



# Investigating and modelling linear associations

## Chapter questions

- ▶ What is linear regression?
- ▶ What is a residual?
- ▶ What is a least squares line of best fit?
- ▶ How do you find the equation of the least squares line using summary statistics?
- ▶ How do you find the equation of the least squares line using technology?
- ▶ How do you interpret the intercept and slope of the least squares line?
- ▶ How do you use the equation of the least squares line to make predictions?
- ▶ How do you use the coefficient of determination in a regression analysis?
- ▶ What is a residual plot and how is it used?
- ▶ How do you report a regression analysis?

Once we identify a linear association between two numerical variables, we can fit a linear model to the data and find its equation. This equation gives us a better understanding of the nature of the relationship between the two variables, and we can also use the linear model to make predictions based on this understanding of the relationship.

## 3A Fitting a least squares regression line to numerical data

### Learning intentions

- ▶ To be able to define linear regression.
- ▶ To be able to define a residual.
- ▶ To introduce the least squares line of best fit.
- ▶ To be able to find the equation of the least squares line using summary statistics.
- ▶ To be able to find the equation of the least squares line using technology.

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**.

The equation of a line relating two variables  $x$  and  $y$  is of the form

$$y = a + bx$$

where  $a$  and  $b$  are constants. When the equation is written in this form:

- $a$  represents the coordinate of the point where the line crosses the  $y$ -axis (the  $y$ -intercept)
- $b$  represents the slope of the line.

In order to summarise any particular  $(x, y)$  data set, numerical values for  $a$  and  $b$  are needed that will ensure the line passes close to the data. There are several ways in which the values of  $a$  and  $b$  can be found.

The easiest way to fit a line to bivariate data is to construct a scatterplot and draw the line ‘by eye’. We do this by placing a ruler on the scatterplot so that it seems to follow the general trend of the data. You can then use the ruler to draw a straight line. Unfortunately, unless the points are very tightly clustered around a straight line, the results you get by using this method will differ a lot from person to person.

The more mathematical approach to fitting a straight line to data is to use the **least squares method**. This method assumes that the variables are linearly related, and works best when there are no clear outliers in the data.

### Some terminology

To explain the least squares method, we need to define several terms.

The scatterplot shows five data points,  $(x_1, y_1)$ ,  $(x_2, y_2)$ ,  $(x_3, y_3)$ ,  $(x_4, y_4)$  and  $(x_5, y_5)$ .

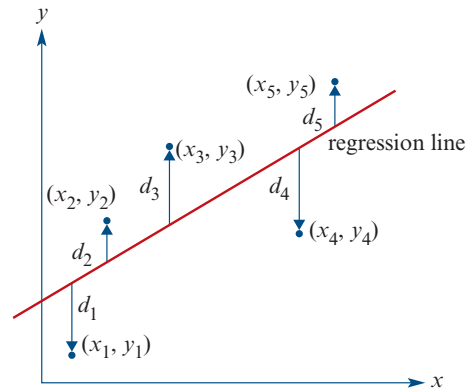
A regression line (not necessarily the least squares line) has also been drawn on the scatterplot.

The vertical distances  $d_1, d_2, d_3, d_4$  and  $d_5$  of each of the data points from the regression line are also shown.

These vertical distances,  $d$ , are known as **residuals**.

Residuals can be positive, negative or zero:

- Data points above the fitted regression line have a positive residual
- Data points below the fitted regression line have a negative residual
- Data points on the fitted regression line have zero residual.



## The least squares line

The least squares line is the line where the sum of the squares of the residuals is as small as possible; that is, it minimises:

$$\text{the sum of the squares of the residuals} = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2$$

Why do we minimise the sum of the *squares* of the residuals and not the sum of the residuals? This is because the sum of the residuals for the least squares line is always zero. The least squares line is like the mean. It balances out the data values on either side of itself. Some residuals are positive and some negative, and in the end they add to zero. Squaring the residuals solves this problem.

### The least squares line

The **least squares line** is the line that minimises the sum of the squares of the residuals.

The assumptions for fitting a least squares line to data are the same as for using the correlation coefficient,  $r$ . These are that:

- the data is numerical
- the association is linear
- there are no clear outliers.

## Determining the equation of the least squares regression line

To determine exactly the equation of the least squares regression line we need to determine the values of the intercept ( $a$ ) and the slope ( $b$ ) that define the line. The mathematics required is beyond the scope of this course, but calculus can be used to give us rules for these values:

### The equation of the least squares regression line

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

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

and

the **intercept** ( $a$ ) 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  $y$ .
- In these formulas  $y$  is the response variable, and  $x$  is the explanatory variable.

Note: The formula for the slope of the least squares regression line can be used to find the value of the correlation coefficient ( $r$ ), when the slope is known.

The **correlation coefficient** ( $r$ ) is given by 
$$r = \frac{bs_x}{s_y}$$

### Warning!

If you do not correctly decide which is the explanatory variable (the  $x$ -variable) and which is the response variable (the  $y$ -variable) before you start calculating the equation of the least squares regression line, you will get the wrong answer.



### Example 1 Determining the equation of the least squares regression line using summary statistics and the correlation coefficient

The height and weight of 11 people have been recorded, and the values of the following statistics determined:

	<i>height</i>	<i>weight</i>
mean	173.3 cm	65.45 kg
standard deviation	7.444 cm	7.594 kg
correlation coefficient	$r = 0.8502$	

Use the formula to determine the equation of the least squares regression line that enables *weight* to be predicted from *height*. Calculate the values of the slope and intercept rounded to two decimal places.

**Explanation**

- 1 Identify and write down the explanatory variable (EV) and the response variable (RV). Label as  $x$  and  $y$ , respectively.
- 2 Write down the given information.
- 3 Calculate the slope.
- 4 Calculate the intercept.
- 5 Use the values of the intercept and the slope to write down the least squares regression line using the variable names.

**Solution**EV: *height* ( $x$ )RV: *weight* ( $y$ )

$$\bar{x} = 173.3 \quad s_x = 7.444$$

$$\bar{y} = 65.45 \quad s_y = 7.594$$

$$r = 0.8502$$

Slope:

$$b = \frac{rs_y}{s_x} = \frac{0.8502 \times 7.594}{7.444}$$

$$= 0.87 \text{ (rounded to two significant figures)}$$

Intercept:

$$a = \bar{y} - b\bar{x}$$

$$= 65.45 - 0.87 \times 173.3$$

$$= -85 \text{ (rounded to two significant figures)}$$

$$y = -85 + 0.87x$$

or

$$\text{weight} = -85 + 0.87 \times \text{height}$$

**Example 2** Determining the correlation coefficient using the slope of the least squares regression line

Use the following information to find the value of the correlation coefficient  $r$ , rounded to three significant figures.

	<i>hours studied</i>	<i>exam score</i>
mean	5.87	68.3
standard deviation	1.34	5.42
least squares equation	$\text{exam score} = 52.7 + 2.45 \times \text{hours studied}$	

**Explanation**

- 1 Identify and write down the explanatory variable (EV) and the response variable (RV). Label as  $x$  and  $y$ , respectively.

**Solution**EV: *hours studied* ( $x$ )RV: *exam score* ( $y$ )

2 Write down the required information.

$$b = 2.45 \quad s_x = 1.34 \quad s_y = 5.42$$

3 Calculate the correlation coefficient.

Correlation coefficient:

$$r = \frac{bs_x}{s_y} = \frac{2.45 \times 1.34}{5.42}$$

$$= 0.61 \text{ (rounded to two significant figures)}$$

### CAS 1: How to determine and graph the equation of a least squares regression line using the TI-Nspire CAS

The following data give the height (in cm) and weight (in kg) of 11 people.

Height ( $x$ )	177	182	167	178	173	184	162	169	164	170	180
Weight ( $y$ )	74	75	62	63	64	74	57	55	56	68	72

Determine and graph the equation of the least squares regression line that will enable weight to be predicted from height. Write the intercept and slope rounded to three significant figures.

#### Steps

- 1 Start a new document by pressing  $\text{ctrl} + \text{N}$ .
- 2 Select **Add Lists & Spreadsheet**. Enter the data into lists named *height* and *weight*, as shown.
- 3 Identify the explanatory variable (EV) and the response variable (RV).

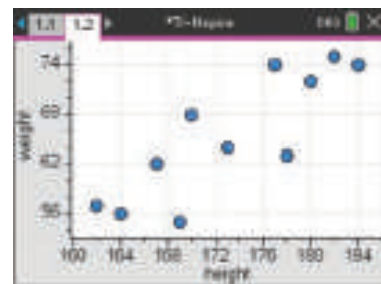
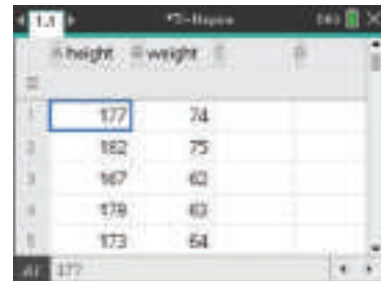
EV: *height*

RV: *weight*

**Note:** In saying that we want to predict *weight* from *height*, we are implying that *height* is the EV.

- 4 Press  $\text{ctrl} + \text{I}$  and select **Add Data & Statistics**. Construct a scatterplot with *height* (EV) on the horizontal (or  $x$ -) axis and *weight* (RV) on the vertical (or  $y$ -) axis.

Press  $\text{menu} > \text{Settings}$  and click the **Diagnostics** box. Select **OK** to activate this feature for *all* future documents. This will show the coefficient of determination ( $r^2$ ) whenever a regression is performed.

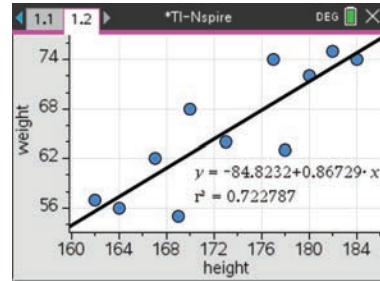


- 5 Press **[menu]**>**Analyze**>**Regression**>**Show Linear (a + bx)** to plot the regression line on the scatterplot.

Note that, simultaneously, the equation of the regression line is shown on the screen.

The equation of the regression line is:

$$\text{weight} = -84.8 + 0.867 \times \text{height}$$



The coefficient of determination is  $r^2 = 0.723$ , rounded to three significant figures.

### CAS 1: How to determine and graph the equation of a least squares regression line using the ClassPad

The following data give the height (in cm) and weight (in kg) of 11 people.

Height ( $x$ )	177	182	167	178	173	184	162	169	164	170	180
Weight ( $y$ )	74	75	62	63	64	74	57	55	56	68	72

Determine and graph the equation of the least squares regression line that will enable weight to be predicted from height. Write the intercept and slope rounded to three significant figures.

#### Steps

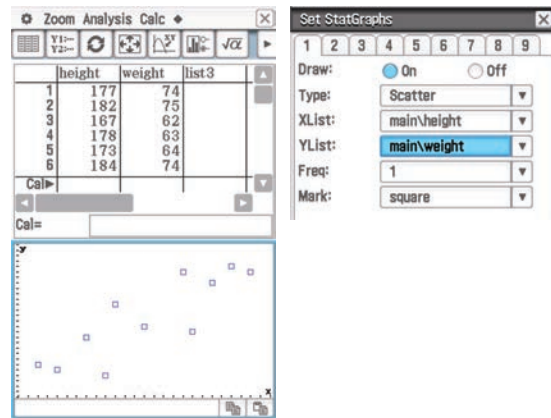
- 1 Open the **Statistics** application

and enter the data into columns labelled **height** and **weight**.

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

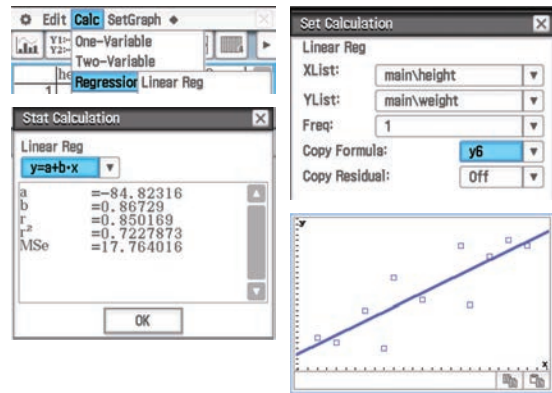
Tap **Set** to confirm your selections.

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





- 4 To calculate the equation of the least squares regression line:
- Tap **Calc** from the menu bar.
  - Tap **Regression** and select **Linear Reg.**
  - Complete the **Set Calculations** dialog box as shown.
  - Tap **OK** to confirm your selections in the **Set Calculations** dialog box. This also generates the regression results shown opposite.



- Tapping **OK** a second time automatically plots and displays the regression line.

**Note:** **y6** as the formula destination is an arbitrary choice.

- 5 Use the values of the intercept  $a$  and slope  $b$  to write the equation of the least squares line in terms of the variables *weight* and *height*.

$weight = -84.8 + 0.867 \times height$  (to three significant figures)

The coefficient of determination is  $r^2 = 0.723$ , rounded to three significant places.



## Exercise 3A

### Basic ideas

- 1 What is a residual?
- 2 The least-squares regression line is obtained by:
  - A minimising the residuals
  - B minimising the sum of the residuals
  - C minimising the sum of the squares of the residuals
  - D minimising the square of the sum of the residuals
  - E maximising the sum of the squares of the residuals.
- 3 Write down the three assumptions we make about the association we are modelling when we fit a least squares line to bivariate data.

### Using a formula to calculate the equation of a least square line

#### Example 1

- 4 A least squares line  $y = a + bx$  is calculated for a set of bivariate data.
  - a Write down the explanatory variable for this least squares line.
  - b Given the following information, determine the equation of the least squares line, giving the values of the intercept and slope rounded to three significant figures.



	$x$	$y$
mean	10.65	19.91
standard deviation	5.162	6.619
correlation coefficient	$r = 0.7818$	

- 5** We wish to find the equation of the least squares regression line that enables *pollution level* beside a freeway to be predicted from *traffic volume*.
- a** Which is the response variable (RV) and which is the explanatory variable (EV)?
- b** Use the formula to determine the equation of the least squares regression line that enables the pollution level to be predicted from the traffic volume where:

	<i>traffic volume</i>	<i>pollution level</i>
mean	11.4	231
standard deviation	1.87	97.9
correlation coefficient	$r = 0.940$	

Write the equation in terms of *pollution level* and *traffic volume* with the intercept and slope rounded to two significant figures.

- 6** We wish to find the equation of the least squares regression line that enables *life expectancy* in a country to be predicted from *birth rate*.
- a** Which is the response variable (RV) and which is the explanatory variable (EV)?
- b** Use the formula to determine the equation of the least squares regression line that enables life expectancy to be predicted from birth rate, where:

	<i>life expectancy</i>	<i>birth rate</i>
mean	55.1	34.8
standard deviation	9.99	5.41
correlation coefficient	$r = -0.810$	

Write the equation in terms of *life expectancy* and *birth rate* with the  $y$ -intercept and slope rounded to two significant figures.

#### Using a formula to calculate the correlation coefficient from the slope

##### Example 2

- 7** The equation of a least squares line  $y = a + bx$  is calculated for a set of bivariate data.
- a** Write down the response variable for this least-squares line.
- b** Use the following information to find the value of the correlation coefficient  $r$ , rounded to three significant figures.

	$x$	$y$
mean	12.51	10.65
standard deviation	4.796	5.162
least squares equation	$y = 16.72 - 0.4847x$	

- 8 The equation of the least squares regression line that enables *distance* travelled by a car (in 1000s of km) to be predicted from its *age* (in years) was found to be:

$$\text{distance} = 15.62 + 11.08 \times \text{age}$$

- a Which is the response variable (RV) and which is the explanatory variable (EV)?  
 b Use the following information to find the value of the correlation coefficient  $r$ , rounded to three significant figures.

	<i>distance</i>	<i>age</i>
mean	78.0	5.63
standard deviation	42.6	3.64

- 9 The following questions relate to the formulas used to calculate the slope and intercept of the least squares regression line.
- a A least squares line is calculated and the slope is found to be negative. What does this tell us about the sign of the correlation coefficient?  
 b The correlation coefficient is zero. What does this tell us about the slope of the least squares regression line?  
 c The correlation coefficient is zero. What does this tell us about the intercept of the least squares regression line?

Using a CAS calculator to determine the equation of the least squares line from data

- 10 The table shows the number of sit-ups and push-ups performed by six students.

<i>Sit-ups</i> ( $x$ )	52	15	22	42	34	37
<i>Push-ups</i> ( $y$ )	37	26	23	51	31	45

Let the number of *sit-ups* be the explanatory ( $x$ ) variable. Use your calculator to show that the equation of the least squares regression line is:

$$\text{push-ups} = 16.5 + 0.566 \times \text{sit-ups} \text{ (rounded to three significant figures)}$$

- 11 The table shows average hours worked and university participation rates (%) in six countries.

<i>Hours</i>	35.0	43.0	38.2	39.8	35.6	34.8
<i>Rate</i>	26	20	36	25	37	55

Use your calculator to show that the equation of the least squares regression line that enables participation *rates* to be predicted from *hours* worked is:

$$\text{rate} = 130 - 2.6 \times \text{hours} \text{ (rounded to two significant figures)}$$

- 12** The table shows the number of *runs* scored and *balls faced* by batsmen in a cricket match.

<i>Runs (y)</i>	27	8	21	47	3	15	13	2	15	10	2
<i>Balls faced (x)</i>	29	16	19	62	13	40	16	9	28	26	6

- a** Use your calculator to show that the equation of the least squares regression line enabling *runs* scored to be predicted from *balls faced* is:

$$y = -2.6 + 0.73x$$

- b** Rewrite the regression equation in terms of the variables involved.

- 13** The table below shows the number of TVs and cars owned (per 1000 people) in six countries.

<i>Number of TVs (y)</i>	378	404	471	354	381	624
<i>Number of cars (x)</i>	417	286	435	370	357	550

We wish to predict the *number of TVs* from the *number of cars*.

- a** Which is the response variable?  
**b** Show that, in terms of  $x$  and  $y$ , the equation of the regression line is:

$$y = 61.2 + 0.930x \text{ (rounded to three significant figures).}$$

- c** Rewrite the regression equation in terms of the variables involved.

### Exam 1 style questions

- 14** A least squares line of the form  $y = a + bx$  is fitted to a scatterplot. Which of the following statements is always true:
- A** The line will divide the data points so that there are as many points above the line as below the line.  
**B** The sum of the vertical distances from the line to each data point will be a minimum.  
**C**  $x$  is the explanatory variable and  $y$  is the response variable.  
**D**  $y$  is the explanatory variable and  $x$  is the response variable.  
**E** Most of the data points will lie on the line.

- 15 The statistical analysis of the set of bivariate data involving variables  $x$  and  $y$  resulted in the information displayed in the table below:

	$x$	$y$
mean	32.5	88.1
standard deviation	3.42	6.84
least squares equation	$y = -2.56 + 1.45x$	

Using this information the value of the correlation coefficient  $r$  for this set of bivariate data is closest to

- A 0.73      B 0.34      C 0.50      D 0.53      E 0.78
- 16 A retailer recorded the number of ice creams sold and the day's maximum temperature over 8 consecutive Saturdays one summer.

Temperature ( $^{\circ}\text{C}$ )	22	25	36	34	21	28	41	31
Number of ice creams sold	145	155	200	198	150	179	230	180

The equation of the least squares regression line fitted to the data is closest to:

- A *number of ice-creams* =  $4.08 + 58.2 \times \text{temperature}$   
 B *number of ice-creams* =  $-12.9 + 0.237 \times \text{temperature}$   
 C *number of ice-creams* =  $58.2 + 4.08 \times \text{temperature}$   
 D *temperature* =  $3.57 + 72.3 \times \text{number of ice-creams}$   
 E *temperature* =  $-12.8 + 0.237 \times \text{number of ice-creams}$

## 3B Using the least squares regression line to model a relationship between two numerical variables

### Learning intentions

- ▶ To be able to interpret the intercept and slope of the least squares line.
- ▶ To be able to use the equation of the least squares line to make predictions.
- ▶ To be able to use the coefficient of determination in a regression analysis.
- ▶ To be able to use a residual plot to investigate the linearity assumption.
- ▶ To be able to report a regression analysis.

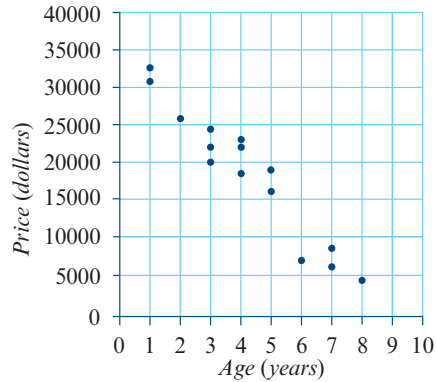
Suppose, for example, that we wish to investigate the nature of the association between the price of a secondhand car and its age. The ultimate aim is to find a mathematical model that will enable the price of a secondhand car to be predicted from its age.

The age (in years) and price (in dollars) of a selection of secondhand cars of the same brand and model have been collected and are recorded in a table (shown).

Age (years)	Price (dollars)	Age (years)	Price (dollars)	Age (years)	Price (dollars)
1	32 500	3	22 000	5	18 400
1	30 500	4	22 000	6	6 500
2	25 600	4	23 000	7	6 400
3	20 000	4	19 200	7	8 500
3	24 300	5	16 000	8	4 200

We start our investigation of the association between price and age by constructing a scatterplot and using it to describe the association in terms of strength, direction and form. In this analysis, *age* is the explanatory variable.

From the scatterplot, we see that there is a strong, negative, linear association between the price of the car and its age. There are no clear outliers. The correlation coefficient is  $r = -0.9643$ .



The equation of the least squares regression line from these data is:

$$price = 35\,100 - 3940 \times age$$

## Interpreting the slope and intercept of a regression line

### Interpreting the slope and intercept of a regression line

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

- the slope ( $b$ ) estimates the average change (increase/decrease) in the *response variable* ( $y$ ) for each one-unit increase in the *explanatory variable* ( $x$ )
- the intercept ( $a$ ) estimates the average value of the *response variable* ( $y$ ) when the *explanatory variable* ( $x$ ) equals 0.

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

Consider again the least squares regression line relating the *age* of a car to its *price*:

$$price = 35\,100 - 3940 \times age$$

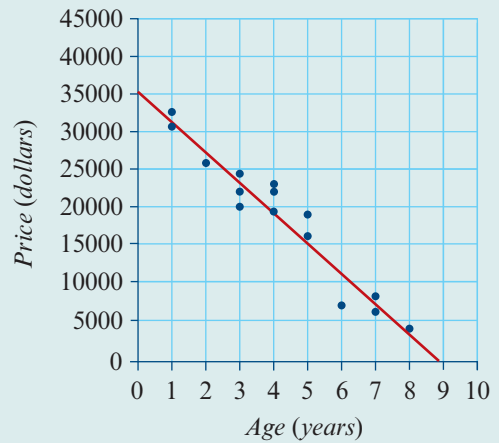
The two key values in this mathematical model are the intercept (35100) and the slope (−3940). The interpretation of these values is discussed in the following example.


**Example 3** Interpreting the slope and intercept of a regression line

The equation of a regression line that enables the *price* of a second-hand car to be predicted from its *age* is:

$$\text{price} = 35\,100 - 3940 \times \text{age}$$

- a** Interpret the slope in terms of the variables *price* and *age*.
- b** Interpret the intercept in terms of the variables *price* and *age*.


**Explanation**

- a** The slope predicts the average change (increase/decrease) in the *price* for each 1-year increase in the *age*. Because the slope is negative, it will be a decrease.
- b** The *intercept* predicts the value of the *price* of the car when *age* equals 0; that is, when the car is new.

**Solution**

On average, for each additional year of age the price of these cars decreases by \$3940.

On average, the price of these cars when new was \$35 100.

## Using the regression line to make predictions


**Example 4** Using the regression line to make predictions

The equation of a regression line that enables the *price* of a second-hand car to be predicted from its *age* is:

$$\text{price} = 35\,100 - 3940 \times \text{age}$$

Use this equation to predict the price of a car that is 5.5 years old.

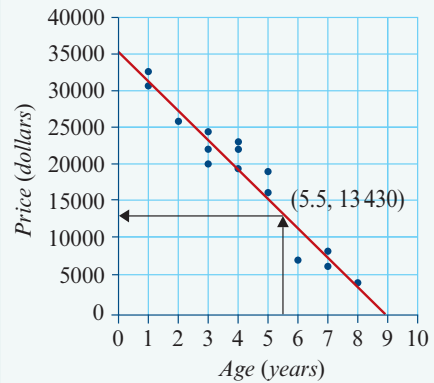
**Explanation**

There are two ways this can be done.

One is to draw a vertical arrow at  $age = 5.5$  up to the graph and then horizontally across to the *price* axis as shown, to get an answer of around \$14 000.

A more accurate answer is obtained by substituting  $age = 5.5$  into the equation to obtain \$13 430, as shown below.

$$\begin{aligned} \text{Price} &= 35\,100 - 3940 \times 5.5 \\ &= \$13\,430 \end{aligned}$$

**Solution****Interpolation and extrapolation**

When using a regression line to make predictions, we must be aware that, strictly speaking, the equation we have found applies only to the range of data values used to derive the equation.

For example, using the equation and rounding to the nearest dollar we would predict that:

- a car which is 2 years old would have a price of \$27 220 (price =  $35\,100 - 3940 \times 2$ )
- a car which is 7 years old would have a price of \$7520 (price =  $35\,100 - 3940 \times 7$ )
- a car which is 12 years old would have a price of \$-12 180 (price =  $35\,100 - 3940 \times 12$ )

This last result,  $-\$12\,180$  points to one of the limitations of substituting into a regression equation without thinking carefully. Using this regression equation, we have predicted a negative price, which is clearly not correct.

The problem is that we are using the regression equation to make predictions well outside the range of values used to calculate this equation. We only have data for cars which are up to 8 years old. Without knowing that the model works equally well for cars older than this, 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, from 1 to 8 years. 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* range of values of the explanatory variable is called **extrapolation**. Extrapolation is generally considered to give an **unreliable** prediction.

## The coefficient of determination

In the previous chapter we define the coefficient of determination as  $r^2$ , where  $r$  is the value of the correlation coefficient. The coefficient of determination can be considered a measure of the predictive power of a regression equation. While the association between the price of a second-hand car and its age does not explain all the variation in price, knowing the age of a car does give us some information about its likely price.

For a perfect relationship, the regression line explains 100% of the variation in prices. In this case, with  $r = -0.964$  we have the:

$$\text{coefficient of determination} = r^2 = (-0.9643)^2 = 0.930 \text{ or } 93.0\%$$

Thus, we can conclude that:

93% of the variation in price of the second-hand cars can be explained by the variation in the ages of the cars.

In this case, the regression equation has good predictive power. As a guide, any relationship with a coefficient of determination greater than 30% can be regarded as having good predictive power. In practice, even much lower values of the coefficient of determination can be useful.



### Example 5 Using the coefficient of determination to compare associations

In a recent study across a number of countries the correlation between educational attainment and the amount spent on education was found to be 0.26, whilst the correlation between educational attainment and the student : teacher ratio was found to be  $-0.38$ .

- Find the values of the coefficient of determination between *educational attainment* and the *amount spent on education*, and *student : teacher ratio* respectively.
- Which of the variables, *amount spent on education* or *student : teacher ratio* is more important in explaining the variation in educational attainment?

#### Solution

- educational attainment*:  $r^2 = 0.26^2 = 6.8\%$   
*student : teacher ratio*:  $r^2 = (-0.38)^2 = 14.4\%$

- b** The variable *student : teacher ratio* explains 14.4% of the variation in *educational attainment*, making it a more important explanatory variable than the *amount spent on education* which explains only 6.8%.

## The residual plot – assessing the appropriateness of fitting a linear model to data

So far all of our analysis has been based on the assumption that the relationship between the two variables of interest is linear. This is why it has been essential to examine the scatterplot before proceeding with any further analyses. However, sometimes the scatterplot is not sensitive enough to reveal the non-linear structure of a relationship. To gain more information we need to investigate the fit of the regression line to the data, and we do this using a **residual plot**.

Residuals are defined as the **vertical** distances between the regression line and the actual data value.

### Residual plot

A residual plot is a graph of the **residuals** (plotted on the vertical axis) against the **explanatory variable** (plotted on the horizontal axis), where:

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

Remember residuals can be positive, negative or zero.

To determine the appropriateness of fitting the least squares regression line to these data we will construct a residual plot. But first, we need to calculate the residual for each value of the explanatory variable, in this case *age*.



### Example 6 Calculating a residual

The actual price of the 6-year-old car is \$6500. Calculate the residual when its price is predicted using the regression equation:  $price = 35100 - 3940 \times age$

#### Explanation

- 1** Write down the actual price.
- 2** Determine the predicted price using the least squares regression equation:  
 $price = 35100 - 3940 \times age$
- 3** Determine the residual.

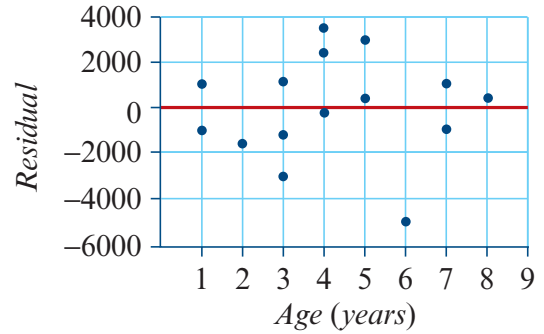
#### Solution

Actual price: \$6500

$$\begin{aligned} \text{Predicted price} &= 35\,100 - 3940 \times 6 \\ &= \$11\,460 \end{aligned}$$

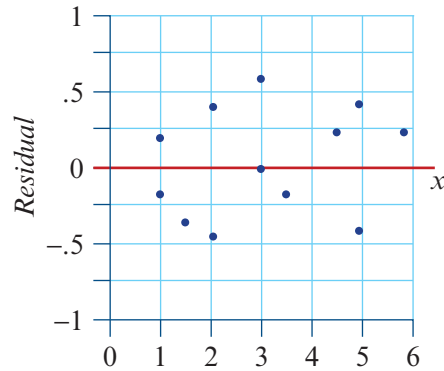
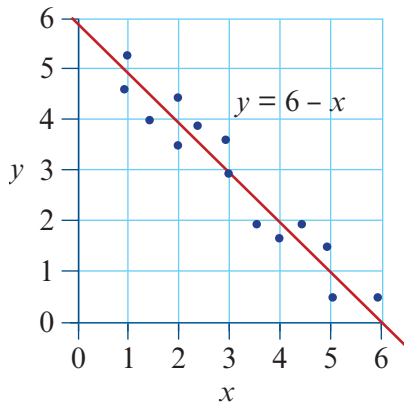
$$\begin{aligned} \text{Residual} &= \text{actual} - \text{predicted} \\ &= \$6500 - \$11\,460 \\ &= -\$4960 \end{aligned}$$

By completing this calculation for all data points, we can construct a residual plot. Because the mean of the residuals is always zero, we will construct the horizontal axis for the plot at zero (indicated by the red line) as shown.

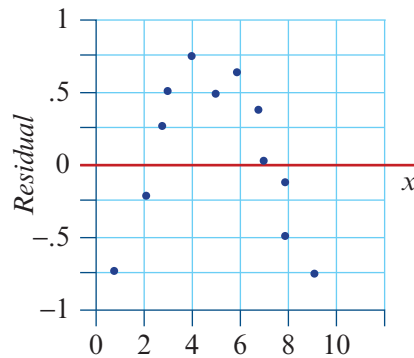
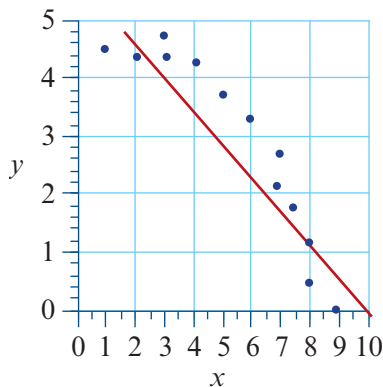


### What are we looking for in a residual plot?

The residual plot is used to check the **linearity assumption** required for a linear regression. The scatterplot below shows a relationship that is clearly linear. When a line is fitted to the data, the resultant residual plot appears to be a random collection of points roughly spread around zero (the horizontal red line in the residual plot).



By contrast, the relationship shown in the following scatterplot is clearly non-linear. Fitting a straight line to the data results in the residual plot shown. While there is some random behaviour, there is also a clearly identifiable curve shown in the scatterplot.

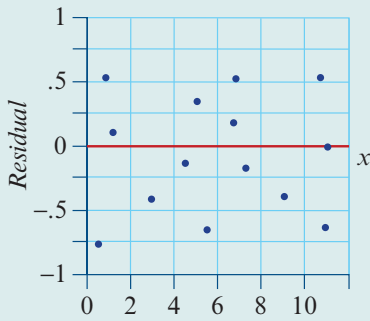


In summary, if a residual plot shows evidence of some sort of systematic behaviour (a pattern), then it is likely that the underlying relationship is non-linear. However, if the residual plot appears to be a random collection of points roughly spread around zero, then we can be happy that our original assumption of linearity was reasonable and that we have appropriately modelled the data. From a visual inspection, it is difficult to say with certainty that a residual plot is random. It is easier to see when it is not random. For present purposes, it is sufficient to say that a clear lack of a pattern in a residual plot is an indication of randomness.

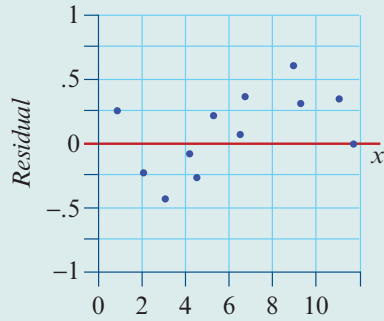


**Example 7** Interpreting a residual plot

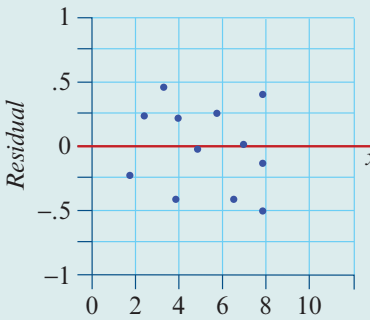
Which of the following residual plots would call into question the assumption of linearity in a regression analysis? Give reasons for your answers.



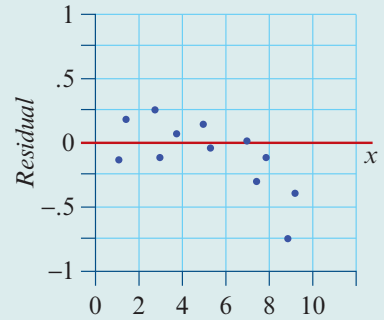
A



B



C



D

**Explanation**

Examine each plot, looking for a pattern or structure in the residual.

**Solution**

Plot *A* – residuals look random, so linearity assumption is met.

Plot *B* – there is a clear curve in the residuals, the linearity assumption is not met.

Plot *C* – residuals look random, so linearity assumption is met.

Plot *B* – there is a clear curve in the residuals, the linearity assumption is not met.

## Performing a regression analysis

A full regression analysis involves all of the following analyses, the results of which are collated in a report.

### Performing a regression analysis

To carry out a **regression analysis** involves several processes, which include:

- constructing a scatterplot to investigate the nature of an association
- calculating the correlation coefficient to indicate the strength of the relationship
- determining the equation of the regression line
- interpreting the coefficients of the  $y$ -intercept ( $a$ ) and the slope ( $b$ ) of the least squares regression line  $y = a + bx$
- calculating and interpreting the coefficient of determination
- using the regression line to make predictions
- calculating residuals and using a residual plot to test the assumption of linearity
- writing a report on your findings.

## Reporting the results of a regression analysis

The final step is to construct a report which brings together all of the analyses which have been described in this section, as shown in the following example.



### Example 8 Reporting the results of a regression analysis

Construct a report to describe the association between the price and age of secondhand cars.

#### Solution

From the scatterplot we see that there is a strong negative, linear association between the price of a second hand car and its age,  $r = -0.964$ . There are no obvious outliers.

The equation of the least squares regression line is:  $\text{price} = 35\,100 - 3940 \times \text{age}$ .

The slope of the regression line predicts that, on average, the price of these second-hand cars decreased by \$3940 each year.

The intercept predicts that, on average, the price of these cars when new was \$35 100.

The coefficient of determination indicates that 93% of the variation in the price of these second-hand cars is explained by the variation in their age.

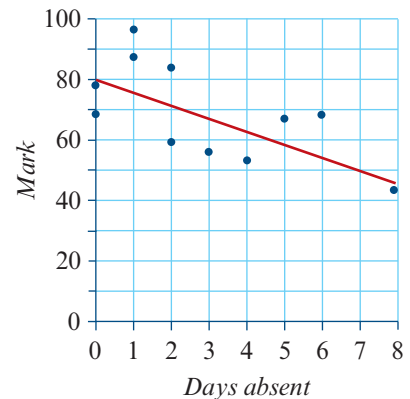
The lack of a clear pattern in the residual plot confirms the assumption of a linear association between the price and the age of these second-hand cars.



### Exercise 3B

#### Some basics

- Use the line on the scatterplot opposite to determine the equation of the regression line in terms of the variables, *mark* and *days absent*. Give the intercept correct to the nearest whole number and the slope correct to one decimal place.



#### Interpreting the intercept and slope of a regression line

##### Example 3

- The equation of a regression line that enables hand span (in cm) to be predicted from height (in cm) is:

$$\text{hand span} = 2.9 + 0.33 \times \text{height}$$

- Write down the value of the intercept, and interpret this value in this context of the variables in the equation.
- Write down the value of the slope, and interpret this value in this context of the variables in the equation.

- 3** The following regression equation can be used to predict a company's weekly sales (\$) from their weekly online advertising expenditure (\$).

$$\text{sales} = 575 + 4.85 \times \text{expenditure}$$

- a** Write down the value of the intercept, and interpret this value in this context of the variables in the equation.
- b** Write down the value of the slope, and interpret this value in this context of the variables in the equation.

#### Using the regression line to make predictions

##### Example 4

- 4** 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 ages. In each case indicate whether you are interpolating or extrapolating.

- a** 20 months old                      **b** 50 months old                      **c** 65 months old
- 5** When preparing between 25 and 100 meals, a hospital's cost (in dollars) is given by the equation:

$$\text{cost} = 487.50 + 6.70 \times \text{meals}$$

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

- a** 0 meals                                  **b** 80 meals                                  **c** 110 meals
- 6** For males of heights from 150 cm to 190 cm tall, the equation relating a *son's height* (in cm) to his *father's height* (in cm) is:

$$\text{son's height} = 83.9 + 0.525 \times \text{father's height}$$

Use this equation to predict (to the nearest cm) the adult height of a male whose father is the following heights. State, with a reason, how reliable your predictions are in each case.

- a** 170 cm tall                              **b** 200 cm tall                              **c** 155 cm tall

#### Using the coefficient of determination to compare associations

##### Example 5

- 7** A teacher found the correlation between her students' scores on an IQ test (*IQ*) and their final examination score in Year 12 (*exam score*) is 0.45, whilst the correlation between the average number of hours they spend each week studying mathematics (*hours*) and their final examination score in Year 12 (*exam score*) is 0.65.

- a** Determine the value of the coefficient of determination between *exam score* and *IQ*, expressed as a percentage rounded to one decimal place.
- b** Determine the value of the coefficient of determination between *exam score* and *hours*, expressed as a percentage rounded to one decimal place.



- c Which of the variables, *IQ* or *hours* is more important in explaining the variation in *exam score*?

Calculating a residual

Example 6

- 8 The equation of a regression line that enables hand span to be predicted from height is:

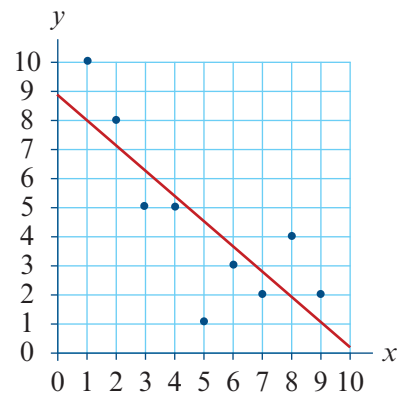
$$\text{hand span} = 2.9 + 0.33 \times \text{height}$$

- a Using this equation, show that the predicted hand span of a person who is 160 cm is 55.7 cm.
  - b This person has an actual hand span of 58.5 cm. Show that the residual value for this person is 2.8 cm.
- 9 For a 100 km trip, the equation of a regression line that enables fuel consumption of a car (in litres) to be predicted from its weight (kg) is:

$$\text{fuel consumption} = -0.1 + 0.01 \times \text{weight}$$

- a Use this equation to predict (to one decimal place) the fuel consumption of a car which weighs 980kg.
  - b This car has an actual fuel consumption of 8.9 litres. What is the residual value for this for this data point?
- 10 From the scatterplot shown determine (to the nearest whole number) the residual values when the value of *x* is equal to:

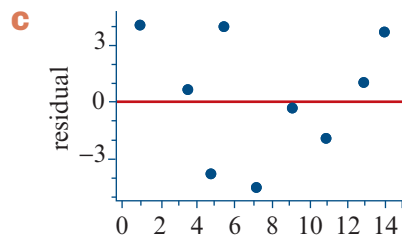
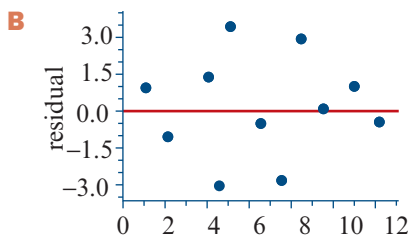
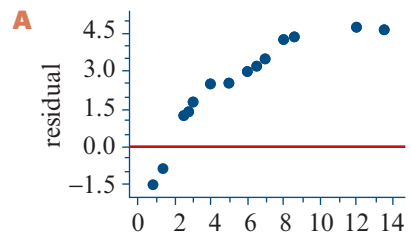
- a 1
- b 3
- c 8



Interpreting a residual plot

Example 7

- 11 Each of the following residual plots has been constructed after a least squares regression line has been fitted to a scatterplot. Which of the residual plots suggest that the use of a linear model to fit the data was inappropriate? Why?



- 12** In an investigation of the association between the food energy content (in calories) and the fat content (in g) in a standard-sized packet of chips, the least squares regression line was found to be:

$$\text{energy content} = 27.8 + 14.7 \times \text{fat content} \quad r^2 = 0.7569$$

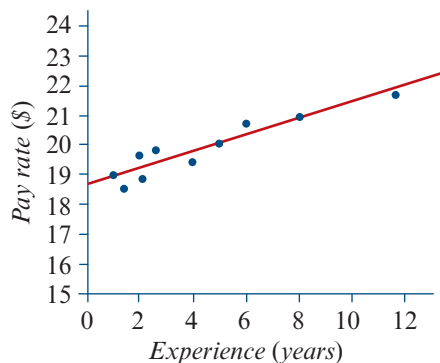
- Write down the value of the intercept, and interpret this value in this context of the variables in the equation.
  - Write down the value of the slope, and interpret this value in this context of the variables in the equation.
  - Interpret the value of the coefficient of determination in terms of the variables in *energy content* and *fat content*.
  - Use this equation to predict the energy content of a packet of chips which contains 8 grams of fat.
  - If the actual energy content of a packet of chips containing 8 grams of fat is 132 calories, what is the value of the residual?
- 13** In an investigation of the association between the success rate (%) of sinking a putt and the distance from the hole (in cm) of amateur golfers, the least squares regression line was found to be:

$$\text{success rate} = 98.5 - 0.278 \times \text{distance} \quad r^2 = 0.497$$

- Write down the slope of this regression equation and interpret.
- Use the equation to predict the success rate when a golfer is 90 cm from the hole.
- At what distance (in metres) from the hole does the regression equation predict an amateur golfer to have a 0% success rate of sinking the putt?
- Calculate the value of  $r$ , rounded to three decimal places.
- Write down the value of the coefficient of determination in percentage terms and interpret.

- 14** The scatterplot opposite shows the pay rate (dollars per hour) paid by a company to workers with different years of work experience. Using a calculator, the equation of the least squares regression line is found to have the equation:

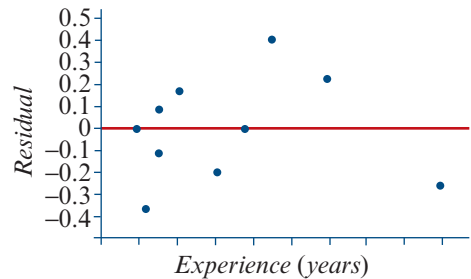
$$y = 18.56 + 0.289x \quad \text{with } r = 0.967$$



- Is it appropriate to fit a least squares regression line to the data? Why?
- Work out the coefficient of determination.
- What percentage of the variation in a person's pay rate can be explained by the variation in their work experience?

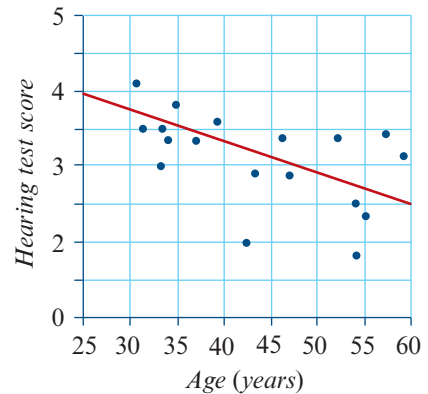
- d** Write down the equation of the least squares line in terms of the variables *pay rate* and years of *experience*.
- e** Interpret the *y*-intercept in terms of the variables *pay rate* and years of *experience*. What does the *y*-intercept tell you?
- f** Interpret the slope in terms of the variables *pay rate* and years of *experience*. What does the slope of the regression line tell you?
- g** Use the least squares regression equation to:
  - i** predict the hourly wage of a person with 8 years of experience
  - ii** determine the residual value if the person's actual hourly wage is \$21.20.

**h** The residual plot for this regression analysis is shown opposite. Does the residual plot support the initial assumption that the relationship between *pay rate* and years of *experience* is linear? Explain your answer.



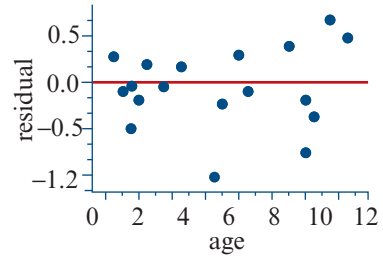
**15** The scatterplot opposite shows scores on a hearing test against age. In analysing the data, a statistician produced the following statistics:

- coefficient of determination:  $r^2 = 0.370$
- least squares line:  $y = 4.9 - 0.043x$
- a** Determine the value of Pearson's correlation coefficient,  $r$ , for the data.
- b** Interpret the coefficient of determination in terms of the variables *hearing test score* and *age*.



- c** Write down the equation of the least squares line in terms of the variables *hearing test score* and *age*.
- d** Write down the slope and interpret.
- e** Use the least squares regression equation to:
  - i** predict the hearing test score of a person who is 20 years old
  - ii** determine the residual value if the person's actual hearing test score is 2.0.
- f** Use the graph to estimate the value of the residual for the person aged:
  - i** 35 years
  - ii** 55 years.

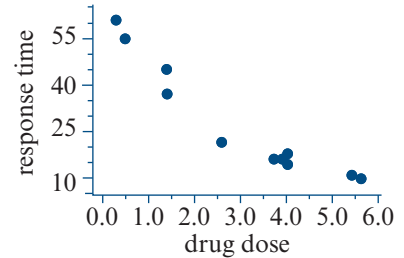
**g** The residual plot for this regression analysis is shown opposite. Does the residual plot support the initial assumption that the relationship between hearing test score and age is essentially linear? Explain your answer.



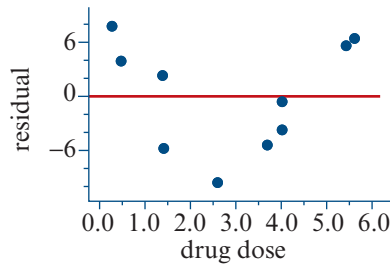
Reporting the results of a residual analysis

Example 8

**16** In a study of the effectiveness of a pain relief drug, the response time (in minutes) was measured for different drug doses (in mg). A least squares regression analysis was conducted to enable response time to be predicted from drug dose. The results of the analysis are displayed.



Regression equation:  $y = a + bx$   
 $a = 55.8947$   
 $b = -9.30612$   
 $r^2 = 0.901028$   
 $r = -0.949225$



Use this information to complete the following report. Call the two variables *drug dose* and *response time*. In this analysis *drug dose* is the explanatory variable.

Report

From the scatterplot we see that there is a strong  relationship between response time and  :  $r =$  . There are no obvious outliers.

The equation of the least squares regression line is:

response time =  +   $\times$  drug dose

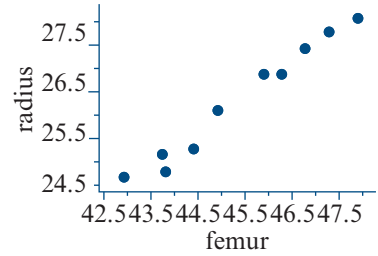
The slope of the regression line predicts that, on average, response time  increases/decreases by  minutes for a 1-milligram increase in drug dose.

The y-intercept of the regression line predicts that, on average, the response time when no drug is administered is  minutes.

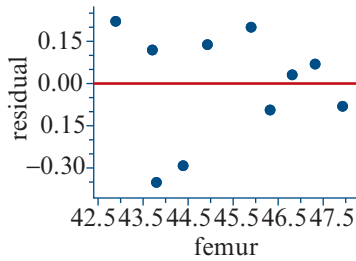
The coefficient of determination indicates that, on average,  % of the variation in  is explained by the variation in .

The residual plot shows a , calling into question the use of a linear equation to describe the relationship between response time and drug dose.

- 17** A regression analysis was conducted to investigate the nature of the association between femur (thigh bone) length and radius (the short thicker bone in the forearm) length in 18-year-old males. The bone lengths are measured in centimetres. The results of this analysis are reported below. In this investigation, femur length was treated as the independent variable.



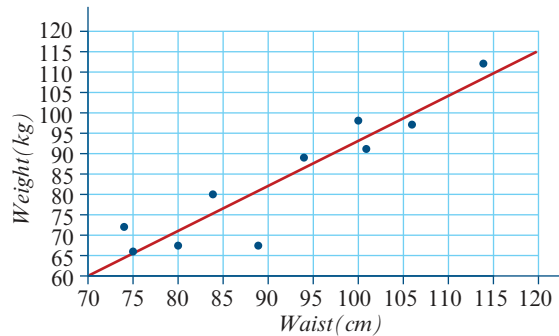
Regression equation  $y = a + bx$   
 $a = -7.24946$   
 $b = 0.739556$   
 $r^2 = 0.975291$   
 $r = 0.987568$



Use the format of the report given in the previous question to summarise findings of this investigation. Call the two variables *femur length* and *radius length*.

**Exam 1 style questions**

- 18** The scatterplot shows the weight (in kg) and waist measurement (in cm) for a group of people. A least squares line had been fitted to the scatterplot with *waist* as the explanatory variable. The equation of the least squares line is closest to:



- A**  $weight = 60.0 + 1.10 \times waist$
- B**  $waist = 60.0 + 0.91 \times weight$
- C**  $weight = 70.0 + 1.10 \times waist$
- D**  $weight = -3.70 + 0.91 \times waist$
- E**  $weight = -17.0 + 1.10 \times waist$

- 19** The table below shows the *life expectancy* in years and the percentage of government expenditure which is spent on health (*health*) in 10 countries.

<i>Health</i>	17.3	10.3	4.7	6.0	20.1	6.0	13.2	7.7	10.1	17.5
<i>Life expectancy (years)</i>	82	76	68	69	83	75	76	76	75	75

A least squares line which enables a country's *life expectancy* to be predicted from their expenditure on *health* is fitted to the data. The number of times that a country's predicted *life expectancy* is greater than their actual *life expectancy* is:

- A** 3
- B** 4
- C** 5
- D** 6
- E** 7

- 20 In a study of the association between the *length* in cm and *weight* in grams of a certain species of fish the following least squares line was obtained:

$$\text{weight} = -329 + 23.3 \times \text{length}$$

Which one of the following is a conclusion that can be made from this least squares line?

- A On average, the *weight* of the fish increased by 23.3 grams for each centimetre increase in *length*.
- B On average, the *length* of the fish increased by 23.3 cm for each one gram increase in *weight*.
- C On average, the *weight* of the fish decreased by 329 grams for each centimetre increase in *length*.
- D The equation cannot be correct as the *weight* of the fish can never be negative.
- E The *weight* of the fish in grams can be determined by subtracting 305.7 from their *length*.

## 3C Conducting a regression analysis using data

In your statistical investigation project you will need to be able to conduct a full regression analysis from data. This section is designed to help you with this task.

### CAS 2: How to conduct a regression analysis using the TI-Nspire CAS

This analysis is concerned with investigating the association between life expectancy (in years) and birth rate (in births per 1000 people) in 10 countries.

<i>Birth rate</i>	30	38	38	43	34	42	31	32	26	34
<i>Life expectancy (years)</i>	66	54	43	42	49	45	64	61	61	66

#### Steps

- Write down the explanatory variable (EV) and response variable (RV). Use the variable names *birth* and *life*.

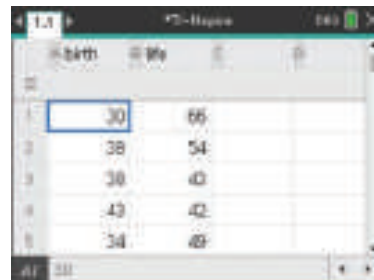
EV: *birth*

RV: *life*

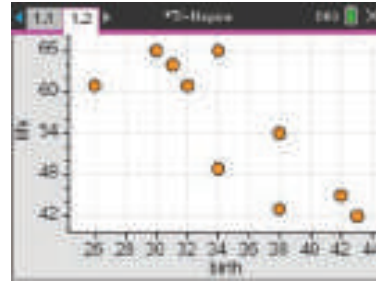
- Start a new document by pressing **ctrl** + **N**.

Select **Add Lists & Spreadsheet**.

Enter the data into the lists named *birth* and *life*, as shown.



**3** Construct a scatterplot to investigate the nature of the relationship between life expectancy and birth rate.

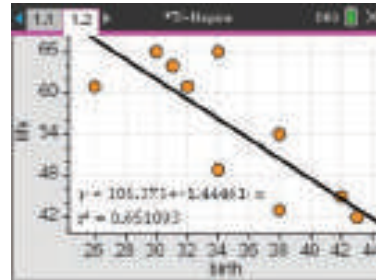


There is a strong, negative, linear relationship between life expectancy and birth rate. There are no obvious outliers.

**4** Describe the association shown by the scatterplot. Mention direction, form, strength and outliers.

**5** Find and plot the equation of the least squares regression line and  $r^2$  value.

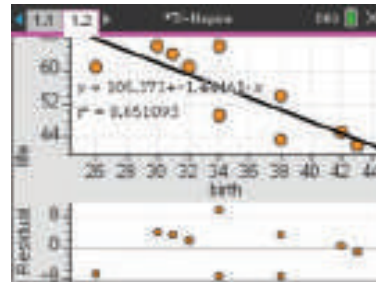
**Note:** Check if **Diagnostics** is activated using **menu** > **Settings**.



**6** Generate a residual plot to test the linearity assumption.

Use **ctrl** + **◀** (or click on the page tab) to return to the scatterplot.

To hide the residual plot press **menu** > **Analyze** > **Residuals** > **Hide Residual Plot**.



**7** Use the values of the intercept and slope to write the equation of the least squares regression line. Also write the values of  $r$  and the coefficient of determination.

Regression equation:

$$life = 105.4 - 1.445 \times birth$$

Correlation coefficient:  $r = -0.8069$

Coefficient of determination:  $r^2 = 0.651$

### CAS 2: How to conduct a regression analysis using the ClassPad

The data for this analysis are shown below.

<i>Birth rate (per thousand)</i>	30	38	38	43	34	42	31	32	26	34
<i>Life expectancy (years)</i>	66	54	43	42	49	45	64	61	61	66

#### Steps

**1** Write down the explanatory variable (EV) and response variable (RV). Use the variable names *birth* and *life*.

EV: *birth*

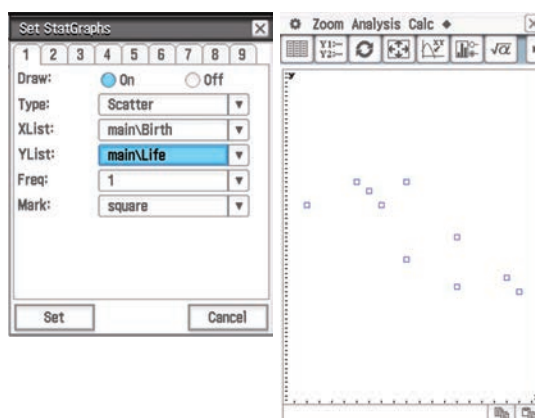
RV: *life*



- 2 Enter the data into lists as shown.
- 3 Construct a scatterplot to investigate the nature of the relationship between life expectancy and birth rate.

	Birth	Life	list3
1	30	66	
2	38	54	
3	38	43	
4	43	42	
5	34	49	
6	42	45	
7	31	64	
8	32	61	
9	26	61	
10	34	66	
11			

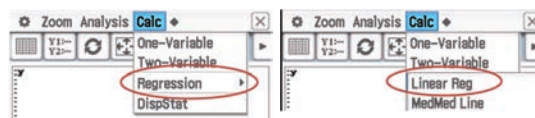
- a Tap and complete the **Set Calculations** dialog box as shown.
- b Tap to view the scatterplot.



- 4 Describe the association shown by the scatterplot. Mention direction, form, strength and outliers.
- 5 Find the equation of the least squares regression line and generate all regression statistics, including residuals.

There is a strong negative, linear association between life expectancy and birth rate. There are no obvious outliers.

- a Tap **Calc** in the toolbar.  
Tap **Regression** and select **Linear Reg**.
- b Complete the **Set Calculations** dialog box as shown.
- Note:** **Copy Residual** copies the residuals to **list3**, where they can be used later to create a residual plot.



- c Tap **OK** in the **Set Calculation** box to generate the regression results.

d Write down the key results.

Regression equation:

$$life = 105.4 - 1.445 \times birth$$

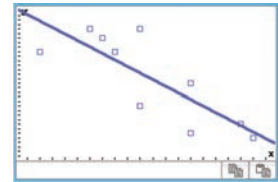
Correlation coefficient:

$$r = -0.8069$$

Coefficient of determination:



$$r^2 = 0.651$$

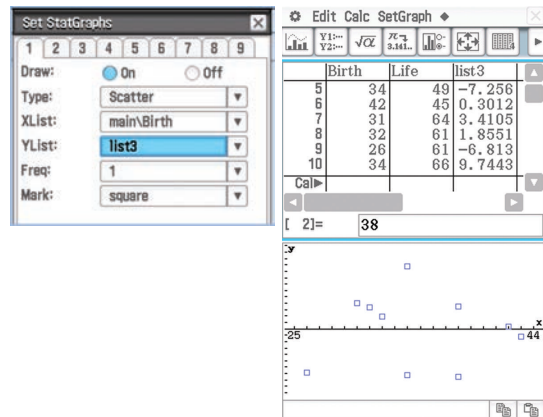
6 Tapping **OK** a second time automatically plots and displays the regression line on the scatterplot.



To obtain a full-screen plot, tap  from the icon panel.

7 Generate a residual plot to test the linearity assumption.

Tap  and complete the **Set Calculations** dialog box as shown. Tap  to view the residual plot.



Inspect the plot and write your conclusion.

The random residual plot suggests linearity.

**Note:** When you performed a regression analysis earlier, the residuals were calculated automatically and stored in **list3**. The residual plot is a scatterplot with **list3** on the vertical axis and **birth** on the horizontal axis.

### Exercise 3C

1 The table below shows the scores obtained by nine students on two tests. We want to be able to predict test B scores from test A scores.

Test A score ( $x$ )	18	15	9	12	11	19	11	14	16
Test B score ( $y$ )	15	17	11	10	13	17	11	15	19

Use your calculator to perform each of the following steps of a regression analysis.

- Construct a scatterplot. Name the variables *test a* and *test b*.
- Determine the equation of the least squares line along with the values of  $r$  and  $r^2$ .

- c** Display the regression line on the scatterplot.  
**d** Obtain a residual plot.
- 2** The table below shows the number of careless errors made on a test by nine students. Also given are their test scores. We want to be able to predict test score from the number of careless errors made.

<i>Test score</i>	18	15	9	12	11	19	11	14	16
<i>Careless errors</i>	0	2	5	6	4	1	8	3	1

Use your calculator to perform each of the following steps of a regression analysis.

- a** Construct a scatterplot. Name the variables *score* and *errors*.  
**b** Determine the equation of the least squares line along with the values of  $r$  and  $r^2$ . Write answers rounded to three significant figures.  
**c** Display the regression line on the scatterplot.  
**d** Obtain a residual plot.
- 3** How well can we predict an adult's weight from their birth weight? The weights of 12 adults were recorded, along with their birth weights. The results are shown.

<i>Birth weight (kg)</i>	1.9	2.4	2.6	2.7	2.9	3.2	3.4	3.4	3.6	3.7	3.8	4.1
<i>Adult weight (kg)</i>	47.6	53.1	52.2	56.2	57.6	59.9	55.3	58.5	56.7	59.9	63.5	61.2

- a** In this investigation, which would be the RV and which would be the EV?  
**b** Construct a scatterplot.  
**c** Use the scatterplot to:
  - comment on the association between adult weight and birth weight in terms of direction, outliers, form and strength
  - estimate the value of Pearson's correlation coefficient,  $r$ .**d** Determine the equation of the least squares regression line, the coefficient of determination and the value of Pearson's correlation coefficient,  $r$ . Write answers rounded to three significant figures.  
**e** Interpret the coefficient of determination in terms of adult weight and birth weight.  
**f** Interpret the slope in terms of adult weight and birth weight.  
**g** Use the regression equation to predict the weight of an adult with a birth weight of:
  - 3.0 kg
  - 2.5 kg
  - 3.9 kg.
 Give answers correct to one decimal place.  
**h** It is generally considered that birth weight is a 'good' predictor of adult weight. Do you think the data support this contention? Explain.  
**i** Construct a residual plot and use it to comment on the appropriateness of assuming that adult weight and birth weight are linearly associated.

## Key ideas and chapter summary



### Bivariate data

**Bivariate data** are data in which each observation involves recording information about two variables for the same person or thing. An example would be the heights and weights of the children in a preschool.

### Linear regression

The process of fitting a line to data is known as **linear regression**. 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**.

### Residuals

The vertical distance from a data point to the straight line is called a **residual**: residual value = data value – predicted value.

### Least squares method

The **least squares method** is one way of finding the equation of a regression line. It minimises the sum of the squares of the residuals. It works best when there are no outliers.

### Determining the values of $a$ and $b$ from the formulas

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

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

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  $y$ .

### Determining the value of $r$ when $b$ is known

The value of the correlation coefficient  $r$  is given by

$$r = \frac{bs_x}{s_y}$$

Here:

- $b$  is the **slope** of the least squares line
- $s_x$  and  $s_y$  are the **standard deviations** of  $x$  and  $y$

**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.

Consider for example the regression line

$$\text{cost} = 1.20 + 0.06 \times \text{number of pages}$$

The slope of the regression line tells us that on average the cost of a textbook increases by 6 cents (\$0.06) for each additional page.

The *intercept* of the line tells us that on average that a book with no pages costs \$1.20 (this might be the cost of the cover).

**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. For example, using the previous equation

$$\text{cost} = 1.20 + 0.06 \times \text{number of pages}$$

predicts that the cost of a 100-page book is:

$$\text{cost} = 1.20 + 0.06 \times 100 = \$7.20$$

**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.

**Coefficient of determination**

The **coefficient of determination** ( $r^2$ ) gives a measure of the predictive power of a regression line. For example, for the regression line above, the coefficient of determination is 0.81.

From this we conclude that 81% of the variation in the cost of a textbook can be explained by the variation in the number of pages.

**Residual plots**

**Residual plots** can be used to test the linearity assumption by plotting the residuals against the EV.

A residual plot that appears to be a random collection of points clustered around zero supports the linearity assumption.

A residual plot that shows a clear pattern indicates that the association is not linear.

## Skills checklist



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

- |  |   |                          |
|--|---|--------------------------|
| <b>3A</b>                                  | <b>1</b> I can determine the equation of the least squares regression line using the formulas.                              | <input type="checkbox"/> |
| See Example 1, and Exercise 3A Question 4  |   |                          |
| <b>3A</b>                                  | <b>2</b> I can determine the correlation coefficient from the slope of the least squares regression line using the formula. | <input type="checkbox"/> |
| See Example 2, and Exercise 3A Question 7  |   |                          |
| <b>3A</b>                                  | <b>3</b> I can determine the equation of the least squares regression line using a CAS calculator.                          | <input type="checkbox"/> |
| See CAS 1, and Exercise 3A Question 10     |   |                          |
| <b>3B</b>                                  | <b>4</b> I can interpret the slope and intercept of a regression line.  | <input type="checkbox"/> |
| See Example 3, and Exercise 3B Question 2  |   |                          |
| <b>3B</b>                                  | <b>5</b> I can use the regression line to make predictions.   | <input type="checkbox"/> |
| See Example 4, and Exercise 3B Question 4  |   |                          |
| <b>3B</b>                                  | <b>6</b> I can use the coefficient of determination to compare associations.  | <input type="checkbox"/> |
| See Example 5, and Exercise 3B Question 7  |   |                          |
| <b>3B</b>                                  | <b>7</b> I can calculate residual values.   | <input type="checkbox"/> |
| See Example 6, and Exercise 3B Question 8  |   |                          |
| <b>3B</b>                                  | <b>8</b> I can interpret a residual plot.   | <input type="checkbox"/> |
| See Example 7, and Exercise 3B Question 11 |   |                          |
| <b>3B</b>                                  | <b>9</b> I can write a report based on a regression analysis.   | <input type="checkbox"/> |
| See Example 8, and Exercise 3B Question 16 |   |                          |
| <b>3C</b>                                  | <b>10</b> I can use a CAS calculator to generate all of the analyses required for a regression analysis.                    | <input type="checkbox"/> |
| See CAS 2, and Exercise 3C Question 1      |   |                          |

## Multiple-choice questions

1 When using a least squares line to model a relationship displayed in a scatterplot, one key assumption is that:

- A** there are two variables                      **B** the variables are related  
**C** the variables are linearly related        **D**  $r^2 > 0.5$   
**E** the correlation coefficient is positive

2 For the least squares regression line  $y = -1.2 + 0.52x$ :

- A** the y-intercept =  $-0.52$             and            slope =  $-1.2$   
**B** the y-intercept =  $0$                     and            slope =  $-1.2$   
**C** the y-intercept =  $0.52$             and            slope =  $-1.2$   
**D** the y-intercept =  $-1.2$             and            slope =  $0.52$   
**E** the y-intercept =  $1.2$                 and            slope =  $-0.52$

3 If the equation of a least squares regression line is  $y = 8 - 9x$  and  $r^2 = 0.25$ :

- A**  $r = -0.5$         **B**  $r = -0.25$         **C**  $r = -0.0625$         **D**  $r = 0.25$         **E**  $r = 0.50$

4 Given that  $b = 1.328$ ,  $s_x = 1.871$  and  $s_y = 3.391$ , the correlation coefficient,  $r$ , is closest to:

- A** 0.357            **B** 0.598            **C** 0.733            **D** 0.773            **E** 1.33

5 The association between the number of errors made in a task, and the time spent practicing the task (in minutes) was found to be approximately linear, and the values of the following statistics were determined:

	<i>time</i>	<i>errors</i>
mean	8.00	34.5
standard deviation	2.40	12.5
correlation coefficient	$r = -0.236$	

The equation of the least squares line that enables errors to be predicted from time is given by

- A**  $errors = 52.2 - 1.23 \times time$                       **B**  $errors = 10.1 - 0.99 \times time$   
**C**  $errors = 24.7 - 1.23 \times time$                       **D**  $errors = 32.6 + 0.24 \times time$   
**E**  $errors = 44.3 - 1.23 \times time$

6 The speed at which a car is travelling (in km/hr), and the distance (in metres) taken by the car to come to a stop when the brakes are applied, were recorded over speeds from 60km/hr to 120km/hr.

	<i>speed</i>	<i>distance</i>
mean	90.5	52.7
standard deviation	1.124	1.349
correlation coefficient	$r = 0.948$	

The association was found to be approximately linear, and the values of the statistics shown were determined.



On average, for each additional km/hr of *speed*, the *distance* taken to come to a stop

- A** decreased by 1.14 metres                      **B** decreased by 0.79 metres  
**C** increased by 1.14 metres                    **D** increased by 0.95 metres  
**E** increased by 0.79 metres

**7** The least squares regression line  $y = 8 - 9x$  predicts that, when  $x = 5$ , the value of  $y$  is:

- A** -45                      **B** -37                      **C** 37                      **D** 45                      **E** 53

**8** A least squares regression line of the form  $y = a + bx$  is fitted to the data set shown.

$x$	25	15	10	5
$y$	10	10	15	25

The equation of the line is:

- A**  $y = -0.69 + 24.4x$                       **B**  $y = 24.4 - 0.69x$                       **C**  $y = 24.4 + 0.69x$   
**D**  $y = 28.7 - x$                               **E**  $y = 28.7 + x$

**9** A least squares regression line of the form  $y = a + bx$  is fitted to the data set shown.

$y$	30	25	15	10
$x$	40	20	30	10

The equation of the line is:

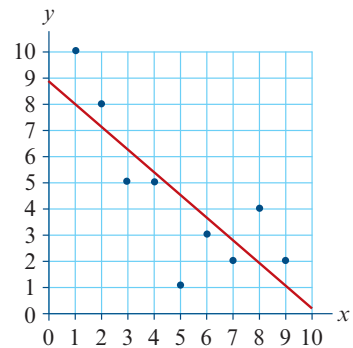
- A**  $y = 1 + 0.5x$                               **B**  $y = 0.5 + x$                               **C**  $y = 0.5 + 7.5x$   
**D**  $y = 7.5 + 0.5x$                               **E**  $y = 30 - 0.5x$

**10** Using a least squares regression line, the predicted value of a data value is 78.6. The residual value is -5.4. The actual data value is:

- A** 73.2                      **B** 84.0                      **C** 88.6                      **D** 94.6                      **E** 424.4

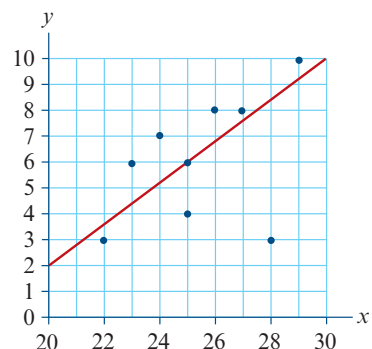
**11** The equation of the least squares line plotted on the scatterplot opposite is closest to:

- A**  $y = 8.7 - 0.9x$   
**B**  $y = 8.7 + 0.9x$   
**C**  $y = 0.9 - 8.7x$   
**D**  $y = 0.9 + 8.7x$   
**E**  $y = 8.7 - 0.1x$



**12** The equation of the regression line plotted on the scatterplot opposite is closest to:

- A**  $y = -14 + 0.8x$   
**B**  $y = 0.8 + 14x$   
**C**  $y = 2.5 + 0.8x$   
**D**  $y = 14 - 0.8x$   
**E**  $y = 17 + 1.2x$



The following information relates to Questions 13 to 16.

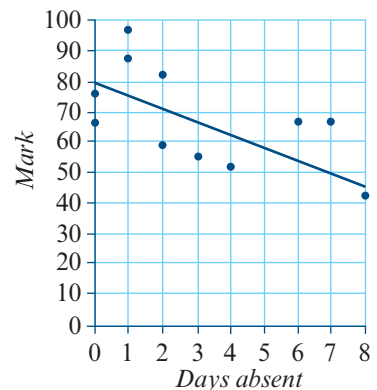
Weight (in kg) can be predicted from height (in cm) from the regression line:

$$\text{weight} = -96 + 0.95 \times \text{height}, \text{ with } r = 0.79$$

- 13** Which of the following statements relating to the regression line is *false*?
- A** The slope of the regression line is 0.95.
  - B** The explanatory variable in the regression equation is *height*.
  - C** The least squares line does *not* pass through the origin.
  - D** The intercept is 96.
  - E** The equation predicts that a person who is 180 cm tall will weigh 75 kg.
- 14** This regression line predicts that, on average, weight:
- A** decreases by 96 kg for each 1 centimetre increase in height
  - B** increases by 96 kg for each 1 centimetre increase in height
  - C** decreases by 0.79 kg for each 1 centimetre increase in height
  - D** decreases by 0.95 kg for each 1 centimetre increase in height
  - E** increases by 0.95 kg for each 1 centimetre increase in height
- 15** Noting that the value of the correlation coefficient is  $r = 0.79$ , we can say that:
- A** 62% of the variation in weight can be explained by the variation in height
  - B** 79% of the variation in weight can be explained by the variation in height
  - C** 88% of the variation in weight can be explained by the variation in height
  - D** 79% of the variation in height can be explained by the variation in weight
  - E** 95% of the variation in height can be explained by the variation in weight
- 16** A person of height 179 cm weighs 82 kg. If the regression equation is used to predict their weight, then the residual will be closest to:
- A** -8 kg      **B** 3 kg      **C** 8 kg      **D** 9 kg      **E** 74 kg

The following information relates to Questions 17 to 21.

The scatterplot shows the association between a student's *mark* on a test, and the number of *days absent* during the term.



- 17** The median *mark* for this group of students is closest to:  
**A** 55                    **B** 60                    **C** 67                    **D** 70                    **E** 72
- 18** The median *days absent* for this group of students is closest to:  
**A** 2                        **B** 3                        **C** 4                        **D** 55                    **E** 62.5
- 19** The coefficient of determination between *mark* and *days absent* is  $r^2 = 0.5$ .  
 The correlation coefficient is closest to:  
**A** -0.7                    **B** -0.25                    **C** 0.25                    **D** 0.5                    **E** 0.7
- 20** There were two students who were absent for 2 days that term. The values of the residuals for these students are  
**A** 0                        **B** 10                        **C** 60 and 80                    **D** -10 and 10                    **E** -10
- 21** Using the graph of the least squares line, we predict that a student who is absent for 4 days would receive a mark of about:  
**A** 48                        **B** 51                        **C** 62                        **D** 65                        **E** 67
- 22** The table below shows the *weight* in grams and the *length* in cm for a certain species of fish.

<i>Length(cm)</i>	13.5	14.3	16.3	17.5	18.4	19.0	19.0	19.8	21.2	23.0
<i>Weight(gm)</i>	55	60	90	120	150	140	170	145	200	273

A least squares line which enables a fish's *weight* to be predicted from their *length* is fitted to the data. The number of times that the fish's predicted *weight* is greater than their actual *weight* is:

- A** 3                        **B** 4                        **C** 5                        **D** 6                        **E** 7
- 23** The value of the correlation coefficient  $r$  for these data is equal to 0.965. The percentage of variation in fish *weight* which is not explained by the *length* of the fish is closest to:  
**A** 96.5%                    **B** 93.1%                    **C** 9.3%                    **D** 6.9%                    **E** 3.5%

### Written response questions

- 1** The table below shows the *age* (in years), the *number of seats*, and the *airspeed* (in km/h), of eight aircraft.

<i>Age</i>	3.5	3.7	4.7	4.9	5.1	7.3	8.7	8.8
<i>number of seats</i>	405	296	288	258	240	193	188	148
<i>airspeed</i>	830	797	774	736	757	765	760	718

- a** Determine to the nearest whole number:
- the median *age* of these aircraft.
  - the mean and standard deviation of the *airspeed* of these aircraft.

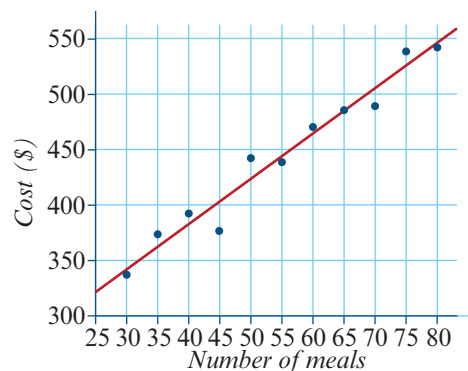
To investigate the association between the *number of seats* and *airspeed*, a least squares line is fitted to the data. The response variable in this investigation is *airspeed*.

- Determine and write down the equation of the least squares line in terms of *number of seats* and *airspeed*. Round the intercept and slope to 3 significant figures.
  - Determine and write down the percentage of variation in the *airspeed* that is explained by the *number of seats*. Write the answer rounded to 1 decimal place.
- 2** In an investigation of the relationship between the hours of sunshine (per year) and days of rain (per year) for 25 cities, the least squares regression line was found to be:

$$\text{hours of sunshine} = 2850 - 6.88 \times \text{days of rain}, \text{ with } r^2 = 0.484$$

Use this information to complete the following sentences.

- In this regression equation, the explanatory variable is .
  - The slope is  and the intercept is .
  - The regression equation predicts that a city that has 120 days of rain per year will have  hours of sunshine per year.
  - The slope of the regression line predicts that the hours of sunshine per year will  by  hours for each additional day of rain.
  - $r =$  , correct to three significant figures.
  - % of the variation in sunshine hours can be explained by the variation in .
  - One of the cities used to determine the regression equation had 142 days of rain and 1390 hours of sunshine.
    - The regression equation predicts that it has  hours of sunshine.
    - The residual value for this city is  hours.
  - Using a regression line to make predictions within the range of data used to determine the regression equation is called .
- 3** The cost of preparing meals, in dollars, and the number of meals prepared are plotted in the scatterplot shown. A least squares line has been fitted to the data which enables the cost of the meals prepared to be predicted from the number of meals prepared.



- a** Which is the response variable?  
**b** Describe the association in terms of strength, direction and form.

The equation of the least squares line that relates the cost of preparing meals to the number of meals produced is:

$$\text{cost} = 222.48 + 4.039 \times \text{number of meals}$$

- c** **i** Use the equation to predict the cost of preparing 21 meals. Round the answer to the nearest cent.  
**ii** In making this prediction, are you interpolating or extrapolating?  
**d** Write down:  
**i** the intercept of the regression line and interpret in terms of *cost* and the *number of meals* prepared.  
**ii** the slope of the regression line and interpret in terms of *cost* and the *number of meals* prepared.  
**e** When the number of meals prepared was 50, the cost of preparation was \$444. Show that, when the least squares line is used to predict the cost of preparing 50 meals, the residual is \$19.57, to the nearest cent.
- 4** We wish to find the equation of the least squares regression line that will enable height (in cm) to be predicted from femur (thigh bone) length (in cm).

- a** Which is the RV and which is the EV?

- b** Use the summary statistics shown to determine the equation of the least squares regression line that will enable *height* to be predicted from *femur length*.

	<i>femur length</i>	<i>height</i>
mean	24.246	166.092
standard deviation	1.873	10.086
correlation coefficient	$r = 0.9939$	

Write the equation in terms of *height* and *femur length*. Give the slope and intercept rounded to three significant figures.

- c** Interpret the slope of the regression equation in terms of *height* and *femur length*.  
**d** Determine the value of the coefficient of determination and interpret in terms of *height* and *femur length*.

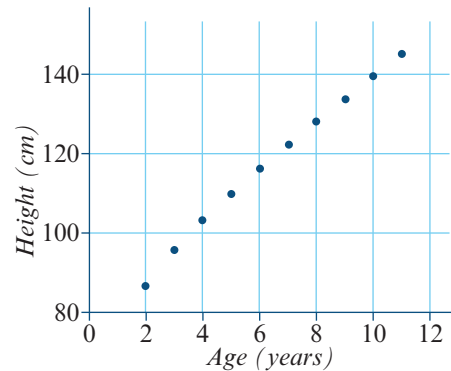
Arm span is also associated with height. A least squares regression line that can be used to model this association is:

$$\text{height} = 0.498 + 0.926 \times \text{arm span}$$

In determining this equation, the summary statistics displayed in the table were also calculated.

	<i>arm span</i>	<i>height</i>
mean	169.615	166.092
standard deviation	10.761	10.086

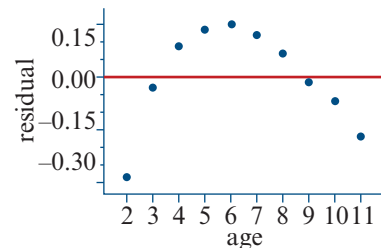
- e Determine the percentage of the variation in *height* explained by the variation in *arm span*. Write the answer as a percentage rounded to one decimal place.
- 5 The scatter plot shows the height (in cm) of a group of 10 children plotted against their age (in years). The data used to generate this scatterplot is shown below.



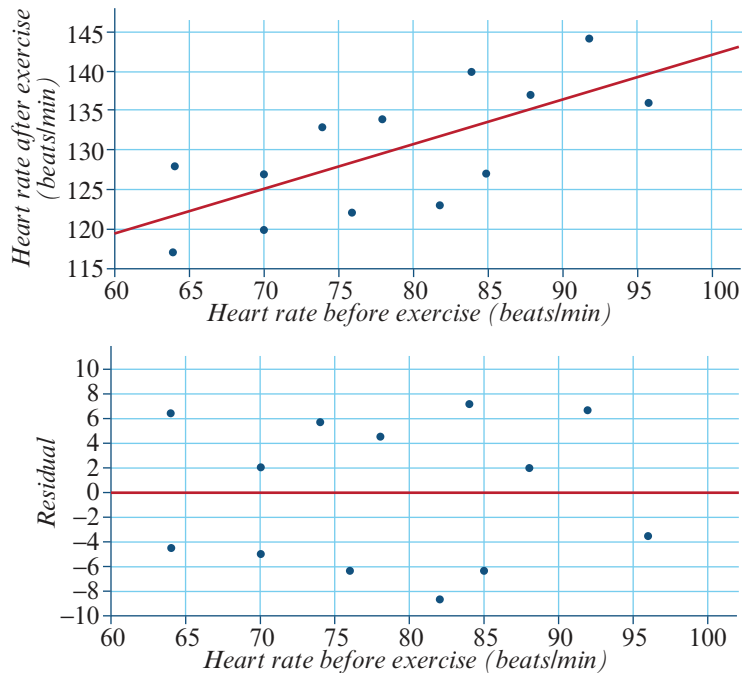
<i>Height (cm)</i>	86.5	95.5	103.0	109.8	116.4	122.4	128.2	133.8	139.6	145.0
<i>Age (years)</i>	2	3	4	5	6	7	8	9	10	11

The task is to determine the equation of a least squares regression line that can be used to predict height from age.

- a In this analysis, which would be the RV and which would be the EV?
- b Use the scatter plot to describe the association between *age* and *height* in terms of strength and direction.
- c Use your calculator to confirm that the equation of the least squares regression line is:  $height = 76.64 + 6.366 \times age$  and  $r = 0.9973$ .
- d i Use the regression line to show that the predicted height of a one-year old is 83.0 cm, rounded to 3 significant figures.  
ii In making this prediction are you extrapolating or interpolating?
- e Interpret the slope of the least squares line in terms of *height* and *age*.
- f Determine the percentage of the variation in *height* of these children explained by their *age*. Round your answer to 1 decimal place.
- g Use the least squares regression equation to:  
i predict the *height* of the 10-year-old child in this sample  
ii determine the residual value for this child.
- h i Confirm that the residual plot for this analysis is shown opposite.  
ii Explain why this residual plot suggests that a linear equation is not the most appropriate model for this association.



- 6 The heart rate (in beats/minute) was measured and recorded for a group of 13 students. The students then completed the same set of exercises and their heart rate measured again immediately on completion. The scatterplot below shows the students' *heart rate after exercise* plotted against their *heart rate before exercise*, with a least squares regression line fitted. Also shown is the residual plot for this line.



- a Describe the association between *heart rate before exercise* and *heart rate after exercise* in terms of strength, direction and form.

The equation of the least squares line is:

$$\text{heart rate after exercise} = 85.671 + 0.561 \times \text{heart rate before exercise}$$

- b i Use the equation to predict heart rate after exercise when heart rate before exercise is 100 beats/minute. Round to the nearest whole number.  
 ii Are you extrapolating or interpolating?
- c The person with a heart rate of 122 beats/minute after exercise had a heartbeat of 76 beats/minute before exercise. If the least squares line is used to predict this person's heart rate after exercise, determine the value of the residual. Give your answer rounded to one decimal place.
- d i What assumption about the form of the association can be tested using a residual plot?  
 ii Referring to the residual plot, explain why this assumption is satisfied.