2 Fitting a linear model to numerical data

2.1 Overview

A mathematical model uses mathematical concepts and language to demonstrate how something in the real world works. These models can then be used to investigate an aspect of the natural world or society without the researcher being present in the situation. A very simple mathematical model is the use of the formula, $V = l \times w \times h$ which describes the volume of a rectangular box. Rather than creating many boxes, and measuring them to determine their volume, this formula can be used to calculate the volume of a rectangular box of any dimensions. A more complicated model, for example the formula required to send a rocket into space, requires information about the energy needed to break through gravity, which depends on the size of the rocket, the starting and ending points of the journey, and the energy provided by the preferred fuel. A trial and error model to determine the variables would compromise safety and waste valuable resources. Mathematical models are first determined and are used to ensure the greatest chance of success.

A statistical model is a particular type of mathematical model. It is created by collecting data from a sample of the population that is under investigation. It then uses this data to create an idealised way of predicting how any future data will behave. The manufacturer of a new running shoe wants to determine the best tread thickness for maximum support. Rather than spending years measuring tread thicknesses and the extent of running injuries across the world, the researchers would take a sample and use the data from the sample to make predictions and, ultimately, designs.

LEARNING SEQUENCE

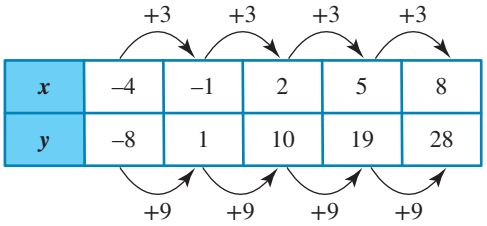
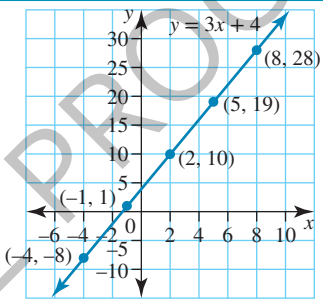
Fully worked solutions for this chapter are available in the Resources section of your eBookPLUS at www.jacplus.com.au.

2.2 Review of the general equation of a straight line

2.2.1 Linear relationships

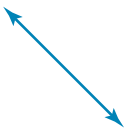
When a line is formed by a set of points on a **Cartesian plane**, there is a consistent relationship between the x-coordinates and the y-coordinates of the points on the line. If the line formed is a straight line, there is a consistent linear relationship between the x-coordinates and the y-coordinates of the points on the line. Linear relationships can be identified by using a table of values, plotting points or looking at an equation.

The general form of a linear equation is $y = mx + c$.

	$y = 3x + 4$	
Each variable has a constant difference.	The equation can be written in the form $y = mx + c$.	The points form a straight line.

2.2.2 Linear relationships

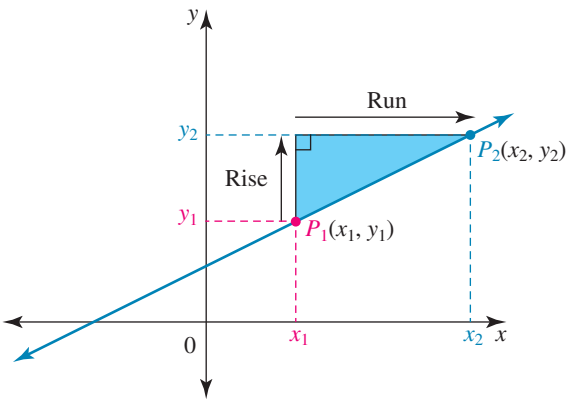
The gradient, m, of the line is a measure of the slope or steepness of the line. The gradient is also a measure of the rate of change in y with respect to x. If the gradient of a line is 3 ($m = 3$), this means that for each increase of 1 in the x-value, there is a corresponding increase of 3 in the y-value.

$m > 0$	$m < 0$	$m = 0$	m is undefined.
$y = mx + c$	$y = mx + c$	Equations: $y = k$ where k is a constant.	Equations: $x = k$ where k is a constant.
			

The gradient of a line can be determined from any two points on the line.

$$m = \frac{\text{rise}}{\text{run}} = \frac{y_2 - y_1}{x_2 - x_1}$$

Subscripts are used to show which points the coordinates belong to. For example, the x- and y-coordinates of point 1 are (x_1, y_1); the x- and y-coordinates of point 2 are (x_2, y_2).



WORKED EXAMPLE 1

Calculate the gradient of the line that passes through the points (3, −2) and (−4, 6).

THINK

1. Let one point be point 1 and the other point be point 2. Record their x- and y-coordinates.
2. Substitute the coordinates into the gradient formula $m = \frac{y_2 - y_1}{x_2 - x_1}$ and simplify.

WRITE

a. Let: $(x_1, y_1) = (3, -2)$
$(x_2, y_2) = (-4, 6)$

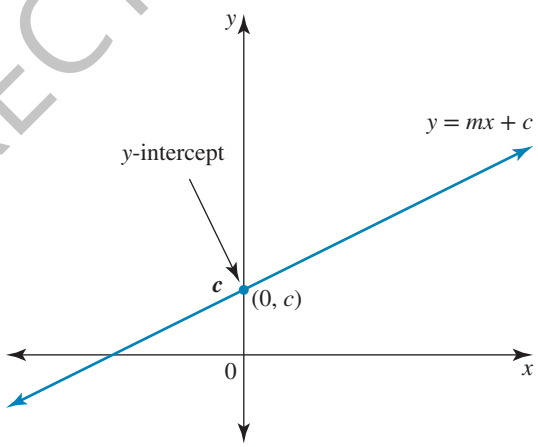
b. $m = \frac{y_2 - y_1}{x_2 - x_1}$
$= \frac{6 - -2}{-4 - 3}$
$= \frac{6 + 2}{-4 - 3}$
$= \frac{8}{-7}$
$= -1\frac{1}{7}$

2.2.3 The *y*-intercept

The y-intercept, c, of the line is the point where the line intersects the y-axis (the line $x = 0$).

The value of the y-intercept can be determined by any of these methods:

- looking at the graph of the line and determining the point at which the line crosses the y-axis
- writing the equation of the line in the form of $y = mx + c$ and identifying the constant value c
- substituting $x = 0$ into the equation of the line, since the y-axis is also the line $x = 0$ and hence the x-coordinate of any y-intercept is 0.



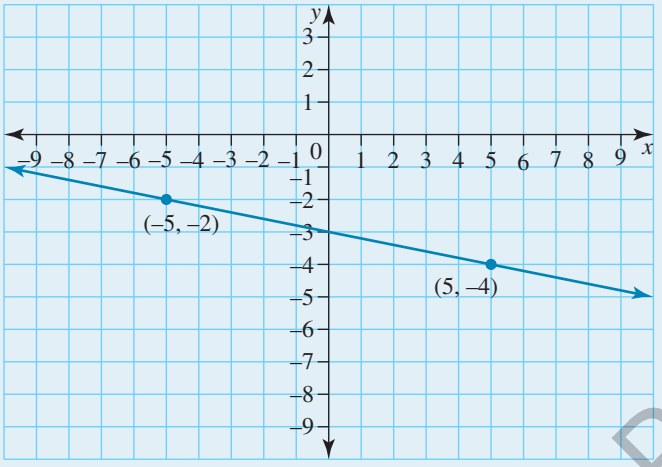
2.2.4 Determining the equation of the line using two points

The coordinates of two points on a line are all that is needed to determine the equation of a line. To determine the equation of the line, the value of both the gradient and the y-intercept are required.

- Calculate the gradient using the two points.
- Use the general form of the equation, the coordinates of one point and the gradient to calculate the value of the y-intercept.
- Write the equation of the line in the form $y = mx + c$

WORKED EXAMPLE 2

Determine the equation of the line that passes through the points $(-5, -2)$ and $(5, -4)$.



THINK

1. To determine the equation of the line, both the gradient and the y-intercept are required. Calculate the gradient first. Let one point be point 1 and the other point be point 2. Record their x- and y-coordinates.
2. Substitute the coordinates into the gradient formula $m = \frac{y_2 - y_1}{x_2 - x_1}$ and simplify.
3. Write the general equation of a straight line. Substitute the value of the gradient m, into the formula. To find the value of the y-intercept, c, substitute the coordinates of one of the points into the equation as the values for x and y.
4. Write the equation of the line.

WRITE

a. Let: $(x_1, y_1) = (-5, -2)$

$(x_2, y_2) = (5, -4)$

b. $m = \frac{y_2 - y_1}{x_2 - x_1}$

$= \frac{-4 - -2}{5 - -5}$

$= \frac{-4 + 2}{5 + 5}$

$= \frac{-2}{10}$

$= -\frac{1}{5}$

c. $y = mx + c$

Let $m = \frac{-1}{5}$

$y = \frac{1}{5}x + c$

Let $(x, y) = (5, -4)$

$-4 = \frac{-1}{5}(5) + c$

$-4 = -1 + c$

$c = -3$

d. $y = \frac{-1}{5}x - 3$ or $y = \frac{-x - 15}{5}$

WORKED EXAMPLE 3

For each of the linear equations below:

i. state the gradient and *y*-intercept

ii. sketch the graph of the equation.

a. $y = 4x - 11$ **b.** $y = -4x$

THINK

a. i. Compare the equation given with the general form of a linear equation: $y = mx + c$. The coefficient of x is m (the gradient), and the constant c is the y-intercept.

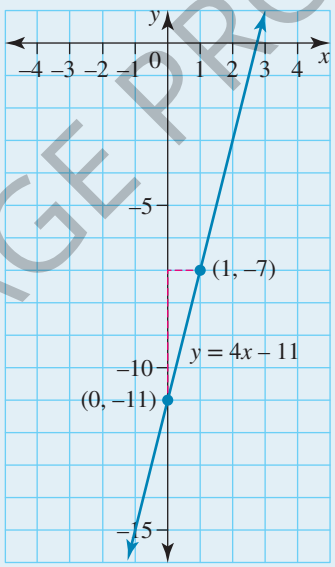
ii. Construct a set of axes and mark the position of the y-intercept.
The y-intercept is -11, as shown in blue.
The gradient is 4, so $\frac{rise}{run} = \frac{4}{1}$.
From the y-intercept, rise 4 and run 1, then mark in a second point $(1, -7)$, as shown in pink.
The two points can now be connected with a straight line.
Write the equation next to the line.

b. i. Compare the equation given with the general form of a linear equation: $y = mx + c$. Identify m as the gradient and c as the y-intercept.

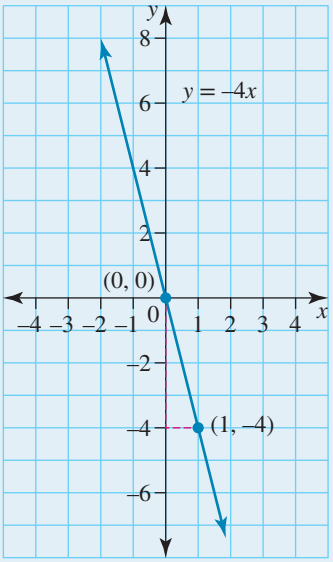
ii. Construct a set of axes.
Mark in the position of the y-intercept at 0, as shown in blue. The gradient is -4, so $\frac{rise}{run} = \frac{-4}{1}$.
From the y-intercept, rise -4 and run 1, then mark in a second point $(1, -4)$, as shown in pink.
The two points can now be connected with a straight line to form the graph.
Write the equation next to the line.

WRITE

a. For $y = 4x - 11$
Gradient $(m) = 4$
y-intercept $(c) = -11$.



b. For $y = -4x$ or $(y = -4x + 0)$
Gradient $(m) = -4$ and
y-intercept $(c) = 0$.



on Resources

Interactivities: Linear graphs (int-6484)
Equations of straight lines (int-6485)

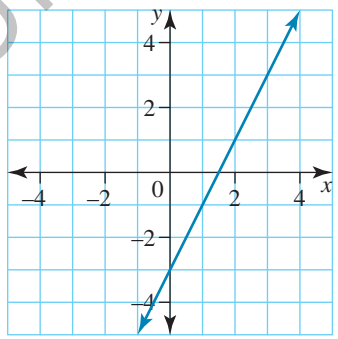
studyon

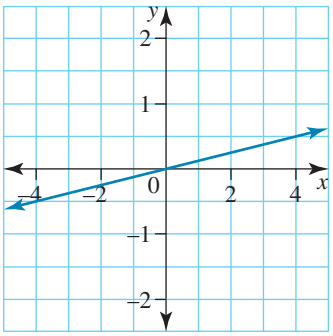
Units 3 & 4 > Area 1 > Sequence 2 > Concept 1

Review of the general equation of a straight line Summary screen and practice questions

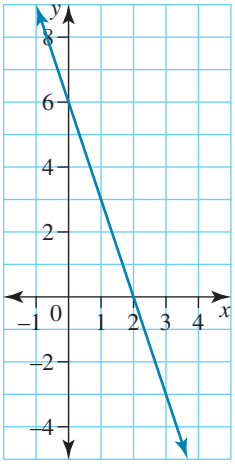
Exercise 2.2 Review of the general equation of a straight line

1. In your own words, explain:
   a. the general form of a straight line
   b. what the letters represent in the general form of a straight line, $y = mx + c$
2. State the value of m and c in each of the following linear relationships.
   a. $y = -5x + 4$ b. $y = 3x + 11$ c. $y = -6x$
   d. $y = \frac{5}{2}x + 5$ e. $y = \frac{2}{3}x - 6$ f. $y = 2x - 1$
3. Complete the sentences below.
   a. The equation $y = 3x - 1$ has a gradient of ____ units. This means that for every increase of 1 unit in the horizontal direction there is an increase of ___ units in the vertical direction.
   b. The equation $y = -x + 1$ has a gradient of ____ units. This means that for every increase of ___ unit in the horizontal direction there is an increase of ___ units in the vertical direction.
   c. The equation $y = \frac{1}{3}x - 1$ has a gradient of ____ units. This means that for every increase of ___ unit in the horizontal direction there is an increase of ___ units in the vertical direction.
4. **WE1** Calculate the gradient between each pair of points for each of the following.
   a. $(3, 4), (1, 0)$ b. $(1, 3), (4, 6)$ c. $(-2, 4), (-4, 2)$
   d. $(6, 3), (2, 5)$ e. $(3, -1), (-1, 0)$ f. $(6, 8), (-3, -4)$
   g. $(-2, -3), (-5, 1)$ h. $(-3, -6), (-10, -4)$ i. $(2, 5), (-1, 5)$
5. Write the linear relationship for the line with the following properties:
   a. gradient $= 2$ $y-$ intercept $= -2$ b. gradient $= 2$ $y-$ intercept $= -1$
   c. gradient $= \frac{-3}{8}$ $y-$ intercept $= 2$ d. gradient $= 0$ $y-$ intercept $= 4$
   e. gradient $= 1$ $y-$ intercept $= 0$ f. gradient $= -0.25$ $y-$ intercept $= -4$
6. **WE2** Determine the equation of the straight lines that join each of the following pairs of points.
   a. $(3, 4), (-2, -3)$ b. $(4, 6), (-5, 1)$ c. $(-2, 4), (-4, 2)$
   d. $(6, 3), (1, 3)$ e. $(-1, 0), (-10, -4)$ f. $(-3, -4), (2, 5)$
7. For each of the linear graphs shown:
   i. state the y-intercept
   ii. calculate the gradient
   iii. write a linear relationship to describe the graph.

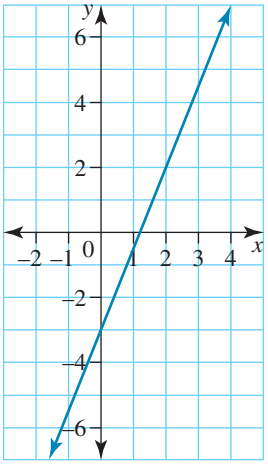
   a.


   b.


c.



d.



8. **WE3** For each of the linear relationships below:

i. state the gradient and the y-intercept

ii. sketch the graph of the equation.

a. $y = 2x + 5$ **b.** $y = -4x - 3$ **c.** $y = -3x + 5$

d. $y = \frac{1}{2}x - 3$ **e.** $y = \frac{-2}{3}x + 3$ **f.** $y = \frac{4}{5}x$

9. Construct graphs for the linear relationships in question **2**.

10. **a.** If t represents the time in hours and C represents cost ($), construct a table of values for 0–3 hours for the cost of playing tenpin bowling at a new alley.

Save $$$ with Supa-Bowl!!!
NEW Tenpin Bowling Alley
Shoe rental just* $2 *(fixed fee)
Rent a lane for ONLY $6/hour!

b. Use your table of values to plot a graph of time versus cost. [*Hint:* Ensure your time axis (horizontal axis) extends to 6 hours and your cost axis (vertical axis) extends to $40.

c. Calculate the gradient.

d. Write a linear equation to describe the relationship between cost and time.

e. Use your linear equation from part **e** to calculate the cost of a 5 hour tournament.

f. Use your graph to check your answer to part **f**.

11. For the equation $y = 4x + 1$, substitute the following values for x to calculate the corresponding values for y.

a. $x = 0$ **b.** $x = -2$ **c.** $x = 6$ **d.** $x = 10$

e. $x = -7$ **f.** $x = -5$ **g.** $x = 4.2$ **h.** $x = \frac{1}{2}$

12. For the equation $y = 4x + 1$, substitute the following values for y to calculate the corresponding values for x.

a. $y = 0$ **b.** $y = 9$ **c.** $y = -11$ **d.** $y = 19$

e. $y = 3$ **f.** $y = 0.6$ **g.** $y = -20$ **h.** $y = 2.5$

13. The equation $C = 3.5g + 55$ represents the cost of hiring online computer games from a gaming club, where C is the cost of hiring g games.

a. How much will it cost for the year to hire:

i. 1 game a week for the year (52 weeks in a year)?

ii. 10 games a month for the year (12 months in a year)?

b. The cost for hire is generated by an annual membership fee plus a small hire fee for each game. How much is:

i. the annual membership fee? ($g = 0$)

ii. the hire fee per game?

c. If a teenager budgets on spending $ 300 on games for the year, how many games will this hire?
d. When considering the annual membership fee, what is the real cost of hiring a game, given the number of games from part c that the teenager intends to hire?
e. The cost of hiring a new game at a regular online subscription service is $5. If the teenager budgets on spending $300, will being a member of the club save money in the long run?

14. A family dental fund charges an annual membership fee of $ 200 and then $ 20 per dental appointment.
a. MC Which of the following equations best represents the cost of the fund in terms of dental appointments, if C is the cost and d is the number of dental appointments in the year?
A. $C = 200$
B. $C = 200 + 20d$
C. $C = 20d$
D. $C = 200d + 20$
b. What would be the cost over the year for the following families if on average each member of the family visited the dentist 2 times in a year?
i. A family of 2
ii. A family of 4
iii. A family of 10
c. If the average cost of a dental appointment is $ 60 without a dental plan, which of the families in part b will have saved money by being in the fund? How much will they have saved?
d. If a family paid $ 340 to the dental fund for the year, how many dental appointments did they have?

2.3 Fitting a least-squares line to data

2.3.1 Lines of best fit

In Chapter 1 scatterplots were constructed from raw data and studied to determine if a linear association existed between the two variables.

If the points on a scatterplot appear to be distributed in a linear pattern, a straight line can be drawn through the data.

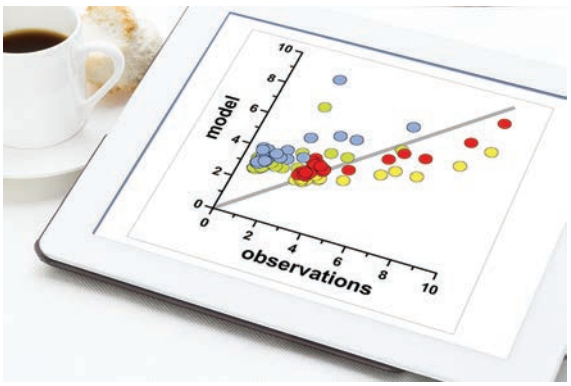
A **line of best fit** is the straight line that is positioned as close as possible to all the data points, that is, the average distance between the data points and the line is minimised. It is used to generalise the relationship between two variables.

There are a number of ways to draw a line of best fit.

This General Mathematics course focuses on the least-squares regression line as a mathematical means of constructing a line of best fit.

2.3.2 The least-squares regression line

Least-squares regression involves an exact mathematical approach to fitting a line of best fit to bivariate data that shows a strong linear relationship.



This line minimises the vertical distances between the data points and the line of best fit. It is called the least-squares regression line because if we took the squares of these vertical distances, this line would represent the smallest possible sum of all these squares.

The equation for the least-squares regression line takes the form: $y = a + bx$, where y is the response (dependent) variable, x is the explanatory (independent) variable, b is the gradient or slope of the line and a is the y-intercept.

Technology can be used to calculate the equation of the least-squares regression line.

To determine the equation of the least-squares regression line, the following summary data is required:
$\bar{x}$ — the mean of the explanatory variable (x-variable)
$\bar{y}$ — the mean of the response variable (y-variable)
s_x — the standard deviation of the explanatory variable
s_y — the standard deviation of the response variable
r — Pearson's correlation coefficient.
The general form of the least-squares regression line is:

$$y = a + bx$$

where:
the slope of the regression line is $b = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = r\dfrac{s_y}{s_x}$
the y-intercept of the regression line is $a = \bar{y} - b\bar{x}$.

WORKED EXAMPLE 4

A study to find a relationship between the height of men and the height of their female partner revealed the following details.

- **Mean height of the men: 180 cm**
- **Mean height of the female partners: 169 cm**
- **Standard deviation of the heights of the men: 5.3 cm**
- **Standard deviation of the heights of the female partners: 4.8 cm**
- **Correlation coefficient: $r = 0.85$**

The form of the least-squares regression line is:
Height of female partner $= b \times$ height of man $+ a$

a. Which variable is the response variable?
b. Calculate the value of b for the regression line (to 2 decimal places).
c. Calculate the value of a for the regression line (to 2 decimal places).
d. Use the equation of the regression line to predict the height of a female whose male partner is 195 cm tall (to nearest cm).

THINK

a. Recall that the response variable is the subject of the equation in $y = a + bx$ form; that is, y.

b. 1. The value of b is the gradient of the regression line. Write the formula and state the required values.
2. Substitute the values into the formula and evaluate b.

c. 1. The value of a is the y-intercept of the regression line. Write the formula and state the required values.
2. Substitute the values into the formula and evaluate a.

WRITE

a. The response variable is the height of the female.

b. $b = r\dfrac{s_y}{s_x}$ r
$= 0.85, s_y = 4.8$ and $s_x = 5.3$
$b = 0.85 \times \dfrac{4.8}{5.3}$
$= 0.7698$

c. $a = \bar{y} - b\bar{x}$
$\bar{y} = 169$, $\bar{x} = 180$ and $b = 0.7698$ (from part **b**)
$a = 169 - 0.7698 \times 180$
$= 30.436$
$= 30.44$

d. 1. State the equation of the regression line, using the values calculated from parts **b** and **c**. In this equation, y represents the height of the female and x represents the height of the male.

d. $y = 0.77x + 30.44$

2. The height of the male is 195 cm, so substitute $x = 195$ into the equation and evaluate.

$y = 0.77 \times 195 + 30.44$
$= 180.59$

3. Write a statement, rounding your answer to the nearest centimetre.

Using the equation of the regression line found, the female's height is predicted to be 181 cm.

2.3.3 Interpreting the intercept and slope

Often data is collected in order to make informed decisions or predictions about a situation. The regression line equation from a scatterplot can be used for this purpose.

Remember that the equation for the regression line is in the form $y = a + bx$, where b is the gradient or slope, a is the y-intercept, and x and y refer to the two variables. Two important pieces of information can be attained from this equation.

1. When the explanatory variable is equal to 0, the value of the response variable is indicated by the y-intercept, a.
2. For each increment of 1 unit of change in the explanatory variable, the change in the response variable is indicated by the value of the slope, b.

WORKED EXAMPLE 5

The following table shows data from Bilbo's Real Estate for house sales in The Shire in November 2020.

House	Number of bedrooms	Number of bathrooms	Size of garage (cars)	Size of land (m^2)	Price ($)
1	2	1	1	117	730 000
2	4	2	1	630	1 875 000
3	3	1	2	688	1 300 000
4	2	1	1	228	790 000
5	3	1	2	858	1 610 000
6	2	1	1	637	670 000
7	3	1	1	588	1 400 000
8	6	4	1	700	2 060 000
9	2	1	1	93	520 000
10	2	1	1	73	639 000
11	3	1	1	242	720 000
12	1	1	1	112	460 000
13	2	1	1	167	737 000

a. Using Excel, or other suitable technology, draw a scatterplot of *S* (size of land) against *P* (price of house)
b. Determine the least-squares regression line.
c. What does the least-squares regression line tell you about property prices in The Shire?

THINK

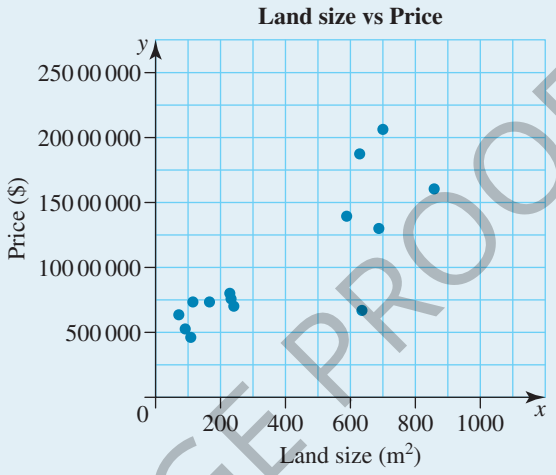
a. Size of land (S) is the explanatory variable. Price of house (P) is the response variable. Plot the table of values using Excel.

b. Determine the regression line using Excel.

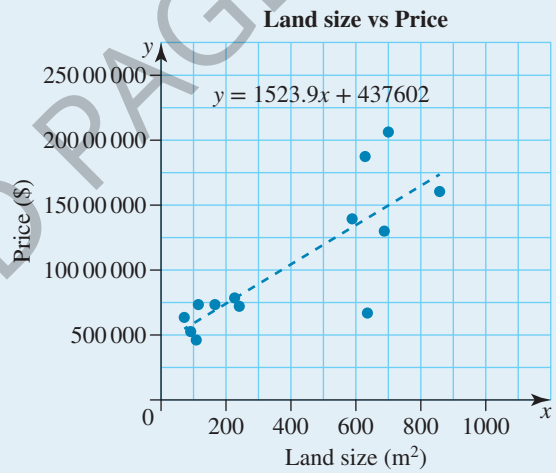
c. Interpret the least-squares regression line by referring to the slope ($1500) and the *y*-intercept ($ 437 602).

WRITE

a.



b.



The regression line calculated is:
$y = 1523.9x + 437\,602$.
In terms of price and size of land, the equation is:
$P = 1523.9 \times S + 437\,602$.

c. Property prices begin at $ 437 602 and increase by over $1500 per square metres after that.

WORKED EXAMPLE 6

The least-squares regression equation for a line is $y = 62 - 8x$.
a. Identify the *y*-intercept.
b. For each unit of change in the explanatory variable, by how much does the response variable change?
c. What does your answer to part b tell you about the direction of the line?

THINK	WRITE
a. Consider the equation in the form $y = a + bx$. Identify the value that represents a.	**a.** y – intercept $(a) = 62$
b. The change in the response variable due to the explanatory variable is reflected in the slope. Identify the b value in the equation.	**b.** slope $(b) = -8$
c. A positive b value indicates a line pointing in a positive direction, while a negative b value indicates a line pointing in a negative direction.	**c.** As the b value is negative, the direction of the line is negative.

2.3.4 Interpolation and extrapolation

The regression line can be used to explore data points both inside and outside of the given scatterplot range. When investigating data inside the variable range, the data is being **interpolated**. Data points that lie above or below the scatterplot range can also be used to make predictions. Predictions outside the range of data is **extrapolation**.

The regression equation can be used to make predictions from the data by substituting in a value for either the explanatory variable (x) or the response variable (y) to calculate the value of the other variable.

WORKED EXAMPLE 7

Flowers with a diameter of 5 – 17 cm were measured and the number of petals for each flower was documented. A regression equation of $N = 0.41 + 1.88d$, where N is the number of petals and d is the diameter of the flower (in cm) was established.

a. Identify the explanatory variable.

b. Determine the number of petals that would be expected on a flower with a diameter of 15 cm. Round to the nearest whole number.

c. Is the value found in part b an example of interpolated or extrapolated data?

d. A flower with 35 petals is found. Use the equation to predict the diameter of the flower, correct to 1 decimal place.

e. Is part d an example of interpolated or extrapolated data?

THINK	WRITE
a. Consider the format of the equation. The variable on the right-hand side of the equation will be the explanatory variable.	**a.** Explanatory variable = flower diameter (d)
b. 1. Using the equation, substitute 15 for d and evaluate.	**b.** $N = 0.41 + 1.88d$ $= 0.41 + 1.88 \times 15$ $= 28.61$
2. Round N to the nearest whole value. *Note*: round down as you can't have .61 of a petal.	28 petals
c. Consider the data range given in the opening statement.	**c.** 15 cm is inside the data range, so this is interpolation, not extrapolation.

d. 1. Using the equation, substitute 35 in place of N.
2. Transpose the equation to solve for d.
3. Round to 1 decimal place.

d. $35 = 0.41 + 1.88d$

$$d = \frac{35 - 0.41}{1.88}$$

$= 18.40$

$= 18.4$ (correct to 1 decimal place)

e. Consider the data range given in the opening statement.

e. 18.4 cm is outside the data range, so this is an example of extrapolated data.

2.3.5 Limitations of regression line predictions

When reviewing predictions drawn from a scatterplot, it is necessary to question the reliability of the results. As with any conclusion or prediction, the results rely heavily on the initial data. If the data was collected from a small sample, then the limited information could contain biases or a lack of diversity that would not be present in a larger sample. The more data that can be provided at the start, the more accurate a result will be produced.

The strength of the association between the variables also provides an indication of the reliability of the data. Data that produces no association or a low association would suggest that any conclusions drawn from the data will be unreliable.

When extrapolating data it is assumed that additional data will follow the same pattern as the data already in use. This assumption means extrapolated data is not as reliable as interpolated data.

on Resources

Interactivity: Fitting a straight line using least-squares regression (int-6254)

studyon

Units 3 & 4 > Area 1 > Sequence 1 > Concept 2

Fitting a least-squares line to data Summary screen and practice questions

Exercise 2.3 Fitting a least-squares line to data

1. WE4 The following summary details were calculated from a study to find a relationship between Mathematics exam marks and English exam marks from the results of 120 Year 10 students.
Mean Mathematics exam mark = 64%
Mean English exam mark = 74%
Standard deviation of Mathematics exam mark = 14.5%
Standard deviation of English exam mark = 9.8%
Correlation coefficient, $r = 0.64$.
The form of the least-squares regression line is:
Mathematics exam mark = $m \times$ English exam mark + c

a. Which variable is the response variable (y-variable)?

b. Calculate the value of m for the least-squares regression line.

c. Calculate the value of c for the least-squares regression line.
d. Use the regression line to predict the expected Mathematics exam mark if a student scores 85% in an English exam (to the nearest percentage).

2. Find the least-squares regression equation, given the following summary data.
a. $\bar{x} = 5.6, s_x = 1.2, \bar{y} = 110.4, s_y = 5.7, r = 0.7$
b. $\bar{x} = 110.4, s_x = 5.7, \bar{y} = 5.6, s_y = 1.2, r = -0.7$
c. $\bar{x} = 25, s_x = 4.2, \bar{y} = 10\,200, s_y = 250, r = 0.88$
d. $\bar{x} = 10, s_x = 1, \bar{y} = 20, s_y = 2, r = -0.5$

3. **WE5** Recall from Chapter 1 that a researcher investigating the proposition that 'tall mothers have tall sons' measures the height of 12 mothers and the height of their adult sons. The results are shown below.

Height of mother (cm)	Height of son (cm)
185	188
155	157
171	172
169	173
170	174
175	180
158	159
156	150
168	172
169	175
179	180
173	190

a. Which variable is the response variable?
b. Using Excel or other suitable technology draw a scatterplot.
c. Fit a least-squares regression line to the data and determine the equation of the line of best fit, expressing the equation in terms of height of mother (M) and height of son (S). Give values correct to 4 significant figures.

4. **WE6** The least-squares regression equation for a line is $y = -1.837 + 1.701x$.
a. Identify the y-intercept.
b. For each unit of change in the explanatory variable, by how much does the response variable change?
c. What does your answer to part **b** tell you about the direction of the line?

5. The least-squares regression equation for a line is $y = 105.90 - 1.476x$.
a. Identify the y-intercept.
b. For each unit of change in the explanatory variable, by how much does the response variable change?
c. What does your answer to part **b** tell you about the direction of the line?

6. **WE7** A brand of medication for babies bases the dosage on the age (in months) of the child. The regression equation for this situation is $M = 0.157 + 0.312A$, where M is the amount of medication in mL and A is the age in months.

a. Identify the explanatory variable.
b. Calculate the amount of medication required for a child aged 6 months.
c. Determine the age of a child who requires 2.5 mL of the medication. Give your answer correct to 1 decimal place.

7. A survey of the nightly room rate for Sydney hotels and their proximity to the Sydney Harbour Bridge produced the regression equation $C = 281.92 - 50.471d$, where C is the cost of a room per night in dollars and d is the distance to the bridge in kilometres.
   a. Identify the response variable.
   b. Based on this equation, calculate the cost of a hotel room 2.5 km from the bridge. Give your answer correct to the nearest cent.
   c. Determine the distance of a hotel room from the bridge if the cost of the room was $\$115$. Give your answer correct to 2 decimal places.
8. An equation for a regression line is $y = 3.2 - 1.56x$. What conclusions about the direction of the regression line can be determined from the equation?
9. a. Use technology to plot the regression line $y = -1.6 + 2.5x$.
   b. Would a data point of $(3, 4)$ be found above or below the regression line?
10. Answer the following questions for the equation $y = 60 - 5x$.
   a. Identify the y-intercept.
   b. For each unit of change in the explanatory variable, by how much does the response variable change?
   c. Is the direction of the data positive or negative?
   d. Calculate the value of y when $x = 40$.
11. Lucy was given the equation $y = -12.9 + 7.32x$ and asked to find the value of x when $y = 15.68$. Her working steps are below:

$$\begin{aligned} y &= -12.9 + 7.32x \\ 15.68 &= -12.9 + 7.32x \\ x &= 12.9 + \frac{15.68}{7.32} \\ &= 15.04 \end{aligned}$$

   Her teacher indicates her answer is wrong.
   a. Calculate the correct value of x. Give your answer correct to 2 decimal places.
   b. Identify and explain Lucy's error.
12. Answer the following questions for the equation $y = -12 + 25x$.
   a. Identify the y-intercept.
   b. For each unit of change in the explanatory variable, by how much does the response variable change?
   c. Is the direction of the data positive or negative?
   d. Calculate the value of y when $x = 3.5$.
13. Answer the following questions for the equation $I = 0.43 + 1.1s$, where I is the number of insects caught and s is the area of a spider's web in cm^2.
   a. Identify the response variable.
   b. For each unit of change in the explanatory variable, by how much does the response variable change?
   c. Is the direction of the data positive or negative?
   d. Determine how many insects are likely to be caught if the area of the spider's web is $60\,\text{cm}^2$. Give your answer correct to the nearest whole number.
14. A data set produced a positive direction and for each incremental increase in the explanatory variable, the response variable increased by 2.5. If $y = 4$ when $x = 0$, determine the equation for the regression line.
15. Use the data given below and appropriate technology to answer the following questions.

x	10	11	12	13	14	15	16	17	18
y	22	18	20	15	17	11	11	7	9

a. Draw a scatterplot and determine the equation of the least-squares regression line. Give values correct to 4 significant figures.
b. Extrapolate the data to predict the value of y when $x = 23$.
c. What assumptions are made when extrapolating data?

16. While camping a mathematician estimated that: number of mosquitoes around the fire $= 10.2 + 0.5 \times$ temperature of the fire (°C).

a. Determine the number of mosquitoes that would be expected if the temperature of the fire was 240°C. Give your answer correct to the nearest whole number.
b. What would be the temperature of the fire if there were only 12 mosquitoes in the area?
c. Identify some factors that could affect the reliability of this equation.

17. Data on 15 people's average monthly income and the amount of money they spend at restaurants was collected.

Average monthly income ($000s)	Money spent at restaurants per month ($)
4.2	620
3.6	395
2.7	185
2.8	150
2.5	130
3.0	220
3.1	245
2.2	100
4.0	400
3.7	380
3.8	200
3.5	360
2.9	175
3.6	350
4.1	600

a. Draw a scatterplot of this data on technology of your choosing.
b. Find the equation of the least-squares regression line in terms of average monthly income in thousands of dollars (I) and money spent at restaurants in dollars (R). Give values correct to 4 significant figures.
c. Predict how much a person who earns $ 5000 a month might spend at restaurants each month.
d. Explain why part c is an example of extrapolation.
e. A person spent $ 265 eating out last month. Estimate their monthly income, giving your answer to the nearest $ 10. Is this an example of interpolation or extrapolation?

18. Data on 15 students' marks in Geography and Music assessments were collected.

Geography	Music
65	91
80	57
72	77
61	89
99	51
54	76
39	62
66	87
78	88
89	64
84	90
73	45
68	60
57	79
60	69

a. Is there an obvious explanatory variable in this situation?
b. Draw a scatterplot of this data on your calculator, using the marks in Geography as the explanatory variable.
c. Find the equation of the line of best fit. Give values correct to 4 significant figures.
d. Based on your equation, if a student received a mark of 85 for Geography, what mark (to the nearest whole number) would you predict they would receive for Music?
e. How confident do you feel about making predictions for this data? Explain your reasons.
f. Calculate Pearson's product–moment correlation coefficient, r for this data. How can you use this value to evaluate the reliability of your data?

19. For three months, Cameron has been wearing an exercise-tracking wristband that records the distance he walks and the number of calories he burns. A graph shows his weekly totals. The regression line equation for the data where y represents the number of calories burned and x represents the distance walked is $y = 14\,301 + 115.02x$.

a. Identify the response variable in this situation.
b. Rewrite the equation in terms of the explanatory and response variables.
c. Using the equation for the regression line, determine the number of calories burned if a person walked 50 km in a week. Is this an example of interpolation or extrapolation? Explain.
d. Due to an injury, in one-week Cameron only walked 10 km. Use the data to determine the number of calories this distance would burn. Is this an example of interpolation or extrapolation? Explain.
e. Pearson's product–moment correlation coefficient for this data is 0.9678. How can you use this value to evaluate the reliability of the data?
f. List at least two other factors that could influence this data set.

20. A phone carrier company used fingerprint biometric technology to collate some results about people using a particular social media application. A graph is drawn that shows their results for people's ages and the number of social media friends they had. The regression line equation for this data set is

$y = -13.613x + 777.84$, where y represents the number of social media friends and x represents the ages of the people.

a. Identify the explanatory variable in this situation.
b. Rewrite the equation in terms of the explanatory and response variables.
c. Pearson's product–moment correlation coefficient for this data set is -0.893. How can you use this value to evaluate the reliability of the data?
d. Using the regression line equation for this data set, determine the number of friends a 35 − year-old person is likely to have on this social media application. Is this an example of interpolation or extrapolation? Explain your response.

2.4 The coefficient of determination and residual plots

2.4.1 The coefficient of determination (r^2)

In Chapter 1, the strength of a linear relationship was determined using Pearson's correlation coefficient (r). **The coefficient of determination**, r^2, is Pearson's correlation coefficient squared.

If a value for r^2 of 0.71 is found, for example, this would indicate that 71% of the variation in the y-variable is explained by the variation in the x-variable and 29% can be explained by other factors.

The coefficient of determination can be calculated using an Excel spreadsheet or other suitable technology.

WORKED EXAMPLE 8

Data was collected on the time it takes students to get to school and their ATAR scores.

Time (mins)	12	35	19	42	33	31	25	46	45	40	14	44	39	31	22
ATAR score	53	75	97	59	87	70	71	66	37	48	94	68	33	59	42

Appropriate technology should be used to answer the following questions.

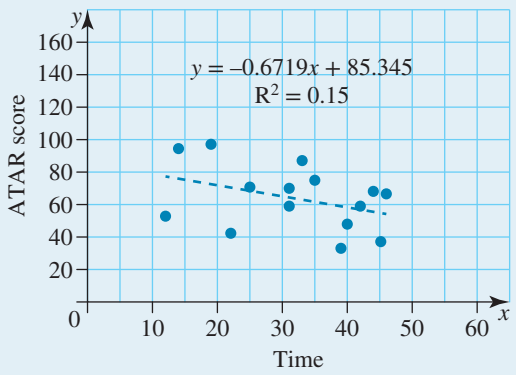
a. **Draw a scatterplot of this data.**
b. **Determine the equation of the line of best fit.**
c. **Determine Pearson's correlation coefficient, r, and the coefficient of determination, r^2.**
d. **Interpret the values of r and r^2.**

THINK

a. Using Excel, draw a scatterplot representing the data collected. Time is the explanatory variable and ATAR score is the response variable.

WRITE

a.



b. Write the equation of the line of best fit using appropriate variable names.

c. Write the value of r^2 and r. r^2 as given by Excel when a line of best fit is added.

d. Interpret the r value.
Interpret the r^2 value

b. ATAR score $= -0.6719 \times \text{Time} + 85.345$

c. $r^2 = 0.15$
$r = \sqrt{0.15}$
$= -0.39$ (negative gradient on line)

d. An r value of -0.39 implies that there is a weak negative linear relationship, which is not a clear indication that an increased distance from school could have a negative effect on your ATAR score.
An r^2 value of 0.15 implies that 15% of the variation in the ATAR score can be explained by the variation in the distance students live from their schools and 85% is explained by other factors.

2.4.2 Residual plots and analysis

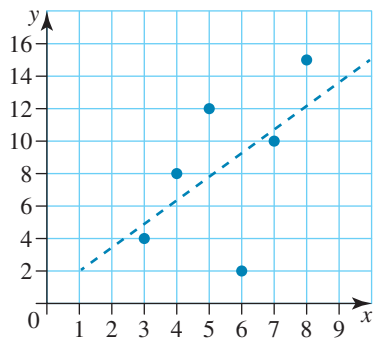
In the world of statistical modelling, it is often not enough to use a scatterplot to determine if a linear model is appropriate. There may be other underlying patterns that are difficult to recognise simply by studying the scatterplot. A **residual plot** can be constructed using the least-squares regression line to highlight any underlying patterns which would indicate that the data does not fit a linear model.

Calculating residuals

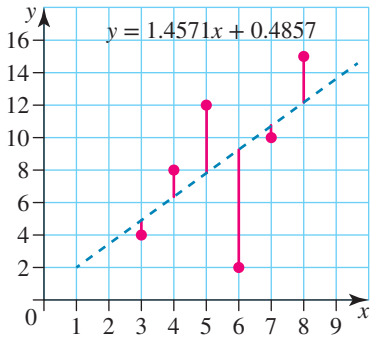
Consider the following set of data.

x	3	4	5	6	7	8
y	4	8	12	2	10	15

The following graph shows the data plotted on a scatterplot and a regression line fitted.
For this set of data, the linear regression equation is $y = 1.4571\,x + 0.4857$



The lengths of the vertical lines joining the data points to the regression line are the **residuals**.



For each value of x (explanatory variable), there is the actual value y (response variable) from the data supplied and there is the predicted value, found from the linear regression equation.

$$\text{Residual value} = \text{actual } y - \text{predicted } y - \text{value}$$

In the data in the previous table for the x-value of 3, the actual y-value is 4 and the predicted y-value using the linear regression equation is:

$$\begin{aligned} y &= 1.4571x + 0.4857 \\ &= 1.4571 \times 3 + 0.4857 \\ &= 4.857 \end{aligned}$$

$$\begin{aligned} \text{Residual value} &= \text{Actual value} - \text{predicted value} \\ &= 4 - 4.857 \\ &= -0.86 \text{ (correct to 2 decimal places)} \end{aligned}$$

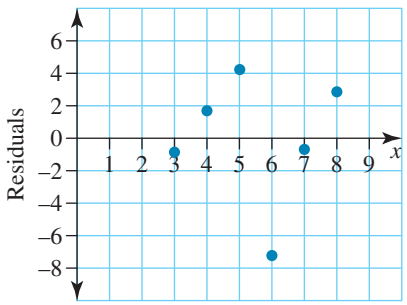
The following table is completed for all the predicted y-values and the residuals.

x	3	4	5	6	7	8
Actual y-value	4	8	12	2	10	15
Predicted y-value	4.86	6.31	7.77	9.23	10.69	12.14
Residual	−0.86	1.69	4.23	−7.23	−0.69	2.86

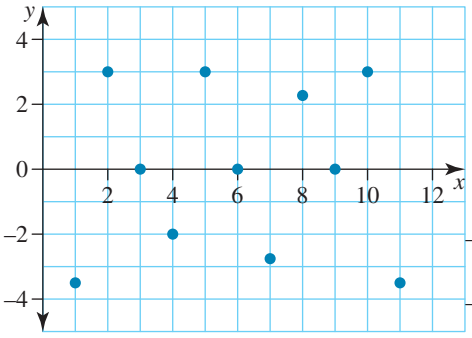
Residual plots

A residual plot is a graph of the points obtained by plotting the residuals on the vertical axis and the explanatory variable (x-value) on the horizontal axis. When the points in a residual plot are randomly spread around the horizontal axis, a linear regression model is appropriate for the data; otherwise a non-linear model should be considered.

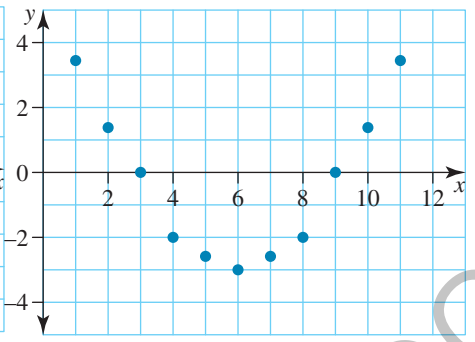
For the data in the previous table a residual plot is completed below.



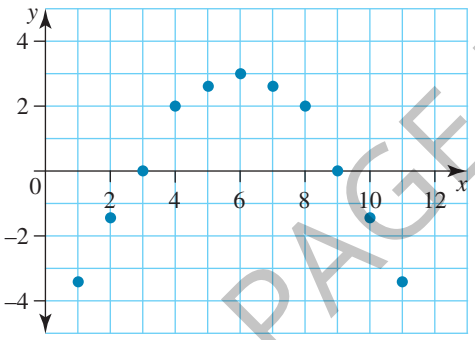
The previous residual plot shows a random pattern, therefore the linear model is seen as a good fit. The following residual plots show some typical patterns for residuals.



Random pattern
Linear model appropriate



Non random
Linear model *not* appropriate



Non-random
Linear model *not* appropriate

If a residual plot showed non-random patterns then a non-linear model for the original data should be investigated. Non-linear models are outside the scope of this course.

WORKED EXAMPLE 9

For the set of data below, the least-squares regression equation is $y = -0.56x + 45.6$. Complete the table.

x	50	55	60	65	70
Actual y-value	20	10	15	8	7
Predicted y-value					
Residuals					

THINK

1. Calculate the predicted y-values:
   $y = -0.56x + 45.6$.
   $y = -0.56 \times 50 + 45.6 = 17.6$
   $y = -0.56 \times 55 + 45.6 = 14.8$
   $y = -0.56 \times 60 + 45.6 = 12$
   $y = -0.56 \times 65 + 45.6 = 9.2$
   $y = -0.56 \times 70 + 45.6 = 6.4$

WRITE

x	50	55	60	65	70
Actual y-value	20	10	15	8	7
Predicted y-value	17.6	14.8	12	9.2	6.4

2. Calculate the residual values:
   Residual value = actual y-value – predicted y-value
   $20 - 17.6 = 2.4$
   $10 - 14.8 = -4.8$
   $15 - 12 = 3$
   $8 - 9.2 = -1.2$
   $7 - 6.4 = 0.6$

x	50	55	60	65	70
Actual y-value	20	10	15	8	7
Predicted y-value	17.6	14.8	12	9.2	6.4
Residuals	2.4	−4.8	3	−1.2	0.6

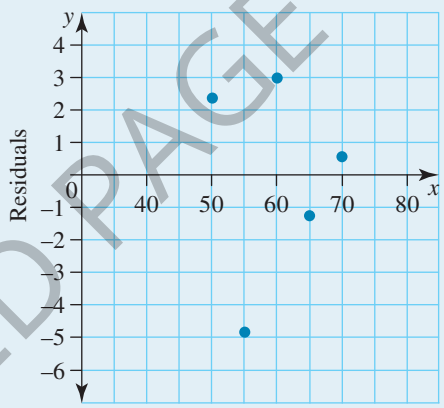
WORKED EXAMPLE 10

For the set of data in Worked example 9, plot the residuals and interpret the graph.

THINK

1. Draw a set of axes with the x value on the horizontal axis and the residual value on the vertical axes. Plot the points on the axes.

WRITE



2. The points do not appear to form a pattern.

As the residuals appear to be randomly spread above and below the x-axis, then a linear model is an appropriate model for this set of data.

on Resources

Interactivity: Pearson's product–moment correlation coefficient and the coefficient of determination (int-2451)

studyon

Units 3 & 4 > Area 1 > Sequence 1 > Concept 3

The coefficient of determination and residual plots Summary screen and practice questions

Exercise 2.4 The coefficient of determination and residual plots

1. **WE8** Data on the daily sales of gumboots and the maximum temperature was collected.

Temperature (°C)	17	16	12	10	14	17	18	22	23	19	17	15	12	15	20
Daily sales (no. of pairs)	2	3	8	16	7	3	2	1	1	2	3	3	12	9	1

Appropriate technology should be used to answer the following questions.

   a. Draw a scatterplot of this data.
   b. Find the equation of the line of best fit.
   c. Find Pearson's product–moment correlation coefficient, r, and the coefficient of determination, r^2.
   d. Interpret these values.

2. **MC** If Pearson's correlation coefficient, r, is found to be -0.7564, then the coefficient of determination, r^2, will be

   A. 0.2436. **B.** -0.7564. **C.** -0.4279. **D.** 0.5721.

3. **MC** If the coefficient of determination, r^2, is found to be 0.5781, the percentage of variation that can be explained by other factors is

   A. 57.81%. **B.** 42.19%. **C.** 76.03%. **D.** 23.97%.

4. Calculate the value of the coefficient of determination, r^2, in the following scenarios.

   a. A study was conducted, and it was found that the association between a child's diet and their health is $r = 0.8923$.
   b. The association between global warming and the amount of water in the oceans was found to be $r = 0.9997$.
   c. Interpret these values.

5. **WE9** For the set of data below, the least- squares regression equation is $y = 1.27x + 7.65$. Complete the following table.

x	13	18	23	28	33	38
Actual y-value	25	31	40	36	48	60
Predicted y-value						
Residuals						

6. **WE10** For the set of data in question 5, plot the residuals and interpret the graph.

7. A woman diagnosed as anaemic has a level of 120 g/L of iron at her initial blood test. She agreed to join a research group to determine how quickly an iron supplement in capsule form, administered daily, would impact on her iron levels.

   Her iron level was measured once a week. The following data were collected.

Week of experiment	1	2	3	4	5	6
Iron level	120	122	130	135	135	140

a. Using appropriate technology, construct a scatterplot and determine the equation of the least-squares regression line.

b. Complete the following table.

Week of experiment	1	2	3	4	5	6
Iron level	120	122	130	135	135	140
Predicted iron level						
Residuals						

c. Draw a residual plot and interpret the plot.

d. Determine the coefficient of determination using technology and hence calculate Pearson's correlation coefficient.

e. Using the information obtained in parts **a**, **b**, **c** and **d**, is this a suitable model to use to determine when the woman will reach a healthy iron count of 155 g/L of iron?

8. A farmer's market is held once a month and features local produce, handicrafts and trash and treasure stalls.

The number of local produce and handicraft stalls varies each weekend and is announced prior to the weekend on the website for the market. The number of local produce and handicraft stalls participating, together with the number of visitors to the market each month, is shown in the following table.

Number of stalls	53	34	61	32	61	25
Number of visitors	501	339	611	300	450	333

The least-squares regression line for this data is

$\text{Number of visitors} = 6.64 \times \text{Number of stalls} + 128.11$

a. Using the least-squares regression line equation, complete the following table.

Number of stalls	53	34	61	32	61	25
Number of visitors	501	339	611	300	450	333
Predicted number of visitors						
Residuals						

b. Draw a residual plot and interpret the plot.

c. If the coefficient of determination for this set of data is 0.7681, determine the Pearson correlation coefficient, correct to two decimal places.

d. Using the information obtained in parts **a**, **b** and **c**, is this a suitable model to use to predict the number of visitors based on the number of stalls?

9. The following table represents the costs for transporting a consignment of surfboards from Brisbane factories.

The cost is given in terms of distance from Brisbane. There are two factories which can be used. The data are summarised below.

Distance from Brisbane (km)	10	20	30	40	50	60	70	80
Factory 1 cost ($)	70	70	90	100	110	120	150	180
Factory 2 cost ($)	70	75	80	100	100	115	125	135

a. Using appropriate technology, construct a scatterplot and determine the equation of the least-squares regression line for each factory.

b. Complete the table.

Distance from Brisbane (km)	10	20	30	40	50	60	70	80
Factory 1 cost ($)	70	70	90	100	110	120	150	180
Predicted cost for factory 1								
Residuals for factory 1								
Factory 2 cost ($)	70	75	80	100	100	115	125	135
Predicted cost for factory 2								
Residuals for factory 2								

c. Draw a residual plot for each factory and interpret the plots.

d. Calculate Pearson's correlation coefficient and the coefficient of determination for each factory.

e. Which factory is likely to have the lower cost to transport to a shop in Brisbane?

f. Which factory is likely to have the lower cost to transport to Mytown, 115 kilometres from Brisbane?

10. The following table contains the age and systolic blood pressure (SBP) for a group of volunteers at a University Health Science department.

Age	37	38	40	42	45	48	50	52	53	55
Systolic blood pressure	130	140	132	149	144	157	161	145	165	162

a. A linear regression equation of the form $y = a + bx$ was calculated for this data and the results were $b = 1.61$, $a = 74.35$, $r = 0.84$. Use this information to calculate the predicted systolic blood pressure and the residuals.

Age	37	38	40	42	45	48	50	52	53	55
Systolic blood pressure	130	140	132	149	144	157	161	145	165	162
Predicted systolic blood pressure										
Residuals										

b. Plot the residuals and comment on the likely linearity of the data.

c. Estimate the systolic blood pressure at age 75 and comment on the reliability of this estimation.

11. Consider the following data set.

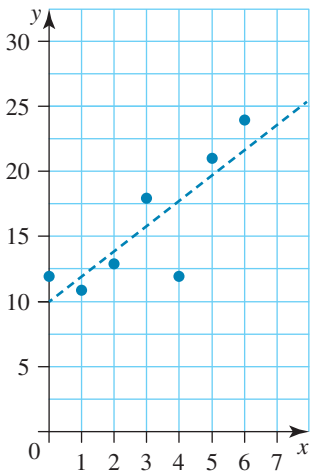
x	1	2	3	4	5	6	7	8	9	10
y	1	5	10	16	26	34	50	62	80	101

a. A linear regression equation of the form $y = a + bx$ was calculated for this data and the results were $a = -21.333$, $b = 10.879$, $r = 0.97$. Use this information to calculate the predicted values of y and the residuals.

b. Plot the residuals and comment on the likely linearity of the data.

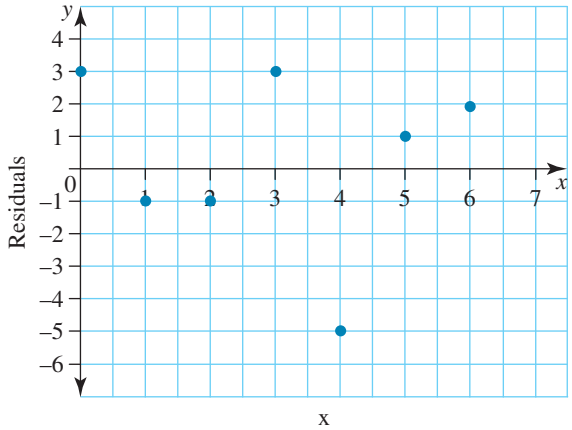
The following information relates to questions 12 and 13.

The following graph shows a least-squares regression line fitted to a set of data. Use this graph to answer questions **12** and **13**.

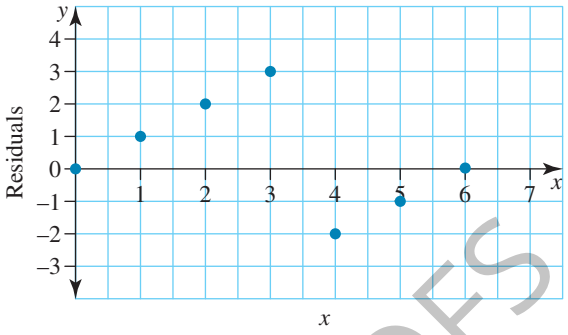


12. From the following options, choose the residual plot that best represents the residuals for this line.

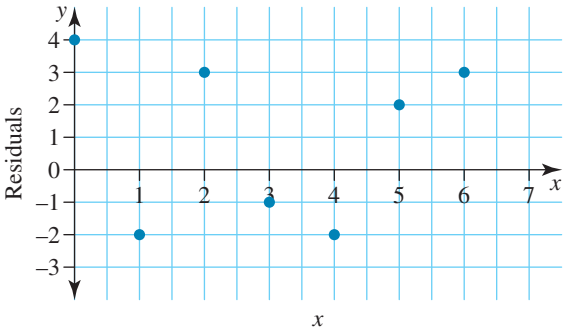
a.



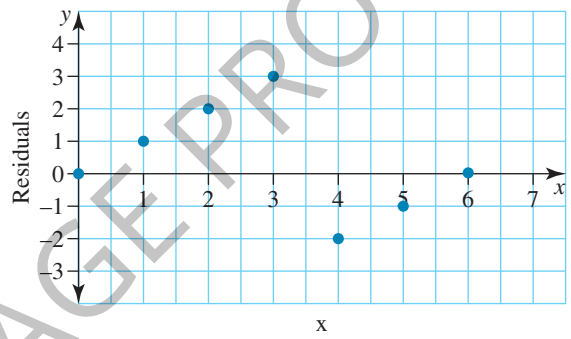
b.



c.



d.



13. MC The value of the product–moment correlation coefficient (r) for this data is closest to

A. −0.81. **B.** −0.33. **C.** 0.33. **D.** 0.81.

14. The coefficient of determination for a set of data is found to be 0.78.

Complete the following sentence.

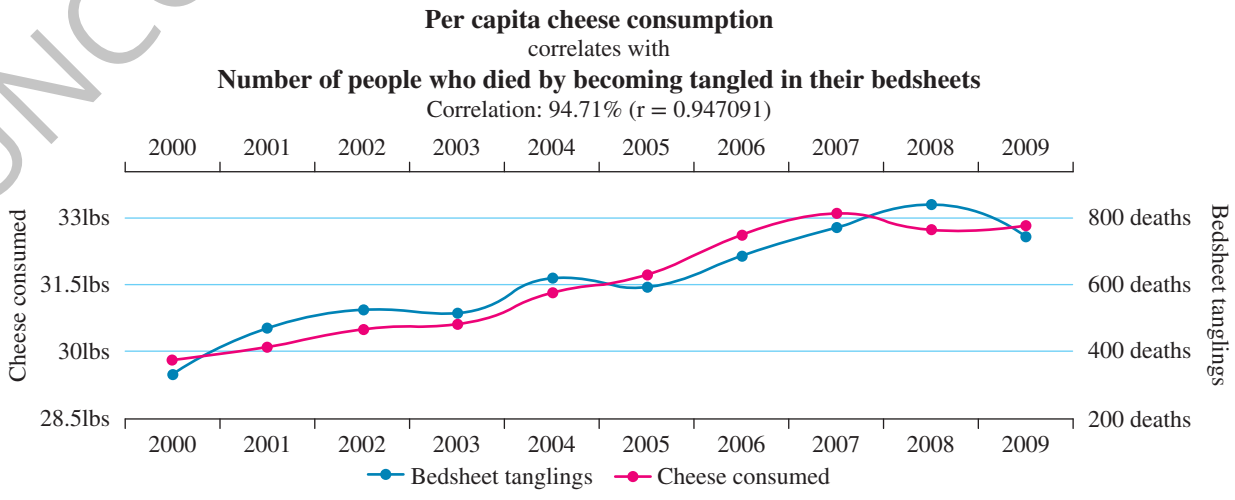
78% of the variation in the _____ variable can be explained by the variation in the _____ variable. _____% of the variation can be explained by other factors.

2.5 Association and causation

2.5.1 Observation and association

Spurious correlations (http://www.tylervigen.com/spurious-correlations) is a website and now a book created by Tyler Vigen. The website shows many examples of graphs of data sets that appear to have a relationship. For example, the following graph shows an association of $r = 0.95$ for the Number of people who died by becoming entangled in their bedsheets and per capita cheese consumption from 2000 to 2009.

This is an extreme example of why **association** does not necessarily imply **causation**.



Establishing a high degree of association between two variables can provide a starting point for an investigation but establishing causation requires further **observation** and **experimentation**. For example, a basketball fan noted that there was a high positive association between the height of a basketballer and the number of points he scored. Does this mean that if a basketball team needs players with greater scoring ability then they just should recruit more tall players? An experienced recruiter knows that height is only one advantage for a basketballer; fitness levels, skill level, hand–eye coordination and ability to read the game are all necessary attributes of a player who can consistently score. The recruiter would observe all of these attributes of potential new players and study the statistics of all the variables, before making a decision.

Therefore, a high degree of association between two variables does not mean that there is a causal relationship between the variables, so further investigation is necessary.

One way to establish causation is to conduct experiments where a control group is used. Agricultural scientists researching the effectiveness of a new fertiliser designed to improve the productivity of a new type of tomato plant split the population (the tomato plants) into two groups. One group of plants will be fed the fertiliser in water at regular intervals, the other group of plants are given the same amount of water without the fertiliser and all other variables are the same across both groups of plants. (Other variables could include the type of soil and weather conditions.) A study of the association between administering the fertiliser and the size of the tomato crop can then establish causation. In this example, the administered water and fertiliser is the explanatory variable, and the size of the tomato crop is the response variable.

2.5.2 Common response, confounding and coincidence explanations

In some cases, the association between two variables can be explained by a **common response** which provides the association. For example, a study may show that there is a strong association between the GDP (gross domestic product) of a country and the infant mortality rate. While a larger GDP will not directly lead to a lower infant mortality rate, a common response — the money spent on childhood vaccination programs — provides a direct link to both variables and is more likely to be the underlying cause for the observed association.

In another example, a study of a group of people may provide a strong association between the number of deaths from heart attacks (response variable) and distance from a major hospital (explanatory variable). But there are other factors that can influence the number of deaths, such as diet, age, starting weight and gender. These are called **confounding variables**. If a study does not take confounding variables into account when setting up the study the results cannot be used to show causality.

Sometimes an association between variables can just be a **coincidence**, as in the example in the previous graph, which shows an association between cheese consumption and deaths from becoming entangled in bed sheets.

When observing an association between two variables it can never be stated that the one variable causes a response in the other variable. In professional research many controlled experiments must be carried out to determine and identify a causal relationship.

WORKED EXAMPLE 11

Data showing the average weekly hours of exercise for a group of 8 people and their LDL-cholesterol reading are as given.

Average weekly hours of moderate exercise	7	1	3	8	3	10	6	2
LDL-C	100	125	130	120	128	90	115	135

a. Construct a scatterplot for the data
b. Comment on the association between the number of hours exercise and the LDL-C reading.
c. Calculate *r*.
d. Based on the value of *r* obtained in part c, would it be appropriate to conclude that the decrease in the LDL-C reading is caused by the increasing hours of exercise?

THINK

a. 1. Determine the explanatory and response variables
2. Sketch a scatterplot

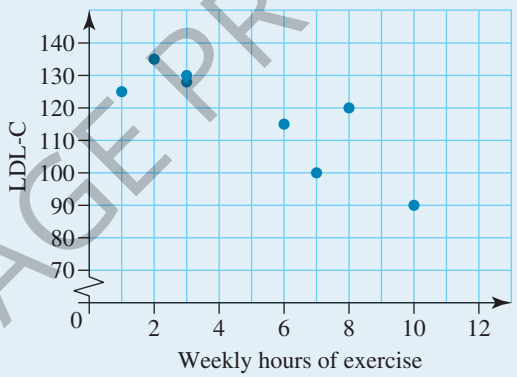
b. A negative association can be observed.

c. Use technology to find the value of *r*.

d. Association does not imply causation.

WRITE

a. Weekly hours of exercise (explanatory)
LDL-C reading (response)



b. From the scatterplot there appears to be a strong negative association between the weekly hours of exercise and the LDL-C reading.

c. $r = -0.85$

d. Just because $r = -0.85$, it cannot be stated that increasing weekly hours of exercise will cause a decrease in the LDL-C reading. Other factors may need to be considered. Further investigation is needed before any conclusions can be drawn.

on Resources

Interactivity: (int-xxxx)

studyon

Units 3 & 4 > Area 1 > Sequence 2 > Concept 1

Association and causation Summary screen and practice questions

Exercise 2.5 Association and causation

1. **WE11** Data showing the number of rose bushes in the gardens of 10 houses and the annual income for each house is given.

Number of roses	4	2	10	5	3	11	6	7
Annual household income ($)	80 000	32 000	120 000	65 000	21 000	122 000	75 000	82 000

   a. Using technology, construct a scatterplot for the data
   b. Comment on the association between the number of rose bushes and the household income.
   c. Calculate r using technology.
   d. Based on the value of r obtained in part **c**, would it be appropriate to conclude that the increase in household income is caused by an increase in the number of rose bushes in the garden?

2. Data was collected on the level of aerobic fitness of 100 15-year-olds and the amount of time they spent playing computer games. The correlation coefficient for this data was -0.86.
   What can be said about the association between the level of aerobic fitness and the amount of time spent playing computer games for this group of 15-year-olds?

3. **MC** During the months of August and September there was a strong positive association ($r = 0.93$) between the number of sightings of whales in the Whitsundays each week and the number of fines issued for boating infringements off Airlie Beach.
   A strong positive association was also found both between the number of whale sightings ($r = 0.96$), the number of fines issued for boating infringements ($r = 0.82$), and the number of visitors to Airlie Beach. Using this information, which of the following statements is true?
   A. It is just a coincidence that there is a strong positive association between the number of sightings of whales in the Whitsundays and the number of fines issued for boating infringements.
   B. During the months of August and September, people are so excited about seeing whales that they forget to follow the rules for driving a boat.
   C. Tourists don't know that there are rules to be followed when at sea.
   D. August and September attract larger numbers of visitors and more visitors means more people taking out boats which leads to more boating infringements. The association between the number of whale sightings and the number of boating infringements can be explained by the common response variable, the number of tourists at Airlie Beach.

The following information applies to questions 4 and 5.
A set of data was collected from a large group of professional sportspeople. They were asked the number of hours they trained per week and the amount of money they earned. The results were recorded, and the value of Pearson's correlation coefficient was found to be 0.87.

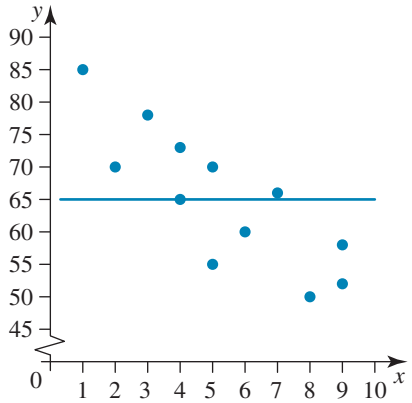
4. MC Which of the following is NOT true?
   - **A.** There is a positive association between the number of hours of training and the amount of money earned.
   - **B.** The association between the number of training hours and the amount of money earned can be classified as strong.
   - **C.** The linear relationship between the two variables suggests that as the number of training hours increase, so too does the amount of money earned.
   - **D.** The increase in the number of training hours causes the increase in the amount of money earned.
5. MC Which of the following is NOT true?
   - **A.** The coefficient of determination is about 0.77.
   - **B.** About 77% of the variation in the number of hours of training can be explained by the variation in the amount of money earned.
   - **C.** Other factors such as the type of sport played can affect the amount of money earned.
   - **D.** The number of training hours is the major factor in predicting the amount of money earned.
6. It has been found that there is a strong positive association between the number of fire fighters that attend a fire and the amount of damage caused by the fire. Does this mean that more fire fighters cause more fire damage? What could be the common cause that links these variables?
7. There is a strong positive association between the shoe size of a child and their academic knowledge. Does a larger shoe size mean a smarter child? What could be the common cause(s) that links these two variables?
8. A group of university students volunteered to be part of a healthy food supplement trial. Their health was monitored every month for 10 months. A positive association was found between improvements in good health statistics such as cholesterol levels and blood pressure and the consumption of the supplement. Does this mean that the consumption of the supplement causes improvement in health? What other confounding variables could also contribute to this association?

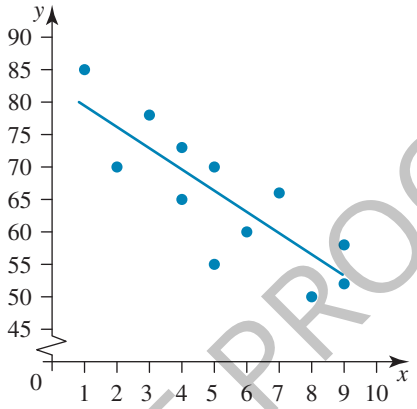
2.6 Review: exam practice

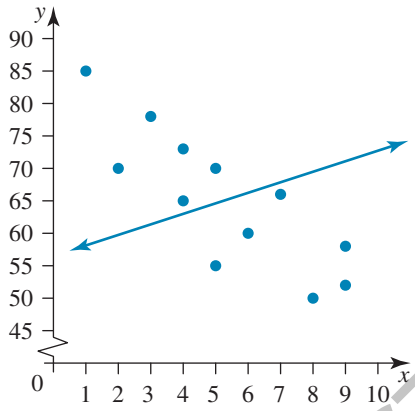
A summary of this chapter is available in the Resources section of your eBookPLUS at www.jacplus.com.au.

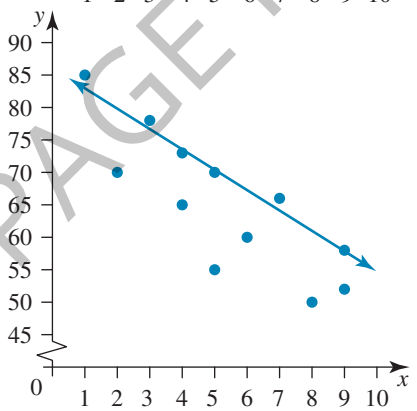
Simple familiar

1. **MC** Which of the following scatterplots best demonstrates a line of best fit?

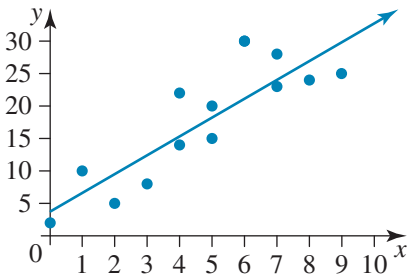
A.


B.


C.


D.


2. **MC** The regression line equation for the graph shown is closest to



A. $y = 3.8 + 2.9x$.
B. $y = -3.8 - 2.9x$.
C. $y = -3.8 + 2.9x$.
D. $y = 3.8 - 2.9x$.

3. **MC** A gardener tracks a correlation coefficient of 0.79 between the growth rate of his trees and the amount of fertiliser used. What can the gardener conclude from this result?

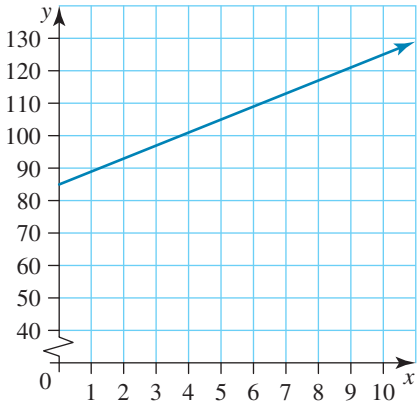
A. An increase in tree growth increases the use of fertiliser.
B. An increase in the use of fertiliser increases the health of the tress.
C. 79% of the variation in the growth rate of his trees can be explained by the variation in the amount of fertiliser used.
D. The growth rate of the trees influences the quality of the fertiliser used.

4. **MC** When $y = 0.54 + 15.87x$, the value of y when $x = 2.5$ is

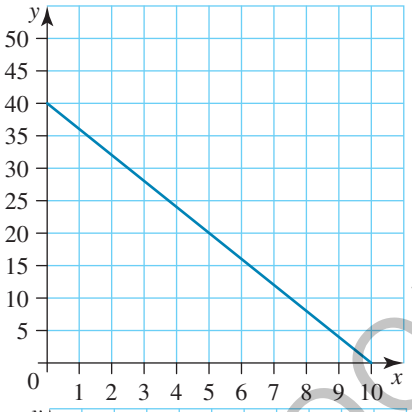
A. 18.91.
B. 40.215.
C. 39.135.
D. 6.888.

5. MC The graph for the regression line equation $y = 85 - 4x$ is most likely to be

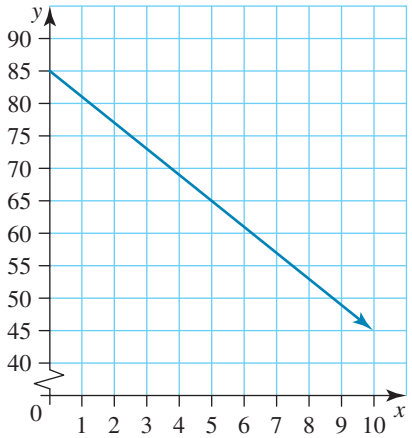
A.



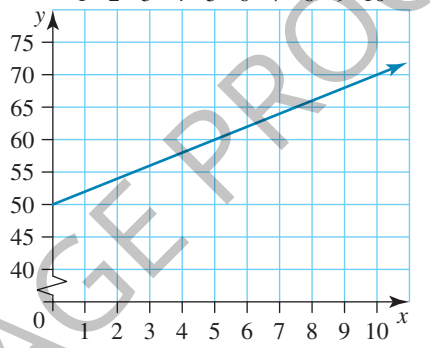
B.



C.



D.



6. MC A series of data points recorded a coefficient of determination value of 0.82. Calculate Pearson's correlation coefficient.

A. 82% **B.** 0.18 **C.** 0.67 **D.** 0.91

7. MC For the following sample data set, which of the following is an example of interpolating data?

x	1	5	15	25
y	10	16	18	22

A. Finding the value of x when $y = -7$
B. Finding the value of y when $x = 17$
C. Finding the value of x when $y = 27$
D. Finding the value of y when $x = 37$

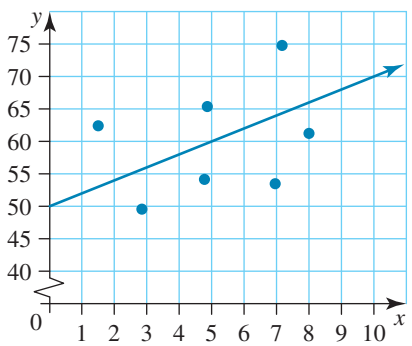
8. MC For the data set $\bar{x} = 11.5, \bar{y} = 16.5, s_x = 10.8, s_y = 5, r = 0.94$ the regression line equation is closest to

A. $y = 10 + x$.
B. $y = 0.435 + 11.496x$.
C. $y = 0.876 + 0.936x$.
D. $y = 11.496 + 0.435x$.

9. MC Consider the regression line drawn on the scatterplot shown.

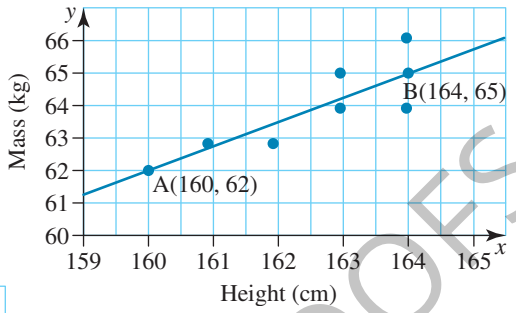
The gradient of the regression line is closest to

A. $\frac{5}{3}$.
B. $-\frac{5}{3}$.
C. $\frac{3}{5}$.
D. $-\frac{3}{5}$.

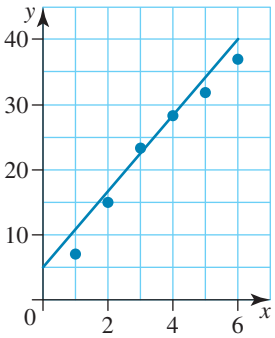


10. MC Consider the regression line drawn on the scatterplot shown below.
The equation of the regression line is

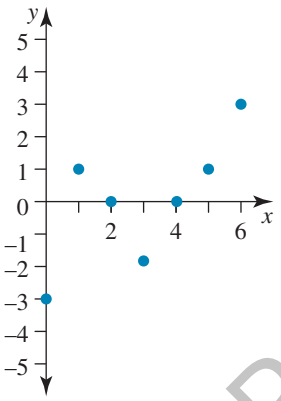
A. height $= 0.75 \times$ mass $- 61$.
B. height $= 0.75 \times$ mass $+ 61$.
C. height $= -0.75 \times$ mass $- 61$.
D. mass $= 0.75 \times$ height $+ 61$.



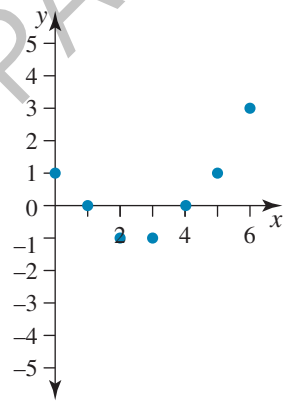
11. MC A least-squares regression line is fitted to the 7 points shown in the figure. Which of the following looks most similar to the plot of residuals?



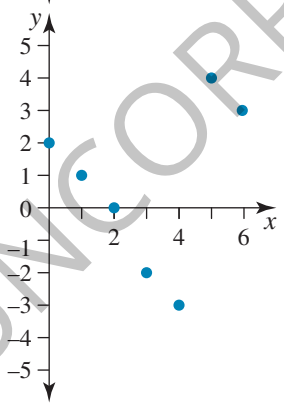
A.



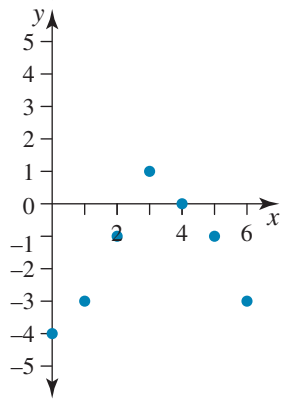
B.



C.



D.



12. MC A study of a group of people found that there was a strong positive association ($r = 0.91$) between their blood pressure reading and the number of times they visited the podiatrist. A strong positive association was also found both between the blood pressure reading ($r = 0.95$), the number of times they visited the podiatrist ($r = 0.82$), and their age.
Using this information, which of the following statements is true?

A. High blood pressure is more common as one ages and problems with feet occur more often with ageing. The association between the blood pressure reading and the number of podiatry visits can be explained by the common response variable, the age of the person.

B. It is just a coincidence that there is a strong positive association between the blood pressure reading and the number of podiatrist visits.

C. No-one likes people touching their feet, so their blood pressure goes up when they visit the podiatrist.

D. Whenever a person has a blood pressure reading, they have to go to the podiatrist.

Complex familiar

13. Consider the following data set.

x	1	2	3	4	5	6	7	8	9	10
y	55	40	42	38	35	43	51	40	47	60

a. Using technology, plot the data and fit a least-squares regression line.

b. Using technology determine the coefficient of determination and interpret its value.

c. Calculate the correlation coefficient and explain its meaning.

d. Calculate the predicted y values and the residuals and hence complete the table.

x	1	2	3	4	5	6	7	8	9	10
y	55	40	42	38	35	43	51	40	47	60
Predicted y-value										
Residuals										

e. Construct the residual plot and use it to comment on the appropriateness of the assumption that the relationship between the variables is linear.

14. Identify the explanatory and response variable for each of the following scenarios:

a. In a junior Science class, students plot the time taken to boil various quantities of water.

b. Extra buses are ordered to transport a number of students to the school athletics carnival.

15. A scatterplot is drawn using the following data.

x	10	9	8	7	6	5	4	3	2	1
y	6	10	4	11	13	18	15	19		26

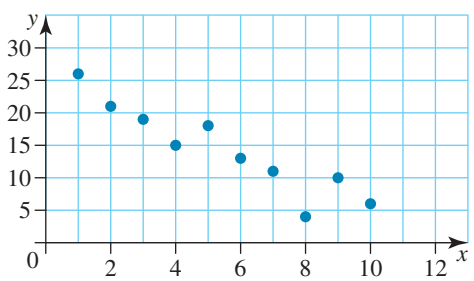
The following summary statistics has been obtained for this data.

$$\bar{x} = 5.5, s_x = 3.03, \bar{y} = 14.3, s_y = 6.86, r = -0.93$$

a. Comment on the form, direction and strength of the data.

b. Determine the least-squares regression line for this data.

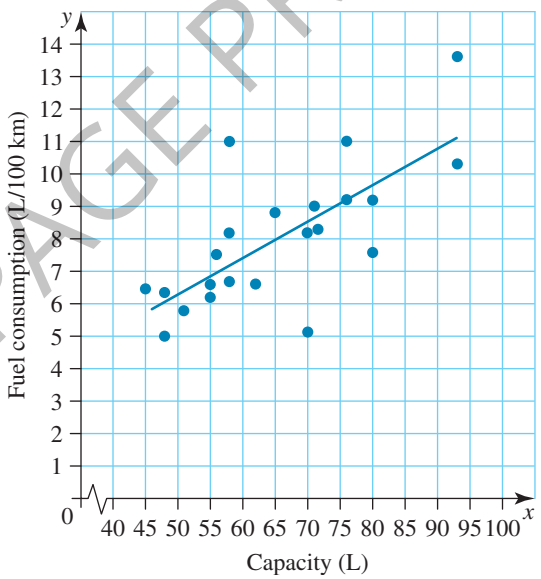
c. Use your answer from part **b** to calculate the value of x when y is 17. Is this interpolation or extrapolation?



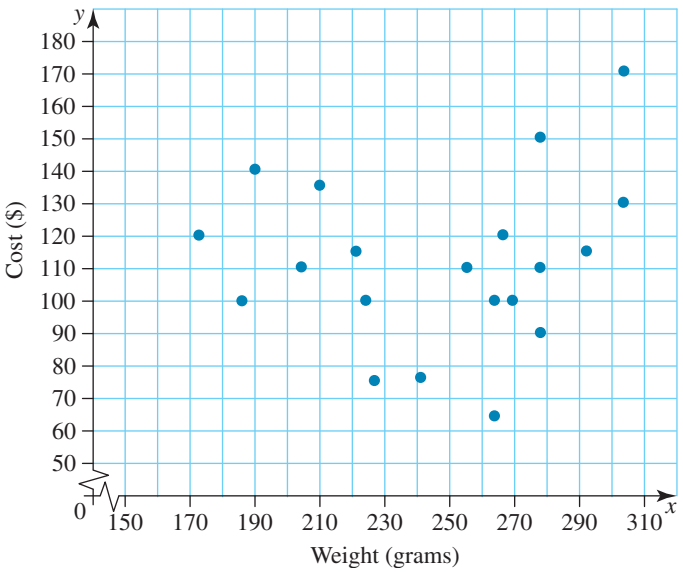
16. Pearson's correlation coefficient for a scatterplot was found to be -0.7564.
    a. Calculate the value of the coefficient of determination.
    b. What would these values indicate to you about the strength of relationship between the two variables?

Complex unfamiliar

17. During an interview investigating the link between the sales of healthy snack foods (functional foods) and the increasing consumer demand for these products, an advertising expert made the following comment:
    'There is an association but it's not causation ... our increasing need for healthy food and our laziness has resulted in mass innovation of functional foods.'
    Explain why he might have stated there is no causative link between the sales of healthy foods and laziness.
18. An independent agency test-drove a random sample of current model vehicles and measured their fuel tank capacity against the average fuel consumption. Along with the following scatterplot, a regression equation of $y = 0.1119x + 0.6968$ was established.
    a. Identify the response variable in this situation.
    b. Rewrite the equation in terms of the explanatory and response variables.
    c. It is often said that smaller vehicles are more economical. Determine correct to 2 decimal places the fuel consumption of a vehicle that had a 40-litre fuel tank.
    d. Is your answer to part **c** an example of interpolation or extrapolation? Explain your response.
    e. Calculate, correct to the nearest whole number, the tank size of a vehicle that had a fuel consumption rate of 10.2 L per 100 km.
    f. Pearson's correlation coefficient for this data is 0.516. How can you use this value to evaluate the reliability of your data?
    g. List at least two other factors that could influence the data.



19. The weight of top brand runners was tracked against the recommended retail price, and the results were recorded in the following scatterplot.
    a. Identify the explanatory variable for this situation.
    b. How would you describe the relationship between these two variables?
    c. The coefficient of determination for this data is $r^2 = 0.018\,72$. What conclusions can be established from this result?
    d. Identify two external factors that could explain the distribution of the data points.



20. The Bureau of Meteorology records data such as maximum temperatures and solar exposure on a daily and monthly basis. The following data table, for the Botanical Gardens in Melbourne, shows the monthly average amount of solar energy that falls on a horizontal surface and the monthly average maximum temperature. (*Note:* The data values have been rounded to the nearest whole number.)

Month	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Average solar exposure (MJ)	25	21	17	11	8	6	7	10	13	18	21	24
Average max daily temp. (°C)	43	41	34	33	24	19	24	24	28	32	25	40

a. Identify the explanatory and response variables for this situation.
b. Using technology, plot the data on a scatterplot.
c. Describe the trend of the data.
d. Calculate Pearson's coefficient and coefficient of determination for this data. What do these values tell you about the reliability of the data?
e. Plot the regression line for this data and write the equation in terms of the variables.
f. Using your equation, calculate the amount of solar exposure for a monthly maximum temperature of 37°C.
g. Extrapolate the data to find the average maximum temperature expected for a month that recorded an average solar exposure of 3 MJ.
h. Explain why part **g** is an example of extrapolation.

Answers

Exercise 2.2 Review of the general equation of a straight line

1. a. $y = mx + c$

b. m is the gradient of the graph and c is the y-intercept.

2. a. $m = -5, c = 4$ **b.** $m = 3, c = 11$

c. $m = -6, c = 0$ **d.** $m = \frac{5}{2}, c = 5$

e. $m = \frac{2}{3}, c = -6$ **f.** $m = 2, c = -1$

3. Complete the sentences below.

a. The equation $y = 3x - 1$ has a gradient of 3 units. This means that for every increase of 1 unit in the horizontal direction there is an increase of 3 units in the vertical direction.

b. The equation $y = -x + 1$ has a gradient of -1units. This means that for every increase of 1units in the horizontal direction there is an increase of -1 units in the vertical direction.

c. The equation $y = \frac{1}{3}x - 1$ has a gradient of $\frac{1}{3}$ units. This means that for every increase of 3 units in the horizontal direction there is an increase of 1 units in the vertical direction.

4. a. 2 **b.** 1 **c.** 1

d. $\frac{-1}{2}$ **e.** $\frac{-1}{4}$ **f.** $\frac{4}{3}$

g. $\frac{-4}{3}$ **h.** $\frac{-2}{7}$ **i.** 0

5. a. $y = 2x - 2$ **b.** $y = -2x - 1$

c. $y = \frac{-3}{8}x + 2$ **d.** $y = 4$

e. $y = x$ **f.** $y = -0.25x - 4$

6. a. $y = \frac{7x}{5} - \frac{1}{5}$ or $5y = 7x - 1$

b. $y = \frac{5x}{9} + \frac{34}{9}$ or $9y = 5x + 34$

c. $y = x + 6$

d. $y = 3$

e. $y = \frac{4x}{9} + \frac{4}{9}$ or $9y = 4x + 4$

f. $y = \frac{9x}{5} + \frac{7}{5}$ or $5y = 9x + 7$

7. a. **i.** $(0, -3)$

ii. $m = 2$

iii. $y = 2x - 3$

b. **i.** $(0, 0)$

ii. $m = \frac{1}{4}$

iii. $y = \frac{1}{4}x$

c. **i.** $(0, 6)$

ii. $m = -3$

iii. $y = -3x + 6$

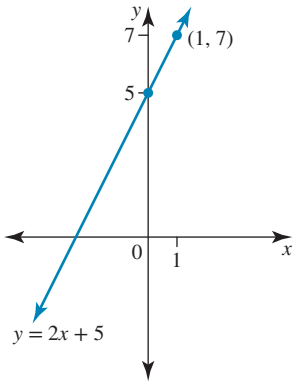
d. **i.** $(0, -3)$

ii. $m = \frac{5}{2}$

iii. $y = \frac{5}{2}x - 3$

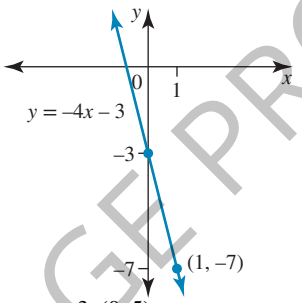
8. a. **i.** $m = 2, (0, 5)$

ii.



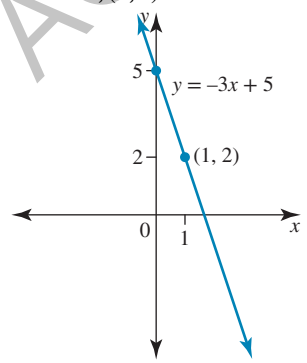
b. **i.** $m = -4, (0, -3)$

ii.



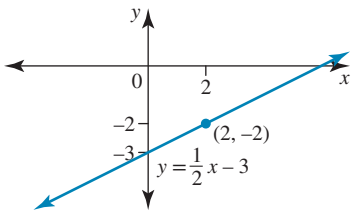
c. **i.** $m = -3, (0, 5)$

ii.



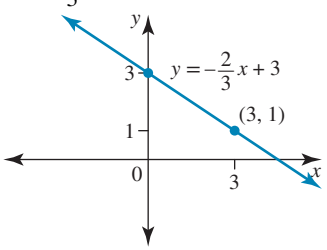
d. **i.** $m = \frac{1}{2}, (0, -3)$

ii.



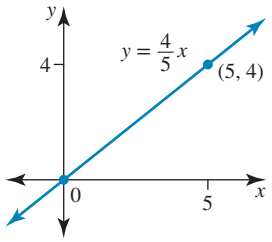
e. **i.** $m = \frac{-2}{3}, (0, 3)$

ii.

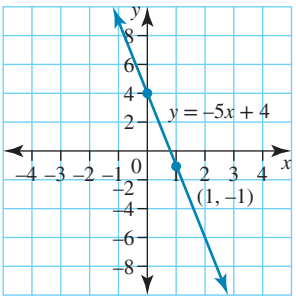


f. **i.** $m = \frac{4}{5}, (0, 0)$

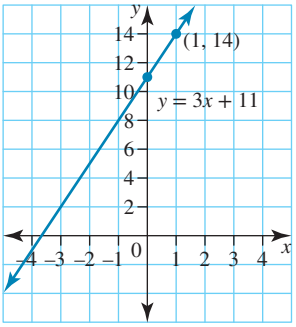
ii.



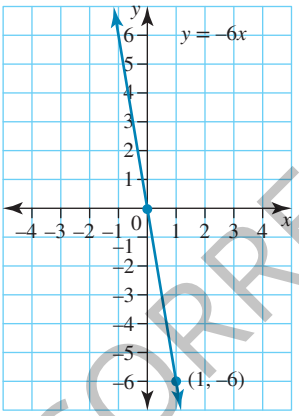
9. a.



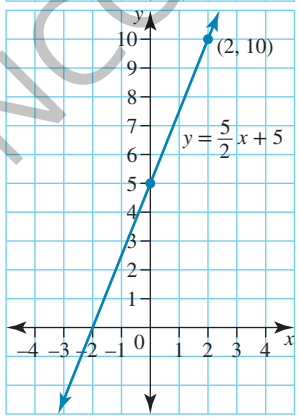
b.



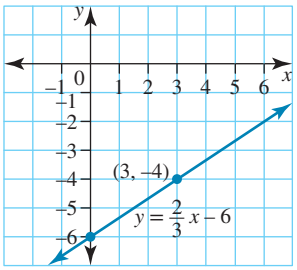
c.



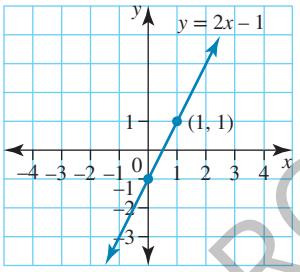
d.



e.



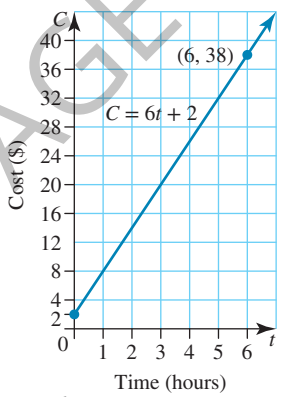
f.



10. a.

t	1	2	3
C	8	14	20

b.



c. $m = 6$

d. $C = 6t + 2$

e. $32

f. Answers will vary.

11. a. $y = 1$ **b.** $y = -7$ **c.** $y = 25$

d. $y = 41$ **e.** $y = -27$ **f.** $y = -19$

g. $y = 17.8$ **h.** $y = 3$

12. a. $x = \frac{-1}{4}$ **b.** $x = 2$ **c.** $x = -3$

d. $x = \frac{18}{4}$ **e.** $x = \frac{1}{2}$ **f.** $x = -0.1$

g. $x = \frac{-21}{4}$ **h.** $x = 0.375$

13. a. **i.** $237

ii. $475

b. **i.** $55

ii. $3.50

c. 70

d. $4.29

e. Yes

14. a. B

b. **i.** $280

ii. $360

iii. $600

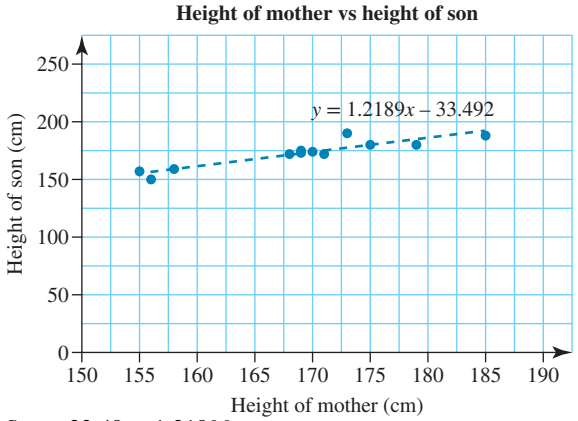
c. **ii.** save $120 and

iii. save $600

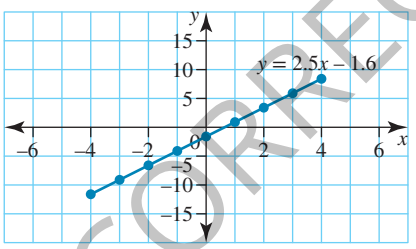
d. 7

Exercise 2.3 Fitting a least-squares line to data

1. a. The Mathematics exam mark
   b. 0.95
   c. −6.07
   d. 75%
2. a. $y = 3.33x + 91.78$
   b. $y = -0.15x + 21.87$
   c. $y = 52.38x + 8890.48$
   d. $y = \frac{1}{8}x + 18\frac{3}{4}$
3. a. The response variable is the height of the son
   b.



   c. $S = -33.49 + 1.219M$
4. a. −1.837 b. 1.701
   c. Positive direction
5. a. 105.9 b. −1.476
   c. Negative direction
6. a. Age b. 2.029 mL
   c. 7.5 months old
7. a. Cost per night b. $155.74
   c. 3.31 km
8. As the b value (gradient) is negative, the direction is negative. The y-intercept is 3.2: therefore, when $x = 0$, $y = 3.2$.
9. a.



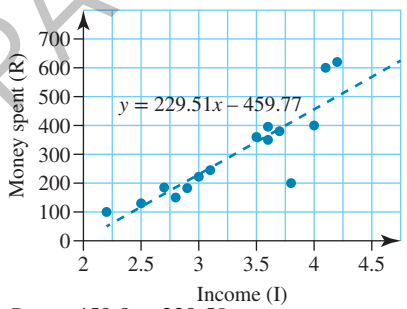
   b. Below the regression line
10. a. 60 b. −5
    c. Negative d. −140
11. a. 3.90
    b. Lucy incorrectly transposed the 12.9. She should have moved this first before dividing by 7.32.
12. a. −12 b. 25
    c. Positive d. 75.5
13. a. Number of insects caught b. 1.1
    c. Positive d. 66
14. $y = 4 + 2.5x$
15. a. $y = -1.783x + 39.41$. See the scatterplot in the worked solutions.
    b. When $x = 23$

$$y = -1.783 \times 23 + 39.41$$
$$= -1.599$$

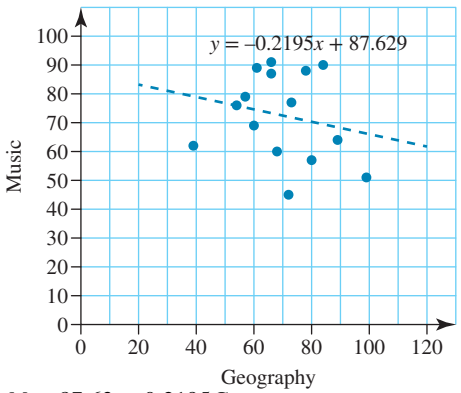
    c. When extrapolating data, it is assumed that the data will continue to follow the same trend line.
16. a. No. of mosquitoes $= 10.2 + 0.5 \times$ temp of fire
$= 10.2 + 0.5 \times 240$
$= 130.2$
The number of mosquitoes is 130 (correct to the nearest whole number)
    b. No. of mosquitoes $= 10.2 + 0.5 \times$ temp of fire
$12 = 10.2 + 0.5 \times$ temp of fire
$1.8 = 0.5 \times$ temp of fire

$$\text{Temperature of fire} = \frac{1.8}{0.5} = 3.6°\text{C}$$

If there were only 12 mosquitoes around the fire the temperature would be 3.6°C.
    c. Mosquitoes are in hibernation in the cooler months of the year, so once the temperature drops below a certain level this model would not be appropriate.
Also, the location of the fire, air temperature, wind conditions, proximity to water, etc. would impact on the mosquito population.
17. a.



    b. $R = -459.8 + 229.5I$
    c. $687.70
    d. Part c asks you to predict outside of the original data set range.
    e. $3160, interpolation
18. a. There is no obvious explanatory variable.
    b.



    c. $M = 87.63 - 0.2195G$
    d. 69
    e. Not very confident. The graph does not indicate a strong correlation between the two variables.

f. $r = -0.2172$. The value of r is low, showing that the line of best fit is poor and not a reliable predictor

19. a. Calories burned

b. Calories burned $= 14301 + 115.02 \times$ Distance walked

c. 20 052. Interpolation, as this data is inside the original range.

d. 15 451.2. Extrapolation, as the explanatory variable provided is outside the original data range.

e. An r value of 0.9678 indicates a very strong positive linear relationship, showing that the relationship between the two variables is very strong and can be used to draw conclusions.

f. Examples: speed of walking, difficulty of walking surface, foods eaten.

20. a. Age

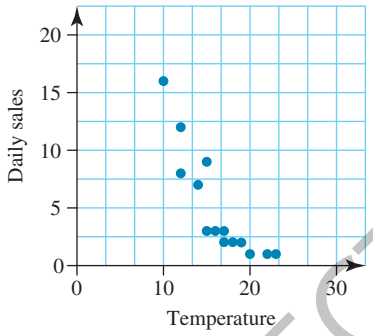
b. Number of social media friends $= -13.613 \times$ age $+ 777.84$

c. Since $r = -0.893$, the relationship between the two variables is a strong negative linear relationship.

d. 301; this is interpolation as the predicted value is within the given data values.

Exercise 2.4 The coefficient of determination and residual plots

1. a.



b. Daily sales $= 22.50 - 1.07 \times$ Temperature

c. $r^2 = 0.74$, $r = -0.86$

d. An r value of -0.86 implies that there is a strong negative linear relationship, which is an indication that an increase in temperature could mean a decrease in daily sales numbers of gumboots.

An r^2 value of 0.74 implies that 74% of the variation in the daily sales of gumboots can be explained by the variation in the temperature. 16% is explained by other factors.

2. D

3. B

4. a. 0.7962

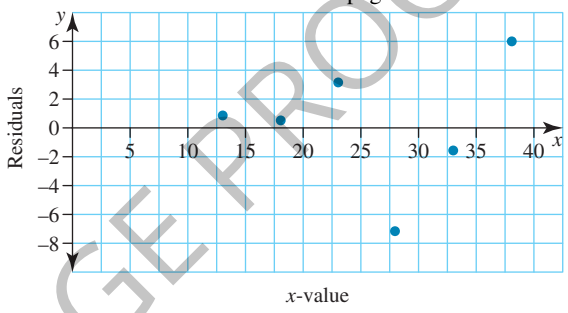
b. 0.9994

c. 79.62% of the variation in a child's health can be explained by the variation in their diet. 20% is explained by other factors.

99.94% of the variation in the amount of water in the oceans can be explained by global warming, 0.06% can be explained by other factors.

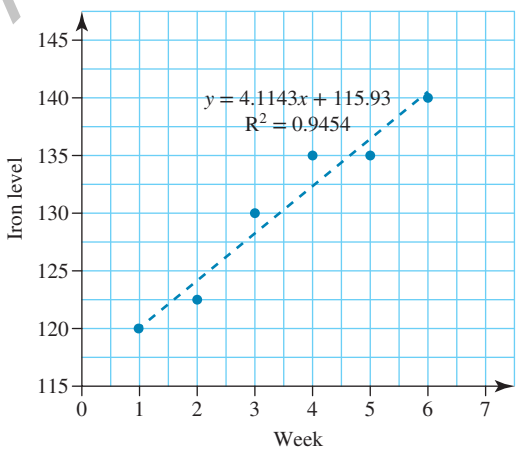
5. * See the table at the bottom of the page.

6.



As the residuals appear to be randomly spread above and below the x-axis, then this linear model is an appropriate model for this set of data.

7. a.



Iron level $= 4.11 \times$ Week $+ 115.93$

b. ** See the table at the bottom of the page.

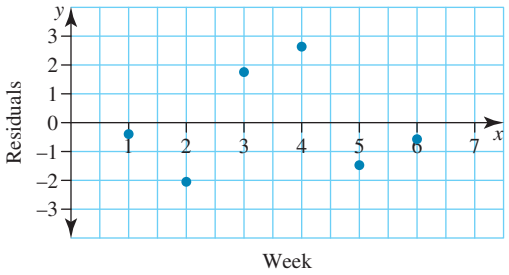
*5.

x	13	18	23	28	33	38
Actual y-value	25	31	40	36	48	60
Predicted y-value	24.16	30.51	36.86	43.21	49.56	55.91
Residuals	0.84	0.49	3.14	−7.21	−1.56	4.09

**7. b.

Week of experiment	1	2	3	4	5	6
Iron level	120	122	130	135	135	140
Predicted iron level	120.04	124.15	128.26	132.37	136.48	140.59
Residuals	−0.04	−2.15	1.74	2.63	−1.48	−0.59

c.



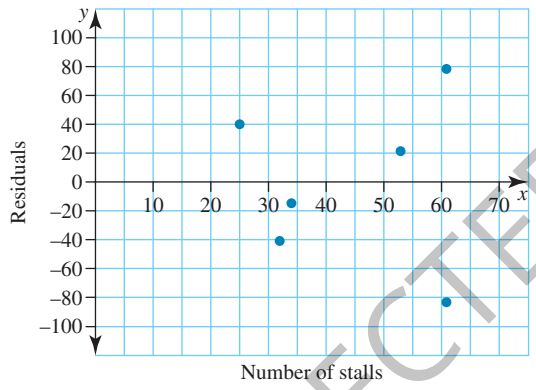
The residuals appear to be scattered randomly about the horizontal axis, so this suggests that the linear model is a suitable model for this data.

d. $r^2 = 0.9454$,

e. The iron level reaches 155 g/L by week 9.5, so according to this model the iron level will reach 155 g/L in week 10. A study of the initial scatterplot of the data suggests that the data shows a linear trend. The residual plot shows no pattern about the horizontal axis and Pearson's product–moment correlation coefficient of 0.9723 indicates a strong positive linear correlation, so the model is suitable to make predictions. However, the iron level of 155 g/L goes outside the range of the given data, so this needs to be considered when making the prediction.

8. a. * See the table at the bottom of the page.

b.



The residuals are scattered randomly above and below the horizontal axis, so this suggests that the linear model is suitable.

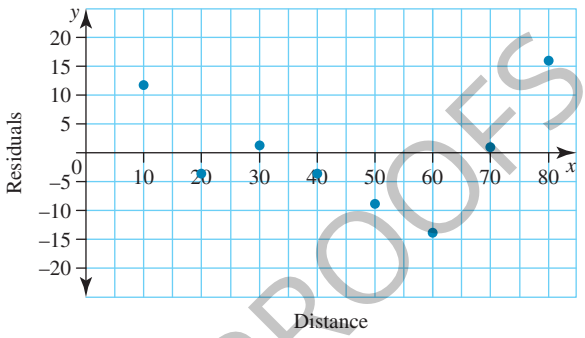
c. $r = 0.88$

d. Yes

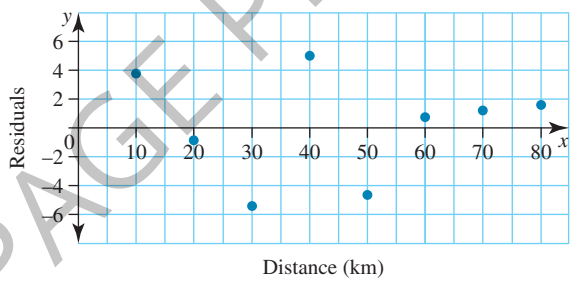
9. a. Cost for factory 1 = 1.51 × Distance from Brisbane + 43.21
Cost for factory 2 = 0.96 × Distance from Brisbane + 56.61
Refer to the worked solutions for the scatterplot.

b. ** See the table at the bottom of the page.

c. Factory 1



Factory 2



The residual plots for both factories appear to be randomly distributed about the horizontal axes, so the linear model is suitable for predictions.

d. Factory 1
$r^2 = 0.9332$, $r = 0.9660$
Factory 2
$r^2 = 0.9763$, $r = 0.9881$

---

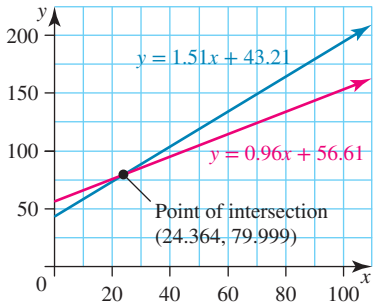
*8. a.

Number of stalls	53	34	61	32	61	25
Number of visitors	501	339	611	300	450	333
Predicted number of visitors	480	354	533	341	533	294
Residuals	21	−15	78	−41	−83	39

**9. b.

Distance from Brisbane (km)	10	20	30	40	50	60	70	80
Factory 1 cost ($)	70	70	90	100	110	120	150	180
Predicted cost for factory 1	58.31	73.41	88.51	103.61	118.71	133.81	148.91	164.01
Residuals for factory 1	11.69	−3.41	1.49	−3.61	−8.71	−13.81	1.09	15.99
Factory 2 cost ($)	70	75	80	100	100	115	125	135
Predicted cost for factory 2	66.21	75.81	85.41	95.01	104.61	114.21	123.81	133.41
Residuals for factory 2	3.79	−0.81	−5.41	4.99	−4.61	0.79	1.19	1.59

e.



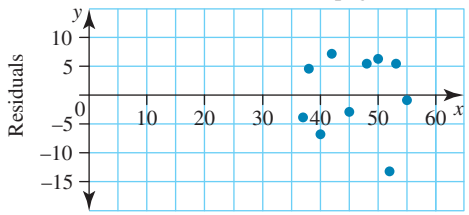
Using the above linear graphs, based on the regression models, if the shop is in Brisbane it is more cost effective to have the surfboards delivered from factory 1.

f. If the shop is more than 24.3 km from Brisbane, as in the case of Mytown, then it is more cost effective to get the surfboards delivered from factory 2.

10. a. $SBP = 74.35 + 1.61 \times Age$

*See the table at the bottom of the page.

b.



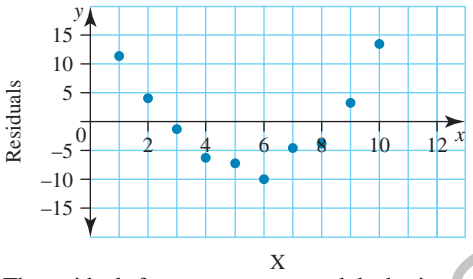
The dots are scattered randomly around the horizontal axis, so the linear model is suitable for this data.

c.
$$\begin{aligned} SBB &= 74.35 + 1.61 \times Age \\ &= 74.35 + 1.61 \times 75 \\ &= 195.10 \end{aligned}$$

Extrapolating outside the range of data given gives a predicted systolic blood pressure for someone aged 75, which could not be considered reliable.

11. a. **See the table at the bottom of the page.

b.



The residuals form a pattern around the horizontal axis, therefore the linear model is not a good model for this set of data.

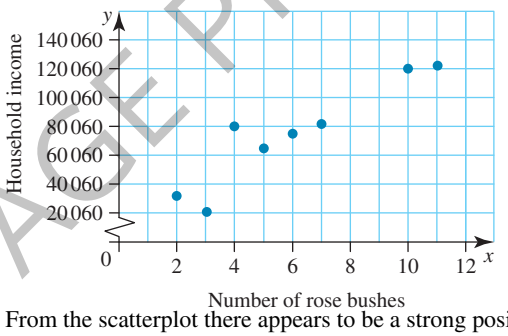
12. A

13. D

14. 78% of the variation in the **response** variable can be explained by the variation in the **explanatory** variable. **22**% of the variation can be explained by other factors.

Exercise 2.5 Association and causation

1. a.



b. From the scatterplot there appears to be a strong positive association between the number of rose bushes in a home garden and the annual household income.

c. $r = 0.93$

d. Just because $r = 0.93$, it cannot be stated that increasing the number of rose bushes in a home garden will increase the annual household income. Other factors may need to be considered. Further investigation is needed before any conclusions can be drawn.

---

*10. a.

Age	37	38	40	42	45	48	50	52	53	55
Systolic blood pressure	130	140	132	149	144	157	161	145	165	162
Predicted systolic blood pressure	133.92	135.53	138.75	141.97	146.8	151.63	154.85	158.07	159.68	162.9
Residuals	−3.92	4.47	−6.75	7.03	−2.8	5.37	6.15	−13.07	5.32	−0.9

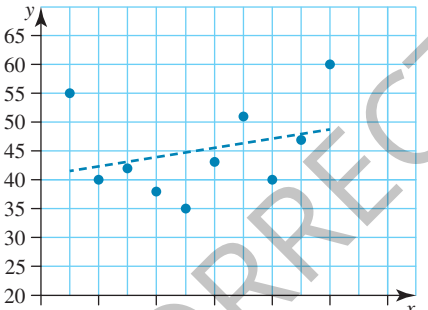
** 11. a.

x	1	2	3	4	5	6	7	8	9	10
y	1	5	10	16	26	34	50	62	80	101
Predicted	−10.454	0.425	11.304	22.183	33.062	43.941	54.82	65.699	76.578	87.457
Residuals	11.454	4.575	−1.304	−6.183	−7.062	−9.941	−4.82	−3.699	3.422	13.543

2. There is a negative association of −0.86 between the amount of time spent playing computer games and the level of aerobic fitness.
3. D
4. D
5. B
6. More fire fighters are called to larger fires, so the damage is due to the size of the fire.
7. Larger shoe sizes are associated with age and so is greater academic knowledge. Age is the common cause.
8. It cannot be said that the consumption of the supplement causes improvement in health. Upon taking the supplement, the students may have also implemented other lifestyle changes, such as giving up smoking, lower alcohol consumption, eating a healthier diet or exercising more. There is no information about the gender or the age of the participants which also may have had an impact on the health improvements.

2.6 Review: exam practice

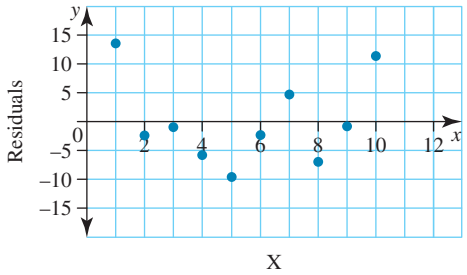
1. B
2. A
3. C
4. B
5. C
6. D
7. B
8. D
9. C
10. D
11. D
12. A
13. a.



Equation of least-squares regression line

$y = 0.7939x + 40.733$

b. $r^2 = 0.0901$
9% of the variation in y can be explained by the variation in x.
c. $r = 0.30$
There is a weak positive linear association between the x and y variables.
d. *See the table at the bottom of the page.
e.



X

The residuals do not appear to be randomly scattered about the horizontal axis. The first point on the left is at about 13, then the points fluctuate above and below the axis finishing at about 11 on the right. Given the value of r is 0.30, indicating a weak association, then it is not appropriate to assume that the relationship between these variables is linear.

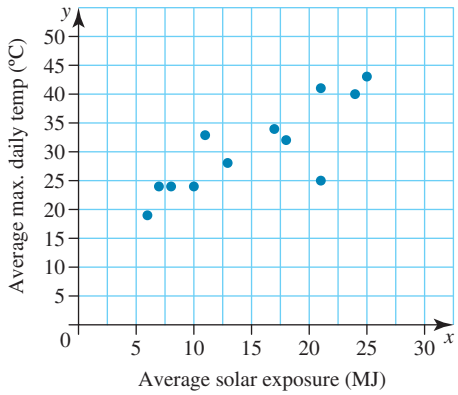
14. a. Time is the explanatory variable and amount of water the response variable.
b. Number of students is the explanatory variable and number of buses is the response variable.
15. a. The data shows a strong, negative linear relationship. The correlation coefficient of −0.93 also indicates a strong, negative linear relationship.
b. $y = 25.85 - 2.10x$
c. 4.21, interpolation
16. a. 0.57
b. A moderate, negative, linear association
17. Although there appears to be a link between the laziness of people and the increase in sales of healthy foods, there are also many other possible factors besides laziness; for example, people are very time poor, unsure of what constitutes healthy food, and are lacking confidence and the skills to cook for themselves. Based on this observation alone, the cause of an increase in sales of healthy foods cannot be concluded to be due to laziness.
18. a. Fuel consumption
b. Fuel consumption = 0.1119 × Capacity + 0.6968
c. 5.17 L/100 km
d. This is an example of interpolation as 40L capacity is within the range of data given.
e. 85 L
f. This value indicates a moderate relationship between the variables. Therefore, the data can be used, but other factors should also be considered.
g. Fuel consumption can be influenced by the size of the engine, the size of the vehicle, the type of driving (city, country) and the age of the vehicle.
19. a. Weight (grams)
b. No correlation
c. This supports the view that there is no correlation between the variables. Based on this value, no conclusions can be made from the data.
d. Various answers are possible, e.g. popularity of the shoe or desired profits.

*13. d.

x	1	2	3	4	5	6	7	8	9	10
y	55	40	42	38	35	43	51	40	47	60
Predicted y value	41.53	42.32	43.11	43.91	44.70	45.50	46.29	47.08	47.88	48.67
Residuals	13.47	−2.32	−1.11	−5.91	−9.70	−2.50	4.71	−7.08	−0.88	11.33

20. a. Explanatory variable = average solar exposure
Response variable = maximum daily temperature

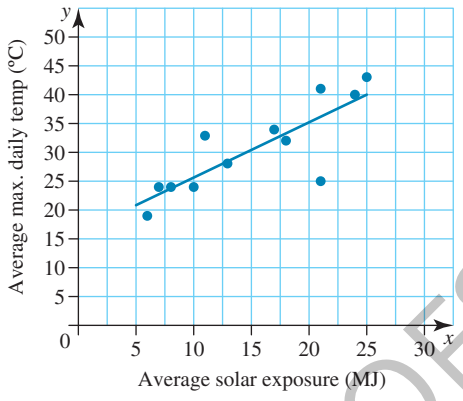
b.



c. Strong positive correlation

d. $r = 0.8242$; $r^2 = 0.6793$
These values indicate a strong relationship between the two variables. The coefficient of determination suggests that nearly 70% of the maximum daily temperature is due to the amount of solar exposure.

e.



Maximum daily temperature = 16.232 + 0.9515 × average solar exposure

f. 22 MJ

g. 19°C

h. An average solar exposure of 3 MJ is outside the original data set.