak, positive, linear correlation between spray concentrations and yield.

c Yes, (15, 82) is an outlier.

d $r \approx 0.994$. Yes it is now reasonable to draw a regression line.

e $y \approx 9.93x + 39.5$

f gradient: ≈ 9.93 ; this indicates that for every additional mL per L the spray concentration increases, the yield of tomatoes per bush increases on average by 9.93.

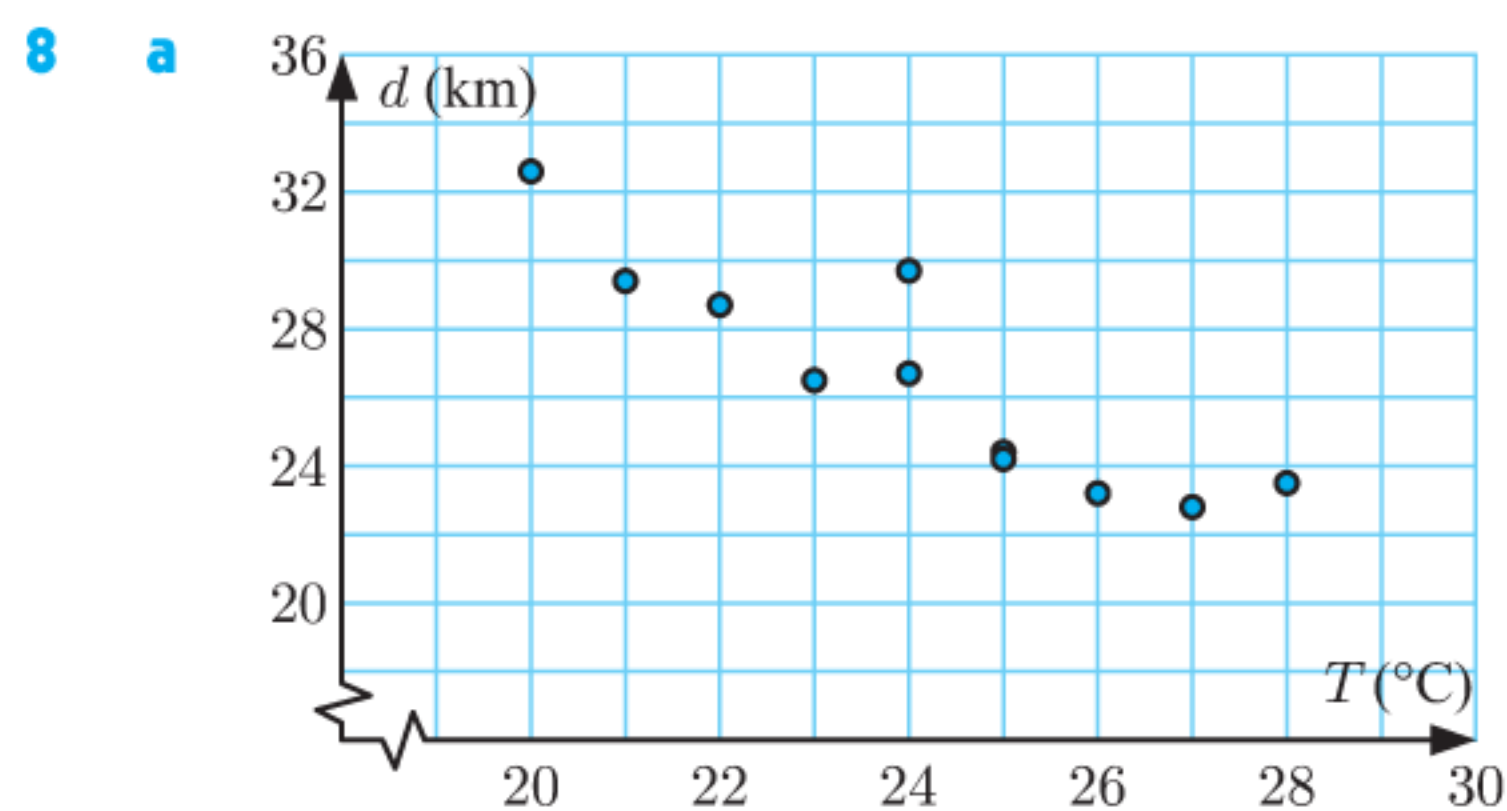
y-intercept: ≈ 39.5 ; this indicates that if the tomato bushes are not sprayed, the average yield per bush is approximately 39.5 tomatoes.

g i ≈ 109 tomatoes per bush

ii ≈ 16.2 mL per L

h In **g i**, this is an interpolation, so this estimate is likely to be reliable.

In **g ii**, this is an extrapolation, so this estimate may not be reliable.



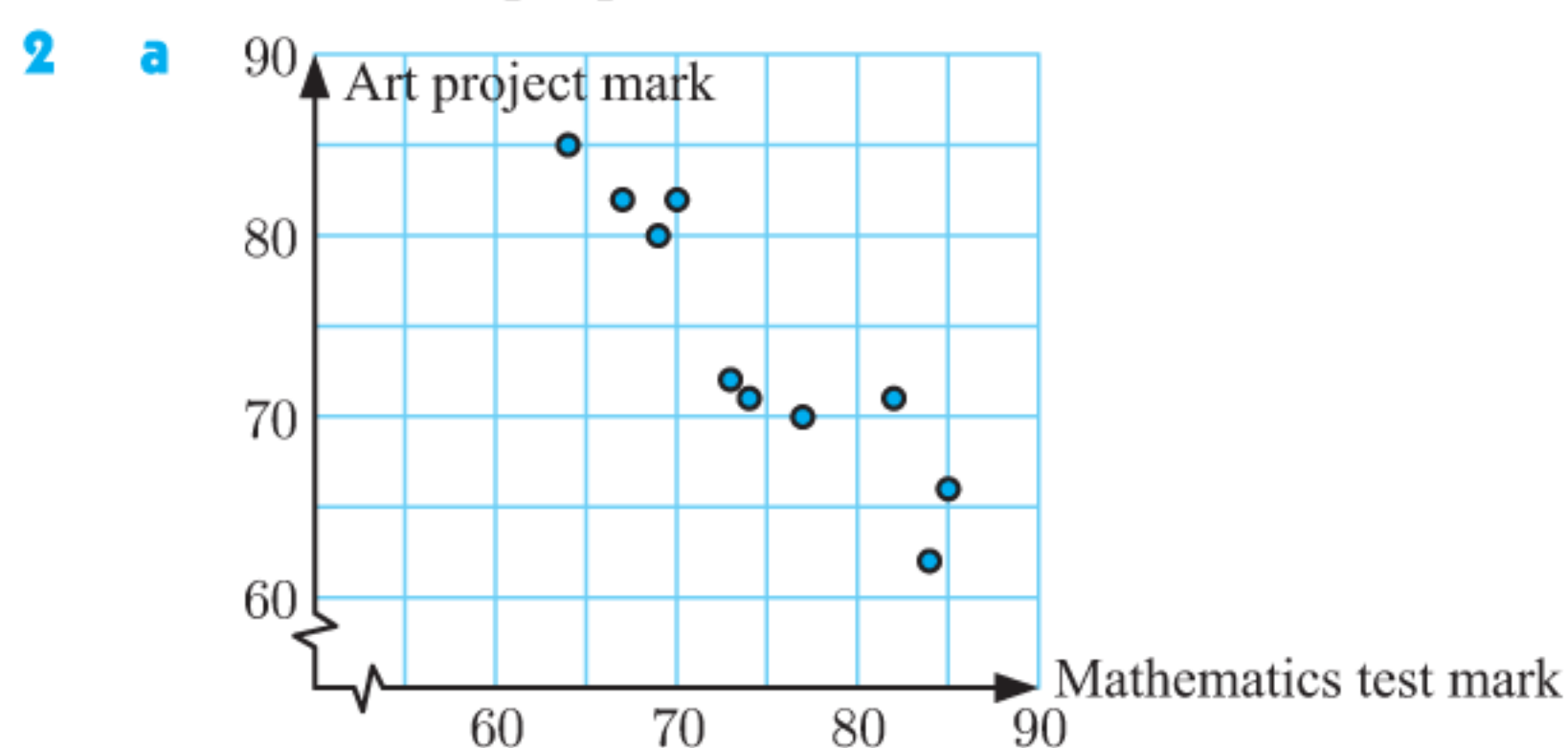
b The values for the distance travelled d are more precisely measured than the daily temperature which Thomas is just estimating.

c $T \approx -0.689d + 42.3$ **d** ≈ 17.9 km

REVIEW SET 26B

1 a Negative correlation. As prices increase, the number of tickets sold is likely to decrease.
Causal. Less people will be able to afford tickets as the prices increase.

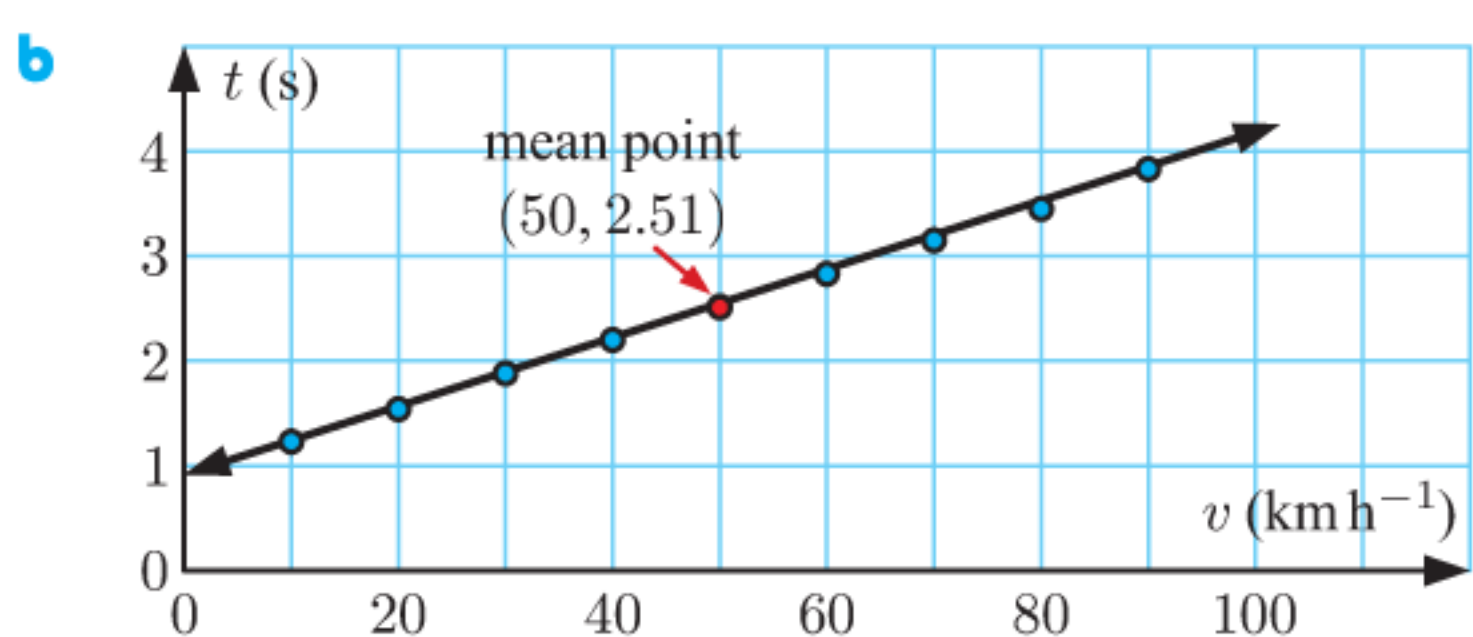
b Positive correlation. As ice cream sales increase, the number of shark attacks is likely to increase.
Not causal. Both of these variables are dependent on the number of people at the beach.



b There is a strong, negative, linear correlation between the Mathematics and Art marks.

c $r \approx -0.930$

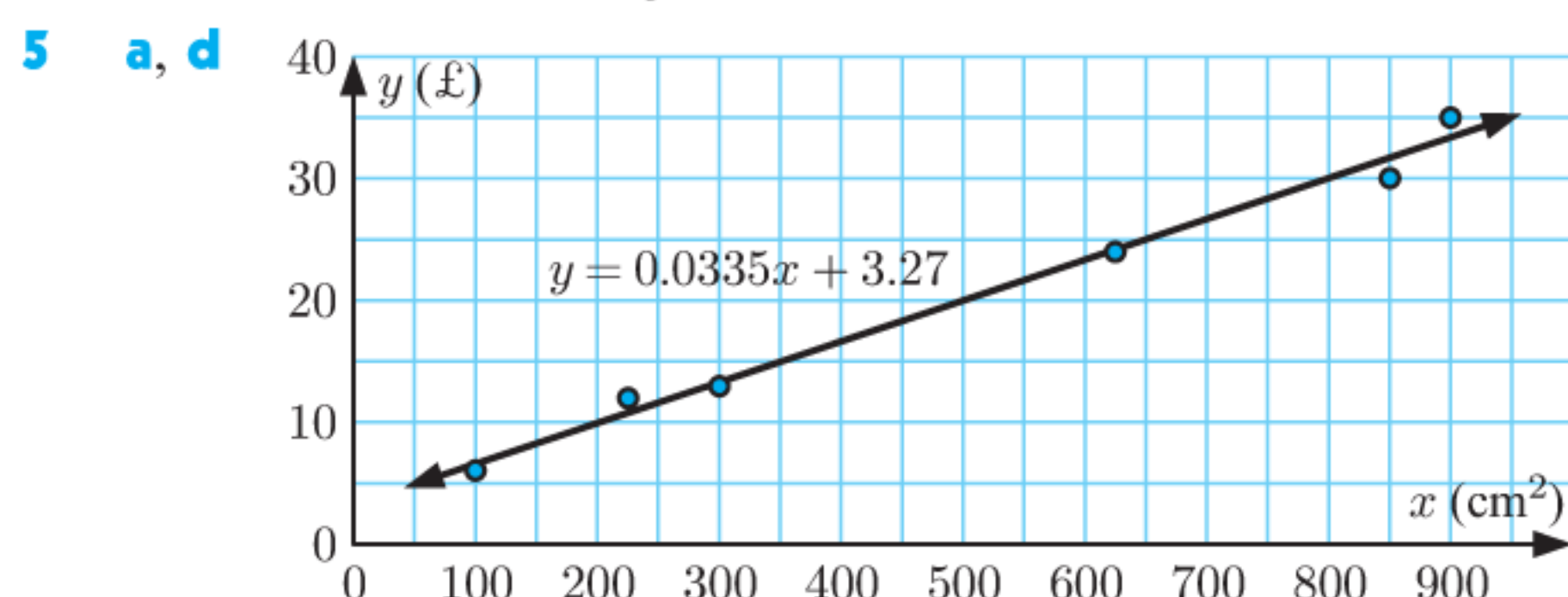
3 a (50, 2.51)



c i ≈ 2.68 seconds **ii** ≈ 4.44 seconds

d The estimate in **c i**, since it is an interpolation.

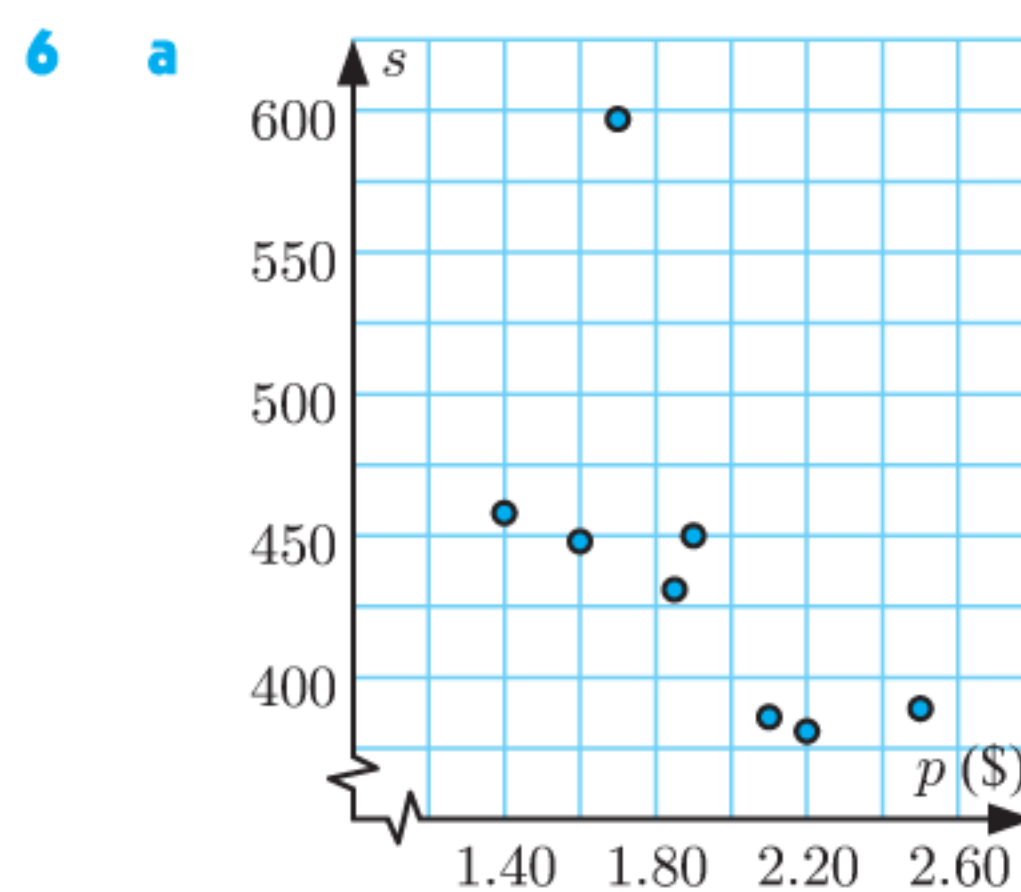
4 a $r \approx 0.983$ **b** $y \approx 3.36x + 8.64$ **c** ≈ 42.2



b $r \approx 0.994$

c There is a very strong, positive correlation between *area* and *price*.

e $\approx £43.42$, this is an extrapolation, so it may be unreliable.

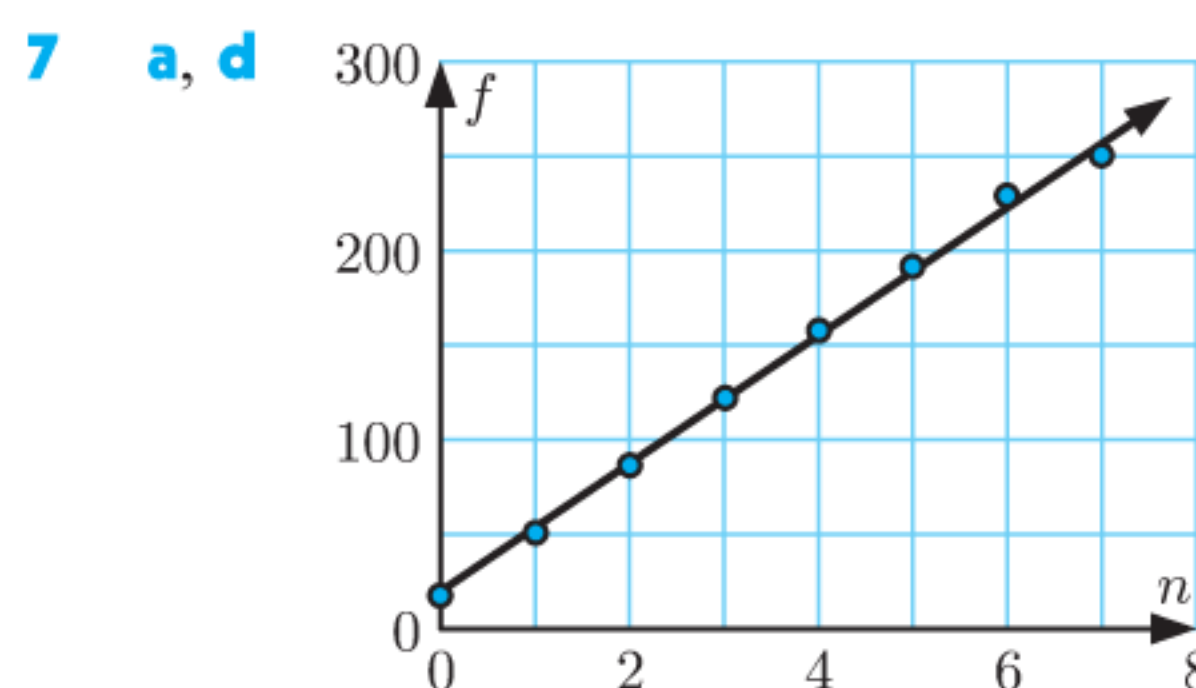


b Yes, the point (1.7, 597) is an outlier. It should not be deleted as there is no evidence that it is a mistake.

c $s \approx -116p + 665$

d ≈ -116 ; this indicates that with every additional dollar the price increases by, the number of sales decreases by 116.

e No, the prediction would not be accurate, as it is an extrapolation.



There is a very strong, positive correlation between number of waterings and flowers produced.

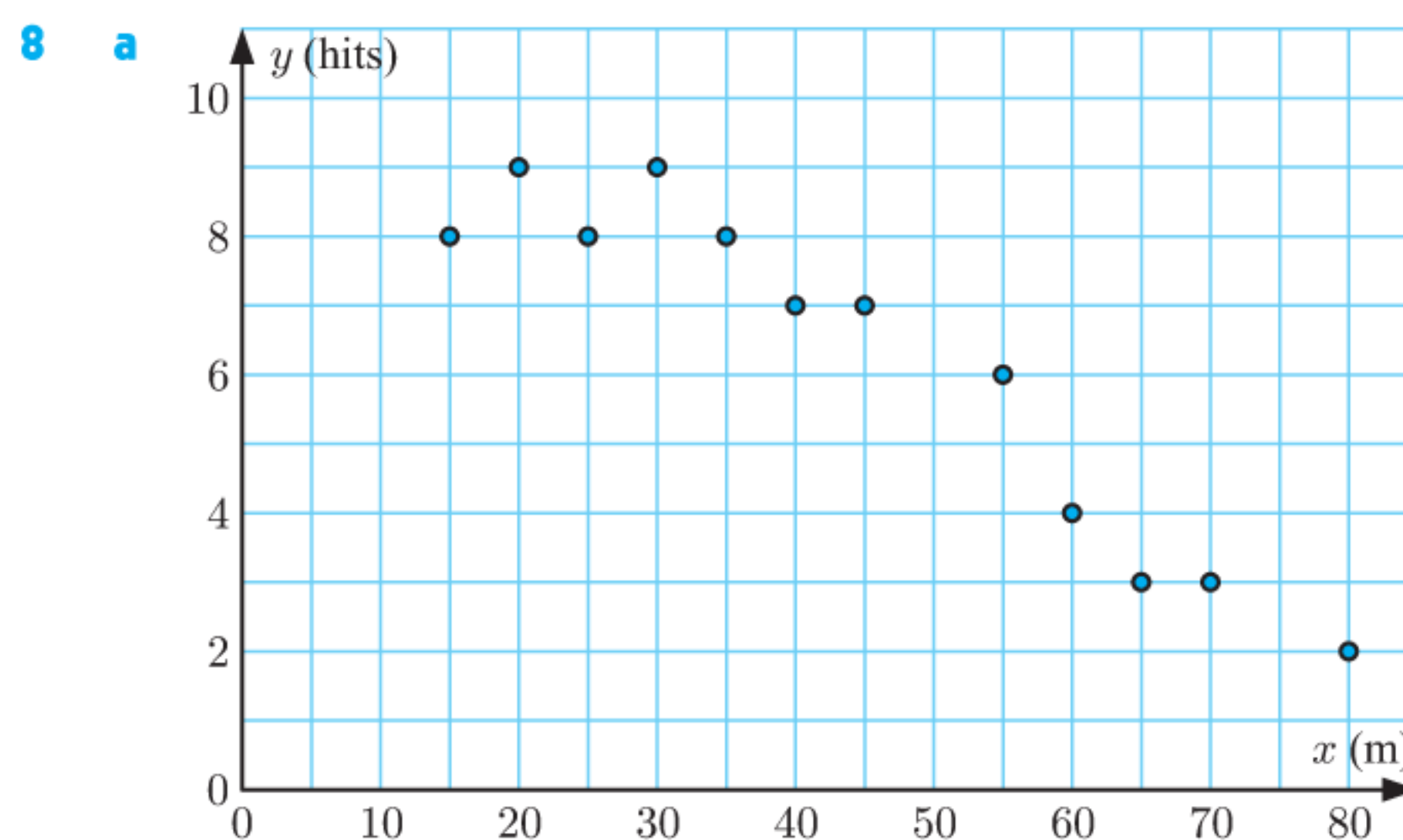
b $f \approx 34.0n + 19.3$

c Yes, plants need water to grow, so it is expected that an increase in watering will result in an increase in flowers.

e i 104 flowers ($n = 2.5$), 359 flowers ($n = 10$)

ii $n = 2.5$ is reliable, as it is an interpolation.

$n = 10$ is unreliable as it is an extrapolation and over-watering could be a problem.



b The number of hits can be measured exactly, while the distance from the target might not be exact.

c $x \approx -7.89y + 93.7$

d ≈ -0.8 shots, but it is impossible to make a negative number of shots. This extrapolation is not valid.

EXERCISE 27A

1 a continuous **b** discrete **c** continuous **d** continuous
e discrete **f** discrete **g** continuous **h** continuous

2 a i X = the height of water in the rain gauge