

Chapter 7

Investigating relationships between two numerical variables

Chapter questions

- ▶ What is bivariate data?
- ▶ What are response and explanatory variables?
- ▶ What is a scatterplot, how is it constructed and what does it tell us?
- ▶ What do we mean when we describe the association between two numerical variables in terms of direction, form and strength?
- ▶ What is Pearson's correlation coefficient, how is it calculated and what does it tell us?
- ▶ What is the difference between association and causation?
- ▶ How do we fit a line of good fit to a scatterplot by eye?
- ▶ How do we fit a line of good fit to a scatterplot using the least squares method?
- ▶ How do we interpret the intercept and slope of a line fitted to a scatterplot in the context of the data?
- ▶ How do we use a line fitted to a scatterplot to make predictions?
- ▶ What is the difference between interpolation and extrapolation?

In this chapter, we begin our study of **bivariate data**; data which is recorded on two variables from the same subject. Measuring the height and the weight of a particular person would be an example of bivariate data.

Bivariate data arises when we consider questions such as: Is the new treatment for a cold more effective than the old treatment? Do younger people spend more time using social media than older people?

Each of these questions is concerned with understanding the association between the two variables. To investigate such relationships will require us to develop some new statistical tools. In this chapter, we will only consider analysis of bivariate data where

both of the variables are classified as numerical

7A Scatterplots

Learning intentions

- ▶ To be able to define **explanatory** and **response** variables.
- ▶ To be able to identify which of the two variables in the data may be the **explanatory variable** and which may be the **response variable**.
- ▶ To be able to construct a **scatterplot** by hand and by using a CAS calculator.

Response and explanatory variables

When we analyse **bivariate data**, we try to answer questions such as: ‘Is there an association between these two variables?’ More specifically, we want to answer the question: ‘Does knowing the value of one of the variables tell us anything about the value of the other variable?’

For example, let us take as our two variables the *mark* a student obtained on a test and the amount of *time* they spent studying for that test. It seems reasonable that the more time one spends studying, the better mark you will achieve. That is, the amount of *time* spent studying may help to **explain** the *mark* obtained. For this reason we call *time* the **explanatory variable (EV)**. And, since the *mark* may go up or down in response to the amount of *time* spent studying, we call *mark* the **response variable (RV)**. In general, we anticipate that the value of the explanatory variable will have some effect on the value of the response variable.

Response and explanatory variables

When investigating associations (relationships) between two variables, the explanatory variable (EV) is the variable we expect to explain or predict the value of the response variable (RV).

Note: The explanatory variable is also sometimes called the independent variable (IV), and the response variable, the dependent variable (DV).

Identifying response and explanatory variables

It is important to be able to identify the explanatory and response variables before starting to explore the association between two numerical variables. Consider the following examples.



Example 1 Identifying the response and explanatory variables

We wish to investigate the question: ‘Do older people sleep less?’ The variables here are *age* and *time spent sleeping*. Which is the response variable (RV), and which is the explanatory variable (EV)?

Explanation

When looking to see if the length of time people spent sleeping is explained by their age, *age* is the EV and *time spent sleeping* is the RV.

Solution

EV: age
RV: time spent sleeping

**Example 2** Identifying the response and explanatory variables

We wish to investigate the association between kilojoule consumption and weight loss. The variables in the investigation are *kilojoule consumption* and *weight loss*. Which is the response variable (RV), and which is the explanatory variable (EV)?

Explanation

Since we are looking to see if the weight loss can be explained by the amount people eat, *kilojoule consumption* is the EV and *weight loss* is the RV.

Solution

EV: kilojoule consumption
RV: weight loss

**Example 3** Identifying the response and explanatory variables

Can we predict a person's height from their wrist circumference? The variables in this investigation are *height* and *wrist circumference*. Which is the response variable (RV), and which is the explanatory variable (EV)?

Explanation

Since we wish to predict height from wrist circumference, *wrist circumference* is the EV. *Height* is then the RV.

Solution

EV: wrist circumference
RV: height

It is important to note that, in Example 3, we could have asked the question the other way around, that is 'Can we predict a person's wrist circumference from their height?' In that case, *height* would be the EV and *wrist circumference* would be the RV. The way we ask our statistical question is an important factor when there is no obvious EV and RV.

Now try this 3 Identifying response and explanatory variables (Example 3)

A teacher is concerned that her students are tired in class. She suspects it is because they spend too much time on social media when they should be sleeping. The variables in the investigation are *amount of sleep* and *time on social media*. Which is the response variable (RV), and which is the explanatory variable (EV)?

Hint 1 Consider the way the teacher has posed the question. Which variable does the teacher think the students should reduce? That is the explanatory variable here.



Constructing a scatterplot manually

The first step in investigating an association between two numerical variables is to construct a visual display of the data, which we call a **scatterplot**.

The scatterplot

- A scatterplot is a plot which enables us to display bivariate data when **both of the variables are numerical**.
- In a scatterplot, each point represents a single case.
- When constructing a scatterplot, it is conventional to use the **vertical** or **y-axis** for the response variable (RV) and the **horizontal** or **x-axis** for the explanatory variable (EV).

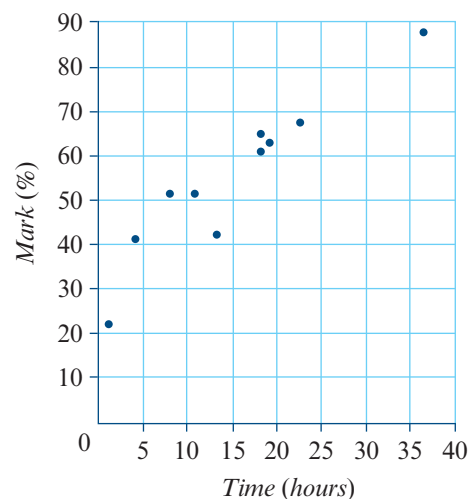
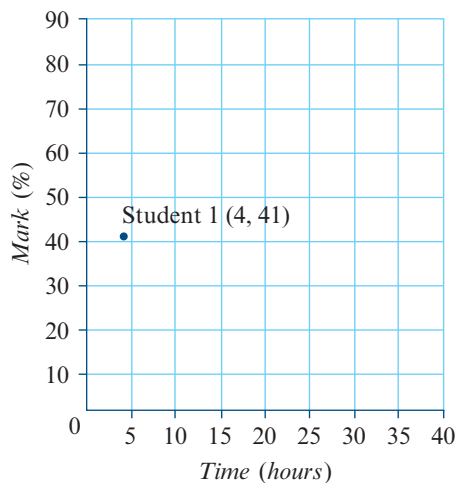
We will illustrate the process by constructing a scatterplot of the marks students obtained on an examination (the RV) and the times they spent studying for the examination (the EV).

<i>Student</i>	1	2	3	4	5	6	7	8	9	10
<i>Time (hours)</i>	4	36	23	19	1	11	18	13	18	8
<i>Mark (%)</i>	41	87	67	62	23	52	61	43	65	52

In this scatterplot, each point will represent an individual student, and:

- The horizontal or x -coordinate of the point represents the time spent studying.
- The vertical or y -coordinate of the point represents the mark obtained.

The following scatterplots show how a scatterplot is constructed. The scatterplot on the left shows the point for Student 1, who studied 4 hours for the examination and obtained a mark of 41. The completed scatterplot on the right shows the data plotted for all students (one point for each student).

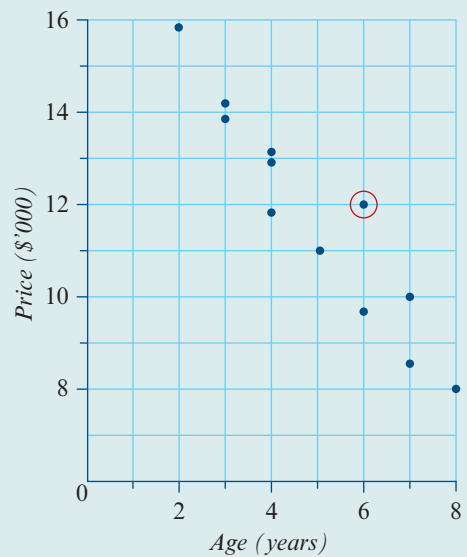


**Example 4** Understanding a scatterplot

The scatterplot shown has been constructed from data collected to investigate the association between the *price* of a second-hand car and its *age*.

Use the scatterplot to answer the following questions.

- 1 Which is the explanatory variable and which is the response variable?
- 2 How many cars are in the data set?
- 3 How old is the car circled? What is its price?

**Explanation**

- 1 The EV will be on the horizontal axis and the RV on the vertical axis.
- 2 The number of cars is equal to the number of points on the scatterplot.
- 3 The x -coordinate of the point will be the car's age and the y -coordinate is its price.

Solution

EV: Age
RV: Price
12 Cars

The car is 6 years old, and its price is \$12 000.

Now try this 4 Understanding a scatterplot (Example 4)

Construct a scatterplot of the maximum temperature on a summer day (the EV) and the number of bottles of water a student drank over the course of that day (the RV) over a 10-day period.

<i>Day</i>	1	2	3	4	5	6	7	8	9	10
<i>Temperature</i>	24	26	22	25	29	32	29	35	33	28
<i>Number of bottles of water</i>	2	3	2	3	4	5	4	6	6	4

Hint 1 Label the horizontal axis: *Temperature*. You can start the scale at 0, but it is acceptable to use a scale from 20 to 35 (for example), as long as it covers the values of the data.

Hint 2 Label the vertical axis: *Number of bottles of water*, with a scale from 0 to 6.

Hint 3 Note that the completed plot shows only 9 dots, as two points in the data set are identical.

Using a CAS calculator to construct a scatterplot

While you need to understand the principles of constructing a scatterplot and maybe need to construct one by hand for a few points, in practice you will use a CAS calculator to complete this task.

How to construct a scatterplot using the TI-Nspire CAS

The data below shows the marks that 10 students obtained on an examination and the time they spent studying for the examination.

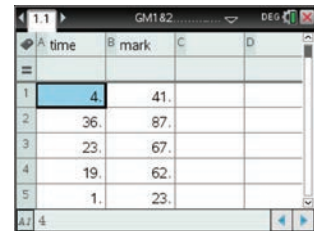
Time (hours)	4	36	23	19	1	11	18	13	18	8
Mark (%)	41	87	67	62	23	52	61	43	65	52

Use a calculator to construct a scatterplot. Use *time* as the explanatory variable.

Steps

- 1 Start a new document ($\text{ctrl} + \text{N}$) and select **Add Lists & Spreadsheet**.

Enter the data into lists named *time* and *mark*.



- 2 Statistical graphing is done through the **Data & Statistics** application.

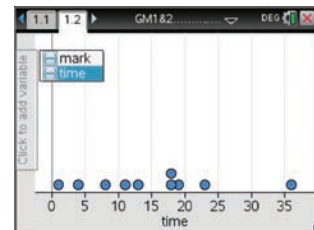
Press $\text{ctrl} + \text{I}$ and select **Add Data & Statistics** (or press $\text{ctrl} + \text{on}$ and arrow to I and press enter).

Note: A random display of dots will appear – this is to indicate list data is available for plotting. It is not a statistical plot.

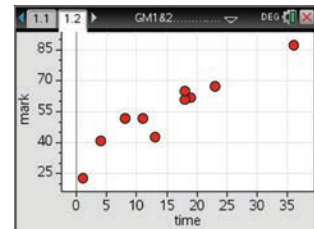


- 3 To construct a scatterplot.

- a Press tab and select the variable *time* from the list. Press enter to paste the variable, *time*, to the x-axis.
- b Press tab again and select the variable *mark* from the list. Press enter to paste the variable, *mark*, to the y-axis to generate the required scatterplot. The plot is automatically scaled.



Note: To change colour, move the cursor over the plot and press $\text{ctrl} + \text{menu} > \text{Colour} > \text{Fill Colour}$.










How to construct a scatterplot using the ClassPad

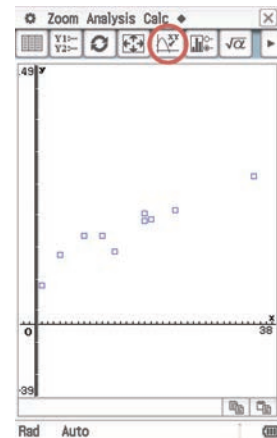
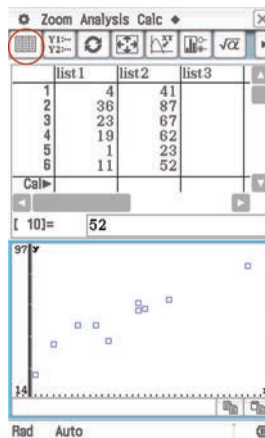
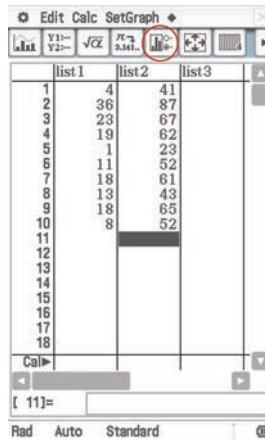
The data below shows the marks that students obtained on an examination and the times they spent studying for the examination.

<i>Time (hours)</i>	4	36	23	19	1	11	18	13	18	8
<i>Mark (%)</i>	41	87	67	62	23	52	61	43	65	52

Use a calculator to construct a scatterplot. Use *time* as the explanatory variable.

Steps

- 1 Open the **Statistics** application .
- 2 Enter the values into lists, with *time* in list1 and *mark* in list2.
- 3 Tap  to open the **Set StatGraphs** dialog box.
- 4 Complete the dialog box as shown and tap SET.
- 5 Tap  to plot a scaled graph in the lower half of the screen.
- 6 Tap  to give a full-screen sized graph. Tap  to return to a half-screen.
- 7 Tap  to place a marker on the first data point: ($x_c = 4, y_c = 41$).
- 8 Use the horizontal cursor arrow  to move from point to point.



Section Summary

- ▶ **Bivariate Data** arises when data on two different variables are collected for each individual or case.
- ▶ Usually, one of the two variables can be identified as the **explanatory** variable and the other as the **response** variable.
- ▶ The explanatory variable (EV) is the variable we expect to explain or predict the value of the response variable (RV).
- ▶ When both variables are numerical, bivariate data can be displayed in a **scatterplot**.
- ▶ When constructing a scatterplot, the EV is plotted on the horizontal axis, and the RV is plotted on the vertical axis.

Exercise 7A

Building understanding

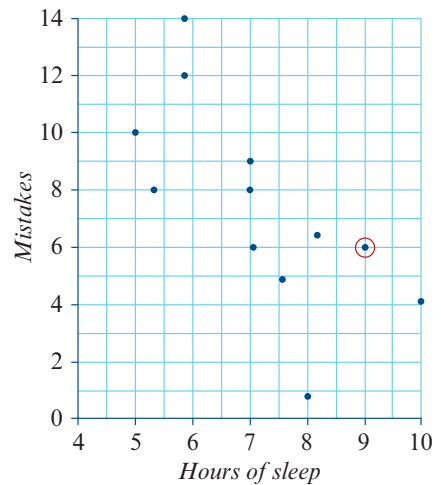
Example 1–3

- 1 Identify the EV and the RV in each of the following situations.
 - a We wish to predict the *diameter* of a certain type of tree from its *age*.
 - b The association between *weight loss* and the number of *weeks* a person is on a diet is to be studied.
 - c Data is collected to investigate the association between *age* of a second-hand textbook and its *selling price*.
 - d The association between the number of *hours* a gas heating system is used and the *amount* of gas used is to be investigated.
 - e A study is to be made of the association between the number of *runs* a cricketer scores and the number of *balls bowled* to them.
- 2 For which of the following pairs of variables would it be appropriate to construct a scatterplot to investigate a possible association?
 - a Car *colour* (blue, green, black, ...) and its *size* (small, medium, large)
 - b A food's *taste* (sweet, sour, bitter) and its *sugar content* (in grams)
 - c The *weight* (in kg) of 12 males and the *weight* (in kg) of 12 females
 - d The *time* people spend exercising each day, in minutes, and their *resting pulse rate* in beats per minute
 - e The *arm span* (in centimetres) and *gender* of a group of students

Example 4

3 This scatterplot has been constructed from data collected to investigate the association between the amount of sleep a person has the night before a test and the number of mistakes they make on the test. Use the scatterplot to answer the following questions.

- Which is the EV and which is the RV?
- How many people are in the data set?
- How much sleep did the individual circled have, and how many mistakes did they make?



Developing understanding

4 The table below shows the heights and weights of eight people.

<i>Height (cm)</i>	190	183	176	178	185	165	185	163
<i>Weight (kg)</i>	77	73	70	65	65	65	74	54

Use your calculator to construct a scatterplot with the variable *height* as the explanatory variable and the variable *weight* as the response variable.

5 The table below shows the ages of 11 couples when they got married.

<i>Age of wife</i>	26	29	27	21	23	31	27	20	22	17	22
<i>Age of husband</i>	29	43	33	22	27	36	26	25	26	21	24

Use your calculator to construct a scatterplot with the variable *wife* (age of wife) as the explanatory variable and the variable *husband* (age of husband) as the response variable.

6 The table below shows the number of seats and airspeeds (in km/h) of eight aircraft.

<i>Airspeed</i>	830	797	774	736	757	765	760	718
<i>Seats</i>	405	296	288	258	240	193	188	148

Use your calculator to construct a scatterplot with the variable *seats* as the explanatory variable and the variable *airspeed* as the response variable.

- 7 The table below shows the response times of 10 patients (in minutes) given a pain relief drug and the drug dosages (in milligrams).
- Which variable is the explanatory variable?
 - Use your calculator to construct an appropriate scatterplot.

<i>Drug dosage</i>	0.5	1.2	4.0	5.3	2.6	3.7	5.1	1.7	0.3	0.6
<i>Response time</i>	65	35	15	10	22	16	10	18	70	50

- 8 The table below shows the number of people in a cinema at 5-minute intervals after the advertisements started.

<i>Number in cinema</i>	87	102	118	123	135	137
<i>Time</i>	0	5	10	15	20	25

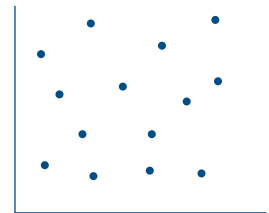
- Which is the explanatory variable?
- Use your calculator to construct an appropriate scatterplot.

7B How to interpret a scatterplot

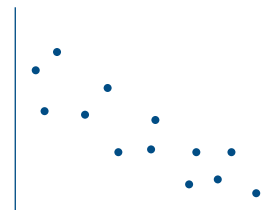
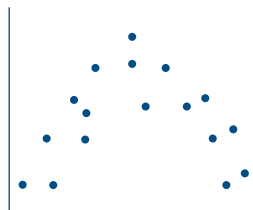
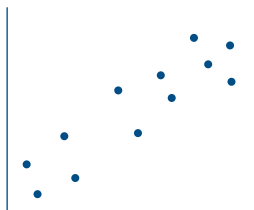
Learning intentions

- ▶ To be able to use a scatterplot to identify an association between two variables.
- ▶ To be able to use the scatterplot to classify an association according to:
 - ▷ **Direction**, which may be **positive** or **negative**
 - ▷ **Form**, which may be **linear** or **non-linear**
 - ▷ **Strength**, which may be **weak**, **moderate** or **strong**.

What features do we look for in a scatterplot to help us identify and describe any associations present? First, we look to see if there is a clear pattern in the scatterplot. In the scatterplot opposite, there is no clear pattern in the points. The points are randomly scattered across the plot, so we conclude that there is no association.



For the three examples below, there is a clear (but different) pattern in each set of points, so we conclude that there is an association in each case.

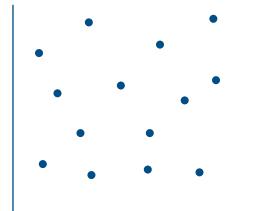
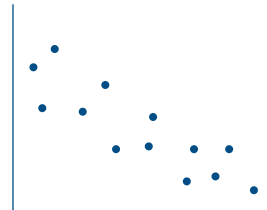
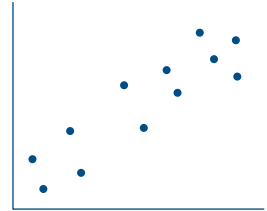


Having found a pattern, we need to be able to describe these associations clearly, as they are obviously quite different. The three features we look for in the pattern of points are **direction**, **form** and **strength**.

Direction of an association

We begin by looking at the overall pattern in the scatterplot.

- If the points in the scatterplot trend upwards as we go from left to right, we say there is a **positive association** between the variables. That is, the values of the explanatory variable and the response variable tend to increase together.
- If the points in the scatterplot trend downwards as we go from left to right, we say there is a **negative association** between the variables. That is, as the values of the explanatory variable increase, the values of the response variable tend to decrease.
- If there is no pattern in the scatterplot, that is, the points just seem to randomly scatter across the plot, we say there is **no association** between the variables.



In general terms, we can classify the direction of an association as follows.

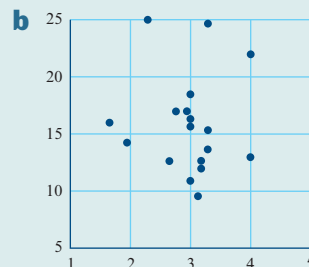
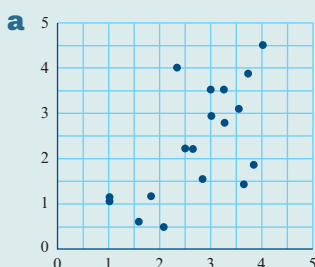
Direction of an association

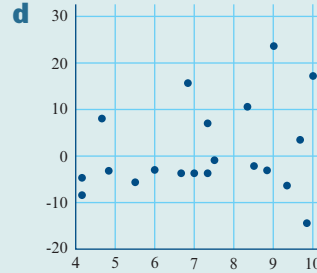
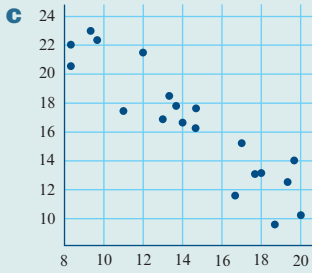
- Two variables have a **positive association** when the value of the response variable tends to increase as the value of the explanatory variable increases.
- Two variables have a **negative association** when the value of the response variable tends to decrease as the value of the explanatory variable increases.
- Two variables have **no association** when there is no consistent change in the value of the response variable when the value of the explanatory variable increases.



Example 5 Direction of an association

Classify each of the following scatterplots as exhibiting positive, negative or no association.





Explanation

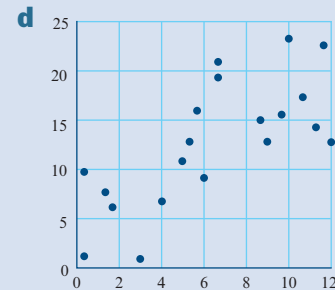
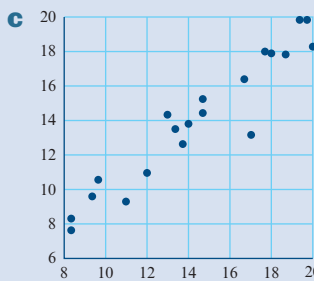
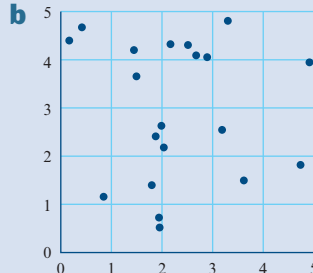
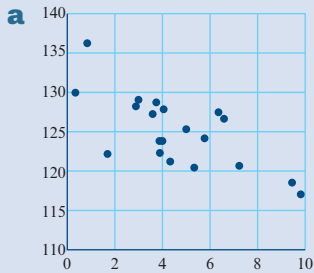
- a** The points trend *upwards* from left to right.
- b** There is *no clear pattern* in the scatterplot.
- c** The points trend *downwards* from left to right.
- d** There is *no clear pattern* in the scatterplot.

Solution

- The direction of the association is **positive**.
- The scatterplot shows **no** association.
- The direction of the association is **negative**.
- The scatterplot shows **no** association.

Now try this 5 Direction of an association (Example 5)

Describe the association in each of the following scatterplots.



Hint 1 Use the scatterplots in Example 5 as a guide.

Once we have identified the direction of an association, we can interpret this specifically in terms of the variables under investigation. So, for example:

- If there is a positive association between *height* and *weight*, then we can say that those individuals who are taller also tend to be heavier.
- If there is a negative association between *hours of sleep* and *reaction time*, then we can say that those individuals who have slept fewer hours tend to have slower reaction times.
- If there is no association between *height* and *reaction time*, then we are saying that the height of an individual does not seem to relate to their reaction time.



Example 6 Interpreting the direction of an association

Write a sentence interpreting each of the following associations:

- a There is a positive association between *study time* and *score on the exam*.
- b There is a negative association between *study time* and *time spent watching TV*.

Solution

- a Those people who spend more time studying tend to score higher marks on the exam.
- b Those people who spend more time studying tend to spend less time watching TV.

Now try this 6 Interpreting the direction of an association (Example 6)

Write a sentence interpreting each of the following associations:

- a There is a negative association between *height above sea level* and *temperature*.
- b There is a positive association between *number of years spent studying* and *salary*.

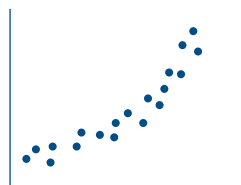
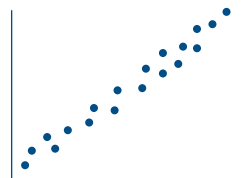
Hint 1 Use the wording of the answers in Example 6 as a model for your answers.

Form of an association

The next feature that interests us in an association is its general form. Do the points in a scatterplot tend to follow a linear pattern or a curved pattern?

For example:

- the association shown in the scatterplot opposite is **linear**.
We can imagine the points in the scatterplot to be scattered around some **straight line**.
- the association shown in the scatterplot opposite is **non-linear**.
We can imagine the points in the scatterplot to be scattered around a **curved line** rather than a straight line.



In general terms, we can describe the **form of an association** as follows.

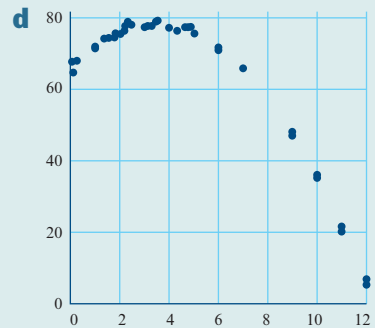
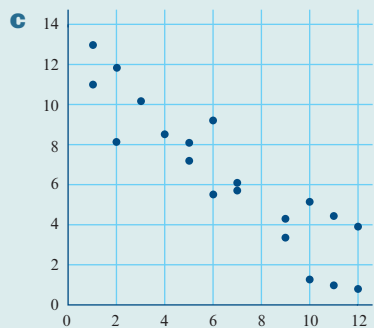
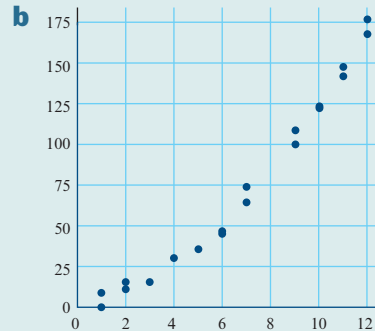
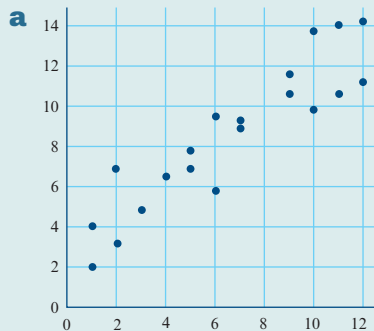
Form

A scatterplot is said to have a **linear form** when the points tend to follow a straight line. A scatterplot is said to have a **non-linear form** when the points tend to follow a curved line.



Example 7 Form of an association

Classify the **form** of the association in each of the following scatterplots as linear or non-linear.



Explanation

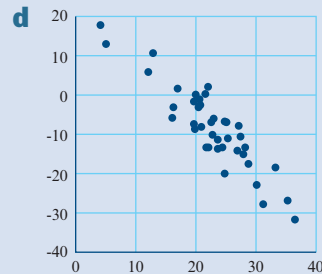
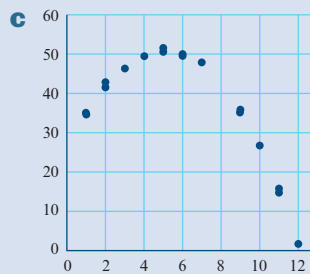
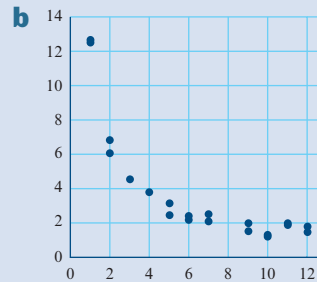
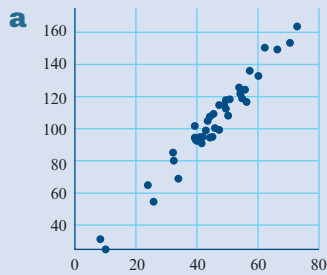
- a** There is a clear straight-line pattern.
- b** There is a clear curved pattern.
- c** There is a clear straight-line pattern.
- d** There is a clear curved pattern.

Solution

- The association is linear.
- The association is non-linear.
- The association is linear.
- The association is non-linear.

Now try this 7 Form of an association (Example 7)

Classify the *form* of the association in each of the following scatterplots.

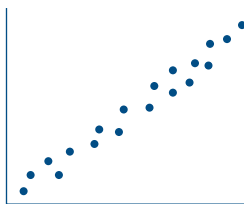


Hint 1 Use the scatterplots in Example 7 as a guide.

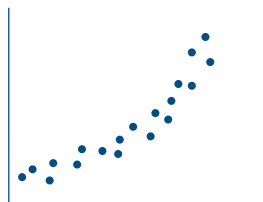
Strength of an association

The **strength of an association** is a measure of how much scatter there is in the scatterplot.

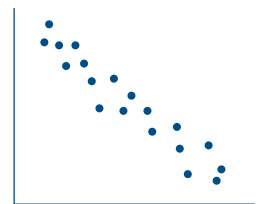
When there is a **strong association** between the variables, there is only a small amount of scatter in the plot, and a pattern is clearly seen.



Strong positive association

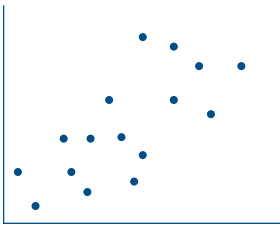


Strong positive association

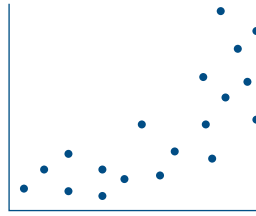


Strong negative association

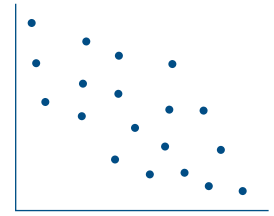
As the amount of scatter in the plot increases, the pattern becomes less clear. This indicates that the association is less strong. In the examples on the following page, we might say that there is a **moderate association** between the variables.



Moderate positive association

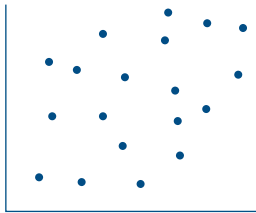


Moderate positive association

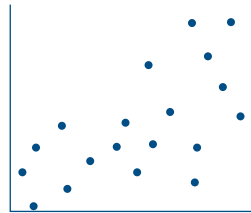


Moderate negative association

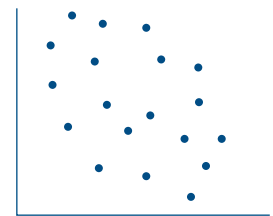
As the amount of scatter increases further, the pattern becomes even less clear. This indicates that any association between the variables is weak. The scatterplots below are examples of **weak associations** between the variables.



Weak positive association

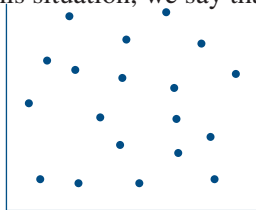


Weak positive association

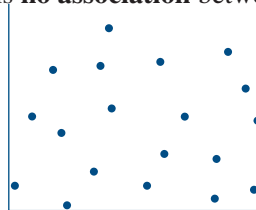


Weak negative association

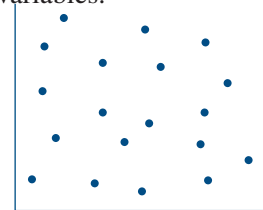
Finally, when all we have is scatter, as seen in the scatterplots below, no pattern can be seen. In this situation, we say that there is **no association** between the variables.



No association



No association



No association

In general terms, we can describe the **strength of an association** as follows.

Strength

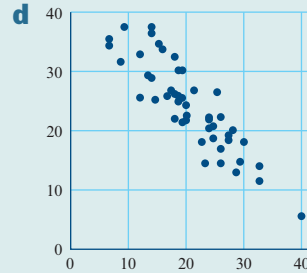
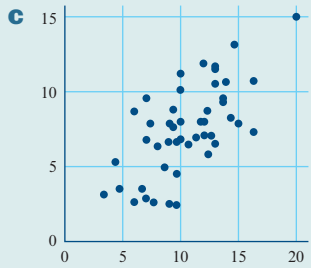
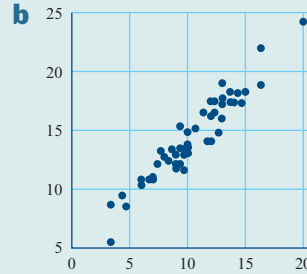
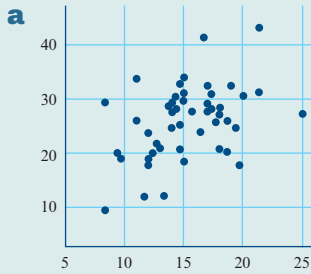
An association is classified as:

- **Strong** if the points on the scatterplot tend to be tightly clustered about a trend line.
- **Moderate** if the points on the scatterplot tend to be moderately clustered about a trend line.
- **Weak** if the points on the scatterplot tend to be loosely clustered about a trend line.



Example 8 Strength of an association

Classify the **strength** of the association in each of the following scatterplots as strong, moderate or weak.



Explanation

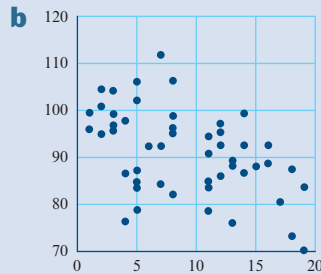
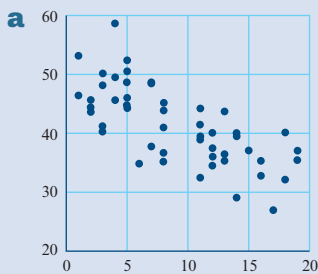
- a** The points are loosely clustered.
- b** The points are tightly clustered.
- c** The points are moderately clustered.
- d** The points are tightly clustered.

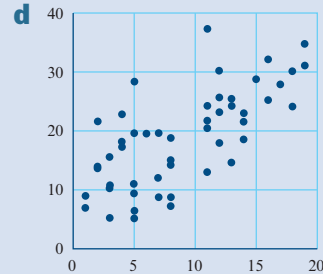
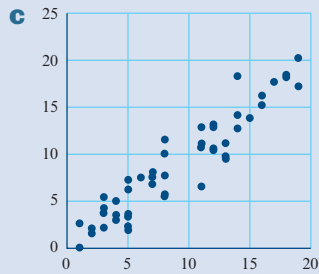
Solution

- The association is weak.
- The association is strong.
- The association is moderate.
- The association is strong.

Now try this 8 Strength of an association (Example 8)

Classify the *strength* of the association in each of the following scatterplots as strong, moderate or weak.





Hint 1 Use the scatterplots in Example 8 as a guide.

At the moment, you only need to be able to estimate the strength of an association, as strong, moderate, weak or none, by comparing it with the standard scatterplots given. However, the previous examples will have shown you that it is sometimes quite difficult to judge the difference between the strength of these associations by merely looking at the scatterplot. In the next section, you will learn about a statistic, the **correlation coefficient**, which can be used to give a value to the strength of linear association from the data values.

Section Summary

From a scatterplot, we can describe key features of a bivariate association.

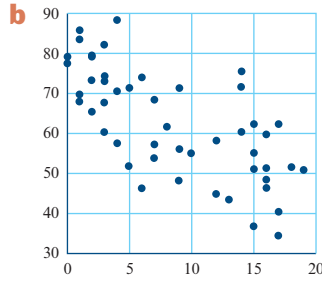
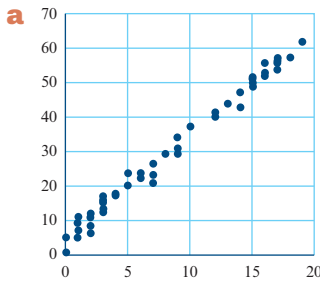
- ▶ **Direction.** The two variables in the scatterplot have:
 - ▷ a positive association when the value of the response variable tends to increase as the value of the explanatory variable increases,
 - ▷ a negative association when the value of the response variable tends to decrease as the value of the explanatory variable increases,
 - ▷ no association when there is no consistent change in the value of the response variable when the values of the explanatory variable increase.
- ▶ **Form.** An association is classified as:
 - ▷ linear when the points tend to follow a straight line,
 - ▷ non-linear when the points tend to follow a curved line.
- ▶ **Strength.** An association is classified as:
 - ▷ Strong if the points on the scatterplot tend to be tightly clustered about a trend line,
 - ▷ Moderate if the points on the scatterplot tend to be moderately clustered about a trend line,
 - ▷ Weak if the points on the scatterplot tend to be loosely clustered about a trend line.

Exercise 7B

Building understanding

Example 5

1 Classify each of the following scatterplots according to **direction** (positive or negative).



Example 7

2 Classify each of the scatterplots in Question 1 according to **form** (linear or non-linear).

Example 8

3 Classify each of the scatterplots in Question 1 according to **strength** (weak, moderate or strong).

Developing understanding

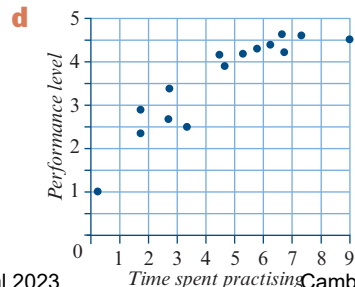
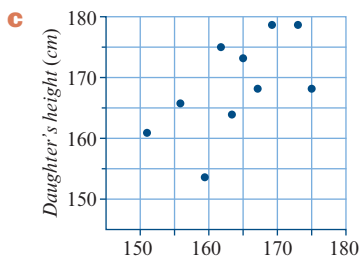
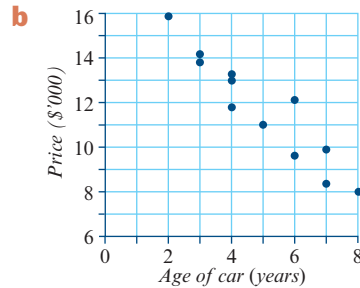
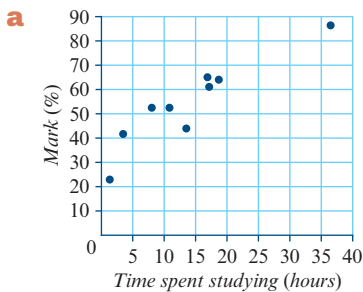
Example 6

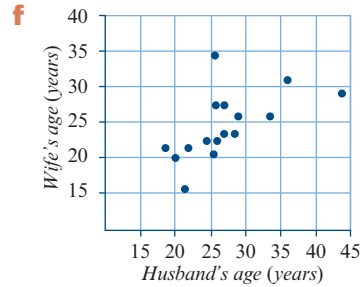
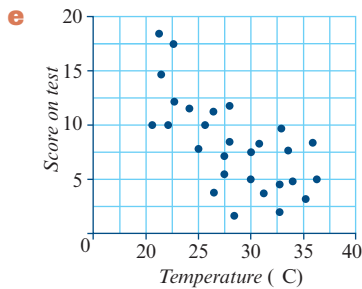
4 Write a sentence interpreting each of the following associations:

- a** There is a positive association between *fitness level* and *amount* of daily exercise.
- b** There is a negative association between *time* taken to run a marathon and *speed* of the runner.

5 The variables in each of the following scatterplots are associated. In each case:

- i** Describe the association in terms of its direction (positive/negative), form (linear/non-linear) and strength (strong/moderate/weak).
- ii** Write a sentence describing the direction of the association in terms of the variables in the scatterplot.





Testing understanding

- 6** In a mathematics class, a group of students were asked to draw circles on squared paper. They measured the diameter of the circles they had drawn, and then estimated the areas of the circles by counting the squares. Their results are given in the following table:

<i>diameter (cm)</i>	3.5	6.2	5.4	3.7	7.3	8.6	3.7	2.9	2.1	9.7	3.7
<i>area (cm²)</i>	9.5	30.0	22.7	10.2	42.6	57.7	10.5	5.7	2.7	74.4	11.0

- Use your calculator to construct a scatterplot of the data, with the variable *diameter* as the explanatory variable and the variable *area* as the response variable.
- Describe the scatterplot in terms of direction, form and strength.
- Create a new column of data in your calculator by squaring values of the diameter (that is, $\text{diameter} \times \text{diameter}$). Construct another scatterplot of the data, this time with the variable diameter^2 as the explanatory variable and *area* as the response variable.
- Describe the second scatterplot in terms of direction, form and strength.
- What has been the effect on the scatterplot of using diameter^2 rather than *diameter* as the explanatory variable?

7C Pearson's correlation coefficient (r)

Learning intentions

- ▶ To be able to understand Pearson's correlation coefficient, r , as a measure of the strength of a linear association between two variables.
- ▶ To be able to use technology to find the value of Pearson's correlation coefficient, r .
- ▶ To be able to classify the strength of a linear association as weak, moderate or strong, based on the value of Pearson's correlation coefficient, r .
- ▶ To be able to define and differentiate the concepts of association and causation.

When an association is linear, the most commonly used measure of strength of the association is Pearson's correlation coefficient, r . It gives a numerical measure of the degree to which the points in the scatterplot tend to cluster around a straight line.

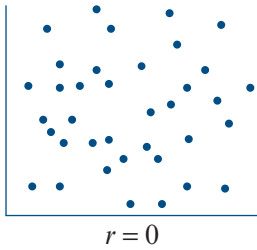
There are two key assumptions when using Pearson’s correlation coefficient, r . These are:

- the data from both variables are numerical
- the association is linear.

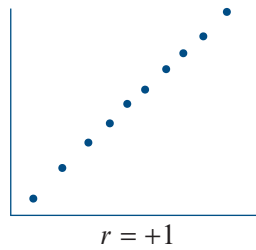
Properties of Pearson’s correlation coefficient (r)

Pearson’s correlation coefficient has the following properties:

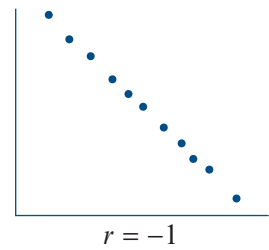
- no linear association,
 $r = 0$



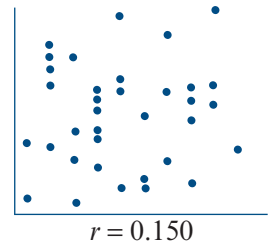
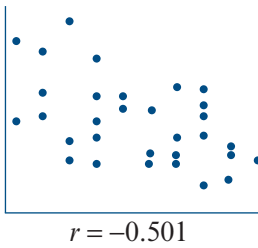
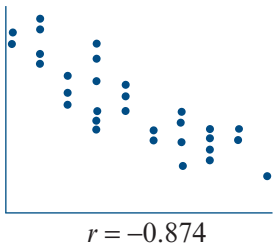
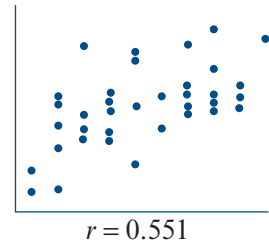
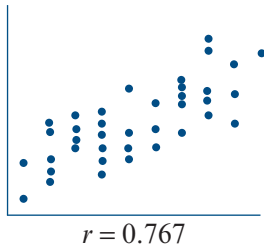
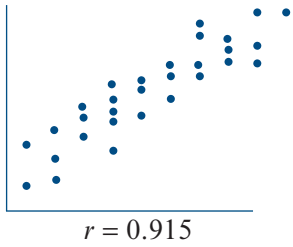
- a perfect positive linear association,
 $r = +1$



- a perfect negative linear association,
 $r = -1$.



In practice, the value of r will be somewhere between $+1$ and -1 and rarely, exactly zero, as shown in the selection of scatterplots below.



These scatterplots illustrate an important point – the stronger the association, the larger the magnitude of Pearson’s correlation coefficient.

Pearson's correlation coefficient

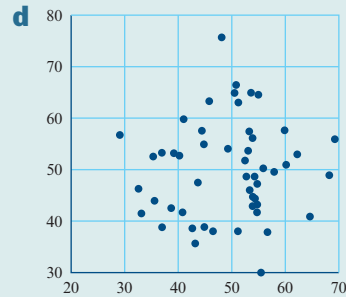
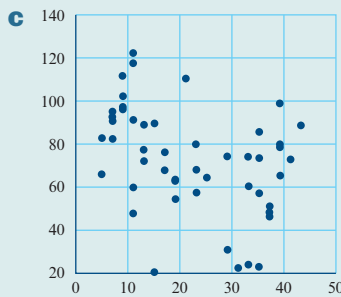
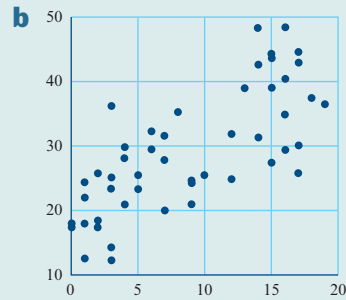
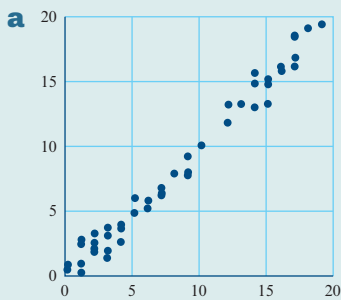
The Pearson's correlation coefficient, r :

- measures the **strength** of a **linear association**, with larger values indicating stronger relationships
- has a value between -1 and $+1$
- is positive if the direction of the linear association is positive
- is negative if the direction of the linear association is negative
- is close to zero if there is no association.



Example 9 Estimating the correlation coefficient from a scatterplot

Estimate the value of the correlation coefficient, r , in each of the following plots, using the plots on page 420 as a guide.



Explanation

- a** Points are tightly clustered, similar to Plot 1, and the direction is positive.
- b** Points look to be more loosely clustered than Plot 2 but not as loose as Plot 3, and the direction is positive.
- c** Points look to be slightly more loosely clustered than Plot 5, and the direction is negative.
- d** The points look random as in Plot 6.

Solution

Estimate: $r \approx 0.9$

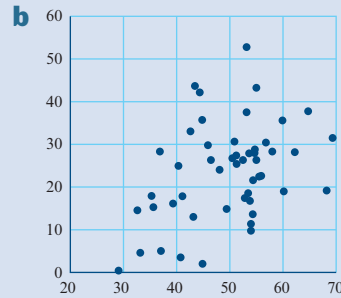
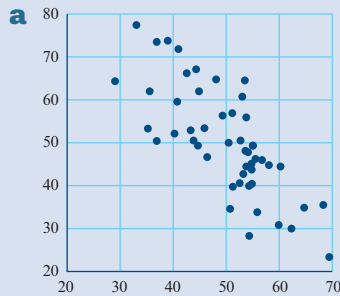
Estimate: $r \approx 0.7$

Estimate: $r \approx -0.4$

Estimate: $r \approx 0$

Now try this 9**Estimating the correlation coefficient from a scatterplot (Example 9)**

Estimate the value of the correlation coefficient, r , in each of the following plots, using the plots on page 420 as a guide.



Hint 1 Firstly decide whether the value of r is positive or negative.

Hint 2 Then use the plots on page 420 to estimate the value of r .

Determining the value of Pearson's correlation coefficient, r

The formula for calculating r is:

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

In this formula, \bar{x} and s_x are the mean and standard deviation of the x scores, and \bar{y} and s_y are the mean and standard deviation of the y scores.

After the mean and standard deviation, Pearson's correlation coefficient is one of the most frequently computed descriptive statistics. The presence of a linear association should always be confirmed with a scatterplot before Pearson's correlation coefficient is calculated. And, like the mean and the standard deviation, Pearson's correlation coefficient can be very sensitive to the presence of outliers, particularly for small data sets.

Pearson's correlation coefficient, r , is rather tedious to calculate by hand and is usually evaluated with the aid of technology.

How to calculate Pearson's correlation coefficient, r , using the TI-Nspire CAS

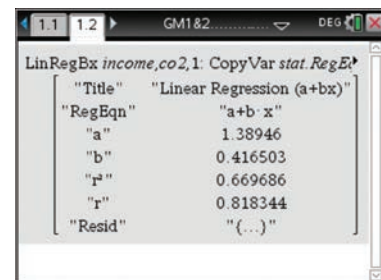
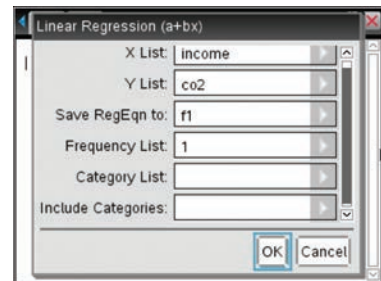
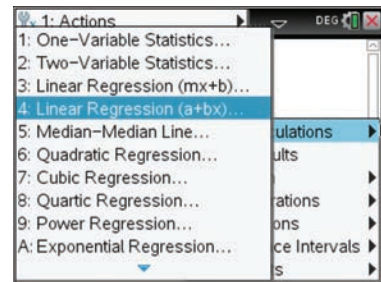
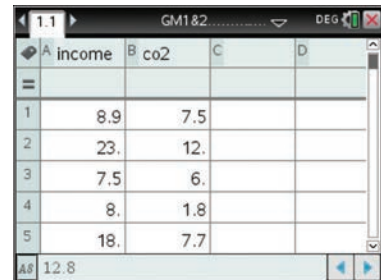
The following data shows the per capita income (in \$'000) and the carbon dioxide emissions (in tonnes) of 11 countries.

Determine the value of Pearson's correlation coefficient, r , for these data.

<i>Income</i> (\$'000)	8.9	23.0	7.5	8.0	18.0	16.7	5.2	12.8	19.1	16.4	21.7
<i>CO₂</i> (tonnes)	7.5	12.0	6.0	1.8	7.7	5.7	3.8	5.7	11.0	9.7	9.9

Steps

- 1 Start a new document by pressing $\text{ctrl} + \text{N}$.
- 2 Select **Add Lists & Spreadsheet**.
Enter the data into lists named *income* and *co2*.
- 3 Statistical calculations can be done in the Calculator application. Press $\text{ctrl} + \text{I}$ and select **Calculator**.
- 4 Press $\text{menu} > \text{Statistics} > \text{Stat Calculations} > \text{Linear Regression (a + bx)}$ to generate the screen opposite.
- 5 Press menu to generate the pop-up screen as shown. To select the variable for the X List entry, use \blacktriangleright and enter to select and paste in the list name, *income*. Press tab to move to the Y List entry, use $\blacktriangleright \blacktriangledown$ and enter to select and paste in the list name, *co2*.
- 6 Press enter to exit the pop-up screen, and generate the results shown in the screen opposite.
- 7 The value of the correlation coefficient is $r = 0.818344\dots$ or 0.818, rounded to three decimal places.




How to calculate Pearson's correlation coefficient, r , using the ClassPad

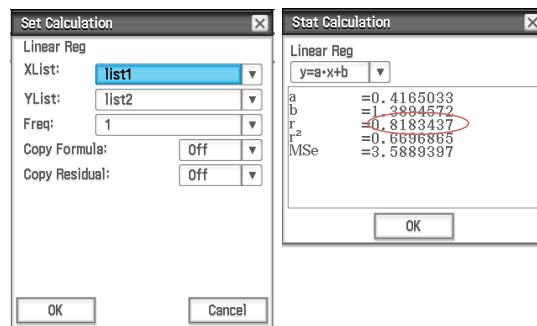
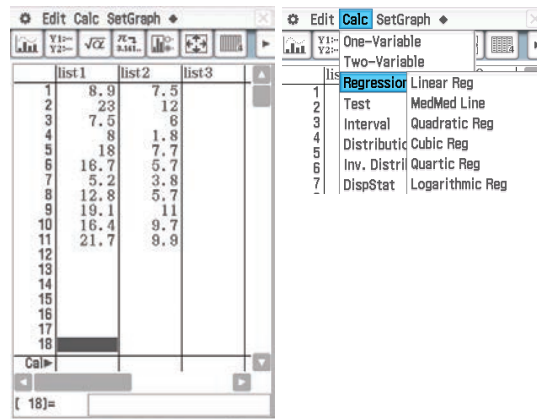
The following data shows the per capita income (in \$'000) and the carbon dioxide emissions (in tonnes) of 11 countries.

Determine the value of Pearson's correlation coefficient, r , for the given data.

<i>Income (\$'000)</i>	8.9	23.0	7.5	8.0	18.0	16.7	5.2	12.8	19.1	16.4	21.7
<i>CO₂ (tonnes)</i>	7.5	12.0	6.0	1.8	7.7	5.7	3.8	5.7	11.0	9.7	9.9

Steps

- Open the **Statistics** application .
- Enter the data into the columns.
 - Income in List1
 - CO₂ in List2
- Select **Calc>Regression>Linear Reg** from the menu bar.
- Press **EXE**.
This opens the **Set Calculation** dialog box, as shown to the right.
- Tap **OK** to confirm your selections.
- The value of the correlation coefficient is $r = 0.818344\dots$ or 0.818, rounded to three decimal places.




Example 10 Calculating the correlation coefficient using a calculator

Scores in two tests for a group of ten students are given in the following table. Determine the value of the correlation coefficient, r , for these data, rounded to four decimal places.

Score test 1 (30)	14	17	26	17	15	13	29	25	17	30
Score test 2 (20)	9	11	15	13	10	9	16	14	12	19

Explanation

- 1 Enter the data into lists named *test1* and *test2*.
- 2 Determine the value of r following the instructions for your calculator.

Solution

$$r=0.9499$$

Now try this 10 Calculating the correlation coefficient using a calculator (Example 10)

The hours spent studying for each of two tests by a group of students are given in the following table. Determine the value of the correlation coefficient, r , for these data, rounded to four decimal places.

Hours studying for test 1	9	13	7	2	8	7	6	3	10	6
Hours studying for test 2	7	12	6	2	8	7	5	6	11	8

Hint 1 Always begin by entering the data into named lists.

Hint 2 Carefully follow the instructions for your calculator.

Guidelines for classifying the strength of a linear association

Pearson's correlation coefficient, r , can be used to classify the strength of a linear association as follows:

$0.75 \leq r \leq 1$	strong positive association
$0.5 \leq r < 0.75$	moderate positive association
$0.25 \leq r < 0.5$	weak positive association
$-0.25 < r < 0.25$	no association
$-0.5 < r \leq -0.25$	weak negative association
$-0.75 < r \leq -0.5$	moderate negative association
$-1 \leq r \leq -0.75$	strong negative association


Example 11 Classifying the strength of a linear association

Classify the strength of each of the following linear associations using the previous table.

a $r = 0.35$

b $r = -0.507$

c $r = 0.992$

d $r = -0.159$

Explanation

a The value 0.35 is more than 0.25 and less than 0.5. That is, $0.25 \leq r < 0.5$.

b The value -0.507 is more than -0.75 and less than -0.5 . That is, $-0.75 < r \leq -0.5$.

c The value 0.992 is more than 0.75 and less than 1. That is, $0.75 \leq r \leq 1$.

d The value -0.159 is more than -0.25 and less than 0.25. That is, $-0.25 < r < 0.25$.

Solution

weak, positive

moderate, negative

strong, positive

no association

Now try this 11 Classifying the strength of a linear association (Example 11)

Classify the strength of each of the following linear associations.

a $r = 0.807$

b $r = -0.818$

c $r = 0.224$

d $r = -0.667$

Hint 1 In each part, compare the value of r to the interval given in the table on the previous page.

Hint 2 Remember: the sign of the correlation coefficient tells you the direction of the association, and the value tells you the strength.

Correlation and causation

A strong correlation between two variables means that they vary together, both increasing together if the correlation is positive, or one decreasing as the other increases if the correlation is negative. The existence of even a strong correlation between two variables is not, in itself, sufficient to imply that altering one variable **causes a change** in the other. It only implies that this **may** be the explanation.

Suppose, for example, we were to find a high correlation between the smoking rate and the incidence of heart disease across a group of countries. We cannot conclude from this correlation coefficient alone that smoking **causes** heart disease. Another possible explanation is that people who smoke neglect lifestyle factors such as exercise and diet. It could well be that people who smoke also tend not to exercise regularly, and it is the lack of exercise which causes heart disease.

- A **correct** interpretation of a high correlation between smoking rate and heart disease across a group of countries would be: "Those countries which have higher rates of smoking also tend to have higher incidence of heart disease".
- An **incorrect** interpretation of a high correlation between smoking rate and heart disease across a group of countries would be: "As the smoking rate increases then the incidence of heart disease will also increase". Also **incorrect** would be to state: "Reducing the smoking rate would also reduce the incidence of heart disease".

It is for this reason that we need to be very careful when interpreting the correlation coefficient.



Example 12 Correlation and causation

The correlation coefficient, r , between the per capita income (in \$'000) and the carbon dioxide emissions (in tonnes) of 11 countries is equal to 0.818. Does this mean that reducing the per capita income would result in decreased carbon dioxide emissions?

Explanation

We cannot **infer** (which means deduce or conclude) causation, even when there is a strong correlation.

Solution

No, we can only conclude that those countries with higher per capita income also tend to have higher carbon dioxide emissions.

Now try this 12 Correlation and causation (Example 12)

If the heights and the scores obtained on a test of mathematical ability by a group of primary school students in Year Prep to Year 6 were recorded, a strong correlation would be found. Can it be inferred from this that taller people are better at mathematics? Give a possible non-causal explanation.

Section Summary

- ▶ **Pearson's correlation coefficient**, r , is a measure of the strength of a linear association.
- ▶ Assumptions when using Pearson's correlation coefficient, r , are:
 - ▶ the data from both variables are **numerical**
 - ▶ the association is **linear**.
- ▶ Pearson's correlation coefficient, r :
 - ▶ has a value between -1 and $+1$, with larger values indicating stronger associations
 - ▶ is close to zero if there is **no** association
 - ▶ is **positive** if the direction of the linear association is positive
 - ▶ is **negative** if the direction of the linear association is negative.

Section Summary

- The strength of the correlation coefficient is classified according to the following table:

$0.75 \leq r \leq 1$	strong positive association
$0.5 \leq r < 0.75$	moderate positive association
$0.25 \leq r < 0.5$	weak positive association
$-0.25 < r < 0.25$	no association
$-0.5 < r \leq -0.25$	weak negative association
$-0.75 < r \leq -0.5$	moderate negative association
$-1 \leq r \leq -0.75$	strong negative association

- The existence of even a strong correlation between two variables is not sufficient to conclude that there is a causal association between them.



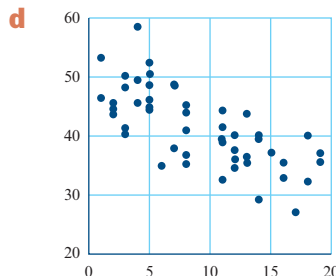
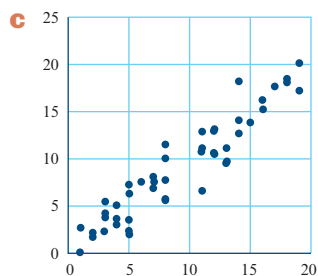
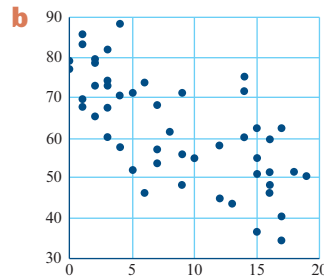
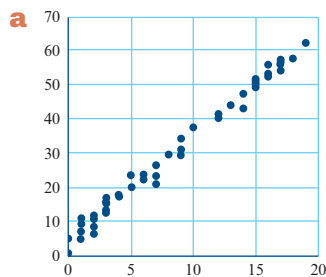
Exercise 7C

Building understanding

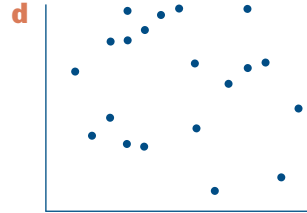
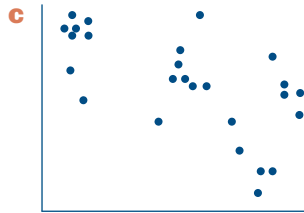
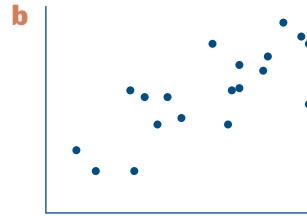
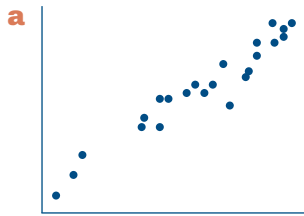
- 1 What are the two key assumptions justifying the use of Pearson's correlation coefficient to quantify the strength of the association between two variables?

Example 9

- 2 Estimate the value of the correlation coefficient, r , in each of the following plots, using the plots on page 420 as a guide.



- 3 Estimate the value of the correlation coefficient, r , in each of the following plots, using the plots on page 420 as a guide.



Developing understanding

Example 10

- 4 Determine the value of Pearson's correlation coefficient, r , for the data in the following table. Give your answer rounded to three decimal places.

X	7	3	7	12	8	17	6	3	10	6
Y	10	2	6	12	8	7	8	6	11	8

- 5 The table below shows the *weight* (in kg) and blood *glucose* level (in mg/100 mL) of eight adults.

<i>Weight</i>	82.1	70.1	76.6	82.1	83.9	73.2	66.0	77.5
<i>Glucose</i>	101	89	98	100	108	104	94	89

Use your calculator to determine the value of Pearson's correlation coefficient for this data set. Give your answer rounded to three decimal places.

- 6 The table below shows the scores which a group of nine students obtained on two class tests, *Test 1* and *Test 2*, as part of their school-based assessment.

<i>Test 1</i>	33	45	27	42	50	38	17	35	29
<i>Test 2</i>	43	46	36	34	48	34	29	41	28

Use your calculator to determine the value of Pearson's correlation coefficient for this data set. Write your answer rounded to three decimal places.

- 7 The table below shows the carbohydrate content (*carbs*) and the fat content (*fat*) in 100 g of nine breakfast cereals.

<i>Carbs (g)</i>	88.7	67.0	77.5	61.7	86.8	32.4	72.4	77.1	86.5
<i>Fat (g)</i>	0.3	1.3	2.8	7.6	1.2	5.7	9.4	10.0	0.7

Use your calculator to determine the value of Pearson's correlation coefficient for this data set. Give your answer rounded to three decimal places.

Example 11

- 8 Use the guidelines on page 428 to classify the strength of a linear association for which Pearson's correlation coefficient is calculated to be:

- a** $r = 0.205$ **b** $r = -0.303$ **c** $r = -0.851$ **d** $r = 0.333$
e $r = 0.952$ **f** $r = -0.740$ **g** $r = 0.659$ **h** $r = -0.240$
i $r = -0.484$ **j** $r = 0.292$ **k** $r = 1$ **l** $r = -1$

Example 12

- 9 There is a strong positive correlation between the number of bars and the number of school teachers in cities around the world. Can we conclude from this that school teachers spend a lot of time in bars? Give a possible non-causal explanation.
- 10 There is a strong negative correlation between birth rate and life expectancy in a country. Can we conclude that decreasing the birth rate in a country will help increase the life expectancy of its citizens? Give a possible non-causal explanation.



- 11** In a survey of nine problem gamblers, the respondents were asked the *amount* (in dollars) they had spent on gambling and the *number of hours* they had spent gambling in the past week. The data collected is recorded in the table below.

<i>Hours</i>	10	11	12	15	20	21	25	35	40
<i>Amount</i>	500	530	300	750	1000	1200	2000	2300	5000

- a** The aim is to predict the amount of money spent on gambling from the time spent gambling. Which is the explanatory variable and which is the response variable?
- b** Construct a scatterplot of these data.
- c** Determine the value of the correlation coefficient, r , to three decimal places.
- d** Describe the association between the variables *amount* and *hours* in terms of strength, direction and form.
- 12** The following data was recorded through the National Health Survey:

<i>Region</i>	<i>Percentage with eye disease</i>	
	<i>Male (%)</i>	<i>Female (%)</i>
Australia	40.7	49.1
Other Oceania countries	46.1	66.2
United Kingdom	74.5	75.0
North-West Europe	71.2	71.5
Southern & Eastern Europe	71.6	74.6
North Africa & the Middle East	52.2	57.5
South-East Asia	47.7	54.8
All other countries	56.0	62.0

- a** Which is the explanatory variable and which is the response variable?
- b** Construct a scatterplot of these data, with *percentage of males* on the horizontal axis and *percentage of females* on the vertical axis.
- c** Determine the value of the correlation coefficient, r , to three decimal places.
- d** Describe the association between the male and female eye disease percentages for these countries in terms of strength, direction and form.

Testing understanding

- 13** The following table gives the educational level (*education*), the number of years the person has worked for the company (*years*) and their current salary to the nearest thousand dollars (*salary*) for a group of current employees of a particular company.

<i>education</i>	<i>years</i>	<i>salary</i>	<i>education</i>	<i>years</i>	<i>salary</i>
Secondary	2	52	Tertiary	2	62
Secondary	3	64	Tertiary	3	69
Secondary	2	56	Tertiary	4	75
Secondary	4	63	Tertiary	5	76
Secondary	7	65	Tertiary	4	72
Secondary	6	64	Tertiary	4	68
Secondary	7	52	Tertiary	1	63
Secondary	10	65	Tertiary	8	85
Secondary	5	59	Tertiary	3	67
Secondary	5	62	Tertiary	6	77

- a**
- Use your calculator to construct a scatterplot of the data for all 20 employees, with the variable *salary* as the response variable and the variable *years* as the explanatory variable.
 - Determine the value of the correlation coefficient, r , to three decimal places.
- b**
- Use your calculator to construct a scatterplot of *salary* against *years* for those employees who have Secondary educational level.
 - Determine the value of the correlation coefficient, r , for these employees to three decimal places.
- c**
- Use your calculator to construct a scatterplot of *salary* against *years* for those employees who have Tertiary educational level.
 - Determine the value of the correlation coefficient, r , for these employees to three decimal places.
- d** Based on the values of the three correlation coefficients determined in parts **a**, **b** and **c**, how would you describe the association between *salary* and *years* for the employees of this company?

7D Fitting a linear model to the data

Learning intentions

- ▶ To be able to fit a linear model to the data **by eye**.
- ▶ To be able to determine the equation of the line fitted by eye from the graph.
- ▶ To be able to use the method of **least squares** for fitting a linear model to the data.

- ▶ To be able to calculate the intercept and slope of the least squares line from the correlation coefficient and summary statistics.
- ▶ To be able to use a CAS calculator to determine the intercept and slope of the least squares line from the bivariate data.

Once we identify a linear association between two numerical variables, we can go one step further by fitting a linear model to the data and finding its equation. This model gives us a better understanding of the association between the two variables and allows us to make predictions. The process of modelling an association with a straight line is known as **linear regression**, and the resulting line is often called the **regression line** or line of good fit.

As discussed in Chapter 5, the equation of the regression line is given by the rule:

$$y = a + bx$$

where a is the y -intercept and b is the slope.

Fitting a line ‘by eye’

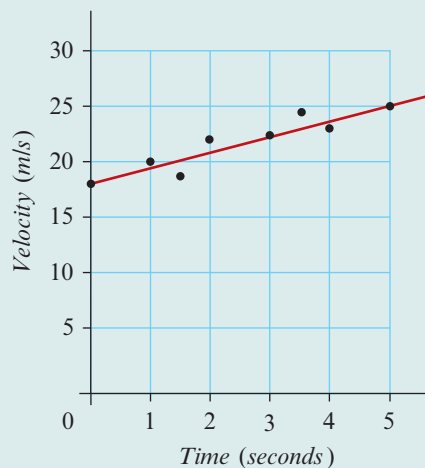
The simplest method of fitting a line is to use a ruler to draw a line on the scatterplot that seems to balance out the points around the line. This is called **fitting a line ‘by eye’**. Once the line has been drawn on the scatterplot, then its equation can be determined from the plot, using the skills that you developed in Chapter 5F, as shown in the following examples.



Example 13 Fitting a line by eye using the intercept and slope

A straight line has been fitted by eye to a set of data that records the velocity, v , (in m/s) of an accelerating car at time, t , seconds.

- a** What is the car’s velocity when *time* = 0?
- b** What is the slope of the line?
- c** Write down the equation of the line in terms of *velocity* and *time*.



Explanation

- a** Where *time* = 0, the y -intercept can be read from the graph.
- b** Calculate the slope by using two points on the graph, say (0, 18) and (5, 25). Any two points can be used.
- c** The general equation of the line is of the form $y = a + bx$.

Solution

y -intercept $a \approx 18$ m/s

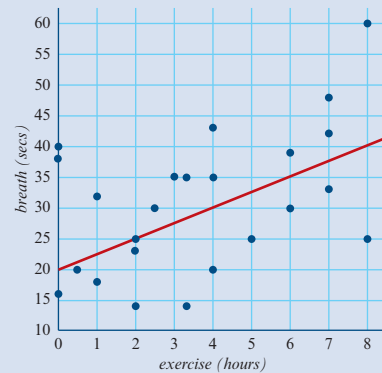
$$\text{slope } b = \frac{\text{rise}}{\text{run}} = \frac{25 - 18}{5 - 0} = 1.4$$

$$\text{velocity} = 18 + 1.4 \times \text{time}$$

Now try this 13 Fitting a line by eye using the intercept and slope (Example 13)

A straight line has been fitted by eye to a scatterplot showing how long a group of students could hold their breath for, in seconds, (*breath*) and the number of hours they spend each week in exercise (*exercise*).

- What is the value of *breath* when *exercise* = 0?
- What is the slope of the line?
- Write down the equation of the line in terms of *exercise* and *breath*.

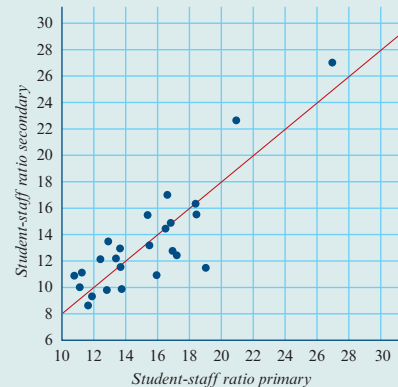


Hint 1 Remember to label the intercept as a and the slope as b .

**Example 14** Fitting a line by eye using two points on the graph

A straight line has been fitted by eye to a scatterplot of the *student-staff ratio for secondary* against the *student-staff ratio for primary* for a group of countries.

Determine the equation of this line in terms of the variables in the question.

**Explanation**

- Find two points on the line where the coordinates of the points can be read easily from the graph. There may be a few, any two points will do.
- Calculate the slope (b) by using the coordinates of the two points.
- To find the value of a , substitute the coordinates of one of the points on the line (either will do) into the general rule $y = a + bx$ and solve for a .
- Write down the equation of the line in terms of the variables in the question.

Solution

Suitable points are (14, 12) and (30, 28).

$$b = \frac{\text{rise}}{\text{run}} = \frac{28 - 12}{30 - 14} = 1.0$$

Using the point (14, 12)

$$12 = a + 1 \times 14$$

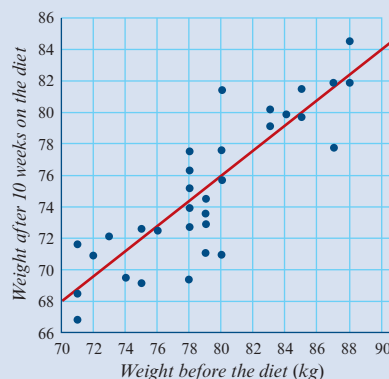
$$\therefore a = 12 - 14 = -2$$

$$\text{student-staff ratio secondary} = -2 + 1 \times \text{student-staff ratio primary}$$

Now try this 14 Fitting a line by eye using two points on the graph (Example 14)

A straight line has been fitted by eye to a scatterplot showing the weights for a group of males before and after they spent 10 weeks on a weight reduction diet.

Determine the equation of this line in terms of the variables in the question.



Hint 1 Find two points on the line where the coordinates are clear.

Hint 2 Find the slope, b , first.

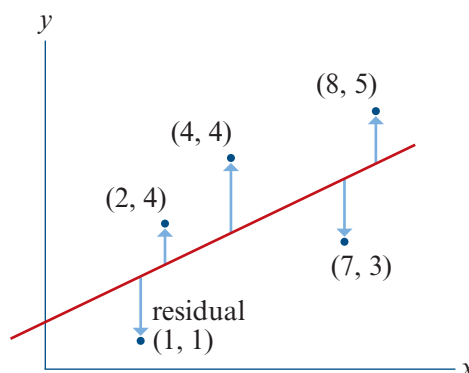
Hint 3 Use either point to substitute into the general rule to find the value of a .

Using the least squares line to model a linear association

A better approach to fitting a line is to use a mathematical strategy known as the **least squares method**.

It is very unlikely that all of the points in the scatterplot will lie exactly on a straight line, so our goal is to find a line that best fits the data in some way. To do this, we start by considering the vertical distances between the line and the actual data points. These are called **residuals** and are shown by the **blue arrows** in the scatterplot below.

One way of fitting the line is to find the line that has the smallest value for the sum of the residuals. However, because some residuals will be positive (because the point is *above* the line) and some will be negative (because the point is *below* the line) then these residuals will tend to cancel out.



We have met this problem before when calculating standard deviation, and we solve it the same way, by squaring the residuals to make them all positive, and then they can be added together.

The least squares line is the equation of the line that minimises the sum of these squared residuals (hence the name of the method: least squares). The exact solution for these values can be found mathematically, using the techniques of calculus; however, this is beyond the mathematics required for General Mathematics.

The equation of the least squares regression line

The equation of the least squares regression line is given by, $y = a + bx$, where:

$$\text{the slope } (b) \text{ is given by: } b = \frac{rs_y}{s_x}$$

and

$$\text{the intercept } (a) \text{ is then given by: } a = \bar{y} - b\bar{x}$$

- r is the correlation coefficient
- s_x and s_y are the standard deviations of x and y
- \bar{x} and \bar{y} are the mean values of x and y .

The assumptions made in using the least squares method to model a linear association are the same as those for Pearson's correlation coefficient. These are:

- the variables are numerical
- the association is linear.



Example 15 Using the formula to find the intercept and slope of the least squares line

Find the equation of the least squares regression line, $y = a + bx$, when:

$$r = 0.600 \quad s_x = 9.80 \quad s_y = 5.40 \quad \bar{x} = 165.0 \quad \bar{y} = 62.3$$

Give the values of the intercept and slope, rounded to three **significant figures**.

Note: To revise significant figures and how to calculate them, see Chapter 10, section 10A.

Explanation

- 1** First, substitute in the formula for b to find the slope.
- 2** Next, substitute in the formula for a to find the intercept.
- 3** Substitute a and b in the formula for a straight line, giving values rounded to three significant figures.

Solution

$$b = \frac{rs_y}{s_x} = \frac{0.600 \times 5.40}{9.80} = 0.3306$$

$$a = \bar{y} - b\bar{x} = 62.3 - 0.3306 \times 165.0 = 7.751$$

$$y = 7.75 + 0.331x$$

Now try this 15 Using the formula to find the intercept and slope of the least squares line (Example 15)

Find the equation of the least squares regression line, $y = a + bx$, when:

$$r = 0.700 \quad s_x = 2.30 \quad s_y = 3.40 \quad \bar{x} = 15.2 \quad \bar{y} = 24.5$$

Give the values of the intercept and slope, rounded to three significant figures.

Hint 1 You must always start by finding the value of b (since this is required in the formula for a).

Hint 2 Work with more significant figures than required, and only round to the number asked for in the answer.

Your CAS calculator can also be used to find the equation for the least squares regression line from the data, as in the following examples.

How to determine and graph the least squares regression line using the TI-Nspire CAS

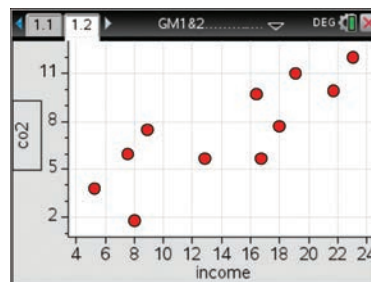
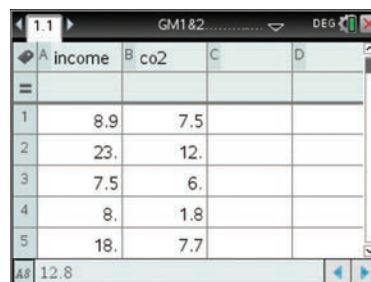
The following data shows the per capita income (*income*) and the carbon dioxide emissions (CO_2) of 11 countries.

<i>Income</i> (\$'000)	8.9	23.0	7.5	8.0	18.0	16.7	5.2	12.8	19.1	16.4	21.7
CO_2 (tonnes)	7.5	12.0	6.0	1.8	7.7	5.7	3.8	5.7	11.0	9.7	9.9

- Construct a scatterplot to display these data, with *income* as the explanatory variable (EV).
- Fit a least squares regression line to the scatterplot and determine its equation.
- Write the equation of the regression line in terms of the variables *income* and CO_2 , with the coefficients given, rounded to three significant figures.
- Determine and write down the value of the correlation coefficient, r , rounded to three significant figures.

Steps

- Start a new document by pressing **ctrl** + **N**.
- Select **Add Lists & Spreadsheet**. Enter the data into lists named *income* and *co₂*.
- Identify the explanatory variable (EV) and the response variable (RV).
EV: *income*
RV: *co₂*
Note: In saying that we want to predict CO_2 from *income*, we are implying that *income* is the EV.
- Press **ctrl** + **I**, select **Data & Statistics** and construct a scatterplot with the *income* (EV) on the horizontal (or x -) axis and *co₂* (RV) on the vertical (or y -) axis.

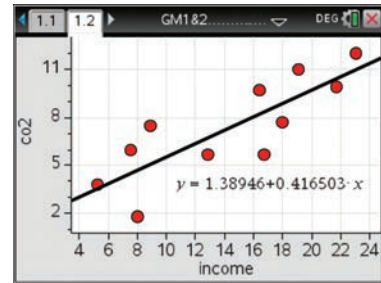


- 5 Press \square > **Analyze** > **Regression** > **Show Linear** (**a+bx**) to display the least squares regression line on the scatterplot. Simultaneously, the equation of the regression line is shown, written using the variables y and x :

$$y = 1.389 \dots + 0.416 \dots \text{ or}$$

$$y = 1.39 + 0.417x \text{ to 3 significant figures}$$

Note: The calculator assumes that the variable on the x -axis is the EV.



- 6 Write down the equation of the least squares regression line in terms of the variables *income* and CO_2 . Write the coefficients, rounded to three significant figures.
- 7 **a** Press \square + \square and select **Calculator** to open the Calculator application.
- b** Now press \square , locate then select **stat.r** and press \square to display the value of r .

$$CO_2 = 1.39 + 0.417 \times \text{income}$$



- 8 Write down the value of the correlation coefficient, rounded to three significant figures.

$$r = 0.818$$

How to determine and graph the least squares regression line using the ClassPad

The following data shows the per capita income (in \$'000) and the carbon dioxide emissions (in tonnes) of 11 countries.

<i>Income</i> (\$'000)	8.9	23.0	7.5	8.0	18.0	16.7	5.2	12.8	19.1	16.4	21.7
CO_2 (tonnes)	7.5	12.0	6.0	1.8	7.7	5.7	3.8	5.7	11.0	9.7	9.9


- a** Determine and graph the equation of the least squares regression line that will enable CO_2 emissions to be predicted from income.
- b** Write the equation in terms of the variables *income* and CO_2 , with the coefficients given, rounded to three significant figures.
- c** Determine and write down the value of the correlation coefficient, r , to three significant figures.


Steps


1 Open the **Statistics** application.

2 Enter the data into columns:

- Income in List1
- CO₂ in List2

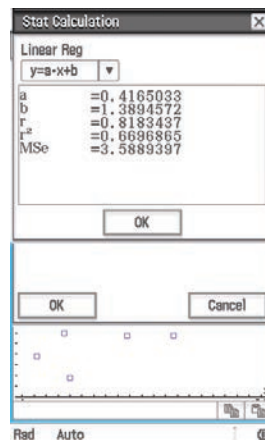
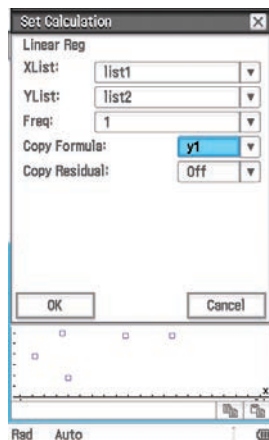
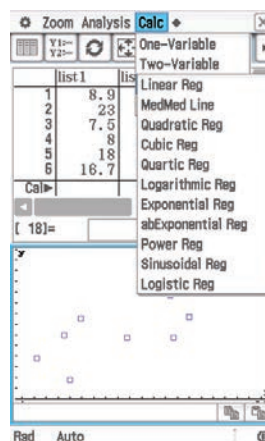
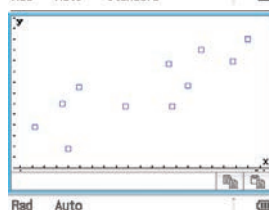
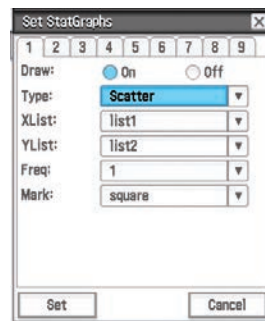
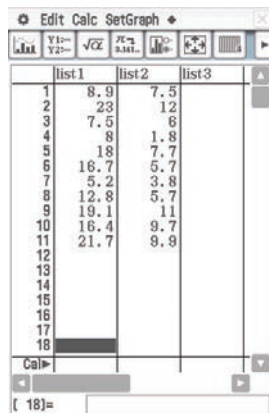
3 Tap  to open the **Set StatGraphs** dialog box and complete as shown.

Tap  to confirm your selections.

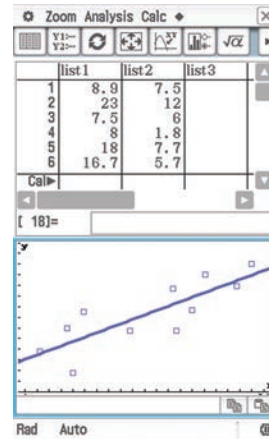
4 Tap  in the toolbar at the top of the screen to plot the scatterplot in the bottom half of the screen.

5 To calculate the equation of the least squares regression line:

- Tap **Calc>Regression>Linreg** from the menu bar.
- Complete the **Set Calculations** dialog box as shown.
- Tap **OK** to confirm your selections in the **Set Calculations** dialog box.
- This generates the regression results in **Stat Calculation**, shown opposite.



- 6 Tapping **OK** a second time automatically plots and displays the regression line on the plot.
- 7 Write down the equation of the least squares line in terms of the variables *income* and CO_2 and the value of the correlation coefficient, to three decimal places.



$$CO_2 = 1.39 + 0.417 \times \text{income}$$

$$r = 0.818$$

Section Summary

- ▶ The simplest method of finding the equation of a linear model is to draw a line **by eye** on the scatterplot.
- ▶ The vertical distance between a point on a bivariate plot and a line fitted to the data is called a **residual**.
- ▶ The method of **least squares** is a method for fitting a straight line to a scatterplot, based on minimising the sum of the squared residuals.
- ▶ The equation of the least squares line is given by $y = a + bx$, where:

$$\text{the slope } (b) \text{ is given by: } b = \frac{rs_y}{s_x}$$

and

$$\text{the intercept } (a) \text{ is then given by: } a = \bar{y} - b\bar{x}$$

Here:

- ▶ r is the correlation coefficient
- ▶ s_x and s_y are the standard deviations of x and y
- ▶ \bar{x} and \bar{y} are the mean values of x and y .

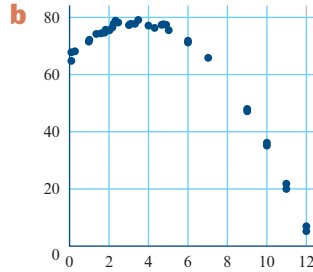
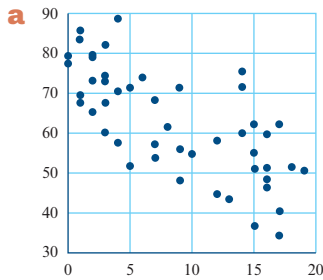
A CAS calculator can be used to determine the intercept and slope of the least squares line from the bivariate data.



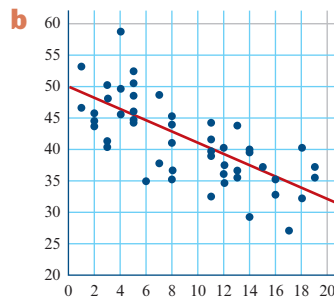
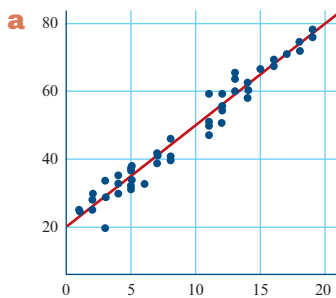
Exercise 7D

Building understanding

- 1 For each of the following plots, indicate whether it would be appropriate or not to fit a least squares regression line to the data.

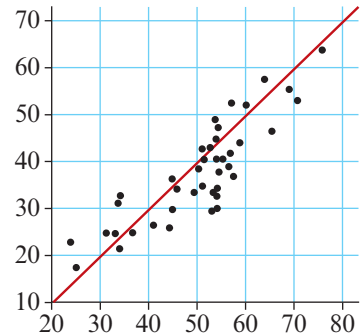


- 2 For each of the following scatterplots, determine the y -intercept (a), the slope (b) and hence the equation of the line $y = a + bx$. Round your answers to 3 significant figures.



Example 13

- 3 For the following scatterplot, determine the equation of the line shown on the scatterplot. Round your answer to two significant figures.



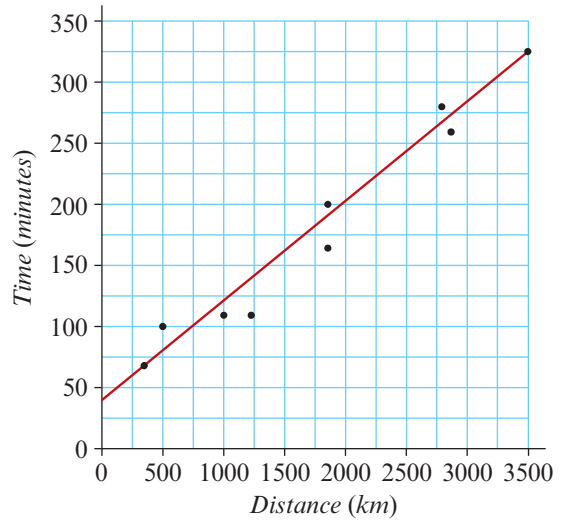
- 4 Use your calculator to find the equation of the least squares regression line, $y = a + bx$, which fits this data. Give your answers to three significant figures.

x	8.9	23.0	7.5	8.0	18.0	16.7	5.2	12.8	19.1	16.4	21.7
y	7.5	12.0	6.0	1.8	7.7	5.7	3.8	5.7	11.0	9.7	9.9

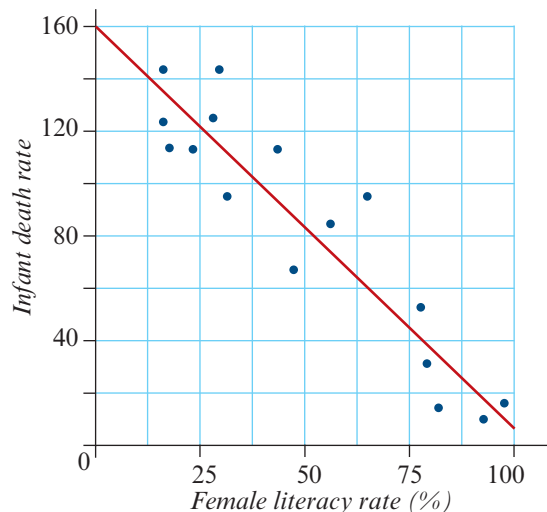
Developing understanding

Example 14

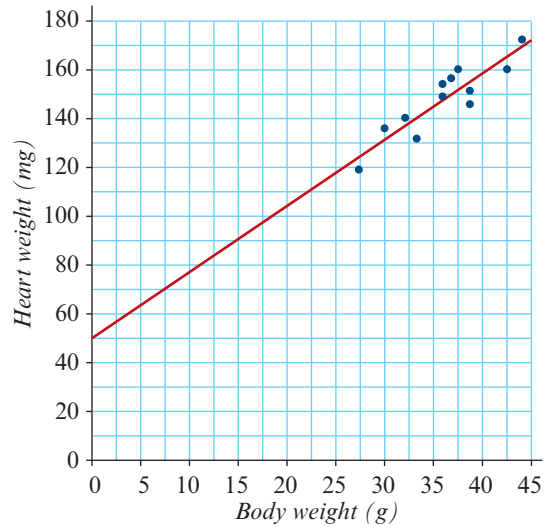
5 A straight line has been fitted by eye to a plot of travelling time, in minutes, (*Time*) against distance travelled, in km, (*Distance*) for nine plane trips between nine different cities. Determine the equation of the line in terms of *Time* and *Distance*.



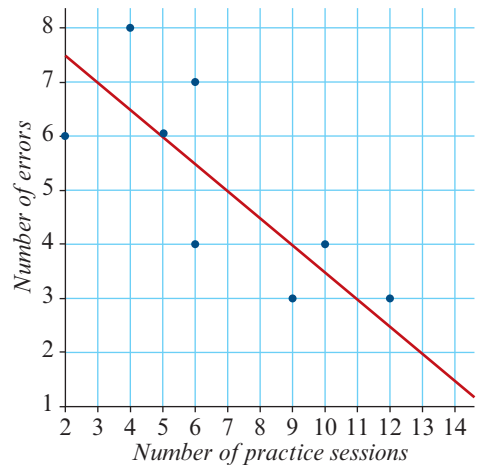
6 A straight line has been fitted by eye to a plot of *Infant death rate* (per 100 000 people) against *Female literacy rate (%)* for a number of countries. Determine the equation of the line in terms of these variables. Give answers correct to one decimal place.



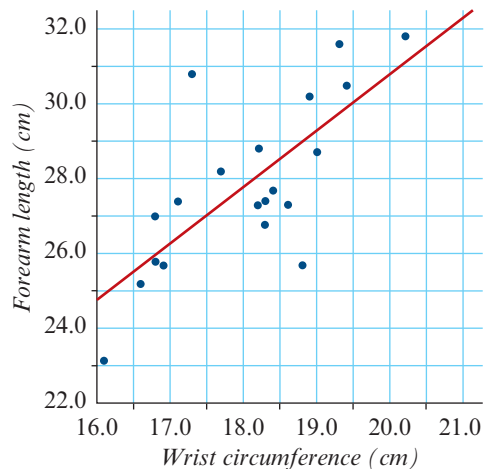
- 7 A straight line has been fitted by eye to a plot of *heart weight (mg)* against *body weight (g)* for twelve laboratory mice. Determine the equation of the line in terms of these variables. Give answers correct to one decimal place.



- 8 A straight line has been fitted by eye to a plot of *number of errors* made on a final assessment against *number of practice sessions* attended before the assessment for a group of university students. Determine the equation of the line shown on the scatterplot in terms of these variables.



- 9 The graph opposite shows the wrist circumference and forearm length, both in centimetres, of 20 people. A least squares line has been fitted to the scatterplot, with wrist circumference as the explanatory variable. Use the scatterplot to find the equation of the line in terms of the variables in the question.



- 10** Find the equation of the least squares regression line, $y = a + bx$, when:
 $r = 0.8$ $s_x = 6.8$ $s_y = 20.5$ $\bar{x} = 115.0$ $\bar{y} = 123.5$
 Give the values of the intercept and slope, rounded to two decimal places.

- 11** Find the equation of the least squares regression line, $y = a + bx$, when:
 $r = -0.600$ $s_x = 2.50$ $s_y = 1.70$ $\bar{x} = 18.0$ $\bar{y} = 15.7$
 Give the values of the intercept and slope, rounded to three significant figures.

Example 15

- 12** The table below shows the weight (in kg) and blood glucose level (in mg/100 mL) of eight adults.

<i>Weight (kg)</i>	82.1	70.1	76.6	82.1	83.9	73.2	66.0	77.5
<i>Glucose (mg/100 mL)</i>	101	89	98	100	108	104	94	89

- a** Construct a scatterplot to display the data, with *weight* as the EV.
b Fit a least squares regression line to the scatterplot and determine its equation.
c Write the equation of the regression line in terms of the variables *glucose* and *weight* with the coefficients given, rounded to three significant figures.
d Determine the correlation coefficient to three significant figures.



- 13** The table below shows the scores which a group of nine students obtained on two class tests, Test 1 and Test 2, as part of their school-based assessment.

<i>Test 1</i>	33	45	27	42	50	38	17	35	29
<i>Test 2</i>	43	46	36	34	48	34	29	41	28

- a** Construct a scatterplot to display these data, with *Test 1* as the EV.
b Fit a least squares regression line to the scatterplot and determine its equation.
c Write the equation of the regression line in terms of the variables *Test 2* and *Test 1* with the coefficients given, rounded to three significant figures.
d Determine the correlation coefficient to three significant figures.

- 14** The table below shows the carbohydrate content, in grams, (*carbs*) and the fat content, in grams, (*fat*) in 100 grams of nine breakfast cereals.

<i>Carbs</i>	88.7	67.0	77.5	61.7	86.8	32.4	72.4	77.1	86.5
<i>Fat</i>	0.3	1.3	2.8	7.6	1.2	5.7	9.4	10.0	0.7

- a** Construct a scatterplot to display these data, with *carbs* as the EV.
b Fit a least squares regression line to the scatterplot and determine its equation.
c Write the equation of the regression line in terms of the variables *fat* and *carbs* with the coefficients given, rounded to three significant figures.
d Determine the correlation coefficient to three significant figures.
- 15** The table below shows the *age* and *height* of six young children.

<i>Age (months)</i>	36	40	44	52	56	60
<i>Height (cm)</i>	84	87	90	92	94	96

- a** Construct a scatterplot to display these data, with *age* as the EV.
b Fit a least squares regression line to the scatterplot and determine its equation.
c Write the equation of the regression line in terms of the variables *height* and *age* with the coefficients given, to three significant figures.
d Determine the correlation coefficient to three significant figures.
- 16** The following table gives the *height* and *arm span*, both in centimetres, for a group of eight people.

<i>Height (cm)</i>	162	170	164	153	171	166	170	163
<i>Arm span (cm)</i>	163	168	165	154	165	164	170	165

- a** Construct a scatterplot to display these data, with *height* as the EV.
b Fit a least squares regression line to the scatterplot and determine its equation.
c Write the equation of the regression line in terms of the variables *arm span* and *height* with the coefficients given, to three significant figures.
d Determine the correlation coefficient to three significant figures.

Testing understanding

- 17** The statistical analysis of a set of bivariate data involving variables x and y resulted in the information displayed in the table below.

Mean	$\bar{x} = 8.97$	$\bar{y} = 4.42$
Standard deviation	$s_x = 4.29$	$s_y = 1.69$
Equation of the least squares line	$y = 1.62 + 0.312x$	

Use this information to determine the value of the correlation coefficient, r .

- 18** In a mathematics class, a group of students were asked to draw circles on squared paper. They measured the diameters of the circles they had drawn and then estimated the areas of the circles by counting the squares. Their results are given in the following table. Their teacher suggested they investigate the association between the area of each circle and the square of its diameter.

<i>diameter (cm)</i>	3.5	6.2	5.4	3.7	7.3	8.6	3.7	2.9	2.1	9.7	3.7
<i>area (cm²)</i>	9.5	30.0	22.7	10.2	42.6	57.7	10.5	5.7	2.7	74.4	11.0

- Construct a scatterplot of the data, with the variable $diameter^2$ as the explanatory variable and the variable $area$ as the response variable.
- Use your calculator to find the intercept and slope of the least squares line for this scatterplot, rounded to three significant figures.
- Complete this equation: $area = \square + \square \times diameter^2$
- Compare this equation with what you know to be the exact association between the diameter and area of a circle.

7E Interpreting and predicting from a linear model

Learning intentions

- ▶ To be able to interpret the slope and intercept of a linear model in the context of the data.
- ▶ To be able to use the equation of a linear model to predict the value of the response variable, based on the value of the explanatory variable.
- ▶ To be able to understand the difference between **interpolation** and **extrapolation** when making predictions.

Interpreting the slope and intercept of a model in the context of the data

Whatever method is used to determine the equation of a straight line, the intercept and slope of the line can be interpreted in the context of the data in the question. This interpretation gives us further insights into the association between the variables.

Interpreting the slope and intercept of a linear model

For the regression line $y = a + bx$:

- the slope (b) tells us on average the change in the response variable (y) for each one-unit increase or decrease in the explanatory variable (x)
- the intercept (a) tells us on average the value of the response variable (y) when the explanatory variable (x) equals 0.

Note: The interpretation of the y -intercept in a data context may not be sensible when $x = 0$ is not within the range of observed x values.


Example 16 Interpreting the slope and intercept of a linear model

A regression line is used to model the association between the *time*, in hours, a group of students spent studying for an examination and their *mark* (%). The equation of the regression line is:

$$\text{mark} = 30.8 + 1.62 \times \text{time}$$

- a i** Write down the value of the intercept.
ii Interpret the intercept in the context of these variables.
- b i** Write down the value of the slope.
ii Interpret the slope in the context of these variables.

Explanation

- a i** In a linear equation of the form $y = a + bx$; the intercept is a .
ii The intercept ($a = 30.8$) gives the average *mark* (y) when the study *time* (x) equals 0.
- b i** In a linear equation of the form $y = a + bx$; the slope is b .
ii The slope ($b = 1.62$) gives the average change in the *mark* (y) associated with a one-unit increase in the variable *time* (x).

Solution

$$\text{intercept} = 30.8$$

On average, students who spend no time studying for the examination will obtain a mark of 30.8.

$$\text{slope} = 1.62$$

On average, students' marks increase by 1.62 for each extra hour of study.

Now try this 16 Interpreting the slope and intercept of a linear model (Example 16)

A regression line is used to model the association between the time, in hours, students spend doing household chores (*chores*) and the hours they spend in part-time work (*work*), which ranged from 0 to 8 hours for this group of students. The equation of the regression line is:

$$\text{chores} = 8.0 - 0.30 \times \text{work}$$

- a i** Write down the value of the intercept.
ii Interpret the intercept in the context of these variables.
- b i** Write down the value of the slope.
ii Interpret the slope in the context of these variables.

Hint 1 Use the wording in Example 16 as a model for your answers.

Hint 2 The sign of the slope tells you if the association between the variables is positive or negative. This is an important consideration when interpreting the slope.

Using the model to make predictions: interpolation and extrapolation

The aim of linear regression is to model the association between two numerical variables by using the equation of a straight line. This equation can then be used to make predictions.

The data below shows the times that 10 students spent studying for an exam and the marks they subsequently obtained.

<i>Time (hours)</i>	4	36	23	19	1	11	18	13	18	8
<i>Mark (%)</i>	41	87	67	62	23	52	61	43	65	52

If we fitted a linear model to this data using the least squares method, we would have an equation close to:

$$\text{mark} = 30.8 + 1.62 \times \text{time}$$

Using this equation and rounding off to the nearest whole number, we would predict that a student who spent:

- 0 hours studying would obtain a mark of 31% (mark = $30.8 + 1.62 \times 0 = 31\%$)
- 8 hours studying would obtain a mark of 44% (mark = $30.8 + 1.62 \times 8 = 44\%$)
- 12 hours studying would obtain a mark of 50% (mark = $30.8 + 1.62 \times 12 = 50\%$)
- 30 hours studying would obtain a mark of 79% (mark = $30.8 + 1.62 \times 30 = 79\%$)
- 80 hours studying would obtain a mark of 160% (mark = $30.8 + 1.62 \times 80 = 160\%$)

This last result, 160%, points to one of the limitations of substituting values into a regression equation without thinking carefully. Using this regression equation, we predict that a student who studies for 80 hours will obtain a mark of more than 100%, which is impossible. Something is wrong!

The problem is that we are using the regression equation to make predictions well outside the range of values used to calculate this equation. The maximum time any student spent studying for this exam was 36 hours; yet, we are using the equation we calculated to try to predict the exam mark for someone who studies for 80 hours. Without knowing that the model works equally well for someone who spends 80 hours studying, which we don't, we are venturing into unknown territory and can have little faith in our predictions.

As a general rule, a regression equation only applies to the range of values of the explanatory variables used to determine the equation. Thus, we are reasonably safe using the line to make predictions that lie roughly within this data range, say from 1 to 36 hours. The process of making a prediction within the range of values of the explanatory variable used to derive the regression equation is called **interpolation**, and we can have some faith in these predictions.

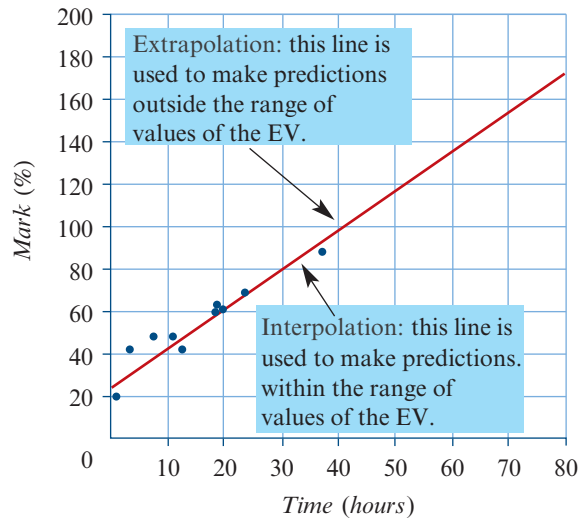
However, we must be extremely careful about how much faith we put into predictions made outside the range of values of the explanatory variable. Making predictions outside the data range is called **extrapolation**.

Interpolation and extrapolation

Predicting within the range of values of the explanatory variable is called **interpolation**. Interpolation is generally considered to give a **reliable** prediction.

Predicting outside the range of values of the explanatory variable is called **extrapolation**. Extrapolation is generally considered to give an **unreliable** prediction.

For example, if we use the regression line to predict the examination mark for 30 hours of studying time, we would be interpolating. However, if we use the regression line to predict the examination mark for 50 hours of studying time, we would be extrapolating. Extrapolation is a less reliable process than interpolation because we are going beyond the range of values of the explanatory variable, and it is quite possible that the association may no longer be linear.




Example 17 Using the linear model to make predictions

The equation relating the weights, in kg, and heights, in cm, of a group of students whose heights ranged from 163 cm to 190 cm, is:

$$\text{weight} = -40 + 0.60 \times \text{height}$$

Use this equation to predict the weight of students with the following heights.

Are you interpolating or extrapolating?

a 170 cm

b 65 cm

Explanation

a Substitute 170 into the equation and evaluate.

b Substitute 65 into the equation and evaluate.

Solution

The weight of a person of height 170 cm is predicted to be:

$$\text{weight} = -40 + 0.60 \times 170 = 62 \text{ kg}$$

Interpolating: predicting within the range of values of the EV.

The weight of a person of height 65 cm is predicted to be:

$$\text{weight} = -40 + 0.60 \times 65 = -1.0 \text{ kg}$$

which is not possible.

Extrapolating: we are predicting well outside the range of values of the EV.

Now try this 17 Using the linear model to make predictions (Example 17)

A regression line is used to model the association between the time, in hours, which students spend doing household chores (*chores*) and the hours they spend in part-time work (*work*), which ranged from 0 to 8 hours for this group of students.

The equation of the regression line is:

$$\text{chores} = 8.00 - 0.30 \times \text{work}$$

Use this equation to predict the hours spent doing chores by a student who works the following hours per week in their part-time job. Are you interpolating or extrapolating?

a 2 hours

b 10 hours

Hint 1 Substitute the value for *work* in the equation given.

Section Summary

- ▶ For the regression line, $y = a + bx$:
 - ▷ on average, the response variable (y) changes by b units for each one-unit increase in the explanatory variable (x).
 - ▷ on average, the intercept (a) predicts the value of the response variable (y) when the explanatory variable (x) equals 0.
- ▶ The linear model, $y = a + bx$, can be used to predict the value of the response variable (y) for a particular value of the explanatory variable (x).
- ▶ Predicting within the range of the values of the explanatory variable used to fit the linear model is called **interpolation**; predicting outside the range of values of the explanatory variable used to fit the linear model is called **extrapolation**.
- ▶ Interpolation is generally considered to give a **reliable** prediction.
- ▶ Extrapolation is generally considered to give an **unreliable** prediction.



Exercise 7E

Building understanding

- 1 The following linear model which allows *height*, in centimetres, to be predicted from *age*, in months, was determined from data collected from a group of children aged from 12 to 36 months.

$$\text{height} = 69 + 0.50 \times \text{age}$$

- a Which is the EV and which is the RV?
- b Write down the values of the intercept and slope.
- c Complete the following sentences:
- i The slope tells us, on average, that *height* increases by cm for each month increase in age.
 - ii The intercept tells us, that, on average, students aged months will be cm tall.



- 2 Complete the following sentences.

Using a linear model to make a prediction:

- a within the range of data is called .
- b outside the range of data is called .

- 3 The linear model relating students' marks (%) on an oral French test (*mark*) to the number of hours they spent practising speaking French in the week before the test (*practice*) has the equation:

$$\text{mark} = 48.1 + 2.20 \times \text{practice}$$

The linear model predicts that:

- a a student who practised for 5 hours will score $48.1 + 2.20 \times \text{ } = \text{ }$ per cent.
- b a student who practised for 8 hours will score $48.1 + 2.20 \times \text{ } = \text{ }$ per cent.

Developing understanding

Example 16

- 4 The equation, $\text{price} = 37\,650 - 4200 \times \text{age}$, can be used to predict the *price* of a used car (in dollars) from its *age* (in years).

- a i For this regression equation, write down the value of the intercept.
ii Interpret the intercept in the context of the variables in the equation.
- b i For this regression equation, write down the value of the slope.
ii Interpret the slope in the context of the variables in the equation.

- 5 The following regression equation can be used to predict the flavour rating of yoghurt from its percentage fat content, (*calories*).

$$\text{flavour rating} = 40 + 2.0 \times \text{calories}$$

- a i For this regression equation, write down the value of the intercept.
ii Interpret the intercept in the context of the variables in the equation.
- b i For this regression equation, write down the value of the slope.
ii Interpret the slope in the context of the variables in the equation.



Example 17

- 6 For children between the ages of 36 and 60 months, the equation relating their *height* (in cm) to their *age* (in months) is:

$$\text{height} = 72 + 0.40 \times \text{age}$$

Use this equation to predict the height (to the nearest cm) of a child with the following age. Are you interpolating or extrapolating?

- a 40 months old b 55 months old c 70 months old

- 7** When preparing between 25 and 100 *meals*, a cafeteria's *cost* (in dollars) is given by the equation:

$$\text{cost} = 175 + 5.80 \times \text{meals}$$

Use this equation to predict the cost (to the nearest dollar) of preparing the following meals. Are you interpolating or extrapolating?

- a** no meals
 - b** 60 meals
 - c** 89 meals
- 8** For women of heights from 150 cm to 180 cm, the equation relating a *daughter's height* (in cm) to her *mother's height* (in cm) is:

$$\text{daughter's height} = 18.3 + 0.910 \times \text{mother's height}$$

Use this equation to predict (to the nearest cm) the adult height of women whose mothers are the following heights. Are you interpolating or extrapolating?

- a** 168 cm tall
 - b** 196 cm tall
 - c** 155 cm tall
- 9** Students sit for two exams, two weeks apart. The following regression equation can be used to predict the students' marks on exam 2 from the marks they obtained on exam 1.

$$\text{mark on exam 2} = 15.7 + 0.650 \times \text{mark on exam 1}$$

- a**
 - i** For this regression equation, write down the value of the intercept.
 - ii** Interpret the intercept in the context of the variables in the equation.
 - b**
 - i** For this regression equation, write down the value of the slope.
 - ii** Interpret the slope in the context of the variables in the equation.
 - c** Use the equation to predict the mark on exam 2 for a student who obtains a mark of 20 on exam 1. Give your answer to the nearest mark.
- 10** It has been suggested that the blood *glucose* level (in mg/100 mL) of adults can be predicted from their *weight* (in kg).

$$\text{glucose} = 51 + 0.62 \times \text{weight}$$

- a**
 - i** For this regression equation, write down the value of the intercept.
 - ii** Interpret the intercept in the context of the variables in the equation.
- b**
 - i** For this regression equation, write down the value of the slope.
 - ii** Interpret the slope in the context of the variables in the equation.
- c** Use the equation to predict the blood glucose level of a person who weighs 75 kg. Give your answer rounded to one decimal place.

Testing understanding

- 11** A study of the association between the average score in an examination in each of 25 schools and the student:staff ratio in that school, resulted in the information below.

Variable	Mean	Stand dev
<i>student:staff ratio</i>	13.404	4.128
<i>score</i>	71.669	12.013
Correlation coefficient	$r = -0.651$	

Use this information to predict the value of average examination scores in a school with a student:staff ratio of 15. Give your answer rounded to one decimal place.

- 12** The following table gives the educational level (*education*), the number of years the person has worked for the company (*years*) and their current salary to the nearest thousand dollars (*salary*) for a group of current employees of a particular company.

<i>education</i>	<i>years</i>	<i>salary</i>	<i>education</i>	<i>years</i>	<i>salary</i>
Secondary	2	52	Tertiary	2	62
Secondary	3	64	Tertiary	3	69
Secondary	2	56	Tertiary	4	75
Secondary	4	63	Tertiary	5	76
Secondary	7	65	Tertiary	4	72
Secondary	6	64	Tertiary	4	68
Secondary	7	52	Tertiary	1	63
Secondary	10	65	Tertiary	8	85
Secondary	5	59	Tertiary	3	67
Secondary	5	62	Tertiary	6	77

- a**
- Find the equation of a linear model which allows *salary* to be predicted from *years* for those with Secondary education, rounding the values of the intercept and slope to three decimal places.
 - Interpret the intercept and slope in the context of the variables in this equation.
 - Use the equation to predict the *salary* for an employee with Secondary education who has worked for the company for 5 years. Round your answer to the nearest hundred dollars.
- b**
- Find the equation of a linear model which allows *salary* to be predicted from *years* for those with Tertiary education, rounding the values of the intercept and slope to three decimal places.
 - Interpret the intercept and slope in the context of the variables in this equation.
 - Use the equation to predict the *salary* for an employee with Tertiary education who has worked for the company for 5 years. Round your answer to the nearest hundred dollars.
- c** Are the predictions in parts **a** **iii** and **b** **iii** reliable? Explain your reasoning.

Key ideas and chapter summary



Explanatory and response variables

The **explanatory variable** is used to explain or predict the value of the **response variable**.

Scatterplot

A two-dimensional data plot where each point represents the value of two related variables in a bivariate data set. In a scatterplot, the **response variable (RV)** is plotted on the vertical axis and the **explanatory variable (EV)** on the horizontal axis.

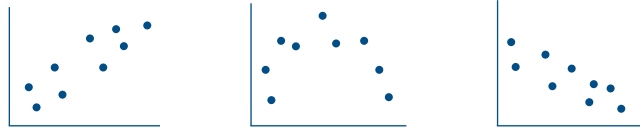
A scatterplot is used to help identify and describe the association between two **numerical** variables.

Identifying associations (relationships) between two numerical variables

A random cluster of points (no clear pattern) indicates that there is **no association** between the variables.



A clear pattern in the scatterplot indicates that there is an **association between the variables**.



Describing associations in scatterplots

Associations are described in terms of:

- **direction** (positive or negative)
- **form** (linear or non-linear)
- **strength** (strong, moderate, weak or none).

Pearson's correlation coefficient (r)

Pearson's correlation coefficient (r) is a statistic that measures the direction and strength of a linear association between a pair of numerical variables. The strength of the linear association can be classified as follows:

$0.75 \leq r \leq 1$	strong positive association
$0.5 \leq r < 0.75$	moderate positive association
$0.25 \leq r < 0.5$	weak positive association
$-0.25 < r < 0.25$	no association
$-0.5 < r \leq -0.25$	weak negative association
$-0.75 < r \leq -0.5$	moderate negative association
$-1 \leq r \leq -0.75$	strong negative association

Correlation and causation

An association (correlation) between two variables does not automatically imply that the observed association between the variables is **causal**.

Linear regression

A straight line can be used to model a linear association between two numerical variables. The association can then be described by a rule of the form, $y = a + bx$.

In this equation:

- y is the **response variable**
- x is the **explanatory variable**
- a is the **y-intercept**
- b is the **slope of the line**.

Fitting a line by eye

The simplest method of fitting a linear model to a scatterplot is to draw in a line **by eye** which follows the trend of the data.

The least squares method

The **least squares method** for fitting a line to a scatterplot minimises the sum of the squares of the residuals.

Interpreting the intercept and slope

For the regression line, $y = a + bx$:

- the slope (b) tells us, on average, the change in the response variable (y) for each one-unit increase or decrease in the explanatory variable (x)
- the intercept (a) tells us, on average, the value of the response variable (y) when the explanatory variable (x) equals 0.

Making predictions

The **regression line**, $y = a + bx$, enables the value of y to be predicted for a given value of x , by substitution into the equation.

Interpolation and extrapolation


Predicting within the range of the values of the explanatory variable is called **interpolation**, and will give a **reliable** prediction.

Predicting outside the range of the values of the explanatory variable is called **extrapolation**, and will give an **unreliable** prediction.

Skills checklist



Check-list

Download this checklist from the Interactive Textbook, then print it and fill it out to check your skills. 

7A

1 Where appropriate, I can identify the EV and RV in bivariate data.

e.g. I wish to predict the length of a person's foot (*foot length*) from their height (*height*), both measured in centimetres. Which is the explanatory variable (EV), and which is the response variable (RV)?

7A

2 Having identified the EV and RV, I can construct an appropriate scatterplot.

e.g. In order to predict the length of a person's foot from their height, a group of 16 students collected the following data set, which gives height (*height*) and the length of their foot (*foot length*) in centimetres. Construct an appropriate scatterplot.

<i>height</i>	<i>foot length</i>	<i>height</i>	<i>foot length</i>
172	24	182	32
172	22	160	26
177	28	153	22
165	20	177	28
176	28	172	26
155	25	173	25
166	22	185	40
166	30	159	24

7B

3 I can use a scatterplot to describe an observed association between two numerical variables in terms of direction, form and strength, and the meaning of the association, within the context of the data.

e.g. Describe the association in the previous scatterplot between *height* and *foot length* in terms of direction, form and strength.

7C

4 I can estimate the value of the correlation coefficient, r , from a scatterplot and calculate its value from data using technology.

e.g. Estimate the value of the correlation coefficient, r , between *height* and *foot length* by comparing the scatterplot to those on page 420, and then use your calculator to find its exact value. Give your answer rounded to three decimal places.

7C

5 I can classify the value of the correlation coefficient, r , as weak, moderate or strong.

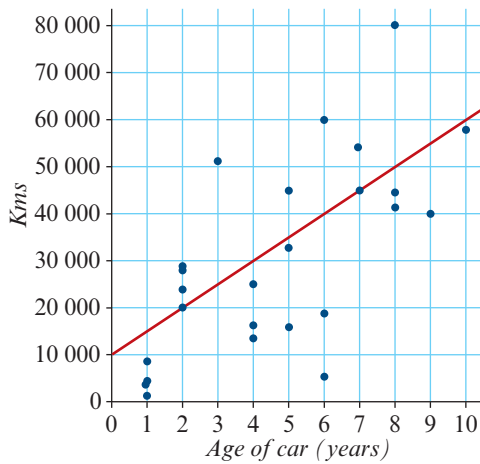
e.g. Use the table on page 428 to classify the value of the correlation coefficient, r , between *height* and *foot length* as weak, moderate or strong.

- 7C** **6** I can recognise that an association (correlation) between two variables does not automatically imply that the observed association between the variables is causal.

e.g. There is a strong positive correlation between sales of umbrellas and traffic accidents. Can we conclude that decreasing the sales of umbrellas will help decrease the number of traffic accidents?

- 7D** **7** I can fit a linear model by eye to a scatterplot and find the equation of the line.

e.g. Determine the equation of the line fitted by eye to the scatterplot, below, which shows the association between the age of a group of cars, in years, (*age of car*) and the number of kilometres they have travelled (*kms*).



- 7D** **8** I can determine the equation of the least squares line fitted to the data to model an observed linear association.

e.g. Find the equation of the least squares line which would enable *foot length* to be predicted from *height*. Give the values of the intercept and slope, rounded to two decimal places.

- 7E** **9** I can interpret the slope and intercept of the linear model in the context of data.

e.g. Interpret the intercept and slope of the regression line relating *height* and *foot length* in terms of these variables.

- 7E** **10** I can use the model to make predictions, differentiating between interpolation and extrapolation.

e.g. Use the linear model line relating *height* and *foot length* to predict the foot length of people with the following heights, indicating in each whether you are interpolating or extrapolating. Give your answers rounded to one decimal place.

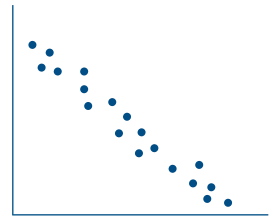
- a** 160 cm tall **b** 100 cm tall

Multiple-choice questions

- 1 For which one of the following pairs of variables would it be appropriate to construct a scatterplot?
- A *eye colour* (blue, green, brown, other) and *hair colour* (black, brown, blonde, other)
 - B *test score* and *sex* (male, female)
 - C *political party preference* (Labor, Liberal, Other) and *age* in years
 - D *age* in years and *blood pressure* in mmHg
 - E *height* in cm and *sex* (male, female)

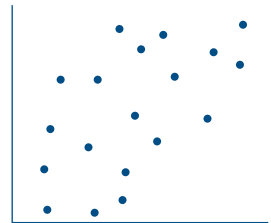
- 2 For the scatterplot shown, the association between the variables is best described as:

- A weak linear negative
- B strong linear negative
- C no association
- D weak linear positive
- E strong linear positive



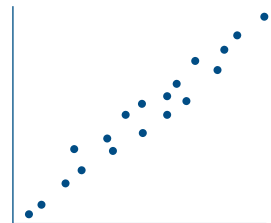
- 3 For the scatterplot shown, the association between the variables is best described as:

- A weak linear negative
- B weak non-linear negative
- C no association
- D weak linear positive
- E strong non-linear positive



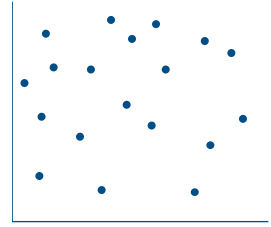
- 4 For the scatterplot shown, the association between the variables is best described as:

- A weak linear positive
- B strong linear positive
- C no association
- D moderate linear positive
- E strong non-linear positive



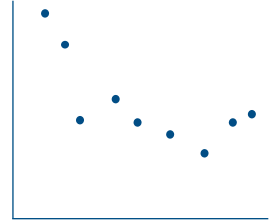
5 For the scatterplot shown, the association between the variables is best described as:

- A weak non-linear negative
- B strong linear negative
- C no association
- D weak non-linear positive
- E weak linear positive



6 For the scatterplot shown, the association between the variables is best described as:

- A weak negative linear
- B strong negative linear
- C no association
- D weak non-linear
- E strong non-linear

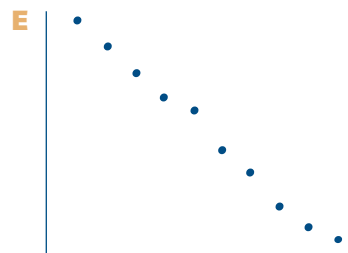
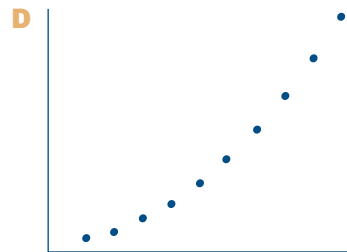
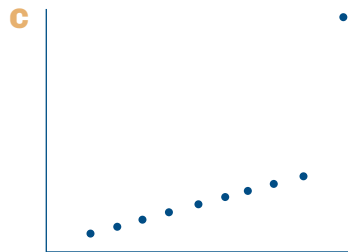
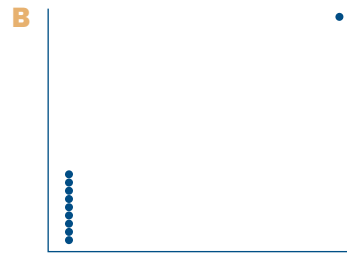
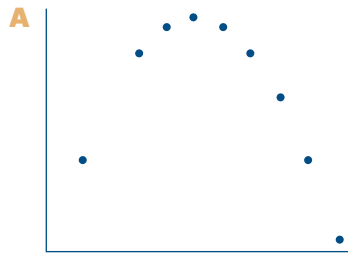


7 The association between birth weight and infant mortality rate is negative. Given this information, it can be concluded that:

- A birth weight and infant mortality rate are not related.
- B infant mortality tends to increase as birth weight increases.
- C infant mortality tends to decrease as birth weight decreases.
- D infant mortality tends to decrease as birth weight increases.
- E the values of infant mortality are, in general, less than the corresponding values of birth weight.

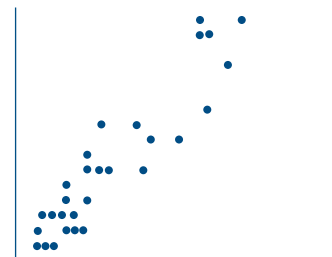


- 8 For which of the following scatterplots would it make sense to calculate the correlation coefficient (r) to indicate the strength of the association between the variables?



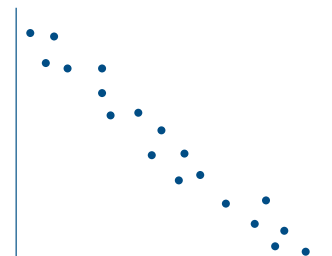
- 9 For the scatterplot shown, the value of Pearson's correlation coefficient, r , is closest to:

A 0.28 **B** 0.41 **C** 0.63
D 0.86 **E** 0.99



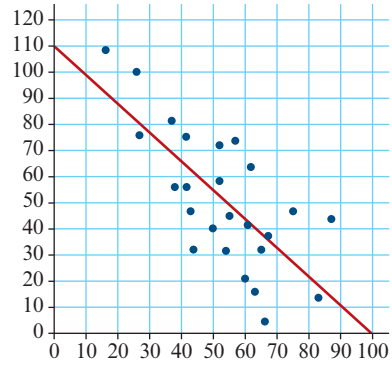
- 10 For the scatterplot shown, the value of Pearson's correlation coefficient, r , is closest to:

A -0.90 **B** -0.64 **C** -0.23
D 0.64 **E** 0.90



- 11** A correlation coefficient of $r = -0.32$ would classify a linear association as:
- A** weak positive **B** weak negative **C** moderately positive
D close to zero **E** moderately weak

- 12** A line is fitted by eye to the scatterplot, as shown. The equation of this line is:



- A** $y = 110 - 1.10x$ **B** $y = -110 + 1.10x$ **C** $y = -110 - 1.10x$
D $y = 1.10 - 110x$ **E** $y = 1.10 + 1.10x$
- 13** The equation of the least squares regression line, $y = a + bx$, when:
 $r = 0.500$ $s_x = 2.40$ $s_y = 12.5$ $\bar{x} = 8.00$ $\bar{y} = 34.5$

is given by:

- A** $y = 13.7 - 2.60x$ **B** $y = 2.60 + 13.7x$ **C** $y = 33.6 + 0.096x$
D $y = 13.7 + 2.60x$ **E** $y = -13.7 + 2.60x$

The following information relates to Questions 14 and 15.

The weekly *income* and weekly *expenditure* on food for a group of 10 university students is given in the following table.

<i>Income (\$/week)</i>	150	250	300	600	300	380	950	450	850	1000
<i>Expenditure (\$/week)</i>	40	60	70	120	130	150	200	260	460	600

- 14** The value of the Pearson correlation coefficient, r , for these data is closest to:
- A** 0.2 **B** 0.4 **C** 0.6 **D** 0.7 **E** 0.8
- 15** The least squares regression line that enables weekly *expenditure* (in dollars) on food to be predicted from weekly *income* (in dollars) is closest to:
- A** $\text{expenditure on food} = 0.482 + 42.9 \times \text{income}$
B $\text{expenditure on food} = 0.482 - 42.9 \times \text{income}$
C $\text{expenditure on food} = -42.9 + 0.482 \times \text{income}$
D $\text{expenditure on food} = 239 + 1.36 \times \text{income}$
E $\text{expenditure on food} = 1.36 + 239 \times \text{income}$

The following information relates to Questions 16 and 17.

The equation of a regression line that enables the weekly *amount* spent on entertainment (in dollars) to be predicted from weekly *income* is given by:

$$\text{amount} = 40 + 0.10 \times \text{income}$$

- 16** Using this equation, the amount spent on entertainment by an individual with a weekly income of \$600 is predicted to be:
- A** \$40 **B** \$46 **C** \$100 **D** \$240 **E** \$24 060
- 17** From the equation of the regression line it can be concluded that, on average:
- A** the weekly *amount* spent on entertainment increases by 40 cents a week for each extra dollar of weekly income.
- B** the weekly *amount* spent on entertainment increases by 10 cents a week for each extra dollar of weekly income.
- C** the weekly *income* increases by 10 cents for each dollar increase in the amount spent on entertainment each week.
- D** \$40 is spent on entertainment each week.
- E** \$40.10 is spent on entertainment each week.



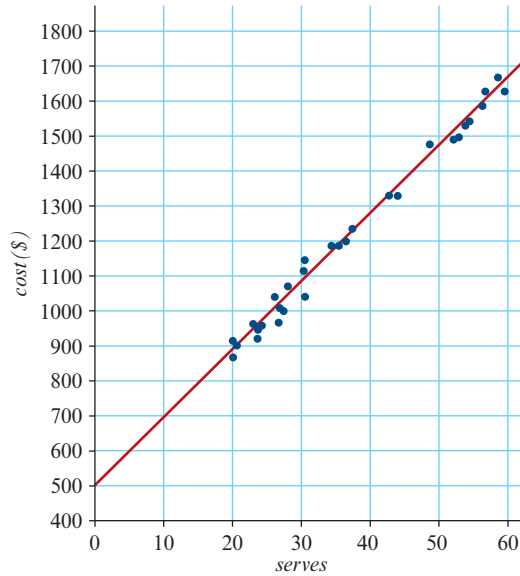
Short-answer questions

- 1** The following table gives the *number* of times the ball was inside the team's 50-metre line in an AFL football game and the team's final score (in points) in that game.

<i>Number</i>	64	57	34	61	51	52	53	51	64	55	58	71
<i>Score (points)</i>	90	134	76	92	93	45	120	66	105	108	88	133

- a** Which variable is the RV?
- b** Construct a scatterplot of *score* against *number*.
- c** Use the scatterplot to describe the association in terms of direction, form and strength.

- 2 The following scatterplot shows the costs of catering a meal, in dollars, (*cost*) with the number of meals served (*serves*). A line fitted by eye is shown on the following scatterplot.



- a Which variable is the RV and which is the EV?
 b Find the equation of the line shown on the plot in terms of *cost* and *serves*. Round your answer to three significant figures.

- 3 The *distance* travelled to work and the *time* taken for ten company employees are given in the opposite table. *Distance* is the response variable.

Distance (km)	Time (min)
12	15
50	75
40	50
25	50
45	80
20	50
10	10
3	5
10	10
30	35

- a Determine the value of the Pearson correlation coefficient, r , for this set of data. Round your answer to three decimal places.
 b Determine the equation of the least squares line for this data, and write the equation in terms of the variables *distance* and *time*. Round your answer to three significant figures.

- 4 From a data set relating *height* (cm) and *weight* (kg) for a group of students it was determined that the correlation coefficient was $r = 0.75$. It was also found that the mean height for the group was 174.5 cm, with a standard deviation of 9.3 cm, and that the mean weight was 65.9 kg, with a standard deviation of 10.8 kg.

- a Find the slope of the least squares regression line which would enable *weight* to be predicted from *height*. Round your answer to three significant figures.
 b Find the intercept for this line. Round your answer to three significant figures.
 c Hence, write down the equation of the least squares regression line in terms of *weight* and *height*.

- 5 The regression equation:
- $$\text{taste score} = -22 + 7.3 \times \text{magnesium content}$$
- can be used to predict the *taste score* of a country town's drinking water from its *magnesium content* (in mg/litre).
- Which variable is the explanatory variable?
 - Write down and interpret the slope of the regression line.
 - Use the regression line to predict the taste score of a country town's drinking water whose magnesium content is 16 milligrams/litre, rounded to one decimal place.
- 6 The *time* taken to complete a task and the number of *errors* on the task were recorded for a sample of 10 primary school children.

<i>Time (s)</i>	22.6	21.7	21.7	21.3	19.3	17.6	17.0	14.6	14.0	8.8
<i>Errors</i>	2	3	3	4	5	5	7	7	9	9

- Determine the equation of the least squares regression line that fits this data, with *errors* as the response variable. Round your answer to three significant figures.
 - Determine the value of Pearson's correlation coefficient to two decimal places.
- 7 Researchers found a strong positive correlation between students' exam results in Mathematics and in French. Can we conclude that encouraging students to study harder in French would improve their scores in Mathematics?

Written-response questions

- 1 A marketing firm wanted to investigate the association between the number of times a song was played on the radio (*played*) and the number of downloads sold the following week (*weekly sales*).

The following data was collected for a random sample of ten songs.

<i>Played</i>	47	34	40	34	33	50	28	53	25	46
<i>Weekly sales</i>	3950	2500	3700	2800	2900	3750	2300	4400	2200	3400

- Which is the explanatory variable and which is the response variable?
- Construct a scatterplot of this data.
- Determine the value of the Pearson correlation coefficient, r , for this data. Round your answer to four decimal places.
- Describe the association between *weekly sales* and *played* in terms of direction, form and strength.
- Determine the equation for the least squares regression line and write it down in terms of the variables *weekly sales* and *played*. Round the values of coefficients to three significant figures.
- Interpret the slope and intercept of the regression line in the context of the problem.
- Use your equation to predict the number of downloads of a song when it was played on the radio 100 times in the previous week.
- In making this prediction, are you interpolating or extrapolating?

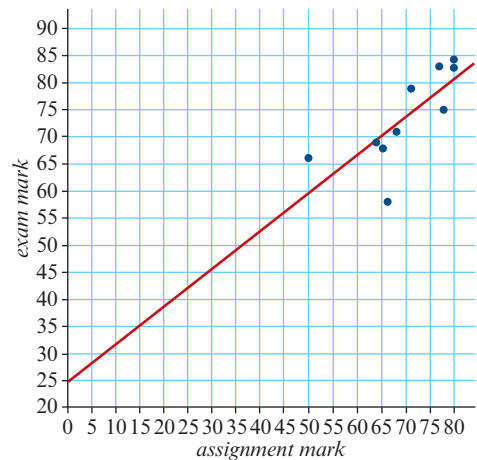
- 2** To test the effect of driving instruction on driving skill, 10 randomly selected learner drivers were given a *score* on a driving skills test. The number of *hours* of instruction for each learner was also recorded. The results are displayed in the table below.

<i>Hours</i>	19	2	5	9	16	4	19	26	14	8
<i>Score</i>	32	12	17	19	23	16	28	36	30	23

- Which is the explanatory variable and which is the response variable?
 - Construct a scatterplot of these data.
 - Determine the correlation coefficient, r , and round your answer to 4 decimal places.
 - Describe the association between *score* and *hours* in terms of direction, strength and form (and outliers, if any).
 - Determine the equation for the least squares regression line and write it down in terms of the variables *score* and *hours*. Give coefficients rounded to three significant figures.
 - Interpret the slope and the intercept (if appropriate) of the regression line.
 - Predict the score after 10 hours of instruction to the nearest whole number.
- 3** To investigate the association between marks, in percentages, on an assignment and the final examination mark, data was collected from 10 students.

- a** The following scatterplot shows this data, with the least squares regression line added.

Use the scatterplot to determine the equation of the least squares regression line, and write it down in terms of the variables, final *exam mark* and *assignment mark*. Write the values in the equation, rounded to two significant figures.



- Interpret the intercept and slope of the least squares regression line in terms of the variables in the study.
- Use your regression equation to predict the final exam mark for a student who scored 50% on the assignment. Give your answer rounded to the nearest mark.
- How reliable is the prediction made in part **c**?
- The scatterplot shows that there is a strong positive linear association between the *assignment mark* and the final *exam mark*. The correlation coefficient is $r = 0.76$. Given this information, a student wrote: 'Good final exam marks are the result of good assignment marks'. Comment on this statement.