TOPIC 8
Bivariate data analysis

8.1 Overview

8.1.1 Introduction

Numerical data is everywhere around us. Have you ever considered what relationships comparing two sets of numerical data could present?

Every day when you use a mobile device, you are providing companies with some data: the number of times you access an application, the location you access the application and how you use the application. How would the company graph the data? Would there be a pattern? What could the data be used to predict?

Both the access and use of digital data raises many issues to do with privacy and ethics — issues that are becoming increasingly important in the modern world.

DISCUSSION

What does your mobile device say about you? Consider how many apps you have given permission to access your data, and what types of data the apps collect about you? What are two sets of data from your mobile device that a company would use for comparison?

LEARNING SEQUENCE

8.1 Overview
8.2 Scatterplots
8.3 Lines of best fit
8.4 The statistical investigation process
8.5 Review

Fully worked solutions are available for this topic in the Resources section of your eBookPLUS.

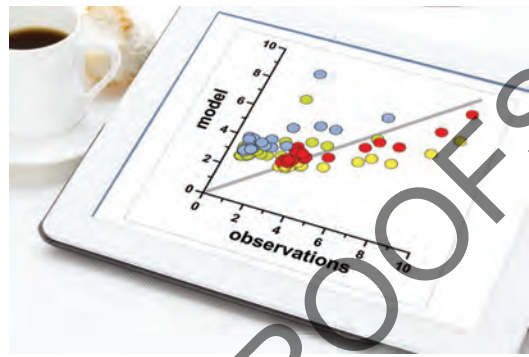
CURRICULUM CONTENT

Students:

- construct a bivariate scatterplot to identify patterns in the data that suggest the presence of an association (ACMGM052) **AAM**
- use bivariate scatterplots (constructing them when needed) to describe the patterns, features and associations of bivariate datasets, justifying any conclusions **AAM**
- model a linear relationship by fitting an appropriate line of best fit to a scatterplot and using it to describe and quantify associations **AAM**
- use the appropriate line of best fit, both found by eye and by applying the equation, to make predictions by either interpolation or extrapolation
- implement the statistical investigation process to answer questions that involve identifying, analysing and describing associations between two numerical variables **AAM**
- construct, interpret and analyse scatterplots for bivariate numerical data in practical contexts **AAM**

8.2 Scatterplots

8.2.1 Bivariate data



- Often when we look at a situation we are trying to assess how much one variable has caused or influenced another to create the end result. **Bivariate data** is the term used for information relating to two different variables.
- When exploring bivariate data, it is necessary to identify which of the two variables is the **independent variable** (represented on the *x*-axis) and which is the **dependent variable** (represented on the *y*-axis).
- The dependent variable is the variable whose value depends on the other variable.
- The independent variable takes on values that do not depend on the value of the other variable.

A set of bivariate data involves two variables where one affects the other. The independent variable is a factor that influences the dependent variable.

For example, we are given the length of a pair of pants (variable 1) and the age of a person (variable 2). In this example, the age of a person is not going to depend on the length of their pants, while the length of the pants will generally be explained by the age of the person. Therefore, the independent variable is the age, while the length of the pants is the dependent variable.

WORKED EXAMPLE 1

Identify the independent and dependent variables in each of the following scenarios.

a. Distance walked in an hour and the age of a person

b. The cost of bananas and the average daily temperature in Queensland

THINK

a. Consider which variable does not respond to the other. The age of a person will not be changed due to the distance they walk; however, their age could explain the distance they have covered.

b. The cost of bananas is influenced by supply and demand. If the growing season has been affected by higher than expected daily temperatures, the number of bananas produced will be less, therefore increasing the price.

WRITE

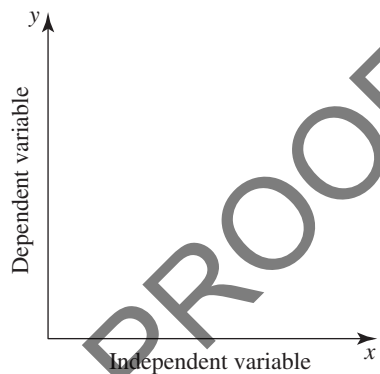
a. Independent variable = age
Dependent variable = distance walked in an hour

b. Independent variable = average daily temperature in Queensland
Dependent variable = cost of bananas

8.2.2 Scatterplots

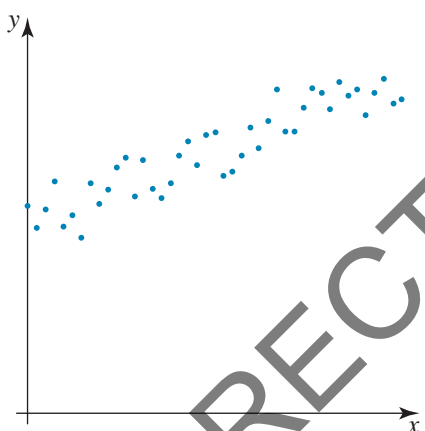
- A common way to interpret bivariate data is through the use of a **scatterplot**.
- Scatterplots provide a visual display of the data and can be used to draw **correlations** between two variables.
- Each data value on the scatterplot is shown by a point on a Cartesian plane.

When constructing a scatterplot, it is important to place the independent variable along the *x*-axis and the dependent variable along the *y*-axis.

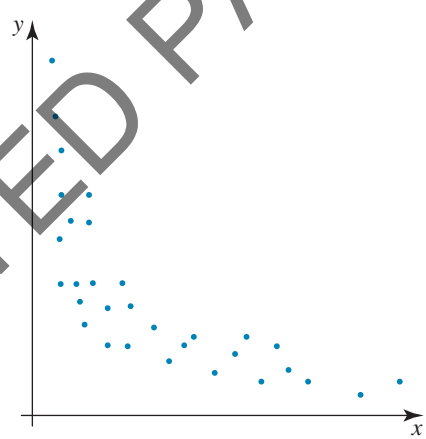


- When we are describing the relationships between two variables displayed on a scatterplot, we need to comment on:
  - the form — whether it is linear or non-linear
  - the direction — whether it is positive or negative
  - the strength — whether it is strong, moderate or weak
  - possible outliers.
- The **form** of a bivariate dataset can be categorised as either **linear** or **non-linear**. The points on the scatterplot will produce a linear pattern or a non-linear pattern.
- In a linear pattern, the points tend to form a straight line, whereas in a non-linear pattern, the points tend to from a curve.

Linear form



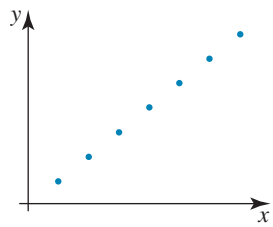
Non-linear form



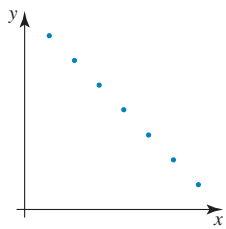
- Linear bivariate datasets can be further described in terms of **direction** and **strength**.
  - Direction describes whether the plot slopes up or down.
  - Strength describes how closely the points form a linear pattern.
- For a linear bivariate dataset:
  - If the points on the scatterplot slope up to the right, we say that there is a **positive correlation**; that is, a positive relationship between the variables. This shows that as the independent variable increases, the dependent variable also increases.
  - If the scatterplot points slope down to the right, we say that there is a **negative correlation**, or a negative association between the two variables. This shows that as the independent variable increases, there is a decrease in the dependent variable.

- When interpreting a scatterplot, the correlation provides an insight into the relationship between the two variables. The correlation is a measure of the strength of the linear relationship between the two variables. There are three classifications for the correlation of data:

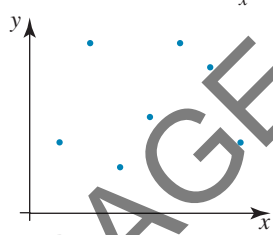
1. **Positive correlation:** as the independent variables (x-axis) increase, the dependent variables (y-axis) also increase, forming an upwards (positive) trend.



2. **Negative correlation:** as the independent variables (x-axis) increase, the dependent variables (y-axis) decrease, forming a downwards (negative) trend.

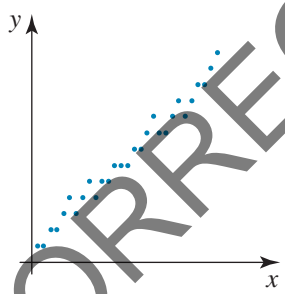


3. **No correlation:** no visible pattern formed by the data points, which appear to be randomly placed.

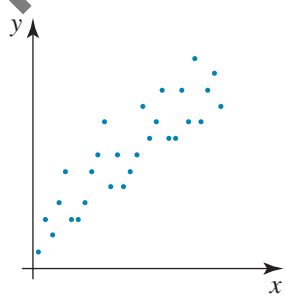


- To measure how strongly the scatterplot points tend to form a straight line, we begin by estimating the strength of the association as strong, moderate or weak based on inspection of the scatterplot. We then calculate and use **Pearson's correlation coefficient, r**, to quantify the strength of a linear association. Pearson's correlation coefficient measures the degree to which the points in a scatterplot tend to cluster around a straight line.
- Here is a gallery of scatterplots showing various patterns to look for.

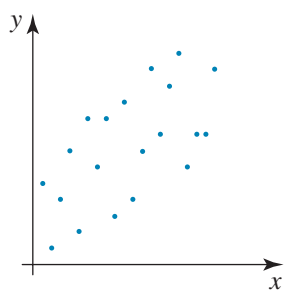
Strong positive correlation



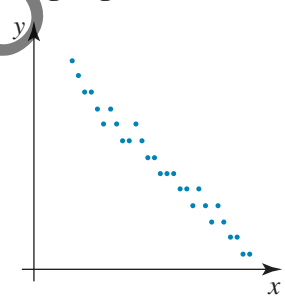
Moderate positive correlation



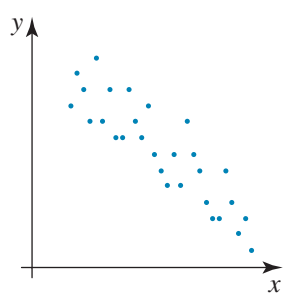
Weak positive correlation



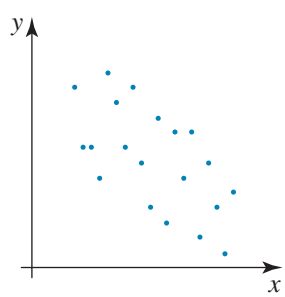
Strong negative correlation



Moderate negative correlation



Weak negative correlation



WORKED EXAMPLE 2

A local café recorded the number of ice-creams sold per day as well as the daily maximum temperature for 12 days.

Temp (°C)	36	32	28	26	30	24	19	25	33	35	37	34
No. of ice-creams sold	162	136	122	118	134	121	65	124	140	154	156	148

a. Identify the dependent and independent variables.
b. Represent the data in a scatterplot.
c. Discuss the strength of the relationship between the variables.

THINK

a. Consider which variable does not rely on the other. This will be the independent variable. Temperature does not rely on the number of ice-creams sold.

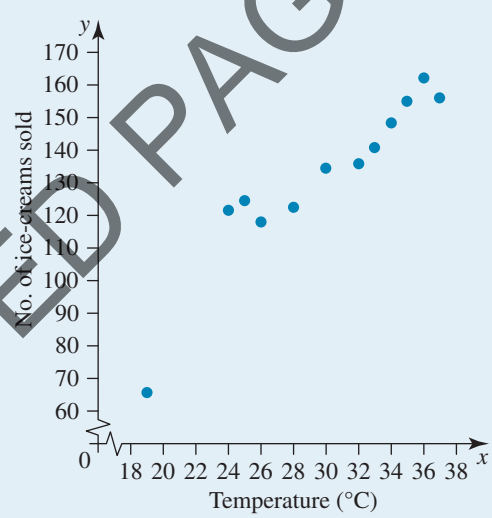
b. Select a reasonable scale for each variable that covers the full range of the data set. Plot the given points, remembering that the independent variable should be represented on the x-axis and the dependent variable should be represented on the y-axis.

c. Look at the pattern of the data points. Do they form a linear pattern? Are they progressing in a similar direction, either positive or negative? How strong is the correlation between the variables?

WRITE

a. Independent variable = temperature
Dependent variable = number of ice-creams sold

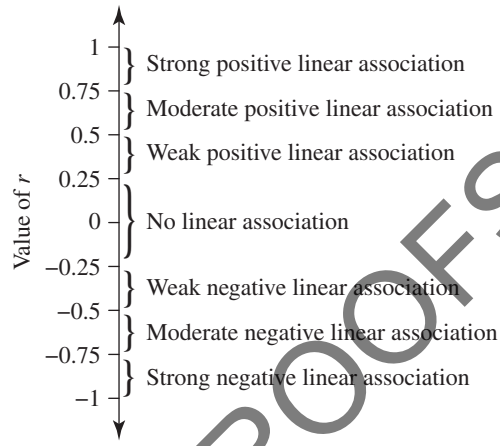
b.



c. There is a linear relationship between the two variables. As the temperature increases so does the number of ice-creams sold. The correlation between the variables is strong. Therefore, this graph could be described as having a strong positive correlation.

8.2.3 Pearson's correlation coefficient

- The strength of the linear relationship can be observed from a scatterplot of the data. However, to determine exactly how strong this relationship is we can use Pearson's correlation coefficient, r, which measures the strength of a linear trend and associates it with a numerical value between -1 and $+1$. A value of either -1 or $+1$ indicates a perfect linear correlation, while a result closer to zero indicates no correlation between the variables. The scale shown is a guide when using r to describe the strength of a linear relationship.



- Pearson's correlation coefficient is calculated using the formula shown below. However, you are not required to use this formula as part of this course. Instead, you will use technology to find values of r, as explained in the examples that follow.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where:
n is the numbers of pieces of data in the data set
x_i is an x-value (independent variable)
y_i is a y-value (dependent variable)
s_x is the standard deviation of the x-values
s_y is the standard deviation of the y-values
$\bar{x}$ is the mean of the x-values
$\bar{y}$ is the mean of the y-values.

WORKED EXAMPLE 3

Below is a list of National Rugby League Holden Cup Player Statistics for 2017.

Player	Team	Games played	Points (P)	Tries (T)	Goals (G)	Field goals (FG)	Minutes played
Kyle Flanagan	Sharks	24	344	19	134	0	1920
Dean Blore	Panthers	25	225	11	90	1	1991
Gerome Burns	Broncos	23	196	14	70	0	1702
Jake Clifford	Cowboys	21	188	9	76	0	1548
Zac Lomax	Dragons	20	186	8	77	0	1582
Jesse Arthars	Storm	21	184	13	66	0	1640
Sean O'Sullivan	Roosters	18	172	6	74	0	1434

Player	Team	Games played	Points (P)	Tries (T)	Goals (G)	Field goals (FG)	Minutes played
Ethan Roberts	Titans	22	152	5	66	0	1723
Jade Anderson	Sea Eagles	17	140	14	42	0	1302
Andre Niko	Raiders	23	128	6	52	0	1838
Dean Hawkins	Rabbitohs	17	122	7	47	0	1360
Adam Keighran	Bulldogs	19	118	9	41	0	1496
James Tautailefua	Wests Tigers	17	116	6	46	0	1348
Jesse Marschke	Roosters	23	116	15	28	0	1672
Adam Doueihi	Rabbitohs	12	116	4	50	0	950
Kainoa Gudgeon	Knights	16	112	7	42	0	1280
Michael Carroll	Cowboys	23	98	18	13	0	1815
Jackson Willis	Dragons	25	97	15	18	1	1872
Tevita Funa	Sea Eagles	10	94	10	27	0	769
Sione Katoa	Sharks	24	92	23	0	0	1875

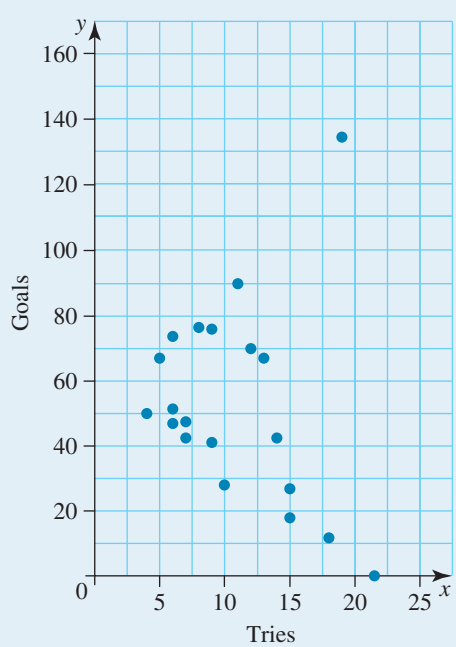
Source: NRL, working link http://www.nrl.com/stats/holdencup/playerstatistics

a. Using technology, draw a scatterplot of the number of tries (T) against the number of goals (G).
b. Describe the form, strength and direction of the scatterplot.
c. Calculate Pearson's correlation coefficient and interpret the strength of the correlation.

THINK

a. T is the independent variable and G is the dependent variable (players get a chance to get a conversion goal after every try).
Plot the table of values using the technology of your choice.

WRITE

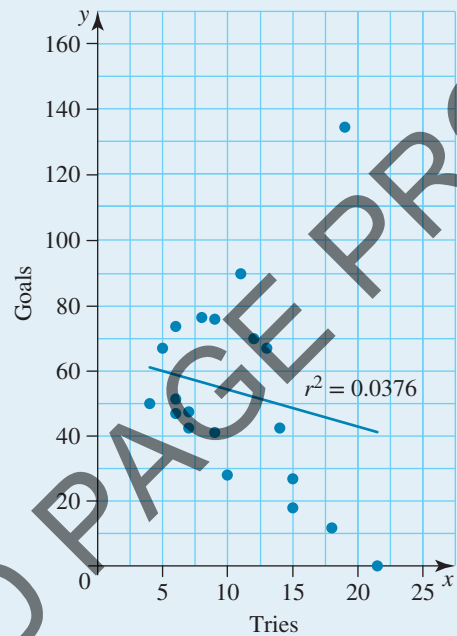


b. Describe the scatterplot in terms of form, strength and direction.

From the scatterplot:
- the form is linear
- the strength is weak
- the direction is negative.

Overall, the scatterplot seems to have a weak negative linear relationship between the number of tries and goals.

c. 1. Some technology will show a value for r^2 (the coefficient of determination) when a line of best fit is applied. Pearson's correlation coefficient (r) is the square root of this number.



$r^2 = 0.0376$

Find the value of r.

Pearson's correlation coefficient (r) is the square root of r^2.

$r^2 = 0.0376$

$r = \pm\sqrt{0.0376}$

$r = \pm 0.194$

Since the direction is negative, $r = -0.194$.

2. Write the answer.

The value of r is between 0 and -0.25, which means that there is no linear relationship between the number of tries and the number of goals.

Exercise 8.2 Scatterplots

Understanding, fluency and communicating

1. **WE 1** A survey was conducted to record how long it takes to eat a pizza and the time of day. Identify the independent and dependent variables.

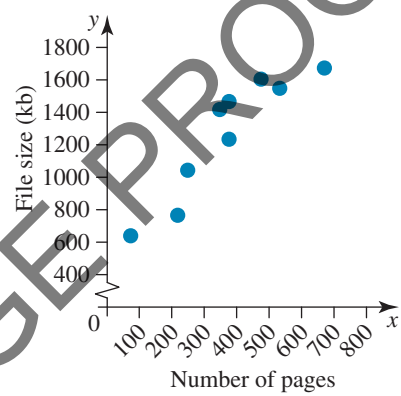
2. A study recorded the amount of data needed on a phone plan and the time spent using phone apps. Identify the independent and dependent variables

3. **WE 2** Consider the data in the following table.

Time (minutes)	2	4	6	8	10	12	14	16	18	20
Weight that can be held (kg)	55	52	46	33	28	25	19	20	17	12

   a. Identify the dependent and independent variables.
   b. Represent the data in a scatterplot.
   c. Identify the type of correlation, if any, that is evident from the scatterplot of these two variables.

4. The scatterplot shown has been established.
   a. Which variable is the dependent variable?
   b. How would you describe the relationship between these variables?



5. **WE 3** The following table outlines the cost of an annual magazine subscription and the number of magazines issues per year.

No. of magazine issues per year	7	9	10	6	8	4	4	5	11	9	10	5	11	3	7	12	7	6	12
Subscription cost ($)	34	40	52	38	50	25	28	40	55	55	45	28	65	24	38	55	50	33	59

   a. Using technology, draw a scatterplot of the number of magazines (n) per year against the subscription cost (s).
   b. Describe the form, strength and direction of the scatterplot.
   c. Calculate Pearson's correlation coefficient, correct to 4 decimal places, and interpret the strength of the correlation.

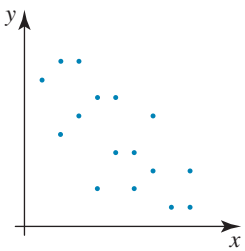
6. For each of the following scenarios, identify the independent and dependent variable.
   a. The age of people (in years) and the number of star jumps they can complete in one minute
   b. The cost of purchasing chocolate and various quantities of chocolate
   c. The number of songs stored on a media player and the memory capacity used
   d. The growth rate of bacterial cells in a laboratory and the quantity of food supplied

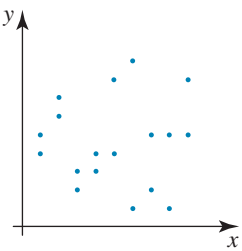
7. The weights and heights of a random sample of people were collected, with the following table displaying the collected data.

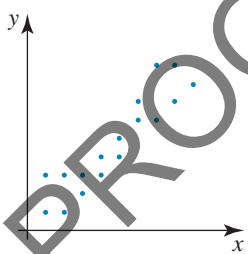
Height (cm)	140	145	150	155	160	165	170	175	180	185
Weight (kg)	58	62	66	70	75	77	78	80	88	90

   a. Identify the independent and dependent variables.
   b. Using a reasonable scale, plot the data.

8. Comment on the type and strength of the correlation displayed in each of the following scatterplots:

   a. 

   b. 

   c. 

9. Suggest a combination of keep an independent variable and a dependent variable that may produce each of the following correlation trends:
   a. negative correlation
   b. no correlation.

10. Use your understanding of Pearson's correlation coefficient to explain what the following results indicate.
   a. $r = 0.68$
   b. $r = -0.97$
   c. $r = -0.1$
   d. $r = 0.30$

11. A survey asked random people for their house number and the combined age of the household members. The following data was collected:

House no.	Total age of household
14	157
65	23
73	77
58	165
130	135
95	110
54	94
122	25
36	68

House no.	Total age of household
101	53
57	64
34	120
120	180
159	32
148	48
22	84
9	69

   a. Using the house number as the independent variable, plot this data.
   b. Comment on the resulting scatterplot.
   c. Determine Pearson's correlation coefficient, correct to 4 decimal places.
   d. Interpret this value in the context of the data.

Problem solving, reasoning and justification

12. A class of Year 12 students were asked to record the amount of time in hours that they spent on a History assignment and the mark out of 100 that they received for the assignment.

Time spent (hours)	Mark(%)
2	72
0.5	52
1.5	76
2.5	82
0.25	36
2	73
2.5	84
2.5	80
2	74
0.5	48

Time spent (hours)	Mark (%)
0.75	58
1.5	69
1	62
2	78
3	90
3.5	94
1	70
3	92
2.5	88
3	97

a. Identify the independent and dependent variables.
b. Draw a scatterplot to represent this data.
c. Comment on the direction and correlation of the data points.
d. Explain why the data is not perfectly linear.
e. Calculate Pearson's correlation coefficient, correct to 3 decimal places.
f. What does the value from part **e** suggest about the relationship between a student's assignment mark and the time they spent on the assignment?

8.3 Lines of best fit

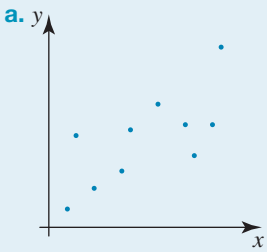
8.3.1 Lines of best fit by eye

- After raw data has been plotted, the scatterplot can be used to determine findings and predictions can be made.
- If the points on a scatterplot appear to lie fairly closely 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 so that it is as close as possible to all the data points, that is, the average distance between the data points and the line is minimised.
- A line of best fit is the straight line that is most representative of the data. It is used to generalise the relationship between two variables.
- There are numerous ways to draw a line of best fit, some more accurate than others.
- A line of best fit can be drawn on a scatterplot **by eye**. This method aims to draw a straight line with approximately the same number of data points above and below the line. The line should follow the direction of the general trend of the data. This method, while quick, leaves significant room for error.

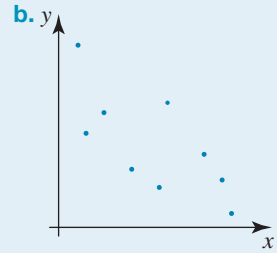
WORKED EXAMPLE 4

For each of the scatterplots below, use a ruler to draw a line of best fit by eye.

a.



b.



THINK

a. 1. Count the number of data points.

2. Consider the direction of the data points.

3. Draw a straight line through the data points using a pencil. Review the line to confirm an even distribution of data points.

b. 1. Count the number of data points.

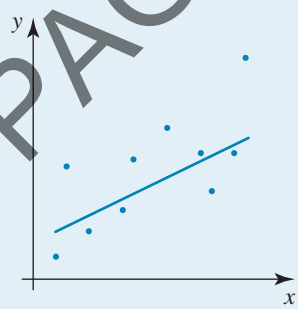
2. Consider the direction of the data points.

3. Draw a straight line through the data points using a pencil. Review the line to confirm an even distribution of data points.

WRITE

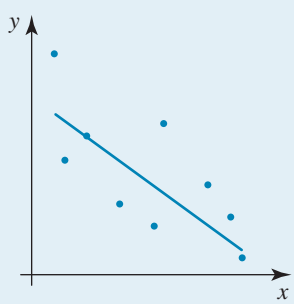
a. In this example there are 10 data points; therefore, the line of best fit should have 5 points on each side.

In this example, the trend is positive. The line of best fit will follow this trend.



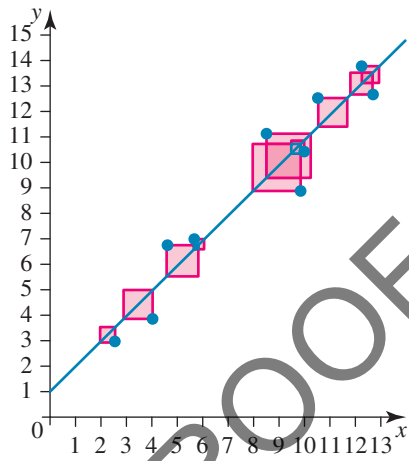
b. In this example there are 9 data points. As there is an odd number of points, either a data point sits on the line or there will be a slight imbalance of points above and below the line.

In this example, the trend is negative. The line of best fit will follow this trend.



8.3.2 Least squares regression

- Sometimes a line of best fit can be drawn by eye; however, in other situations it is necessary to be more accurate. When there are no outliers in a scatterplot, we can generate an equation using the **least squares regression line**.
- Least squares regression involves an exact mathematical approach to fitting a line of best fit to bivariate data that show a strong 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 of these squares.



The equation for the least squares regression line takes the form:

$$y = a + bx,$$

where y is the dependent variable, x is the independent variable, b is the gradient or slope of the line and a is the y-intercept.

- We can use technology 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 independent variable (x-variable)
  $\bar{y}$ the mean of the dependent variable (y-variable)
  s_x the standard deviation of the independent variable
  s_y the standard deviation of the dependent variable
  r Pearson's product–moment correlation coefficient.

The general form of the least-squares regression line is:

$$y = a + bx$$

where:

the slope of the regression line is $b = r\dfrac{s_y}{s_x}$

the y-intercept of the regression line is $a = \bar{y} - b\bar{x}$.

WORKED EXAMPLE 5

A study to find relationship between the height of husbands and the height of their wives revealed the following details.

- **Mean height of the husbands: 180 cm**
- **Mean height of the wives: 169 cm**
- **Standard deviation of the height of the husband: 5.3 cm**
- **Standard deviation of the height of the wives: 4.8 cm**
- **Correction coefficient, $r = 0.85$**

The form of the least-squares regression line is to be:

$$\text{Height of wife} = b \times \text{height of husband} + a$$

a. Which variable is the dependent 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 wife whose husband is 195 cm tall (to the nearest cm).

THINK	WRITE
a. Recall that the dependent variable is the subject of the equation in $y = bx + a$ form; that is, y.	**a.** The dependent variable is the height of the wife.
b. 1. The value of b is gradient of the regression line. Write the formula and state the required values.	**b.** $r = 0.85$, $s_y = 4.8$ and $s_x = 5.3$ $b = r\frac{s_y}{s_x}$ $a = 0.85 \times \frac{4.8}{5.3}$ $= 0.7698$ $= 0.77$
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.	**c.** $\bar{y} = 169$, $\bar{x} = 180$ and $b = 0.7698$ (from part b) $a = \bar{y} - b\bar{x}$ $a = 169 - 0.7698 \times 180$ $= 30.436$ $= 30.44$
2. Substitute the values into the formula and evaluate a.	
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 wife and x represents the height of the husband.	**d.** $y = 0.77x + 30.44$
2. The height of the husband 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 cm.	Using the equation of the regression line found, the wife's height would be 181 cm.

8.3.3 Interpreting the intercept and gradient

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

WORKED EXAMPLE 6

The table below shows Ramsay Real Estate's data for house sales in Quakers Hill in November 2017.

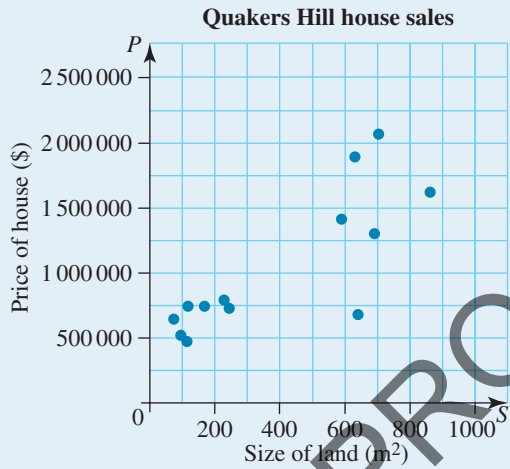
House	Number of bedrooms	Number of bathrooms	Size of garage (cars)	Size of land (m^2)	Price ($)
1	2	1	1	117	730 000
2	4	2	1	630	1 875 000
3	3	1	2	688	1300 000
4	2	1	1	228	790 000
5	3	1	2	858	1 610 000
6	2	1	1	637	670 000
7	3	1	1	588	1400 000
8	6	4	1	700	2 060 000
9	2	1	1	93	520 000
10	2	1	1	73	639 000
11	3	1	1	242	720 000
12	1	1	1	112	460 000
13	2	1	1	167	737 000

a. Using 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 Quakers Hill?

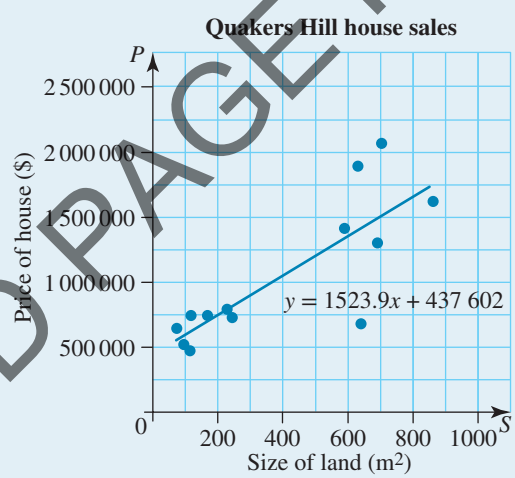
THINK

a. Size of land (S) is the independent variable.
Price of house (P) is the dependent variable.
Plot the table of values using technology of your choice.

WRITE



b. Determine the regression line using your chosen technology.



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. Interpret the least squares regression line by referring to the gradient and the y-intercept.

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

WORKED EXAMPLE 7

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 independent variable, by how much does the dependent 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.	y-intercept $= 62$
b. The change in the dependent variable due to the independent variable is reflected in the slope. Identify the b value in the equation.	$b = -8$
c. A positive b value indicates a positive trending line, while a negative b value indicates a negative trending line.	As the b value is negative, the trend of the line is negative.

8.3.4 Interpolation and extrapolation

- The regression line can be used to explore data points both inside and outside of the 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 prediction. Prediction 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 independent variable (x) or the dependent variable (y) in order to find the value of the other variable.

WORKED EXAMPLE 8

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 independent 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 $y = a + bx$. The variable on the right-hand side of the equation will be the independent variable.	**a.** Independent variable = flower diameter (d)
b. 1. Using the equation, substitute 15 in place of 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.	29 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.	**d.** $35 = 0.41 + 1.88d$
2. Transpose the equation to solve for d.	$d = \frac{35 - 0.41}{1.88}$ $= 18.40$
3. Round to 1 decimal place.	$= 18.4$ (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.

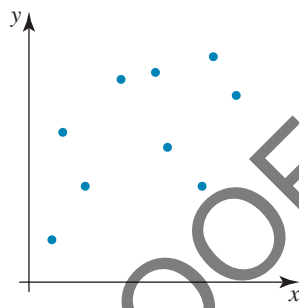
8.3.5 Limitations of regression line predictions

- 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 correlation between the variables also provides an indication of the reliability of the data. Data that produces no correlation or a low correlation 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.

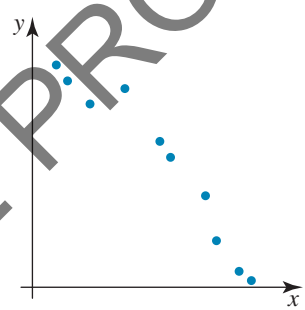
Exercise 8.3 Lines of best fit

Understanding, fluency and communicating

1. **WE 4** For the scatterplot below, use a ruler to draw a line of best fit by eye.

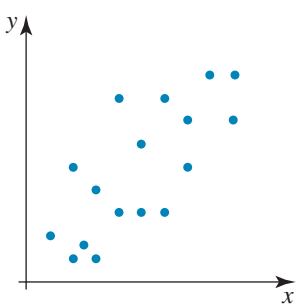


2. For the scatterplot below, use a ruler to draw a line of best fit by eye.

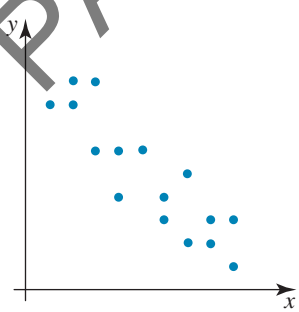


3. For each of the scatterplots, draw a line of best fit by eye.

a.



b.



4. For each of scatterplots in question 3, comment on the type and strength of the correlation displayed.

5. 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%
- Correction coefficient, $r = 0.64$

The form of the least-squares regression line is to be:

$$\text{Mathematics exam mark} = m \times \text{English exam mark} + c$$

a. Which variable is the dependent 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 score 85% in an English exam (to the nearest percentage).

6. 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. $\bar{x} = 10$ $s_x = 1$ $\bar{y} = 20$ $s_y = 2$ $r = -0.5$

7. **WE 5 + 6** 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

Height of mother (cm)	Height of son (cm)
158	159
156	150
168	172
169	175
179	180
173	190

a. Which variable is the dependent variable?
b. Draw a scatterplot and a line of best fit.
c. 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.

8. **WE 7** 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 independent variable, by how much does the dependent variable change?
c. What does your answer to part **b** tell you about the direction of the line?

9. 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 independent variable, by how much does the dependent variable change?
c. What does your answer to part **b** tell you about the direction of the line?

10. **WE 8** 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 independent 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.

11. 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 dependent 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

12. An equation for a regression line is $y = 3.2 - 1.56x$. What conclusions about the trend of the regression line can be determined from the equation?

13. Data on the daily sales of gumboots and the maximum daily temperature were collected.

Temp (°C)	Daily sales (no. of pairs)
17	2
16	3
12	8
10	16
14	7
17	3
18	2
22	1

Temp (°C)	Daily sales (no. of pairs)
23	1
19	2
17	3
15	3
12	12
15	9
20	1

a. Draw a scatterplot of this data.

b. Find the equation of the line of best fit, expressed in terms of temperature (T) and daily sales (D). Give values correct to 4 significant figures.

c. Calculate Pearson's correlation coefficient, correct to 4 significant figures.

d. Interpret these values in the context of the data.

14. 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?

15. Answer the following questions for the equation $y = 60 - 5x$.

a. Identify the y-intercept.

b. For each unit of change in the independent variable, by how much does the dependent variable change?

c. Is the trend of the data positive or negative?

d. Calculate the value of y when $x = 40$.

16. 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:

$$y = -12.9 + 7.32x$$
$$15.68 = -12.9 + 7.32x$$
$$x = 12.9 + \frac{15.68}{7.32}$$
$$= 15.04$$

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.

17. Answer the following questions for the equation $y = -12 + 25x$.

a. Identify the y-intercept.

b. For each unit of change in the independent variable, by how much does the dependent variable change?

c. Is the trend of the data positive or negative?

d. Calculate the value of y when $x = 3.5$.

18. 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 dependent variable.
- **b.** For each unit of change in the independent variable, by how much does the dependent variable change?
- **c.** Is the trend 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 $60\ cm^2$. Give your answer correct to the nearest whole number.

19. a. Use the data given to draw a scatterplot and a line of best fit by eye.

x	1	2	3	4	5	6	7	8	9	10
y	35.3	35.9	35.7	36.2	37.3	38.6	38.4	39.1	40.0	41.1

- **b.** Find the equation of the line of best fit and use it to predict the value of y when $x = 15$. Give your answers correct to 4 significant figures.

20. Use the data given below to complete the following questions.

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

- **a.** Draw a scatterplot and a line of best fit by eye.
- **b.** Determine the equation of the line of best fit. Give values correct to 2 significant figures.
- **c.** Predict the value of y when $x = 20$.
- **d.** Predict the value of x when $y = 9$. Give your answer correct to 2 decimal places.

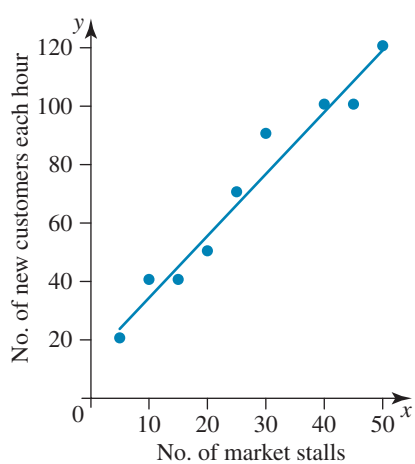
Problem solving, reasoning and justification

21. A data set produced a positive trend and for each incremental increase in the independent variable, the dependent variable increased by 2.5. If $y = 4$ when $x = 0$, determine the equation for the regression line

22. Use technology to design a data set that meets the following criteria:

- contains 10–15 data points
- produces a positive trend
- has a b value between 2 and 5 in its regression equation.

23. In the scatterplot shown, the line of best fit has the equation of $c = 13.33 + 2.097m$, where c is the number of new customers each hour and m is the number of market stalls.



- **a.** Using the line of best fit, interpolate the data to find the number of new customers expected if there are 30 market stalls.
- **b.** Use the formula to extrapolate the number of market stalls required in order to expect 150 new customers.
- **c.** Explain why part **a** is an example of interpolating data, while part **b** demonstrates extrapolation.

24. Use the data given below to complete the following questions.

x	10	11	12	13	14	15	16	17	18	19
y	22	18	20	15	17	11	11	7	9	8

a. Draw a scatterplot and a line of best fit by eye.
b. Determine the equation of the line of best fit. Give values correct to 4 significant figures.
c. Extrapolate the data to predict the value of y when $x = 23$.
d. What assumptions are made when extrapolating data?

25. While camping a mathematician estimated that:
number of mosquitos around 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.

26. Data on people's average monthly income and the amount of money they spend at restaurants was collected as shown in the table.

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

a. Draw a scatterplot of this data on technology of your choosing.
b. Find the equation of the line of best fit 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?

27. Data on students' marks in Geography and Music were collected.

Geography	Music
65	91
80	57
72	77
61	89
99	51
54	76
39	62
66	87

Geography	Music
78	88
89	64
84	90
73	45
68	60
57	79
60	69

a. Is there an obvious independent variable in this situation?

b. Draw a scatterplot of this data on your calculator, using the marks in Geography as the independent variable.

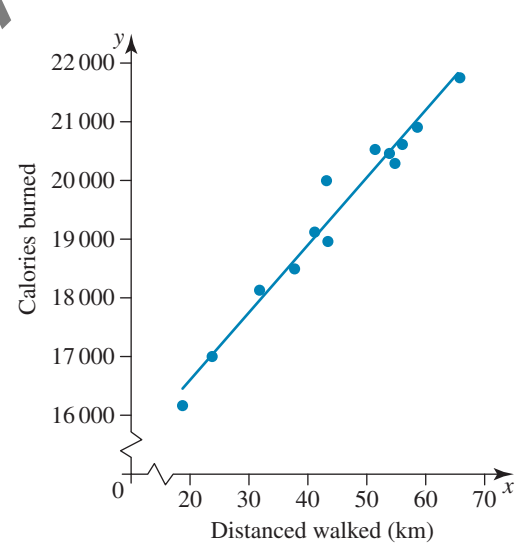
c. Find the equation of the line of best fit. Give values correct to 4 significant figures.

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 dependent.

f. Calculate Pearson's correlation coefficient for this data. How can you use this value to evaluate the reliability of your data?

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



a. Identify the dependent variable in this situation.

b. Rewrite the equation in terms of the independent and dependent 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.

8.4 The statistical investigation process

8.4.1 Privacy and ethics

- Data collected from digital devices are known as **digital data**.
- The use of digital data can raise **privacy** and **ethical** issues.
- Most software packages have terms and conditions that users must agree to before installing the software. However, very few users actually read these terms and conditions. By agreeing to an installation or sign-up policy, users are giving companies complete access to their data via the applications on their devices, without any control on how their data may be used.
- Digital data can be collected from a variety of sources, including mobile phones; fitness monitoring devices; social applications, such as gambling and dating applications; web-based media, such as games and movies; web browsers; and navigational devices. Users' personal digital data can also be obtained from government or commercial sources, such as medical, legal, financial and travel records.
- When researchers collect and analyse digital data, they need to be mindful of their ethical obligations to users. They need to make sure that their reports do not misrepresent the data or show any bias.
- The main ethical issues that researchers must consider are:
  - consent. Have users agreed to the collection and use of their data?
  - privacy. Can the data be used to identify users?
  - ownership. Who owns the data? Who has the right to determine what the data can be used for?
  - data sharing and reuse. Can researchers or companies share the data they collect? Can they use data for different purposes than the purpose for which the data were originally collected?
- A privacy concern is when an individual or individuals can be identified from a data source. Both companies and statistical researchers need to be aware of privacy issues when collecting, publishing and storing personal data.
- When you install an application or download software, there is almost always a 'Terms and conditions' policy. Most people click on the 'I agree' button without reading the details — they trust the company that has provided the application or software. Unfortunately for users, some applications have written permissions in their agreement policies for applications to create and save files in various locations on their device. This gives the company access to a lot of personal data — data that you may have thought was safe on your own device. In other words, clicking on 'I agree' means that you have given your consent for the company to use your data in the ways set out in the terms and conditions.
- Some applications such as Google Maps can provide your exact location. Social applications such as Snapchat and Facebook can access your pictures and contacts as well as your location. When you register with social or store applications, you are also providing details about your date of birth, your email accounts and your credit card.
- Because of the digital world we live in, application and software companies can monitor users' online browsing activities and social media interactions. Computer algorithms have been built into applications to use people's personal data. The algorithms then send advertisements and

recommendations based on users' preferences and lifestyle habits. For example, have you ever used a web-based media application such as Netflix and then received movie recommendations the next time you used the application? Have you ever shopped at an online store and then searched on another site, only to find the online store advertising in your browser? When you agreed to the terms and conditions of the application, you may have given your consent for the company to use your data in this way.

8.4.2 Guidelines for ethical research

- In Australia, research on human subjects is covered by the National Statement on Ethical Conduct in Human Research (2007). Researchers collecting data need to consider and explain the processes used to protect research participants and ensure anonymity of data. The Privacy Act 1988 requires Commonwealth agencies to comply with the Information Privacy Principles (IPPs) regarding personal information. Researchers must consider relevant privacy legislation when storing data and ensure that the collection, storage and use is permitted by law.
- Research involving different cultures may contain other ethical challenges. The Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) provides guidelines for ethical research and other educational resources for such situations involving Indigenous Australian groups.

WORKED EXAMPLE 9

A supermarket chain offers a loyalty rewards program. The sign-up form for the rewards program requests your personal contact information and additional demographic data. The rewards card allows the chain to monitor your spending habits, and the chain will use the data to bombard you with offers by automatically sending you 'suggestions' for your most frequently purchased items. The company can also sell the data to anyone else it wishes.

What privacy or ethical issues do you need to consider before you sign up to the program?

THINK

What privacy and ethical issues need to be considered before signing up to the program?

WRITE

The main considerations to signing up to the program are privacy and confidential issues such as 'Where is the personal information stored?' and 'Can I be identified?'
As the supermarket chain can monitor your buying habits and offer 'suggestions', you are being identified through the email system. The following are some ethical issues that relate to this:

- The company knows your demographic data, which can be an intrusion on your personal data.
- The company can sell your information to an unknown company, which means that you have no control over who sees that information.

It is important to read the company's privacy policy before you sign up.

8.4.3 Bias

- Researchers can consider ways to protect research participants and ensure anonymity of data. However, ethical problems can still arise for other reasons. **Bias** can occur throughout the investigative process, from data collection through to statistical analysis and reporting. Bias can result from biased questions, biased samples, biased answers and biased reporting. Bias can be prejudice in favour of or against an outcome. Bias can skew data and distort the truth of findings. Bias influences the validity and reliability of conclusions, which in turn can have consequences in how decisions are made.
- In a statistical investigation, bias can result from the statistical researcher themselves. A researcher may select biased sample groups (i.e. groups that do not represent the target population). If a researcher is collecting data directly, they may bias people's opinions by their behaviour (e.g. mannerisms, style of dress, speaking tone or body language). They may ask biased questions, or phrase questions in ways that influences participants' answers. The researcher may also have bias due to their own personal opinion, which can affect their analysis and reporting.

WORKED EXAMPLE 10

A computer company uses fingerprint access for workers to log on and off each day. Monthly personal data from a particular department within the company were collected by Human Resources to analyse. The results are shown in the table.

Based on the data, Human Resources created a report to the Managing Director. The report stated that all company employees are lazy and only 60% of employees work the 8 hours required per day. Is this report biased? Explain.

Name	Age	Average hours worked each day
Harry Stanton	52	8.5
Gia Spazianu	43	8
Ming Chen	35	8.5
Jyoti Lal	48	6.5
Pierre Ouboure	55	10
Mark Smith	37	8
Paul Ryan	29	7.5
Mary Lee	25	6
Gary Ng	33	7.5
Yin Fong	26	9

THINK

Think about how bias affects the validity and reliability of findings.

WRITE

From the data, only 60% of workers (6 out of 10 in this department) averaged 8 or more hours per day in that month. However, the company cannot assume that all company employees are lazy. The four other employees may have different work conditions — they could be employed as casual or part-time workers. The data is based on login times using a computer. The four employees that were not logged in for over 8 hours may also have job descriptions that do not require them to be at a computer workstation all day. This report is biased.

8.4.4 Biometric data

- **Biometric technology** is technology used to authenticate, validate or identify a characteristic of an individual for security purposes. Devices that check fingerprints, facial features, voices, written signatures, and iris colours and patterns are all examples of biometric technology.
- **Biometric data** is the specific measurement of a unique characteristic of an individual. Digital data such as passwords and identification numbers can be shared between individuals, but biometric data cannot be shared. For example, some financial institutions require fingerprint recognition to access a bank account, rather than an identification number, as a person's fingerprints are unique.
- Biometric data can be used by organisations to allow access to high-security information or restricted locations. Biometric data can also be collected by government agencies for law enforcement and border control. For example, facial and iris recognition are commonly used in international airports to identify individuals.
- The idea of biometric authentication is to enhance privacy by preventing unauthorised access to personal data. However, some biometric systems make use of personal data. Security and preventative measures need to be implemented in such databases to ensure that there is no possibility of an individual's privacy being at risk. Companies, government agencies and universities using biometric data need to be aware of the damage that can happen if the data is attacked by an outside source or unwanted intruder.

WORKED EXAMPLE 11

The table shows more data from the company department discussed in Worked example 10.

Name	Age	Average hours worked each day	Number of sick days	Annual wage ($)
Harry Stanton	52	8.5	4	98 000
Gia Spazianu	43	8	7	82 000
Ming Chen	35	8.5	2	84 500
Jyoti Lal	48	6.5	3	75 300
Pierre Ouboure	55	10	0	101 500
Mark Smith	37	8	3	77 500
Paul Ryan	29	7.5	1	58 500
Mary Lee	25	6	8	43 000
Gary Ng	33	7.5	4	65 250
Yin Fong	26	9	4	68 000

a. **All workers in the department are expected to work 8 hours each day. Mary Lee's average hours of work each day and her number of sick days taken for the month were forwarded in an email to all staff in her department. Was this ethical? Explain.**

b. **Using technology:**

- **i.** **plot the age against the number of hours worked for each employee on a scatterplot**
- **ii.** **describe the trend of the data**
- **iii.** **plot the regression line for this data set and write the equation in terms of the variables**
- **iv.** **calculate Pearson's correlation coefficient for this data set.**

c. **Is the following statement true? 'The older you are, the harder you work.' Explain your answer.**

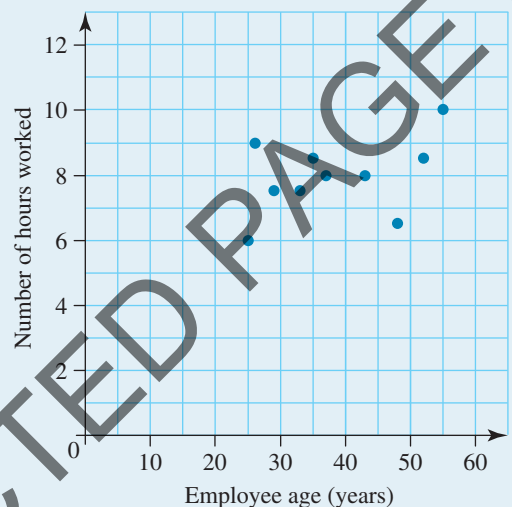
THINK

a. Think about the types of ethical issues.

WRITE

To forward and share someone's personal data is not ethical. Mary can be identified by the sharing of her data. Her consent was not given, and her privacy has been breached by the company.

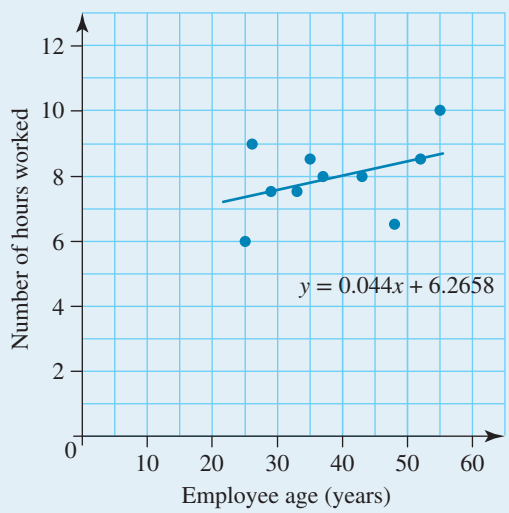
b. i. Using the technology of your choice, plot the data of age against the number of hours worked for each employee.



ii. Describe the trend of the data.

The data seems to have a weak positive linear correlation.

iii. Find the regression line.



$y = 0.044x + 6.2658$

Average number of hours worked daily $= 0.04 \times \text{Age} + 6.27$

iv. Use technology to calculate the correlation coefficient.	Pearson's correlation coefficient is the square root of r^2. $r^2 = 0.1642$ $r = \pm\sqrt{0.1642}$ $r = \pm 0.41$ Since the direction is positive, $r = 0.41$.
c. Explain the statement 'The older you are, the harder you work' using the data.	The equation of the regression line is Average number of hours worked daily $= 0.04 \times$ Age $+ 6.27$. The gradient of 0.04 means that for every year a person is older, the average number of hours worked increases by 0.04 per day. However, the statement 'the older you are the harder you work' may only apply to this sample of 10 people. It is a very generalised statement, not taking into account other variables such as job descriptions and conditions. Logging in to a computer cannot measure how hard someone works.

Exercise 8.4 The statistical investigation process

Understanding, fluency and communicating

1. Answer True or False for each of the following privacy statements.
   - **a.** If a website has a privacy policy, it means that the site cannot share information about you with other companies, unless you give the website your permission.
   - **b.** If a website has a privacy policy, it means that the website must delete information it has about you, such as your name and address, if you ask them to do so.
   - **c.** If a company wants to follow your internet use across multiple sites on the internet, it must first obtain your permission.
2. **WE 9** List any ethical or privacy issues that are involved when a user gives consent by clicking on 'I agree' to an online policy in an application or software download.
3. Rewrite the following biased questions for use in surveys.
   - **a.** Should concerned parents vaccinate their children?
   - **b.** Have you stopped bullying people on social media?
4. A department chain store throughout Australia surveys customers for their postal code. Is this an ethical concern?

5. To sign into a social club, all visitors over 18 had to use their fingerprints for identification, as well as scanning their driver's licence. The social club sold their identification to other companies for marketing. Are there any ethical or privacy issues involved here? Explain.
6. What is an example of a privacy risk associated with using biometric data?
7. Jasjit, a university student, applies to go on a 6-month exchange from Australia to Denmark. As part of his VISA approval process, he has to submit fingerprint biometric data. What are the privacy issues associated with this process?
8. **WE 10** A computer company uses fingerprint access for workers to log on and off each day. Monthly personal data from a particular department within the company was collected by Human Resources to analyse. The results are shown in the table.

Name	Age	Average hours worked each day	Number of sick days	Annual wage ($)
Harry Stanton	52	8.5	4	98 000
Gia Spazianu	43	8	7	82 000
Ming Chen	35	8.5	2	84 500
Jyoti Lal	48	6.5	3	75 300
Pierre Ouboure	55	10	0	101 500
Mark Smith	37	8	3	77 500
Paul Ryan	29	7.5	1	58 500
Mary Lee	25	6	8	43 000
Gary Ng	33	7.5	4	65 250
Yin Fong	26	9	4	68 000

   a. All workers are entitled to have 12 sick days per year. The number of sick days Gia and Mary took for the month were forwarded in an email to all staff in their department. Was this ethical? Explain.
   b. Based on the data, Human Resources created a report to the company director. The report stated that all company employees are slack and 90% of employees have used more than their average entitlement for sick leave for the month. Is this report biased? Explain.
   c. Using technology:
      i. plot the age against the annual wage for each employee on a scatterplot
      ii. describe the trend of the data
      iii. plot the regression line for this data set and write the equation in terms of the variables
      iv. calculate Pearson's correlation coefficient for this data set.
   d. Is the following statement true? 'The older you are, the more money you earn.' Explain your answer.

9. A cryptography lecturer at a university wanted to study the relationship between the number of class absences a student had in a given semester and the student's final course grade. The course was delivered online, and students accessed it using voice recognition. The data shown were collected from a sample of students studying the cryptography course.

Number of absences	Final course grade (%)
0	89.2
1	86.4
2	83.5
3	81.1
4	78.2
5	73.9
6	64.3
7	71.8
8	65.5
9	66.2

a. How did the lecturer make sure that the sample data collected was not biased?

b. Using technology:

i. plot the number of absences against the final course grade for each student on a scatterplot

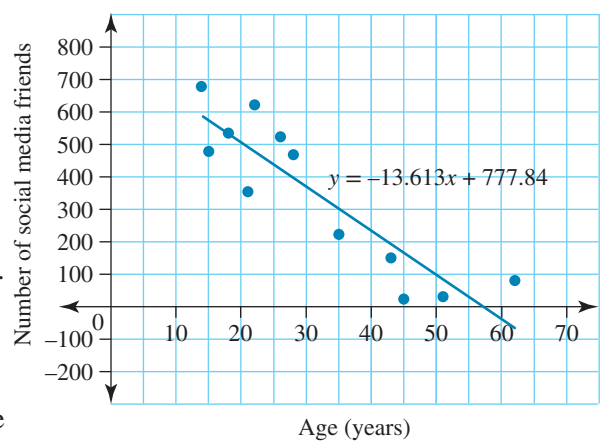
ii. describe the trend of the data

iii. plot the regression line for this data set and write the equation in terms of the variables

iv. calculate Pearson's correlation coefficient determination for this data set.

c. The lecturer wrote a report on his investigation and published the data, including the students' personal details. Was this ethical? Explain.

10. A phone carrier company used fingerprint biometric technology to collate some results about people using a particular social media application. The graph 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 $y = -13.613x + 777.84$.



a. Identify the independent variable in this situation.

b. Rewrite the equation in terms of the independent and dependent variables.

c. Pearson's 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.

e. Is it ethical for the phone carrier to publish these results? Explain.

11. A high-tech company provides each of its employees with a mobile phone and computer. Biometric software is installed to monitor employees' email and text usage. The following data are collected on a daily basis about each employee. The company also collects the employee's personal data.

Age (years)	Number of text messages received	Number of emails received	Number of emails sent
22	132	88	34
24	124	70	25
27	72	74	24
31	85	63	20
32	105	52	19
36	48	30	12
37	64	50	20
43	52	35	16
48	33	40	15
54	24	45	10

a. Using technology:
- **i.** create a scatterplot of employees' ages against the number of text messages they received
- **ii.** describe the trend of the data
- **iii.** calculate Pearson's correlation coefficient for this data set
- **iv.** plot the regression line for this data set and write the equation in terms of the variables
- **v.** interpret the regression line.

b. The company sends a survey question to all employees asking how much time they spend on social media during work hours. What is wrong with the question?

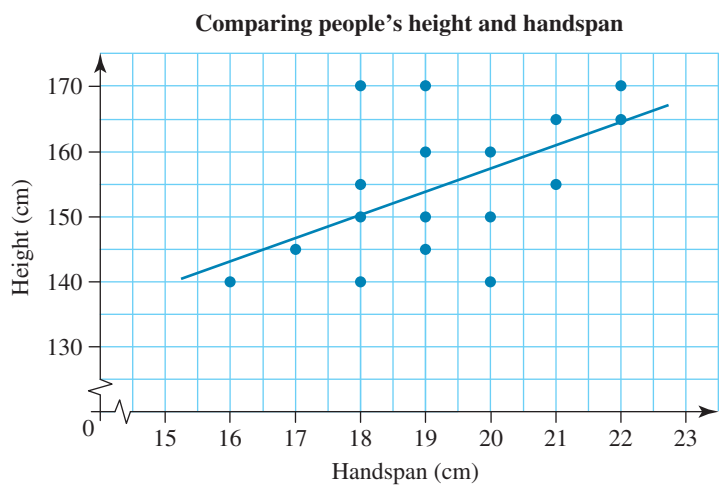
c. The company publishes their research to all employees, including all the data that was collected. Is there an ethical concern?

12. Data were collected on the heights and handspans of 17 students at school.

a. Determine an approximate regression line for this data set and write the equation in terms of the variables.

b. Interpret the regression line.

c. What is the handspan of a student who is 160 cm tall?



Problem solving, reasoning and justification

13. To access free wi-fi at a café, customers need to click 'I agree' on the café website. Give an example of a privacy or ethical concern that would be involved if a customer clicks 'I agree' on the café website without reading the terms and conditions.
14. In a published article, a group of researchers released the profile data collected from the social application accounts of an entire university. What ethical and privacy issues are involved in the release of the data?

15. Digital applications on phones and wearable devices enable users to track their health statistics.
    a. List some of the different types of personal data these applications can record about the users.
    b. List any ethical or privacy concerns about using these types of applications.
16. The data below shows the total crowd attendance at home and away games for Greater Western Sydney Giants AFL club in the years 2012-2017. (Attendance figures for finals and grand finals are not included.)

	Total crowd attendance	
Year	**Home games**	**Away games**
2012	119 073	230 797
2013	106 715	239 015
2014	101 491	252 333
2015	118 651	283 367
2016	135 664	267 543
2017	145 152	286 823

Source: AFL Tables, https://afltables.com/afl/crowds/gws.html

    a. Using technology, draw a scatterplot of total crowd attendance at home games versus total crowd attendance for away games in each year.
    b. Describe the trend of the data.
    c. Find the equation of the line of best fit and write the equation in terms of the variables.
    d. Calculate Pearson's correlation coefficient for this data set. What does this mean?
    e. From the data, it can be said that 'People who live in Western Sydney don't like AFL football.' Explain if the statement is biased.

8.5 Review

8.5.1 Summary

In this topic you have learnt:

- how to construct a bivariate scatterplot in order to identify patterns in the data that suggest the presence of an association
- to use bivariate scatterplots to describe the patterns, features and associations of bivariate datasets,
  - describe bivariate datasets in terms of form (linear/non-linear) and, in the case of linear, the direction (positive/negative) and strength of any association (strong/moderate/weak)
  - identify the dependent and independent variables within bivariate datasets where appropriate
  - describe and interpret a variety of bivariate datasets involving two numerical variables
  - calculate and interpret Pearson's correlation coefficient (r) using technology to quantify the strength of a linear association of a sample
- to model a linear relationship by fitting an appropriate line of best fit to a scatterplot and using it to describe and quantify associations
  - fit a line of best fit both by eye and by using technology to the data
  - fit a least squares regression line to the data using technology
  - interpret the intercept and gradient of the fitted line
- to use the appropriate line of best fit, both found by eye and by applying the equation, to make predictions by either interpolation or extrapolation
  - recognise the limitations of interpolation and extrapolation, and interpolate from plotted data to make predictions where appropriate
- implement the statistical investigation process to answer questions that involve identifying, analysing and describing associations between two numerical variables
  - construct, interpret and analyse scatterplots for bivariate numerical data in practical contexts while demonstrating awareness of issues of privacy and bias, ethics, and responsiveness to diverse groups and cultures
  - investigate using biometric data.

eBookplus RESOURCES

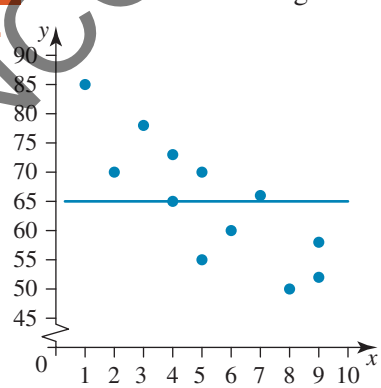
Digital doc: Topic 8 summary — a comprehensive summary of key learning points (doc-26477)

Exercise 8.5 Review

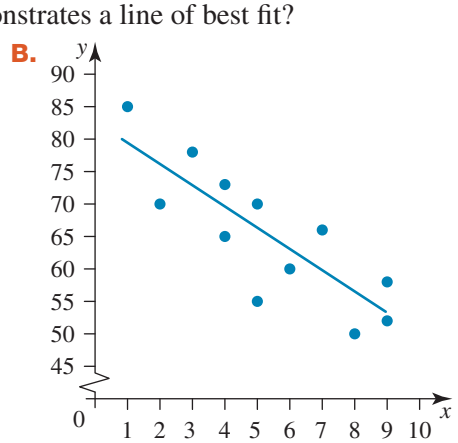
Understanding, fluency and communicating

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

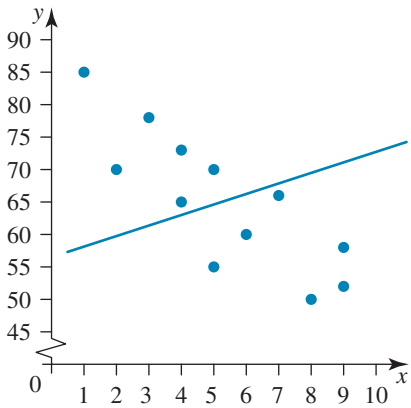
A.



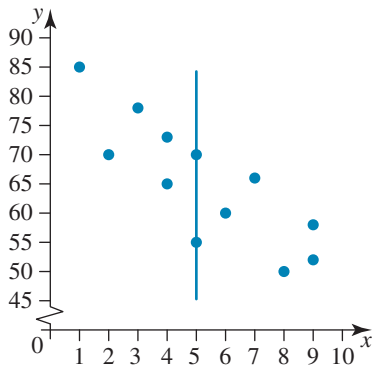
B.



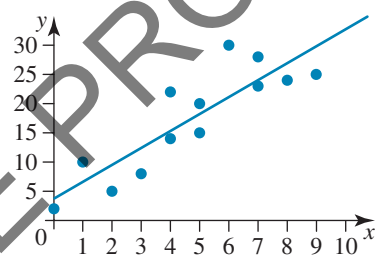
C.



D.



2. MC The regression line equation for the following graph is closest to:



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

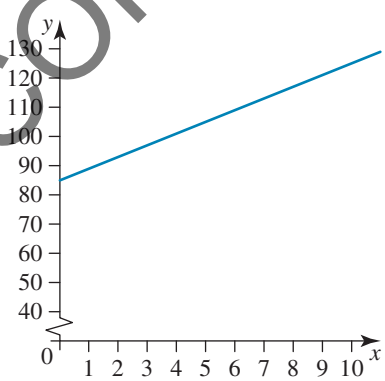
3. MC The type of correlation shown in the graph for question 2 would best be described as:
A. weak, positive correlation
B. moderate, positive correlation
C. strong, positive correlation
D. no correlation

4. MC What type of correlation does an r-value of 0.64 indicate?
A. Strong, positive correlation
B. Strong, negative correlation
C. Moderate, positive correlation
D. Moderate, negative correlation

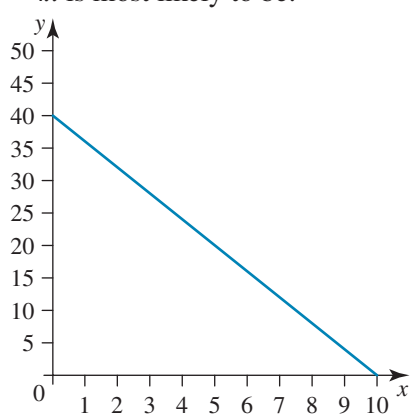
5. MC A gardener tracks a correlation coefficient of 0.79 between the growth rate of his trees and the amount of fertiliser used. What can the gardener conclude from this result?
A. An increase in tree growth increases the use of fertiliser.
B. There is no correlation between the growth rate of the trees and the amount of fertiliser used.
C. The growth rate of the trees is influenced by the amount of fertiliser used.
D. The growth rate of the trees influences the quality of the fertiliser used.

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

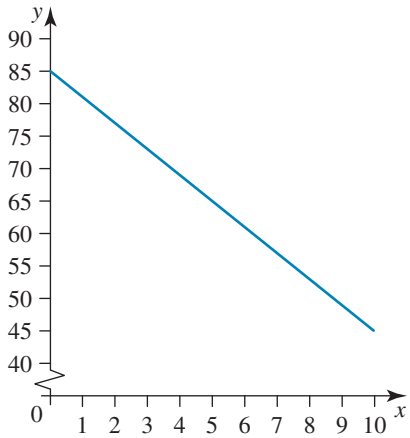
A.



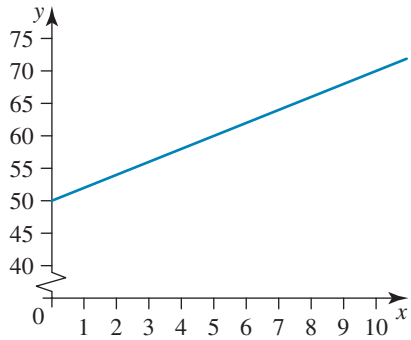
B.



C.



D.



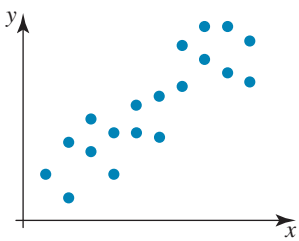
7. MC For the sample data set shown in the table, which of the following is an example of interpolating data?

x	1	5	15	25
y	10	16	18	22

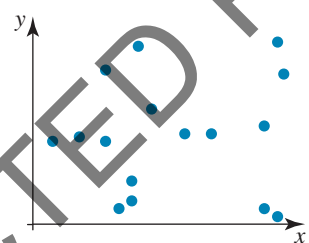
A. Finding the value of x when $y = -7$
B. Finding the value of y when $x = 17$
C. Finding the value of x when $y = 27$
D. Finding the value of y when $x = 37$

8. For each of the following graphs, describe the strength of correlation between the independent and dependent variables.

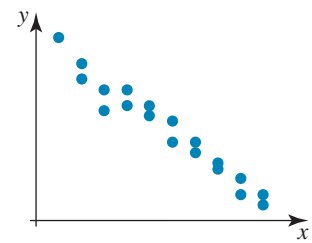
a.



b.



c.



9. For each of the graphs in question **8**, draw a line of best fit where possible.

10. Identify the independent and dependent 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.

11. Complete the following for the data shown in the table.

x	10	9	8	7	6	5	4	3	2	1
y	6	10	4	11	13	18	15	19	21	26

a. Plot the data on a scatterplot.
b. Comment on the direction and strength of the data.
c. Find Pearson's correlation coefficient for the data.
d. Use you answer from part **c** to further discuss the relationship between the variables.

Problem solving, reasoning and justification

12. Data were collected on 15 people's shoe size and hair length.

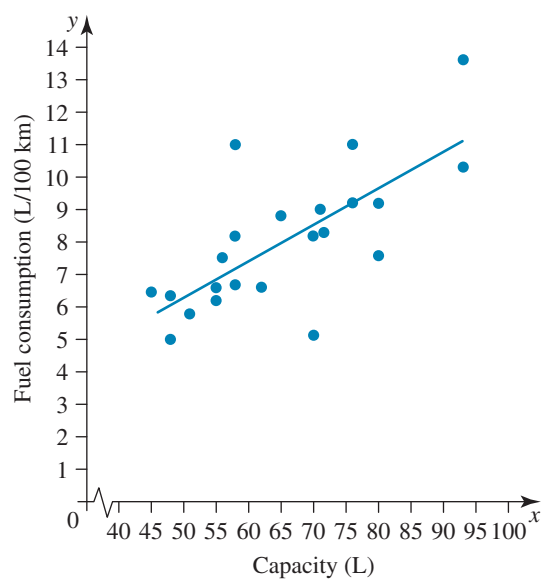
Shoe size	Hair length (cm)
6	9
8	14
7	12
8	1
9	7
6	8
7	5
12	22

Shoe size	Hair length (cm)
8	15
9	8
10	18
12	4
7	5
9	9
11	3

a. Draw a scatterplot of the data.
b. Find the equation of the line of best fit.
c. Find Pearson's correlation coefficient for the data.
d. What conclusions could you draw from the data?

13. An independent agency test-drove a random sample of current model vehicles and measured their fuel tank capacities against their average fuel consumption rates. The following scatterplot was drawn, and a regression equation of $y = 0.1119x + 0.6968$ was established.

a. Identify the dependent variable in this situation.
b. Rewrite the equation in terms of the independent and dependent variables.
c. It is often said that smaller vehicles are more economical. Determine, correct to 2 decimal places, the fuel consumption of a vehicle with a 40-litre fuel tank.
d. Is your answer to part **c** an example of interpolation or extrapolation? Explain your dependent.
e. Calculate, correct to the nearest whole number, the tank size of a vehicle that has a fuel consumption rate of 10.2 L per 100 km.
f. Pearson's correlation coefficient for this data set is 0.516. How can you use this value to evaluate the reliability of the data?
g. List at least two other factors that could influence the data.

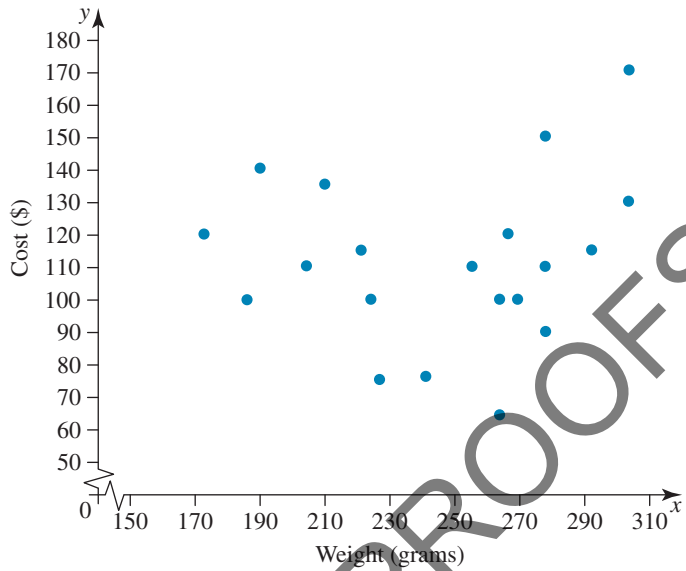


14. The weights of top brand running shoes were tracked against their recommended retail prices, and the results were recorded in the following scatterplot.

a. Identify the independent variable for this situation.

b. How would you describe the relationship between these two variables?

c. Identify two external factors that could explain the distribution of the data points.



15. The following data show the average adult weight and gestation period of a selection of mammals.

Mammal	Average adult weight (kg)	Usual gestation period (weeks)
Human	70	39
Dolphin	250	50
Elephant	6000	90
Cat	4	9
Horse	800	48
Rabbit	4	4
Giraffe	1800	62
Elk	520	34
Bison	1000	34

a. Using technology, draw the scatterplot of average adult weight against gestation period.

b. Describe the trend of the data.

c. Calculate Pearson's correlation coefficient for this data set.

d. Find the equation of line of best fit and write the equation in terms of the variables.

16. The Bureau of Meteorology records data such as maximum temperatures and solar exposure on a daily and monthly basis. The data table below, 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 independent and dependent variables for this situation.

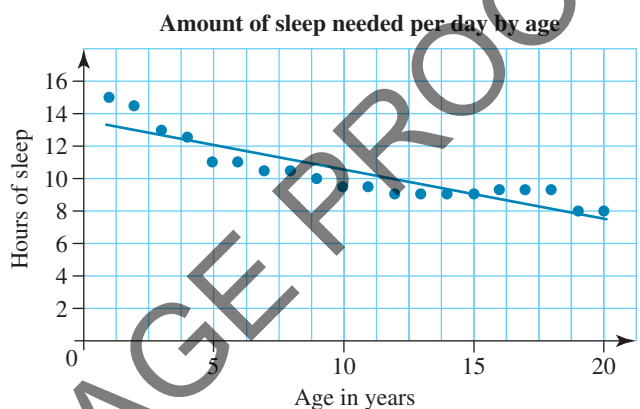
b. Using technology, plot the data on a scatterplot.

c. Describe the trend of the data.

d. Calculate Pearson's correlation coefficient for this data set.

e. Plot the regression line for the data and write the equation in terms of the variables.
f. Using your equation from part e, 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.

17. Below is a scatterplot showing the amount of sleep needed per day, by age. The regression line equation for this data set is $y = -0.2995x + 13.489$.
a. Identify the dependent variable in this situation.
b. Rewrite the equation in terms of the independent and dependent variables.
c. Pearson's correlation coefficient for this data set is −0.8942. 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 hours of sleep required for a 22-year-old. Is this an example of interpolation or extrapolation? Explain your dependent.



18. The Hawaiian–Emperor seamount chain is a range of volcanoes in the Pacific. Most of the volcanoes are underwater, but the southeast end of the chain reaches above the water to form the islands of Hawaii. The chain was formed by the movement of the Pacific Plate (a section of the Earth's crust) over a hot spot (a place where hot rock rises up from below the crust). By measuring the age and location of the volcanoes and seamounts, geologists can estimate the speed and direction of movement of Pacific Plate.

Volcano name	Age (million years)	Distance from the Hawaii hot spot (km)
Kilauea	0.20	0
Mauna Kea	0.38	54
Kohala	0.43	100
East Maui	0.75	182
Kahoolawe	1.03	185
West Maui	1.32	221
Lanai	1.28	226
East Molokai	1.76	256
West Molokai	1.90	280
Koolau	2.60	339
Waianae	3.70	374
Kauai	5.10	519
Niihau	4.89	565
Nihoa	7.20	780
Necker	10.3	1058
La Perouse	12.0	1209
Brooks Bank	13.0	1256

Volcano name	Age (million years)	Distance from the Hawaii hot spot (km)
Gardner	12.3	1435
Laysan	19.9	1818
Northampton	26.6	1841
Pearl and Hermes	20.6	2291
Midway	27.7	2432
Colahan	38.6	3128
Abbott	38.7	3280
Daikakuji	42.4	3493
Yuryaku	43.4	3520
Kimmei	39.9	3668
Koko	48.1	3758
Ojin	55.2	4102
Jingu	55.4	4175
Nintoku	56.2	4452
Suiko 1	59.6	4794
Suiko 2	64.7	4860

a. Using technology, draw a scatterplot of age versus distance for the Hawaii–Emperor chain.
b. Describe the trend of the data.
c. Find the equation of the line of best fit and write the equation in terms of the variables.
d. Estimate the rate at which the Pacific Plate has been moving over the last 65 million years.
e. Calculate Pearson's correlation coefficient for this data set. What does this mean?

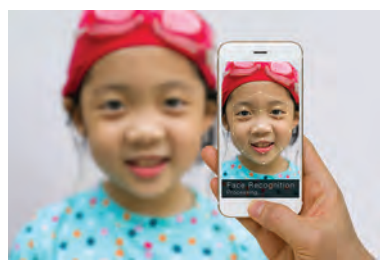
19. The table below shows data from 2014 comparing Uber and taxi fares for journeys in different cities in America. The journeys in each city cover the same distance.

City	Uber fare ($)	Taxi fare ($)
New York	17.75	15.50
Philadelphia	15.25	14.20
Portland	15.05	15.00
Cleveland	13.00	13.95
Dallas	10.30	11.25
Miami	13.25	14.50
Indianapolis	11.65	13.00
Phoenix	11.00	12.50
Minneapolis	12.15	14.25
Baltimore	10.75	13.05
Columbus	10.20	12.85
Denver	10.35	13.75
Detroit	12.30	16.50
Seattle	11.70	16.00
San Francisco	12.30	17.25
Chicago	9.50	14.00
Boston	11.10	16.60
Atlanta	10.00	15.00
Houston	9.00	13.75
San Diego	11.35	17.80
Los Angeles	9.40	16.30

Source: Uber, taxifarefinder.com

a. Using technology, draw a scatterplot of Uber fare price versus taxi fare price for the same distance travelled.
b. Describe the trend of the data.
c. Find the equation of the line of best fit and write the equation in terms of the variables.
d. Calculate Pearson's correlation coefficient for this data set. What does this mean?
e. Which is cheaper to use, Uber or taxi? Explain.

20. A social application on a digital device allows facial biometric data to be used in its photo tagging features. What are the privacy issues associated with this feature?



21. A computer company uses fingerprint access for workers to log on and off each day. Monthly personal data from a particular department within the company was collected by Human Resources to analyse. The results are shown in the table.

Name	Age	Average hours worked each day	Number of sick days	Annual wage ($)
Harry Stanton	52	8.5	4	98 000
Gia Spazianu	43	8	7	82 000
Ming Chen	35	8.5	2	84 500
Jyoti Lal	48	6.5	3	75 300
Pierre Ouboure	55	10	0	101 500
Mark Smith	37	8	3	77 500
Paul Ryan	29	7.5	1	58 500
Mary Lee	25	6	8	43 000
Gary Ng	33	7.5	4	65 250
Yin Fong	26	9	4	68 000

a. All employees' names and their annual wages were forwarded in an email to all staff in their department. Is this ethical? Explain.

b. Using technology:

i. plot the average hours worked against the annual wage for each employee on a scatterplot

ii. describe the trend of the data

iii. plot the regression line for this data set and write the equation in terms of the variables

iv. calculate Pearson's correlation coefficient for this data set.

22. As part of airport security, data are collected from iris recognition scanners for all international arrivals at Sydney airport. On a particular morning, the number of passengers on each flight and the number of Australian residents on each flight were processed in customs. The data are shown in the table.

Number of passengers on flight	Number of Australian residents on flight
350	150
189	30
290	47
451	212
366	174
525	408
220	63
172	94

a. List some security concerns with collection of these data.

b. Using technology:

i. plot the data on a scatterplot

ii. describe the trend of the data

iii. write the regression line equation and write the equation in terms of the variables

iv. state how many passengers would be expected to be Australian residents if a plane had 310 passengers

v. calculate Pearson's correlation coefficient for this data set.

23. To monitor work productivity, a company uses fingerprint scanner devices for employees to log in to their workstations. The data for one week is shown below. In a monthly report, it was stated that all employees are less productive on a Wednesday. Explain if this statement is biased.

Day	Number of employee logins	Average hours of computer activity
Monday	25	7.2
Tuesday	20	6.3
Wednesday	23	4.1
Thursday	21	6.8
Friday	18	6.5

24. For marketing purposes, a technology company analyses a sample of the number of applications purchased on a personal device and the total cost spent for these applications for a group of teenagers in one month.

Number of applications purchased	Total spent on applications ($)
4	22.35
2	3.50
1	7.50
1	1.75
5	16.85
3	6.00
2	5.60
1	2.25
3	10.95

a. Based on the data, what is the average amount spent on one application for a personal device by a teenager?

b. The company accesses its complete database and sells the teenagers' details and all information regarding their purchases to a software developer. The software developer then sends spam emails to the teenagers. Is this ethical?

25. An online travel website collects personal data from customers who purchase flights or accommodation through the website. Visa and passport numbers, credit card details, date of birth and address details are all recorded. The travel website sells the information to a third party which manages a rewards loyalty program and insurance policies. Is this ethical?

26. A video streaming company records the amount of movies and TV series watched each week and the age of the person with the account.

Number of hours spent watching movies and TV series	Age of account holder
42	73
27	14
21	23
15	18
31	49
19	36
34	65

a. Using technology:

i. plot the data on a scatterplot

ii. describe the trend of the data

iii. plot the regression line for this data set and write the equation in terms of the variables.

b. i. If a person views movies and/or TV series for 10 hours a week, how old would the account holder be?

ii. Is this extrapolation or interpolation? Explain.

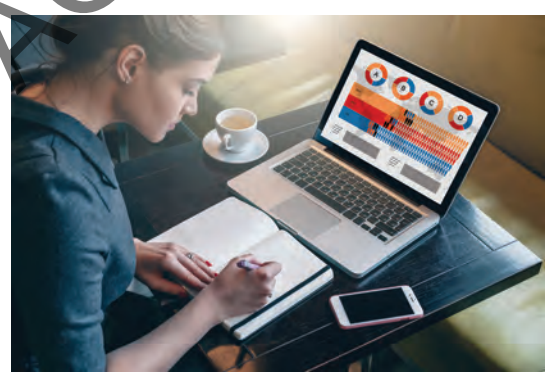
c. The video streaming company sells the account holders' information to a third party. The third party company starts to send spam emails relating to what the account holders like to view on the video service. Is this ethical?

27. A passport service study researched the real-time speed for individual verification at a biometric device station in customs. The study found that it took on average of 1 minute and 30 seconds for fingerprint verification of each individual. (This included the time needed for the user's interaction with the biometric device.) In the report, some individuals took up to 5 minutes and 15 seconds. These individual outliers were categorised as disabled users and were deemed to be 'understandably slower'. Is it ethical to write this statement in a report?

28. An adaptive learning tool has been designed to assist students with their studies in a course. The amount of time some of the students spent using the tool and the number of tasks they completed are shown in the table. There are 40 tasks to complete in the course.

Student name	Time spent (minutes)	Number of tasks completed
Dylan Bradley	5	3
Ben Hillman	94	10
Sophie Krieger	66	28
Ming Yi	17	23
Henry He	340	32
Surya Hamid	223	40
Jesse Sondhu	38	16
Tyler Aaron	171	35

a. Each day, the system automatically emails all students who have completed less than 15 tasks. All students' names along with number of tasks completed are listed in the email. Is the generation of a student list to be sent to these students ethical?



b. Using technology:
- **i.** plot the data on a scatterplot
- **ii.** describe the trend of the data
- **iii.** plot the regression line for this data set and write the equation in terms of the variables.

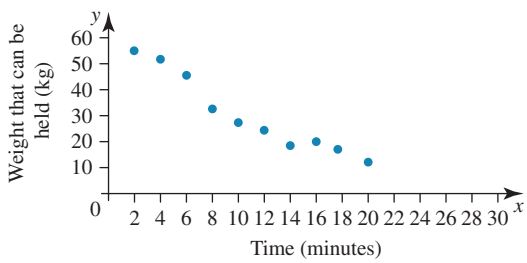
c. **i.** If a student completes 27 tasks, approximately what is the time spent using the adaptive tool?
- **ii.** Is this extrapolation or interpolation? Explain.

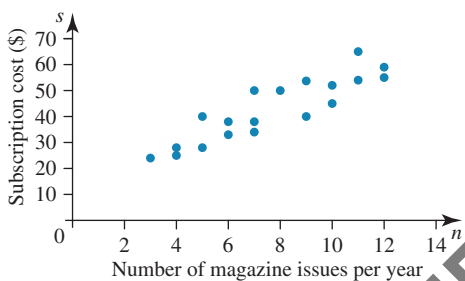
Answers

Topic 8 Bivariate data analysis

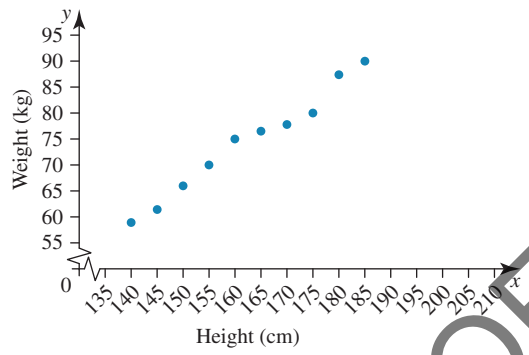
Exercise 8.2 Scatterplots

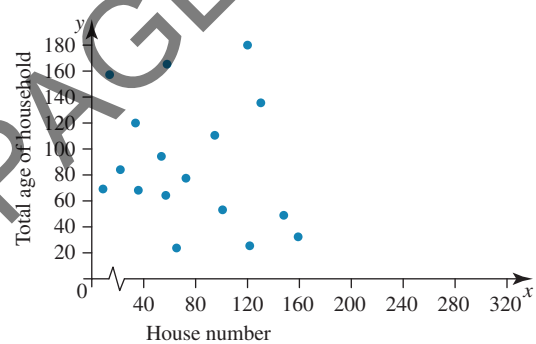
1. Independent variable: time of day
   Dependent variable: time taken to eat a pizza
2. Independent variable: time spent using phone apps
   Dependent variable: amount of data required
3. a. Independent variable: time
   Dependent variable: weight that can be held (kg)

   b.


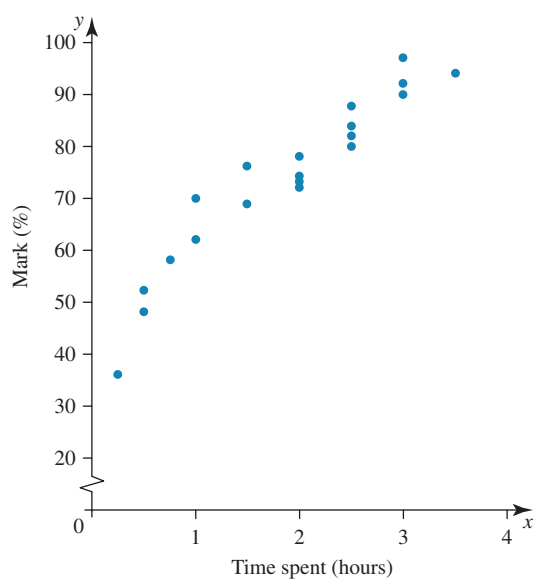
   c. Strong negative correlation
4. a. File size
   b. Strong positive correlation
5. a.


   b. The scatterplot has a strong positive linear relationship between the number of magazines and the subsciption cost.
   c. $r = 0.8947$, which indicates a strong positive linear association
6. a. Independent variable: age
   Dependent variable: number of star jumps
   b. Independent variable: quantity of chocolate
   Dependent variable: cost
   c. Independent variable: number of songs
   Dependent variable: memory used
   d. Independent variable: food supplied
   Dependent variable: growth rate of bacterial cells
7. a. Independent variable: height
   Dependent variable: weight

   b.


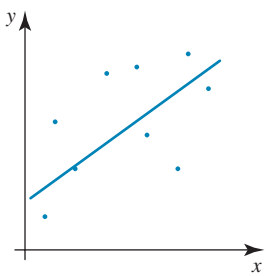
8. a. Weak negative correlation
   b. No correlation
   c. Moderate positive correlation
9. Various possible answers, for example:
   a. Loss of money over time
   b. Temperature and number of shoes owned
10. a. A moderate positive linear association
    b. A strong negative linear association
    c. No linear association
    d. A weak positive linear association
11. a.


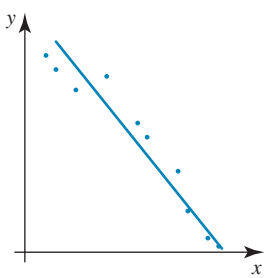
    b. There is no correlation.
    c. $r = -0.2135$
    d. There is no relationship between the house number and the age of the household.
12. a. Independent variable: time spent
    Dependent variable: mark

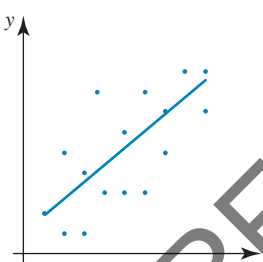
    b.


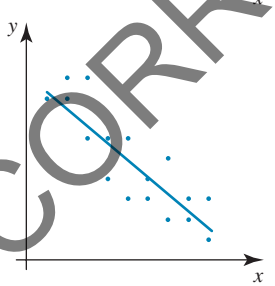
c. Strong positive linear correlation
d. Each person's understanding of the topic is different and their study habits are unique. Therefore 1 hour spent on the assignment does not guarantee a consistent result. Individual factors will also influence the assignment mark.
e. $r = 0.952$
f. There is a strong relationship between the time spent on an assignment and the resulting grade. As the time spent increased, so did the mark.

Exercise 8.3 Lines of best fit

1.


2.


3. a.


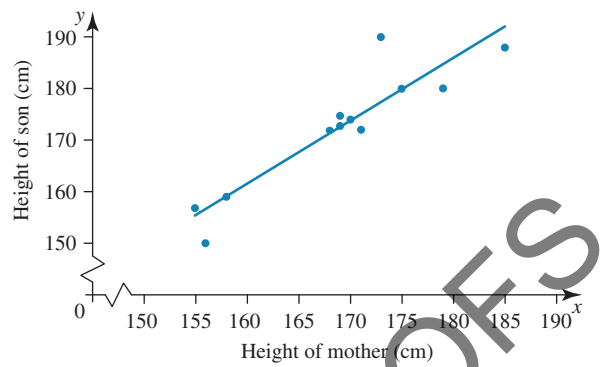
b.


4. a. Weak positive correlation
b. Moderate negative correlation

5. a. The Mathematics exam mark
b. 0.95
c. −6.07
d. 75%

6. a. $y = 3.33x + 91.78$ **b.** $y = -0.15x + 21.87$
c. $y = 52.38x + 8890.48$ **d.** $y = -x + 30$

7. a. Dependent variable: height of son

b.


c. $S = -33.49 + 1.219M$

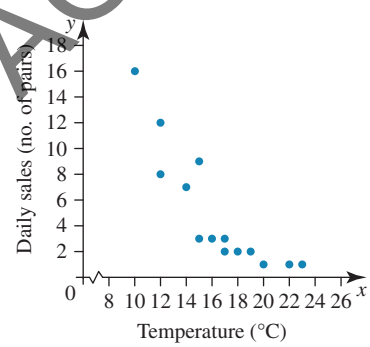
8. a. −1.837 **b.** 1.701
c. Positive trend

9. a. 105.9 **b.** −1.476
c. Negative trend

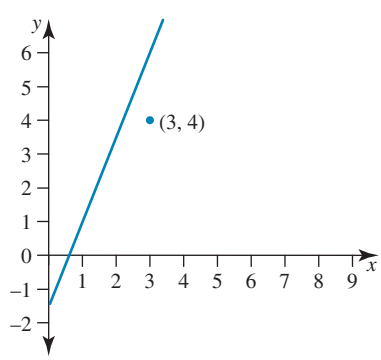
10. a. Age **b.** 2.029 mL
c. 7.5 months old

11. a. Cost per night **b.** $155.74
c. 3.31 km

12. As the b value (gradient) is negative, the trend is negative. The y-intercept is 3.2; therefore, when $x = 0$, $y = 3.2$.

13. a.


b. $D = 22.50 - 1.071T$
c. $r = -0.8621$
d. There is a moderate negative relationship between the number of gumboots sold and the temperature. The data indicates that 74% of the sales are due to the temperature; therefore 26% of sales are due to other factors.

14. a.


b. Below the regression line

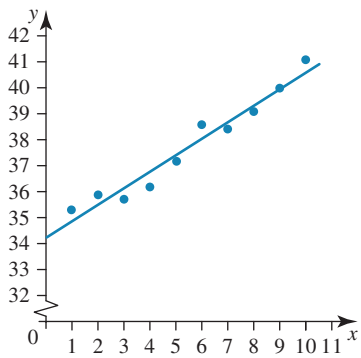
15. a. 60 **b.** −5 **c.** Negative **d.** −140

16. a. 3.90
b. Lucy incorrectly transposed the 12.9. She should have moved this first before dividing by 7.32.

17. **a.** −12 **b.** 25 **c.** Positive **d.** 75.5

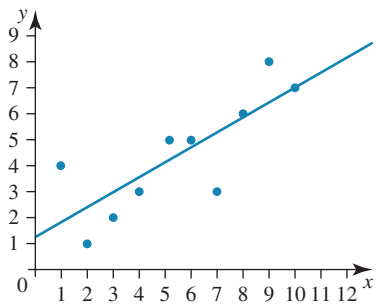
18. **a.** Number of insects caught **b.** 1.1
c. Positive **d.** 66

19. **a.**



b. $y = 34.23 + 0.641x$, When $x = 15$, $y = 43.85$.

20. **a.**



b. $y = 1.2 + 0.58x$
c. 12.8
d. 13.45

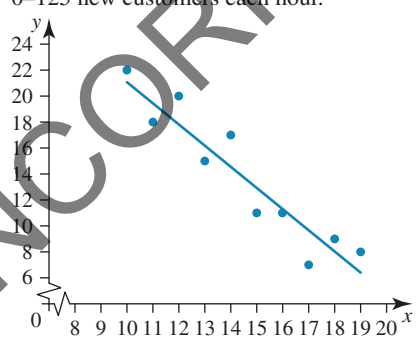
21. $y = 4 + 2.5x$

22. Various answers are possible.
An example data set would be:

x	1	2	2	3	4	5	5	6	7	7	8	9	10	12	14
y	2	5	6	8	9	7	12	15	16	22	25	29	32	33	35

23. **a.** 76
b. 65
c. Part **a** looks at data within the original data set range, while part **b** predicts data outside of the original data set range of 0–125 new customers each hour.

24. **a.**



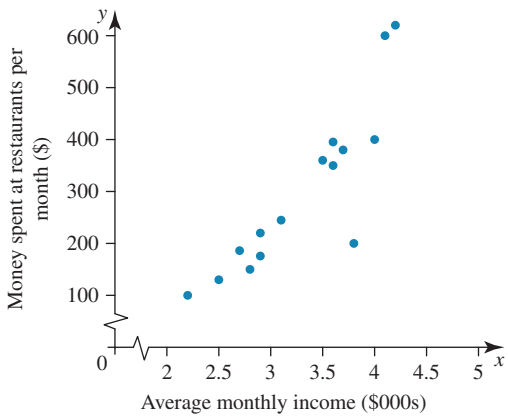
b. $y = 37.70 - 1.648x$
c. −0.204
d. It is assumed the data will continue to behave in the same manner as the data originally supplied.

25. a. 130

b. 3.6 °C

c. The location of the fire, air temperature, proximity to water, etc.

26. a.



b. $R = -459.8 + 229.5I$

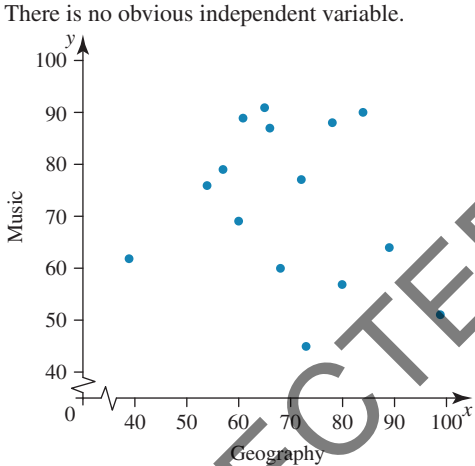
c. $687.70

d. Part **c** asks you to predict data outside of the original data set range.

e. $3160, interpolation

27. a. There is no obvious independent 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. $r = -0.2172$. This indicates very weak correlation between the data, which supports the view that conclusions cannot be drawn from this data.

28. a. Calories burned

b. Calories burned = 14 301 + 115.02 × Distance walked

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

d. 15 451.2. Extrapolation, as the independent 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.

Exercise 8.4 The statistical investigation process

1. a. False **b.** False **c.** False

2. Sample responses can be found in the worked solutions in the eBookPLUS. Participants need to read and check the privacy issues relating to their personal data when they tick the 'I agree' box online. Participants also need to check the storage and the use of such data by the company.

3. a. 'Should parents vaccinate their children?' Including 'concerned' in the question implies bias.

b. 'Do you bully people on social media?' Saying 'have you stopped' implies that you already have bullied and you are a bully on social media.

4. People often live in areas that are associated with particular ethnic identities, so using an area code in a data mining study runs the risk of building models that are based on race, even though racial information has been explicitly excluded from the data.

5. The privacy issues concerned with the social club selling the information is that the other companies may be able to identify a person associated with the data.

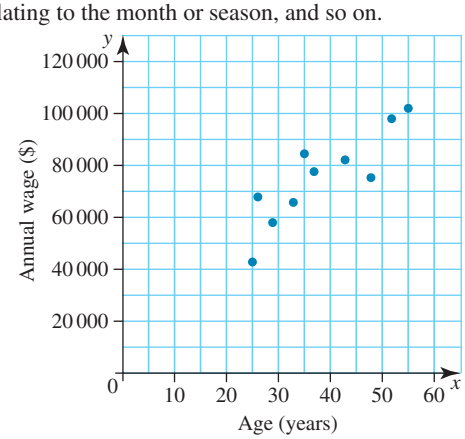
6. The privacy risk with using biometric data is the storage of the data and the potential damage that can be caused if the data is hacked, as the identification of the data can then be exposed.

7. The privacy risk for Jasjit in using biometric data is the storage of the data and the potential damage that can be caused if the data storage is hacked or misused. Jasjit could be identified from his data, or the data could be used in identity theft. However, in Denmark and Australia, the use of biometric data to identify a person is important for security.

8. a. It is not ethical to publish the number of sick days for employees and share this data, as it breaches the employees' privacy.

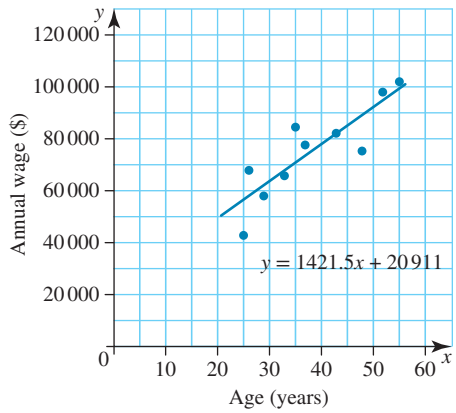
b. Saying that all company employees are slack and 90% of employees have used more than their average entitlement for sick leave for the month is biased, as the statement has not considered the whole company. The department chosen may not be a representative sample. Also, the statement has not considered other factors such as work entitlements, job descriptions, factors relating to the month or season, and so on.

c. i.



ii. The data follows a strong positive linear correlation.

iii.



$y = 1421.5x + 20\,911$

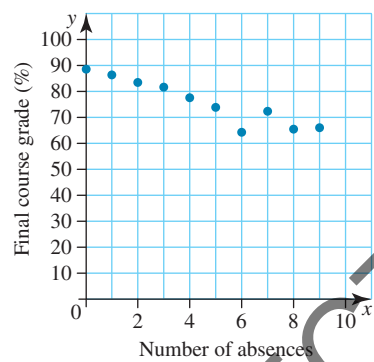
Annual income $= 1421.5 \times$ age $+ 20\,911$

iv. $r = 0.86$

d. From the regression line, 'the older you are, the more money you earn' is true for this data set. However, not everyone earns more as their age increases. Other factors need to be considered, such as experience and qualifications.

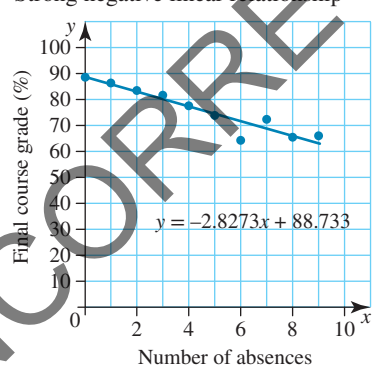
9. a. The data was not biased as it was collected through the use of biometric voice recognition. No other students could access the course.

b. i.



ii. Strong negative linear relationship

iii.



$y = -2.8273x + 88.733$

Final grade $= -2.83 \times$ number of absences $+ 88.7$

iv. $r = -0.95$

c. It is not ethical to publish the student data and details as the students would be identifiable.

10. a. Age

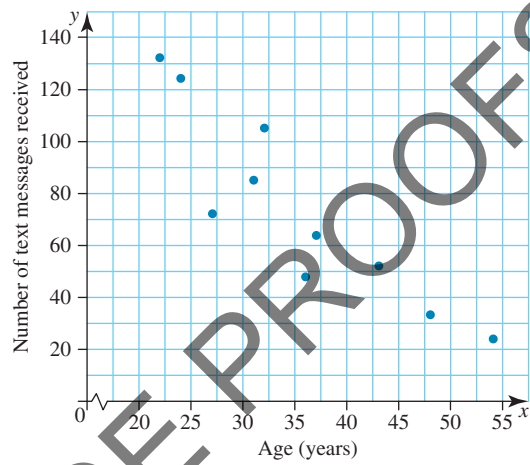
b. Number of social media friends $=$ $-13.613 \times$ age $+ 777.84$

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.

e. It is only ethical for the phone carrier to publish these results if the data of the people involved cannot be identified.

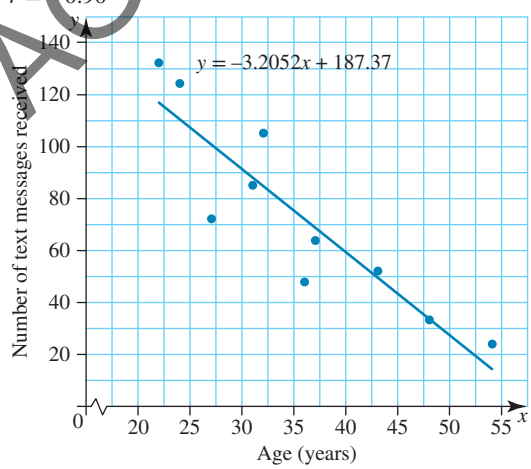
11. a. i.



ii. Strong negative linear relationship

iii. $r = -0.90$

iv.



$y = -3.2052x + 187.37$

Number of text messages $= -3.2052 \times$ age $+ 187.37$

v. As the age increases by 1 year, the number of text messages is reduced by approximately 3 messages.

b. The question is biased, as it is implying that their employees are already using social media during work hours.

c. The ethical concern in publishing all the data is that the privacy of individuals has been breached.

12. a. $y = 3.6x + 86$

Height $= 3.6 \times$ handspan $+ 86$

b. As handspan increases by 1 cm , the height of a student increases by 3.6 cm.

c. 20.56 cm (2 decimal places)

13. Sample responses can be found in the worked solutions in the eBookPLUS. Sample answer: The café can monitor the user's online browsing activities.

14. Sample responses can be found in the worked solutions in the eBookPLUS. Sample answer: Researchers breached a number of ethical obligations, including failing to gain specific consent from the students whose data were being harvested for the study. Researchers failed to ensure the students' expectations of privacy, and insufficient attention was paid to ensuring the efficacy of anonymisation techniques before the data were released. .

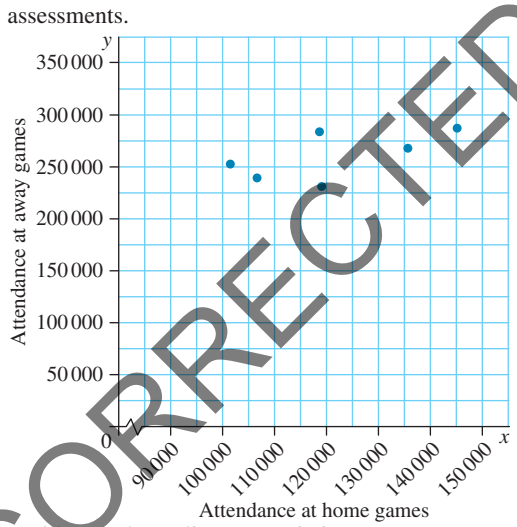
15. Sample responses can be found in the worked solutions in the eBookPLUS. Sample answers shown:

a. Data could include location, age, height, weight, heart rate, steps walked in a day etc.

b. Ethical concerns about the use of this data focus on who collects the data, how it is handled, and what privacy protections are given. Research that uses mobile phones to collect data may collect the data with the knowledge of the mobile phone owners, or the data may be collected without the consent or knowledge of the person who generated it.

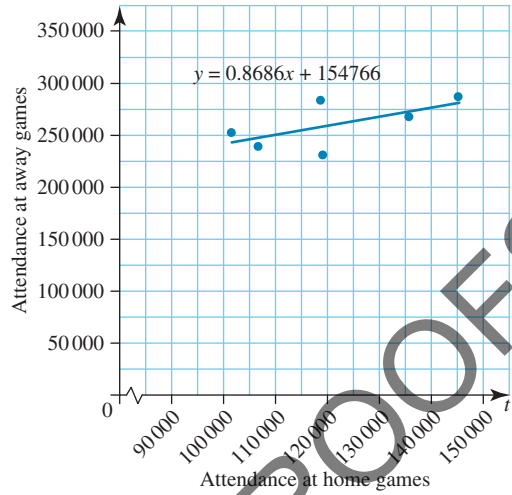
Information gathered from mobile phones may be stored remotely, for example in the cloud, with the phone manufacturers, and/or within the application on the phones. Concerns about these forms of storage include security and the extent to which the data may become available to other parties. These digital data can be used and shared by commercial entities for research, financial gain and/or for advertising purposes. Companies such as health insurers are beginning to encourage their clients to use self-tracking devices and are using the data produced to make insurance assessments.

16. a.



b. Positive moderate linear association

c.



$y = 0.8686x + 154\,766$

Away crowd attendance

$= 0.8686 \times$ home crowd attendance $+ 154\,766$

d. $r = 0.63$; moderate positive association

e. The statement is biased. The data shows people's attendance, not their geographical location.

Exercise 8.5 Review

1. B

2. A

3. B

4. C

5. C

6. C

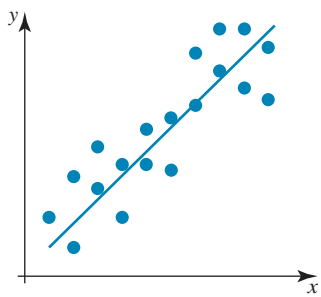
7. B

8. a. Moderate

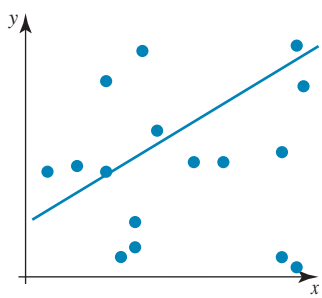
b. Weak

c. Strong

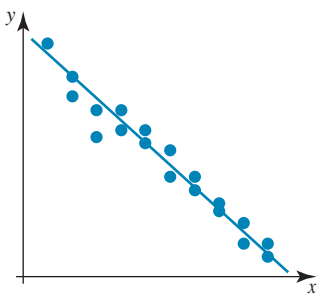
9. a.



b.



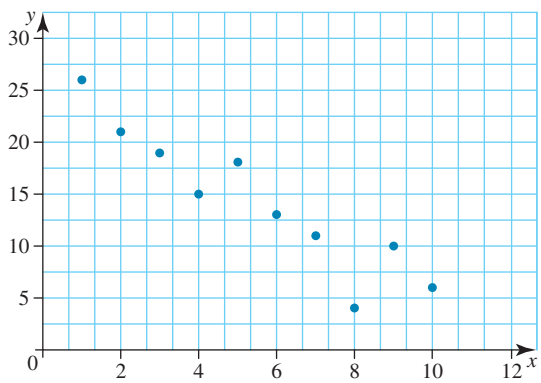
c.



10. a. Independent variable: amount of water, dependent variable: time

b. Independent variable: amount of students, dependent variable: number of buses

11. a.

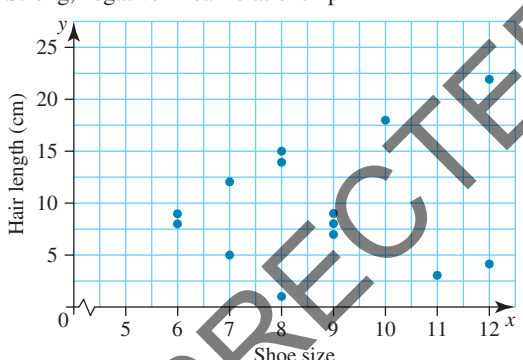


b. Strong, negative association

c. $r = -0.93$

d. Strong, negative linear relationship

12. a.



b. $y = 0.6157x + 4.0386$

Hair length = 0.6157 × shoe size + 4.0386

c. $r = 0.21$

d. Weak, positive linear relationship

13. a. Fuel consumption

b. Fuel consumption = 0.1119 × capacity + 0.6968

c. 5.17 L per 100 km

d. This is extrapolation, as it is a prediction outside the data range.

e. 84.93 L

f. There is a moderate positive relationship between fuel consumption and capacity.

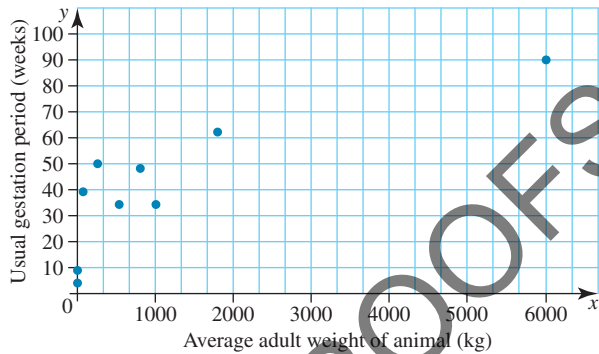
g. Sample responses can be found in the worked solutions in the eBookPLUS. But could include the surface of the road and the steepness of the road.

14. a. Weight

b. Weak relationship

c. Sample responses can be found in the worked solutions in the eBookPLUS. But could include the place where the runners are manufactured and the type of material used to manufacture the runner.

15. a.



b. Moderate positive trend

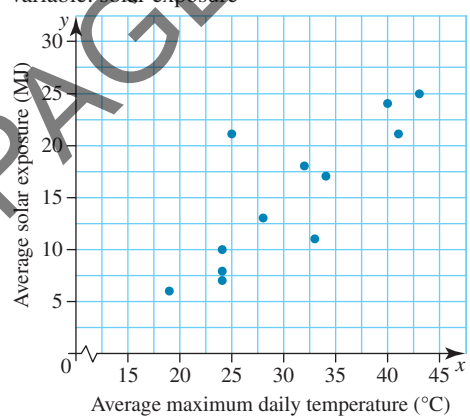
c. $r = 0.82$

d. $y = 0.0113x + 28.044$

Gestation period = 0.0113 × animal weight + 28.044

16. a. Independent variable: daily temperature, dependent variable: solar exposure

b.



c. Strong positive relationship

d. $r = 0.82$

e. $y = 0.7139x - 6.7511$

Solar exposure = 0.7139 × daily temperature − 6.7511

f. 19.7 °C

g. 13.7 °C

h. It involves predicting a value outside the given data values.

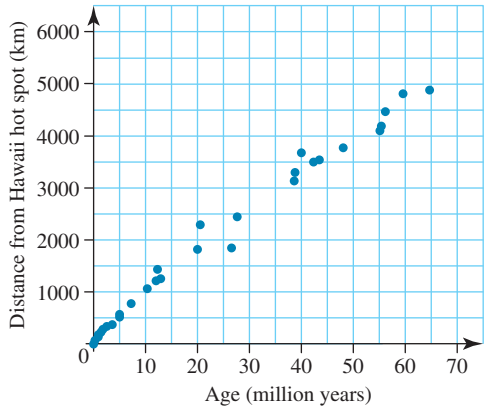
17. a. Hours of sleep

b. Hours of sleep = −0.2995 × age + 13.489

c. Strong negative relationship

d. 6.9 hours. This is an example of extrapolation, as it involves prediction outside the given data values.

18. a.



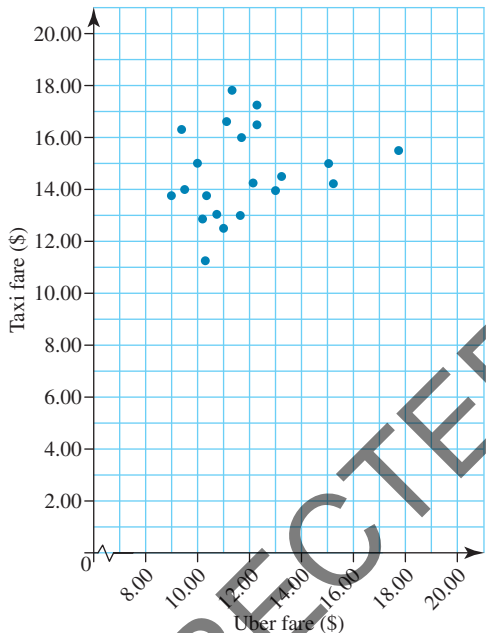
b. Strong positive association

c. $y = 75.989x + 186.55$
Distance $= 75.989 \times$ age $+ 186.55$

d. 75.989 km/million years

e. $r = 0.99$; strong positive linear relationship

19. a.



b. Weak positive association

c. $y = 0.1614x + 12.718$
Taxi fare $= 0.1614 \times$ Uber fare $+ 12.718$

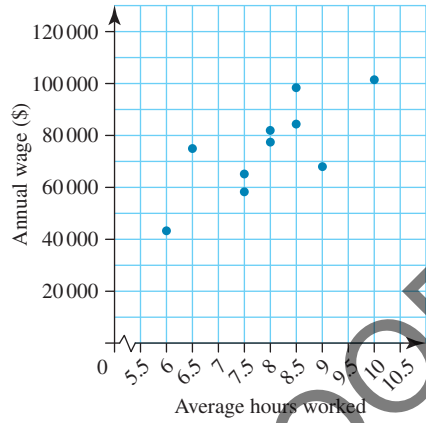
d. $r = 0.21$; weak positive linear relationship

e. It is cheaper to use Uber. For every $0.01 spent on an Uber, it costs $0.16 for a taxi.

20. Sample responses can be found in the worked solutions in the eBookPLUS. Privacy issues could include that people in the background of photos or in photos with friends who use social media may have not given consent to have their facial biometric data used. Photo tagging people without their consent is a privacy issue. There could also be concerns with how the company owning the application stores and uses the data.

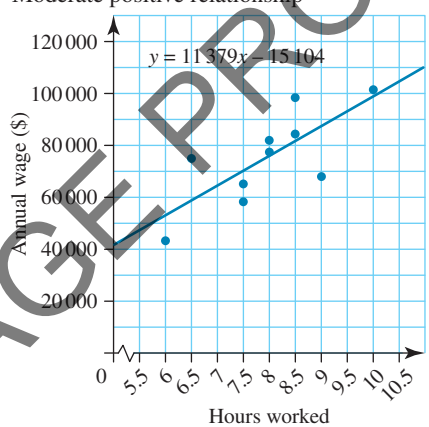
21. a. It is not ethical to send an email to all employees including the names of employees and their annual wages. Staff have not given consent, and their privacy has been breached, as they can be identified by their data.

b. i.



ii. Moderate positive relationship

iii.

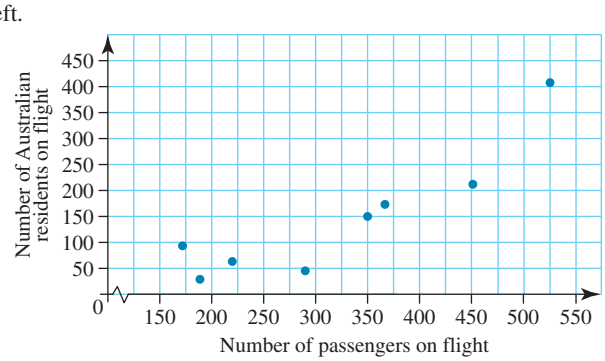


Annual wage $= 11\,379 \times$ number of hours worked $- 15\,104$

iv. $r = 0.75$; Moderate positive linear relationship

22. a. Security concerns include how the data are stored, and the potential damage that can be caused if the data storage is hacked or misused, as the identification of people will be known and hacking could lead to identity theft.

b. i.



ii. Moderate positive linear relationship

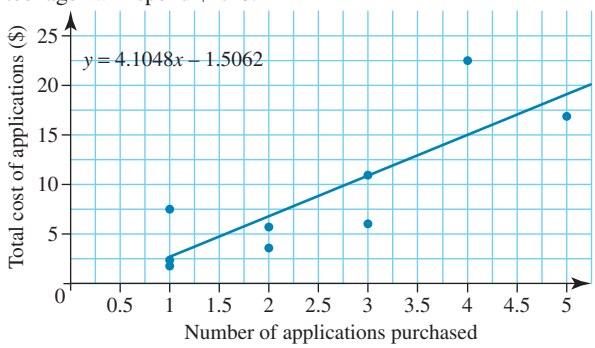
iii. $y = 0.873x - 132.44$
Number of Australian residents
$= 0.873 \times$ number of passengers $- 132.44$

iv. 138 passengers

v. $r = 0.90$; strong positive linear relationship

23. Based on the data, the value for Wednesday could be considered an outlier. The statement that all employees are less productive on a Wednesday is biased. It generalises to all Wednesdays when the data only include values for 1 week. Other factors may need to be considered, for example whether meetings or professional learning sessions occurred on that day.

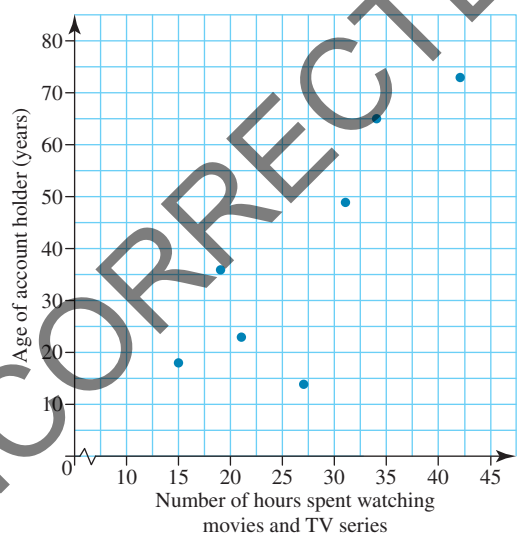
24. a. From the trend line, for every 1 application purchased, a teenager will spend $4.10.



b. Sample responses can be found in the worked solutions in the eBookPLUS. Sample answer:
If the teenagers have agreed for the company to sell their data to third parties, then they have no rights to their data. However, if the teenagers have not agreed, it is not ethical for the company to sell the teenagers' details and all information regarding to purchases to a software developer, as this is a breach of the teenagers' privacy.

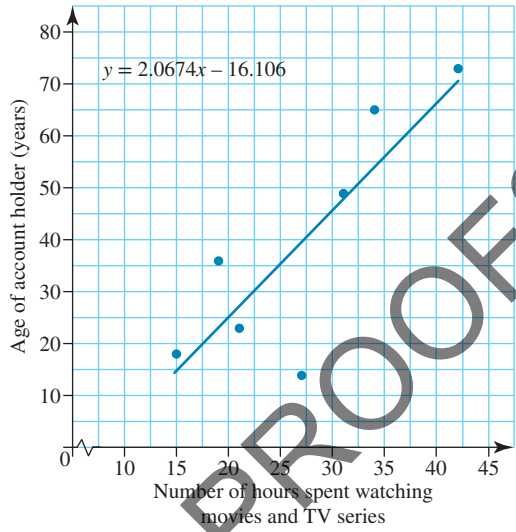
25. Sample responses can be found in the worked solutions in the eBookPLUS. Sample answer: It is not ethical for the travel website to sell the information to a third party, as the customers' privacy has been breached. Visa and passport numbers, credit card details, date of birth and address details can all be used to identify individuals.

26. a. i.



ii. Moderate positive relationship

iii. $y = 2.0674x - 16.106$
Age = 2.0674 × number of hours − 16.106



b. i. 4.6 years

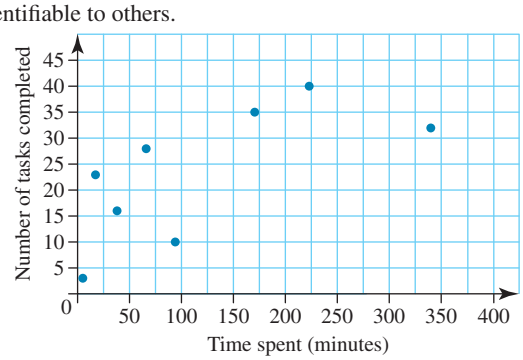
ii. This is extrapolation, as the predicted value is outside the given data values.

c. Sample responses can be found in the worked solutions in the eBookPLUS. Sample answer:
It is not ethical for the video streaming company to sell account holders' information to a third party if the account holders have not consented, particularly if account holders can be identified from their data, as privacy will have been breached. However, if account holders have given consent in their contract, then the company can use for any legal purpose.

27. Sample responses can be found in the worked solutions in the eBookPLUS. Sample answer:
It is not ethical to write this, and the report is quite biased. It assumes a lot about the user's physical and mental capabilities just based on the time taken to use the device. Other factors may be involved, for example the device itself could have had mechanical and software issues.

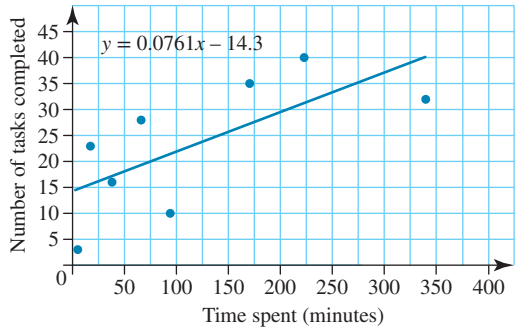
28. a. Sample responses can be found in the worked solutions in the eBookPLUS. Sample answer:
It is not ethical for all students' names along with number of tasks completed to be sent to students, as students' privacy has been breached by making them identifiable to others.

b. i.



ii. Moderate positive relationship

iii.



$y = 0.0761x + 14.3$

Tasks completed = $0.0761 \times$ time + 14.3

c. i. 167 minutes

ii. This is interpolation, as the predicted value is within the given data values.