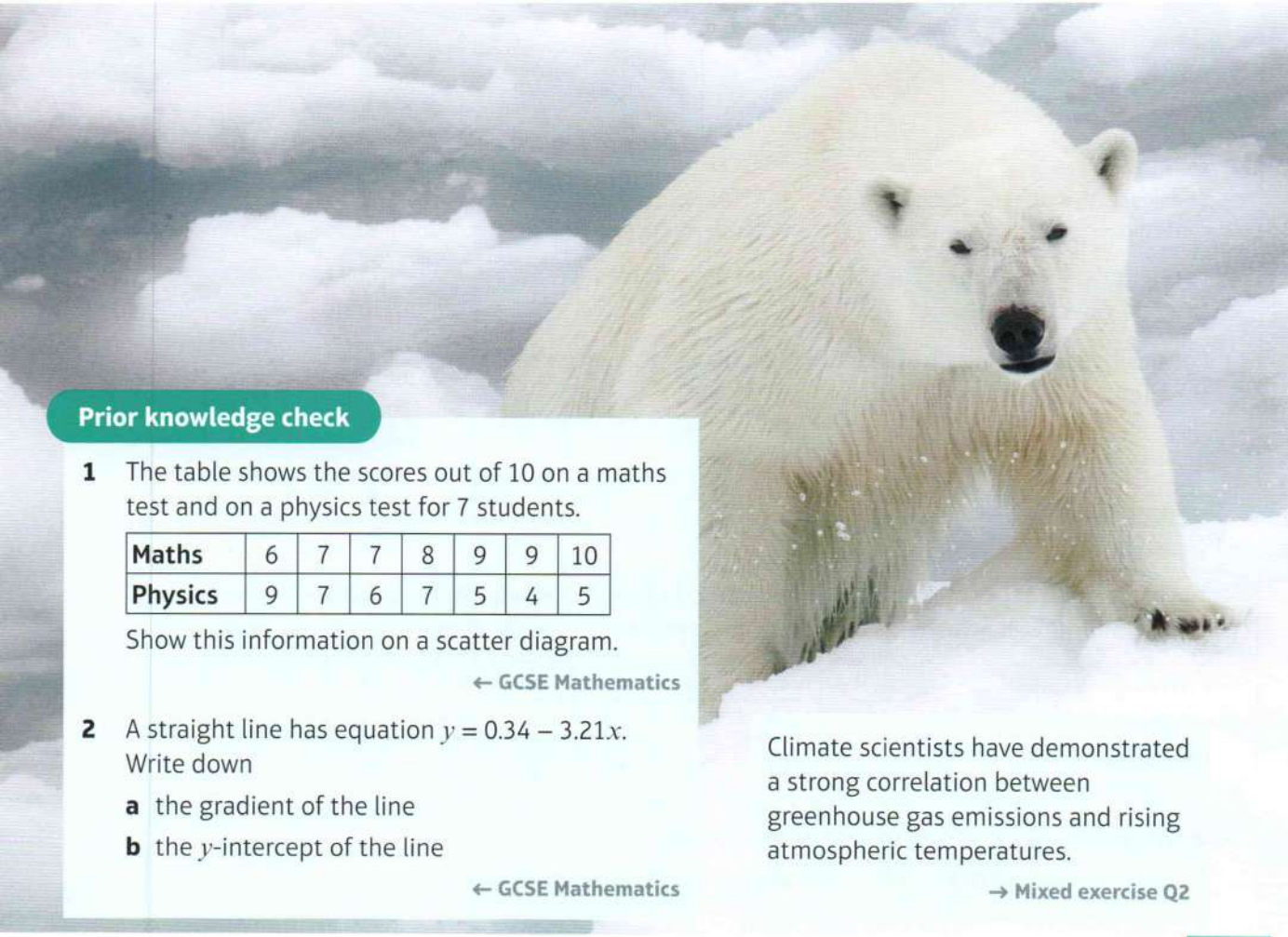
# Correlation

# 4

## Objectives

After completing this chapter you should be able to:

- Draw and interpret scatter diagrams for bivariate data → pages 60–61
- Interpret correlation and understand that it does not imply causation → pages 61–62
- Interpret the coefficients of a regression line equation for bivariate data → pages 63–64
- Understand when you can use a regression line to make predictions → pages 64–66



## Prior knowledge check

**1** The table shows the scores out of 10 on a maths test and on a physics test for 7 students.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Maths** | 6 | 7 | 7 | 8 | 9 | 9 | 10 |
| **Physics** | 9 | 7 | 6 | 7 | 5 | 4 | 5 |

Show this information on a scatter diagram.

← GCSE Mathematics

**2** A straight line has equation $y = 0.34 - 3.21x$.
Write down

**a** the gradient of the line

**b** the $y$-intercept of the line

← GCSE Mathematics

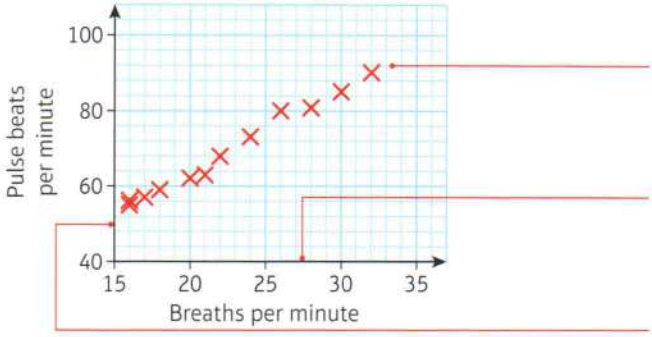
Climate scientists have demonstrated a strong correlation between greenhouse gas emissions and rising atmospheric temperatures.

→ Mixed exercise Q2

## 4.1 Correlation

- **Bivariate data is data which has pairs of values for two variables.**

You can represent bivariate data on a **scatter diagram**. This scatter diagram shows the results from an experiment on how breath rate affects pulse rate:



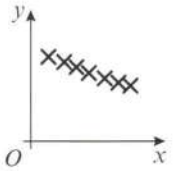

Each cross represents a data point. This subject had a breath rate of 32 breaths per minute and a pulse rate of 89 beats per minute.

The researcher could control this variable. It is called the **independent** or **explanatory variable**. It is usually plotted on the horizontal axis.

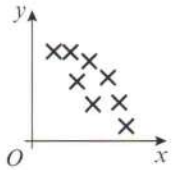
The researcher measured this variable. It is called the **dependent** or **response variable**. It is usually plotted on the vertical axis.

The two different variables in a set of bivariate data are often related.

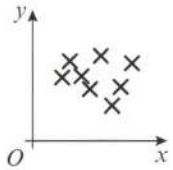
- **Correlation describes the nature of the linear relationship between two variables.**



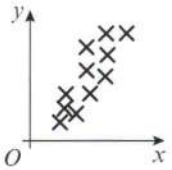

Strong negative correlation



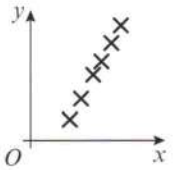

Weak negative correlation




No (or zero) linear correlation




Weak positive correlation




Strong positive correlation

For negatively correlated variables, when one variable increases the other decreases.

For positively correlated variables, when one variable increases, the other also increases.

**Watch out** You should only use correlation to describe data that shows a linear relationship. Variables with no linear correlation could still show a non-linear relationship.

### Example 1

In the study of a city, the population density, in people/hectare, and the distance from the city centre, in km, was investigated by picking a number of sample areas with the following results.

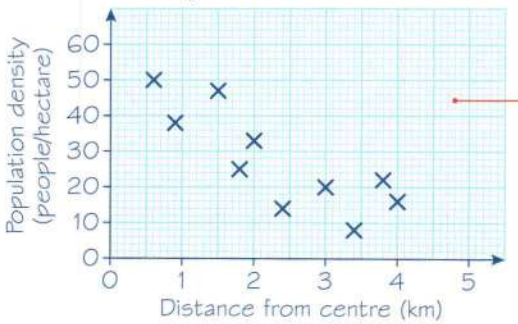
| Area | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Distance (km) | 0.6 | 3.8 | 2.4 | 3.0 | 2.0 | 1.5 | 1.8 | 3.4 | 4.0 | 0.9 |
| Population density (people/hectare) | 50 | 22 | 14 | 20 | 33 | 47 | 25 | 8 | 16 | 38 |

**a** Draw a scatter diagram to represent this data.

**b** Describe the correlation between distance and population density.

**c** Interpret your answer to part **b**.

a




b There is weak negative correlation.

c As distance from the centre increases, the population density decreases.

There are ten data points, so check that there are ten crosses on your scatter diagram.
Make sure that you include units with your axis labels.

Describe the strength of the correlation as well as whether it is negative or positive.

**Problem-solving**

Make sure you interpret results **in the context** of the question.

Two variables have a **causal relationship** if a change in one variable causes a change in the other. Just because two variables show correlation it does not necessarily mean that they have a causal relationship.

- **When two variables are correlated, you need to consider the context of the question and use your common sense to determine whether they have a causal relationship.**

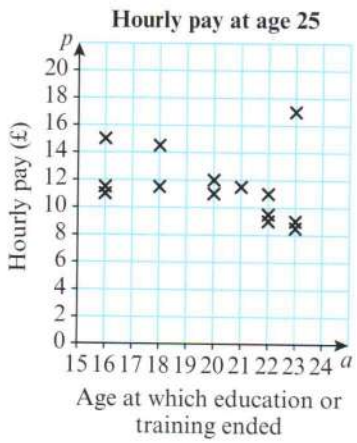
## Example 2

Hideko was interested to see if there was a relationship between what people earn and the age at which they left education or training. She asked 14 friends to fill in an anonymous questionnaire and recorded her results in a scatter diagram.

**a** Describe the type of correlation shown.

Hideko says that her data supports the conclusion that more education causes people to earn a lower hourly rate of pay.

**b** Give one reason why Hideko's conclusion might not be valid.




a Weak negative correlation.

b Respondents who left education later would have significantly less work experience than those who left education earlier. This could be the cause of the reduced income shown in her results.

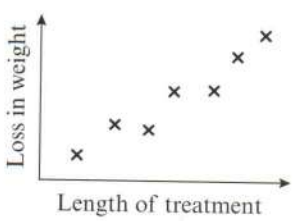
You could also say that Hideko's conclusion is not valid because she used a small, opportunistic sample. ← Section 1.1

## Exercise 4A

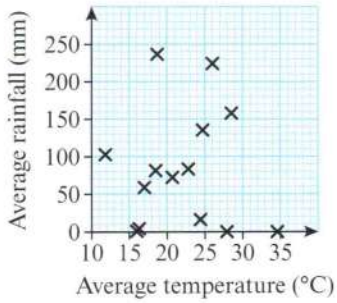
**1** Some research was done into the effectiveness of a weight-reducing drug. Seven people recorded their weight loss and this was compared with the length of time for which they had been treated. A scatter diagram was drawn to represent this data.

**a** Describe the type of correlation shown by the scatter diagram.

**b** Interpret the correlation in context.



**2** The average temperature and rainfall were collected for a number of cities around the world.
The scatter diagram shows this information.




**a** Describe the correlation between average temperature and average rainfall.

**b** Comment on the claim that hotter cities have less rainfall.

**3** Eight students were asked to estimate the mass of a bag of sweets in grams. First they were asked to estimate the mass without touching the bag and then they were told to pick the bag up and estimate the mass again. The results are shown in the table below.

| **Student** | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| **Estimate of mass not touching bag (g)** | 25 | 18 | 32 | 27 | 21 | 35 | 28 | 30 |
| **Estimate of mass holding bag (g)** | 16 | 11 | 20 | 17 | 15 | 26 | 22 | 20 |

**a** Draw a scatter diagram to represent this data.

**b** Describe and interpret the correlation between the two variables.

**4** Donal was interested to see whether there was a relationship between the value of a house and the speed of its internet connection, as measured by the time taken to download a 100 megabyte file. The table shows his results.

| **Time taken (s)** | 5.2 | 5.5 | 5.8 | 6.0 | 6.8 | 8.3 | 9.3 | 13 | 13.6 | 16.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| **House value (£1000s)** | 300 | 310 | 270 | 200 | 230 | 205 | 208 | 235 | 175 | 180 |

**a** Draw a scatter diagram to represent this data.

**b** Describe the type of correlation shown.

Donal says that his data shows that a slow internet connection reduces the value of a house.

**c** Give one reason why Donal's conclusion may not be valid.

(E) **5** The table shows the daily total rainfall, $r$ mm, and daily total hours of sunshine, $s$, in Leuchars for a random sample of 11 days in August 1987, from the large data set.

| $r$ | 0 | 6.8 | 0.9 | 4.8 | 0 | 21.7 | 1.7 | 4.9 | 0.1 | 2.2 | 0.1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $s$ | 8.4 | 4.9 | 10.2 | 4.5 | 3.3 | 3.9 | 5.4 | 1.8 | 9.7 | 1 | 4.6 |



The median and quartiles for the rainfall data are: $Q_1 = 0.1$ $Q_2 = 1.7$ $Q_3 = 4.85$

An outlier is defined as a value which lies either 1.5 × the interquartile range above the upper quartile or 1.5 × the interquartile range below the lower quartile.

**a** Show that $r = 21.7$ is an outlier. **(1 mark)**

**b** Give a reason why you might:
**i** include **ii** exclude this day's readings. **(2 marks)**

**c** Exclude this day's readings and draw a scatter diagram to represent the data for the remaining ten days. **(3 marks)**

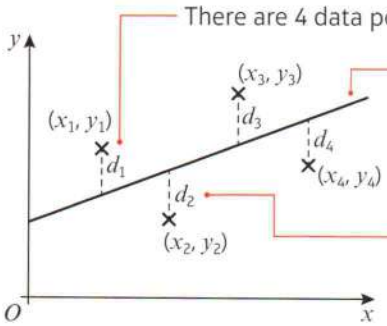
**d** Describe the correlation between rainfall and hours of sunshine. **(1 mark)**

**e** Do you think there is a causal relationship between the amount of rain and the hours of sunshine on a particular day? Explain your reasoning. **(1 mark)**

## 4.2 Linear regression

When a scatter diagram shows correlation, you can draw a **line of best fit**. This is a linear model that approximates the relationship between the variables. One type of line of best fit that is useful in statistics is a **least squares regression line**. This is the straight line that minimises the sum of the squares of the distances of each data point from the line.

**Notation** The least squares regression line is usually just called the regression line.




There are 4 data points on this scatter diagram.

The regression line of $y$ on $x$ is the straight line that minimises the value of $d_1^2 + d_2^2 + d_3^2 + d_4^2$. In general, if each data point is a distance $d_i$ from the line, the regression line minimises the value of $\Sigma d_i^2$.

The point $(x_2, y_2)$ is a vertical distance $d_2$ from the line.

- **The regression line of $y$ on $x$ is written in the form $y = a + bx$.**

**Watch out** The order of the variables is important. The regression line of $y$ on $x$ will be different from the regression line of $x$ on $y$.

You can use a calculator to find the values of the coefficients $a$ and $b$ for a given set of bivariate data. You will not be required to do this in your exam.

- **The coefficient $b$ tells you the change in $y$ for each unit change in $x$.**
  - **If the data is positively correlated, $b$ will be positive.**
  - **If the data is negatively correlated, $b$ will be negative.**

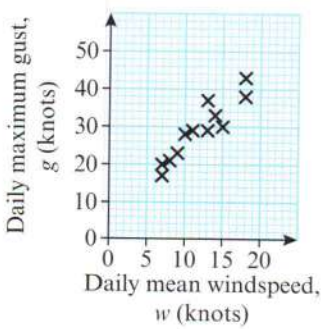
### Example 3

From the large data set, the daily mean windspeed, $w$ knots, and the daily maximum gust, $g$ knots, were recorded for the first 15 days in May in Camborne in 2015.

| $w$ | 14 | 13 | 13 | 9 | 18 | 18 | 7 | 15 | 10 | 14 | 11 | 9 | 8 | 10 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $g$ | 33 | 37 | 29 | 23 | 43 | 38 | 17 | 30 | 28 | 29 | 29 | 23 | 21 | 28 | 20 |



The data was plotted on a scatter diagram:




**a** Describe the correlation between daily mean windspeed and daily maximum gust.

The equation of the regression line of $g$ on $w$ for these 15 days is $g = 7.23 + 1.82w$.

**b** Give an interpretation of the value of the gradient of this regression line.

**c** Justify the use of a linear regression line in this instance.

a There is a strong positive correlation between daily mean windspeed and daily maximum gust.

b If the daily mean windspeed increases by 10 knots the daily maximum gust increases by approximately 18 knots.

c The correlation suggests that there is a linear relationship between $g$ and $w$ so a linear regression line is a suitable model.

**Online** Explore the regression line and analysis using technology.

Make sure your interpretation refers to both the context and your numerical value of the gradient. Try to phrase your answer as a complete, clear sentence.

**Problem-solving**

A regression line is a valid model when the data shows linear correlation. The stronger the correlation, the more accurately the regression line will model the data.

If you know a value of the **independent variable** from a bivariate data set, you can use the regression line to make a prediction or estimate of the corresponding value of the **dependent variable**.

- **You should only use the regression line to make predictions for values of the dependent variable that are within the range of the given data.**

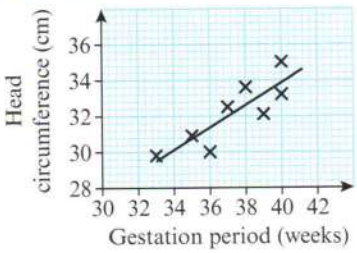
**Notation** This is called **interpolation**. Making a prediction based on a value outside the range of the given data is called **extrapolation**, and gives a much less reliable estimate.

## Example 4

The head circumference, $y$ cm, and gestation period, $x$ weeks, for a random sample of eight newborn babies at a clinic were recorded.

| Gestation period ($x$ weeks) | 36 | 40 | 33 | 37 | 40 | 39 | 35 | 38 |
|---|---|---|---|---|---|---|---|---|
| Head circumference ($y$ cm) | 30.0 | 35.0 | 29.8 | 32.5 | 33.2 | 32.1 | 30.9 | 33.6 |

The scatter graph shows the results.




The equation of the regression line of $y$ on $x$ is $y = 8.91 + 0.624x$.

The regression equation is used to estimate the head circumference of a baby born at 39 weeks and a baby born at 30 weeks.

**a** Comment on the reliability of these estimates.

A nurse wants to estimate the gestation period for a baby born with a head circumference of 31.6 cm.

**b** Explain why the regression equation given above is not suitable for this estimate.

a The prediction for 39 weeks is within the range of the data (interpolation) so is more likely to be accurate.

The prediction for 30 weeks is outside the range of the data (extrapolation) so is less likely to be accurate.

You could also comment on the sample. The sample was randomly chosen which would improve the accuracy of the predictions, but the sample size is small which would reduce the accuracy of the predictions.

b The independent (explanatory) variable in this model is the gestation period, $x$. You should not use this model to predict a value of $x$ for a given value of $y$.

**Watch out** You should only make predictions for the **dependent** variable. If you needed to predict a value of $x$ for a given value of $y$ you would need to use the regression line of $x$ on $y$.

## Exercise 4B

**1** An accountant monitors the number of items produced per month by a company together with the total production costs. The table shows this data.

| **Number of items, $n$ (1000s)** | 21 | 39 | 48 | 24 | 72 | 75 | 15 | 35 | 62 | 81 | 12 | 56 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Production costs, $p$ (£1000s)** | 40 | 58 | 67 | 45 | 89 | 96 | 37 | 53 | 83 | 102 | 35 | 75 |

**a** Draw a scatter diagram to represent this data.

The equation of the regression line of $p$ on $n$ is $p = 21.0 + 0.98n$.

**b** Draw the regression line on your scatter diagram.

**c** Interpret the meaning of the figures 21.0 and 0.98.

The company expects to produce 74 000 items in June, and 95 000 items in July.

**d** Comment on the suitability of this regression line equation to predict the production costs in each of these months.

**2** The relationship between the number of coats of paint applied to a boat and the resulting weather resistance was tested in a laboratory. The data collected is shown in the table.

| **Coats of paint ($x$)** | **Protection (years) ($y$)** |
|---|---|
| 1 | 4.4 |
| 2 | 5.9 |
| 3 | 7.1 |
| 4 | 8.8 |
| 5 | 10.2 |

**a** Draw a scatter diagram to represent this data.

The equation of the regression line is $y = 2.93 + 1.45x$.

Helen says that a gradient of 1.45 means that if 10 coats of paint are applied the protection will last 14.5 years.

**b** Comment on Helen's statement.

**3** The table shows the ages of some chickens and the number of eggs that they laid in a month.

| **Age of chicken, $a$ (months)** | 18 | 32 | 44 | 60 | 71 | 79 | 99 | 109 | 118 | 140 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Number of eggs laid in a month, $n$** | 16 | 18 | 13 | 7 | 12 | 7 | 11 | 13 | 6 | 9 |

**a** Draw a scatter diagram to show this information.

Robin calculates the regression line of $n$ on $a$ as $n = 16.1 + 0.063a$.

**b** Without further calculation, explain why Robin's regression equation is incorrect.

**4** Aisha collected data on the numbers of bedrooms, $x$, and the values, $y$ (£1000s), of the houses in her village. She calculates the regression equation of $y$ on $x$ to be $y = 190 + 50x$.

She states that the value of the constant in her regression equation means that a house with no bedrooms in her village would be worth £190 000. Explain why this is not a reasonable statement.

(E) **5** The table shows the daily maximum relative humidity, $h$ (%), and the daily mean visibility, $v$ decametres (Dm), in Heathrow for the first two weeks in September 2015, from the large data set.

| $h$ | 94 | 95 | 92 | 80 | 97 | 94 | 93 | 90 | 87 | 95 | 93 | 92 | 91 | 98 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $v$ | 2600 | 2900 | 3900 | 4300 | 2800 | 2400 | 2700 | 3500 | 3000 | 2200 | 2200 | 3300 | 2800 | 2200 |



The equation of the regression line of $v$ on $h$ is $v = 12\,700 - 106h$

**a** Give an interpretation of the value of the gradient of the regression line. **(1 mark)**

**b** Use your knowledge of the large data set to explain whether there is likely to be a causal relationship between humidity and visibility. **(2 marks)**

**c** Give reasons why it would not be reliable to use this regression equation to predict:

**i** the mean visibility on a day with 100% humidity **(2 marks)**

**ii** the humidity on a day with visibility of 3000 dm. **(2 marks)**

**d** State two ways in which better use could be made of the large data set to produce a model describing the relationship between humidity and visibility. **(2 marks)**

## Mixed exercise 4

**1** A survey of British towns recorded the number of serious road accidents in a week ($x$) in each town, together with the number of fast food restaurants ($y$). The data showed a strong positive correlation. Katie states that this shows that building more fast food restaurants in her town will cause more serious road accidents. Explain whether the data supports Katie's statement.

**2** The following table shows the mean $CO_2$ concentration in the atmosphere, $c$ (ppm), and the increase in average temperature compared to the 30-year period 1951–1980, $t$ (°C).

| Year | 2015 | 2013 | 2011 | 2009 | 2007 | 2005 | 2003 | 2001 | 1999 | 1997 | 1995 | 1994 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $c$ (ppm) | 401 | 397 | 392 | 387 | 384 | 381 | 376 | 371 | 368 | 363 | 361 | 357 |
| $t$ (°C) | 0.86 | 0.65 | 0.59 | 0.64 | 0.65 | 0.68 | 0.61 | 0.54 | 0.41 | 0.47 | 0.45 | 0.24 |

Source: Earth System Research Laboratory ($CO_2$ data); GISS Surface Temperature Analysis, NASA (temperature data)

**a** Draw a scatter diagram to represent this data.

**b** Describe the correlation between $c$ and $t$.

**c** Interpret your answer to part **b**.

(E) **3** The table below shows the packing times for a particular employee for a random sample of orders in a mail order company.

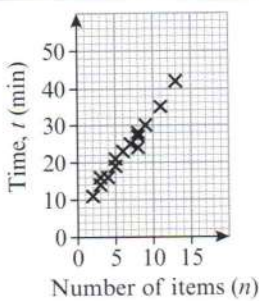
| Number of items ($n$) | 2 | 3 | 3 | 4 | 5 | 5 | 6 | 7 | 8 | 8 | 8 | 9 | 11 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Time ($t$ min) | 11 | 14 | 16 | 16 | 19 | 21 | 23 | 25 | 24 | 27 | 28 | 30 | 35 | 42 |

A scatter diagram was drawn to represent the data.

**a** Describe the correlation between number of items packed and time taken. **(1 mark)**

The equation of the regression line of $t$ on $n$ is $t = 6.3 + 2.64n$.

**b** Give an interpretation of the value 2.64. **(1 mark)**



E **4** Energy consumption is claimed to be a good predictor of Gross National Product.
An economist recorded the energy consumption ($x$) and the Gross National Product ($y$) for eight countries. The data is shown in the table.

| **Energy consumption ($x$)** | 3.4 | 7.7 | 12.0 | 75 | 58 | 67 | 113 | 131 |
|---|---|---|---|---|---|---|---|---|
| **Gross National Product ($y$)** | 55 | 240 | 390 | 1100 | 1390 | 1330 | 1400 | 1900 |

The equation of the regression line of $y$ on $x$ is $y = 225 + 12.9x$.

The economist uses this regression equation to estimate the energy consumption of a country with a Gross National Product of 3500.

Give two reasons why this may not be a valid estimate. **(2 marks)**

E **5** The table shows average monthly temperature, $t$ (°C), and the number of pairs of gloves, $g$, a shop sells each month.

| $t$ (°C) | 6 | 6 | 50 | 10 | 13 | 16 | 18 | 19 | 16 | 12 | 9 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $g$ | 81 | 58 | 50 | 42 | 19 | 21 | 4 | 2 | 20 | 33 | 58 | 65 |

The following statistics were calculated for the data on temperature:
mean = 15.2, standard deviation = 11.4

An outlier is an observation which lies ±2 standard deviations from the mean.

**a** Show that $t = 50$ is an outlier. **(1 mark)**

**b** Give a reason whether or not this outlier should be omitted from the data. **(1 mark)**

The equation of the regression line of $g$ on $t$ for the remaining data is $g = 99.6 - 5.2t$.

**c** Give an interpretation of the value −5.2 in this regression equation. **(1 mark)**

E **6** James placed different masses ($m$) on a spring and measured the resulting length of the spring ($s$) in centimetres. The smallest mass was 20 g and the largest mass was 100 g.

He found the equation of the regression line of $s$ on $m$ to be $s = 44 + 0.2m$.

**a** Interpret the values 44 and 0.2 in this context. **(2 marks)**

**b** Explain why it would not be sensible to use the regression equation to work out:
**i** the value of $s$ when $m = 150$ **ii** the value of $m$ when $s = 60$. **(2 marks)**

E **7** A student is investigating the relationship between the price ($y$ pence) of 100 g of chocolate and the percentage ($x$%) of cocoa solids in the chocolate.

The data obtained is shown in the table.

| **Chocolate brand** | **$x$ (% cocoa)** | **$y$ (pence)** |
|---|---|---|
| A | 10 | 35 |
| B | 20 | 55 |
| C | 30 | 40 |
| D | 35 | 100 |
| E | 40 | 60 |
| F | 50 | 90 |
| G | 60 | 110 |
| H | 70 | 130 |

**a** Draw a scatter diagram to represent this data. **(2 marks)**

The equation of the regression line of $y$ on $x$ is $y = 17.0 + 1.54x$.

**b** Draw the regression line on your diagram. **(2 marks)**

The student believes that one brand of chocolate is overpriced and uses the regression line to suggest a fair price for this brand.

**c** Suggest, with a reason, which brand is overpriced. **(1 mark)**

**d** Comment on the validity of the student's method for suggesting a fair price. **(1 mark)**

## Large data set

You will need access to the large data set and spreadsheet software to answer these questions.

**1** Investigate the relationship between daily mean windspeed, $w$, and daily maximum gust, $g$, in Leeming in 2015.

- **a** Draw a scatter diagram of $w$ against $g$ for the entire data set for Leeming in 2015.
- **b** Describe the correlation shown.
- **c** Comment on whether there is likely to be a causal relationship between mean windspeed and maximum gust.

The equation of the regression line of $g$ on $w$ is given by $g = 4.97 + 2.15w$.

- **d** Use the equation of the regression line to predict the maximum gust on a day when the mean windspeed is:
  **i** 0.5 knots **ii** 5 knots **iii** 12 knots **iv** 40 knots.
- **e** Comment on the accuracy of each prediction in part **d**.
- **f** Calculate the equation of the regression line of $w$ on $g$, and use it to predict the mean windspeed on a day when the maximum gust was 30 knots.

**2** Use a similar approach to investigate the daily total sunshine and daily mean total cloud cover in Heathrow in 1987.

- **a** Use a regression model to suggest values for the missing total sunshine data in the first half of May.
- **b** Do you think there is a causal relationship between these two variables? Give a reason for your answer.

**Hint** You can use the SLOPE and INTERCEPT functions in some spreadsheets to find the values of $a$ and $b$ in a regression equation.

## Summary of key points

**1** **Bivariate data** is data which has pairs of values for two variables.

**2** **Correlation** describes the nature of the linear relationship between two variables.

**3** When two variables are correlated, you need to consider the context of the question and use your common sense to determine whether they have a causal relationship.

**4** The **regression line** of $y$ on $x$ is written in the form $y = a + bx$.

**5** The coefficient $b$ tells you the change in $y$ for each unit change in $x$.

- If the data is positively correlated, $b$ will be positive.
- If the data is negatively correlated, $b$ will be negative.

**6** You should only use the regression line to make predictions for values of the dependent variable that are within the range of the given data.