3

Introduction to regression

3.1 Kick off with CAS

Lines of best fit with CAS

Least-squares regression allows us to fit a line of best fit to a scatterplot. We can then use this line of best fit to make predictions about the data.

1 Using CAS, plot a scatterplot of the following data set, which indicates the temperature (x) and the number of visitors at a popular beach (y).

x	21	26	33	24	35	16	22	30	39	34	22	19
y	95	154	212	141	173	40	104	193	177	238	131	75

2 If there appears to be a linear relationship between x and y, use CAS to add a least-squares regression line of best fit to the data set.

3 What does the line of best fit tell you about the relationship between the x- and y-values?

4 Use the line of best fit to predict y-values given the following x-values:

a 37 b 24 c 17.

Are there any limitations on the data points?

5 The line of best fit can be extended beyond the limits of the original data set. Would you feel comfortable making predictions outside of the scope of the original data set?

6 Use CAS to plot scatterplots of the following sets of data, and if there appears to be a linear relationship, plot a least-squares regression line of best fit:

a

x	62	74	59	77	91	104	79	85	55	74	90	83
y	108	83	127	90	62	55	86	70	141	92	59	77

b

x	2.2	1.7	0.4	2.6	–0.3	1.5	3.1	1.1	0.8	2.9	0.7	–0.8
y	45	39	22	50	9	33	66	34	21	56	27	6

c

x	40	66	38	55	47	61	34	49	53	69	43	58
y	89	112	93	90	75	106	101	77	86	120	81	99

Please refer to the Resources tab in the Prelims section of your **eBookPLUS** for a comprehensive step-by-step guide on how to use your CAS technology.

3.2 Response (dependent) and explanatory (independent) variables

studyon
Unit 3
AOS DA
Topic 6
Concept 1
Defining two variable data
Concept summary
Practice questions

A set of data involving two variables where one affects the other is called bivariate data. If the values of one variable 'respond' to the values of another variable, then the former variable is referred to as the response (dependent) variable. So an **explanatory (independent) variable** is a factor that influences the response (dependent) variable.

When a relationship between two sets of variables is being examined, it is important to know which one of the two variables responds to the other. Most often we can make a judgement about this, although sometimes it may not be possible.

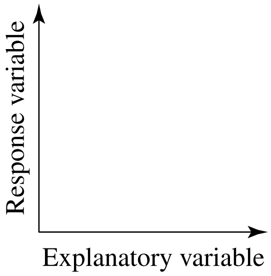
Consider the case where a study compared the heights of company employees against their annual salaries. Common sense would suggest that the height of a company employee would not respond to the person's annual salary nor would the annual salary of a company employee respond to the person's height. In this case, it is not appropriate to designate one variable as explanatory and one as response.

In the case where the ages of company employees are compared with their annual salaries, you might reasonably expect that the annual salary of an employee would depend on the person's age. In this case, the age of the employee is the explanatory variable and the salary of the employee is the response variable.

It is useful to identify the explanatory and response variables where possible, since it is the usual practice when displaying data on a graph to place the explanatory variable on the horizontal axis and the response variable on the vertical axis.

The explanatory variable is a factor that influences the response variable.

When displaying data on a graph place the explanatory variable on the horizontal axis and the response variable on the vertical axis.



WORKED EXAMPLE 1

For each of the following pairs of variables, identify the explanatory (independent) variable and the response (dependent) variable. If it is not possible to identify this, then write 'not appropriate'.

a **The number of visitors at a local swimming pool and the daily temperature**

b **The blood group of a person and his or her favourite TV channel**

THINK	WRITE
a It is reasonable to expect that the number of visitors at the swimming pool on any day will respond to the temperature on that day (and not the other way around).	**a** Daily temperature is the explanatory variable; number of visitors at a local swimming pool is the response variable.
b Common sense suggests that the blood type of a person does not respond to the person's TV channel preferences. Similarly, the choice of a TV channel does not respond to a person's blood type.	**b** Not appropriate

EXERCISE 3.2 Response (dependent) and explanatory (independent) variables

PRACTISE

1 WE1 For each of the following pairs of variables, identify the explanatory (independent) and the response (dependent) variable. If it is not possible to identify this, then write ‘not appropriate’.

a The number of air conditioners sold and the daily temperature
b The age of a person and their favourite colour

2 For each of the following pairs of variables, identify the explanatory (independent) and the response (dependent) variable. If it is not possible to identify the variables, then write ‘not appropriate’.

a The size of a crowd and the teams that are playing
b The net score of a round of golf and the golfer’s handicap

CONSOLIDATE

3 For each of the following pairs of variables, identify the explanatory variable and the response variable. If it is not possible to identify this, then write ‘not appropriate’.

a The age of an AFL footballer and his annual salary
b The growth of a plant and the amount of fertiliser it receives
c The number of books read in a week and the eye colour of the readers
d The voting intentions of a woman and her weekly consumption of red meat
e The number of members in a household and the size of the house

4 For each of the following pairs of variables, identify the explanatory variable and the response variable. If it is not possible to identify this, then write ‘not appropriate’.

a The month of the year and the electricity bill for that month
b The mark obtained for a maths test and the number of hours spent preparing for the test
c The mark obtained for a maths test and the mark obtained for an English test
d The cost of grapes (in dollars per kilogram) and the season of the year

5 In a scientific experiment, the explanatory variable was the amount of sleep (in hours) a new mother got per night during the first month following the birth of her baby. The response variable would most likely have been:

A the number of times (per night) the baby woke up for a feed
B the blood pressure of the baby
C the mother’s reaction time (in seconds) to a certain stimulus
D the level of alertness of the baby
E the amount of time (in hours) spent by the mother on reading

6 A paediatrician investigated the relationship between the amount of time children aged two to five spend outdoors and the annual number of visits to his clinic. Which one of the following statements is not true?

A When graphed, the amount of time spent outdoors should be shown on the horizontal axis.
B The annual number of visits to the paediatric clinic is the response variable.
C It is impossible to identify the explanatory variable in this case.
D The amount of time spent outdoors is the explanatory variable.
E The annual number of visits to the paediatric clinic should be shown on the vertical axis.

7 Alex works as a personal trainer at the local gym. He wishes to analyse the relationship between the number of weekly training sessions and the weekly weight loss of his clients. Which one of the following statements is correct?

A When graphed, the number of weekly training sessions should be shown on the vertical axis, as it is the response variable.

B When graphed, the weekly weight loss should be shown on the vertical axis, as it is the explanatory variable.

C When graphed, the weekly weight loss should be shown on the horizontal axis, as it is the explanatory variable.

D When graphed, the number of weekly training sessions should be shown on the horizontal axis, as it is the explanatory variable.

E It is impossible to identify the response variable in this case.

Answer questions 8 to 12 as true or false.

8 When graphing data, the explanatory variable should be placed on the x-axis.

9 The response variable is the same as the dependent variable.

10 If variable A changes due to a change in variable B, then variable A is the response variable.

11 When graphing data the response variable should be placed on the x-axis.

12 The independent variable is the same as the response variable.

MASTER

13 If two variables investigated are the number of minutes on a basketball court and the number of points scored:

a which is the explanatory variable

b which is the response variable?

14 Callum decorated his house with Christmas lights for everyone to enjoy. He investigated two variables, the number of Christmas lights he has and the size of his electricity bill.

a Which is the response variable?

b If Callum was to graph the data, what should be on the x-axis?

c Which is the explanatory variable?

d On the graph, what variable should go on the y-axis?

3.3 Scatterplots

Fitting straight lines to bivariate data

The process of ‘fitting’ straight lines to bivariate data enables us to analyse relationships between the data and possibly make predictions based on the given data set.

We will consider the most common technique for fitting a straight line and determining its equation, namely least squares.

The **linear relationship** expressed as an equation is often referred to as the *linear regression equation* or line. Recall that when we display bivariate data as a **scatterplot**, the explanatory variable is placed on the horizontal axis and the response variable is placed on the vertical axis.

studyon

Unit 3

AOS DA

Topic 6

Concept 5

Scatterplots
Concept summary
Practice questions

eBookplus

Interactivity
Scatterplots
int-6250

eBookplus

Interactivity
Create scatterplots
int-6497

Scatterplots

We often want to know if there is a relationship between two numerical variables. A scatterplot, which gives a visual display of the relationship between two variables, provides a good starting point.

Consider the data obtained from last year's 12B class at Northbank Secondary College. Each student in this class of 29 students was asked to give an estimate of the average number of hours they studied per week during Year 12. They were also asked for the ATAR score they obtained.

The figure below shows the data plotted on a scatterplot.

Average hours of study	ATAR score
18	59
16	67
22	74
27	90
15	62
28	89
18	71
19	60
22	84
30	98
14	54
17	72
14	63
19	72
20	58

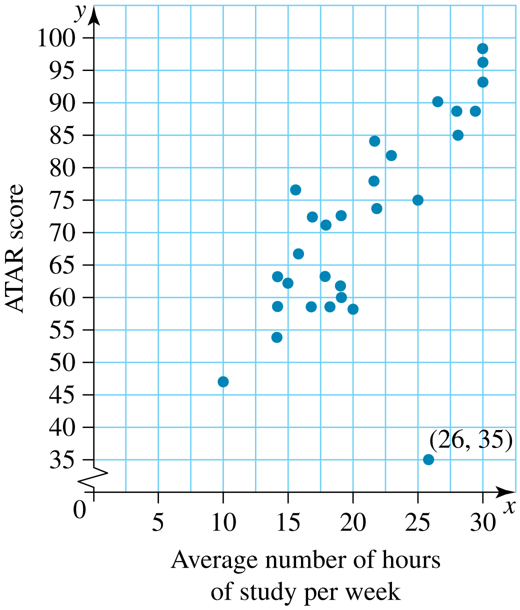
Average hours of study	ATAR score
10	47
28	85
25	75
18	63
19	61
17	59
16	76
14	59
29	89
30	93
30	96
23	82
26	35
22	78

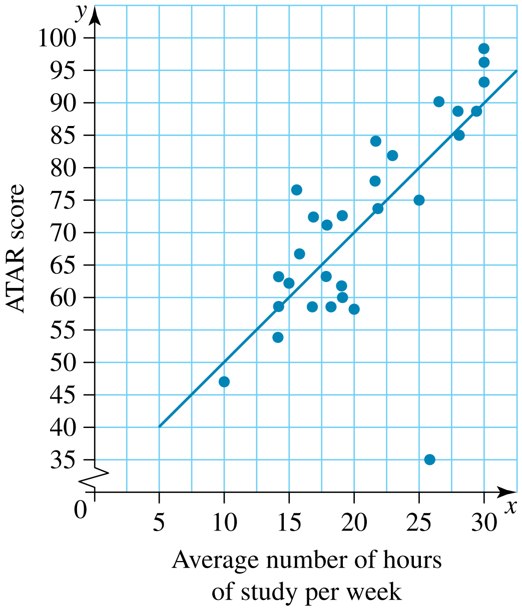
It is reasonable to think that the number of hours of study put in each week by students would affect their ATAR scores and so the number of hours of study per week is the explanatory (independent) variable and appears on the horizontal axis. The ATAR score is the response (dependent) variable and appears on the vertical axis.

There are 29 points on the scatterplot. Each point represents the number of hours of study and the ATAR score of one student.

In analysing the scatterplot we look for a pattern in the way the points lie. Certain patterns tell us that certain relationships exist between the two variables. This is referred to as **correlation**. We look at what type of correlation exists and how strong it is.

In the diagram we see some sort of pattern: the points are spread in a rough corridor from bottom left to top right. We refer to data following such a direction as having a *positive relationship*. This tells us that as the average number of hours studied per week increases, the ATAR score increases.





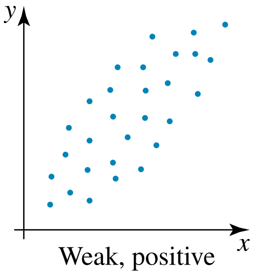
The point (26, 35) is an outlier. It stands out because it is well away from the other points and clearly is not part of the 'corridor' referred to previously. This outlier may have occurred because a student exaggerated the number of hours he or she worked in a week or perhaps there was a recording error. This needs to be checked.

We could describe the rest of the data as having a *linear* form as the straight line in the diagram indicates.

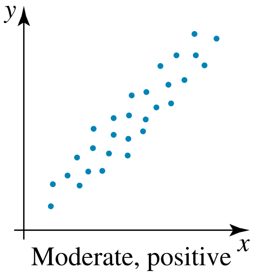
When describing the relationship between two variables displayed on a scatterplot, we need to comment on:

(a) the direction — whether it is positive or negative
(b) the form — whether it is linear or non-linear
(c) the strength — whether it is strong, moderate or weak
(d) possible outliers.

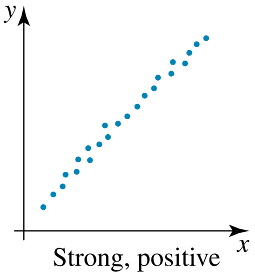
Here is a gallery of scatterplots showing the various patterns we look for.



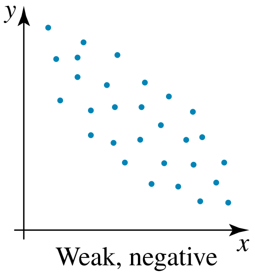
Weak, positive linear relationship



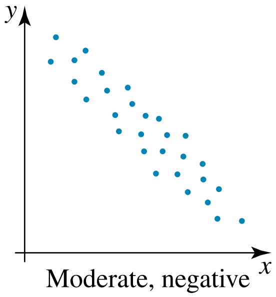
Moderate, positive linear relationship



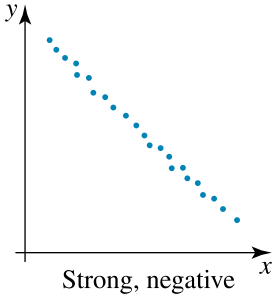
Strong, positive linear relationship



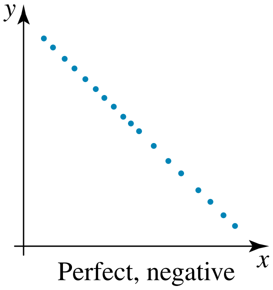
Weak, negative linear relationship



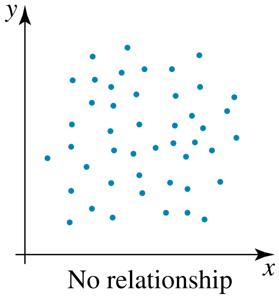
Moderate, negative linear relationship



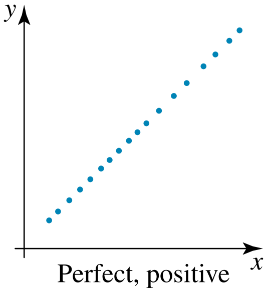
Strong, negative linear relationship



Perfect, negative linear relationship



No relationship

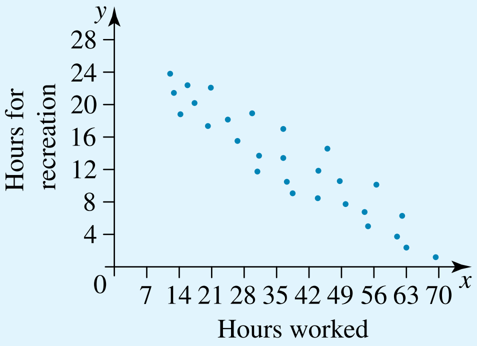


Perfect, positive linear relationship

WORKED EXAMPLE 2

The scatterplot shows the number of hours people spend at work each week and the number of hours people get to spend on recreational activities during the week.

Decide whether or not a relationship exists between the variables and, if it does, comment on whether it is positive or negative; weak, moderate or strong; and whether or not it has a linear form.



THINK

1 The points on the scatterplot are spread in a certain pattern, namely in a rough corridor from the top left to the bottom right corner. This tells us that as the work hours increase, the recreation hours decrease.

2 The corridor is straight (that is, it would be reasonable to fit a straight line into it).

3 The points are neither too tight nor too dispersed.

4 The pattern resembles the central diagram in the gallery of scatterplots shown previously.

WRITE

There is a moderate, negative linear relationship between the two variables.

WORKED EXAMPLE 3

Data showing the average weekly number of hours studied by each student in 12B at Northbank Secondary College and the corresponding height of each student (correct to the nearest tenth of a metre) are given in the table.

Average hours of study	18	16	22	27	15	28	18	20	10	28	25	18	19	17
Height (m)	1.5	1.9	1.7	2.0	1.9	1.8	2.1	1.9	1.9	1.5	1.7	1.8	1.8	2.1

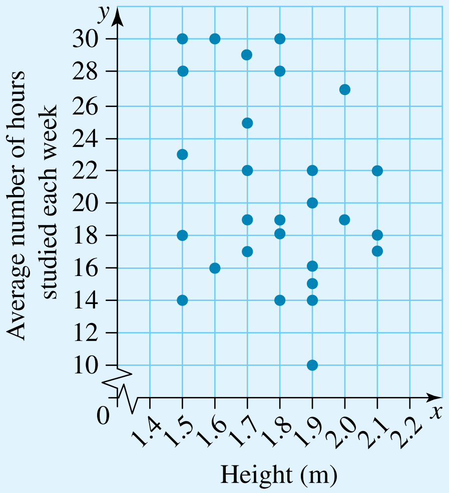
Average hours of study	19	22	30	14	17	14	19	16	14	29	30	30	23	22
Height (m)	2.0	1.9	1.6	1.5	1.7	1.8	1.7	1.6	1.9	1.7	1.8	1.5	1.5	2.1

Construct a scatterplot for the data and use it to comment on the direction, form and strength of any relationship between the number of hours studied and the height of the students.

THINK

1 CAS can be used to assist you in drawing a scatterplot.

WRITE/DRAW

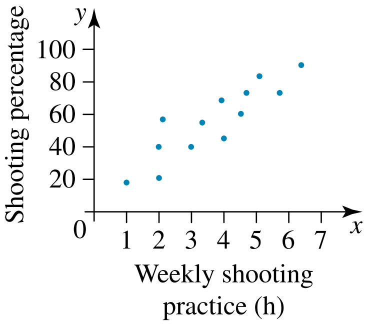


2 Comment on the direction of any relationship.	There is no relationship; the points appear to be randomly placed.
3 Comment on the form of the relationship.	There is no form, no linear trend, no quadratic trend, just a random placement of points.
4 Comment on the strength of any relationship.	Since there is no relationship, strength is not relevant.
5 Draw a conclusion.	Clearly, from the graph, the number of hours spent studying for VCE has no relation to how tall you might be.

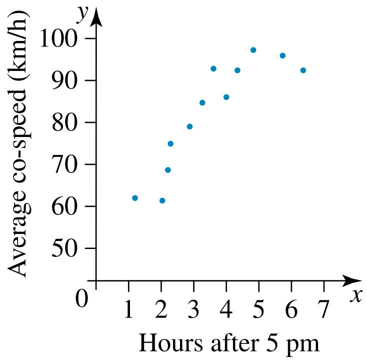
EXERCISE 3.3 Scatterplots

PRACTISE

1 WE2 The scatterplot shown represents the number of hours of basketball practice each week and a player's shooting percentage. Decide whether or not a relationship exists between the variables and, if it does, comment on whether it is positive or negative; weak, moderate or strong; and whether or not it is linear form.



2 The scatterplot shown shows the hours after 5 pm and the average speed of cars on a freeway. Explain the direction, form and strength of the relationship of the two variables.



3 WE3 Data on the height of a person and the length of their hair is shown. Construct a scatterplot for the data and use it to comment on the direction, form and strength of any relationship between the height of a person and the length of their hair.

Height (cm)	158	164	184	173	194	160	198	186	166
Hair length (cm)	18	12	5	10	7	3	10	6	14

4 The following table shows data on hours spent watching television per week and your age. Use the data to construct a scatterplot and use it to comment on the direction, form and strength of any relationship between the two variables.

Age (yr)	12	25	61	42	18	21	33	15	29
TV per week (h)	23	30	26	18	12	30	20	19	26

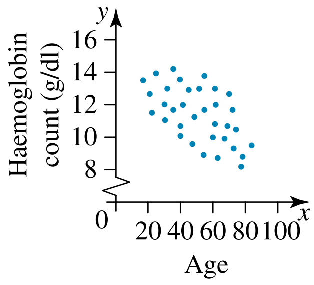
CONSOLIDATE

5 For each of the following pairs of variables, write down whether or not you would reasonably expect a relationship to exist between the pair and, if so, comment on whether it would be a positive or negative association.

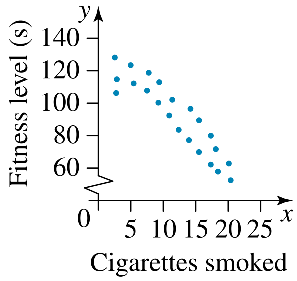
a Time spent in a supermarket and total money spent
b Income and value of car driven
c Number of children living in a house and time spent cleaning the house
d Age and number of hours of competitive sport played per week
e Amount spent on petrol each week and distance travelled by car each week
f Number of hours spent in front of a computer each week and time spent playing the piano each week
g Amount spent on weekly groceries and time spent gardening each week

6 For each of the scatterplots, describe whether or not a relationship exists between the variables and, if it does, comment on whether it is positive or negative, whether it is weak, moderate or strong and whether or not it has a linear form.

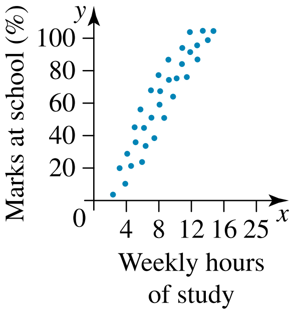
a



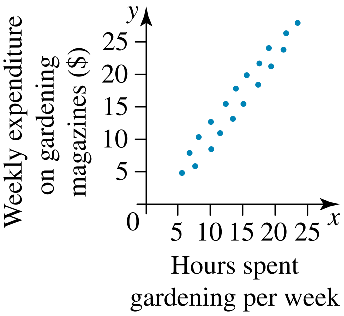
b



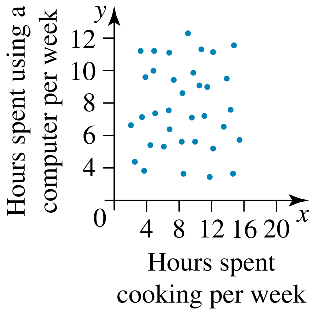
c



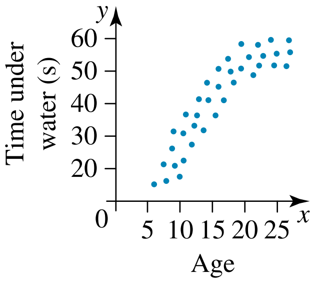
d



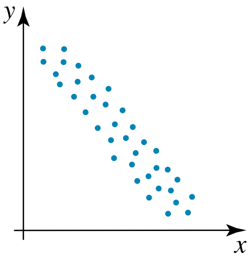
e



f



7 From the scatterplot shown, it would be reasonable to observe that:



A as the value of x increases, the value of y increases
B as the value of x increases, the value of y decreases
C as the value of x increases, the value of y remains the same
D as the value of x remains the same, the value of y increases
E there is no relationship between x and y

8 The population of a municipality (to the nearest ten thousand) together with the number of primary schools in that particular municipality is given below for 11 municipalities.

Population (× 1000)	110	130	130	140	150	160	170	170	180	180	190
Number of primary schools	4	4	6	5	6	8	6	7	8	9	8

Construct a scatterplot for the data and use it to comment on the direction, form and strength of any relationship between the population and the number of primary schools.

9 The table contains data for the time taken to do a paving job and the cost of the job.

Construct a scatterplot for the data. Comment on whether a relationship exists between the time taken and the cost. If there is a relationship describe it.

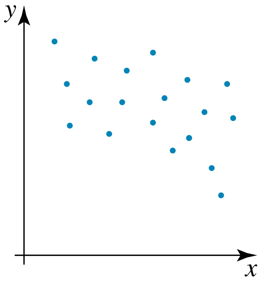
Time taken (hours)	Cost of job ($)
5	1000
7	1000
5	1500
8	1200
10	2000
13	2500
15	2800
20	3200
18	2800
25	4000
23	3000

10 The table shows the time of booking (how many days in advance) of the tickets for a musical performance and the corresponding row number in A-reserve seating.

Time of booking	Row number	Time of booking	Row number	Time of booking	Row number
5	15	14	12	25	3
6	15	14	10	28	2
7	15	17	11	29	2
7	14	20	10	29	1
8	14	21	8	30	1
11	13	22	5	31	1
13	13	24	4		

Construct a scatterplot for the data. Comment on whether a relationship exists between the time of booking and the number of the row and, if there is a relationship, describe it.

11 The correlation of this scatterplot is:



A weak, positive, linear
B no correlation
C strong, positive linear
D weak, negative, linear
E strong, negative, linear

12 Draw a scatterplot to display the following data:

a by hand **b** using CAS.

Number of dry-cleaning items	1	2	3	4	5	6	7
Cost ($)	12	16	19	20	22	24	25

13 Draw a scatterplot to display the following data:

a by hand **b** using CAS.

Maximum daily temperature (°C)	26	28	19	17	32	36	33	23	24	18
Number of drinks sold	135	156	98	87	184	133	175	122	130	101

14 Describe the correlation between:

a the number of dry-cleaning items and the cost in question **12**
b the maximum daily temperature and the number of drinks sold in question **13**.

MASTER

15 Draw a scatterplot and describe the correlation for the following data.

	NSW	VIC	QLD	SA	WA	TAS	NT	ACT
Population	7 500 600	5 821 000	4 708 000	1 682 000	2 565 000	514 000	243 000	385 000
Area of land (km^2)	800 628	227 010	1 723 936	978 810	2 526 786	64 519	1 335 742	2 358

16 The table at right contains data giving the time taken to engineer a finished product from the raw recording (of a song, say) and the length of the finished product.

a Construct a scatterplot for these data.
b Comment on whether a relationship exists between the time spent engineering and the length of the finished recording.

Time spent engineering in studio (hours)	Finished length of recording (minutes)
1	3
2	4
3	10
4	12
5	20
6	16
7	18
8	25
9	30
10	28
11	35
12	36
13	39
14	42
15	45

3.4 Pearson's product–moment correlation coefficient

studyon

Unit 3

AOS DA

Topic 6

Concept 6

Pearson's product–moment correlation coefficient (*r*)
Concept summary
Practice questions

eBook*plus*

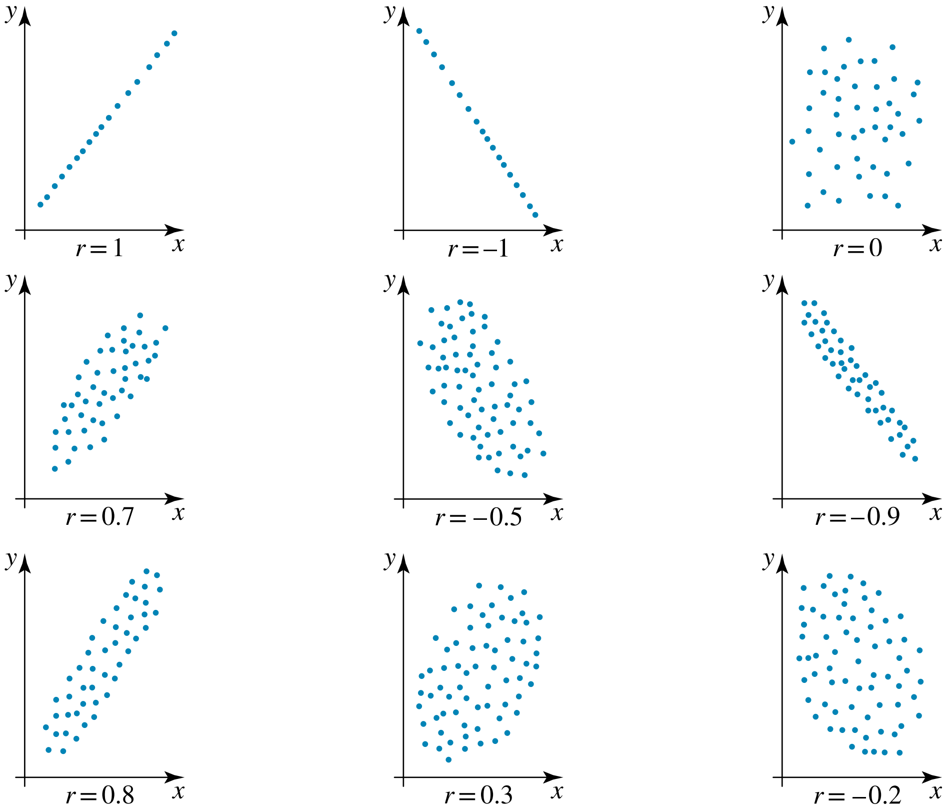
Interactivity
Pearson's product–moment correlation coefficient and the coefficient of determination
int-6251

In the previous section, we estimated the strength of association by looking at a scatterplot and forming a judgement about whether the correlation between the variables was positive or negative and whether the correlation was weak, moderate or strong.

A more precise tool for measuring correlation between two variables is **Pearson's product–moment correlation coefficient**. This coefficient is used to measure the strength of *linear relationships* between variables.

The symbol for Pearson's product–moment correlation coefficient is r. The value of r ranges from -1 to 1; that is, $-1 \leq r \leq 1$.

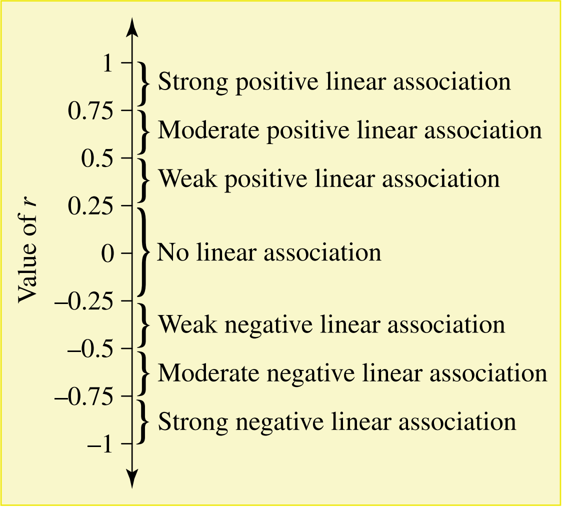
Following is a gallery of scatterplots with the corresponding value of r for each.



The two extreme values of r (1 and -1) are shown in the first two diagrams respectively.

From these diagrams we can see that a value of $r = 1$ or -1 means that there is perfect linear association between the variables.

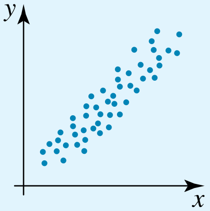
The value of the Pearson's product–moment correlation coefficient indicates the strength of the linear relationship between two variables. The diagram at right gives a rough guide to the strength of the correlation based on the value of r.


Value of r
1
0.75
0.5
0.25
0
–0.25
–0.5
–0.75
–1
Strong positive linear association
Moderate positive linear association
Weak positive linear association
No linear association
Weak negative linear association
Moderate negative linear association
Strong negative linear association

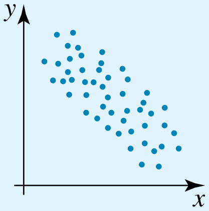
WORKED EXAMPLE 4

For each of the following:

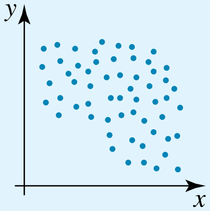
a



b



c



i Estimate the value of Pearson's product–moment correlation coefficient (r) from the scatterplot.

ii Use this to comment on the strength and direction of the relationship between the two variables.

THINK	WRITE
a i Compare these scatterplots with those in the gallery of scatterplots shown previously and estimate the value of r.	a i $r \approx 0.9$
ii Comment on the strength and direction of the relationship.	ii The relationship can be described as a strong, positive, linear relationship.
b Repeat parts i and ii as in a.	b i $r \approx -0.7$
	ii The relationship can be described as a moderate, negative, linear relationship.
c Repeat parts i and ii as in a.	c i $r \approx -0.1$
	ii There is almost no linear relationship.

Note that the symbol ≈ means 'approximately equal to'. We use it instead of the = sign to emphasise that the value (in this case r) is only an estimate.

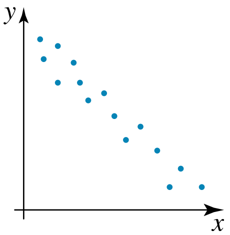
In completing Worked example 4 above, we notice that estimating the value of r from a scatterplot is rather like making an informed guess. In the next section, we will see how to obtain the actual value of r.

EXERCISE 3.4 Pearson's product–moment correlation coefficient

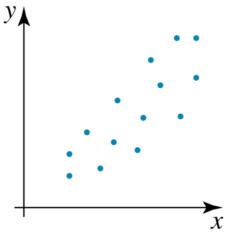
PRACTISE

1 WE4 For each of the following:

a



b



i estimate the Pearson's product–moment correlation coefficient (r) from the scatterplot.

ii use this to comment on the strength and direction of the relationship between the two variables.

2 What type of linear relationship does each of the following values of r suggest?

a 0.85

b −0.3

CONSOLIDATE

3 What type of linear relationship does each of the following values of r suggest?

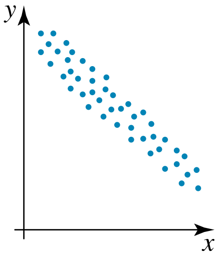
a 0.21 **b** 0.65 **c** −1 **d** −0.78

4 What type of linear relationship does each of the following values of r suggest?

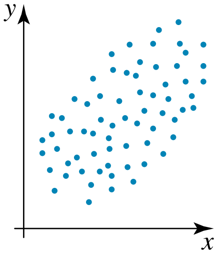
a 1 **b** 0.9 **c** −0.34 **d** −0.1

5 For each of the following:

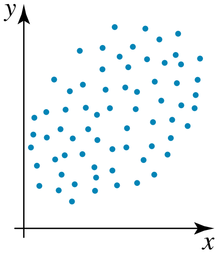
a



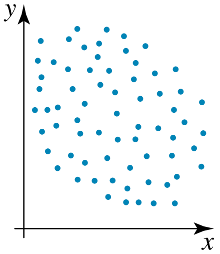
b



c



d

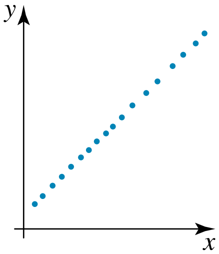


i estimate the value of Pearson's product–moment correlation coefficient (r), from the scatterplot.

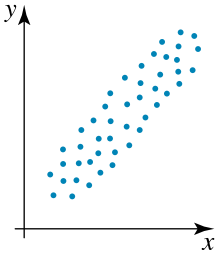
ii use this to comment on the strength and direction of the relationship between the two variables.

6 For each of the following:

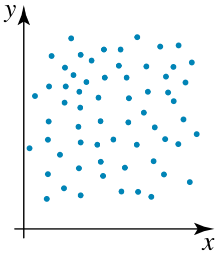
a



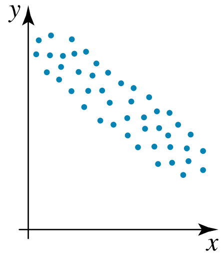
b



c



d

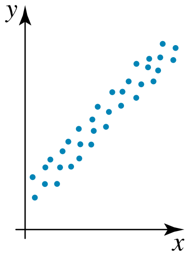


i estimate the value of Pearson's product–moment correlation coefficient (r), from the scatterplot.

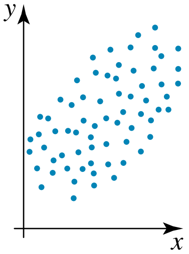
ii use this to comment on the strength and direction of the relationship between the two variables.

7 A set of data relating the variables x and y is found to have an r value of 0.62. The scatterplot that could represent the data is:

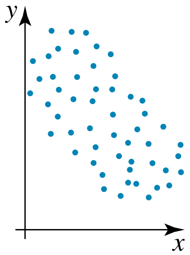
A



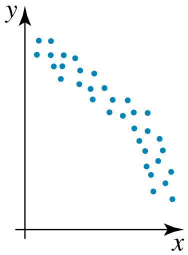
B



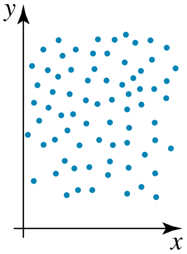
C



D



E



8 A set of data relating the variables x and y is found to have an r value of −0.45. A true statement about the relationship between x and y is:

A There is a strong linear relationship between x and y and when the x-values increase, the y-values tend to increase also.

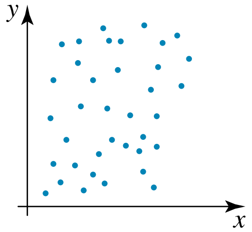
B There is a moderate linear relationship between x and y and when the x-values increase, the y-values tend to increase also.

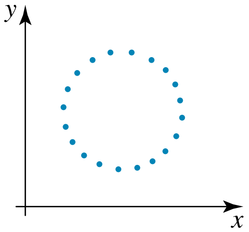
C There is a moderate linear relationship between x and y and when the x-values increase, the y-values tend to decrease.

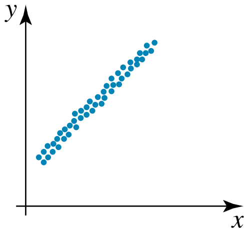
D There is a weak linear relationship between x and y and when the x-values increase, the y-values tend to increase also.

E There is a weak linear relationship between x and y and when the x-values increase, the y-values tend to decrease.

9 From the scatterplots shown estimate the value of r and comment on the strength and direction of the relationship between the two variables.

a


b


c


10 A weak, negative, linear association between two variables would have an r value closest to:

A -0.55 B 0.55 C -0.65 D -0.45 E 0.45

11 Which of the following is *not* a Pearson product–moment correlation coefficient?

A 1.0 B 0.99 C -1.1 D -0.01 E 0

12 Draw a scatterplot that has a Pearson product–moment correlation coefficient of approximately -0.7.

MASTER

13 If two variables have an r value of 1, then they are said to have:

A a strong positive linear relationship
B a strong negative linear relationship
C a perfect positive relationship
D a perfect negative linear relationship
E a perfect positive linear relationship

14 Which is the correct ascending order of positive values of r?

A Strong, Moderate, Weak, No linear association
B Weak, Strong, Moderate, No linear association
C No linear association, Weak, Moderate, Strong
D No linear association, Moderate, Weak, Strong
E Strong, Weak, Moderate, No linear association

3.5 Calculating *r* and the coefficient of determination

Pearson's product–moment correlation coefficient (*r*)

The formula for calculating Pearson's correlation coefficient r is as follows:

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

where n is the number of pairs of data in the set
s_x is the standard deviation of the x-values
s_y is the standard deviation of the y-values
$\bar{x}$ is the mean of the x-values
$\bar{y}$ is the mean of the y-values.

The calculation of r is often done using CAS.

studyon

Unit 3

AOS DA

Topic 6

Concept 7

Use and interpretation of *r*
Concept summary
Practice questions

There are two important limitations on the use of *r*. First, since *r* measures the strength of a linear relationship, it would be inappropriate to calculate *r* for data which are not linear — for example, data which a scatterplot shows to be in a quadratic form.

Second, outliers can bias the value of *r*. Consequently, if a set of linear data contains an outlier, then *r* is not a reliable measure of the strength of that linear relationship.

The calculation of *r* is applicable to sets of bivariate data which are known to be linear in form and which do not have outliers.

With those two provisos, it is good practice to draw a scatterplot for a set of data to check for a linear form and an absence of outliers before *r* is calculated. Having a scatterplot in front of you is also useful because it enables you to estimate what the value of *r* might be — as you did in the previous exercise, and thus you can check that your workings are correct.

WORKED EXAMPLE 5

The heights (in centimetres) of 21 football players were recorded against the number of marks they took in a game of football. The data are shown in the following table.

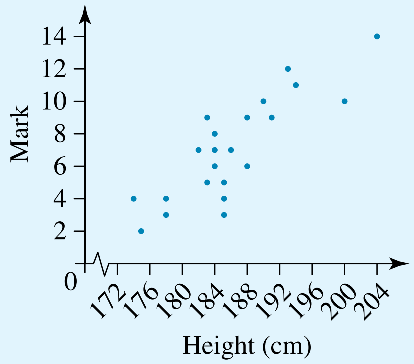
a **Construct a scatterplot for the data.**

b **Comment on the correlation between the heights of players and the number of marks that they take, and estimate the value of *r*.**

c **Calculate *r* and use it to comment on the relationship between the heights of players and the number of marks they take in a game.**

Height (cm)	Number of marks taken
184	6
194	11
185	3
175	2
186	7
183	5
174	4
200	10
188	9
184	7
188	6

Height (cm)	Number of marks taken
182	7
185	5
183	9
191	9
177	3
184	8
178	4
190	10
193	12
204	14

THINK	WRITE/DRAW
a Height is the explanatory variable, so plot it on the *x*-axis; the number of marks is the response variable, so show it on the *y*-axis.	**a** 
b Comment on the correlation between the variables and estimate the value of *r*.	**b** The data show what appears to be a linear form of moderate strength. We might expect $r \approx 0.8$.
c 1 Because there is a linear form and there are no outliers, the calculation of *r* is appropriate.	**c**
2 Use CAS to find the value of *r*. Round correct to 2 decimal places.	$r = 0.859\,311...$ ≈ 0.86
3 The value of $r = 0.86$ indicates a strong positive linear relationship.	$r = 0.86$. This indicates there is a strong positive linear association between the height of a player and the number of marks he takes in a game. That is, the taller the player, the more marks we might expect him to take.

Correlation and causation

In Worked example 5 we saw that $r = 0.86$. While we are entitled to say that there is a strong association between the height of a footballer and the number of marks he takes, we cannot assert that the height of a footballer causes him to take a lot of marks. Being tall might assist in taking marks, but there will be many other factors which come into play; for example, skill level, accuracy of passes from teammates, abilities of the opposing team, and so on.

So, while establishing a high degree of correlation between two variables may be interesting and can often flag the need for further, more detailed investigation, it in no way gives us any basis to comment on whether or not one variable *causes* particular values in another variable.

As we have looked at earlier in this topic, *correlation* is a statistic measure that defines the size and direction of the relationship between two variables. **Causation** states that one event is the result of the occurrence of the other event (or variable). This is also referred to as **cause and effect**, where one event is the cause and this makes another event happen, this being the effect.

An example of a cause and effect relationship could be an alarm going off (cause — happens first) and a person waking up (effect — happens later). It is also important to realise that a high correlation does not imply causation. For example, a person smoking could have a high correlation with alcoholism but it is not necessarily the cause of it, thus they are different.

One way to test for causality is experimentally, where a control study is the most effective. This involves splitting the sample or population data and making one a control group (e.g. one group gets a placebo and the other get some form of medication). Another way is via an observational study which also compares against a control variable, but the researcher has no control over the experiment (e.g. smokers and non-smokers who develop lung cancer). They have no control over whether they develop lung cancer or not.

Non-causal explanations

studyon

Unit 3
AOS DA
Topic 6
Concept 8

Non-causal explanations
Concept summary
Practice questions

Although we may observe a strong correlation between two variables, this does not necessarily mean that an association exists. In some cases the correlation between two variables can be explained by a common response variable which provides the association. For example, a study may show that there is a strong correlation between house sizes and the life expectancy of home owners. While a bigger house will not directly lead to a longer life expectancy, a common response variable, the income of the house owner, provides a direct link to both variables and is more likely to be the underlying cause for the observed correlation.

In other cases there may be hidden, confounding reasons for an observed correlation between two variables. For example, a lack of exercise may provide a strong correlation to heart failure, but other hidden variables such as nutrition and lifestyle might have a stronger influence.

Finally, an association between two variables may be purely down to coincidence. The larger a data set is, the less chance there is that coincidence will have an impact.

When looking at correlation and causation be sure to consider all of the possible explanations before jumping to conclusions. In professional research, many similar tests are often carried out to try to identify the exact cause for a shown correlation between two variables.

The coefficient of determination (r^2)

studyon

Unit 3
AOS DA
Topic 7
Concept 4

The coefficient of determination (r^2)
Concept summary
Practice questions

The **coefficient of determination** is given by r^2. It is very easy to calculate, we merely square Pearson's product–moment correlation coefficient (r). The value of the coefficient of determination ranges from 0 to 1; that is, $0 \leq r^2 \leq 1$.

The coefficient of determination is useful when we have two variables which have a linear relationship. It tells us the proportion of variation in one variable which can be explained by the variation in the other variable.

The coefficient of determination provides a measure of how well the linear rule linking the two variables (x and y) predicts the value of y when we are given the value of x.

WORKED EXAMPLE 6

A set of data giving the number of police traffic patrols on duty and the number of fatalities for the region was recorded and a correlation coefficient of $r = -0.8$ was found.

a **Calculate the coefficient of determination and interpret its value.**

b **If it was a causal relationship state the most likely variable to be the cause and effect.**

THINK	WRITE
a 1 Calculate the coefficient of determination by squaring the given value of *r*.	**a** Coefficient of determination $= r^2$ $= (-0.8)^2$ $= 0.64$
2 Interpret your result.	We can conclude from this that 64% of the variation in the number of fatalities can be explained by the variation in the number of police traffic patrols on duty. This means that the number of police traffic patrols on duty is a major factor in predicting the number of fatalities.
b 1 The two variables are: i the number of police on traffic patrols, and ii the number of fatalities.	**b**
2 Cause happens first and it has an effect later on.	FIRST: Number of police cars on patrol (cause) LATER: Number of fatalities (effect)

Note: In Worked example 6, 64% of the variation in the number of fatalities was due to the variation in the number of police cars on duty and 36% was due to other factors; for example, days of the week or hour of the day.

EXERCISE 3.5 Calculating *r* and the coefficient of determination

PRACTISE

1 WE5 The heights (cm) of basketball players were recorded against the number of points scored in a game. The data are shown in the following table.

Height (cm)	Points scored	Height (cm)	Points scored
194	6	201	13
203	4	196	10
208	18	205	20
198	22	215	14
195	2	203	3

a Construct a scatterplot of the data.

b Comment on the correlation between the heights of basketballers and the number of points scored, and estimate the value of *r*.

c Calculate the *r* value and use it to comment on the relationship between heights of players and the number of points scored in a game.

2 The following table shows the gestation time and the birth mass of 10 babies.

Gestation time (weeks)	31	32	33	34	35	36	37	38	39	40
Birth mass (kg)	1.08	1.47	1.82	2.06	2.23	2.54	2.75	3.11	3.08	3.37

a Construct a scatterplot of the data.

b Comment on the correlation between the 'gestation time' and 'birth mass', and estimate the value of r.

c Calculate the r value and use it to comment on the relationship between gestation time and birth mass.

3 WE6 Data on the number of booze buses in use and the number of drivers registering a blood alcohol reading over 0.05 was recorded and a correlation coefficient of $r = 0.77$ was found.

a Calculate the coefficient of determination and interpret its value.

b If there was a causal relationship, state the most likely variable to be the cause and the variable to be the effect.

4 An experiment was conducted that looked at the number of books read by a student and their spelling skills. If this was a cause and effect relationship, what variable most likely represents the cause and what variable represents the effect?

CONSOLIDATE

5 The yearly salary (× $1000) and the number of votes polled in the Brownlow medal count are given below for 10 footballers.

Yearly salary (× $1000)	360	400	320	500	380	420	340	300	280	360
Number of votes	24	15	33	10	16	23	14	21	31	28

a Construct a scatterplot for the data.

b Comment on the correlation of salary and the number of votes and make an estimate of r.

c Calculate r and use it to comment on the relationship between yearly salary and number of votes.

6 A set of data, obtained from 40 smokers, gives the number of cigarettes smoked per day and the number of visits per year to the doctor. The Pearson's correlation coefficient for these data was found to be 0.87. Calculate the coefficient of determination for the data and interpret its value.

7 Data giving the annual advertising budgets (× $1000) and the yearly profit increases (%) of 8 companies are shown below.

Annual advertising budget (× $1000)	11	14	15	17	20	25	25	27
Yearly profit increase (%)	2.2	2.2	3.2	4.6	5.7	6.9	7.9	9.3

a Construct a scatterplot for these data.

b Comment on the correlation of the advertising budget and profit increase and make an estimate of r.

c Calculate r.

d Calculate the coefficient of determination.

e Write the proportion of the variation in the yearly profit increase that can be explained by the variation in the advertising budget.

8 Data showing the number of tourists visiting a small country in a month and the corresponding average monthly exchange rate for the country's currency against the American dollar are as given.

Number of tourists (× 1000)	2	3	4	5	7	8	8	10
Exchange rate	1.2	1.1	0.9	0.9	0.8	0.8	0.7	0.6

a Construct a scatterplot for the data.

b Comment on the correlation between the number of tourists and the exchange rate and give an estimate of *r*.

c Calculate *r*.

d Calculate the coefficient of determination.

e Write the proportion of the variation in the number of tourists that can be explained by the exchange rate.

9 Data showing the number of people in 9 households against weekly grocery costs are given below.

Number of people in household	2	5	6	3	4	5	2	6	3
Weekly grocery costs ($)	60	180	210	120	150	160	65	200	90

a Construct a scatterplot for the data.

b Comment on the correlation of the number of people in a household and the weekly grocery costs and give an estimate of *r*.

c If this is a causal relationship state the most likely variable to be the cause and which to be the effect.

d Calculate *r*.

e Calculate the coefficient of determination.

f Write the proportion of the variation in the weekly grocery costs that can be explained by the variation in the number of people in a household.

10 Data showing the number of people on 8 fundraising committees and the annual funds raised are given in the table.

Number of people on committee	3	6	4	8	5	7	3	6
Annual funds raised ($)	4500	8500	6100	12 500	7200	10 000	4700	8800

a Construct a scatterplot for these data.

b Comment on the correlation between the number of people on a committee and the funds raised and make an estimate of *r*.

c Calculate *r*.

d Based on the value of *r* obtained in part c, would it be appropriate to conclude that the increase in the number of people on the fundraising committee causes the increase in the amount of funds raised?

e Calculate the coefficient of determination.

f Write the proportion of the variation in the funds raised that can be explained by the variation in the number of people on a committee.

The following information applies to questions **11** and **12**. A set of data was obtained from a large group of women with children under 5 years of age. They were asked the number of hours they worked per week and the amount of money they spent on child care. The results were recorded and the value of Pearson's correlation coefficient was found to be 0.92.

11 Which of the following is not true?

A The relationship between the number of working hours and the amount of money spent on child care is linear.

B There is a positive correlation between the number of working hours and the amount of money spent on child care.

C The correlation between the number of working hours and the amount of money spent on child care can be classified as strong.

D As the number of working hours increases, the amount spent on child care increases as well.

E The increase in the number of hours worked causes the increase in the amount of money spent on child care.

12 Which of the following is not true?

A The coefficient of determination is about 0.85.

B The number of working hours is the major factor in predicting the amount of money spent on child care.

C About 85% of the variation in the number of hours worked can be explained by the variation in the amount of money spent on child care.

D Apart from number of hours worked, there could be other factors affecting the amount of money spent on child care.

E About $\frac{17}{20}$ of the variation in the amount of money spent on child care can be explained by the variation in the number of hours worked.

13 To experimentally test if a relationship is a cause and effect relationship the data is usually:

A randomly selected

B split to produce a control study

C split into genders

D split into age categories

E kept as small as possible

14 A study into the unemployment rate in different Melbourne studies found a negative correlation between the unemployment rate in a suburb and the average salary of adult workers in the same suburb.

Using your knowledge of correlation and causation, explain whether this is an example of cause and effect. If not, what non-causal explanations might explain the correlation?

MASTER

15 The main problem when using an observational study to determine causality is:

A collecting the data

B splitting the data

C controlling the control group

D getting a high enough coefficient of determination

E none of the above

16 An investigation is undertaken with people following the Certain Slim diet to explore the link between weeks of dieting and total weight loss. The data are shown in the table.

Total weight loss (kg)	Number of weeks on the diet
1.5	1
4.5	5
9	8
3	3
6	6
8	9
3.5	4
3	2
6.5	7
8.5	10
4	4
6.5	6
10	9
2.5	2
6	5

a Display the data on a scatterplot.

b Describe the association between the two variables in terms of direction, form and strength.

c Is it appropriate to use Pearson's correlation coefficient to explain the link between the number of weeks on the Certain Slim diet and total weight loss?

d Estimate the value of Pearson's correlation coefficient from the scatterplot.

e Calculate the value of this coefficient.

f Is the total weight loss affected by the number of weeks staying on the diet?

g Calculate the value of the coefficient of determination.

h What does the coefficient of determination say about the relationship between total weight loss and the number of weeks on the Certain Slim diet?

3.6 Fitting a straight line — least-squares regression

studyon

Unit 3

AOS DA

Topic 7

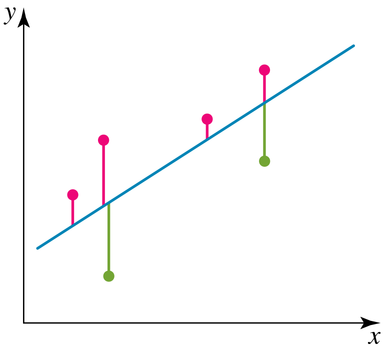
Concept 1

Least-squares regression
Concept summary
Practice questions

A method for finding the equation of a straight line which is fitted to data is known as the method of **least-squares regression**. It is used when data show a linear relationship and have no obvious outliers.

To understand the underlying theory behind least-squares, consider the regression line shown.

We wish to minimise the total of the vertical lines, or 'errors' in some way. For example, balancing the errors above and below the line. This is reasonable, but for sophisticated mathematical reasons it is preferable to minimise the sum of the *squares* of each of these errors. This is the essential mathematics of least-squares regression.



The calculation of the equation of a least-squares regression line is simple using CAS.

WORKED EXAMPLE 7

A study shows the more calls a teenager makes on their mobile phone, the less time they spend on each call. Find the equation of the linear regression line for the number of calls made plotted against call time in minutes using the least-squares method on CAS. Express coefficients correct to 2 decimal places, and calculate the coefficient of determination to assess the strength of the association.

Number of minutes (x)	1	3	4	7	10	12	14	15
Number of calls (y)	11	9	10	6	8	4	3	1

THINK	WRITE
1 Enter the data into CAS to find the equation of least squares regression line.	$y = 11.7327 - 0.634\,271x$
2 Write the equation with coefficients expressed to 2 decimal places.	$y = 11.73 - 0.63x$
3 Write the equation in terms of the variable names. Replace x with number of minutes and y with number of calls.	Number of calls = 11.73 − 0.63 × no. of minutes
4 Read the r^2 value from your calculator.	$r^2 = 0.87$, so we can conclude that 87% of the variation in y can be explained by the variation in x. Therefore the strength of the linear association between y and x is strong.

Calculating the least-squares regression line by hand

eBook*plus*

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

The least-squares regression equation minimises the average deviation of the points in the data set from the line of best fit. This can be shown using the following summary data and formulas to arithmetically determine the least-squares regression equation.

Summary data needed:

$\bar{x}$ the mean of the explanatory variable (x-variable)

$\bar{y}$ the mean of the response variable (y-variable)

s_x the standard deviation of the explanatory variable

s_y the standard deviation of the response variable

r Pearson's product–moment correlation coefficient.

Formula to use:

The general form of the least-squares regression line is

$$y = a + bx$$

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

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

Alternatively, if the general form is given as $y = mx + c$, then $m = r\dfrac{s_y}{s_x}$ and $c = \bar{y} - m\bar{x}$.

WORKED EXAMPLE 8

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

Mean height of the husbands: 180 cm

Mean height of the wives: 169 cm

Standard deviation of the height of the husbands: 5.3 cm

Standard deviation of the height of the wives: 4.8 cm

Correlation coefficient, $r = 0.85$

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

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

a **Which variable is the response variable?**

b **Calculate the value of b (correct to 2 significant figures).**

c **Calculate the value of a (correct to 4 significant figures).**

d **Use the equation of the regression line to predict the height of a wife whose husband is 195 cm tall (correct to the nearest cm).**

THINK	WRITE
a Recall that the response variable is the subject of the equation in $y = a + bx$ form; that is, y.	a The response variable is the height of the wife.
b 1 The value of b is the gradient of the regression line. Write the formula and state the required values.	b $b = r\dfrac{s_y}{s_x}$ $\quad r = 0.85$, $s_y = 4.8$ and $s_x = 5.3$
2 Substitute the values into the formula and evaluate b.	$= 0.85 \times \dfrac{4.8}{5.3}$ $= 0.7698$ ≈ 0.77
c 1 The value of a is the y-intercept of the regression line. Write the formula and state the required values.	c $a = \bar{y} - b\bar{x}$ $\bar{y} = 169$, $\bar{x} = 180$ and $b = 0.7698$ (from part b)
2 Substitute the values into the formula and evaluate a.	$= 169 - 0.7698 \times 180$ $= 30.436$ ≈ 30.44
d 1 State the equation of the regression line, using the values calculated from parts b and c. In this equation, y represents the height of the wife and x represents the height of the husband.	d $y = 30.44 + 0.77x$ or height of wife $= 30.44 + 0.77 \times$ height of husband
2 The height of the husband is 195 cm, so substitute $x = 195$ into the equation and evaluate.	$= 30.44 + 0.77 \times 195$ $= 180.59$
3 Write a statement, rounding your answer correct to the nearest cm.	Using the equation of the regression line found, the wife's height would be 181 cm.

EXERCISE 3.6 Fitting a straight line — least-squares regression

PRACTISE

1 WE7 A study shows that as the temperature increases the sales of air conditioners increase. Find the equation of the linear regression line for the number of air conditioners sold per week plotted against the temperature in °C using the least-squares method on CAS. Also find the coefficient of determination. Express the values correct to 2 decimal places and comment on the association between temperature and air conditioner sales.

Temperature °C (*x*)	21	23	25	28	30	32	35	38
Air conditioner sales (*y*)	3	7	8	14	17	23	25	37

2 Consider the following data set: x represents the month, y represents the number of dialysis patients treated.

x	1	2	3	4	5	6	7	8
y	5	9	7	14	14	19	21	23

Using CAS find the equation of the linear regression line and the coefficient of determination, with values correct to 2 decimal places.

3 WE8 A study was conducted to find the relationship between the height of Year 12 boys and the height of Year 12 girls. The following details were found.

Mean height of the boys: 182 cm

Mean height of the girls: 166 cm

Standard deviation of the height of boys: 6.1 cm

Standard deviation of the height of girls: 5.2 cm

Correlation coefficient, $r = 0.82$

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

Height of Year 12 Boy $= a + b \times$ height of Year 12 girl

a Which is the explanatory variable?
b Calculate the value of b (correct to 2 significant figures).
c Calculate the value of a (correct to 2 decimal places).

4 Given the summary details $\bar{x} = 4.4$, $s_x = 1.2$, $\bar{y} = 10.5$, $s_y = 1.4$ and $r = -0.67$, find the value of b and a for the equation of the regression line $y = a + bx$.

CONSOLIDATE

5 Find the equation of the linear regression line for the following data set using the least-squares method, and comment on the strength of the association.

x	4	6	7	9	10	12	15	17
y	10	8	13	15	14	18	19	23

6 Find the equation of the linear regression line for the following data set using the least-squares method.

x	1	2	3	4	5	6	7	8	9
y	35	28	22	16	19	14	9	7	2

7 Find the equation of the linear regression line for the following data set using the least-squares method.

x	−4	−2	−1	0	1	2	4	5	5	7
y	6	7	3	10	16	9	12	16	11	21

8 The following summary details were calculated from a study to find a relationship between mathematics exam marks and English exam marks from the results of 120 Year 12 students.

Mean mathematics exam mark = 64%

Mean English exam mark = 74%

Standard deviation of mathematics exam mark = 14.5%

Standard deviation of English exam mark = 9.8%

Correlation coefficient, $r = 0.64$

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

$$\text{Mathematics exam mark} = a + b \times \text{English exam mark}$$

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

b Calculate the value of b for the least-squares regression line (correct to 2 decimal places).

c Calculate the value of a for the least-squares regression line (correct to 2 decimal places).

d Use the regression line to predict the expected mathematics exam mark if a student scores 85% in an English exam (correct to the nearest percentage).

9 Find the least-squares regression equations, given the following summary data.

a	$\bar{x} = 5.6$	$s_x = 1.2$	$\bar{y} = 110.4$	$s_y = 5.7$	$r = 0.7$
b	$\bar{x} = 110.4$	$s_x = 5.7$	$\bar{y} = 5.6$	$s_y = 1.2$	$r = -0.7$
c	$\bar{x} = 25$	$s_x = 4.2$	$\bar{y} = 10200$	$s_y = 250$	$r = 0.88$
d	$\bar{x} = 10$	$s_x = 1$	$\bar{y} = 20$	$s_y = 2$	$r = -0.5$

10 Repeat questions **5**, **6** and **7**, collecting the values for $\bar{x}$, s_x, $\bar{y}$, s_y and r from CAS. Use these data to find the least-squares regression equation. Compare your answers to the ones obtained earlier from questions **5**, **6** and **7**. What do you notice?

11 A mathematician is interested in the behaviour patterns of her kitten, and collects the following data on two variables. Help her manipulate the data.

x	1	2	3	4	5	6	7	8	9	10
y	20	18	16	14	12	10	8	6	4	2

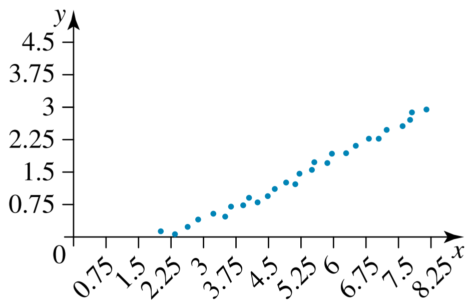
a Fit a least-squares regression line.

b Comment on any interesting features of this line.

c Now fit the 'opposite regression line', namely:

x	20	18	16	14	12	10	8	6	4	2
y	1	2	3	4	5	6	7	8	9	10

12 The best estimate of the least-squares regression line for the scatterplot is:



A $y = 2x$

B $y = \frac{1}{2}x$

C $y = 2 + \frac{1}{2}x$

D $y = -2 + \frac{1}{2}x$

E $y = -1 + \frac{1}{2}x$

13 The life span of adult males in a certain country over the last 220 years has been recorded.

Year	1780	1800	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
Life span (years)	51.2	52.4	51.7	53.2	53.1	54.7	59.9	62.7	63.2	66.8	72.7	79.2

a Fit a least-squares regression line to these data.
b Plot the data and the regression line on a scatterplot.
c Do the data really look linear? Discuss.

14 The price of a long distance telephone call changes as the duration of the call increases. The cost of a sample of calls from Melbourne to Slovenia are summarised in the table.

Cost of call ($)	1.25	1.85	2.25	2.50	3.25	3.70	4.30	4.90	5.80
Duration of call (seconds)	30	110	250	260	300	350	420	500	600

Cost of call ($)	7.50	8.00	9.25	10.00	12.00	13.00	14.00	16.00	18.00
Duration of call (seconds)	840	1000	1140	1200	1500	1860	2400	3600	7200

a What is the explanatory variable likely to be?
b Fit a least-squares regression line to the data.
c View the data on a scatterplot and comment on the reliability of the regression line in predicting the cost of telephone calls. (That is, consider whether the regression line you found proves that costs of calls and duration of calls are related.)
d Calculate the coefficient of determination and comment on the linear relationship between duration and cost of call.

MASTER

15 In a study to find a relationship between the height of plants and the hours of daylight they were exposed to, the following summary details were obtained.

Mean height of plants = 40 cm

Mean hours of daylight = 8 hours

Standard deviation of plant height = 5 cm

Standard deviation of daylight hours = 3 hours

Pearson's correlation coefficient = 0.9

The most appropriate regression equation is:

A height of plant (cm) = −13.6 + 0.54 × hours of daylight
B height of plant (cm) = −8.5 + 0.34 × hours of daylight
C height of plant (cm) = 2.1 + 0.18 × hours of daylight
D height of plant (cm) = 28.0 + 1.50 × hours of daylight
E height of plant (cm) = 35.68 + 0.54 × hours of daylight

16 Consider the following data set.

x	1	2	3	4	5	6
y	12	16	17	21	25	29

a Perform a least-squares regression on the first two points only.
b Now add the 3rd point and repeat.
c Repeat for the 4th, 5th and 6th points.
d Comment on your results.

3.7 Interpretation, interpolation and extrapolation

Interpreting slope and intercept (*b* and *a*)

studyon

Unit 3
AOS DA
Topic 7
Concept 2

Interpretation of slope and intercepts
Concept summary
Practice questions

Once you have a linear regression line, the **slope** and **intercept** can give important information about the data set.

The slope (b) indicates the change in the response variable as the explanatory variable increases by 1 unit.

The y-intercept indicates the value of the response variable when the explanatory variable = 0.

WORKED EXAMPLE 9

In the study of the growth of a species of bacterium, it is assumed that the growth is linear. However, it is very expensive to measure the number of bacteria in a sample. Given the data listed, find:

a the equation, describing the relationship between the two variables
b the rate at which bacteria are growing
c the number of bacteria at the start of the experiment.

Day of experiment	1	4	5	9	11
Number of bacteria	500	1000	1100	2100	2500

THINK	WRITE
a 1 Find the equation of the least-squares regression line using CAS.	a
2 Replace x and y with the variables in question.	Number of bacteria = 202.5 + 206.25 × day of experiment

b The rate at which bacteria are growing is given by the gradient of the least-squares regression.	**b** b is 206.25, hence on average, the number of bacteria increases by approximately 206 per day.
c The number of bacteria at the start of the experiment is given by the y-intercept of the least-squares regression line.	**c** The y-intercept is 202.5, hence the initial number of bacteria present was approximately 203.

studyon

Unit 3

AOS DA

Topic 7

Concept 3

Interpolation and extrapolation
Concept summary
Practice questions

Interpolation and extrapolation

As we have already observed, any linear regression method produces a linear equation in the form:

$$y = a + bx$$

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

This equation can be used to 'predict' the y-value for a given value of x. Of course, these are only approximations, since the regression line itself is only an estimate of the 'true' relationship between the bivariate data. However, they can still be used, in some cases, to provide additional information about the data set (that is, make predictions).

There are two types of prediction: **interpolation** and **extrapolation**.

Interpolation

Interpolation is the use of the regression line to predict values within the range of data in a set, that is, the values that are *in between the values* already in the data set. If the data are highly linear (r near +1 or −1) then we can be confident that our interpolated value is quite accurate. If the data are not highly linear (r near 0) then our confidence is duly reduced. For example, medical information collected from a patient every third day would establish data for day 3, 6, 9, . . . and so on. After performing regression analysis, it is likely that an interpolation for day 4 would be accurate, given a high r value.

Extrapolation

Extrapolation is the use of the regression line to predict values outside the range of data in a set, that is, values that are *smaller than the smallest value* already in the data set or *larger than the largest value*.

Two problems may arise in attempting to extrapolate from a data set. Firstly, it may not be reasonable to extrapolate too far away from the given data values. For example, suppose there is a weather data set for 5 days. Even if it is highly linear (r near +1 or −1) a regression line used to predict the same data 15 days in the future is highly risky. Weather has a habit of randomly fluctuating and patterns rarely stay stable for very long.

Secondly, the data may be highly linear in a narrow band of the given data set. For example, there may be data on stopping distances for a train at speeds of between 30 and 60 km/h. Even if they are highly linear in this range, it is unlikely that things are similar at very low speeds (0–15 km/h) or high speeds (over 100 km/h).

Generally, one should feel *more confident about the accuracy of a prediction derived from interpolation* than one derived from extrapolation. Of course, it still depends upon the correlation coefficient (r). The closer to linearity the data are, the more confident our predictions in all cases.

Interpolation is the use of the regression line to predict values within the range of data in a set.

Extrapolation is the use of the regression line to predict values outside the range of data in a set.

WORKED EXAMPLE 10

Using interpolation and the following data set, predict the height of an 8-year-old girl.

Age (years)	1	3	5	7	9	11
Height (cm)	60	76	115	126	141	148

THINK	WRITE
1 Find the equation of the least-squares regression line using your calculator. (Age is the explanatory variable and height is the response one.)	$y = 55.63 + 9.23x$
2 Replace x and y with the variables in question.	Height = 55.63 + 9.23 × age
3 Substitute 8 for age into the equation and evaluate.	When age = 8, Height = 55.63 + 9.23 × 8 = 129.5 (cm)
4 Write the answer.	At age 8, the predicted height is 129.5 cm.

WORKED EXAMPLE 11

Use extrapolation and the data from Worked example 10 to predict the height of the girl when she turns 15. Discuss the reliability of this prediction.

THINK	WRITE
1 Use the regression equation to calculate the girl's height at age 15.	Height = 55.63 + 9.23 × age = 55.63 + 9.23 × 15 = 194.08 cm
2 Analyse the result.	Since we have extrapolated the result (that is, since the greatest age in our data set is 11 and we are predicting outside the data set) we cannot claim that the prediction is reliable.

EXERCISE 3.7 Interpretation, interpolation and extrapolation

PRACTISE

1 WE9 A study on the growth in height of a monkey in its first six months is assumed to be linear. Given the data shown, find:

a the equation, describing the relationship between the two variables

b the rate at which the monkey is growing

c the height of the monkey at birth.

Month from birth	1	2	3	4	5	6
Height (cm)	15	19	23	27	30	32

2 The outside temperature is assumed to increase linearly with time after 6 am. Given the data shown, find:

a the equation, describing the relationship between the two variables
b the rate at which the temperature is increasing
c the temperature at 6 am.

Hours after 6 am	0.5	1.5	3	3.5	5
Temperature (°C)	15	18	22	23	28

3 WE10 Using interpolation and the following data set, predict the height of a 10-year-old boy.

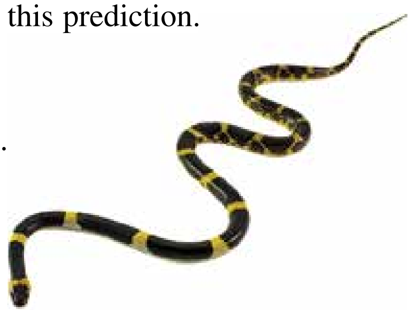
Age (years)	1	3	4	8	11	12
Height (cm)	65	82	92	140	157	165

4 Using interpolation and the following data set, predict the length of Matt's pet snake when it is 15 months old.

Age (months)	1	3	5	8	12	18
Length (cm)	48	60	71	93	117	159

5 WE11 Use extrapolation and the data from question 3 to predict the height of the boy when he turns 16. Discuss the reliability of this prediction.

6 Use extrapolation and the data from question 4 to predict the length of Matt's pet snake when it is 2 years old. Discuss the reliability of the prediction.



CONSOLIDATE

7 A drug company wishes to test the effectiveness of a drug to increase red blood cell counts in people who have a low count. The following data are collected.

Day of experiment	4	5	6	7	8	9
Red blood cell count	210	240	230	260	260	290

Find:

a the equation, describing the relationship between the variables in the form $y = a + bx$
b the rate at which the red blood cell count was changing
c the red blood cell count at the beginning of the experiment (that is, on day 0).

8 A wildlife exhibition is held over 6 weekends and features still and live displays. The number of live animals that are being exhibited varies each weekend. The number of animals participating, together with the number of visitors to the exhibition each weekend, is as shown.

Number of animals	6	4	8	5	7	6
Number of visitors	311	220	413	280	379	334

Find:

a the rate of increase of visitors as the number of live animals is increased by 1

b the predicted number of visitors if there are no live animals.

9 An electrical goods warehouse produces the following data showing the selling price of electrical goods to retailers and the volume of those sales.

Selling price ($)	60	80	100	120	140	160	200	220	240	260
Sales volume (× 1000)	400	300	275	250	210	190	150	100	50	0

Perform a least-squares regression analysis and discuss the meaning of the gradient and y-intercept.

10 A study of the dining-out habits of various income groups in a particular suburb produces the results shown in the table.

Weekly income ($)	100	200	300	400	500	600	700	800	900	1000
Number of restaurant visits per year	5.8	2.6	1.4	1.2	6	4.8	11.6	4.4	12.2	9

Use the data to predict:

a the number of visits per year by a person on a weekly income of $680

b the number of visits per year by a person on a weekly income of $2000.

11 Fit a least-squares regression line to the following data.

x	0	1	2	4	5	6	8	10
y	2	3	7	12	17	21	27	35

Find:

a the regression equation

b y when $x = 3$

c y when $x = 12$

d x when $y = 7$

e x when $y = 25$.

f Which of b to e above are extrapolations?

12 The following table represents the costs for shipping a consignment of shoes from Melbourne factories. The cost is given in terms of distance from Melbourne. There are two factories that can be used. The data are summarised in the table.

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

a Find the least-squares regression equation for each factory.

b Which factory is likely to have the lowest cost to ship to a shop in Melbourne (i.e. distance from Melbourne = 0 km)?

c Which factory is likely to have the lowest cost to ship to Mytown, 115 kilometres from Melbourne?

d Which factory has the most 'linear' shipping rates?

13 A factory produces calculators. The least-squares regression line for cost of production (C) as a function of numbers of calculators (n) produced is given by:

$$C = 600 + 7.76n$$

Furthermore, this function is deemed accurate when producing between 100 and 1000 calculators.

a Find the cost to produce 200 calculators.

b How many calculators can be produced for $2000?

c Find the cost to produce 10 000 calculators.

d What are the 'fixed' costs for this production?

e Which of a to c above is an interpolation?

14 A study of the relationship between IQ and results in a mathematics exam produced the following results. Unfortunately, some of the data were lost. Copy and complete the table by using the least-squares equation with the data that were supplied.

Note: Only use (x, y) pairs if both are in the table.

IQ	80		92	102	105		107	111	115	121
Test result (%)	56	60	68	65		74	71	73		92

15 The least-squares regression line for a starting salary (s) as a function of number of years of schooling (n) is given by the rule: $s = 37\,000 + 1800n$.

a Find the salary for a person who completed 10 years of schooling.

b Find the salary for a person who completed 12 years of schooling.

c Find the salary for a person who completed 15 years of schooling.

d Mary earned $60 800. What was her likely schooling experience?

e Discuss the reasonableness of predicting salary on the basis of years of schooling.

16 Fit a least-squares regression to the following data.

q	0	1	3	7	10	15
r	12	18	27	49	64	93

Find:

a the regression equation.

b r when $q = 4$

c r when $q = 18$

d q when $r = 100$.

e Which of b to d is extrapolation?

MASTER

17 A plumbing company's charges follow the least-squares regression line:

$$C = 180 + 80n$$

where C is the total cost and n is the number of hours of work. This function is accurate for a single 8 hour day.

a Find the total cost if the plumber worked for 3 hours.

b If the total charge was $1250, how long did the plumber work, correct to 2 decimal places?

c Find the total cost if the plumber worked 9 hours and 30 minutes.

d What is the call-out fee (the cost to come out before they start any work)?

e Which of a to c is an extrapolation?

18 A comparison was investigated between AFL memberships sold and the amount of money spent on advertising by the club.

Advertising (in millions $)	2.3	1.8	1.2	0.8	1.6	0.6	1.0
Members	81 363	67 947	58 846	55 597	62 295	54 946	57 295

a Find the least-squares regression equation. Round the coefficient to the nearest whole number.

b Using the least-squares equation if $2 million was spent on advertising, how many members would you expect to have? Is this extrapolation or interpolation?

c If you wanted 70 000 members, how much would you expect to have to pay on advertising?

d Calculate the coefficient of determination and use it to explain the association between membership numbers and the amount of money spent on advertising.

3.8 Residual analysis

studyon

Unit 3

AOS DA

Topic 7

Concept 5

Residual analysis

Concept summary

Practice questions

There are situations where the mere fitting of a regression line to some data is not enough to convince us that the data set is *truly* linear. Even if the correlation is close to +1 or −1 it still may not be convincing enough.

The next stage is to analyse the **residuals**, or deviations, of each data point from the straight line. A residual is the vertical difference between each data point and the regression line.

Calculating residuals

A sociologist gathers data on the heights of brothers and sisters in families from different cultural backgrounds. He enters his records in the table shown.

x	2	3	5	8	9	9
y	3	7	12	10	12	16

He then plots each point, and fits a regression line as shown in Figure 1, which follows. He then decides to calculate the residuals.

The *residuals* are simply the vertical distances from the line to each point. These lines are shown as green and pink bars in Figure 2.

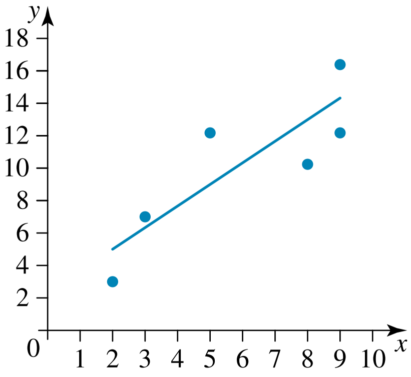


Figure 1

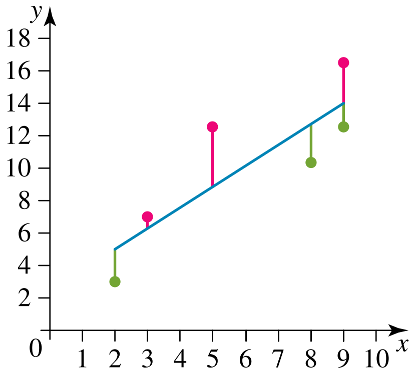


Figure 2

Finally, he calculates the residuals for each data point. This is done in two steps.

Step 1. He calculates the *predicted* value of y by using the regression equation.

Step 2. He calculates the *difference* between this predicted value and the original value.

A residual = actual y-value − predicted y-value

WORKED EXAMPLE 12

Consider the data set shown. Find the equation of the least-squares regression line and calculate the residuals.

x	1	2	3	4	5	6	7	8	9	10
y	5	6	8	15	24	47	77	112	187	309

THINK

1 Find the equation of a least-squares regression line using a calculator.

2 Use the equation of the least-squares regression line to calculate the predicted y-values (these are labelled as y_{pred}) for every x-value in the table. That is, substitute each x-value into the equation and evaluate record results in the table.

3 Calculate residuals for each point by subtracting predicted y-values from the actual y-value. (That is, residual = observed y-value − predicted y-value). Record results in the table.

WRITE

$y = -78.7 + 28.7x$

x-values	1	2	3	4	5
y-values	5.0	6.0	8.0	15.0	24.0
Predicted y-values	−50.05	−21.38	7.3	35.98	64.66
Residuals ($y - y_{\text{pred}}$)	55.05	27.38	0.7	−20.98	−40.66

x-values	6	7	8	9	10
y-values	47.0	77.0	112.0	187.0	309.0
Predicted y-values	93.34	122.02	150.7	179.38	208.06
Residuals ($y - y_{\text{pred}}$)	−46.34	−45.02	−38.7	7.62	100.94

Notes

1. The residuals may be determined by ($y - y_{\text{pred}}$); that is, the actual values minus the predicted values.
2. The *sum* of all the residuals *always* adds to 0 (or very close to 0 after rounding), when least-squares regression is used. This can act as a check for our calculations.

Introduction to residual analysis

As we observed in Worked example 12, there is not really a good fit between the data and the least-squares regression line; however, there seems to be a pattern in the residuals. How can we observe this pattern in more detail?

The answer is to plot the *residuals* themselves against the *original x-values*. If there is a pattern, it should become clearer after they are plotted.

Types of residual plots

studyon

Unit 3

AOS DA

Topic 7

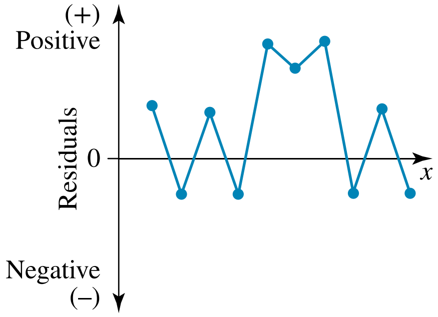
Concept 6

Residual plots
Concept summary
Practice questions

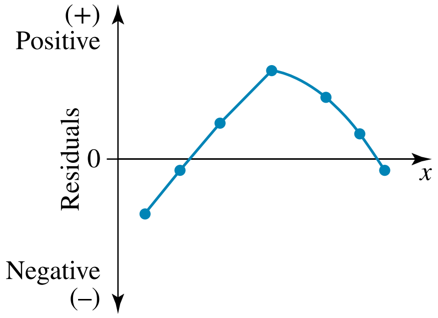
There are three basic types of **residual plots**. Each type indicates whether or not a linear relationship exists between the two variables under investigation.

Note: The points are joined together to see the patterns more clearly.

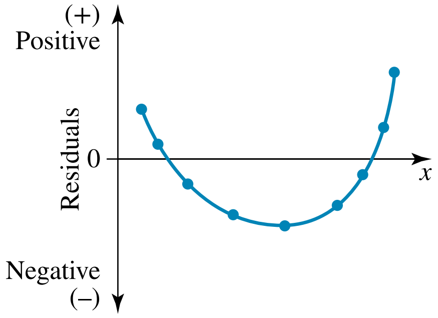
The points of the residuals are randomly scattered above and below the x-axis. The original data probably have a *linear* relationship.



The points of the residuals show a curved pattern (∩), with a series of negative, then positive and back to negative residuals along the x-axis. The original data probably have a **non-linear relationship**. Transformation of the data may be required.



The points of the residuals show a curved pattern (∪), with a series of positive, then negative and back to positive residuals along the x-axis. The original data probably have a *non-linear* relationship. Transformation of the data may be required.



The transformation of data suggested in the last two residual plots will be studied in more detail in the next section.

WORKED EXAMPLE 13

Using the same data as in Worked example 12, plot the residuals and discuss the features of the residual plot.

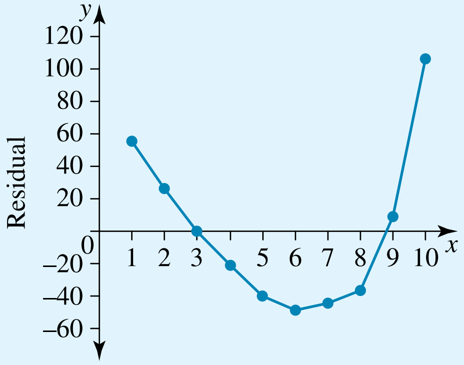
THINK

WRITE/DRAW

1 Generate a table of values of residuals against x.

x-values	1	2	3	4	5
Residuals ($y - y_{\text{pred}}$)	55.05	27.38	0.7	−20.98	−40.66
x-values	6	7	8	9	10
Residuals ($y - y_{\text{pred}}$)	−46.34	−45.02	−38.7	7.62	100.94

2 Plot the residuals against x. To see the pattern clearer, join the consecutive points with straight line segments.



3 If the relationship was linear the residuals would be scattered randomly above and below the line. However, in this instance there is a pattern which looks somewhat like a parabola. This should indicate that the data were not really linear, but were more likely to be quadratic. Comment on the residual plot and its relevance.

The residual plot indicates a distinct pattern suggesting that a non-linear model could be more appropriate.

EXERCISE 3.8 Residual analysis

PRACTISE

1 WE12 Consider the data set shown. Find the equation of the least-squares regression line and calculate the residuals.

x	1	2	3	4	5	6	7	8	9
y	12	20	35	40	50	67	83	88	93

2 From the data shown, find the equation of the least-squares regression line and calculate the residuals.

x	5	7	10	12	15	18	25	30	40
y	45	61	89	122	161	177	243	333	366

3 WE13 Using the same data from question 1, plot the residuals and discuss the features of the residual plot. Is your result consistent with the coefficient of determination?

4 Using the same data from question 2, plot the residuals and discuss the features of the residual plot. Is your result consistent with the coefficient of determination?

CONSOLIDATE

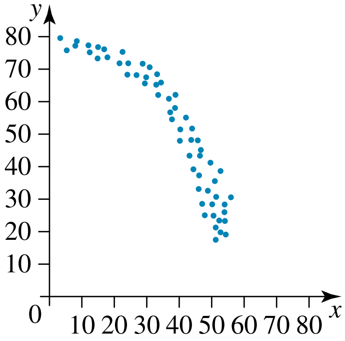
5 Find the residuals for the following data.

x	1	2	3	4	5	6
y	1	9.7	12.7	13.7	14.4	14.5

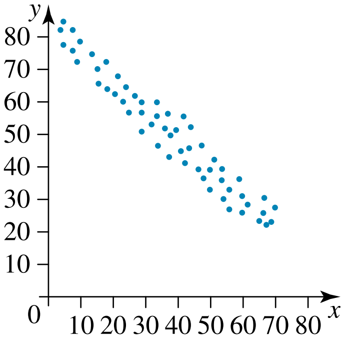
6 For the results of question 5, plot the residuals and discuss whether the relationship between x and y is linear.

7 Which of the following scatterplots shows linear relationship between the variables?

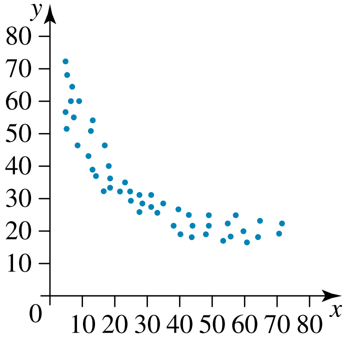
i



ii



iii



A All of them
B None of them
C i and iii only
D ii only
E ii and iii only

8 Consider the following table from a survey conducted at a new computer manufacturing factory. It shows the percentage of defective computers produced on 8 different days after the opening of the factory.

Day	2	4	5	7	8	9	10	11
Defective rate (%)	15	10	12	4	9	7	3	4

a The results of least-squares regression were: $b = -1.19$, $a = 16.34$, $r = -0.87$. Given $y = a + bx$, use the above information to calculate the predicted defective rates (y_{pred}).
b Find the residuals ($y - y_{\text{pred}}$).
c Plot the residuals and comment on the likely linearity of the data.
d Estimate the defective rate after the first day of the factory's operation.
e Estimate when the defective rate will be at zero. Comment on this result.

9 The following data represent the number of tourists booked into a hotel in central Queensland during the first week of a drought. (Assume Monday = 1.)

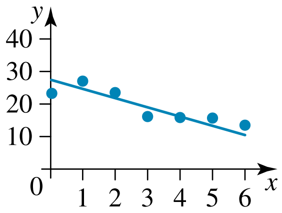
Day	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Sun.
Bookings in hotel	158	124	74	56	31	35	22

The results of least-squares regression were:
$b = -22.5$, $a = 161.3$, $r = -0.94$, where $y = a + bx$.

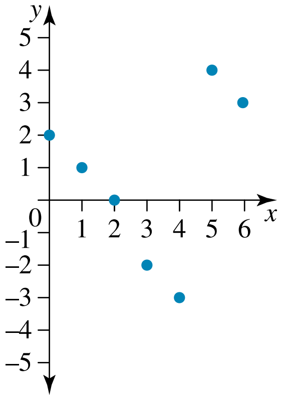
a Find the predicted hotel bookings (y_{pred}) for each day of the week.
b Find the residuals ($y - y_{\text{pred}}$).
c Plot the residuals and comment on the likely linearity of the data.
d Would this regression line be a typical one for this hotel?

10 A least-squares regression is fitted to the points shown in the scatterplot.

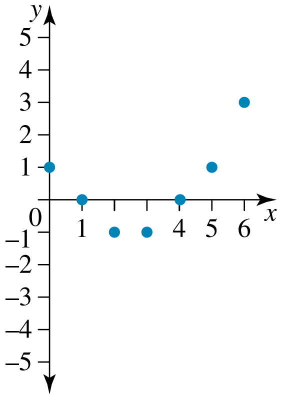
Which of the following looks most similar to the residual plot for the data?



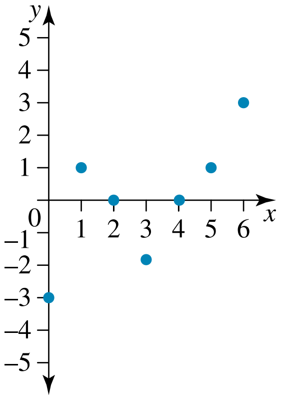
A



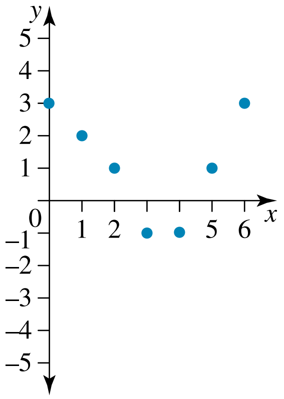
B



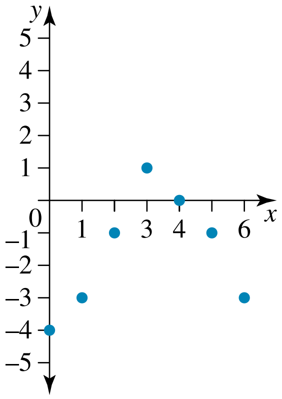
C



D



E



11 From each table of residuals, decide whether or not the relationship between the variables is likely to be linear.

a

x	y	Residuals
1	2	−1.34
2	4	−0.3
3	7	−0.1
4	11	0.2
5	21	0.97
6	20	2.3
7	19	1.2
8	15	−0.15
9	12	−0.9
10	6	−2.8

b

x	y	Residuals
23	56	0.12
21	50	−0.56
19	43	1.30
16	41	0.20
14	37	−1.45
11	31	2.16
9	28	−0.22
6	22	−3.56
4	19	2.19
3	17	−1.05

c

x	y	Residuals
1.2	23	0.045
1.6	25	0.003
1.8	24	−0.023
2.0	26	−0.089
2.2	28	−0.15
2.6	29	−0.98
2.7	34	−0.34
2.9	42	−0.01
3.0	56	0.45
3.1	64	1.23

12 Consider the following data set.

x	0	1	2	3	4	5	6	7	8	9	10
y	1	4	15	33	60	94	134	180	240	300	390

a Plot the data and fit a least-squares regression line.

b Find the correlation coefficient and interpret its value.

c Calculate the coefficient of determination and explain its meaning.

d Find the residuals.

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

13 Find the residuals for the following data set.

m	12	37	35	41	55	69	77	90
P	2.5	21.6	52.3	89.1	100.7	110.3	112.4	113.7

14 For the data in question 13, plot the residuals and comment whether the relationship between x and y is linear.

MASTER

15 Calculate the residuals of the following data.

k	1.6	2.5	5.9	7.7	8.1	9.7	10.3	15.4
D	22.5	37.8	41.5	66.9	82.5	88.7	91.6	120.4

16 For the data in question 15:

a plot the residuals and comment whether the relationship between x and y is linear.

b calculate the coefficient of determination and explain is meaning.

3.9 Transforming to linearity

studyon

Unit 3
AOS DA
Topic 7
Concept 7

Transformation of data
Concept summary
Practice questions

Although linear regression might produce a 'good' fit (high r value) to a set of data, the data set may still be non-linear. To remove (as much as is possible) such *non-linearity*, the data can be transformed.

Either the x-values, y-values, or both may be transformed in some way so that the transformed data are more linear. This enables more accurate predictions (extrapolations and interpolations) from the regression equation. In Further Mathematics, six transformations are studied:

Logarithmic transformations:	y **versus** $\log_{10}(x)$	$\log_{10}(y)$ **versus** x
Quadratic transformations:	y **versus** x^2	y^2 **versus** x
Reciprocal transformations:	y **versus** $\frac{1}{x}$	$\frac{1}{y}$ **versus** x

eBook*plus*

Interactivity
Transforming to linearity
int-6253

Choosing the correct transformations

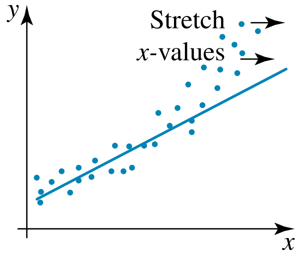
To decide on an appropriate transformation, examine the points on a scatterplot with high values of x and/or y (that is, away from the origin) and decide for each axis whether it needs to be stretched or compressed to make the points line up. The best way to see which of the transformations to use is to look at a number of 'data patterns'.

eBook*plus*

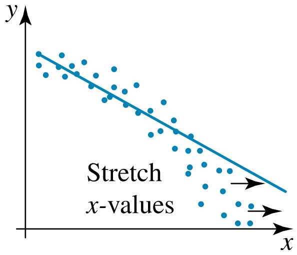
Interactivity
Linearising data
int-6491

Quadratic transformations

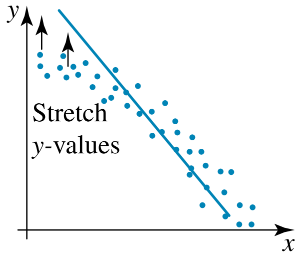
1. Use y versus x^2 transformation.



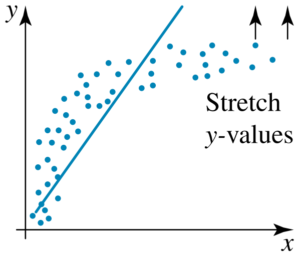
2. Use y versus x^2 transformation.



3. Use y^2 versus x transformation.

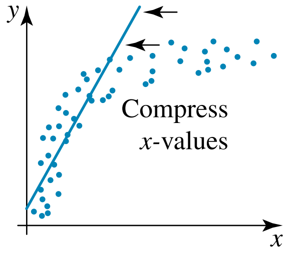


4. Use y^2 versus x transformation.

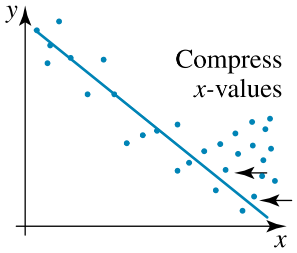


Logarithmic and reciprocal transformations

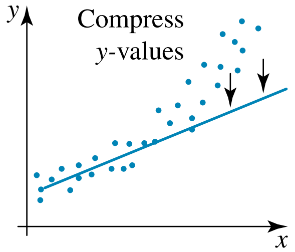
1. Use y versus $\log_{10}(x)$ or y versus $\frac{1}{x}$ transformation.



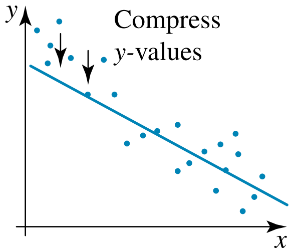
2. Use y versus $\log_{10}(x)$ or y versus $\frac{1}{x}$ transformation.



3. Use $\log_{10}(y)$ versus x or $\frac{1}{y}$ versus x transformation.



4. Use $\log_{10}(y)$ versus x or $\frac{1}{y}$ versus x transformation.



Testing transformations

As there are at least two possible transformations for any given non-linear scatterplot, the decision as to which is the best comes from the coefficient of correlation. The least-squares regression equation that has a Pearson correlation coefficient closest to 1 or −1 should be considered as the most appropriate. However, there may be very little difference so common sense needs to be applied. It is sometimes more useful to use a linear function rather than one of the six non-linear functions.

WORKED EXAMPLE 14

a **Plot the following data on a scatterplot, consider the shape of the graph and apply a quadratic transformation.**

b **Calculate the equation of the least-squares regression line for the transformed data.**

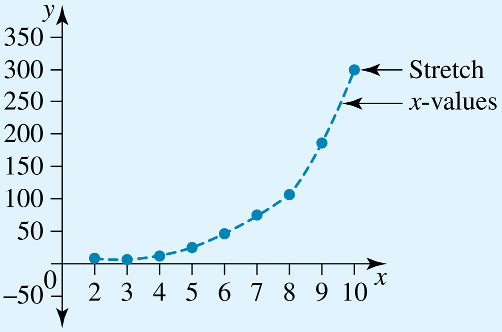
x	1	2	3	4	5	6	7	8	9	10
y	5	6	8	15	24	47	77	112	187	309

THINK

WRITE/DRAW

1 Plot the data to check that a quadratic transformation is suitable.

Looking at the shape of the graph, the best option is to stretch the x-axis. This requires an x^2 transformation.



2 Square the x-values to give a transformed data set by using CAS.

x^2	1	4	9	16	25	36	49	61	81	100
y	5	6	8	15	24	47	77	112	187	309

3 Find the equation of the least-squares regression line for the transformed data.

Using CAS:

y-intercept (a) = −28.0

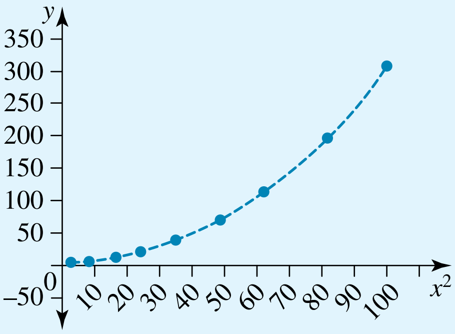
gradient (b) = 2.78

correlation (r) = 0.95.

$y = a + bx$
$y = -28.0 + 2.78x_{\text{T}}$ where $x_{\text{T}} = x^2$; that is,
$y = -28.0 + 2.78x^2$

4 Plot the new transformed data.

Note: These data are still not truly linear, but are 'less' parabolic. Perhaps another transformation would improve things even further. This could involve transforming the y-values, such as $\log_{10}(y)$, and applying another linear regression.



WORKED EXAMPLE 15

a Tranform the data by applying a logarithmic transformation to the *y*-variable.

b Calculate the equation of the least squares regression line for the transformed data.

c Comment on the value of *r*.

Time after operation (h)	x	1	2	3	4	5	6	7	8
Heart rate (beats/min)	y	100	80	65	55	50	51	48	46

THINK

WRITE

a Transform the y data by calculating the log of y-values or, in this problem, the log of heart rate.

a

Time	x	1	2	3	4	5	6	7	8
log (heart rate)	log y	2	1.903	1.813	1.740	1.694	1.708	1.681	1.663

b 1 Use a calculator to find the equation of least-squares regression line for x and log y.

b $\log_{10}(y) = 1.98 - 0.05x$

2 Rewrite the equation in terms of the variables in question.

$\log_{10}$ (heart rate) $= 1.98 - 0.05 \times$ time (i.e. time = number of hours after the operation.)

c State the value of r and comment on the result.

c $r = -0.93$

There is a slight improvement of the correlation coefficient that resulted from applying logarithmic transformation.

Further investigation

Often all appropriate transformations need to be performed to choose the best one. Extend Worked example 15 by compressing the y data using the reciprocals of the y data or even compress the x data. Go back to the steps for transforming the data. Did you get a better r value and thus a more reliable line of best fit? (*Hint:* The best transformation gives $r = -0.98$.)

Using the transformed line for predictions

Once the appropriate model has been established and the equation of least-squares regression line has been found, the equation can be used for predictions.

WORKED EXAMPLE 16

a Using CAS, apply a reciprocal transformation to the following data.

b Use the transformed regression equation to predict the number of students wearing a jumper when the temperature is 12 °C.

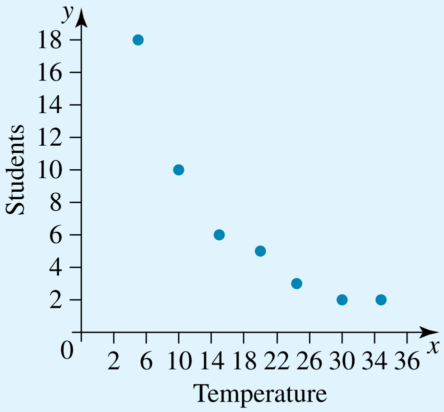
Temperature (°C)	x	5	10	15	20	25	30	35
Number of students in a class wearing jumpers	y	18	10	6	5	3	2	2

THINK | **WRITE/DRAW**

a **1** Construct the scatterplot. Temperature is the explanatory variable, while the number of students wearing jumpers is the response one.

Therefore, put *temperature* on the horizontal axis and *students* on the vertical axis.

a



2 The x-values should be compressed, so it may be appropriate to transform the x-data by calculating the reciprocal of temperature. Reciprocate each x-value $\left(\text{that is, find } \frac{1}{x}\right)$.

$\frac{1}{\text{Temperature}}$	$\frac{1}{x}$	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{15}$	$\frac{1}{20}$	$\frac{1}{25}$	$\frac{1}{30}$	$\frac{1}{35}$
Number of students wearing jumpers	y	18	10	6	5	3	2	2

3 Use CAS to find the equation of the least-squares regression line for $\frac{1}{x}$ and y.

$y = -0.4354 + 94.583x_T$, where $x_T = \frac{1}{x}$ or

$y = -0.4354 + \frac{94.583}{x}$

4 Replace x and y with the variables in question.

The number of students in class wearing jumpers

$= -0.4354 + \frac{94.583}{\text{Temperature}}$

b **1** Substitute 12 for x into the equation of the regression line and evaluate.

b Number of students wearing jumpers

$= -0.4354 + \frac{94.583}{\text{Temperature}}$

$= -0.4354 + \frac{94.583}{12}$

$= 7.447$

2 Write your answer to the nearest whole number.

7 students are predicted to wear jumpers when the temperature is 12 °C.

EXERCISE 3.9 Transforming to linearity

PRACTISE

1 WE14 **a** Plot the data on a scatterplot, consider the shape of the graph and apply a quadratic transformation.

b Calculate the equation of the least-squares regression line for the transformed data.

x	1	2	3	4	5	6	7	8	9
y	12	19	29	47	63	85	114	144	178

2 a Plot the following data on a scatterplot, consider the shape of the graph and apply a quadratic transformation.

b Calculate the equation of the least-squares regression line for the transformed data.

x	3	5	9	12	16	21	24	33
y	5	12	38	75	132	209	291	578

3 WE15 Apply a logarithmic transformation to the following data, which represents a speed of a car as a function of time, by transforming the y-variable.

Time (s)	1	2	3	4	5	6	7	8
Speed (ms^{-1})	90	71	55	45	39	35	32	30

4 Apply a logarithmic transformation to the following data by transforming the y-variable.

x	5	10	15	20	25	30	35	40
y	1000	500	225	147	99	70	59	56

5 WE16 a Using CAS, apply a reciprocal transformation to the x-variable of the following data.

Time after 6 pm (h)	1	2	3	4	5	6	7	8
Temperature (°C)	32	22	16	11	9	8	7	7

b Use the transformed regression equation to predict the temperature at 10.30 pm.

6 a Using CAS, apply a reciprocal transformation to the x-variable of the following data.

x	2	5	7	9	10	13	15	18
y	120	50	33	15	9	5	2	1

b Use the transformed regression equation to predict y when $x = 12$.

c Use the transformed regression equation to predict x when $y = 20$.

CONSOLIDATE

7 Apply a quadratic (x^2) transformation to the following data set. The regression line has been determined as $y = 186 - 27.7x$ with $r = -0.91$.

x	2	3	4	5	7	9
y	96	95	92	90	14	−100

8 The *average* heights of 50 girls of various ages were measured as follows.

Age group (years)	9	10	11	12	13	14	15	16	17	18
Average height (cm)	128	144	148	154	158	161	165	164	166	167

The original linear regression yielded:

$$\text{Height} = 104.7 + 3.76 \times \text{age, with } r = 0.92.$$

a Plot the original data and regression line.

b Apply a $\log_{10}(x)$ transformation.

c Perform regression analysis on the transformed data and comment on your results.

9 a Use the transformed data from question 8 to predict the heights of girls of the following ages:

i 7 years old

ii 10.5 years old

iii 20 years old.

b Which of the predictions in part a were obtained by interpolating?

10 Comment on the suitability of transforming the data of question 8 in order to improve predictions of heights for girls under 8 years old or over 18.

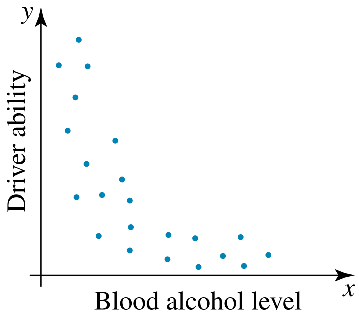
11 a Apply a reciprocal transformation to the following data obtained by a physics student studying light intensity.

Distance from light source (metres)	1	2	3	4	5	10
Intensity (candlepower)	90	60	28	22	20	12

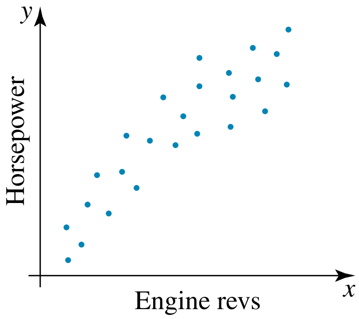
b Use the transformed regression equation to predict the intensity at a distance of 20 metres.

12 For each of the following scatterplots suggest an appropriate transformation(s).

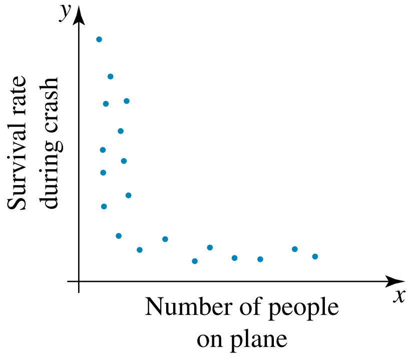
a



b



c



13 Use the equation $y = -12.5 + 0.2x^2$, found after transformation, to predict values of y for the given x-value (correct to 2 decimal places):

a $x = 2.5$ b $x = -2.5$.

14 Use the equation $y = -25 + 1.12 \log_{10}(x)$, found after transformation, to predict values of y for the given x-value (correct to 2 decimal places):

a $x = 2.5$ b $x = -2.5$ c $x = 0$.

15 Use the equation $\log_{10}(y) = 0.03 + 0.2x$, found after transformation, to predict values of y for the given x-value (correct to 2 decimal places):

a $x = 2.5$ b $x = -2.5$.

16 Use the equation $\frac{1}{y} = 12.5 + 0.2x$, found after transformation, to predict values of y for the given x-value (correct to 2 decimal places):

a $x = 2.5$ b $x = -2.5$.

MASTER

17 The seeds in the sunflower are arranged in spirals for a compact head. Counting the number of seeds in the successive circles starting from the centre and moving outwards, the following number of seeds were counted.

Circle	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Number of seeds	3	5	8	13	21	34	55	89	144	233

a Plot the data and fit a least-squares regression line.

b Find the correlation coefficient and interpret its value.

c Using the equation of the regression line, predict the number of seeds in the 11th circle.

d Find the residuals.

e Construct the residual plot. Is the relation between the number of the circle and the number of seeds linear?

f What type of transformation could be applied to:

i the x-values? Explain why.

ii the y-values? Explain why.

18 Apply a $\log_{10}(y)$ transformation to the data used in question 17.

a Fit a least-squares regression line to the transformed data and plot it with the data.

b Find the correlation coefficient. Is there an improvement? Why?

c Find the equation of the least-squares regression line for the transformed data.

d Calculate the coefficient of determination and interpret its value.

e Using the equation of the regression line for the transformed data, predict the number of seeds for the 11th circle.

f How does this compare with the prediction from question 17?

eBook*plus* ONLINE ONLY

3.10 Review

www.jacplus.com.au

The Maths Quest Review is available in a customisable format for you to demonstrate your knowledge of this topic.

The Review contains:

- **Multiple-choice** questions — providing you with the opportunity to practise answering questions using CAS technology
- **Short-answer** questions — providing you with the opportunity to demonstrate the skills you have developed to efficiently answer questions using the most appropriate methods
- **Extended-response** questions — providing you with the opportunity to practise exam-style questions.

A summary of the key points covered in this topic is also available as a digital document.

REVIEW QUESTIONS

Download the Review questions document from the links found in the Resources section of your eBookPLUS.

eBook*plus* ONLINE ONLY

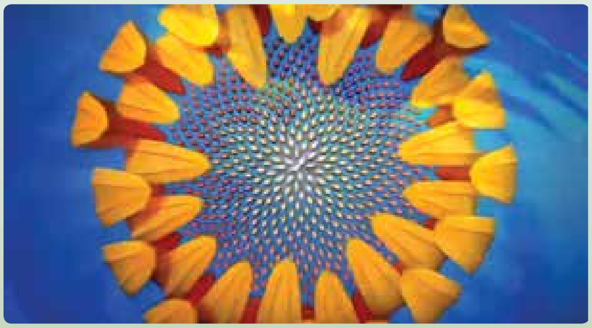
Activities

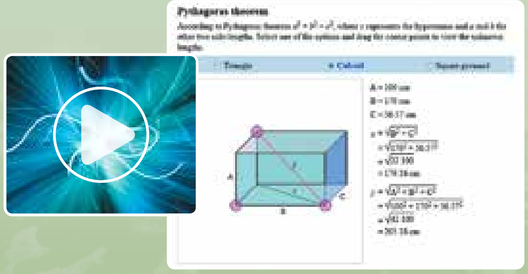
To access eBookPLUS activities, log on to

www.jacplus.com.au

Interactivities

A comprehensive set of relevant interactivities to bring difficult mathematical concepts to life can be found in the Resources section of your eBookPLUS.





studyON

studyON is an interactive and highly visual online tool that helps you to clearly identify strengths and weaknesses prior to your exams. You can then confidently target areas of greatest need, enabling you to achieve your best results.

3 Answers

EXERCISE 3.2

1 a Daily temperature = explanatory variable, air conditioners sold = response variable

b Not appropriate

2 a Size of the crowd = response variable, teams playing = explanatory variable

b Net score of a round of golf = response variable, golfer's handicap = explanatory variable

3 a Explanatory — age, response — salary

b Explanatory — amount of fertiliser, response — growth

c Not appropriate

d Not appropriate

e Explanatory — number in household, response — size of house

4 a Explanatory — month of the year, response — size of electricity bill

b Explanatory — number of hours, response — mark on the test

c Not appropriate

d Explanatory — season, response — cost

5 C

6 C

7 D

8 True

9 True

10 True

11 False

12 False

13 a Minutes on the court: explanatory variable

b Points scored: response variable

14 a Response variable: electricity bill

b Number of Christmas lights should be on the *x*-axis.

c Explanatory variable: number of Christmas lights

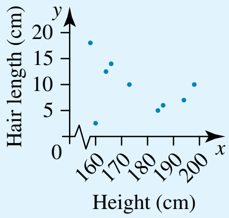
d Electricity bill should be on the *y*-axis.

EXERCISE 3.3

1 Moderate positive linear relationship

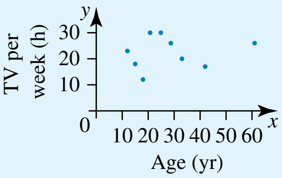
2 Moderate positive linear relationship

3



No relationship

4



No relationship

5 a Yes — positive association

b Yes — positive association

c Yes — positive association

d Yes — negative association

e Yes — positive association

f Yes — negative association

g No — no association

6 a Weak, negative association of linear form

b Moderate, negative association of linear form

c Moderate, positive association of linear form

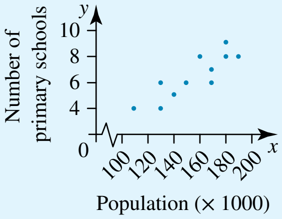
d Strong, positive association of linear form

e No association

f Non-linear association

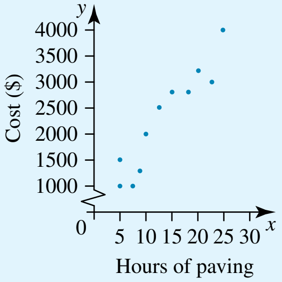
7 B

8



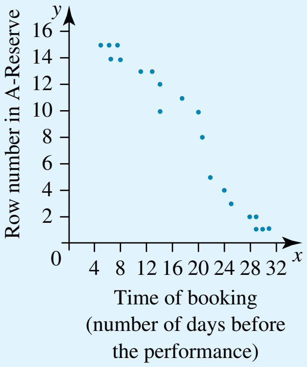
Moderate positive association of linear form, no outliers

9



Strong positive association of linear form, no outliers

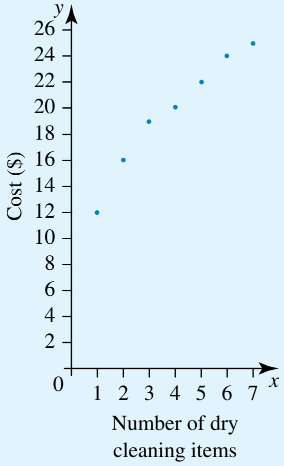
10



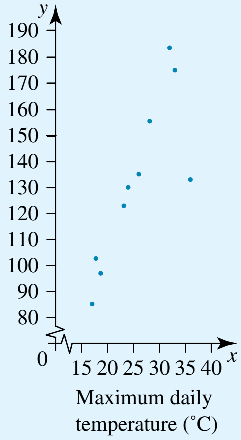
Strong negative association of linear form, no outliers

11 D

12 a, b



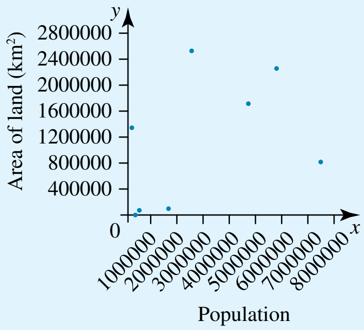
13 a, b



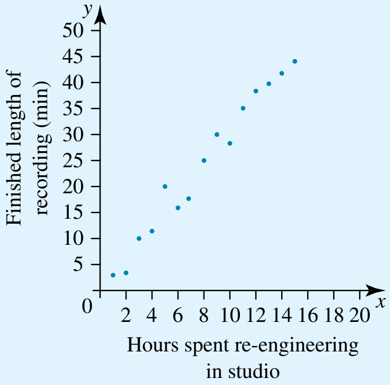
14 a There is a strong positive correlation between the dry cleaning items and the cost.

b There is a strong positive correlation between the maximum daily temperature and the number of drinks sold.

15 There appears to be no correlation between population and area of the various states and territories.



16 a



b There is a relationship. It is strong, positive and linear.

EXERCISE 3.4

1 a i $r = -0.9$

ii Moderate, negative, linear

b i $r = 0.7$

ii Strong, positive, linear

2 a Strong, positive, linear **b** Weak, negative, linear

3 a No association **b** Moderate positive

c Strong negative **d** Strong negative

4 a Strong positive **b** Strong positive

c Weak negative **d** No association

5 a i $r \approx -0.8$

ii Strong, negative, linear association

b i $r \approx 0.6$

ii Moderate, positive, linear association

c i $r \approx 0.2$

ii No linear association

d i $r \approx -0.2$

ii No linear association

6 a i $r = 1$

ii Perfect, positive, linear association

b i $r \approx 0.8$

ii Strong, positive, linear association

c i $r \approx 0$

ii No linear association

d i $r \approx -0.7$

ii Moderate, negative, linear association

7 B

8 E

9 a $r = 0.1$: no linear relationship

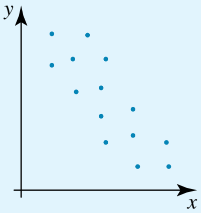
b $r = 0.2$: no linear relationship

c $r = 0.95$: strong, positive linear relationship

10 D

11 C

12

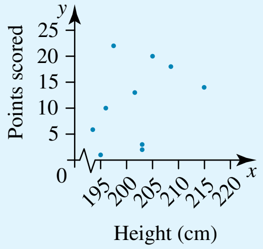


13 E

14 C

EXERCISE 3.5

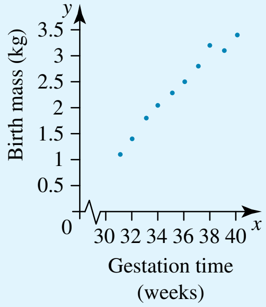
1 a



b No linear relationship, $r \approx 0.3$

c $r \approx 0.36$, weak, positive linear relationship

2 a



b Strong, positive, linear relationship. $r \approx 0.95$

c $r \approx 0.99$, very strong linear relationship

3 a Coefficient of determination = 0.59

We can then conclude that 59% of the variation in the number of people found to have a blood alcohol reading over 0.05 can be explained by the variation in the number of booze buses in use. Thus the number of booze buses in use is a factor in predicting the number of drivers with a reading over 0.05.

b First: Number of booze buses in use (CAUSE)

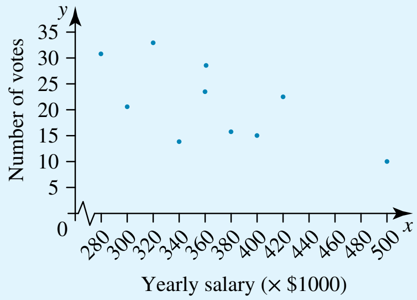
Later: Number of drivers with a blood alcohol reading over 0.05 (EFFECT)

4 First: Number of books read (CAUSE)

Later: Spelling ability (EFFECT)

Note: This is the most likely case, but there could be an argument that spelling ability affects the number of books read.

5 a

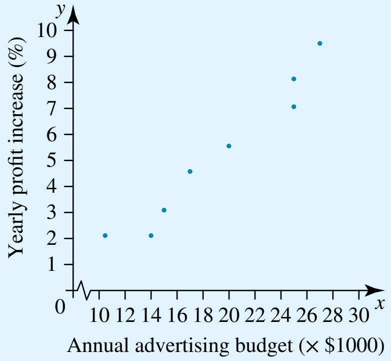


b There is moderate, negative linear association. r is approximately −0.6.

c $r = -0.66$. There is a moderate negative linear association between the yearly salary and the number of votes. That is, the larger the yearly salary of the player, the fewer the number of votes we might expect to see.

6 Coefficient of determination is 0.7569. The portion of variation in the number of visits to the doctor that can be explained by the variation in the number of cigarettes smoked is about 76%.

7 a



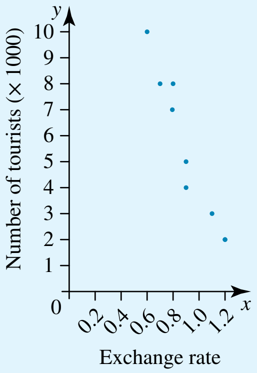
b There is strong, positive linear association. r is approximately 0.8.

c $r = 0.98$

d Coefficient of determination is 0.96.

e The proportion of the variation in the yearly profit increase that can be explained by the variation in the advertising budget is 96%.

8 a



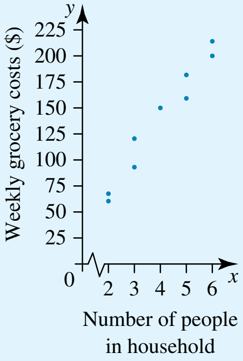
b There is strong, negative association of a linear form and r is approximately -0.9.

c $r = -0.96$

d Coefficient of determination is 0.92.

e The proportion of the variation in the number of tourists that can be explained by the variation in the exchange rate is 92%.

9 a



b There is strong, positive association of a linear form and r is approximately 0.9.

c First: Number of people in household (CAUSE)

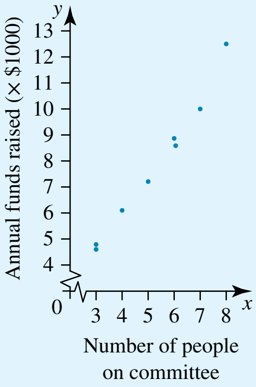
Later: Weekly grocery costs (EFFECT)

d $r = 0.98$

e Coefficient of determination is 0.96.

f The proportion of the variation in the weekly grocery costs that can be explained by the variation in the number of people in the household is 96%.

10 a



b There is almost perfect positive association of a linear form and r is nearly 1.

c $r = 0.99$

d No. High degree of correlation does not mean we can comment on whether one variable causes particular values in another.

e Coefficient of determination is 0.98.

f The proportion of the variation in the funds raised that can be explained by the variation in the number of people on the committee is 98%.

11 E

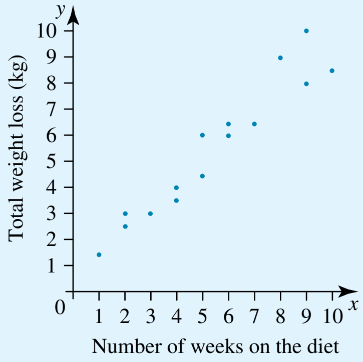
12 C

13 B

14 This is not an example of cause and effect. A common response variable, the education level of adults in a suburb, would provide a direct link to both variables.

15 C

16 a



b The scatterplot shows strong, positive association of linear form.

c It is appropriate since the scatterplot indicates association showing linear form and there are no outliers.

d $r \approx 0.9$

e $r = 0.96$

f We cannot say whether total weight loss is affected by the number of weeks people stayed on the Certain Slim diet. We can only note the degree of correlation.

g $r^2 = 0.92$

h The coefficient of determination tells us that 92% of the variation in total weight loss can be explained by the variation in the number of weeks on the Certain Slim diet.

EXERCISE 3.6

1 $y = -37.57 + 1.87x$ or air conditioner sales = –37.57 + 1.87 × (temperature), $r^2 = 0.97$, strong linear association

2 $y = 2.11 + 2.64x$ or number of dialysis patients = 2.11 + 2.64 × (month), $r^2 = 0.95$, strong linear association

3 a Independent variable = ‘height of Year 12 girl’.

b $b = 0.96$

c $a = 22.64$

4 $b = -0.78167$, $a = 13.94$

$y = 13.94 - 0.78x$

5 $y = 4.57 + 1.04x$

6 $y = 35.47 - 3.72x$

7 $y = 9.06 + 1.20x$

8 a The mathematics exam mark

b 0.95 **c** -6.07 **d** 75%

9 a $y = 91.78 + 3.33x$

b $y = 21.87 - 0.15x$

c $y = 8890.48 + 52.38x$

d $y = 30 - x$

10 The least-squares regression equations are exactly the same as obtained in questions **5**, **6** and **7**.

11 a $y = 22 - 2x$

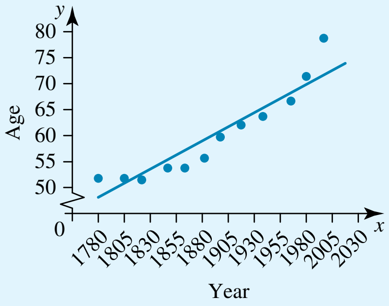
b A ‘perfect’ fit

c $y = 11 - 0.5x$

12 E

13 a $y = -164.7 + 0.119x$

b

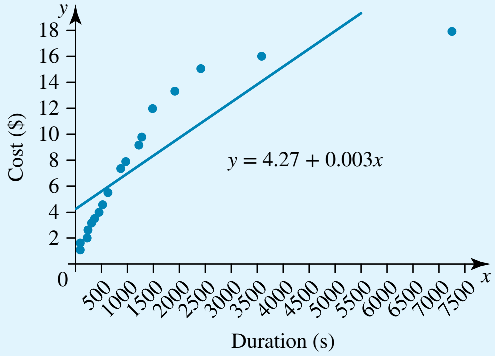


c The data definitely are not linear; there are big increases from 1880–1920, 1940–2000.

14 a Duration of call

b Cost of call ($) = $4.27 + $0.002 57 × duration of call (sec)

c



The line does not fit closely for all data points. The equation is not reliable due to outliers. If you eliminate the last two calls then there is a direct relationship.

d $r^2 = 0.73$. Therefore the linear association is only moderate.

15 D

16 a $y = 8 + 4x$, perfect fit, but meaningless

b $y = 10 + 2.5x$, good fit, but almost meaningless

c $y = 9.5 + 2.8x$, $y = 8.9 + 3.1x$, $y = 8.4 + 3.3x$, good fit.

d The answers appear to be converging towards a ‘correct’ line.

EXERCISE 3.7

1 a $y = 12.13 + 3.49x$ or monkey height (cm) = 12.13 + 3.49 × (month from birth)

b 3.49 cm/month

c 12.13 cm

2 a $y = 13.56 + 2.83x$ or temperature (°C) = 13.56 + 2.83 × (time after 6 a.m.)

b 2.83°C/hr **c** 13.56°C

3 $y = 56.03 + 9.35x$ or height (cm) = 56.03 + 9.35 × (age)

Height (cm) = 149.53 cm

4 $y = 40.10 + 6.54x$ or length (cm) = 40.10 + 6.54 × (months)

Length (cm) = 138.20 cm

5 Height (cm) = 205.63 cm

We cannot claim that the prediction is reliable, as it uses extrapolation.

6 Length (cm) = 197.06 cm

We cannot claim that the prediction is reliable, as it uses extrapolation.

7 a $y = 157.3 + 14x$

b 14 cells per day

c 157

8 a 48.5, or 49 people per extra animal

b 31.8, or 32 visitors

9 $y = 464 - 1.72x$, $r = -0.98$. Gradient shows a drop of 1720 sales for every $1 increase in the price of the item. Clearly, the y-intercept is nonsensical in this case since an item is not going to be sold for $0! This is a case where extrapolation of the line makes no sense.

10 a 7

b 18

11 a $y = 0.286 + 3.381x$ **b** 10.4

c 40.9 **d** 1.99

e 7.31 **f** c

12 a Factory 1: $y = 43.21 + 1.51x$
Factory 2: $y = 56.61 + 0.96x$

b Factory 1 is cheaper at $43.21 (compared to Factory 2 at $56.61).

c Factory 2 is cheaper at $167.47 (compared to Factory 1 at $216.86).

d Factory 2 is marginally more linear (Factory 1: $r = 0.97$; Factory 2: $r = 0.99$).

13 a $2152 **b** 180

c $78 200 **d** $600

e **a**, **b** only

14

IQ	Test result (%)
80	56
87	60
92	68
102	65
105	73
106	74
107	71
111	73
115	80
121	92

15 a $55 000 b $58 600

c $64 000 d About 13 years

e Various answers

16 a $r = 11.73 + 5.35q$ b $r = 33.13$

c $r = 108.03$ d $q = 16.50$

e c and d

17 a $C = \$420$ b $n = 13.38$ hours

c $C = \$940$ d Call out fee = $180

e b and c

18 a $y = 42956 + 14.795 \times$ (millions spent on advertising)
or
Members = 42956 + 14795
× (millions spent on advertising)

b 72 546 members

Interpolation

c $1.83 million

d $r^2 = 0.90$, strong linear relationship

EXERCISE 3.8

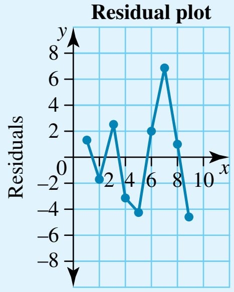
1 Using CAS: $y = -0.03 + 10.85x$

See table at foot of the page.*

2 Using CAS: $y = 0.69 + 9.82x$

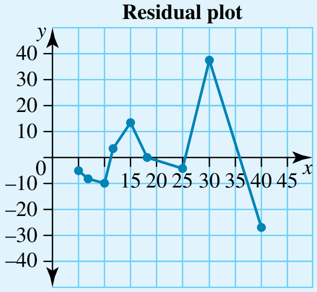
See table at foot of the page.*

3



$r^2 = 0.98$, consistent with a linear relationship

4



$r^2 = 0.98$, consistent with a linear relationship

5

x	y	y_{pred}	Residuals
1	1	5.1	−4.1
2	9.7	7.46	2.24
3	12.7	9.82	2.88
4	13.7	12.18	1.52
5	14.4	14.54	−0.14
6	14.5	16.9	−2.4

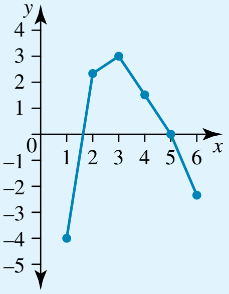
*1

x	1	2	3	4	5	6	7	8	9
y	12	20	35	40	50	67	83	88	93
predicted y	10.82	21.67	32.52	43.37	54.22	65.07	75.92	86.77	97.62
Residual $(y - y_{\text{predicted}})$	1.18	−1.67	2.48	−3.37	−4.22	1.93	7.08	1.23	−4.62

*2

x	5	7	10	12	15	18	25	30	40
y	45	61	89	122	161	177	243	333	366
predicted y	49.79	69.43	98.89	118.53	147.99	177.45	246.19	295.29	393.49
Residual $(y - y_{\text{predicted}})$	−4.79	−8.43	−9.39	3.47	13.01	−0.45	−3.19	37.71	−27.49

6 By examining the original scatterplot, and residual plot, data are clearly not linear.

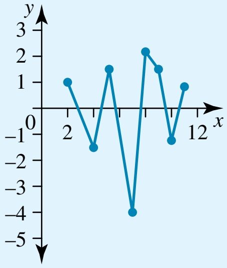


7 D

8 a, b

Day	Defective rate (%)	y_{pred}	Residuals
2	15	13.96	1.04
4	10	11.58	−1.58
5	12	10.39	1.61
7	4	8.01	−4.01
8	9	6.82	2.18
9	7	5.63	1.37
10	3	4.44	−1.44
11	4	3.25	0.75

c



No apparent pattern in the residuals — likely to be linear

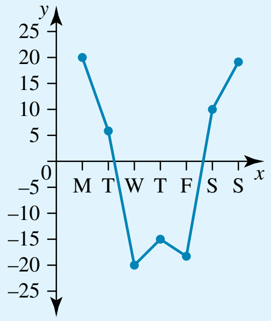
d 15.15%

e 13.7 days. Unlikely that extrapolation that far from data points is accurate. Unlikely that there would be 0% defectives.

9 a, b

Day	Bookings in hotel	y_{pred}	Residuals
1	158	138.8	19.2
2	124	116.3	7.7
3	74	93.8	−19.8
4	56	71.3	−15.3
5	31	48.8	−17.8
6	35	26.3	8.7
7	22	3.8	18.2

c



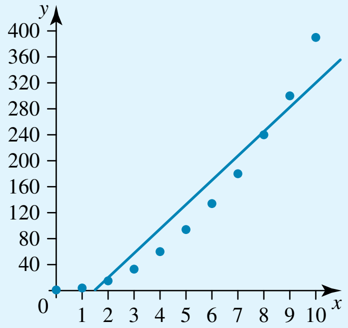
Slight pattern in residuals — may not be linear

d Decline in occupancy likely due to drought — an atypical event.

10 C

11 a Non-linear **b** Linear **c** Non-linear

12 a $y = -57.73 + 37.93x$



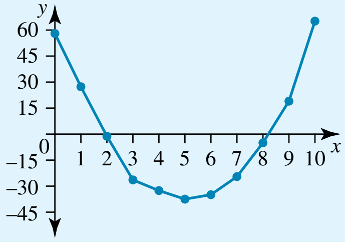
b $r = 0.958$. This means that there is a strong positive relationship between variables x and y.

c 0.9177, therefore 91.8% of the variation in y can be explained by the variation in x.

d

x	Residuals
0	58.7
1	23.8
2	−3.1
3	−23.1
4	−34.0
5	−37.9
6	−35.8
7	−27.8
8	−5.7
9	16.4
10	68.5

e

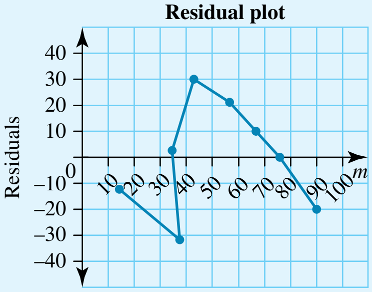


There is a clear pattern; the relationship between the variables is non-linear.

13 $P = -3.94 + 1.52m$

See table at foot of the page.*

14

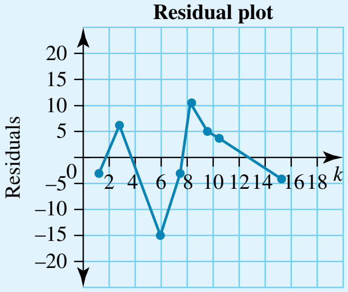


Since the data shows a curved pattern, the original data probably have a non-linear relationship.

15 Using CAS: $D = 13.72 + 7.22k$

See table at foot of the page.*

16 a

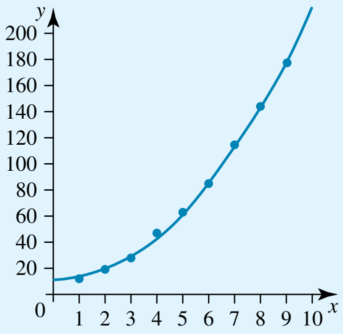


Since the data randomly jumps from above to below the k-axis; the data probably has a linear relationship.

b $r^2 = 0.94$; 94% of variation in D can be explained by variation in k.

EXERCISE 3.9

1 a



Apply an x^2 transformation to stretch the x-axis.

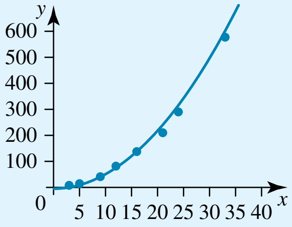
b

x^2	1	4	9	16	25	36	49	64	81
y	12	19	29	47	63	85	114	144	178

$y = 11.14 + 2.07x^2$ with $r = 0.99976$

This transformation has improved the correlation coefficient from 0.97 to 0.99976; thus the transformed equation is a better fit of the data.

2 a



Apply an x^2 transformation to stretch the x-axis.

b

x^2	9	25	81	144	256	441	576	1089
y	5	12	38	75	132	209	291	578

$y = -4.86 + 0.53x^2$ with $r = 0.9990$

This transformation has improved the correlation coefficient from 0.96 to 0.9990, thus the transformed equation is a better fit of the data.

3

Time (sec)	1	2	3	4	5	6	7	8
Log_{10} Speed (ms^{-1})	1.95	1.85	1.74	1.65	1.59	1.54	1.51	1.48

$\log_{10}(\text{speed}) = 1.97 - 0.07 \times \text{time}$ with $r = -0.97$

Therefore there is an improvement of the correlation coefficient that resulted from applying a logarithmic transformation.

4

x	5	10	15	20	25	30	35	40
$\log_{10} y$	3	2.70	2.35	2.17	2.00	1.85	1.77	1.75

$\log_{10} y = 3.01 - 0.04x$ with $r = -0.96$

Therefore there is a significant improvement of the correlation coefficient that resulted from applying a logarithmic transformation.

*13

m	12	37	35	41	55	69	77	90
P	2.5	21.6	52.3	89.1	100.7	110.3	112.4	113.7
$P_{\text{predicted}}$	14.30	52.30	49.26	58.38	79.66	100.94	113.10	132.86
Residual ($P - P_{\text{predicted}}$)	−11.8	−30.7	3.04	30.72	21.04	9.36	−0.7	−19.16

*15

k	1.6	2.5	5.9	7.7	8.1	9.7	10.3	15.4
D	22.5	37.8	41.5	66.9	82.5	88.7	91.6	120.4
$D_{\text{predicted}}$	25.27	31.77	56.32	69.31	72.20	83.75	88.09	124.91
Residual ($D - D_{\text{predicted}}$)	−2.77	6.03	−14.82	−2.41	10.30	4.95	3.51	−4.51

5 a

$\frac{1}{\text{Time after 6 pm (hr)}}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	$\frac{1}{8}$
Temperature (°C)	32	22	16	11	9	8	7	7

$y = 3.85 + 29.87x_T$, where $x_T = \dfrac{1}{x}$

or

$$\text{Temperature} = 3.85 + \frac{29.87}{\text{Time after 6 pm}}$$

b Temperature = 10.49 °C

6 a

$\frac{1}{x}$	$\frac{1}{2}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{9}$	$\frac{1}{10}$	$\frac{1}{13}$	$\frac{1}{15}$	$\frac{1}{18}$
y	120	50	33	15	9	5	2	1

$y = -13.51 + 273.78x_T$, where $x_T = \dfrac{1}{x}$

or

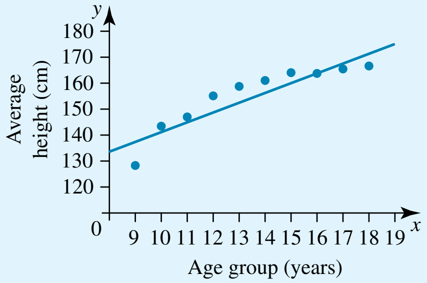
$$y = -13.51 + \frac{273.78}{x}$$

b $y = 9.31$

c $y = 8.17$

7 $y = 128.15 - 2.62x_T$ where $x_T = x^2$, $r = -0.97$, which shows some improvement.

8 a



b

log (age group)	Average height (cm)
0.954	128
1	144
1.041	148
1.079	154
1.114	158
1.146	161
1.176	165
1.204	164
1.230	166
1.255	167

c $y = 24.21 + 117.2x_T$ where $x_T = \log_{10}(x)$, $r = 0.95$, most non-linearity removed.

9 a i 123.3 cm **ii** 143.9 cm **iii** 176.7 cm

b a ii

10 Normal growth is linear only within given range; eventually the girl stops growing. Thus logarithmic transformation is a big improvement over the original regression.

11 a $y = 2.572 + 90.867x_T$ where

$x_T = \dfrac{1}{x}(r = 0.9788)$

b The intensity is 7.1 candlepower.

12 a Compress the y- or x-values using logs or reciprocals.

b Stretch the y-values using y^2 or compress the x-values using logs or reciprocals.

c Compress the y- or x-values using logs or reciprocals.

13 a −11.25 **b** −11.25

14 a −24.55

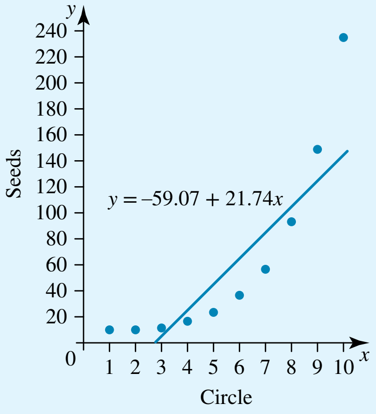
b Cannot take the log of a negative number.

c Cannot take the log of zero.

15 a 3.39 **b** 0.34

16 a −0.08 **b** −0.08

17 a



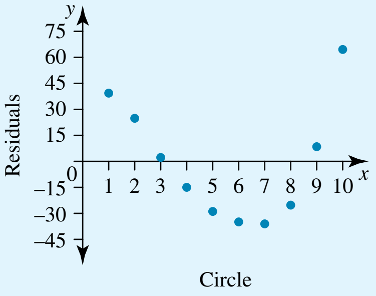
b $r = 0.87$, which means it is a strong and positive relationship.

c 180

d

Circle	Seeds	Residual
1	3	40.33
2	5	20.59
3	8	1.85
4	13	−14.89
5	21	−28.63
6	34	−37.37
7	55	−38.11
8	89	−25.84
9	144	7.41
10	233	74.67

e

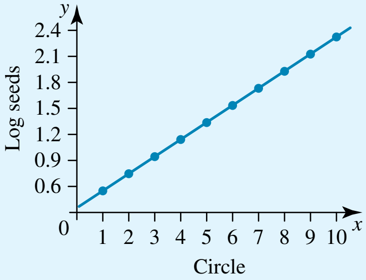


No, the relationship is not linear.

f **i** We can stretch the x-values towards linearity by using an x^2 transformation.

ii We can compress the y-values towards linearity by using either a $\log_{10}(y)$ or a $\frac{1}{y}$ transformation.

18 **a**



b $r = 0.9999$, this is an almost perfect relation.

c $\log_{10}(y) = 0.2721 + 0.2097x$

$\log_{10}$ (number of seeds) = 0.2721 + 0.2097 × circle number

d 0.9999, 99.99% (100.0%) of variation in number of seeds is due to number of circles. This is a perfect relation, often found in nature (see the Golden Ratio).

e 378

f This is a much better prediction as it follows a steep upward trend.