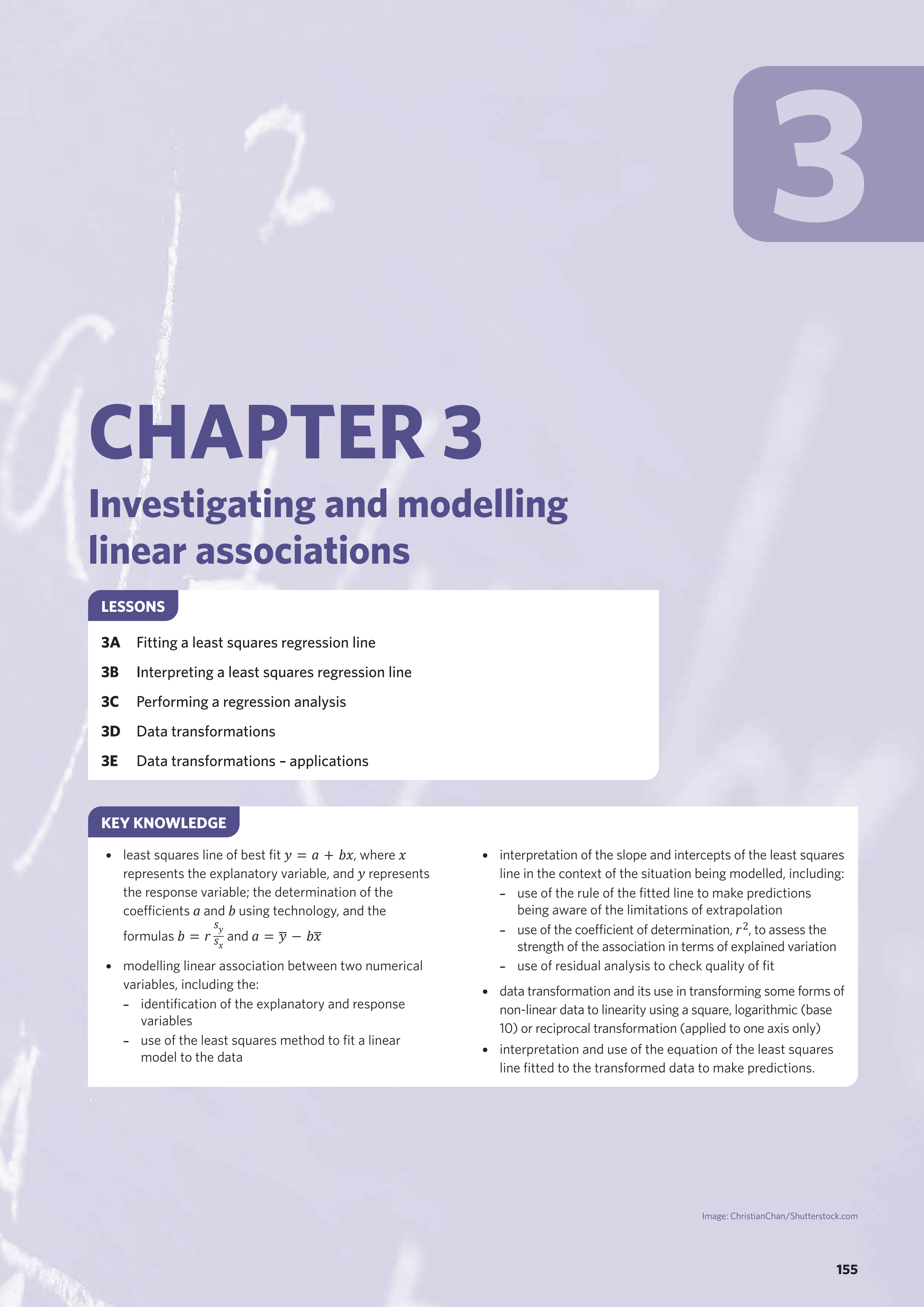


CHAPTER 3

Investigating and modelling linear associations

LESSONS

KEY KNOWLEDGE

- least squares line of best fit $y = a + bx$, where x represents the explanatory variable, and y represents the response variable; the determination of the coefficients a and b using technology, and the formulas $b = r\frac{s_y}{s_x}$ and $a = \overline{y} - b\overline{x}$
- modelling linear association between two numerical variables, including the:
  - identification of the explanatory and response variables
  - use of the least squares method to fit a linear model to the data
- interpretation of the slope and intercepts of the least squares line in the context of the situation being modelled, including:
  - use of the rule of the fitted line to make predictions being aware of the limitations of extrapolation
  - use of the coefficient of determination, r^2, to assess the strength of the association in terms of explained variation
  - use of residual analysis to check quality of fit
- data transformation and its use in transforming some forms of non-linear data to linearity using a square, logarithmic (base 10) or reciprocal transformation (applied to one axis only)
- interpretation and use of the equation of the least squares line fitted to the transformed data to make predictions.

Image: ChristianChan/Shutterstock.com

3A Fitting a least squares regression line

STUDY DESIGN DOT POINTS

- least squares line of best fit $y = a + bx$, where x represents the explanatory variable and y represents the response variable; the determination of the coefficients a and b using technology, and the formulas $b = r\frac{s_y}{s_x}$ and $a = \overline{y} - b\overline{x}$
- modelling linear association between two numerical variables, including the:
  - identification of the explanatory and response variables
  - use of the least squares method to fit a linear model to the data

3A 3B 3C 3D 3E

KEY SKILLS

During this lesson, you will be:

- using technology to determine the least squares regression equation
- determining the least squares regression equation from a graph
- calculating the least squares regression equation from summary statistics
- sketching a least squares regression line from its equation.

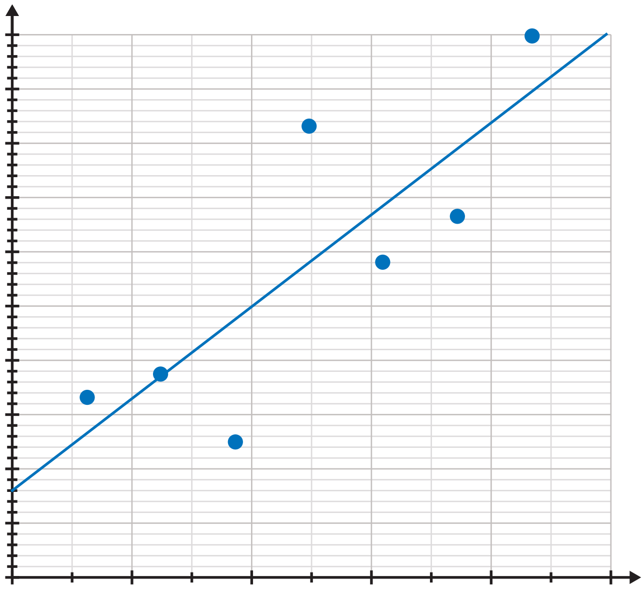
KEY TERMS

- Residual
- Least squares regression line

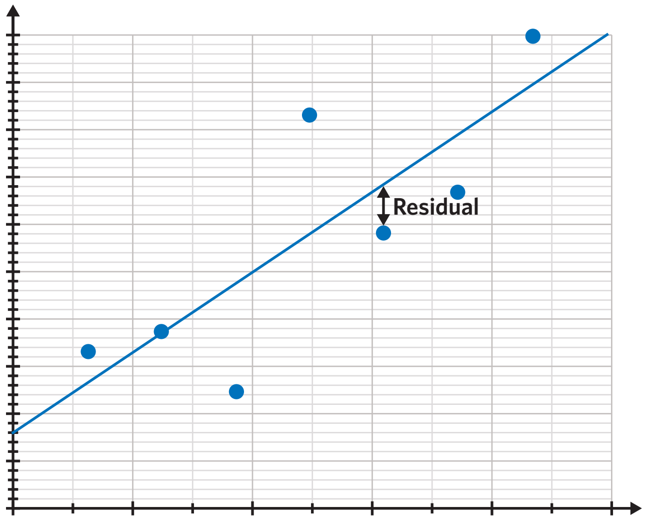
When bivariate numerical data is represented using scatterplots, there is often an underlying linear trend that can be observed. In these cases, it is useful to show this trend by fitting a straight line onto the scatterplot. While the least squares line will rarely be a perfect fit for the data, it can be useful for making inferences and predictions later on.

Using technology to determine the least squares regression equation

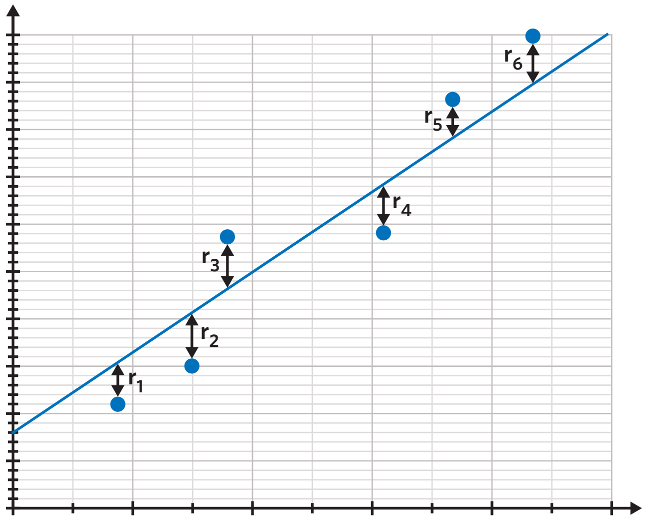
To fit a line to a scatterplot representing a set of data, a visual approach is often taken. Through this process, the aim is to construct the line to be as close as possible to all of the data points collectively, and to represent the direction of the scatterplot.



By hand, a reasonable line can usually be achieved. However, there is a mathematical process by which this line can be optimised. This process involves the minimisation of residual values, or residuals. A **residual** is the vertical distance between the straight line and any given point of the scatterplot, as shown.



The **least squares regression line** is the line which creates the minimum sum of the squares of the residuals. It is highly accurate because it can be determined mathematically rather than by eye. Visually, this will represent the underlying trend of the corresponding scatterplot.



The equation of the least squares regression line is generally in the form $y = a + bx$, where b represents the slope of the line, and a represents the y-intercept. This line is used to show the general trend of numerical bivariate data. When using a least squares regression line, some assumptions about the data are made:

- The data is numerical
- The relationship between the variables is linear
- There are no clear outliers present

A calculator can be used to determine the least squares regression equation.

Worked example 1

The following table gives the heights, in cm, of fathers and their sons when they were the same age.

height of father **(cm)**	182	187	174	177	162	171	188	165	170	185
height of son **(cm)**	184	179	176	168	161	173	187	163	170	181

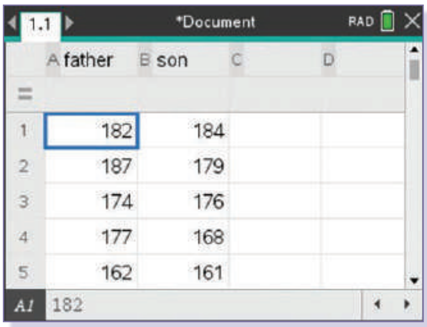
Determine the equation of the least squares regression line that will allow the *height of son* to be predicted from the *height of father*. Give values correct to three significant figures.

Continues →

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select ‘1: New’ → ‘4: Add Lists & Spreadsheet’.

Step 2: Name a list ‘father’ and another list ‘son’ and enter the data as shown.



	A father	B son	C	D
=				
1	182	184		
2	187	179		
3	174	176		
4	177	168		
5	162	161		

A1 182

Step 3: Identify the explanatory and response variables.

As *height of father* is being used to predict *height of son*, *height of father* is the explanatory variable.

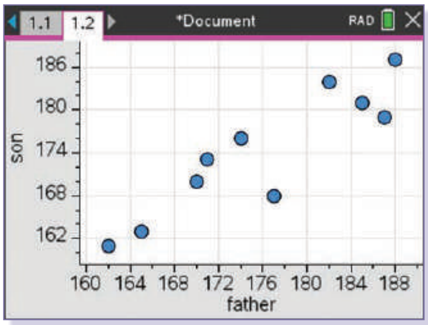
EV: *height of father*

RV: *height of son*

Step 4: Press ctrl + doc▸ and select ‘5: Add Data & Statistics’.

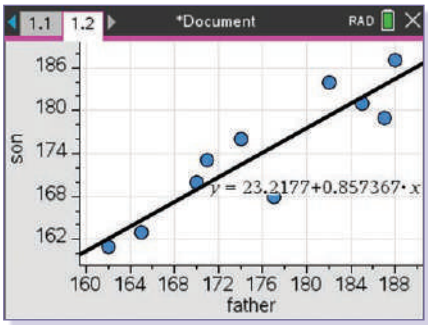
Step 5: Add the variables on each axis using the ‘Click to add variable’ function.

The *RV* will be positioned on the vertical axis and the *EV* will be positioned on the horizontal axis.



Step 6: Press menu. Select → ‘4: Analyse’ → ‘6: Regression’ →’2: Show Linear (a+bx)’ to plot the least squares regression line.

The least squares regression line and its equation in the form $y = a + bx$ will appear.

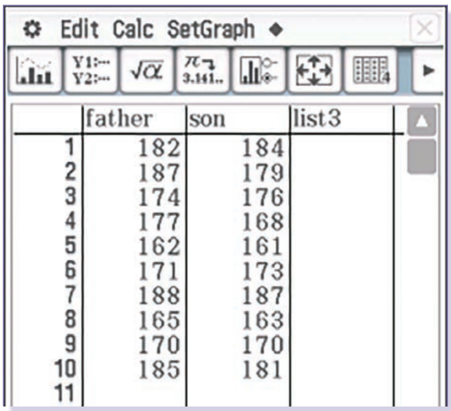


Step 7: Rewrite the equation in terms of the variables in the question and round as specified.

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Step 2: Name a list ‘father’ and another list ‘son’ and enter the data as shown.



	father	son	list3
1	182	184	
2	187	179	
3	174	176	
4	177	168	
5	162	161	
6	171	173	
7	188	187	
8	165	163	
9	170	170	
10	185	181	
11			

Step 3: Identify the explanatory and response variables.

As *height of father* is being used to predict *height of son*, *height of father* is the explanatory variable.

EV: *height of father*

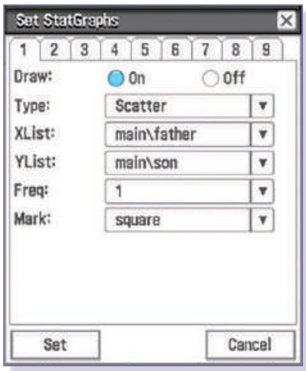
RV: *height of son*

Continues →

Step 4: Configure the settings of the graph by tapping in the icon bar.

Create a scatterplot by selecting 'Type' as 'Scatter'.

Specify the data set by changing 'XList:' to 'main\father' and 'YList:' to 'main\son'.



Set StatGraphs

Draw: ● On ○ Off

Type: Scatter

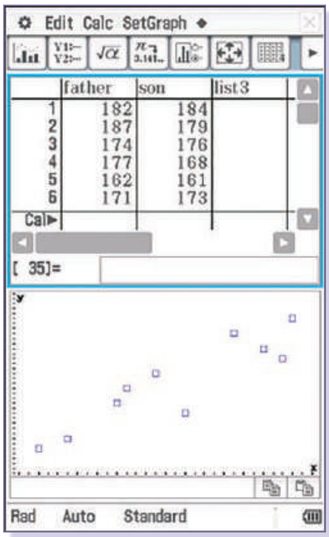
XList: main\father

YList: main\son

Freq: 1

Mark: square

Step 5: Tap 'Set' to confirm and then to plot the scatterplot.



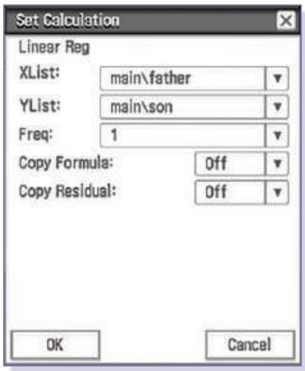
	father	son	list3
1	182	184	
2	187	179	
3	174	176	
4	177	168	
5	162	161	
6	171	173	

Cal▶

[35]=

Step 6: Fit a least squares regression line to the scatterplot by tapping 'Calc' → 'Regression' → 'Linear Reg'. Specify the data set by changing 'XList:' to 'main\father' and 'YList:' to 'main\son'.

Tap 'OK' to confirm.



Set Calculation

Linear Reg

XList: main\father

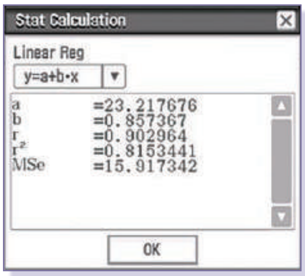
YList: main\son

Freq: 1

Copy Formula: Off

Copy Residual: Off

Step 7: The Stat Calculation window shows the values of a and b which will be used to write the least squares regression line.



Stat Calculation

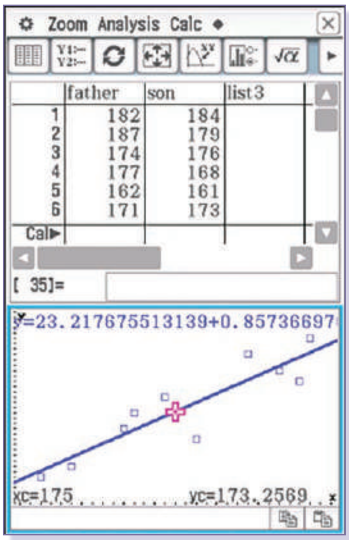
Linear Reg

y=a+b·x

```
a     =23.217676
b     =0.857367
r     =0.902964
r²    =0.8153441
MSe   =15.917342
```

Note: If $y = a \bullet x + b$ is selected, the a and b values need to be switched when writing the equation in $y = a + bx$ form.

Step 8: To visualise the data, tap 'OK', and the regression line will be generated on the scatterplot.



	father	son	list3
1	182	184	
2	187	179	
3	174	176	
4	177	168	
5	162	161	
6	171	173	

Cal▶

[35]=

y=23.21767551339+0.85736697

xc=175 yc=173.2569

Step 9: Rewrite the equation in terms of the variables in the question and round as specified.

Answer – Method 1 and 2

height of son $= 23.2 + 0.857 \times$ *height of father*

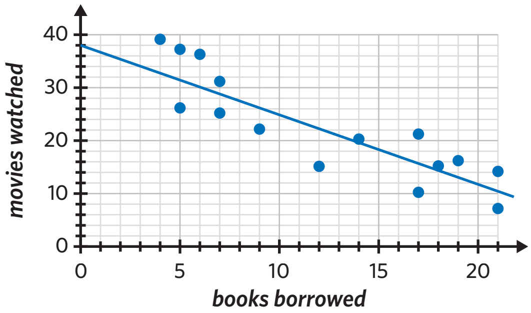
Determining the least squares regression equation from a graph

When the least squares regression line is already sketched on a scatterplot, its equation $y = a + bx$ can be found by determining the y-intercept (a) and the slope of the line (b).

Worked example 2

Consider the scatterplot shown.

From the scatterplot, estimate the equation of the least squares regression line. Give values correct to 3 significant figures.



Explanation

Step 1: Determine the explanatory and response variable.

On a scatterplot, the explanatory variable is positioned on the horizontal axis and the response variable on the vertical axis.

EV: *books borrowed*

RV: *movies watched*

Step 2: Write the equation in terms of the variables in the question.

$\textit{movies watched} = a + b \times \textit{books borrowed}$

Step 3: Determine the vertical axis intercept of the line, a.

The point at which the line intersects the vertical axis, *movies watched*, is (0, 38).

$a = 38$

Step 4: Determine the slope of the line, b.

The slope of the line can be determined using two points that the line clearly passes through.

Two points that can be used are (0, 38) and (21, 10).

The formula for the gradient of a straight line between two points is $b = \frac{rise}{run} = \frac{y_2 - y_1}{x_2 - x_1}$.

$b = \frac{10 - 38}{21 - 0} = \frac{-28}{21}$

$= -1.333...$

Step 5: Write the equation using the values for a and b, rounded to 3 significant figures.

Answer

$\textit{movies watched} = 38.0 - 1.33 \times \textit{books borrowed}$

Calculating the least squares regression equation from summary statistics

For a least squares line of the form $y = a + bx$, a and b can also be found using the formulas

- $b = r \times \frac{s_y}{s_x}$
- $a = \bar{y} - b\bar{x}$

where:

- a is the y-intercept
- b is the gradient
- r is Pearson's correlation coefficient
- $\bar{x}$ is the mean of the explanatory variable (x)
- $\bar{y}$ is the mean of the response variable (y)
- s_x is the standard deviation of the explanatory variable (x)
- s_y is the standard deviation of the response variable (y)

Worked example 3

Consider the following summary statistics for a set of bivariate data involving the variables x and y.

$r = 0.845, \quad \bar{x} = 11.0, \quad \bar{y} = 29.2, \quad s_x = 6.06, \quad s_y = 16.8$

Find the equation of the least squares regression line that allows y to be predicted from x. Give all values correct to three significant figures.

Explanation

Step 1: Calculate the slope, b in the equation $y = a + bx$, using the formula $b = r \times \frac{s_y}{s_x}$

$b = 0.845 \times \frac{16.8}{6.06}$

$= 2.342...$

Step 2: Calculate the y-intercept, a using the formula

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

$a = 29.2 - 2.342... \times 11.0$

$= 3.431...$

Step 3: Write the equation of the least squares regression line, correct to three significant figures.

Answer

$y = 3.43 + 2.34x$

Sketching a least squares regression line from its equation

When the equation of a least squares regression line is given, it can be sketched onto a scatterplot using the two-point method.

Worked example 4

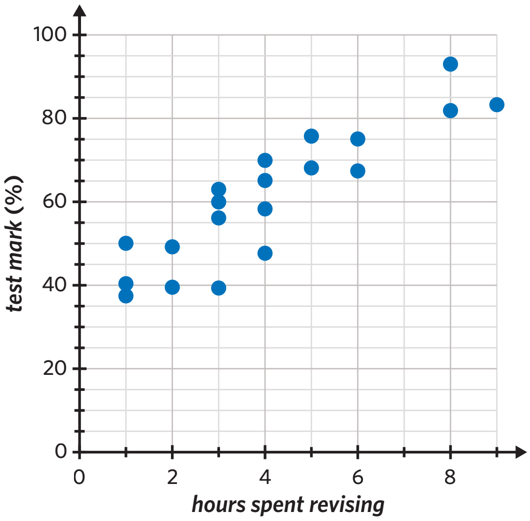
20 students in Ms Benton's geography class took their end of year test. Before the test, each member of the class was surveyed to find out how many hours they spent revising for the test. Each student's *hours spent revising* and *test mark* (%) are depicted in the following scatterplot.

Ms Benton noticed a strong linear association between the two variables.

The least squares regression line for this data is approximated by the following equation:

test mark = 36.0 + 6.11 × *hours spent revising*

From the equation, sketch the least squares line for the scatterplot.



Explanation

Step 1: Decide on two horizontal axis values to substitute into the equation.

It is best to use two values for *hours spent revising* that are on either end of the visible plane for accuracy.

The best values to use are 0 and 9.

Step 2: Substitute both values for *hours spent revising* into the equation to find their corresponding *test mark* values.

hours spent revising $= 0$

test mark $= 36.0 + 6.1 \times 0 = 36$

hours spent revising $= 9$

test mark $= 36.0 + 6.1 \times 9 = 90.9$

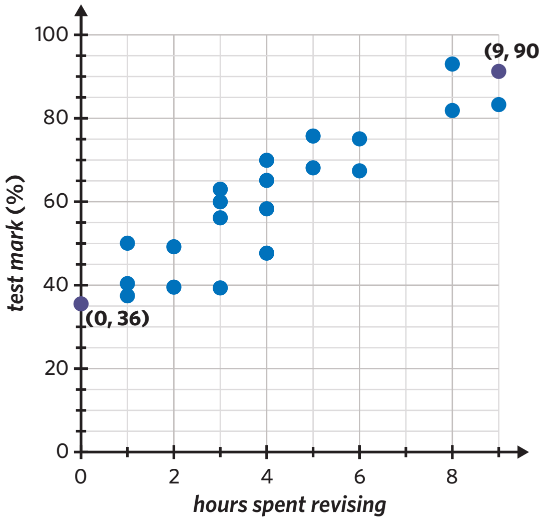
Continues →

Step 3: Determine the points to plot on the graph.

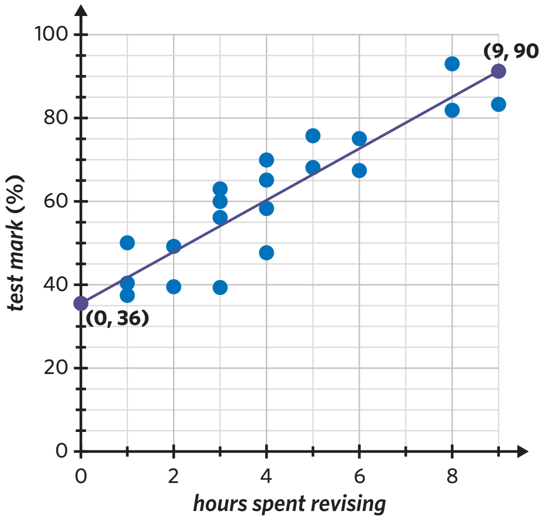
Point 1: (0, 36)

Point 2: (9, 90.9)

Step 4: Draw a straight line passing through the two points.



Answer



Exam question breakdown

VCAA 2017 Exam 1 Data analysis Q8

The following scatterplot shows the *wrist* circumference and *ankle* circumference, both in centimetres, of 13 people. A least squares line has been fitted to the scatterplot with *ankle* circumference as the explanatory variable.

The equation of the least squares line is closest to

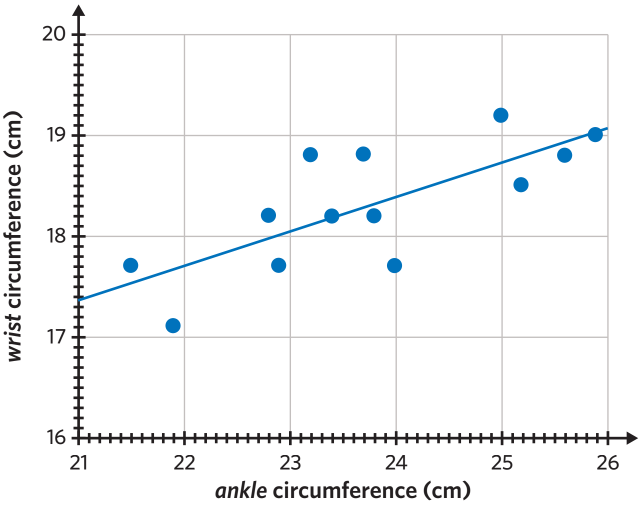
A. *ankle* = 10.2 + 0.342 × *wrist*

B. *wrist* = 10.2 + 0.342 × *ankle*

C. *ankle* = 17.4 + 0.342 × *wrist*

D. *wrist* = 17.4 + 0.342 × *ankle*

E. *wrist* = 17.4 + 0.731 × *ankle*



Continues →

Explanation

Step 1: Identify the explanatory and response variable.

The question specifies that *ankle* circumference is the explanatory variable, which means *wrist* circumference is the response variable.

EV: *ankle*

RV: *wrist*

Step 2: Write the equation in terms of the variables in the question.

$wrist = a + b \times ankle$

Step 3: Determine the vertical axis intercept, a, by inspecting the graph.

At first, the intercept value appears to be 17.4. However, the horizontal axis begins at 21, not 0. Therefore, the actual intercept occurs further to the left, and will be a lower value.

Observing the available options, since a is less than 17.4, a must equal 10.2.

Step 4: Determine the slope, b, by inspecting the graph.

As there are only two possible b values, an approximation of the slope will be sufficient.

The least squares regression line shows that *wrist* circumference increases by approximately 1 cm for every 3 cm increase in *ankle* circumference. $\frac{rise}{run} \approx \frac{1}{3}$

$\frac{1}{3}$ is closest to 0.342 so $b = 0.342$

Step 5: Write the final equation.

$wrist = 10.2 + 0.342 \times ankle$

Answer

B

43% of students answered this question correctly.

39% of students incorrectly answered D, due to the assumption that the horizontal axis begins at 0. This led to an incorrectly obtained vertical axis intercept of 17.4.

3A Questions

Using technology to determine the least squares regression equation

1. Consider the following table.

number of beaches	1	3	4	4	4	5	5	7	7	8
number of surf shops	2	3	4	18	5	6	6	7	9	10

A least squares regression line would not be appropriate for this data because

A. the data is categorical.

B. there are not enough data points.

C. both variables are numerical.

D. there is a clear outlier.

2. Consider the following table.

number of shops	8	12	15	19	20	21	25	27	30	34
population **(000's)**	4.7	6.9	8.4	7.8	10.8	8.1	13.0	11.6	19.0	16.2

A least squares line is fitted to the data, with the aim of predicting *population* from *number of shops*. Which of the following statements is true?

A. $a = 0.91$ and $b = 0.82$

B. $a = 0.49$ and $b = 0.24$

C. $a = 0.24$ and $b = 0.49$

D. $a = 0.82$ and $b = 0.91$

3. A group of students were asked how many hours they had slept the night before a General Mathematics exam.

Their *number of hours slept* and *result* (%) are shown in the table.

number of hours slept	7.2	7.8	8.3	5.3	6	9	2.3	6.8	7.7	8	8.6	9.2
result **(%)**	88	79	83	67	70	92	56	74	63	84	98	85

A least squares regression line has been fitted to the data to predict *result*, from *number of hours slept*.

Write the equation of the regression line, giving values correct to three decimal places.

Determining the least squares regression equation from a graph

4. The least squares line fitted to the scatterplot shown is of the form $y = a + bx$.

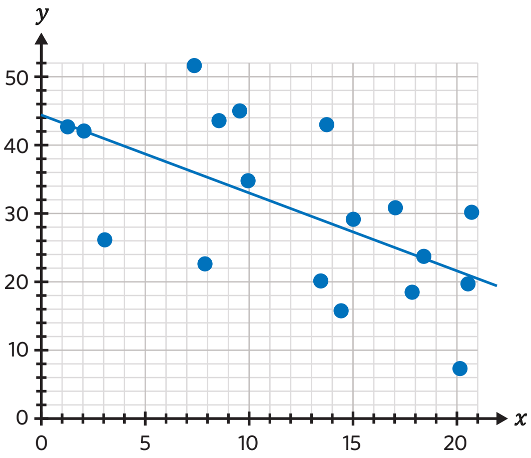
Which of the following statements is false?

A. The value of a is positive.

B. The value of b is positive.

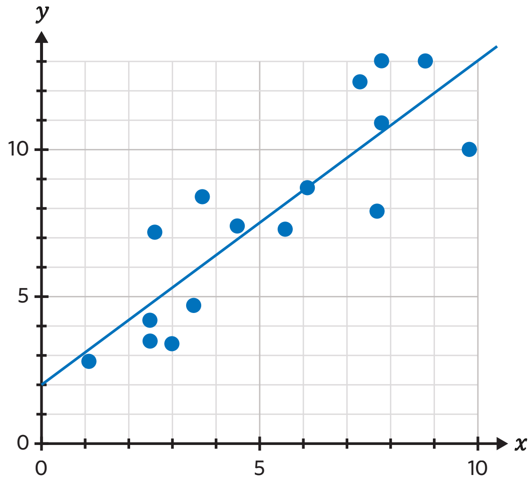
C. x is the explanatory variable.

D. y is the response variable.

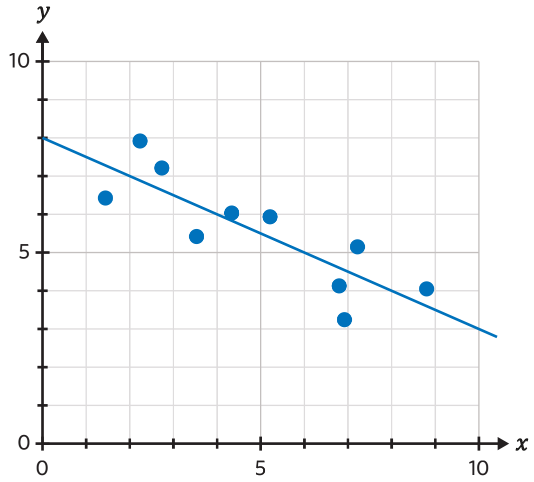


5. Estimate the equation of the least squares regression line for the following scatterplots.

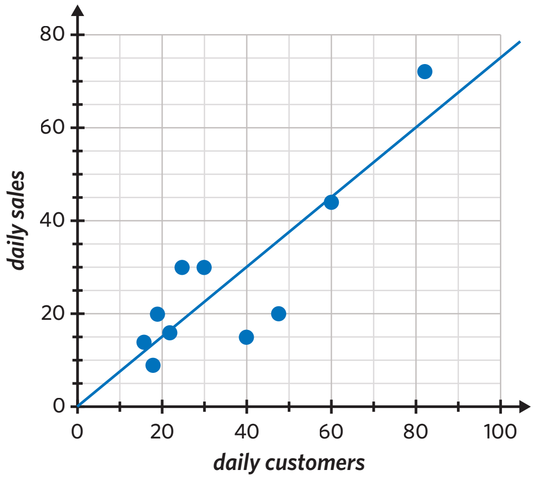
a.



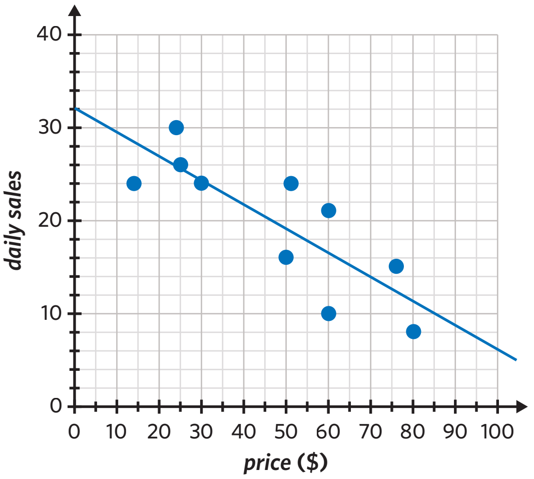
b.



c.



d.



Calculating the least squares regression equation from summary statistics

6. Consider the following summary statistics for a set of bivariate data involving the variables x and y.

$r = -0.96, \quad \bar{x} = 15.3, \quad \bar{y} = 43.8, \quad s_x = 12.2, \quad s_y = 13.0$

The equation of the least squares regression line that allows y to be predicted from x, with values given correct to 3 significant figures, is closest to

A. $y = 59.5 + 1.02x$

B. $y = 1.02 + 59.5x$

C. $y = 59.5 - 1.02x$

D. $y = 1.02 - 59.5x$

7. Given the following information:

$a = 31.4, \quad b = 1.7, \quad r = 0.96, \quad s_x = 9.2, \quad \bar{x} = 15.1$

Calculate s_y and $\bar{y}$ correct to one decimal place.

8. Joe the farmer wants to investigate the relationship between *time* (days) and the *height* (cm) of his growing wheat crops. Joe hires a data analyst, who is able to provide certain statistics based on the data gathered from a patch of growing wheat crops.

$\bar{x} = 55.0, \quad \bar{y} = 63.8, \quad s_x = 30.3, \quad s_y = 37.4, \quad r = 0.99$

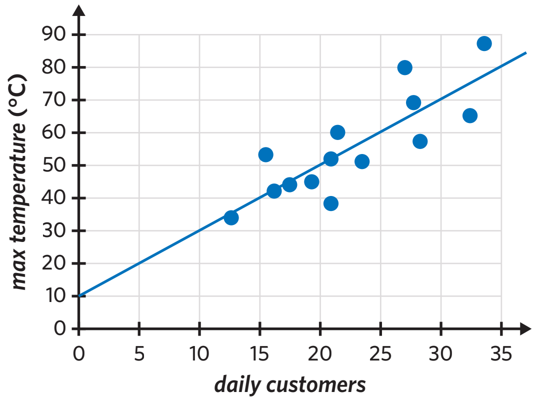
Joe wants to model a least squares regression line to his data, so that he can predict the *height* of any given wheat plant based on the amount of *time* that has passed. Determine the equation of this line, giving values correct to one decimal place.

Sketching a least squares regression line from its equation

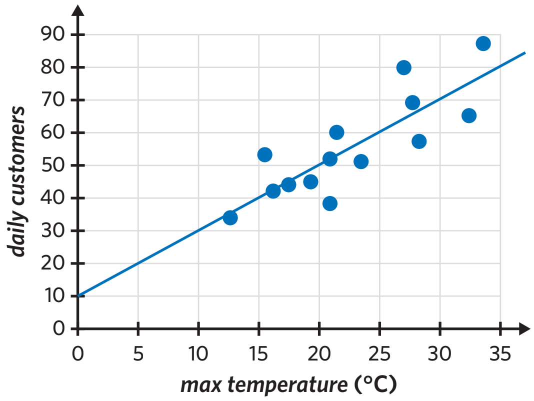
9. A least squares line is fitted to data with the aim of predicting the number of *daily customers* of an ice-cream shop from the *max temperature* (°C). The equation of the line is *daily customers* = 10 + 2 × *max temperature*.

A possible scatterplot that this equation could be obtained from is

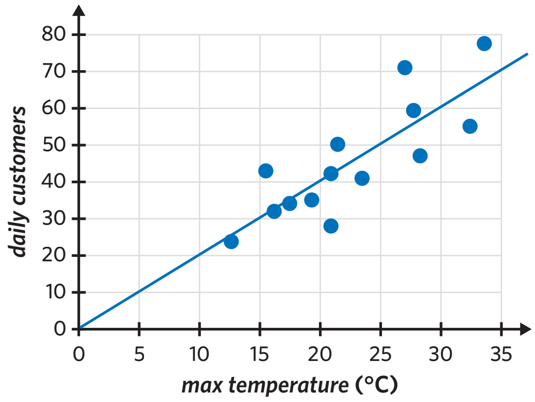
A.



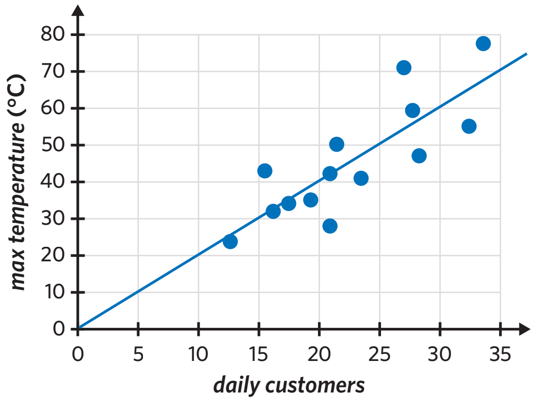
B.



C.



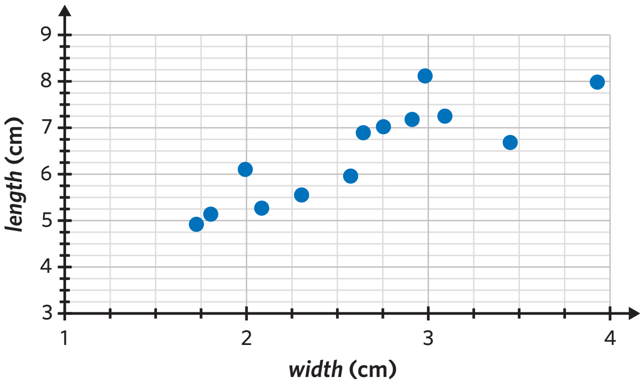
D.



10. The *width* (cm) and *length* (cm) of 13 insects were plotted on a scatterplot.

The least squares regression line for this data is given by the equation *length* = 2.85 + 1.37 × *width*

From this equation, plot the line on the scatterplot.



Joining it all together

11. Which of the following statements regarding the least squares regression line is false?

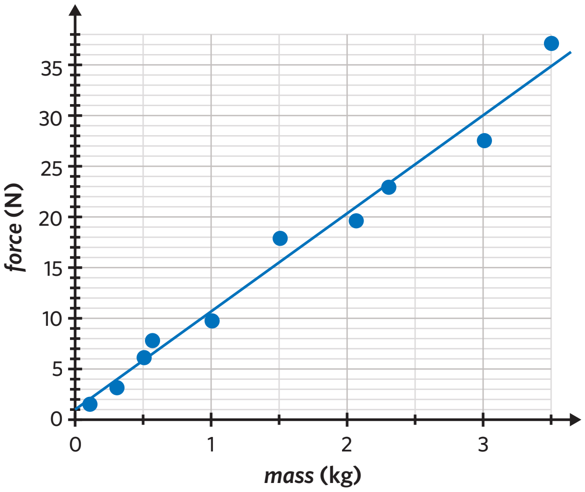
A. A least squares regression line minimises the sum of the squares of the residuals.

B. A least squares regression line is not appropriate if there is a clear outlier.

C. A least squares regression line is generally written in the form $y = a + bx$.

D. A least squares regression line can be fitted to the variables *day of the week* and *hours spent watching TV*.

12. During a high school physics experiment, the *mass* (kg) and gravitational *force* (N) acting on 10 different objects were recorded and displayed using a scatterplot.



a. Estimate the equation of the least squares regression line from the graph, giving values correct to 1 decimal place.

b. Use the corresponding data provided to determine the least squares equation, giving values accurate to 2 decimal places.

mass (kg)	0.1	0.3	0.5	0.56	1.0	1.5	2.06	2.3	3.0	3.5
force (N)	1.5	3.1	6.1	7.8	9.7	17.9	19.6	22.9	27.5	37.1

c. Using the values for a and b from the equation obtained in part **b**, and knowing that $\bar{x} = 1.482$, show that $\bar{y} = 15.32$.

Exam practice

13. The statistical analysis of a set of bivariate data involving variables x and y resulted in the information displayed in the table shown.

mean	$\bar{x} = 27.8$	$\bar{y} = 33.4$
standard deviation	$s_x = 2.33$	$s_y = 3.24$
equation of the least squares line	$y = -2.84 + 1.31x$	

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

A. 0.88 **B.** 0.89 **C.** 0.92
D. 0.94 **E.** 0.97

VCAA 2018 Exam 1 Data analysis Q13

49% of students answered this question correctly.

14. The following table shows the *weight* in kilograms, and the *height*, in centimetres, of 10 adults.

***weight* (kg)**	59	67	69	84	64	74	76	56	58	66
***height* (cm)**	173	180	184	195	173	180	192	169	164	180

A least squares line is fitted to the data.

The least squares line enables an adult's *weight* to be predicted from their *height*.

The number of times that the predicted value of an adult's *weight* is greater than the actual value of their *weight* is

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

VCAA 2021 Exam 1 Data analysis Q11

40% of students answered this question correctly.

15. The *number of male moths* caught in a trap set in a forest and the *egg density* (eggs per square metre) in the forest are shown in the following table.

number of male moths	35	37	45	49	65	74	77	86	95
***egg density* (eggs per square metre)**	471	635	664	997	1350	1100	2010	1640	1350

a. Determine the equation of the least squares line that can be used to predict the *egg density* in the forest from the *number of male moths* caught in the trap.

Write the values of the intercept and slope of this least squares line in the appropriate boxes provided.

Round your answers to one decimal place.

egg density = ☐ + ☐ × *number of male moths* (2 MARKS)

b. The *number of female moths* caught in a trap set in a forest and the *egg density* (eggs per square metre) in the forest can also be examined.

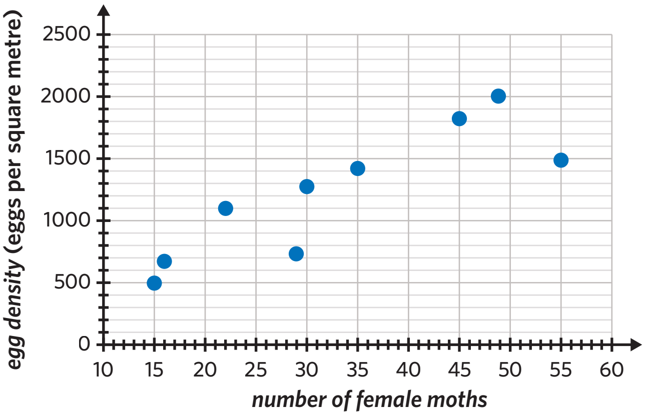
A scatterplot of the data is shown.

The equation of the least squares line is
egg density = 191 + 31.3 × *number of female moths*

Draw the graph of this least squares line on the scatterplot. (1 MARK)

VCAA 2017 Exam 2 Data analysis Q3a, bi

Part **a**: The average mark on this question was **1.5**.
Part **b**: **26%** of students answered this question correctly.



Questions from multiple lessons

Data analysis *Year 11 content*

16. The marks obtained by 59 students in a maths test are shown in the following histogram.

The median mark for the test was

A. greater than or equal to 15 but less than 20.

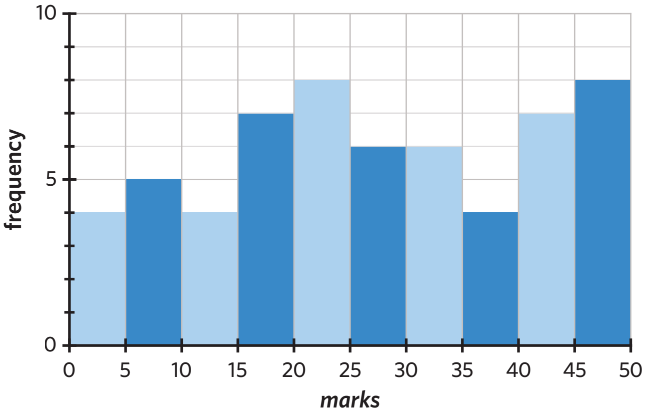
B. greater than or equal to 20 but less than 25.

C. greater than or equal to 25 but less than 30.

D. greater than or equal to 30 but less than 35.

E. greater than or equal to 35 but less than 40.

Adapted from VCAA 2017NH Exam 1 Data analysis Q3



Recursion and financial modelling *Year 11 content*

17. Victoria wants to buy a jumbo trampoline from her friend, Shaun. She agrees to pay Shaun back over several months. In the first month, she will pay Shaun \$300, and in each subsequent month she will pay \$50 less than she did the month before.

Let p_n be the amount of money that Victoria pays Shaun during the n^{th} month.

A recurrence relation that can be used to model this situation for $1 \leq n \leq 6$ is

A. $p_1 = 300, \quad p_{n+1} = 0.83p_n$

B. $p_1 = 250, \quad p_{n+1} = p_n - 50$

C. $p_0 = 300, \quad p_{n+1} = p_n - 50$

D. $p_0 = 250, \quad p_{n+1} = 300 - 50p_n$

E. $p_1 = 300, \quad p_{n+1} = p_n - 50$

Adapted from VCAA 2014 Exam 1 Number patterns Q4

Data analysis

18. The weight of ducks in a large population is approximately normally distributed with a mean of 1.2 kg and a standard deviation of 120 g.

Which one of the following statements relating to this population of ducks is **not** true?

A. Approximately half of the ducks will weigh greater than 1.2 kg.

B. Approximately 16% of the ducks will weigh less than 1.08 kg.

C. Approximately 81.5% of the ducks will weigh between 1.08 kg and 1.44 kg.

D. Approximately 13.5% of the ducks will weigh between 1.32 kg and 1.44 kg.

E. No duck will weigh less than 1 kg.

Adapted from VCAA 2017NH Exam 1 Data analysis Q8

3B Interpreting a least squares regression line

STUDY DESIGN DOT POINT

- interpretation of the slope and intercepts of the least squares line in the context of the situation being modelled, including:
  - use of the rule of the fitted line to make predictions being aware of the limitations of extrapolation
  - use of the coefficient of determination, r^2, to assess the strength of the association in terms of explained variation
  - use of residual analysis to check quality of fit

3A 3B 3C 3D 3E

KEY SKILLS

During this lesson, you will be:

- interpreting a least squares regression line
- making predictions using a least squares regression line.

KEY TERMS

- y-intercept
- Slope
- Interpolation
- Extrapolation

A least squares regression line is often used to make predictions. It is therefore useful to understand and be able to interpret the least squares regression line and its equation. When making predictions, it is important to be aware of the reliability of these predictions.

Interpreting a least squares regression line

The **y-intercept** is the approximate value of the response variable (y) when the explanatory variable (x) is equal to 0. When the value of the explanatory variable cannot equal 0, this value has no useful meaning.

The **slope** is the average change in the response variable for every one-unit increase in the explanatory variable.

For a least squares regression line with the rule: $y = a + bx$,

- a is the y-intercept, and
- b is the slope

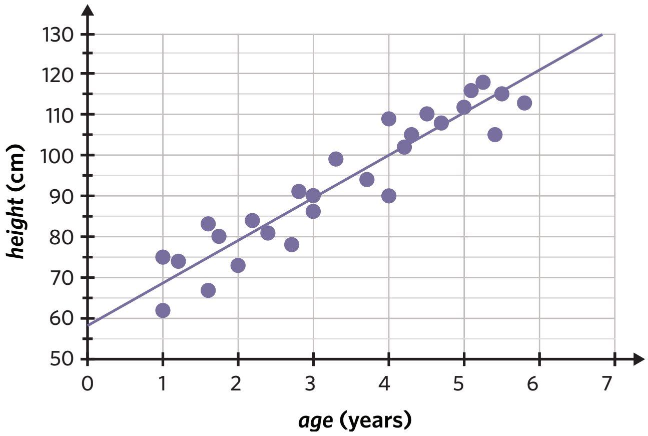
Worked example 1

The heights of a group of children aged between one and six years old were measured.

The scatterplot and regression equation show the relationship between *height* (cm) and *age* (years).

Based on the regression equation,

$height = 58.22 + 10.46 \times age$



Continues →

a. How much on average does a child grow each year, correct to two decimal places?

Explanation

Step 1: Identify which part of the equation requires interpretation.

The question is asking about the change in the response variable (*height*) for each one-unit (one year) increase in the explanatory variable (*age*). This is the slope.

Step 2: Determine the slope.

In the rule $y = a + bx$, the slope is represented by b.

$b = 10.46$

Step 3: Interpret the slope.

For every one year increase, there will be an increase in *height* of 10.46 cm.

Answer

On average, a child grows 10.46 cm each year.

b. Interpret the y-intercept in this regression equation.

Explanation

Step 1: Identify the y-intercept.

In the rule $y = a + bx$, a is the y-intercept.

$a = 58.22$

Step 2: Interpret the y-intercept.

The y-intercept is the approximate *height* when *age* equals 0.

Answer

On average, the height of children when they are born is 58.22 cm.

Making predictions using a least squares regression line

Least squares regression lines can be used to predict the value of the response variable from the explanatory variable or vice versa. **Interpolation** involves making predictions that are within the range of the data set, and are the most reliable predictions. **Extrapolation** involves making predictions that are outside the range of the data set. These predictions have limited reliability because an assumption is made that the relationship between the two variables continues outside the range of the data set. However, it is unknown whether this relationship will in fact continue.

Suppose the heights of children aged between 5 and 15 years old are measured and the height of a 7 year old is then predicted. This is interpolation. If the height of a 3 year old is predicted, this is extrapolation since it is outside the range of the data set.

Predictions can be made by eye using a graph of the regression line on a scatterplot.

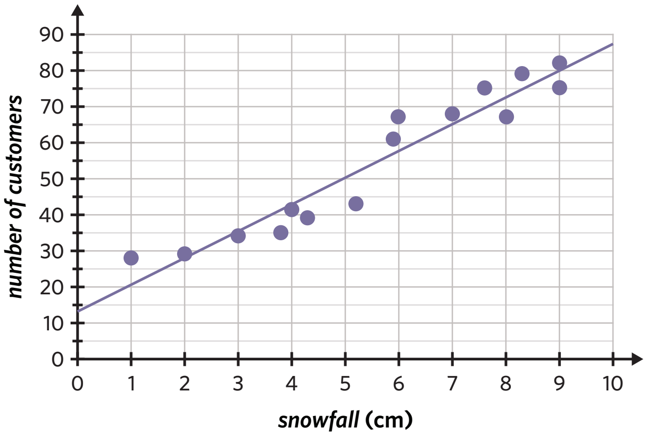
See worked example 2

Predictions can also be made by substituting the value of the explanatory variable or response variable into the regression equation and solving. These predictions are more accurate than ‘by eye’ predictions made from the regression line on a scatterplot.

See worked example 3

Worked example 2

A ski hire shop has recorded the daily *snowfall* (cm) and the corresponding *number of customers* that hire from them. The relationship between *number of customers* and *snowfall* is displayed on the scatterplot.



a. What is the predicted *number of customers* when there is 5 cm of *snowfall*? Is this interpolation or extrapolation?

Explanation

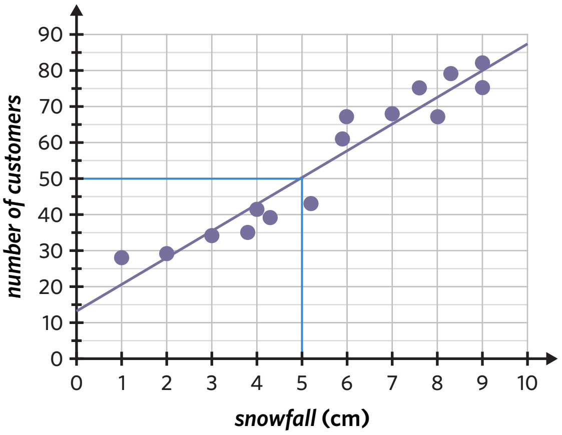
Step 1: Identify whether the prediction is interpolation or extrapolation.

The data set shows the *number of customers* for varying levels of *snowfall* ranging from 1 to 9 cm.

5 cm of *snowfall* is within this range. This is interpolation.

Step 2: Predict the response variable.

Use the regression line to predict the *number of customers* when *snowfall* = 5.



Answer

50 customers

Interpolation

b. How many customers are expected to hire from the ski hire shop when there is no snowfall?

Explanation

Determine the *number of customers* when *snowfall* = 0. This is the y-intercept.

Note that 0 cm of *snowfall* lies outside the range of the data set, so this prediction involves extrapolating the data.

Each interval on the vertical axis represents 5 customers. The y-intercept is just below the third interval, which is approximately 14.

Answer

14 customers

Worked example 3

The following table shows data collected on the *distance from the CBD* of Melbourne apartments and their *weekly rent.*

***distance from the CBD* (km)**	3.5	7	2.5	4	10	9.5	6.5	1.5	3	2.75	1	8.25	5.5	11.5	4.75
***weekly rent* ($)**	350	275	325	290	195	230	245	425	335	375	440	220	310	170	345

The least squares regression equation from the data is:

$\textit{weekly rent} = 427.92 - 23.25 \times \textit{distance from the CBD}$

a. If an apartment has a *weekly rent* of $250, predict its *distance from the CBD*, correct to two decimal places.

Explanation

Step 1: Substitute *weekly rent* = 250 into the regression equation.

$250 = 427.92 - 23.25 \times \textit{distance from the CBD}$

Step 2: Solve the equation.

In this question, the value of the response variable is given. The equation needs to be solved for the explanatory variable.

$250 = 427.92 - 23.25 \times \textit{distance from the CBD}$

$-177.92 = -23.25 \times \textit{distance from the CBD}$

$\textit{distance from the CBD} = 7.6524...$

Answer

7.65 km

b. Predict the *weekly rent* of an apartment that is 12 km from the CBD, correct to the nearest dollar.

Explanation

Step 1: Substitute *distance from the CBD* = 12 into the regression equation.

$\textit{weekly rent} = 427.92 - 23.25 \times 12$

Step 2: Solve the equation.

$$\begin{aligned}\textit{weekly rent} &= 427.92 - 279\\ &= 148.92\\ &\approx 149\end{aligned}$$

Answer

$149

c. Is the prediction from part **b** reliable? Justify your answer.

Explanation

Identify whether the prediction is an interpolation or extrapolation.

The range of the data set is apartments between 1 and 11.5 km from the CBD.

12 km falls outside the range of the data set, meaning it is an extrapolation.

Answer

No, it may not be reliable because the prediction is an example of extrapolation. The regression equation is not always reliable for values outside the data set.

Exam question breakdown

VCAA 2018 Exam 1 Data analysis Q10

In a study of the association between a person's *height*, in centimetres, and *body surface area*, in square metres, the following least squares line was obtained.

body surface area = −1.1 + 0.019 × *height*

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

A. An increase of 1 m^2 in *body surface area* is associated with an increase of 0.019 cm in *height*.

B. An increase of 1 cm in *height* is associated with an increase of 0.019 m^2 in *body surface area*.

C. The correlation coefficient is 0.019

D. A person's *body surface area*, in square metres, can be determined by adding 1.1 cm to their *height*.

E. A person's *height*, in centimetres, can be determined by subtracting 1.1 from their *body surface area*, in square metres.

Explanation

To solve this question, check whether each option is true or false.

A: This is false. 0.019 is the value of the slope which indicates the average change in the response variable for every one-unit increase in the explanatory variable. The response variable is *body surface area* and the explanatory variable is *height*. ✗

B: This is true. The slope indicates the average change in *body surface area* for every one-unit (1 cm) increase in *height*. ✓

C: This is false. There is not enough information to determine the correlation coefficient. ✗

D: This is false. A person's *body surface area* can be predicted by multiplying their *height* by 0.019 and subtracting 1.1. ✗

E: This is false. A person's *height* can be predicted by adding 1.1 to their *body surface area* and dividing by 0.019. ✗

Answer

B

51% of students answered this question correctly.

20% of students incorrectly answered option A, as they misinterpreted the slope as the average change in the explanatory variable (*height*) for a one-unit increase in the response variable (*body surface area*) instead of the other way around.

3B Questions

Interpreting a least squares regression line

1. The following regression equation shows the relationship between the *number of bedrooms* and *price* ($000's) of houses in Melbourne:

price = 593.2 + 98.4 × *number of bedrooms*

By how much, on average, do house prices increase as the *number of bedrooms* increases?

A. $98.40 **B.** $593.20 **C.** $98 400.00 **D.** $593 200.00

2. The *price* ($) of a second-hand car depends on its *age* (years).

An approximate relationship between the car's *price* and *age* is given in the following equation.

price = 24 000 − 3500 × *age*

a. What was the price of the car when it was brand new?

b. By how much does the car's value decrease each year?

3. A group of runners are training for the next Melbourne Marathon. During their training, they record the *time* and *distance* of each run. They then calculated the equation of the regression line that could predict the *time* (minutes) taken to complete a run from the *distance* (km).

time = 0.32 + 5.61 × *distance*

a. How long on average does it take a runner to complete an additional kilometre?

b. Interpret the value of the y-intercept. Does this make sense?

Making predictions using a least squares regression line

4. The relationship between the daily *maximum temperature* (°C) and the daily *minimum temperature* (°C) is given by the following least squares regression line:

maximum temperature = 10.42 + 0.89 × *minimum temperature*

What is the expected maximum temperature on Tuesday if the minimum temperature is 18 °C?

A. 16.02 °C **B.** 25.29 °C **C.** 26.44 °C **D.** 28.42 °C

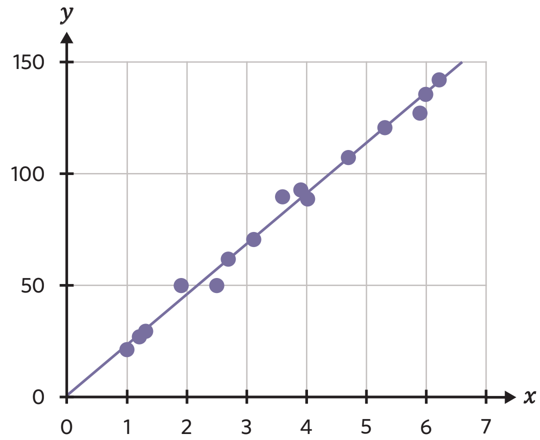
5. The scatterplot shown has been fitted with a regression equation:

$y = -1.7 + 23.2x$

Find the value of y when x equals

a. 2

b. 9

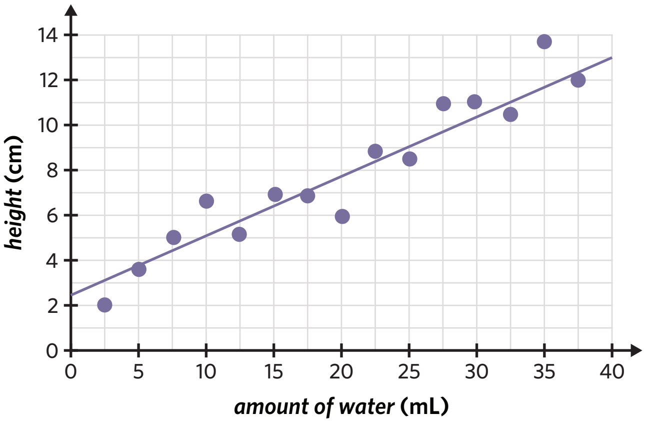


6. 15 seedlings were planted and given different amounts of water. The heights of the seedlings were then recorded after two weeks. The scatterplot shows the *amount of water* (mL) each seedling received when planted and the *height* (cm) after two weeks. A regression line has been fitted to the data.

a. What is the expected amount of water that a seedling received initially if it is 5 cm tall after two weeks?

b. How tall is a seedling predicted to be after 2 weeks if it receives 40 mL of water initially?

c. Why might the prediction in part **b** not be reliable?



7. The manager of a swimming pool investigated the effect that the *temperature* has on *attendance* at his pool. He recorded the weather and the number of people that attended the pool for two weeks.

The results are displayed in the scatterplot.

The regression equation that allows *attendance* to be predicted from *temperature* (°C) is as follows.

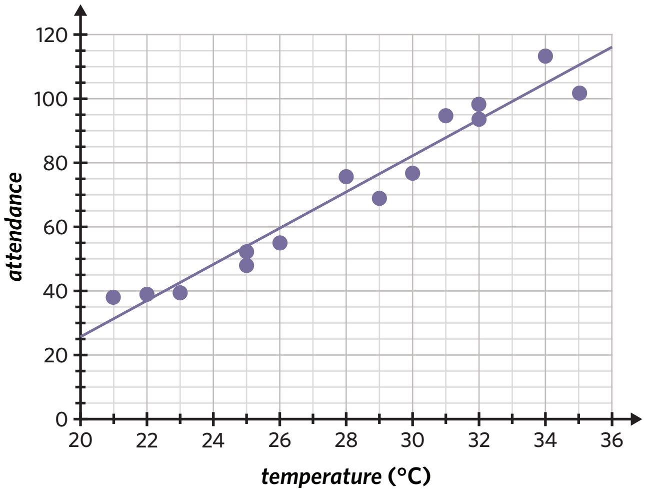
attendance = −88.53 + 5.7 × *temperature*

a. Use the regression equation to predict the number of people that will attend the pool if the *temperature* is 31 °C. Round to the nearest whole number.

b. Is the prediction from part **a** reliable? Explain briefly.

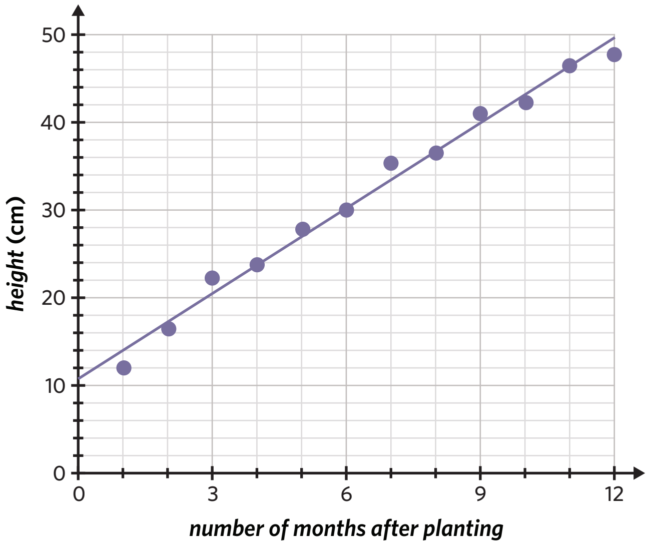
c. Use the regression equation to predict the number of people that will attend the pool if the *temperature* is 16 °C. Round to the nearest whole number.

d. Is the prediction from part **c** reliable? Explain briefly.



Joining it all together

8. A group of gardeners have developed a special soil to make cacti grow faster. In order to test it, they planted a small cactus in their soil and measured its *height* each month for a year. The gardeners displayed their results in a scatterplot and calculated the equation of a regression line that could predict the *height* (cm) of the cactus from the *number of months after planting*.



height = 10.89 + 3.23 × *number of months after planting*

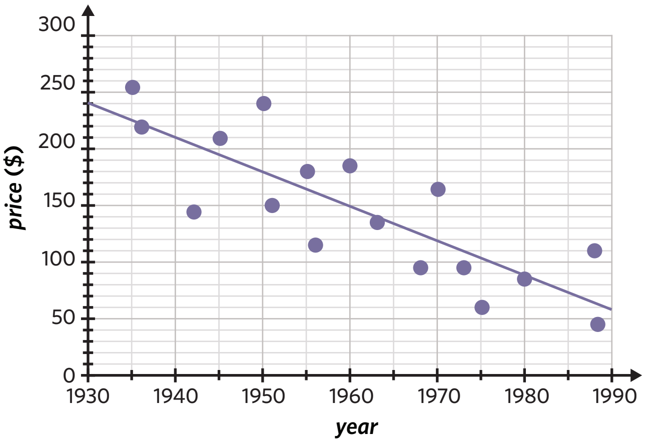
a. The gardeners forgot to measure the *height* of the cactus when they planted it into their soil.

Use the regression equation to find its predicted *height* when it was planted. Give the value correct to two decimal places.

b. On average, how much does the cactus grow in *height* each month? Give the value correct to two decimal places.

c. Predict what the *height* of the cactus was five and a half months after being planted. Give the value correct to two decimal places.

9. Anna owns a vintage dress shop and is trying to create an efficient way of pricing her dresses by basing the selling price on the year they were made. She usually prices the older dresses higher because they are more valuable. She has recorded the *year* that some of her dresses were made and their corresponding *price* ($). The results are shown in the scatterplot provided.



The equation of the regression line that Anna calculated is given.

price = 6079.28 − 3.03 × *year*

a. Complete the following sentence:

On average, for each unit increase in *year*, the *price* of the dress ___________ by ___________.

b. Predict the *price* of a dress that was made in 1955, correct to the nearest dollar.

c. Predict the *price* of a dress that was made in 1982, correct to the nearest dollar.

d. Is the prediction from part **b** reliable? Explain briefly.

10. The *maximum daily temperature* (°C) and *precipitation* (mm) were recorded over a two-week period. The results are shown in the following table.

***maximum daily temperature* (°C)**	20	15	11	19	22	16	14	18	20	21	10	22	15	16
***precipitation* (mm)**	2.2	4.9	5.7	1.2	1.3	3.4	2.2	1.1	1.2	0.5	6.2	0.4	2.3	2.1

The regression equation that allows *precipitation* to be predicted from *maximum daily temperature* is:

precipitation = 9.77 − 0.43 × *maximum daily temperature*

a. Interpret the slope of the least squares regression equation.

b. What is the *maximum daily temperature* that would result in no precipitation? Round to two decimal places.

c. Predict the *precipitation*, correct to two decimal places, if the daily temperature reaches a maximum of 25 °C. Is this prediction reliable?

d. Which of the following predictions would be most reliable?

A. *precipitation* when *maximum daily temperature* is 25 °C.

B. *precipitation* when *minimum daily temperature* is 15 °C.

C. *maximum daily temperature* when *precipitation* is 7 mm.

D. *maximum daily temperature* when *precipitation* is 1 mm.

Exam practice

11. The scatterplot provided shows the atmospheric pressure, in hectopascals (hPa), at 3 pm (*pressure 3 pm*) plotted against the atmospheric pressure, in hectopascals, at 9 am (*pressure 9 am*) for 23 days in November 2017 at a particular weather station.

A least squares line has been fitted to the scatterplot as shown.

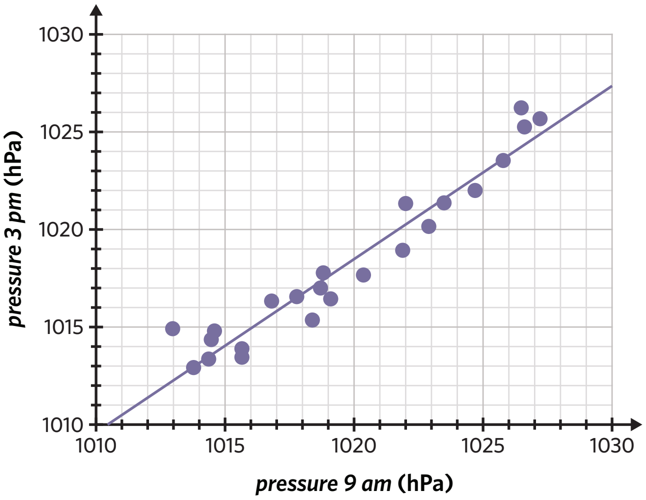
The equation of this line is

pressure 3 pm = 111.4 + 0.8894 × *pressure 9 am*

The equation of the least squares line is used to predict the atmospheric pressure at 3 pm when the atmospheric pressure at 9 am is 1025 hPa.

Is this prediction an example of extrapolation or interpolation? (1 MARK)

Adapted from VCAA 2019 Exam 2 Data analysis Q5c



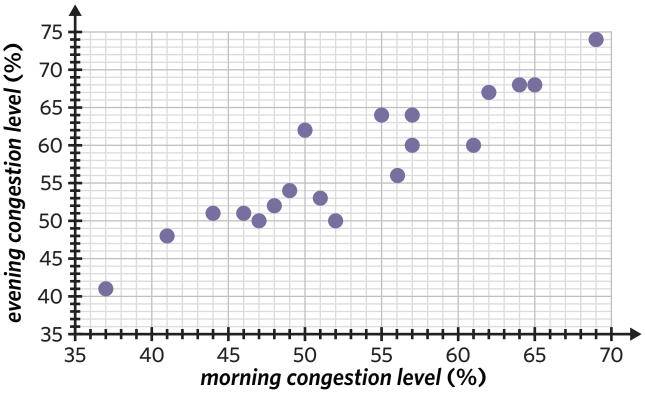
Data: Australian Government, Bureau of Meteorology, <www.bom.gov.au/>

87% of students answered this type of question correctly.

12. The congestion level in a city can be recorded as the percentage increase in travel time due to traffic congestion in peak periods (compared to non-peak periods).

This is called the percentage congestion level.

The percentage congestion levels for the morning and evening peak periods for 19 large cities are plotted on the scatterplot shown.



A least squares line is to be fitted to the data with the aim of predicting *evening congestion level* from *morning congestion level*.

The equation of this line is

evening congestion level = 8.48 + 0.922 × *morning congestion level*

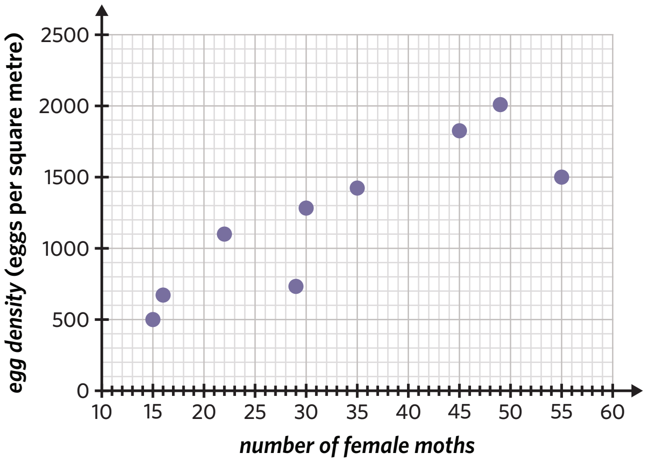
Use the equation of the least squares line to predict the *evening congestion level* when the *morning congestion level* is 60%. (1 MARK)

VCAA 2018 Exam 2 Data analysis Q2c

78% of students answered this question correctly.

13. The *number of female moths* caught in a trap set in a forest and the *egg density* (eggs per square metre) in the forest are examined.

A scatterplot of the data is shown.



A least squares regression line is fitted to the data. The equation of the least squares line is

egg density = 191 + 31.3 × *number of female moths*

Interpret the slope of the regression line in terms of the variables *egg density* and *number of female moths* caught in the trap. (1 MARK)

Adapted from VCAA 2017 Exam 2 Data analysis Q3bii

39% of students answered this type of question correctly.

14. The following table shows the yearly average traffic congestion levels in two cities, Melbourne and Sydney, during the period 2008 to 2016.

	***congestion level* (%)**								
year	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
Melbourne	25	26	26	27	28	28	28	29	33
Sydney	28	30	32	33	34	34	35	36	39

A least squares line is used to model the trend in the time series plot for Sydney. The equation is

congestion level = −2280 + 1.15 × *year*

The equation when a least squares line is used to model the trend in the data for Melbourne is

congestion level = −1515 + 0.7667 × *year*

Since 2008, the equations of the least squares lines for Sydney and Melbourne have predicted that future traffic congestion levels in Sydney will always exceed future traffic congestion levels in Melbourne.

Explain why, quoting the values of appropriate statistics. (2 MARKS)

Adapted from VCAA 2018 Exam 2 Data analysis Q3e

The average mark on this type of question was **0.4**.

Questions from multiple lessons

Data analysis

15. Which of the following graphical representations could be used to best identify and describe the association between the variables *age*, in years, and *favourite colour* (red, green, blue)?

A. a histogram

B. a scatterplot

C. a bar graph

D. a back-to-back stem plot

E. parallel boxplots

Adapted from VCAA 2017NH Exam 1 Data analysis Q6

Data analysis *Year 11 content*

16. A group of students is conducting research on the participation numbers of sports at their school, and decide to focus on the two following variables.

- *number of participants* (less than 25, 25–75, more than 75)
- *sport* (football, soccer, hockey, netball)

These variables are

A. both nominal variables.

B. both ordinal variables.

C. a nominal variable and ordinal variable respectively.

D. an ordinal variable and nominal variable respectively.

E. a numerical variable and a categorical variable respectively.

Adapted from VCAA 2017 Exam 1 Data analysis Q7

Recursion and financial modelling *Year 11 content*

17. Jack decides to deposit some money into a savings account that will pay interest every day so that he can go on a holiday.

The balance in Jack's account, in dollars after n days, V_n, can be modelled by the recurrence relation shown.

$V_0 = 15\,000, \quad V_{n+1} = 1.000041 \times V_n$

A rule of the form $V_n = a \times b^n$ can be used to determine the balance of Jack's account after n days.

a. Determine the values of a and b. (1 MARK)

b. What would be the value of n if Jack wants to determine the value of his investment after six years? Assume that there are no leap years. (1 MARK)

Adapted from VCAA 2018 Exam 2 Recursion and financial modelling Q4ci,ii

3C Performing a regression analysis

STUDY DESIGN DOT POINT

- interpretation of the slope and intercepts of the least squares line in the context of the situation being modelled, including:
  – use of the rule of the fitted line to make predictions being aware of the limitations of extrapolation
  – use of the coefficient of determination, r^2, to assess the strength of the association in terms of explained variation
  – use of residual analysis to check quality of fit

3A 3B 3C 3D 3E

KEY SKILLS

During this lesson, you will be:

- calculating and interpreting the coefficient of determination
- performing residual calculations and constructing a residual plot
- performing a residual analysis.

KEY TERMS

- Coefficient of determination
- Residual plot

Although a linear model can fit all data sets, it is not always the best fit. This can result in inaccurate predictions that impact the reliability of long-term modelling for a data set. The coefficient of determination and a residual analysis are tools that can help determine whether the current linear model is a quality fit for the data set being analysed.

Calculating and interpreting the coefficient of determination

The **coefficient of determination**, r^2, is the degree to which the variation in the response variable can be predicted from the variation in the explanatory variable for a given linear relationship. It is calculated by squaring the correlation coefficient, r. As it is the square of the correlation coefficient, the r^2 value will always be between 0 and 1.

$0 \leq r^2 \leq 1$

In order to interpret the coefficient of determination, it must first be converted into a percentage.

A lower coefficient of determination could indicate that a linear model might not be the best fit for the data set.

Worked example 1

A study was conducted to explore the effect of the time in minutes spent playing *video games* on time spent *studying* in minutes. It was determined that the correlation coefficient, r, for the data collected during the study is 0.6584.

a. Determine the coefficient of determination, r^2, correct to four decimal places.

Explanation

Calculate the coefficient of determination.

$(0.6584)^2 = 0.43349...$

Continues →

Answer

0.4335

b. Interpret the coefficient of determination in terms of time spent *studying* and time spent playing *video games*.

Explanation

Step 1: Determine the explanatory and response variables.

Since the time spent *studying* is affected by the time spent playing *video games*, *studying* is the response variable.

EV: *video games*

RV: *studying*

Step 2: Convert the coefficient of determination into a percentage.

$0.4335 \times 100 = 43.35$

Step 3: Interpret the coefficient of determination in terms of time spent playing *video games* and time spent *studying*.

Answer

43.35% of the variation in the time spent *studying* can be explained by the variation in the time spent playing *video games*.

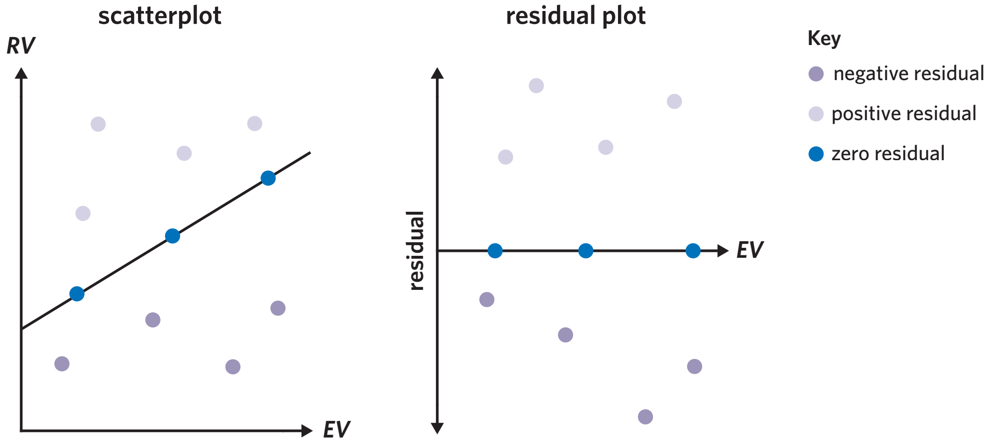
Performing residual calculations and constructing a residual plot

Recall that a residual is the vertical distance between the actual data value and the value predicted from the regression line. A residual can be calculated using the formula

residual = *actual data value* − *predicted data value*

Data points that sit above the regression line will have a positive residual value. Those that fall below the regression line will have a negative residual value. Those that lie on the regression line will have a residual value of zero.

A **residual plot** is a graph of the residual values against the explanatory variable. It can be created once all of the residual values are calculated.



Worked example 2

A least squares regression equation has been used to predict the *total cost* ($) of a table bill at a restaurant from the *number of people* eating together on the table, such that

total cost $= 7 + 29 \times$ *number of people*

a. If the actual *total cost* for a table of six people is $166, calculate the residual value when the *total cost* is predicted using the regression equation.

Explanation

Step 1: Determine the actual data value.

The actual data value is the cost given in the question.

actual data value $= 166$

Step 2: Calculate the predicted data value.

Substitute *number of people* $= 6$ into the regression equation.

total cost $= 7 + 29 \times 6$

$= 181$

predicted data value $= 181$

Step 3: Calculate the residual.

Substitute *actual data value* $= 166$ and *predicted data value* $= 181$ into the residual formula.

residual $= 166 - 181$

Answer

-15

b. There were 8 other groups of people at the restaurant. The *number of people* at each table and the *total cost* for each group were recorded and displayed in a table.

number of people	3	4	1	6	5	2	2	4
actual *total cost* ($)	90	133	29	174	154	73	62	129
predicted *total cost* ($)								
residual value ($)								

Use the regression equation to complete the table by filling in the predicted *total cost* and residual values.

Explanation

Step 1: Fill in the table with the predicted data values.

Substitute each value for *number of people* from the table into the regression equation.

3 people: *total cost* $= 7 + 29 \times 3 = 94$

4 people: *total cost* $= 7 + 29 \times 4 = 123$

number of people	3	4	1	6	5	2	2	4
actual *total cost* ($)	90	133	29	174	154	73	62	129
predicted *total cost* ($)	94	123	36	181	152	65	65	123
residual value ($)								

Step 2: Fill in the table with the residual values.

Substitute the actual and predicted *total cost* for each group into the residual formula.

3 people: *residual* $= 90 - 94 = -4$

4 people: *residual* $= 133 - 123 = 10$

Continues →

Answer

number of people	3	4	1	6	5	2	2	4
actual *total cost* ($)	90	133	29	174	154	73	62	129
predicted *total cost* ($)	94	123	36	181	152	65	65	123
residual value ($)	−4	10	−7	−7	2	8	−3	6

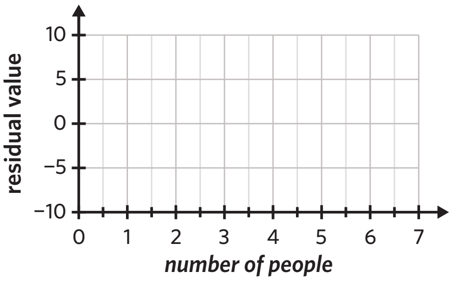
c. Construct a residual plot for this data set.

Explanation – Method 1: By hand

Step 1: Draw a set of axes with an appropriate scale and labels.

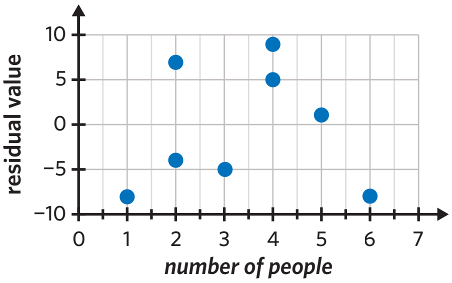
The label for the horizontal axis will be the explanatory variable, *number of people*.

The label for the vertical axis will be ‘residual value’.



Step 2: Plot each residual value against the number of people.

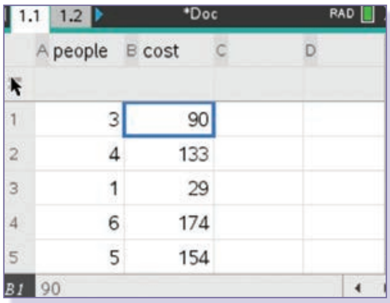
Answer



Explanation – Method 2: TI-Nspire

Step 1: From the home screen, select ‘1: New’ → ‘4: Add Lists & Spreadsheet’.

Step 2: Name column A ‘people’ and column B ‘cost’ and enter the data from the table.



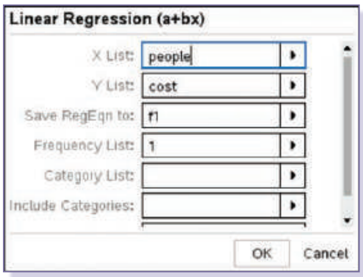
	A people	B cost	C	D
1	3	90		
2	4	133		
3	1	29		
4	6	174		
5	5	154		

B1 90

Continues →

Step 3: Determine the least squares regression line. Press menu → '4: Statistics' → '1: Stat Calculations' → '4: Linear Regression (a+bx)'.

Step 4: On the settings window, change 'X List' to 'people' and 'Y List' to 'cost'. Check that '1st Result Column' is set to 'c[ ]'. Select 'OK'.



Linear Regression (a+bx)

X List: people

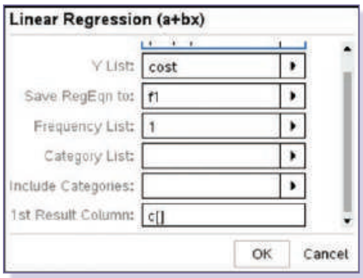
Y List: cost

Save RegEqn to: f1

Frequency List: 1

Category List:

Include Categories:



Linear Regression (a+bx)

Y List: cost

Save RegEqn to: f1

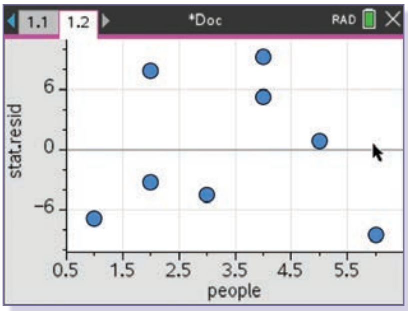
Frequency List: 1

Category List:

Include Categories:

1st Result Column: c[]

Answer



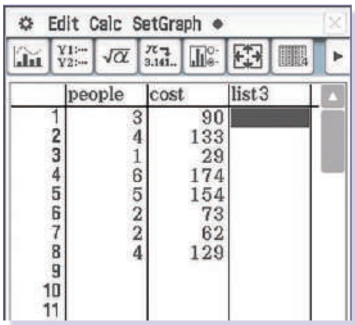
Step 5: Press ctrl + doc and select 'Add Data & Statistics'. Add the variables to each axis using the 'Click to add variable' function.

Select 'people' on the horizontal axis and 'stat.resid' on the vertical axis.

Explanation – Method 3: Casio ClassPad

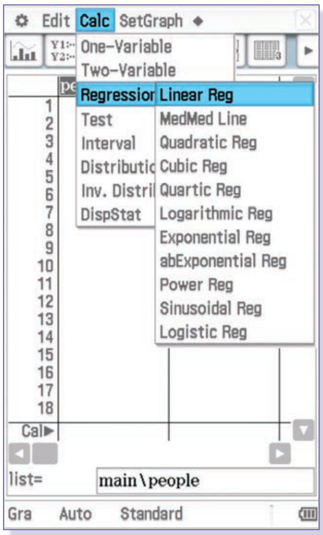
Step 1: From the main menu, tap Statistics.

Step 2: Name list1 'people' and list2 'cost' and enter the data from the table.



	people	cost	list3
1	3	90	
2	4	133	
3	1	29	
4	6	174	
5	5	154	
6	2	73	
7	2	62	
8	4	129	
9			
10			
11			

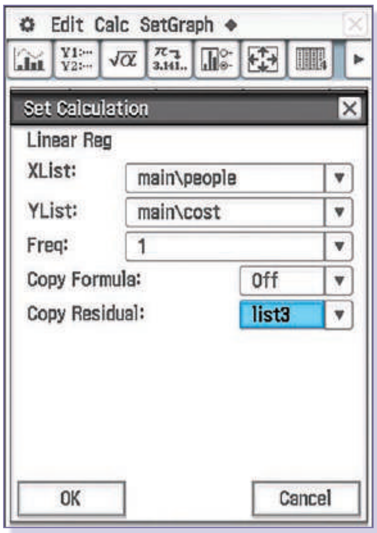
Step 3: Tap 'Calc' → 'Regression' → 'Linear Reg'.



list= main\people

Continues →

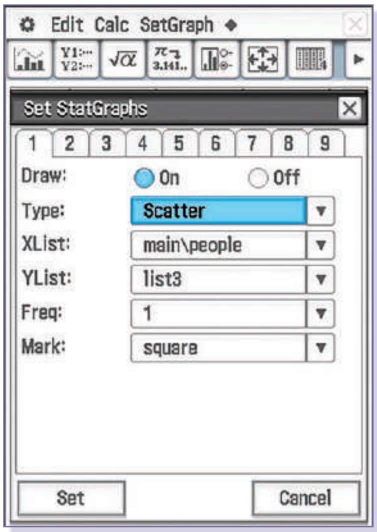
Step 4: Specify the data set by changing 'XList:' to 'main\people' and 'YList:' to 'main\cost'. Change 'Copy Residual' to 'list3'.



Linear Reg
XList: main\people
YList: main\cost
Freq: 1
Copy Formula: Off
Copy Residual: list3

Step 5: Tap 'OK' and the 'Stat Calculations' window will appear. Tap 'OK' to close this window.

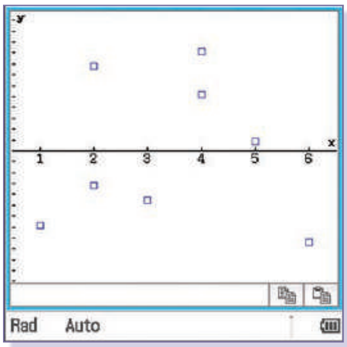
Step 6: Tap and the 'Set StatGraphs' window will appear. Change 'Type 'to 'Scatter'. Change 'XList:' to 'main\people' and 'YList:' to 'list3'. Tap set to confirm.



Draw: ● On ○ Off
Type: Scatter
XList: main\people
YList: list3
Freq: 1
Mark: square

Step 7: Tap to display the residual plot.

Answer

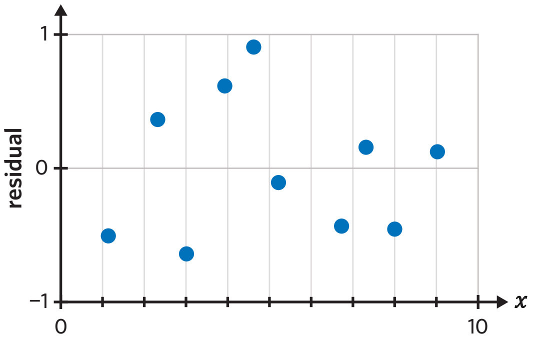


Performing a residual analysis

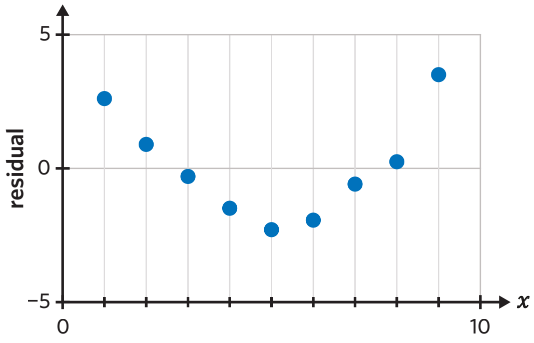
Residual plots can be analysed to help determine if a linear relationship exists between the explanatory and response variables.

If a linear relationship is present, the residual plot will show an approximately equal number of points randomly scattered above and below the horizontal axis, with no clear pattern. These plots support the assumption of a linear relationship between the variables.

If a linear relationship is not present, the residual plot might show an unequal number of points above and below the horizontal axis or an ordered scattering of the points, showing a clear pattern. These plots do not support the assumption of a linear relationship between the variables.



This residual plot supports the assumption of linearity.

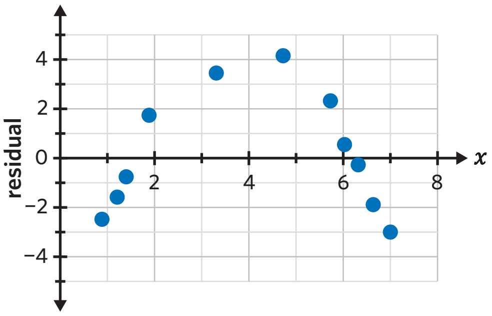


This residual plot does not support the assumption of linearity.

Worked example 3

Determine if the following residual plots support the assumption of linearity.

a.



Explanation

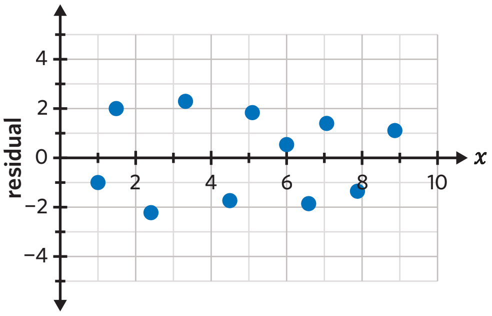
Check if the residual plot meets the conditions to support the assumption of linearity.

The residual plot appears to show a clear curved pattern.

Answer

The residual plot does not support the assumption of linearity.

b.



Explanation

Check if the residual plot meets the conditions to support the assumption of linearity.

The residual plot shows an approximately equal number of points above and below the horizontal axis. The points also appear to be randomly scattered and do not show a pattern.

Answer

The residual plot supports the assumption of linearity.

Exam question breakdown

VCAA 2021 Exam 2 Data analysis Q2b

In the sport of heptathlon, athletes compete in seven events. The two running events in the heptathlon are the 200 m and the 800 m run. The times taken by the athletes in these two events, *time200* and *time800*, are linearly related.

The mean and standard deviation for each variable, *time200* and *time800*, are shown in the table.

statistic	*time200* (seconds)	*time800* (seconds)
mean	24.6492	136.054
standard deviation	0.96956	8.2910

The equation of the least squares line is

$time800 = 0.03931 + 5.2756 \times time200$

Use this information to calculate the coefficient of determination as a percentage.

Round to the nearest percentage. (2 MARKS)

Explanation

Step 1: Identify the correct method to calculate the coefficient of determination.

As the coefficient of determination is r^2, the only way to calculate it from the given information is to calculate the correlation coefficient (r) first and square it.

The equation that can be used is

$$b = \frac{r \times s_y}{s_x}$$

Step 2: Identify the values needed to calculate the slope using summary statistics.

*time*800 is the response variable and *time*200 is the explanatory variable. This will help in determining which statistic is s_x and which is s_y.

$s_y = 8.2910,\ s_x = 0.96956,\ b = 5.2756$

Step 3: Substitute the known values and solve for r.

$$b = \frac{r \times s_y}{s_x}$$

$$5.2756 = \frac{r \times 8.2910}{0.96956}$$

$r = 0.6169...$

Step 4: Calculate the coefficient of determination.

$$r^2 = (0.6169...)^2$$
$$= 0.3806...$$

Step 5: Convert the coefficient of determination into a percentage.

$0.3806... \times 100 = 38.06...$

Answer

38%

The average mark on this question was **0.6**.

A number of students struggled to calculate Pearson’s correlation coefficient from the summary statistics formulas. This is likely because they did not remember how the correlation coefficient is related to the slope of a least squares regression equation. Errors could have also been made by misinterpreting which variables were explanatory and response, which could have led to the incorrect use of the summary statistics given.

3C Questions

Calculating and interpreting the coefficient of determination

1. The *price of burgers* and *revenue* (per annum) for seven different burger stores are collected and studied. The r value for the collected data is 0.785.

The coefficient of determination for the study is closest to

A. 0.616 **B.** 0.617 **C.** 0.785 **D.** 0.886

2. A study was conducted to research the impacts of *hours worked* on *hours slept*. The results are summarised in the following table.

hours worked	4	5	6	7	8	9	10
hours slept	9.0	8.5	8.3	8.0	7.5	7.0	6.8

a. What is the coefficient of determination, rounded to two decimal places?

b. Interpret the coefficient of determination in terms of the amount of *hours worked* and *hours slept*.

Performing residual calculations and constructing a residual plot

3. Rachel's Health and Human Development class is studying the average growth in the *height* of children and young teenagers. They determine the following least squares regression equation, which can be used to predict *height* (cm) from *age* (years).

$\textit{height} = 47.32 + 7.13 \times \textit{age}$

Rachel is 14 years old and 152 cm tall. What is the residual value when her *height* is predicted from the least squares regression equation, correct to two decimal places?

A. −4.86 **B.** 0.68 **C.** 4.86 **D.** 152.00

4. A group of scientists collected the IQ of 5 participants. The scientists then determined a least squares regression equation that can be used to predict *test score* on a particular intelligence test from *IQ*. The predicted test scores were calculated and recorded.

The participants then completed the intelligence test and the scientists recorded their actual *test scores*, before calculating the residual values of each participant. However, they lost the predicted *test score* and residual value for the participant with an *IQ* of 91.

IQ	101	87	94	106	91
actual *test score* (%)	76	54	60	80	54
predicted *test score* (%)	72.90	50.92	61.91	80.75	
residual value (%)	3.10	3.08	−1.91	−0.75	

a. Find the regression equation used by the scientists. Round values to two decimal places.

b. Use the regression equation from part **a** to find the predicted *test score* and residual value for the participant with an *IQ* of 91.

5. The following table shows the *length* (cm) and *width* (cm) of seven rectangular cakes at 'Chloe's Cake Shop'.

***length* (cm)**	32	45	50	28	40	35	56
***width* (cm)**	26	35	50	20	25	35	44

A regression equation has been calculated based on these results and can be used to predict the width of a cake from its length.

$\textit{width} = -3.64 + 0.91 \times \textit{length}$

Use the regression equation to fill in the predicted widths and residual values in the following table.

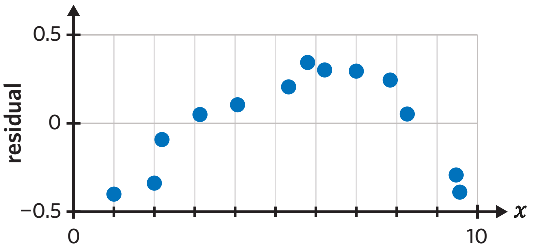
Give values correct to two decimal places.

***length* (cm)**	32	45	50	28	40	35	56
***width* (cm)**	26	35	50	20	25	35	44
predicted *width* (cm)							
residual *value* (cm)							

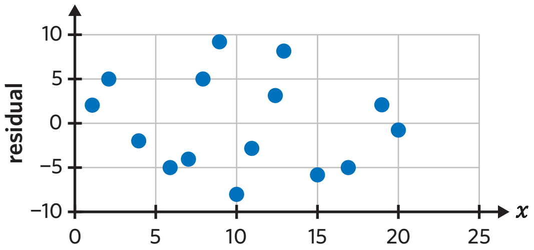
Performing a residual analysis

6. For each of the following residual plots, determine whether the assumption of linearity is supported.

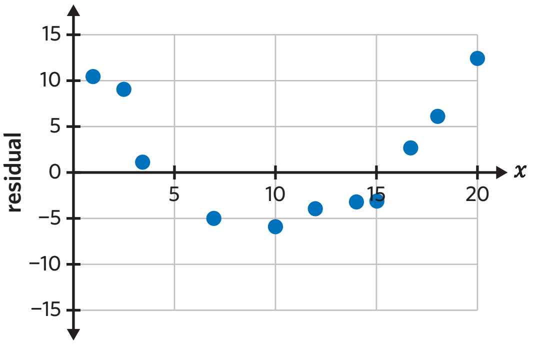
a.



b.



c.



7. The *height* (m) of a small tree was recorded at the end of every year from the time it was planted. The results are displayed in a table.

years after planting	1	2	3	4	5	6	7	8
***height* (m)**	0.76	1.58	2.15	2.54	2.98	3.21	3.45	3.55

Use a calculator to construct a residual plot and use it to test the assumption of linearity.

Joining it all together

8. A Year 10 Maths class recently completed a statistics test and a calculus test. Six students recorded their results from both tests. They are displayed in the following table.

	Sophie	**Isaac**	**Mayu**	**Deanne**	**Selby**	**Emily**
***statistics test result* (%)**	75	61	88	43	94	64
***calculus test result* (%)**	64	53	95	40	92	59

The students calculated the following least squares regression equation that could predict a *calculus test result* from a *statistics test result*.

calculus test result = −12.66 + 1.13 × *statistics test result*

a. Use the least squares regression equation to determine each student's predicted *calculus test result* (%), correct to one decimal place.

b. Using the rounded answers from part **a**, calculate the residual values when the *calculus test result* (%) is predicted from the least squares regression equation. Give values correct to one decimal place.

c. Which student(s) didn't perform as well as was predicted on the calculus test?

d. Construct a residual plot by hand and use it to determine if the assumption of linearity is supported for the relationship between *statistics test result* and *calculus test result* (%).

9. A scientific study was conducted measuring *age* (years) against *reaction time* (milliseconds).

The results of ten people are displayed in the following table.

age **(years)**	19	42	73	56	23	85	61	49	34	58
reaction time **(ms)**	3.2	5.5	9.6	9.1	6.0	14.1	11.7	8.3	6.6	8.0

a. What is the value of the coefficient of determination, correct to two decimal places?

b. Interpret the coefficient of determination in terms of *age* and *reaction time*.

c. Determine the least squares regression equation that can be used to predict *reaction time* (milliseconds) from *age* (years). Give values correct to two decimal places.

d. Fill in the following table by calculating the residual values. Give values correct to two decimal places.

age **(years)**	19	42	73	56	23	85	61	49	34	58
reaction time **(ms)**	3.2	5.5	9.6	9.1	6.0	14.1	11.7	8.3	6.6	8.0
residual value (ms)										

e. Construct a residual plot on a calculator and use it to determine if the assumption of linearity for the relationship between *age* and *reaction time* is supported.

10. Sebastian thinks that there may be a relationship between the amount of time he spends on his phone and the amount of sleep that he gets. He records his results for 8 days and is summarised in the table.

time spent on phone **(min)**	175	115	75	205	155	220	80	165
sleep **(hours)**	6.0	8.0	9.5	5.0	6.2	5.4	9.0	6.1

a. What is the value of the coefficient of determination, correct to three significant figures?

b. Interpret the coefficient of determination in terms of *time spent on phone* and *sleep*.

c. Determine the least squares regression line that can be used to predict *sleep* (hours) from *time spent on phone* (min). Give values correct to four significant figures.

d. Fill in the following table by calculating the residual values. Give values correct to 2 decimal places.

time spent on phone **(min)**	175	115	75	205	155	220	80	165
sleep **(hours)**	6.0	8.0	9.5	5.0	6.2	5.4	9.0	6.1
residual value (hours)								

e. Construct a residual plot on a calculator and use it to determine if the assumption of linearity for the relationship between *time spent on phone* and *sleep* is supported.

Exam practice

11. The congestion level in a city can be recorded as the percentage increase in travel due to traffic congestion in peak periods (compared to non-peak periods).

This is called the percentage congestion level.

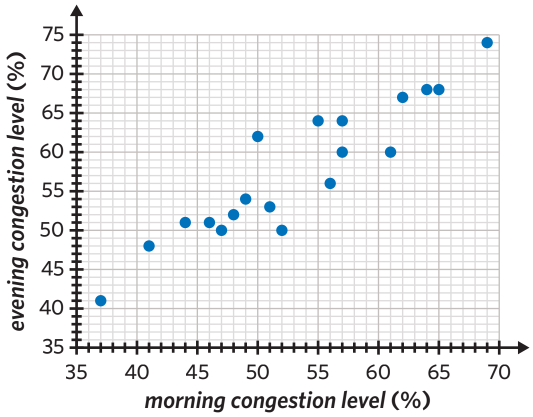
The percentage congestion levels for the morning and evening peak periods for 19 large cities are plotted on the scatterplot.

The value of the correlation coefficient r is 0.92.

What percentage of the variation in the *evening congestion level* can be explained by the variation in the *morning congestion level*? Round to the nearest percentage. (1 MARK)

VCAA 2018 Exam 2 Data analysis Q2e

59% of students answered this question correctly.



12. The following scatterplot displays the *resting pulse rate*, in beats per minute, and the *time spent exercising* in hours per week, of 16 students. A least squares line has been fitted to the data.

Using this least squares line to model the association between *resting pulse rate* and *time spent exercising*, the residual for the student who spent four hours per week exercising is closest to

A. −2.0 beats per minute.

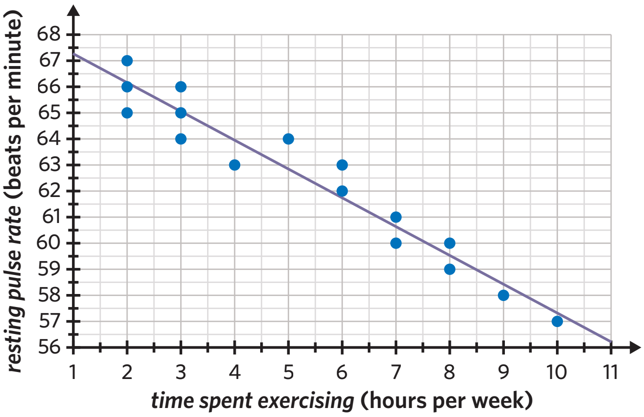
B. −1.0 beats per minute.

C. −0.3 beats per minute.

D. 1.0 beats per minute.

E. 2.0 beats per minute.

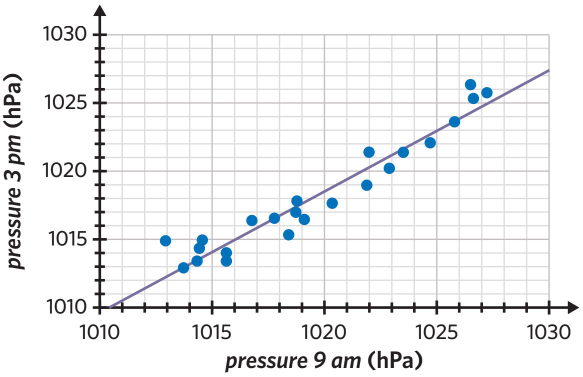
VCAA 2018 Exam 1 Data analysis Q7



73% of students answered this question correctly.

13. The scatterplot shows the atmospheric pressure, in hectopascals (hPa), at 3 pm (*pressure 3 pm*) plotted against the atmospheric pressure, in hectopascals, at 9 am (*pressure 9 am*) for 23 days in November 2017 at a particular weather station.

A least squares line has been fitted to the scatterplot as shown.

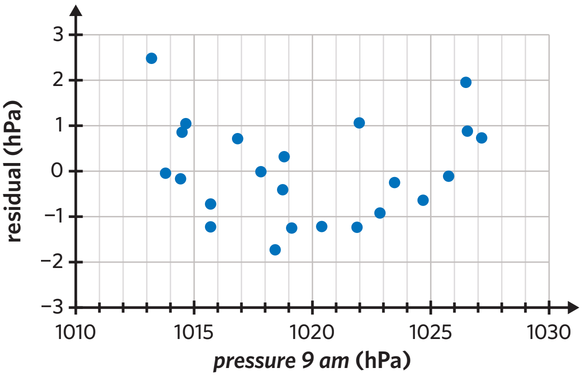


Data: Australian Government, Bureau of Meteorology, <www.bom.gov.au/>

The equation of this line is

pressure 3 pm = 111.4 + 0.8894 × *pressure 9 am*

The residual plot associated with the least squares line is shown.



a. The residual plot can be used to test one of the assumptions about the nature of the association between the atmospheric pressure at 3 pm and the atmospheric pressure at 9 am.

What is this assumption? (1 MARK)

b. The residual plot shown does not support this assumption.

Explain why. (1 MARK)

VCAA 2019 Exam 2 Data analysis Q5f

Part **a**: **41%** of students answered this question correctly.

Part **b**: **36%** of students answered this question correctly.

14. The following plot shows the *winning time*, in seconds, for the women's 100 m freestyle swim plotted against *year*, for each year that the Olympic Games were held during the period 1956 to 2016.

A least squares line has been fitted to the plot to model the decreasing trend in the *winning time* over this period.

The equation of the least squares line is

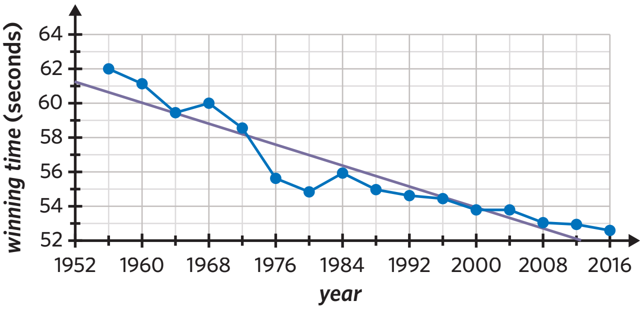
winning time = 357.1 − 0.1515 × *year*

The coefficient of determination is 0.8794.

Determine the value of the correlation coefficient (r).

Round to three decimal places. (1 MARK)

Adapted from VCAA 2021 Exam 2 Data analysis Q3b



Data: International Olympic Committee, <https://olympics.com/en/olympic-games/olympic-results>

19% of students answered this type of question correctly.

Questions from multiple lessons

Data analysis *Year 11 content*

15. The variables *time* (less than 8 minutes, 8–10 minutes, over 10 minutes) and *size* (small, medium, large) are

A. both numerical variables.

B. both nominal variables.

C. both ordinal variables.

D. a numerical and ordinal variable respectively.

E. an ordinal and nominal variable respectively.

Adapted from VCAA 2018NH Exam 1 Data analysis Q5

Data analysis *Year 11 content*

16. The following scatterplot displays the association between the *size of house*, in square metres, and *distance from CBD*, in kilometres for 15 similarly-priced Melbourne houses.

A least squares regression line has been fitted to the data with *distance from CBD* as the explanatory variable.

The equation of the least squares regression line could be

A. *distance from CBD* = 74.5 + 29.2 × *size of house*

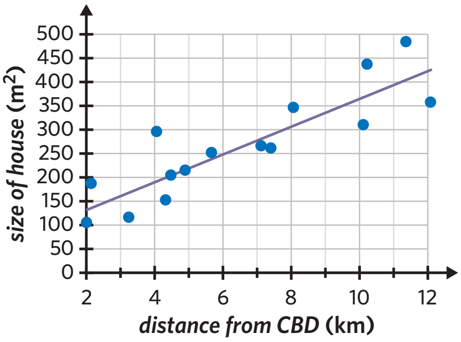
B. *distance from CBD* = 133 + 29.2 × *size of house*

C. *size of house* = 74.5 + 29.2 × *distance from CBD*

D. *size of house* = 133 + 29.2 × *distance from CBD*

E. *size of house* = 74.5 + 85.1 × *distance from CBD*

Adapted from VCAA 2017 Exam 1 Data analysis Q8



Data analysis

17. The following segmented bar chart displays the sport played by the students at three schools.

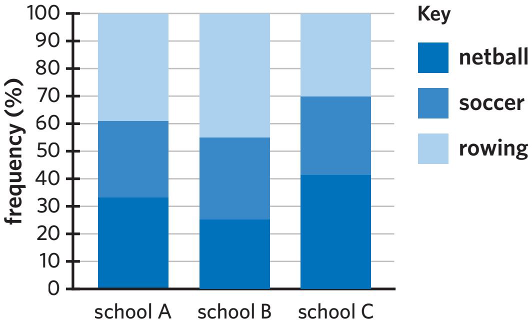
a. What percentage of students from school B play netball? (1 MARK)

b. There are 350 students at school C who play a sport. How many of these students do rowing? (1 MARK)

c. From this bar chart, it can be concluded that there is no association between the percentage of students who play soccer and the school which they attend.

Explain why this can be concluded and quote any appropriate percentages. (1 MARK)

Adapted from VCAA 2014 Exam 2 Data analysis Q1



3D Data transformations

STUDY DESIGN DOT POINT

- data transformation and its use in transforming some forms of non-linear data to linearity using a square, logarithmic (base 10), or reciprocal transformation (applied to one axis only)

3A 3B 3C **3D** 3E

KEY SKILLS

During this lesson, you will be:

- choosing an appropriate data transformation
- applying a squared transformation
- applying a log transformation
- applying a reciprocal transformation.

KEY TERMS

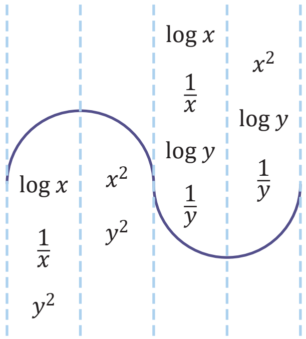
- Linearise
- x-squared transformation
- y-squared transformation
- $\log x$ transformation
- $\log y$ transformation
- Reciprocal
- x-reciprocal transformation
- y-reciprocal transformation

A least squares regression line should not be fitted to data if it is not linear, as any interpretations or predictions will not be accurate. If data is not linear, it may be possible to linearise it by applying a transformation to one of the variables. Three possible transformations are a squared transformation, a log transformation, and a reciprocal transformation.

Choosing an appropriate data transformation

To **linearise** data is to use a transformation to make non-linear data linear. There are three main types of transformations used to linearise data:

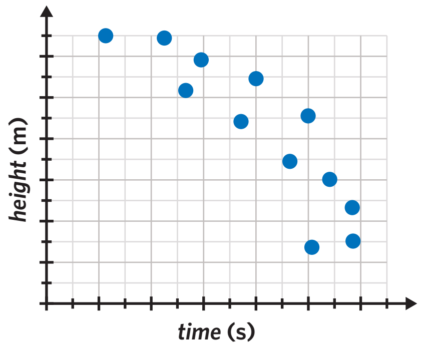
- squared transformations $(x^2 \text{ and } y^2)$
- log (base 10) transformations ($\log x$ and $\log y$)
- reciprocal transformations $\left(\frac{1}{x} \text{ and } \frac{1}{y}\right)$



The transformation wave can help identify which transformations are most appropriate to linearise a distribution of data. The shape of the relationship between the variables should be compared to each of the segments of the transformation wave. All transformation options provided within the most similarly shaped segment can be used to linearise the data.

Worked example 1

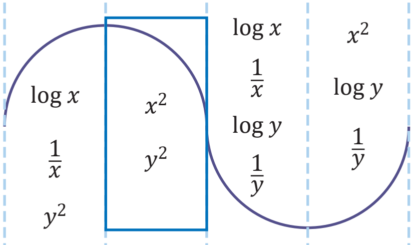
Determine the transformations that could be used to linearise the data in the scatterplot.



Continues →

Explanation

Step 1: Identify which segment of the transformation wave this scatterplot resembles.



The scatterplot most closely resembles the second segment of the transformation wave.

Step 2: Identify which transformations may be applied to this segment.

The x^2 or y^2 transformations may be applied.

Step 3: Express the transformations in terms of the variables given.

The x variable is *time* and the y variable is *height* in this instance. The transformations become $time^2$ and $height^2$.

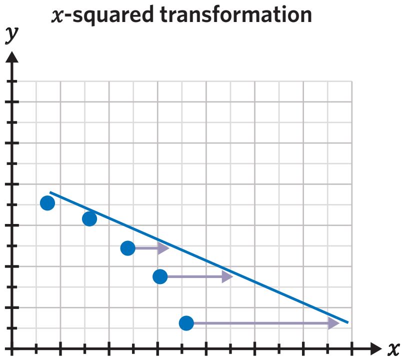
Answer

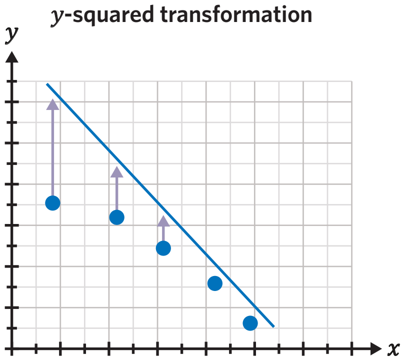
The $time^2$ or $height^2$ transformations could be applied.

Applying a squared transformation

The ***x*-squared transformation** involves 'stretching' the larger x values more than the smaller x values. The y values remain the same.

The ***y*-squared transformation** involves 'stretching' the larger y values more than the smaller y values. The x values remain the same.





Worked example 2

A scatterplot was constructed from the following data.

x	3	4	5	6	7
y	8	17	24	33	52

Apply an x-squared transformation and plot the transformed data.

Explanation – Method 1: By hand

Step 1: Calculate the square of all the x values.

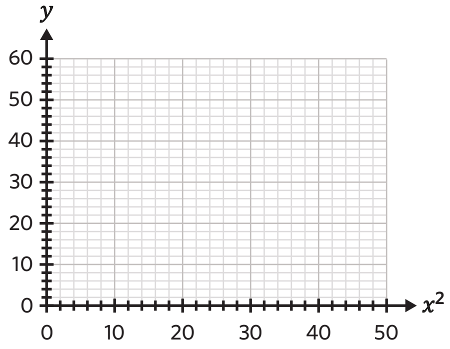
x	3	4	5	6	7
y	8	17	24	33	52
x^2	9	16	25	36	49

Continues →

Step 2: Construct a set of axes.

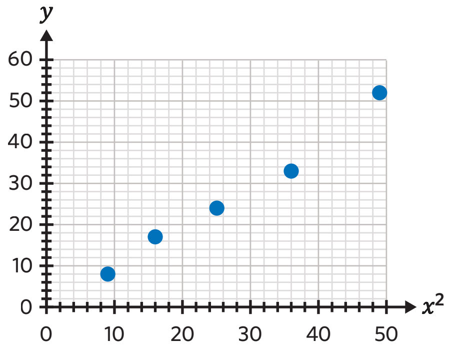
A scale from 0 to 50 is appropriate for the horizontal axis, while the vertical axis needs to extend from 0 to at least 52.

The horizontal axis should be labelled x^2.



Step 3: Plot the data points using the x^2 values rather than the x values.

Answer



Explanation – Method 2: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Name column A 'x' and column B 'y'.

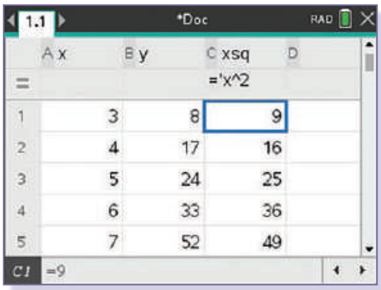
Enter the x values into column A, starting from row 1.

Enter the y values into column B, starting from row 1.

Step 3: Name column C 'xsq' (short for x squared).

Enter '=x^2' into the cell below the 'xsq' heading.

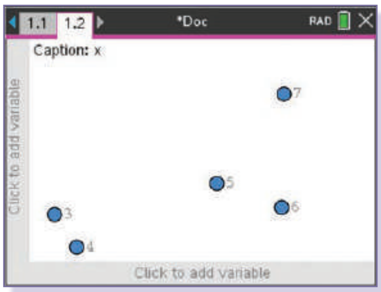
Select 'Variable Reference' → 'OK'.



	A x	B y	C xsq	D
=			='x^2	
1	3	8	9	
2	4	17	16	
3	5	24	25	
4	6	33	36	
5	7	52	49	

C1 =9

Step 4: Press ctrl + doc ▾, and select '5: Add Data & Statistics'.



Step 5: Move the cursor to the horizontal axis and select 'Click to add variable'.

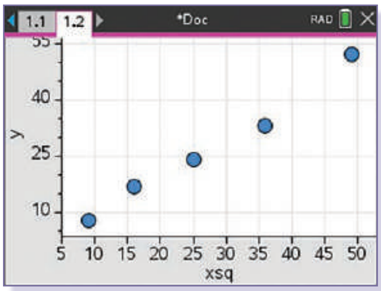
Select 'xsq'.

Move the cursor to the vertical axis and select 'Click to add variable'.

Select 'y'.

Continues →

Answer



Explanation – Method 3: Casio ClassPad

Step 1: From the main menu, tap Statistics.

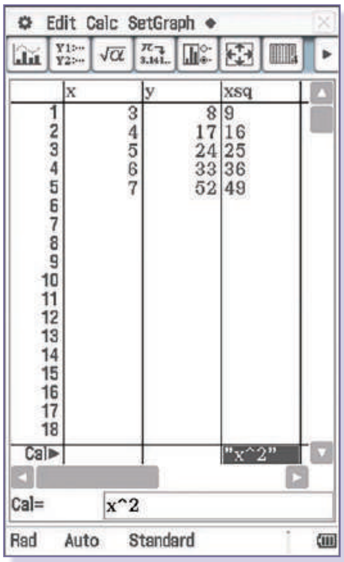
Step 2: Name the first list ‘x’ and the second list ‘y’.

Enter the x values into list ‘x’, starting from row 1.

Enter the y values into list ‘y’, starting from row 1.

Step 3: Name the third list ‘xsq’ (short for x squared).

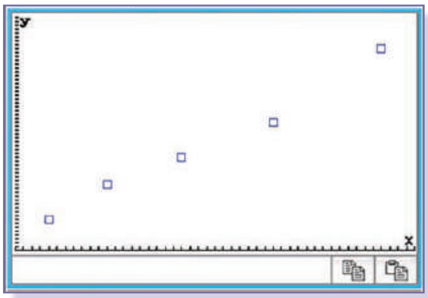
In the third list, go down to the calculation cell Cal▶ and enter ‘x^2’.



	x	y	xsq
1	3	8	9
2	4	17	16
3	5	24	25
4	6	33	36
5	7	52	49
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
Cal▶			"x^2"

Cal= x^2

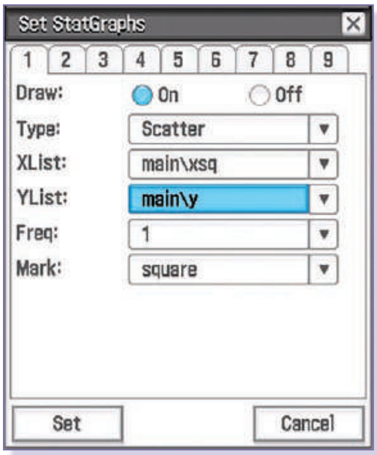
Answer



Step 4: Configure the settings of the graph by tapping .

Create a scatterplot by changing ‘Type’ to ‘Scatter’.

Specify the data set by changing ‘XList:’ to ‘main\xsq’ and ‘YList:’ to ‘main\y’.



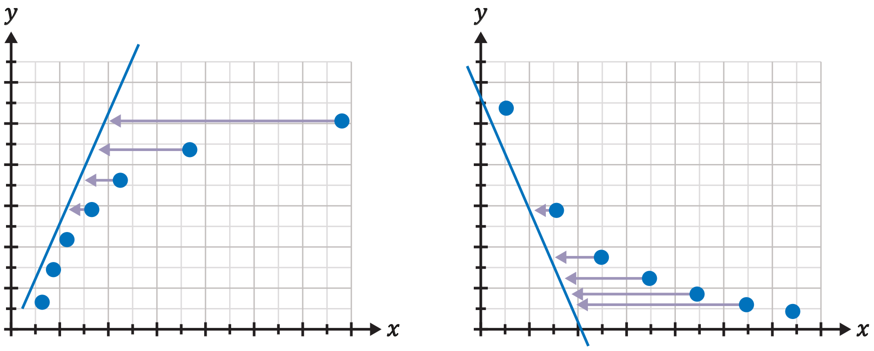
Draw: ● On ○ Off
Type: Scatter
XList: main\xsq
YList: main\y
Freq: 1
Mark: square

Tap ‘Set’ to confirm.

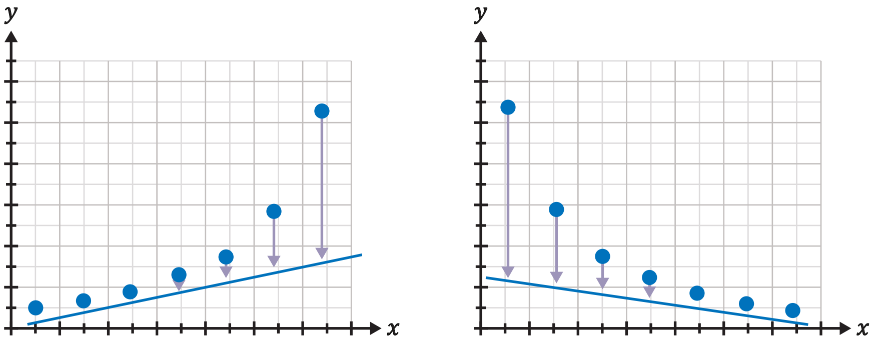
Step 5: Tap to plot the graph.

Applying a log transformation

The **log *x* transformation** involves 'compressing' the larger *x* values more than the smaller *x* values. The *y* values remain the same.



The **log *y* transformation** involves 'compressing' the larger *y* values more than the smaller *y* values. The *x* values remain the same.



Worked example 3

Augustus participated in a sushi eating competition and noted his progress at certain stages of the 14 minute event. His results are shown in the table.

time **(min)**	2	3	5	7	8	11	13	14
sushi rolls eaten	21	35	51	57	62	64	67	67

Apply a log *x* transformation and plot the transformed data.

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

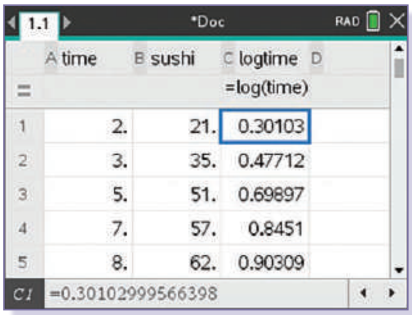
Step 2: Name column A 'time' and column B 'sushi'.

Enter the *time* values into column A, starting from row 1.

Enter the *sushi rolls eaten* values into column B, starting from row 1.

Step 3: Name column C 'logtime'.

Enter '=log(time)' into the cell below the 'logtime' heading.

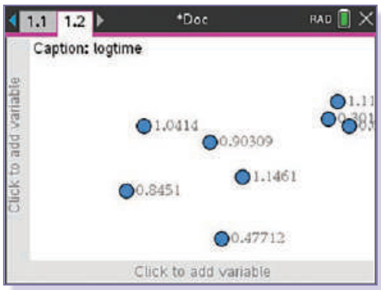


	A time	B sushi	C logtime	D
=			=log(time)	
1	2.	21.	0.30103	
2	3.	35.	0.47712	
3	5.	51.	0.69897	
4	7.	57.	0.8451	
5	8.	62.	0.90309	

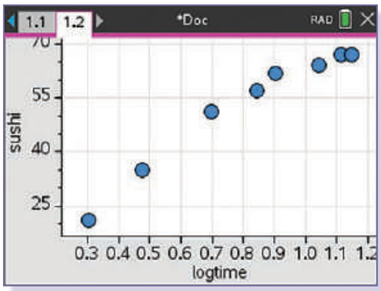
C1 =0.30102999566398

Continues →

Step 4: Press ctrl + doc▾, and select '5: Add Data & Statistics'.



Answer



Step 5: Move the cursor to the horizontal axis and select 'Click to add variable'.

Select 'logtime'.

Move the cursor to the vertical axis and select 'Click to add variable'.

Select 'sushi'.

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

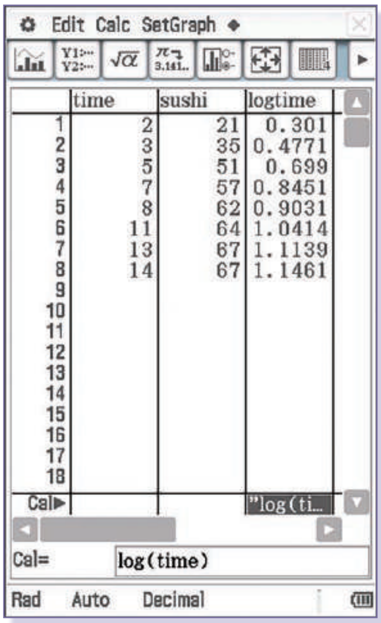
Step 2: Name the first list 'time' and the second list 'sushi'.

Enter the *time* values into list 'time', starting from row 1.

Enter the *sushi rolls eaten* values into list 'sushi', starting from row 1.

Step 3: Name the third list 'logtime'.

In the third list, go down to the calculation cell Cal▶ and enter 'log(time)'.



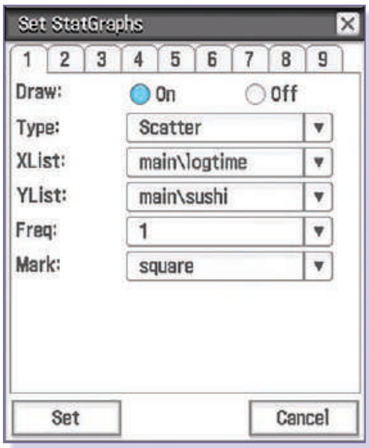
	time	sushi	logtime
1	2	21	0.301
2	3	35	0.4771
3	5	51	0.699
4	7	57	0.8451
5	8	62	0.9031
6	11	64	1.0414
7	13	67	1.1139
8	14	67	1.1461
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
Cal▶			"log(ti…

Cal= log(time)

Step 4: Configure the settings of the graph by tapping .

Create a scatterplot by changing 'Type' to 'Scatter'.

Specify the data set by changing 'XList:' to 'main\logtime' and 'YList:' to 'main\sushi'.



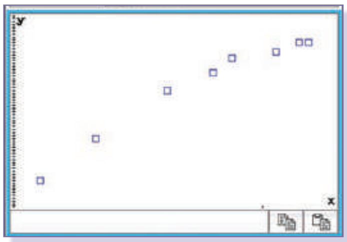
Draw: ● On ○ Off
Type: Scatter
XList: main\logtime
YList: main\sushi
Freq: 1
Mark: square

Tap 'Set' to confirm.

Step 5: Tap to plot the graph.

Continues →

Answer

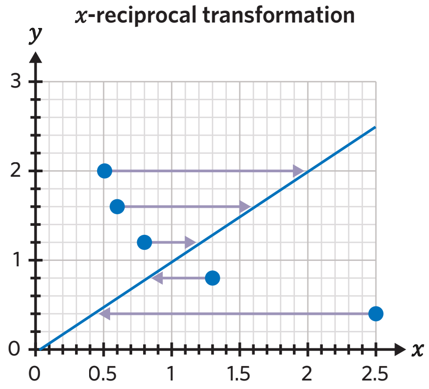


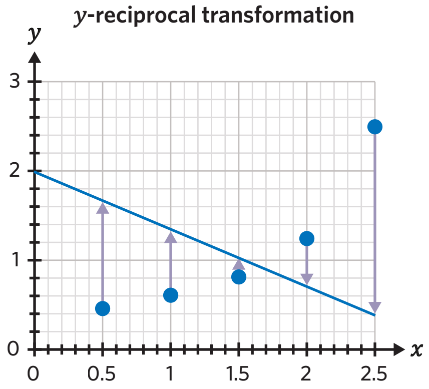
Applying a reciprocal transformation

A **reciprocal** of a number is the value found by dividing one by that number. The reciprocal of a number between zero and one becomes larger than the original number, whereas the reciprocal of a number greater than one is smaller than the original number.

The ***x*-reciprocal transformation** involves 'compressing' *x* values that are greater than one, whilst 'stretching' *x* values that are less than one. Values closer to one are compressed/stretched less than numbers that are further away. The *y* values remain the same.

The ***y*-reciprocal transformation** involves 'compressing' *y* values that are greater than one, whilst 'stretching' *y* values that are less than one. Values closer to one are compressed/stretched less than numbers that are further away. The *x* values remain the same.


x-reciprocal transformation


y-reciprocal transformation

Worked example 4

A class of ten students recorded the time they spent studying for their psychology test. The number of questions they got incorrect are shown in the table.

hours studied	1	2	2.5	3	4	4	5	6	6.5	7
questions incorrect	63	32	23	17	16	14	8	7	5	5

Apply a *y*-reciprocal transformation $\left(\frac{1}{y}\right)$ and plot the transformed data.

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

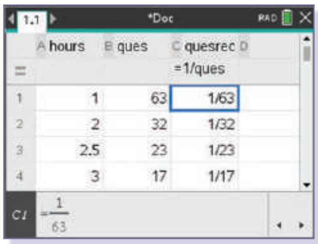
Step 2: Name column A 'hours' and column B 'ques'.

Enter the *hours studied* values into column A, starting from row 1.

Enter the *questions incorrect* values into column B, starting from row 1.

Step 3: Name column C 'quesrec' (short for *questions incorrect* reciprocal).

Enter '=1/ques' into the cell below the 'quesrec' heading.

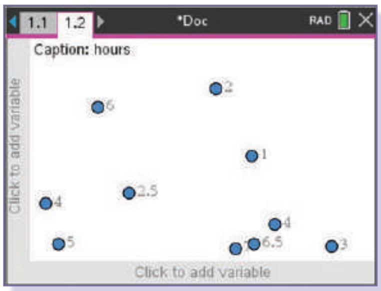


	A hours	B ques	C quesrec	D
=			=1/ques	
1	1	63	1/63	
2	2	32	1/32	
3	2.5	23	1/23	
4	3	17	1/17	

C1 $=\frac{1}{63}$

Continues →

Step 4: Press ctrl + doc ▾, and select '5: Add Data & Statistics'.



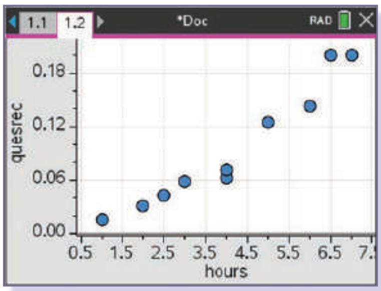
Step 5: Move the cursor to the horizontal axis and select 'Click to add variable'.

Select 'hours'.

Move the cursor to the vertical axis and select 'Click to add variable'.

Select 'quesrec'.

Answer



Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

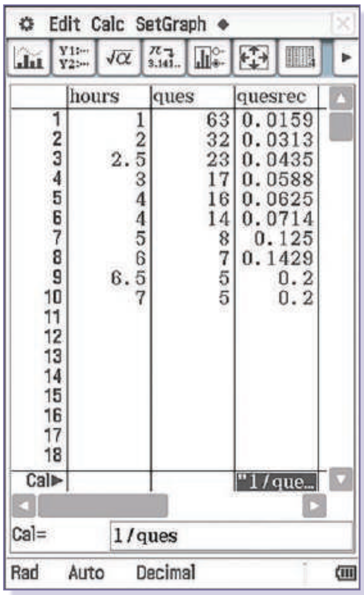
Step 2: Name the first list 'hours' and the second list 'ques'.

Enter the *hours studied* values into list 'hours', starting from row 1.

Enter the *questions incorrect* values into list 'ques', starting from row 1.

Step 3: Name the third list 'quesrec' (short for *questions incorrect* reciprocal).

In the third list, go down to the calculation cell Cal▶ and enter '1/ques'.



	hours	ques	quesrec
1	1	63	0.0159
2	2	32	0.0313
3	2.5	23	0.0435
4	3	17	0.0588
5	4	16	0.0625
6	4	14	0.0714
7	5	8	0.125
8	6	7	0.1429
9	6.5	5	0.2
10	7	5	0.2
11			
12			
13			
14			
15			
16			
17			
18			
Cal▶			"1/que...

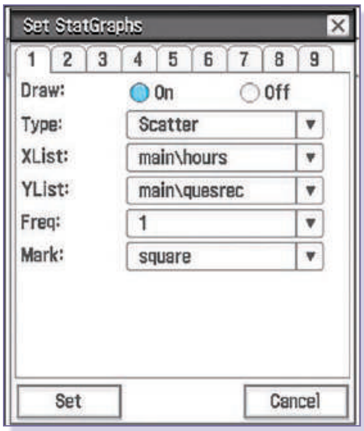
Cal= 1/ques

Rad Auto Decimal

Step 4: Configure the settings of the graph by tapping .

Create a scatterplot by changing 'Type' to 'Scatter'.

Specify the data set by changing 'XList:' to 'main\hours' and 'YList:' to 'main\quesrec'.



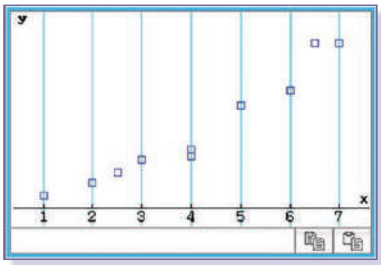
Draw: ● On ○ Off
Type: Scatter
XList: main\hours
YList: main\quesrec
Freq: 1
Mark: square

Tap 'Set' to confirm.

Step 5: Tap to plot the graph.

Continues →

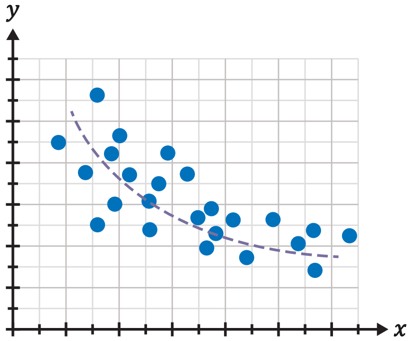
Answer



3D Questions

Choosing an appropriate data transformation

1. Which transformations could be used to linearise the data in the scatterplot?



A. log x, x-reciprocal and y-squared

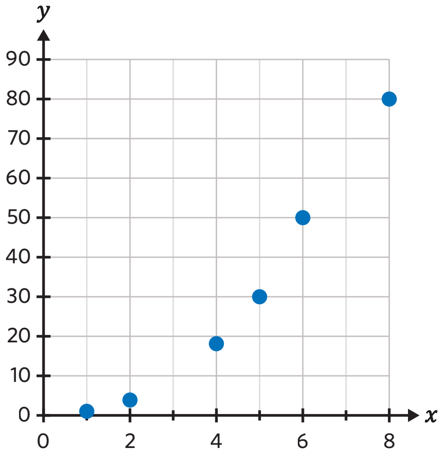
B. x-squared and y-squared

C. log x, x-reciprocal, log y and y-reciprocal

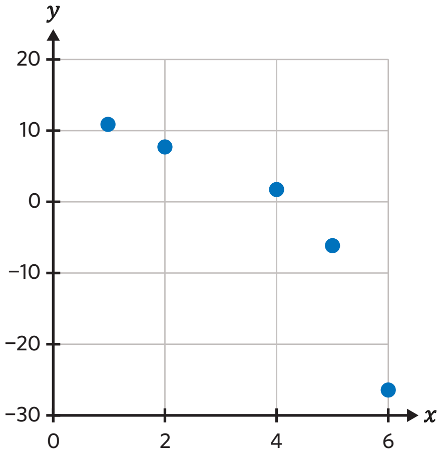
D. x-squared, log y and y-reciprocal

2. Which square transformation(s) can be used to linearise each data set?

a.

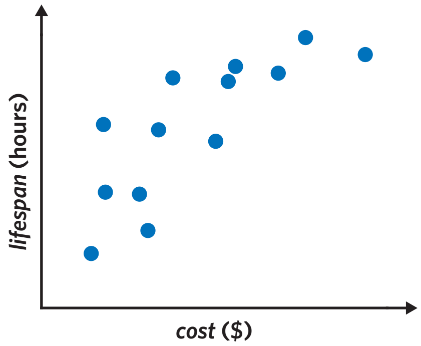


b.



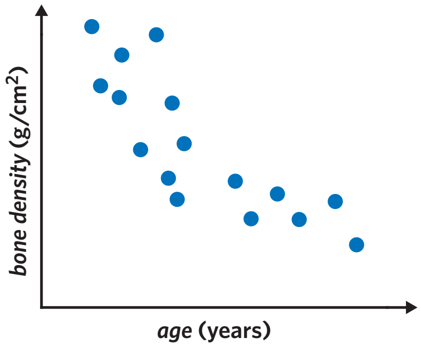
3. Determine the transformations that could be used to linearise the data in each of the following scatterplots.

a.



A. $lifespan^2$ **B.** $\log_{10}(cost)$ **C.** $\frac{1}{cost}$ **D.** All of the above

b.



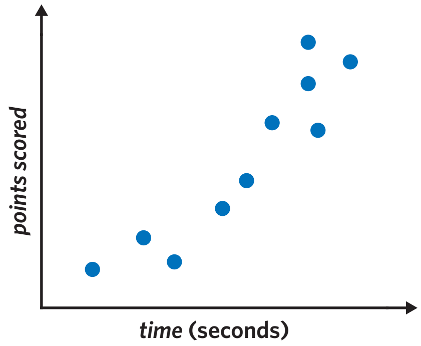
A. age^2 and $(bone\ density)^2$

B. $\frac{1}{bone\ density}$, $\log_{10}(bone\ density)$, $\frac{1}{age}$ and $\log_{10}(age)$

C. age^2, $\log_{10}(bone\ density)$ and $\frac{1}{age}$

D. $(bone\ density)^2$, $\log_{10}(age)$ and $\frac{1}{bone\ density}$

c.



A. $time^2$ **B.** $\log_{10}(points\ scored)$ **C.** $\frac{1}{points\ scored}$ **D.** All of the above

Applying a squared transformation

4. An x-squared transformation was applied to a data set, but one value is missing.

x	1	3	6	8	12
y	3	12	32	81	132
x^2	1	9	36		144

What is the missing value?

A. 8 **B.** 9 **C.** 64 **D.** 81

5. Consider the following table.

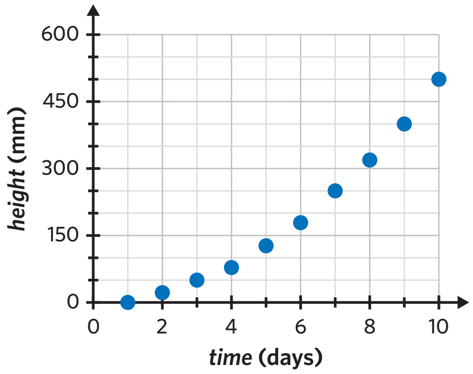
x	1	2	3	4	5
y	1.44	1.36	1.30	1.20	1.00

a. Apply an x-squared transformation to the data in the table by hand, and plot the transformed data.

b. Use a calculator to apply a y-squared transformation to the data in the table, and plot the transformed data.

6. One keen young science student monitored the growth of a plant over 10 days. He recorded the results and constructed a scatterplot from the data shown in the table.

time **(days)**	1	2	3	4	5	6	7	8	9	10
height **(mm)**	7	20	50	80	130	180	250	320	400	500



Use a calculator to apply a squared transformation to the variable *time*, and plot the transformed data.

Applying a log transformation

7. A log y transformation was applied to a data set, but one value is missing.

x	10	20	30	40	50
y	10	84	732	5045	38 720
log y	1	1.92	2.86	3.70	

What is the missing value?

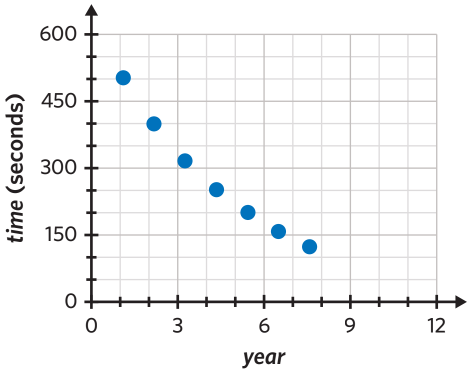
A. 4.59 **B.** 5.34 **C.** 7.81 **D.** 10.56

8. Use a calculator to apply a log x transformation to the data in the table, and plot the transformed data.

x	1	2	3	4	5	6	7
y	24	16	10	7	4	3	2

9. Ed has participated in an annual cheese rolling competition for the past seven years. The *time* taken for him to reach the bottom of the hill was recorded each *year* and the results are shown in the table and scatterplot.

year	1	2	3	4	5	6	7
time **(seconds)**	501	398	316	251	200	159	126



A log transformation can be applied to the variable *time* to linearise the scatterplot. Use a calculator to apply the transformation and plot the transformed data.

Applying a reciprocal transformation

10. An x-reciprocal transformation was applied to a data set, but one value is missing.

x	12	16	20	24	28
y	43	94	111	118	120
$1/x$	0.0833	0.0625	0.0500		0.0357

What is the missing value?

A. 0.0042 **B.** 0.0085 **C.** 0.0406 **D.** 0.0417

11. Consider the following table.

x	2	4	6	8	10
y	0.53	0.27	0.16	0.13	0.10

a. Use a calculator to apply an x-reciprocal transformation to the data in the table, and plot the transformed data.

b. Use the same sheet to apply a y-reciprocal transformation to the data in the table, and plot the transformed data.

12. Donald Trump started his career with a 'small loan' of $1 million dollars. The following table shows his *net worth* at the start of each *year* for his first ten years in the business world.

year	1	2	3	4	5	6	7	8	9	10
net worth **($mil)**	1.0	1.2	1.5	1.7	2.2	2.6	4.0	5.6	10.4	31.7

Use a calculator to apply a reciprocal transformation to the variable *net worth*, and plot the transformed data.

Joining it all together

13. Danny, an Olympic discus thrower, recorded the distance of 15 of his throws at a training session in the following table and scatterplot. The relationship between the *throw number* and *distance* (m) is non-linear.

throw number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
distance **(m)**	14.7	15.4	14.9	16.3	18.9	19.6	21.7	21.5	24.7	26.8	31.6	32.4	35.9	43.6	65.2

a. Which transformation **cannot** linearise the scatterplot if applied?

A. x-reciprocal transformation

B. y-reciprocal transformation

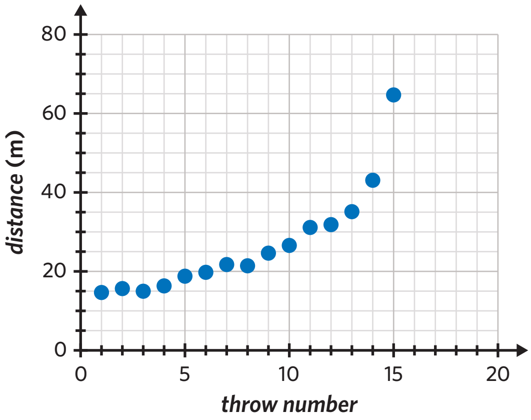
C. x-squared transformation

D. log y transformation

b. Use a calculator to apply a squared transformation to the variable *throw number*, and plot the transformed data.

c. Use a calculator to apply a log transformation to the variable *distance*, and plot the transformed data.

d. Use a calculator to apply a reciprocal transformation to the variable *distance*, and plot the transformed data.



Exam practice

14. The *mathematics achievement* level (TIMSS score) for grade 8 students and the general rate of *internet use* (%) for 10 countries are displayed in the following scatterplot.

To linearise the data, it would be best to plot

A. *mathematics achievement* against *internet use*.

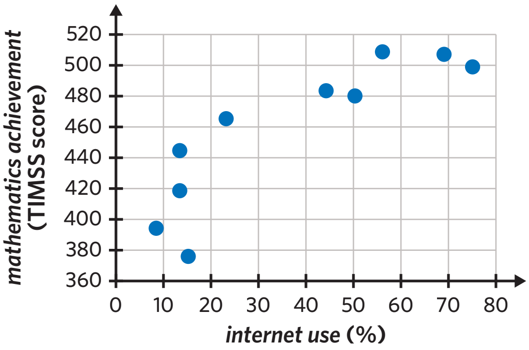
B. $\log_{10}$(*mathematics achievement*) against *internet use*.

C. *mathematics achievement* against $\log_{10}$(*internet use*).

D. *mathematics achievement* against (*internet use*)^2.

E. $\dfrac{1}{\textit{mathematics achievement}}$ against *internet use*.

VCAA 2009 Exam 1 Data analysis Q12

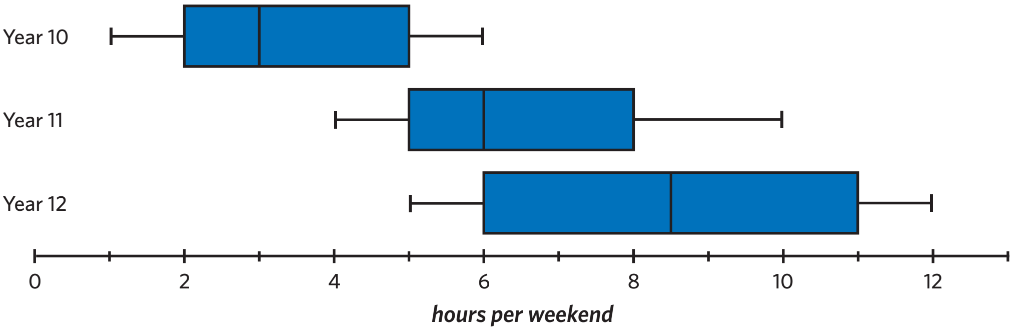


59% of students answered this question correctly.

Questions from multiple lessons

Data analysis *Year 11 content*

15. The following parallel boxplots display the distribution of hours spent studying for three groups of students (Year 10s, Year 11s and Year 12s).



Which one of the following statements is **not** true?

A. Year 12s have the most variable hours spent studying.

B. 75% of Year 10s study less than or equal to all Year 12s.

C. More than 50% of Year 10s study less than all of the other students.

D. In terms of the median, Year 10 students spend the least amount of hours studying.

E. 50% of Year 10s study the same amount as 75% of Year 11s.

Adapted from VCAA 2018 Exam 1 Data analysis Q6

Recursion and financial modelling *Year 11 content*

16. Chris plans to complete his collection of game tokens in four years.

Each year, he collects 120 more game tokens than the previous year.

There are a total of 2000 game tokens to collect.

If he is going to finish collecting in four years, the number of game tokens he needs to collect in the first year is

A. 200 **B.** 230 **C.** 320 **D.** 760 **E.** 1320

Adapted from VCAA 2014 Exam 1 Number patterns Q5

Data analysis

17. Eddie's monthly salary of $2500 during winter was divided according to the type of *expenditure* (food, drinks, clothes, savings).

The percentage of his salary divisions were calculated and are displayed in the following table.

expenditure	**%**
drinks	16
food	23
clothes	11
savings	50
total	100

a. How much of his salary, in dollars, was spent on clothes? (1 MARK)

b. Use the percentages to construct a percentage segmented bar chart. Use a key to indicate the segment of the bar chart that corresponds to each type of *expenditure*. (3 MARKS)

In order to investigate a possible association between his spending habits in summer and winter, Eddie's salary is divided according to the type of *expenditure*, and the season in which it was earned.

	season	
expenditure	**winter**	**summer**
food	23	16
drinks	16	31
clothes	11	29
savings	50	24
total	100	100

c. Does the information support the theory that *expenditure* is associated with *season*? Justify your answer by quoting appropriate percentages. Note that it is sufficient to quote only one type of *expenditure* in your answer. (2 MARKS)

Adapted from VCAA 2018NH Exam 2 Data analysis Q4

3E Data transformations – applications

STUDY DESIGN DOT POINT

- interpretation and use of the equation of the least squares line fitted to the transformed data to make predictions

3A 3B 3C 3D 3E

KEY SKILLS

During this lesson, you will be:

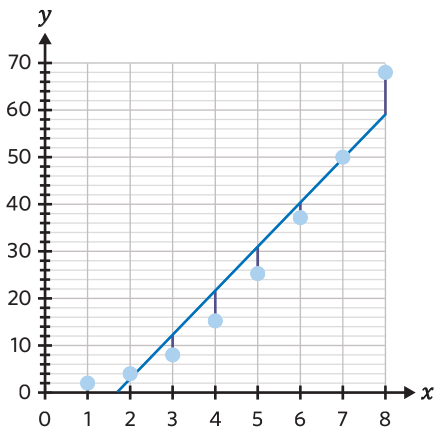
- calculating the equation of the least squares regression line for transformed data
- making predictions using the regression equation of transformed data.

Least squares regression lines are most accurate when fitted to linear data. This means that when data is non-linear, a transformation should be applied to a variable before fitting the least squares regression line. A least squares regression line that is fitted to transformed data can be used to make predictions about non-linear relationships.

Calculating the equation of the least squares regression line for transformed data

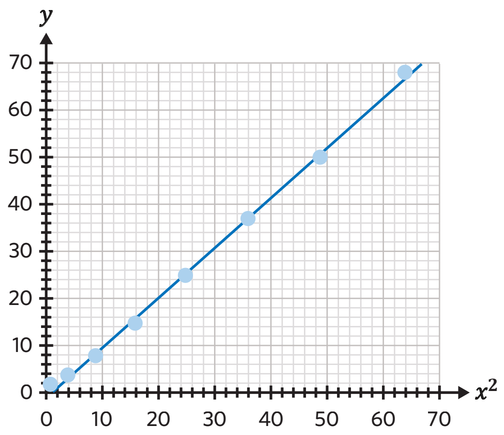
The process of fitting a least squares regression line to transformed data is no different to when fitting to linear data, except that one of the original variables is replaced by a transformed variable in the least squares regression equation. Appropriately transforming data before fitting a least squares regression line will improve accuracy for non-linear relationships.

Non-linear data



$y = -16.14 + 9.39x$

Linearised data (x^2 transformation)



$y = -0.73 + 1.05x^2$

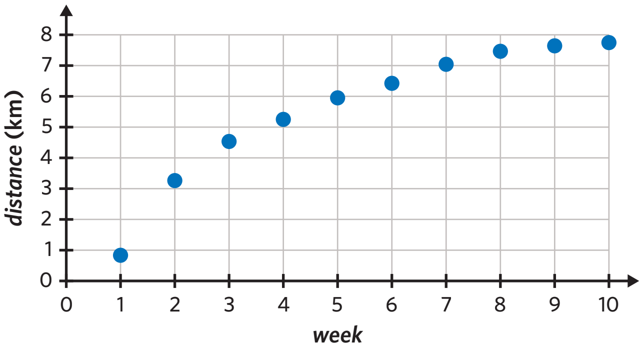
The residuals are smaller for the least squares regression line fitted on the linearised data compared to the non-linear data. This means that the least squares regression line is more accurate after the x^2 transformation.

As there is more than one way to linearise a data set, it is important to be able to determine which transformation is best. This can be achieved by determining the transformation with the largest coefficient of determination, or r^2, value.

Worked example 1

Akin is training for an Ironman Triathlon. Each Sunday, for ten weeks, he woke up early and ran ten kilometres. He recorded the *distance*, in kilometres, he ran before his first break. The results are shown in the following table and scatterplot.

week	1	2	3	4	5	6	7	8	9	10
***distance* (km)**	0.8	3.2	4.5	5.2	5.9	6.4	7	7.4	7.6	7.7



a. Apply a squared transformation to the variable *distance* and calculate the equation of the least squares regression line for the transformed data. Round the values of the intercept and slope to three decimal places.

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select ‘1: New’ → ‘4: Add Lists & Spreadsheet’.

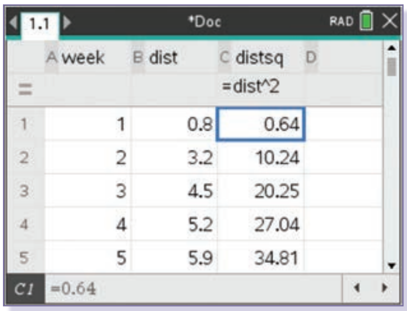
Step 2: Name column A ‘week’ and column B ‘dist’.

Enter the *week* values into column A, starting from row 1.

Enter the *distance* values into column B, starting from row 1.

Step 3: Name column C ‘distsq’ (short for *distance* squared).

Enter ‘=dist^2’ into the cell below the ‘distsq’ heading.



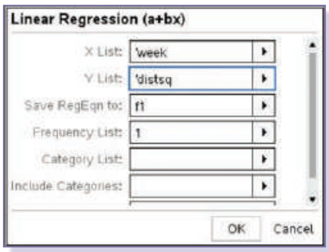
	A week	B dist	C distsq	D
=			=dist^2	
1	1	0.8	0.64	
2	2	3.2	10.24	
3	3	4.5	20.25	
4	4	5.2	27.04	
5	5	5.9	34.81	

C1 =0.64

Select ‘OK’.

Step 4: Press menu and select ‘4: Statistics’ → ‘1: Stat Calculations’ → ‘4: Linear Regression (a+bx)’.

Select ‘week’ in ‘X List:’ and ‘distsq’ in ‘Y List:’



Linear Regression (a+bx)

X List: 'week

Y List: 'distsq

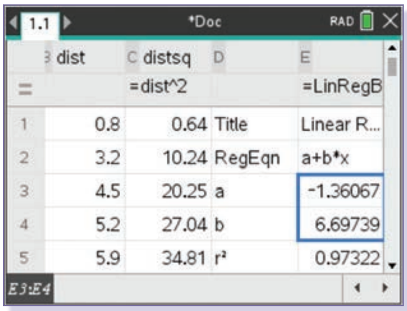
Save RegEqn to: f1

Frequency List: 1

Category List:

Include Categories:

Step 5: Round the values of *a* and *b* to three decimal places.



	B dist	C distsq	D	E
=		=dist^2		=LinRegB
1	0.8	0.64	Title	Linear R...
2	3.2	10.24	RegEqn	a+b*x
3	4.5	20.25	a	-1.36067
4	5.2	27.04	b	6.69739
5	5.9	34.81	r²	0.97322

E3:E4

$a = -1.361$

$b = 6.697$

Step 6: Write the equation in terms of the variables in the question.

The explanatory variable is *week*.

The response variable is $(distance)^2$.

Continues →

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

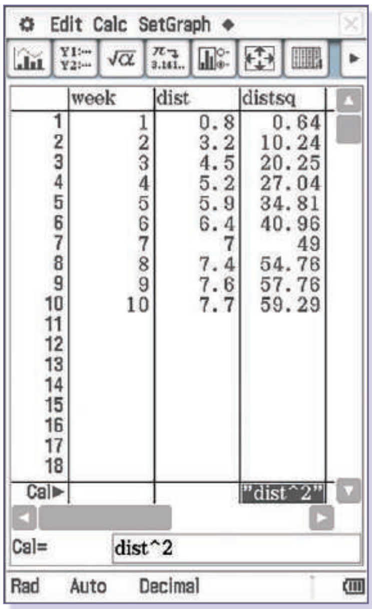
Step 2: Name the first list 'week' and the second list 'dist'.

Enter the *week* values into list 'week', starting from row 1.

Enter the *distance* values into list 'dist', starting from row 1.

Step 3: Name the third list 'distsq' (short for *distance* squared).

In the third list, go down to the calculation cell Cal▶ and enter 'dist^2'.

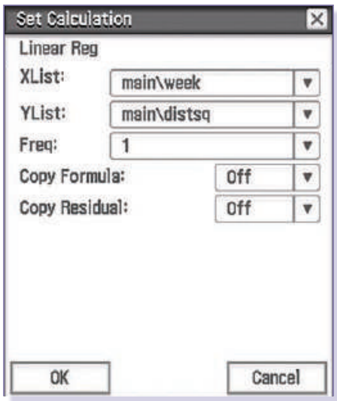


	week	dist	distsq
1	1	0.8	0.64
2	2	3.2	10.24
3	3	4.5	20.25
4	4	5.2	27.04
5	5	5.9	34.81
6	6	6.4	40.96
7	7	7	49
8	8	7.4	54.76
9	9	7.6	57.76
10	10	7.7	59.29
11			
12			
13			
14			
15			
16			
17			
18			
Cal▶			"dist^2"

Cal= dist^2

Step 4: Tap 'Calc' → 'Regression' → 'Linear Reg'.

Specify the data set by changing 'XList:' to 'main\week' and 'YList:' to 'main\distsq'.

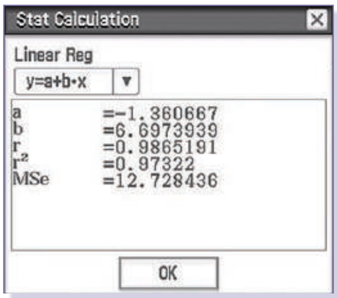


Linear Reg
XList: main\week
YList: main\distsq
Freq: 1
Copy Formula: Off
Copy Residual: Off

Tap 'OK' to confirm.

Step 5: Round the values of a and b to three decimal places.

Change the form of the regression equation to $y = a + bx$.



Linear Reg
y=a+b·x

```
a     =-1.360667
b     =6.6973939
r     =0.9865191
r²    =0.97322
MSe   =12.728436
```

$a = -1.361$

$b = 6.697$

Step 6: Write the equation in terms of the variables in the question.

The explanatory variable is *week*.

The response variable is $(distance)^2$.

Answer – Method 1 and 2

$(distance)^2 = -1.361 + 6.697 \times week$

b. Apply a log transformation to the variable *week* and calculate the equation of the least squares regression line for the transformed data. Round the values of the intercept and slope to three decimal places.

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Name column A 'week' and column B 'dist'.

Enter the *week* values into column A, starting from row 1.

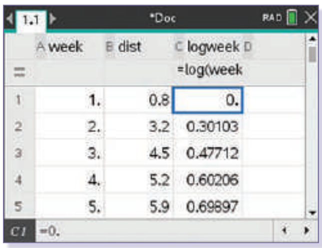
Enter the *distance* values into column B, starting from row 1.

Note: The data can be reused from part **a**.

Continues →

Step 3: Name column C 'logweek'.

Enter '=log(week)' into the cell below the 'logweek' heading.

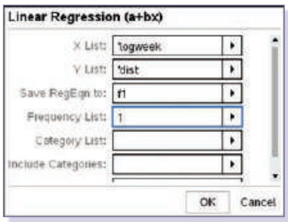


	A week	B dist	C logweek	D
=			=log(week	
1	1.	0.8	0.	
2	2.	3.2	0.30103	
3	3.	4.5	0.47712	
4	4.	5.2	0.60206	
5	5.	5.9	0.69897	

C1 =0.

Step 4: Press menu and select '4: Statistics' → '1: Stat Calculations' → '4: Linear Regression (a+bx)'.

Select 'logweek' in 'X List:' and 'dist' in 'Y List:'



Linear Regression (a+bx)

X List: logweek

Y List: dist

Save RegEqn to: f1

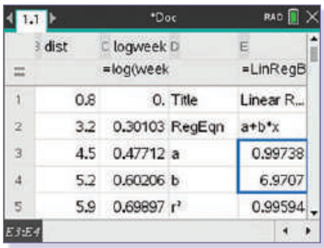
Frequency List: 1

Category List:

Include Categories:

Select 'OK'.

Step 5: Round the values of a and b to three decimal places.



	dist	C logweek	D	E
=		=log(week		=LinRegB
1	0.8	0.	Title	Linear R...
2	3.2	0.30103	RegEqn	a+b*x
3	4.5	0.47712	a	0.99738
4	5.2	0.60206	b	6.9707
5	5.9	0.69897	r²	0.99594

E3:E4

$a = 0.997$

$b = 6.971$

Step 6: Write the equation in terms of the variables in the question.

The explanatory variable is $\log_{10}(week)$.

The response variable is *distance*.

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Step 2: Name the first list 'week' and the second list 'dist'.

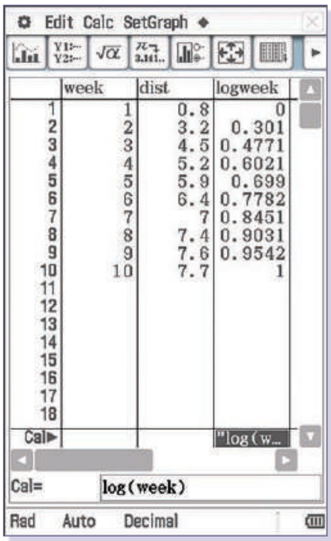
Enter the *week* values into list 'week', starting from row 1.

Enter the *distance* values into list 'dist', starting from row 1.

Note: The data can be reused from part a.

Step 3: Name the third list 'logweek'.

In the third list, go down to the calculation cell Cal▶ and enter 'log(week)'.



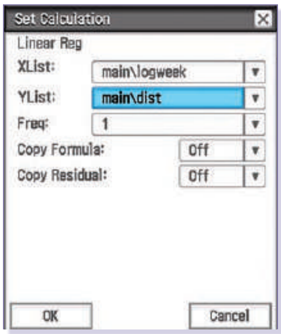
	week	dist	logweek
1	1	0.8	0
2	2	3.2	0.301
3	3	4.5	0.4771
4	4	5.2	0.6021
5	5	5.9	0.699
6	6	6.4	0.7782
7	7	7	0.8451
8	8	7.4	0.9031
9	9	7.6	0.9542
10	10	7.7	1
11			
12			
13			
14			
15			
16			
17			
18			
Cal▶			"log(w...

Cal= log(week)

Rad Auto Decimal

Step 4: Tap 'Calc' → 'Regression' → 'Linear Reg'.

Specify the data set by changing 'XList:' to 'main\logweek' and 'YList:' to 'main\dist'.



Set Calculation

Linear Reg

XList: main\logweek

YList: main\dist

Freq: 1

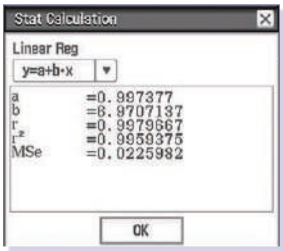
Copy Formula: Off

Copy Residual: Off

Tap 'OK' to confirm.

Step 5: Round the values of a and b to three decimal places.

Change the form of the regression equation to $y = a + bx$.



Stat Calculation

Linear Reg

y=a+b·x

```
a     =0.997377
b     =6.9707137
r     =0.9979667
r²    =0.9959375
MSe   =0.0225982
```

$a = 0.997$
$b = 6.971$

Continues →

Step 6: Write the equation in terms of the variables in the question.

The explanatory variable is $\log_{10}(week)$.

The response variable is *distance*.

Answer – Method 1 and 2

$distance = 0.997 + 6.971 \times \log_{10}(week)$

c. Apply a reciprocal transformation to the variable *week* and calculate the equation of the least squares regression line for the transformed data. Round the values of the intercept and slope to three decimal places.

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Name column A 'week' and column B 'dist'.

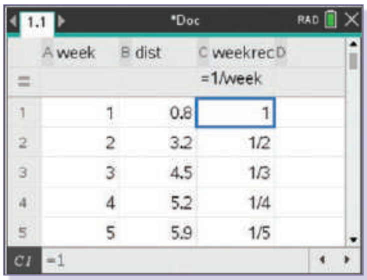
Enter the *week* values into column A, starting from row 1.

Enter the *distance* values into column B, starting from row 1.

Note: The data can be reused from part a.

Step 3: Name column C 'weekrec' (short for *week* reciprocal).

Enter '=1/week' into the cell below the 'weekrec' heading.

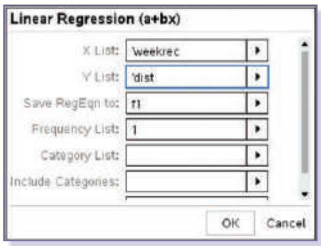


	A week	B dist	C weekrec	D
=			=1/week	
1	1	0.8	1	
2	2	3.2	1/2	
3	3	4.5	1/3	
4	4	5.2	1/4	
5	5	5.9	1/5	

C1 =1

Step 4: Press menu and select '4: Statistics' → '1: Stat Calculations' → '4: Linear Regression (a+bx)'.

Select 'weekrec' in 'X List:' and 'dist' in 'Y List:'

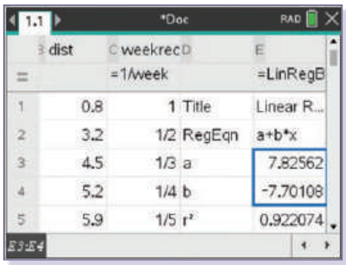


Linear Regression (a+bx)

X List: weekrec
Y List: 'dist
Save RegEqn to: f1
Frequency List: 1
Category List:
Include Categories:

Select 'OK'.

Step 5: Write down the equation for the least squares regression line as displayed on the screen, and round the values of *a* and *b* to three decimal places.



	B dist	C weekrec	D	E
=		=1/week		=LinRegB
1	0.8	1	Title	Linear R...
2	3.2	1/2	RegEqn	a+b*x
3	4.5	1/3	a	7.82562
4	5.2	1/4	b	-7.70108
5	5.9	1/5	r²	0.922074

E3:E4

$a = 7.826$

$b = -7.701$

Step 6: Write the equation in terms of the variables in the question.

The explanatory variable is $\frac{1}{week}$.

The response variable is *distance*.

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Step 2: Name the first list 'week' and the second list 'dist'.

Enter the *week* values into list 'week', starting from row 1.

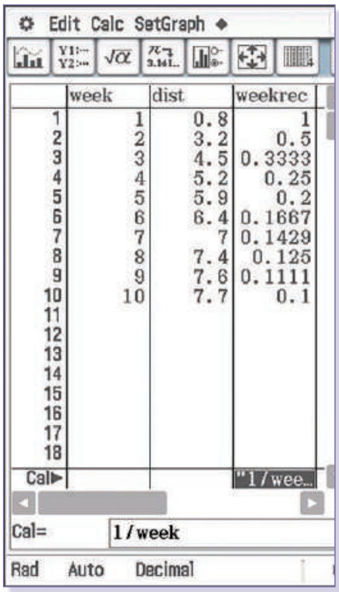
Enter the *distance* values into list 'dist', starting from row 1.

Note: The data can be reused from part **a**.

Continues →

Step 3: Name the third list 'weekrec' (short for *week* reciprocal).

In the third list, go down to the calculation cell Cal▶ and enter '1/week'.



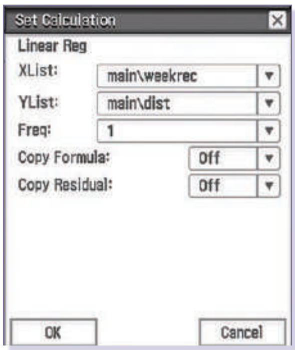
	week	dist	weekrec
1	1	0.8	1
2	2	3.2	0.5
3	3	4.5	0.3333
4	4	5.2	0.25
5	5	5.9	0.2
6	6	6.4	0.1667
7	7	7	0.1429
8	8	7.4	0.125
9	9	7.6	0.1111
10	10	7.7	0.1
11			
12			
13			
14			
15			
16			
17			
18			
Cal▶			"1/wee...

Cal= 1/week

Rad Auto Decimal

Step 4: Tap 'Calc' → 'Regression' → 'Linear Reg'.

Specify the data set by changing 'XList:' to 'main\weekrec' and 'YList:' to 'main\dist'.


Set Calculation

Linear Reg

XList: main\weekrec

YList: main\dist

Freq: 1

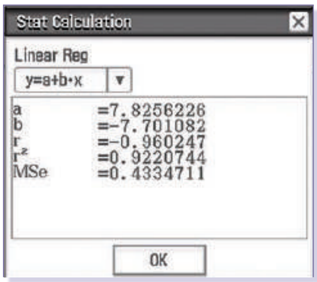
Copy Formula: Off

Copy Residual: Off

Tap 'OK' to confirm.

Step 5: Round the values of a and b to three decimal places.

Change the form of the regression equation to $y = a + bx$.


Stat Calculation

Linear Reg

y=a+b·x

```
a     =7.8256226
b     =-7.701082
r     =-0.960247
r²    =0.9220744
MSe   =0.4334711
```

$a = 7.826$

$b = -7.701$

Step 6: Write the equation in terms of the variables in the question.

The explanatory variable is $\frac{1}{week}$.

The response variable is *distance*.

Answer – Method 1 and 2

$$distance = 7.826 - 7.701 \times \frac{1}{week}$$

d. Determine which transformation is the best fit for the data set.

Explanation

Step 1: Determine the r^2 value for the least squares regression line of each transformation.

This can be done using the data entered in parts **a**–**c**.

Remember that the r^2 value can be located on the same screen as the a and b values.

$(distance)^2 = -1.361 + 6.697 \times week \qquad r^2 = 0.9732$

$distance = 0.997 + 6.971 \times \log_{10}(week) \qquad r^2 = 0.9959$

$distance = 7.826 - 7.701 \times \frac{1}{week} \qquad r^2 = 0.9221$

Step 2: Identify the largest value of r^2.

The largest value of r^2 is 0.9959.

Therefore, the log transformation on the variable *week* has the best fit.

Continues →

Answer

$distance = 0.997 + 6.971 \times \log_{10}(week)$

Making predictions using the regression equation of transformed data

A least squares regression line on transformed data can still be used to interpolate and extrapolate. However, when making predictions it is important to consider the transformation that has been applied to one of the variables.

The limitations of extrapolation are also present with a least squares regression line fitted to a transformed data set. When extrapolating, it is assumed that the shape of the relationship between the variables will continue outside of the range of the data set. This assumption has limited reliability.

Worked example 2

Akin recorded the *distance*, in kilometres, he ran before his first break for the last ten weeks, whilst training for an Ironman Triathlon. As the data he recorded was non-linear, a transformation had to be applied to one of the variables before estimating the equation of the least squares regression line. Use the following regression equations to predict how many kilometres Akin will be able to run before his first break in his 20th week of training. Give answers correct to three decimal places.

a. $(distance)^2 = -1.361 + 6.697 \times week$

Explanation

Step 1: Substitute the known value into the regression equation.

Akin wants to predict the value of *distance* in the 20th week, so let $week = 20$.

Step 2: Solve for the unknown value.

$$(distance)^2 = -1.361 + 6.697 \times 20$$

$$(distance)^2 = 132.579$$

$$distance = 11.514$$

Answer

11.514 km

b. $distance = 0.997 + 6.971 \times \log_{10}(week)$

Explanation

Step 1: Substitute the known value into the regression equation.

Akin wants to predict the value of *distance* in the 20th week, so let $week = 20$.

Step 2: Solve for the unknown value.

$$distance = 0.997 + 6.971 \times \log_{10}(20)$$

$$distance = 10.066$$

Answer

10.066 km

Continues →

c. $distance = 7.826 - 7.701 \times \frac{1}{week}$

Explanation

Step 1: Substitute the known value into the regression equation.

Akin wants to predict the value of *distance* in the 20^{th} week, so let $week = 20$.

Step 2: Solve for the unknown value.

$distance = 7.826 - 7.701 \times \frac{1}{20}$

$distance = 7.441$

Answer

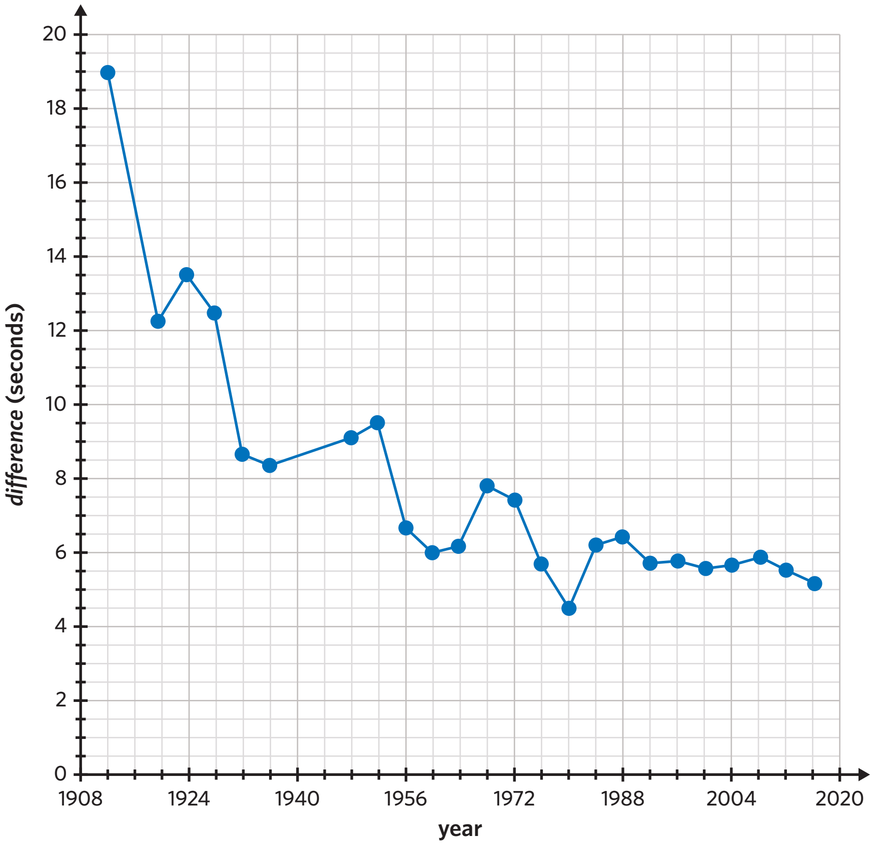
7.441 km

Exam question breakdown

VCAA 2021 Exam 2 Data analysis Q5a

A method for predicting future time differences in the 100 m freestyle swim is to use the formula $difference = winning\ time\ women - winning\ time\ men$

The resulting data and time series plot are shown. The plot is clearly non-linear.



year	*difference* (seconds)
1912	18.8
1920	12.2
1924	13.4
1928	12.4
1932	8.6
1936	8.3
1948	9.0
1952	9.4
1956	6.6
1960	6.0
1964	6.1
1968	7.8
1972	7.4
1976	5.7
1980	4.4
1984	6.1
1988	6.3
1992	5.6
1996	5.8
2000	5.5
2004	5.7
2008	5.9
2012	5.5
2016	5.1

Note: No Olympic Games were held in 1916, 1940 and 1944.

Apply a reciprocal transformation to the variable *difference* to linearise the data. Fit a least squares line to the transformed data and write its equation.

Round the values of the intercept and the slope to four significant figures. (2 MARKS)

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Name column A 'year' and column B 'dif'.

Enter the *year* values into column A, starting from row 1.

Enter the *difference* values into column B, starting from row 1.

Continues →

Step 3: Name column C 'difrec' (short for *difference* reciprocal).

Enter '=1/diff' into the cell below the 'difrec' heading.

Step 4: Press menu and select '4: Statistics' → '1: Stat Calculations' → '4: Linear Regression (a+bx)'.

Select 'year' in 'X List:' and 'difrec' in 'Y List:'

Select 'OK'.

Step 5: Round the values of a and b to four significant figures.

$a = -2.234$

$b = 0.001209$

Step 6: Write the equation in terms of the variables in the question.

The explanatory variable is *year*.

The response variable is $\frac{1}{difference}$.

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Step 2: Name the first list 'year' and the second list 'dif'.

Enter the *year* values into list 'year', starting from row 1.

Enter the *difference* values into list 'dif', starting from row 1.

Step 3: Name the third list 'difrec' (short for *difference* reciprocal).

In the third list, go down to the calculation cell Cal▶ and enter '1/dif'.

Step 4: Tap 'Calc' → 'Regression' → 'Linear Reg'.

Specify the data set by changing 'XList:' to 'main\year' and 'YList:' to 'main\difrec'.

Tap 'OK' to confirm.

Step 5: Round the values of a and b to four significant figures.

$a = -2.234$

$b = 0.001209$

Step 6: Write the equation in terms of the variables in the question.

The explanatory variable is *year*.

The response variable is $\frac{1}{difference}$.

Answer – Method 1 and 2

$$\frac{1}{difference} = -2.234 + 0.001209 \times year$$

The average mark on this question was **0.5**.

Few students received full marks on this question. Students who calculated the regression equation correctly often wrote the variables incorrectly, or rounded the slope and intercept values incorrectly.

3E Questions

Calculating the equation of the least squares regression line for transformed data

1. An x-squared transformation was applied to the data in the following table, and a least squares regression line was fitted.

x	5	9	14	20	24
y	22	23	25	30	35

The equation of the least squares regression line is closest to

A. $y = 17.269 + 0.676x^2$

B. $y = 20.975 + 0.024x^2$

C. $y = 206.701 + 37.910x^2$

D. $y = -880.917 + 43.093x^2$

2. Apply a $\log_{10}y$ transformation to the data in the following table, and calculate the equation of the least squares regression line for the transformed data. Round the values of the intercept and slope to three decimal places.

x	2	4	5	8	10	14	15	17	19	20
y	32	34	38	53	67	102	118	134	168	182

3. Loki uploads soccer videos to Youtube daily. The number of new views on his latest video, on each day in the first week after uploading, are shown in the table.

day	1	2	3	4	5	6	7
views	321	113	54	36	28	24	21

Apply a reciprocal transformation to the variable *views*, and calculate the equation of the least squares regression line for the transformed data. Round the values of the intercept and slope to four decimal places.

4. Amanda competed in a 400 m swimming race. Her time for each lap of the 50-metre pool was recorded and the results are shown in the table.

lap	1	2	3	4	5	6	7	8
***time* (seconds)**	31	40	44	50	53	55	58	60

a. Apply a log transformation to the variable *lap* and calculate the equation of the least squares regression line for the transformed data. Round the values of the intercept and slope to one decimal place.

b. Apply a squared transformation to the variable *time* and calculate the equation of the least squares regression line for the transformed data. Round the values of the intercept and slope to one decimal place.

c. Which transformation is the best fit for the data set?

Making predictions using the regression equation of transformed data

5. An x-reciprocal transformation was used to linearise a set of non-linear bivariate data. A least squares line was then fitted to the transformed data. The equation of this least squares line is

$$y = 843 - 165 \times \frac{1}{x}$$

This equation is used to predict the value of y when $x = 5$.

The value of y is

A. 0

B. 18

C. 810

D. 876

6. The relationship between the *value* ($) and *age* (years) of an antique vase has been linearised using a log transformation. A least squares regression line is fitted to the transformed data and its equation is

$$value = 37\,000 + 13\,000 \times \log_{10}(age)$$

If the value of the vase is $49 400, its age is closest to

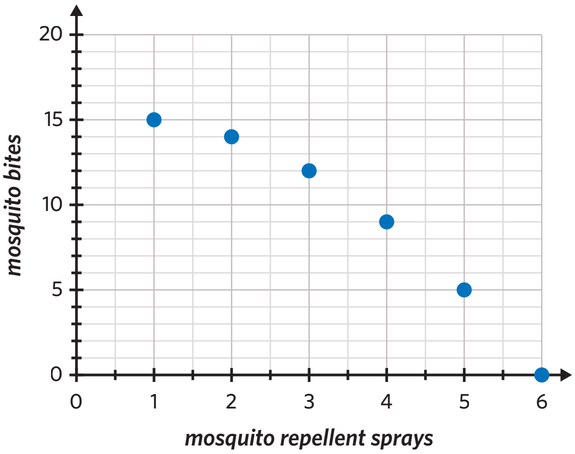
A. 1 year

B. 5 years

C. 8 years

D. 9 years

7. A group of medicine students went camping with a limited amount of mosquito repellent. The number of *mosquito repellent sprays* and the number of *mosquito bites* received for each student is shown in the scatterplot.



A squared transformation was applied to the variable *mosquito bites* to linearise the data. A least squares regression line was then fitted to the transformed data and its equation is

$(\textit{mosquito bites})^2 = 282 - 49 \times \textit{mosquito repellent sprays}$

a. Using the equation, estimate the number of *mosquito repellent sprays* a student would have applied if they received 6 mosquito bites. Round your answer to the nearest whole number.

b. Is the estimate in part **a** reliable? Justify your answer.

8. In an attempt to investigate claims that the sea levels are rising, a climate scientist decides to measure the increase in the depth of the sea just off a remote island every year, for ten years. Her results are shown in the following table.

year	1	2	3	4	5	6	7	8	9	10
***sea level rise* (cm)**	0.21	0.10	0.13	0.25	0.12	0.20	0.19	0.27	0.20	0.42

A reciprocal transformation was applied to the variable *sea level rise* to linearise the data. A least squares regression line was then fitted to the transformed data and its equation is

$$\frac{1}{\textit{sea level rise}} = 8.14 - 0.46 \times \textit{year}$$

a. Fill out the following table using the equation to predict the *sea level rise* in each of the next five years. Round each estimate to two decimal places.

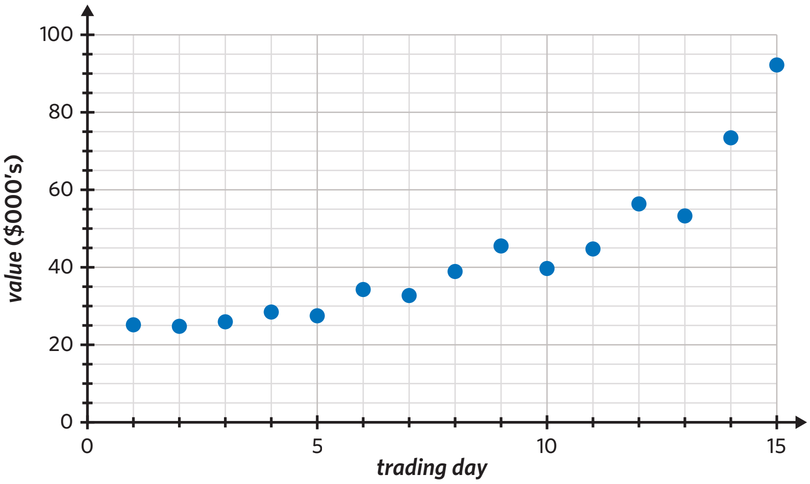
year	11	12	13	14	15
***estimated sea level rise* (cm)**					

b. Are the predictions in part **a** reliable? Justify your answer.

Joining it all together

9. Johnny is a day trader who buys and sells futures contracts. He tracks the *value* ($000's) of his portfolio at the end of each *trading day*, over three weeks. The data is shown in the following table and scatterplot.

trading day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
***value* ($000's)**	25.0	24.5	26.0	28.4	27.3	34.1	32.5	38.7	45.2	39.4	44.8	56.2	53.2	73.4	92.4



a. Apply all appropriate transformations for the data set. Give the least squares regression equation for each transformation, rounding the values of the intercept and slope to three decimal places.

b. Which transformation is the best fit for the data set?

c. Consider the least squares regression equation for the transformation chosen in part **b**.

i. Predict the *value*, in dollars, of Johnny's portfolio at the end of the 20th *trading day*. Round your answer to the nearest dollar.

ii. Predict the *trading day* in which the *value* of Johnny's portfolio first closes at more than $200 000.

iii. Estimate the *value* ($) of Johnny's portfolio at the end of the *trading day* 4 days before he started recording. Round your answer to the nearest dollar.

d. Which of the previous predictions made are reliable? Justify your answer.

Exam practice

10. In a study, the association between the *number of tasks* completed on a test and the *time* allowed for the test, in hours, was found to be non-linear.

The data can be linearised using a $\log_{10}$ transformation applied to the variable *number of tasks*.

The equation of the least squares line for the transformed data is

$\log_{10}(\textit{number of tasks}) = 1.160 + 0.03617 \times \textit{time}$

This equation predicts that the *number of tasks* completed when the *time* allowed for the test is three hours is closest to

A. 13

B. 16

C. 19

D. 25

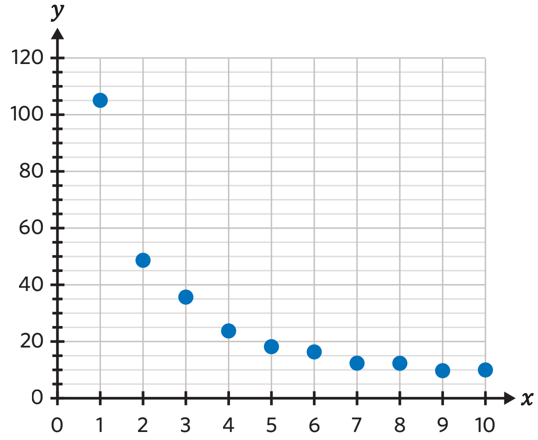
E. 26

VCAA 2020 Exam 1 Data analysis Q14

74% of students answered this question correctly.

11. Freya uses the data in the table to generate the following scatterplot.

x	1	2	3	4	5	6	7	8	9	10
y	105	48	35	23	18	16	12	12	9	9



The scatterplot shows that the data is non-linear.

To linearise the data, Freya applies a reciprocal transformation to the variable y.

She then fits a least squares line to the transformed data.

With x as the explanatory variable, the equation of this least squares line is closest to

A. $\frac{1}{y} = -0.0039 + 0.012x$

B. $\frac{1}{y} = -0.025 + 1.1x$

C. $\frac{1}{y} = 7.8 - 0.082x$

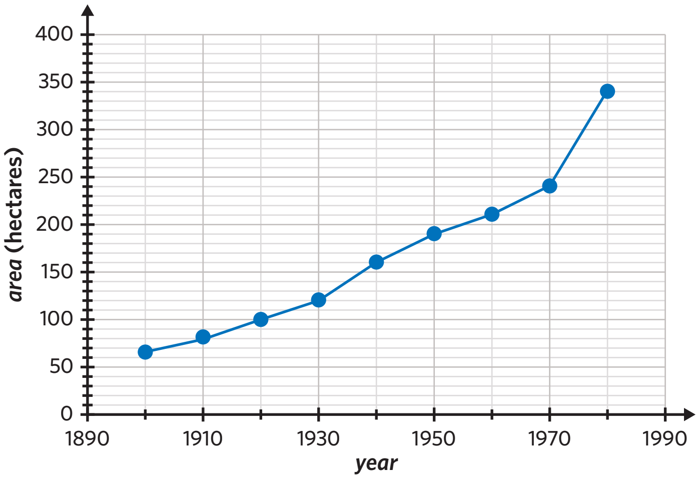
D. $y = 45.3 + 59.7 \times \frac{1}{x}$

E. $y = 59.7 + 45.3 \times \frac{1}{x}$

VCAA 2018 Exam 1 Data analysis Q11

58% of students answered this question correctly.

12. The following time series plot shows the total *area*, in hectares, of forest eaten by the caterpillars in a rural area during the period 1900 to 1980. The data used to generate this plot is also given.



year	1900	1910	1920	1930	1940	1950	1960	1970	1980
area (hectares)	66	80	100	120	160	190	210	240	340

The association between *area* of forest eaten by the caterpillars and *year* is non-linear.

A $\log_{10}$ transformation can be applied to the variable *area* to linearise the data.

a. Perform the $\log_{10}$ transformation to the variable *area* and determine the equation of the least squares line that can be used to predict $\log_{10}(area)$ from *year*.

Round the values of the intercept and slope to three significant figures. (2 MARKS)

b. The least squares line predicts that the $\log_{10}(area)$ of forest eaten by the caterpillars by the year 2020 will be approximately 2.85.

Using this value of 2.85, calculate the expected area of forest that will be eaten by the caterpillars by the year 2020.

Round your answer to the nearest hectare. (1 MARK)

c. Give a reason why this prediction may have limited reliability. (1 MARK)

VCAA 2017 Exam 2 Data analysis Q4b-cii

Part **a**: The average mark on this question was **0.9**.

Part **b**: **29%** of students answered this question correctly.

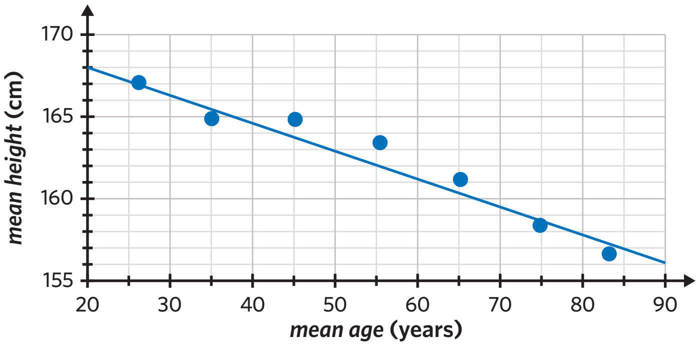
Part **c**: **53%** of students answered this question correctly.

13. The following table shows the *mean age*, in years, and the *mean height*, in centimetres, of 648 women from seven different age groups.

	age group						
	twenties	**thirties**	**fourties**	**fifties**	**sixties**	**seventies**	**eighties**
mean age (years)	26.3	35.2	45.2	55.3	65.1	74.8	83.1
mean height (cm)	167.1	164.9	164.8	163.4	161.2	158.4	156.7

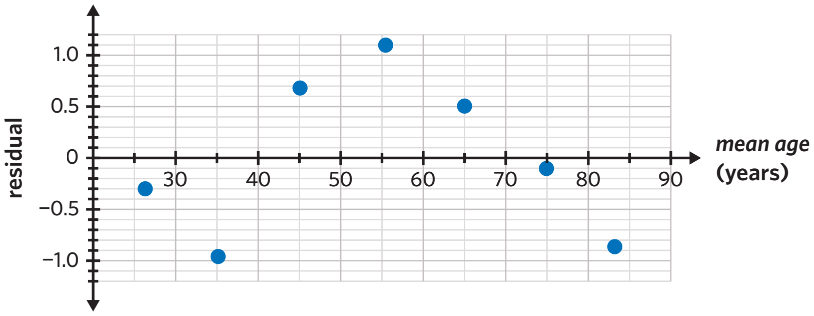
Data: J Sorkin et al., 'Longitudinal change in height of men and women: Implications for interpretation of the body mass index', American Journal of Epidemiology, vol. 150, no. 9, 1999, p. 971

A scatterplot displaying this data shows an association between the *mean height* and the *mean age* of these women. In an initial analysis of the data, a line is fitted to the data by eye, as shown.



In a further analysis of the data, a least squares line was fitted.

The associated residual plot that was generated is shown.



The residual plot indicates that the association between the *mean height* and the *mean age* of women is non-linear.

Apply an appropriate transformation to the variable *mean age* to linearise the data.
Fit a least squares line to the transformed data and write its equation.

Round the values of the intercept and the slope to four significant figures. (2 MARKS)

VCAA 2020 Exam 2 Data analysis Q6d

The average mark on this question was **0.6**.

Questions from multiple lessons

Data analysis

14. A least squares line is fitted to a set of bivariate data. If the response and explanatory variables are reversed on the set of axes, which of the following statistics would not change?

A. The slope of the least squares line

B. The residual values

C. The equation of the least squares line

D. The correlation coefficient

E. The y-intercept of the least squares line

Adapted from VCAA 2018 Exam 1 Data analysis Q14

Recursion and financial modelling *Year 11 content*

15. A sequence can be generated by the following recurrence relation.

$V_{n+1} = 1.5 \times V_n, \quad V_0 = 10$

Which of the following rules can be used to find the term in the sequence after n iterations of the rule?

A. $V_n = 1.5 \times 10\text{ n}$

B. $V_n = 1.5 + 10^n$

C. $V_n = 10 \times 1.5n$

D. $V_n = 10 \times 1.5^n$

E. $V_n = 10 + 1.5^n$

Adapted from VCAA 2013 Exam 1 Number patterns Q5

Data analysis

16. The following table shows the number of *words written* and *time spent*, in minutes, on eight English essays.

***time spent* (mins)**	93	144	58	220	195	138	87	104
words written	811	997	440	1543	1230	1114	690	758

a. A least squares regression line is fitted to the data and has a y-intercept of 159.058781.

Round this value to four significant figures. (1 MARK)

b. Use the values in the table to calculate the equation of the least squares regression line, where *time spent* is the explanatory variable. Round all values correct to four significant figures. (2 MARKS)

Adapted from VCAA 2018 Exam 2 Data analysis Q3c,d