STATISTICAL ANALYSIS

8.

CORRELATION AND REGRESSION

This chapter looks at comparisons between variables and uses correlation to measure the strength of these relationships. We also look at regression and lines of best fit to make predictions from this data using interpolation and extrapolation.

CHAPTER OUTLINE

8.01 Bivariate data
8.02 Correlation
8.03 Line of best fit
8.04 Least-squares regression line



IN THIS CHAPTER YOU WILL:

- interpret scatterplots of bivariate data
- look for correlation in bivariate data and calculate Pearson's correlation coefficient
- apply lines of best fit, including the least-squares regression line
- interpolate and extrapolate from data

Shutterstock.com/PopTika

TERMINOLOGY

bivariate data: Data relating to 2 variables that have been measured from the same data set.
extrapolation: Making predictions from a model using values outside the range of the original data set.
interpolation: Making predictions from a model using values lying within the range of the original data set.
least-squares regression line: A line of best fit where the squares of the distances from each point in the scatterplot to the line are minimised.
line of best fit: A line drawn through a scatterplot that best models the relationship between 2 variables in bivariate data.
Pearson's correlation coefficient: A calculated value, r, that measures how closely 2 variables are related in a linear relationship. Its value is always between -1 and 1.
scatterplot: A graph showing the value of 2 variables in a bivariate data set.

WS Scatterplots

WS A page of scatterplots

WS Body measurements

8.01 Bivariate data

In Chapter 7, *Statistics*, we looked at data with one variable.

Bivariate data measures 2 variables on the same data set to see if they correlate with (are related to) each other. For example, we might want to see if a person's level of education correlates with the amount of money the person earns.

We draw **scatterplots** to graph bivariate data.



Shutterstock.com/michaeljung

EXAMPLE 1

This table shows the bivariate data for the number of days each year the temperature in a city was over 30°C and the number of people admitted to the local hospital for heat stroke.

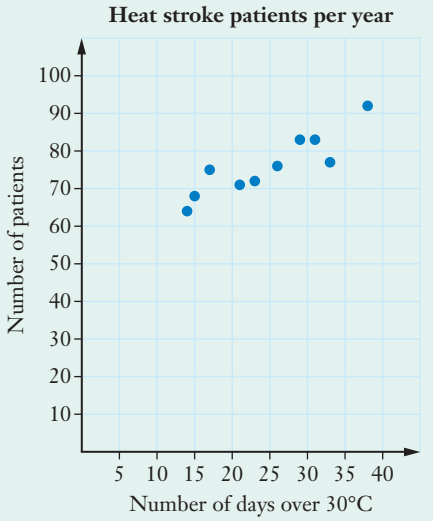
Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Number of days over 30°C	14	17	15	23	21	26	31	33	29	38
Number of patients	64	75	68	72	71	76	83	77	83	92

Draw a scatterplot for this data.

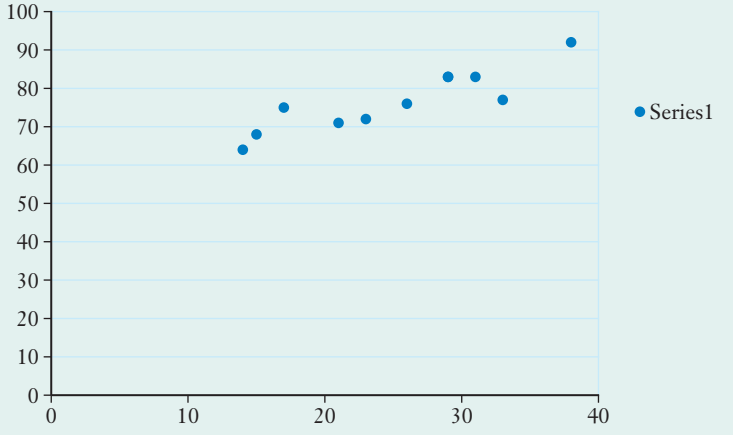
ISBN 9780170413220

Solution

We draw the graph with number of days over 30°C on the horizontal axis and number of patients on the vertical axis.



You could also put the data from the table into a spreadsheet and choose the scatterplot chart.

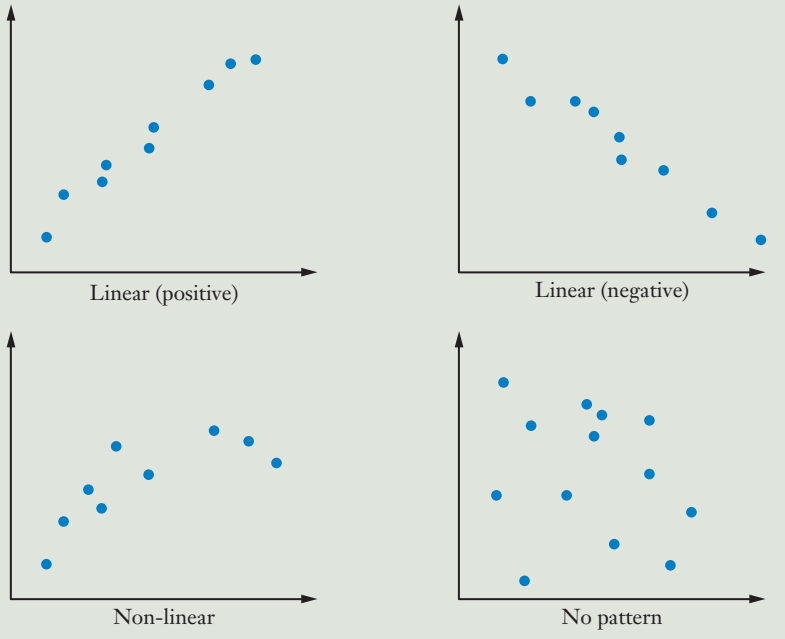


We look for patterns in the scatterplot to see if the 2 variables are related. For instance, in the example above there seems to be a linear pattern (it is roughly a straight-line graph).

Not all relationships are linear. We can have scatterplots that have curves or non-linear shapes, or no shape at all.

ISBN 9780170413220

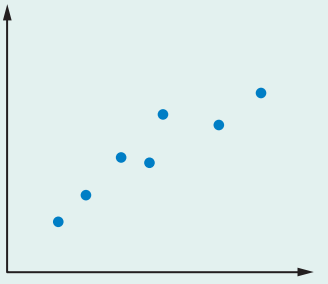
Shape of scatterplots



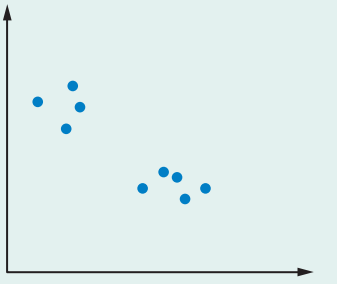
EXAMPLE 2

Describe the shape of each scatterplot.

a



b



Solution

a This data is linear (positive).

b This data is linear (negative).

ISBN 9780170413220

Exercise 8.01 Bivariate data

1 **Biometric data** is bivariate data drawn from body measurements such as arm length or height. This table shows the heights and weights of 10 people surveyed.

Height (m)	161.8	175.3	159.5	182.3	166.4	167.9	186.4	164.7	154.8	171.2
Weight (kg)	59.4	73.7	55.3	74.8	63.5	68.2	73.5	62.1	49.9	83.6

a Draw a scatterplot to show this data.

b Describe the shape of the scatterplot.

c Describe the relationship between height and weight (if possible).

2 A group of 8 people were surveyed at random about how many people in their family lived at home. They were also asked about the number of bedrooms in their home.

Number of people in family	4	3	6	7	2	1	5	8
Number of bedrooms	3	2	5	4	2	3	4	6

a Draw a scatterplot showing the data.

b Describe the shape of the scatterplot.

c Describe the relationship (if any) between the number of people in the family and the number of bedrooms in the home.

3 A survey of 10 people measured their IQ with the amount they earned each week.

IQ	114	127	95	130	123	141	136	83	109	96
Amount earned per week ($)	689	945	510	874	751	769	553	350	1250	884

a Draw a scatterplot for this data.

b Describe the shape of the scatterplot.

c Describe any relationship between IQ and amount earned.

4 The table shows the amount of sleep students have and their exam results.

Hours of sleep	6.5	9	7.5	6	7	10	11.5	4	8	8.5
Exam results (%)	87	76	43	87	69	55	60	78	94	72

a Draw a scatterplot of this data.

b Describe the shape of the scatterplot.

c What is the relationship between the amount of sleep and exam results?

ISBN 9780170413220

5 The table shows the amount of study time and exam results for some students.

Hours of study/week	15	26	12	17	5	10	2	8	20	3
Exam results (%)	78	86	70	80	63	67	43	77	92	58

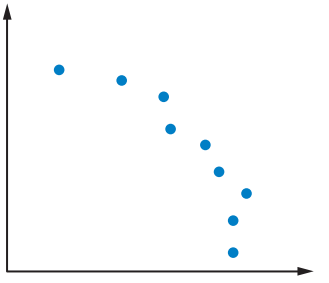
a Draw a scatterplot of the data.

b Describe the form of the scatterplot.

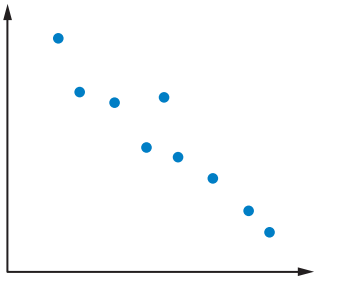
c What is the relationship between the amount of study and the exam results?

6 Describe the pattern of each set of bivariate data.

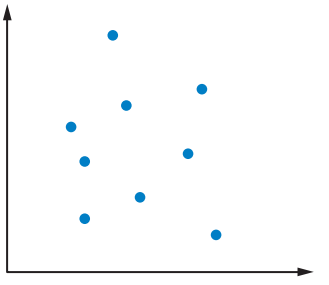
a



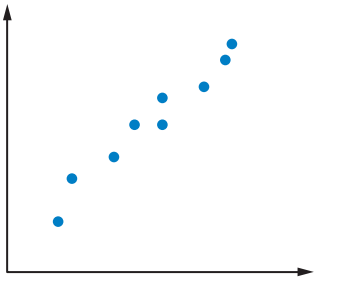
b



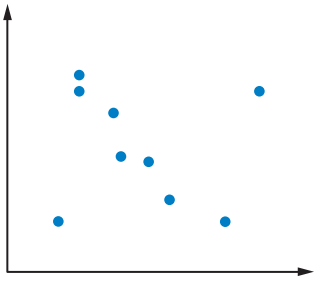
c



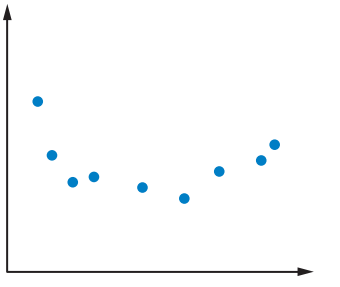
d



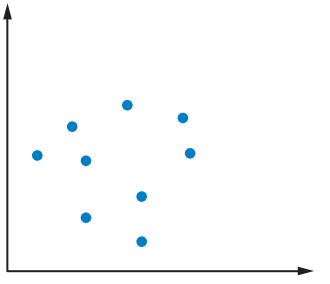
e



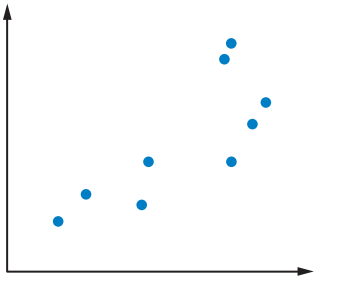
f



g

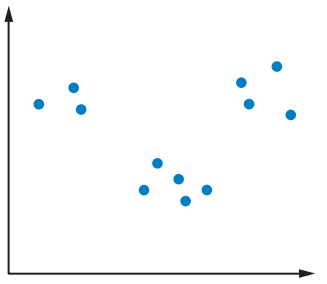


h

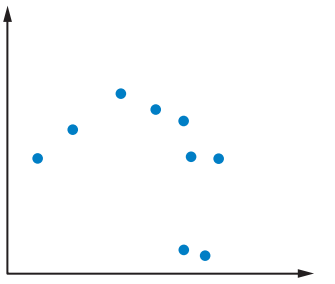


ISBN 9780170413220

i



j



7 a Draw a scatterplot for this set of bivariate data.

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

b Describe the shape of the data.

8 State whether each scatterplot has:

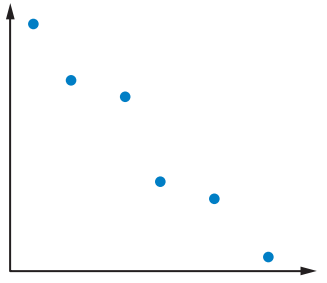
A a positive linear relationship

B a negative linear relationship

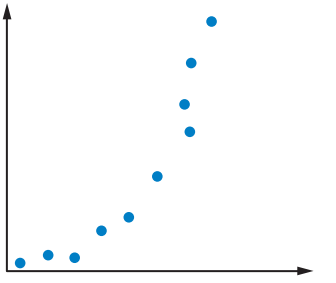
C a non-linear relationship

D little or no relationship

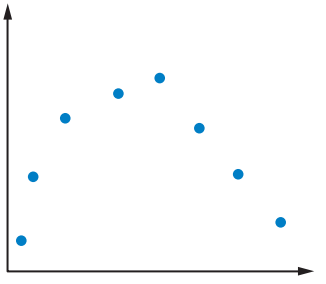
a



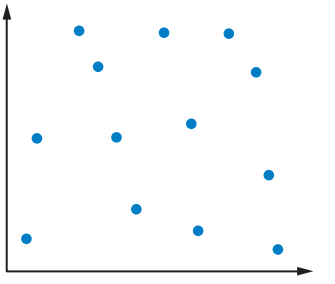
b



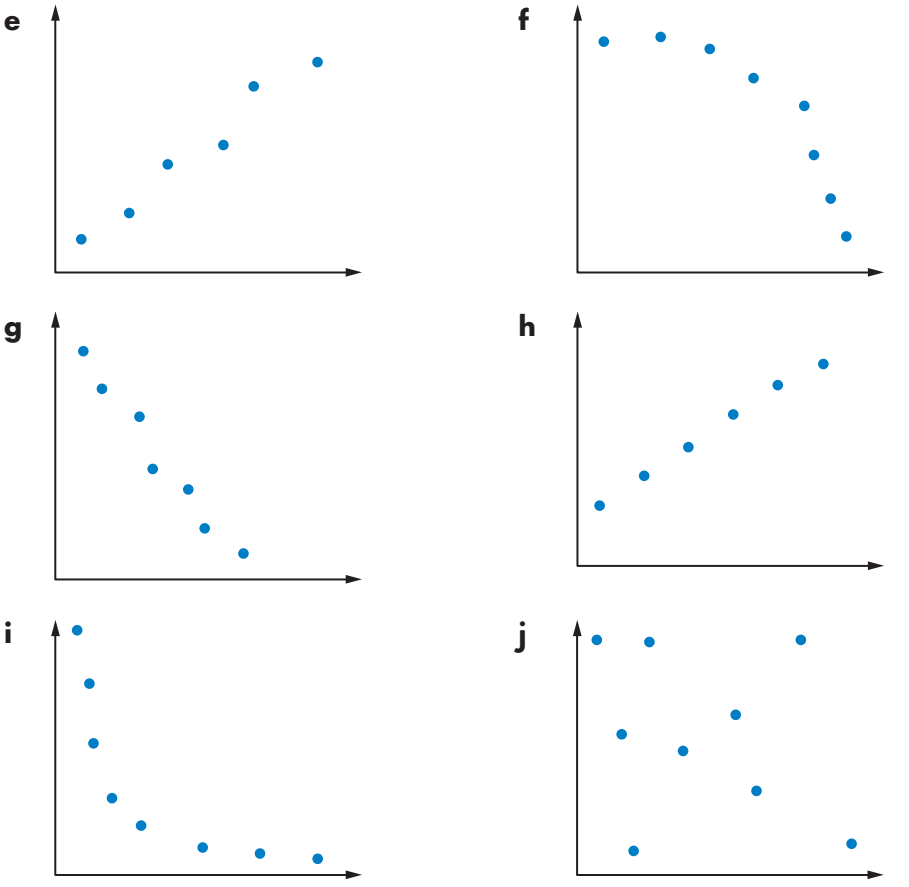
c



d



ISBN 9780170413220



Correlation

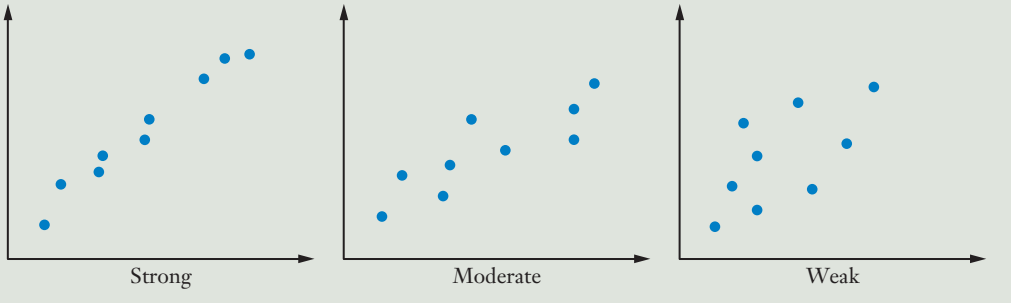
Correlations matching game

Correlation

8.02 Correlation

Correlation measures how well 2 variables are related if there seems to be a linear relationship between them. The relationship could be strong, moderate or weak.

Linear scatterplots

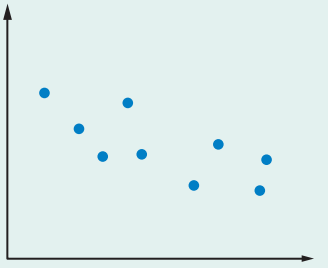


ISBN 9780170413220

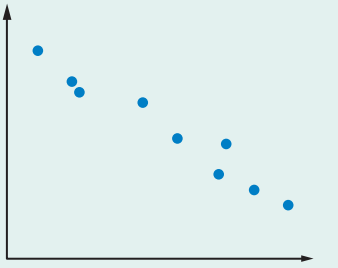
EXAMPLE 3

Describe the strength of the linear pattern in each scatterplot as strong, moderate or weak.

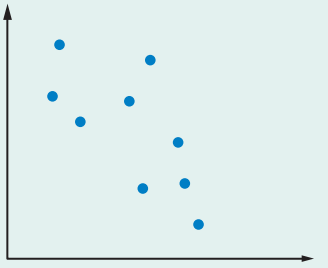
a



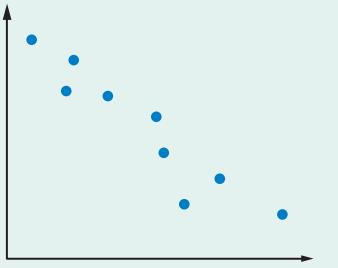
b



c



d



Solution

a The linear pattern is moderate.

b The linear pattern is strong.

c The linear pattern is weak.

d The linear pattern is strong.

Pearson's correlation coefficient

The terms strong, moderate or weak are very general and not very accurate. We use a measurement (r) called the **Pearson's correlation coefficient** to determine how closely related variables are in a linear relationship.

The formula for r is complex, but you can use a calculator or spreadsheet to calculate it.

The correlation coefficient always lies between -1 and 1.

Correlation coefficient

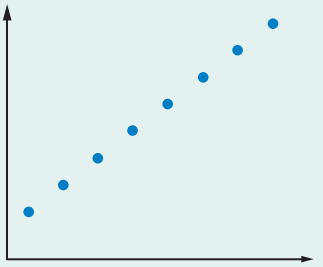
- $-1 \leq r \leq 1$ for all correlation coefficients
- $0 < r \leq 1$ for a scatterplot with positive direction, where 1 is perfect positive correlation
- $-1 \leq r < 0$ for a scatterplot with negative direction, where -1 is perfect negative correlation
- $r = 0$ means no correlation

ISBN 9780170413220

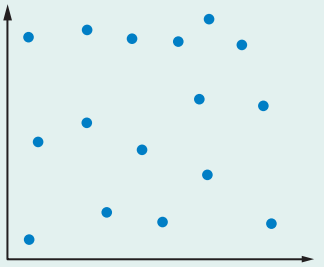
EXAMPLE 4

Match each scatterplot with its correct correlation coefficient.

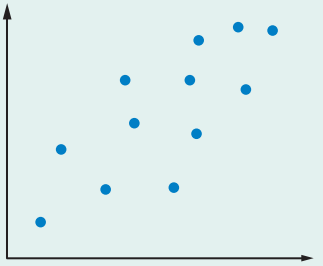
a



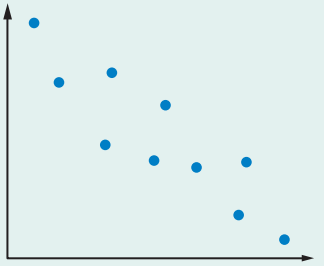
b



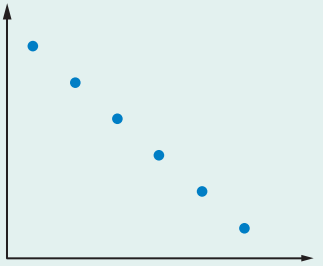
c



d



e



A 0.5 **B** 1 **C** −1 **D** −0.4 **E** 0

Solution

a B **b** E **c** A **d** D **e** C

DID YOU KNOW?

Karl Pearson

Karl Pearson (1857–1936), an Englishman, developed the formula for the correlation coefficient r. The coefficient's full name is **Pearson's product moment correlation coefficient**. Karl Pearson was a mathematician and statistician, but he also studied history, law and German literature.

Research Karl Pearson to find out more about his life and studies.

ISBN 9780170413220

EXAMPLE 5

A group of students was surveyed for the number of hours that they studied and their result in a maths exam. Find the correlation coefficient for this bivariate data.

Number of hours studied	6	5	12	8	15	9	14
Exam results (%)	71	46	74	67	76	77	83

Solution

Operation	Casio Scientific	Sharp Scientific
Place your calculator in statistical mode.	MODE 2: STAT 2: A+BX	MODE 1 STAT 1 LINE
Clear the statistical memory.	SHIFT 1 3 : Edit 2 : Del-A	2ndF DEL
Enter data.	SHIFT 1 2 : Data 6 = 5 = etc. for 1st column 71 = 46 = etc. for 2nd column AC	6 2ndF STO 71 M+ 5 2ndF STO 46 M+ etc.
Calculate r.	SHIFT 1 5 : Reg 3 : r =	ALPHA r =
Change back to normal mode.	MODE 1 : COMP	MODE 0

$r = 0.738$ (correct to 3 decimal places)



Newpix/Ian Currie

ISBN 9780170413220

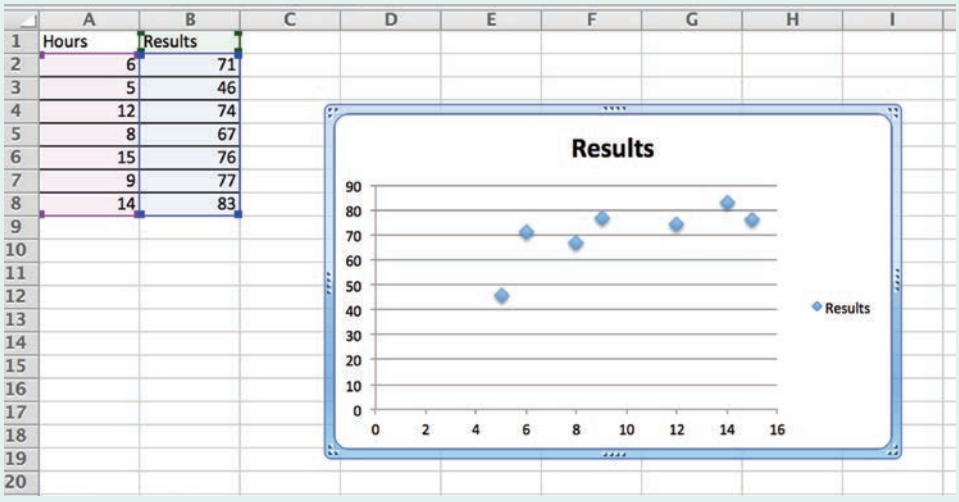
EXAMPLE 6

a Enter the data from Example 5 above into a spreadsheet and draw a scatterplot.

b Use the spreadsheet to find the Pearson's correlation coefficient for the data.

Solution

a Put the values from the table into a spreadsheet and select scatterplot from the charts.



b In an empty cell, type =**PEARSON(A2:A8,B2:B8)**

$r = 0.738$ (to 3 decimal places)

Causality

Two variables can have a high correlation without one *causing* the other. For example, does a person's height cause them to weigh more? Possibly, since most tall people would have heavier, longer bones. However, there are other causes for higher weight that aren't related to height.

EXAMPLE 7

For each set of bivariate data, find which ones have a causal relationship.

a Number of people in a family and the number of TVs

b Speed of a boat and time taken to travel across a lake

Solution

a Not causal: the number of people in a family doesn't determine how many TVs there are.

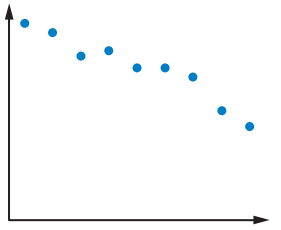
b Causal: The speed of a boat will determine the time taken to travel across a lake (higher speed means less time).

ISBN 9780170413220

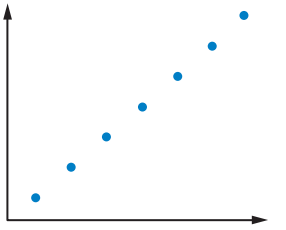
Exercise 8.02 Correlation

1 Match each graph with the correct correlation coefficient.

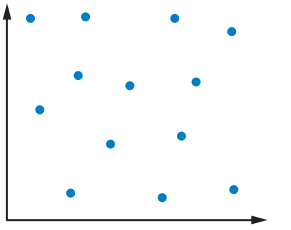
a



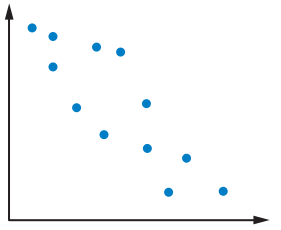
b



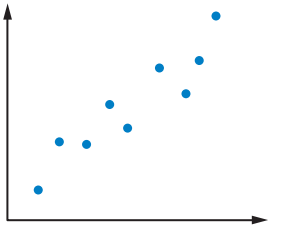
c



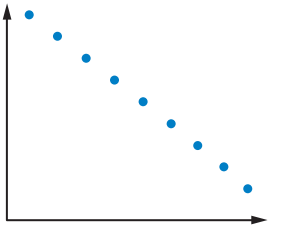
d



e



f



A –0.5 **B** 1 **C** 0.8 **D** –1

E –0.9 **F** 0 **G** –0.7

2 For each table of values:

i draw a scatterplot **ii** find the correlation coefficient

a

Height (m)	1.72	1.85	1.61	1.74	1.59	1.79
Weight (kg)	91.3	85.2	58.3	61.9	74.5	102.6

b

Height (cm)	167	180	174	171	154	190
Speed (m s^{-1})	3.7	2.1	3.4	2.2	2.8	4.1

c

Study time (h)	13	21	8	11	18	17
Results (%)	45	89	81	67	74	53

d

Temperature (°C)	15	18	21	24	26	30	35
Attendance at beach	28	19	54	88	190	245	108

e

Forest cleared (ha)	41	58	87	99	132	168
Number of birds	1200	854	530	201	157	92

f

Number of cars in car park	1100	1450	1809	2004	2234	2569	2871	2906
Pollution (ppm)	1.21	1.54	1.78	2.34	2.99	3.35	4.76	5.97

ISBN 9780170413220

g

Age	15	19	27	34	49	57
Annual income ($)	2 851	12 600	27 890	38 740	41 834	29 450

h

Exercise (h/week)	14	8	2	10	6	4	32
Weight (kg)	51.8	87.2	74.8	68.4	62.1	63.9	58.9

i

Height (m)	1.59	1.77	1.64	1.78	1.89	1.42
Shoe size	5	7	6	10	9.5	4

j

Exam results (%)	68	92	38	51	77	84
Hours of sleep	7	6	8	6.5	9	7.5

3 For the bivariate data in question 2, which do you think have causality?

4 Find the correlation coefficient of each set of data, correct to 2 decimal places.

a

x	y
3	7
5	9
4	3
11	7
15	12
8	4
9	1

b

x	y
5	67
6	49
3	81
9	23
11	55
8	91
4	61

c

x	y
8	11
4	8
7	11
2	4
9	12
14	16
23	23

d

x	y
5	21
3	28
6	19
5	17
9	21
4	26
11	15
15	18
9	12

5 Determine whether each pair of variables are likely to have a causal relationship.

a Population of a city and pollution

b Head circumference and weight

c Hours training and fitness

d Weight and health

e Size of house and number of pets

f Size of house and selling price

ISBN 9780170413220

INVESTIGATION

CAUSALITY

Discuss whether each pair of variables have a high correlation and, if they do, whether one variable causes the other.

1 A person's height and shoe size

2 A person's smoking and lung cancer

3 Amount of study and success in an exam

4 Mathematical and musical ability

5 The number of people at a party and the amount of food and drink consumed

6 The amount of time sunbaking and the incidence of skin cancer

7 The amount of time practising basketball and the number of baskets scored in a game

8 Results in English and Maths exams

9 The length of a person's leg and their walking speed

10 The temperature and the number of people swimming at the beach



Shutterstock.com/FRITZ16

ISBN 9780170413220

Discuss other relationships between variables. Can you find other examples of highly correlated variables where one causes the other? Can you find examples of highly correlated variables where there is no causality?

Discuss causality in each situation described below.

1 The time a sales representative has been with a company and the number of sales gives a correlation coefficient of –0.6.

2 Height of basketball players and number of baskets scored have a correlation coefficient of 0.87.

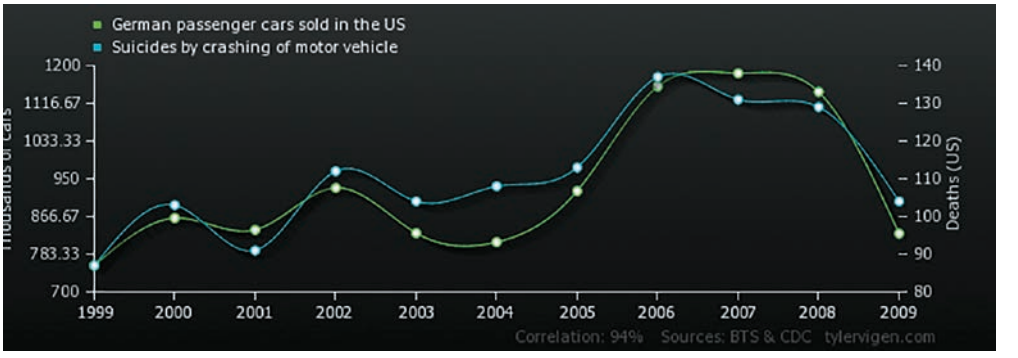
3 Height and self-esteem have a correlation coefficient of 0.32.

4 Temperature and growth of grass have a correlation coefficient of –0.75.

5 Number of hours study and results in the HSC have a correlation coefficient of 0.85.

Collect data from the Internet, newspapers or magazines, or do your own experiments to compare two sets of data. Draw a scatterplot and find the correlation coefficient, then, if there is a high correlation investigate causality. Is the correlation positive or negative? Is it linear?

Here is an example of a high correlation of totally unrelated variables: German cars sold vs suicides by car crashes in the US each year.



http://tylervigen.com/page?page=2

There are many other examples of correlated data that have no causality.

ISBN 9780170413220

8.03 Line of best fit

Statistical data is rarely perfect, but we can often see trends in a scatterplot. If there seems to be a linear correlation, we can draw a **regression line** and find its equation. We can then use this line to make predictions.

The easiest regression line to find is the **line of best fit**.

Using a ruler, we draw the line that represents as many points as possible. We try to draw a line where about half the points are above the line and half are below it, so that the distance between the line and the points is kept to a minimum.

WS Lines of best fit

WS Lines of fit

WS Height vs shoe size

EXAMPLE 8

A ball is rolled down a ramp and its velocity is measured over time.

t (s)	1	2	3	4	5	6
v (m s^{-1})	0.7	1.5	3.6	6.1	7.8	9.9

a Draw a scatterplot of the data and draw a line of best fit.

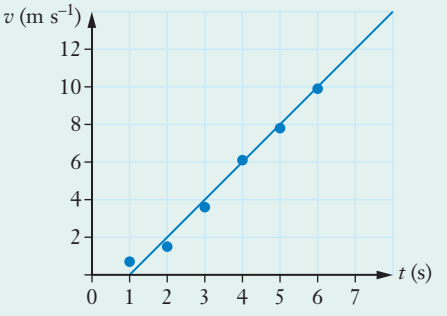
b Use the line of best fit to find the velocity after 3.5 seconds.

c Find the equation of this line.

d Use the equation to find the velocity after 10 seconds.

Solution

a



b After 3.5 seconds, the velocity is 5 m s^{-1}.

c Choose 2 points on the line, say, (2, 2) and (4, 6).

$$\text{Gradient} = \frac{y_2 - y_1}{x_2 - x_1}$$
$$= \frac{6-2}{4-2}$$
$$= \frac{4}{2}$$
$$= 2$$

Equation: $y - y_1 = m(x - x_1)$

$v - 2 = 2(t - 2)$ using $m = 2$ and (2, 2)

$= 2t - 4$

$v = 2t - 2$

ISBN 9780170413220

d When $t = 10$:

$v = 2(10) - 2$

$= 18$

So after 10 seconds, the velocity is 18 m s^{-1}.

Interpolation and extrapolation

Interpolation is using a model to make predictions about values lying **within** the range of the original data set.

Extrapolation is using a model to make predictions about values **outside** the range of the original data set.

In the above example, finding v when $t = 3.5$ is interpolation while finding v when $t = 10$ is extrapolation.

DID YOU KNOW?

Regression

The word **regression** comes from the Latin *regressio*, meaning 'a return'.

Sir Francis Galton (1822–1911) was the first to use this name. He created the concepts of correlation and regression, and used statistics to develop questionnaires and surveys on human differences.

Research correlation, regression and Sir Francis Galton.

CLASS DISCUSSION

EXTRAPOLATION

Extrapolation is not always accurate. For example, a formula may work well at first, then the conditions may change in a way that means the formula is no longer a good model.

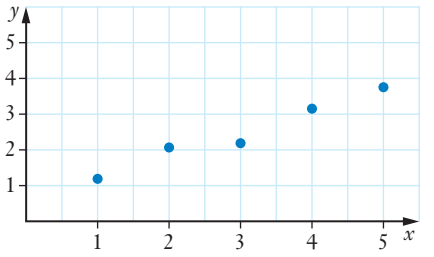
Think of examples of bivariate data from the previous examples. Can you always extrapolate answers from a line of best fit? Why? What are the risks?

ISBN 9780170413220

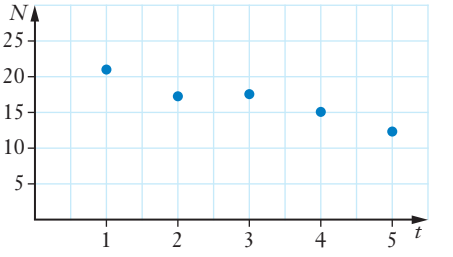
Exercise 8.03 Line of best fit

1 Copy each scatterplot, draw a line of best fit and find its equation.

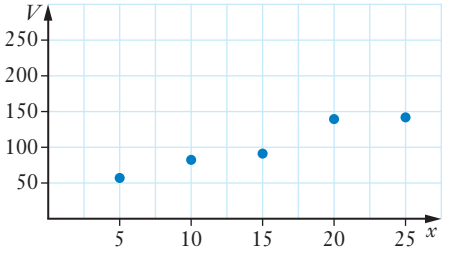
a



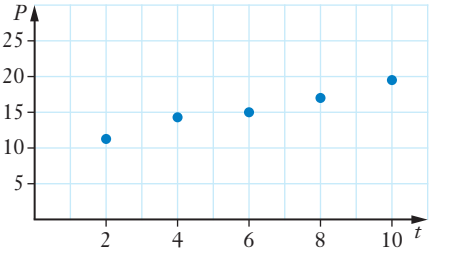
b



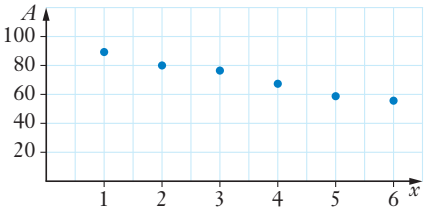
c



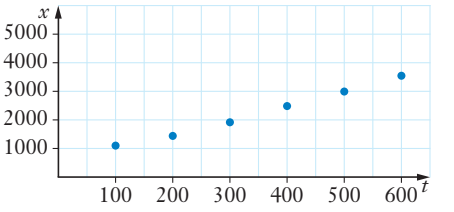
d



e



f



2 The following table shows the results of an experiment testing the temperature of a liquid as it cools down.

t (min)	5	10	15	20	25	30
T (°C)	87	78	69	56	53	41

a Plot this data on a number plane and draw a line of best fit.

b Find the equation of the line.

c Use the equation to find the temperature after:

i 17 minutes

ii 35 minutes

d If room temperature is 23°C, is the line of best fit a good model for the cooling of the liquid? Why?

ISBN 9780170413220

3 The population of birds in a particular area was sampled over several years.

t (years)	1	2	3	4	5	6
P	1030	983	968	954	915	899

a Plot this data on a number plane and draw a line of best fit.

b Find the equation of the line.

c Use the equation to find the population of birds after 7 years.

d At this rate of decline, after how many years would you expect there to be no more birds in this area?

4 The effect of a dose of medicine on a child's temperature over time was measured from a sample of children, with the following results.

t (min)	2	4	6	8	10
T (°C)	39.5	39.1	38.9	38.2	37.7

a Plot this data on a number plane and draw a line of best fit.

b Find the equation of the line.

c Use the equation to find the child's temperature after 15 minutes.

d Is this equation reliable as a measure of temperature after a longer time, say, 1 hour?

5 This table shows the results of a survey into the number of people who attend a new restaurant over a number of weeks.

t (weeks)	1	2	3	4	5	6
No. of people	76	114	163	187	228	274

a Draw a scatterplot and sketch a line of best fit.

b Find the equation of this line.

c Use the equation to find the number of people you would expect to attend the restaurant after 10 weeks.

d Is this equation a good model for the number of people attending the restaurant?

ISBN 9780170413220

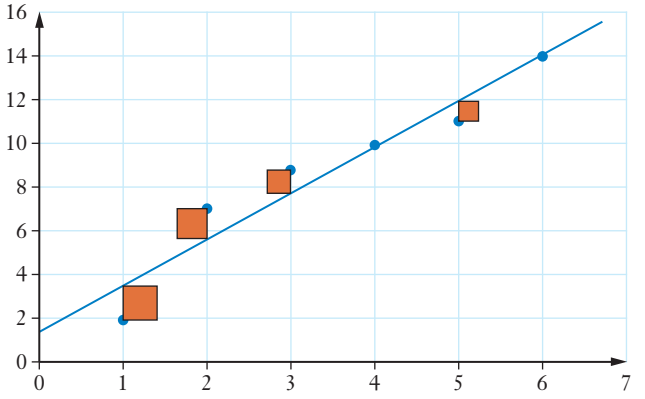
8.04 Least-squares regression line

WS Least-squares regression line

WS Least-squares regression line

The line of best fit relies on our eyes and ruler for its accuracy.

There are several different models of regression lines that try to give a more accurate result. The most popular model is called the **least-squares regression line**. It uses a line of best fit for which the squares of the distances from each point in the scatterplot to the line are minimised (see the diagram below). Squaring the distances takes away any negative values (a similar technique to finding standard deviation).



The equation of the least-squares regression line is given by $y = mx + c$ where $m = r\dfrac{s_y}{s_x}$ and $c = \overline{y} - m\overline{x}$, where r = correlation coefficient, $\overline{x}$ and $\overline{y}$ are the sample means and s_x and s_y are the sample standard deviations.

However, you don't need to use these formulas because the regression line can be found using a scientific calculator, graphics calculator, online calculator or spreadsheet.

EXAMPLE 9

The table shows the results of a survey into the number of cigarettes people smoke during the year and the number of days they are absent from work.

Cigarettes/year	2000	3000	4000	5000	6000	7000	8000	9000	10 000
Absences	23	27	54	49	63	81	107	128	147

Find the least-squares regression line by using:

a a calculator

b a spreadsheet

ISBN 9780170413220

Solution

a

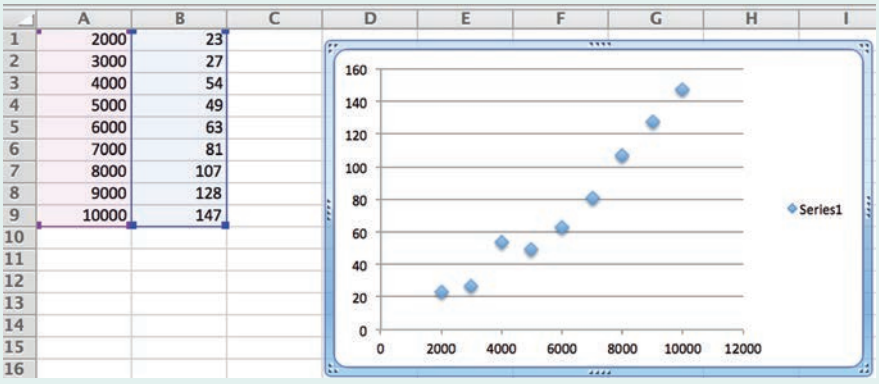
Operation	Casio scientific	Sharp scientific
Enter data.	SHIFT 1 2 : Data 2000 = 3000 = etc. for 1st column 23 = 27 = etc. for 2nd column AC	2000 2ndF STO 23 M+ 3000 2ndF STO 27 M+ etc.
Calculate a.	SHIFT 1 5 : Reg 1 : A =	ALPHA a =
Calculate b.	SHIFT 1 5 : Reg 2 : B =	ALPHA b =

$a = -18.26, \quad b = 0.0156$

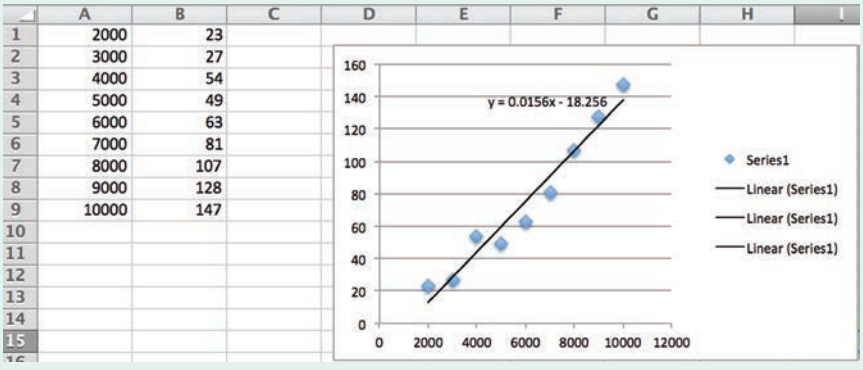
On a scientific calculator, the equation is in the form $y = a + bx$ or $y = bx + a$

So $y = 0.0156x - 18.26$

b Enter the 2 columns in a spreadsheet and draw a scatterplot.



In Chart Layout select Trendline, then Linear trendline. You can display the equation of the line by going to Trendline again, selecting Trendline options and selecting Display equation on chart.



The equation of the least-squares regression line is $y = 0.0156x - 18.256$.

ISBN 9780170413220

Exercise 8.04 Least-squares regression line

1 For each data set, find:

i the correlation coefficient

ii the gradient of the least-squares regression line

a

x	1	2	3	4	5	6	7
y	3	4	7	10	11	15	17

b

x	2	4	6	8	10	12	14	16
y	8	11	19	29	34	41	45	67

2 For each data set:

i find the equation of the least-squares regression line

ii sketch the scatterplot and regression line on the same axes

a Age of machine and breakdown rates

Age (years)	1	2	3	4	5
Breakdowns	0	2	5	9	15

b Length of time in office and popularity of a political party

Time (years)	1	2	3	4	5	6
Popularity (%)	52.3	43.8	43.7	42.1	37.9	37.6

c Length of drought and yield of crops

Time (years)	1	2	3	4	5	6
Yield (t)	107.3	101.8	100.2	87.6	63.5	47.1

d Engine size of cars and number of accidents

Size (L)	1.3	1.8	2.0	2.1	2.4	3.8
Accidents	459	447	513	519	506	625

e Age and number wearing glasses

Age	10	20	30	40	50	60	70	80
Glasses	34	28	41	56	87	105	156	209

ISBN 9780170413220

3 A block of ice was taken out of a freezer and left to thaw. The results are in the table below.

Time t (min)	5	10	15	20	25	30	35	40
Mass m (kg)	23.7	18.8	11.3	8.7	6.2	5.5	2.3	1.5

a Find the correlation coefficient. Is there a high correlation? Is it positive or negative? What does this mean?

b Find the equation of the least-squares regression line.

c Use the equation to estimate the mass of the ice after:

i 18 minutes **ii** an hour

d Discuss why extrapolation may not be useful in this situation.

4 This table shows the weight of gemstones sold at an auction and their selling price.

Weight (carat)	0.05	0.8	1.5	1.7	2.5	2.8	3.1	4.0
Price ($)	144	672	1245	1478	2100	2500	2881	3215

a Draw a scatterplot for this data.

b Find the correlation coefficient. Is there a high correlation between the weight of a gemstone and its cost? Is it positive or negative?

c Find the equation of the least-squares regression line.

d How much would you expect to pay for a 2 carat gemstone?

e How much would you expect a gemstone to weigh if it cost $10 000?

5 The table shows the results of a survey into ages and earnings of a group of people.

Age	15	32	19	28	43	67
Earnings/week ($)	689	1205	840	1154	1587	986

a Draw a scatterplot for this data.

b Find the correlation coefficient.

c Find the equation of the least-squares regression line.

d From this equation find the earnings of a 50-year-old.

e Is this equation a good model to extrapolate?

ISBN 9780170413220

8. TEST YOURSELF

Practice quiz

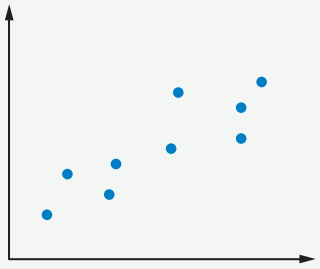
Fitting the data find-a-word

For Questions **1** to **5**, select the correct answer **A**, **B**, **C** or **D**.

1 Which variables are not correlated?

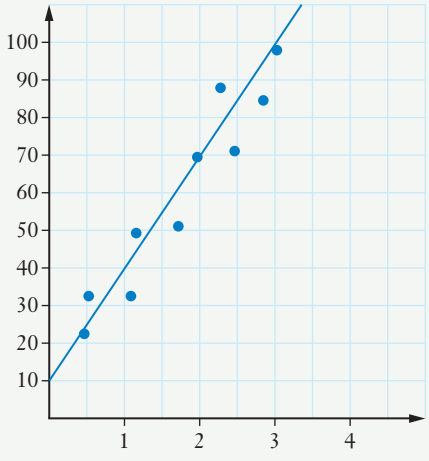
A Hand size and height
B Hours of study and exam results
C Height and hours of employment
D Speed of car and fuel economy

2 Describe the correlation in the scatterplot shown.



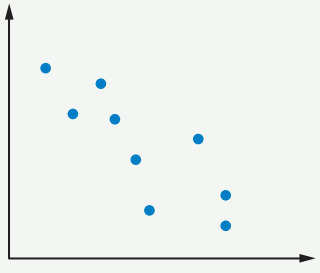
A Weak negative correlation
B Moderate positive correlation
C Moderate negative correlation
D Strong positive correlation

3 Find the equation of the line of best fit.



A $y = 60x + 10$
B $y = 10x + 30$
C $y = 30x - 10$
D $y = 30x + 10$

4 Estimate the correlation coefficient of this bivariate data.



A -0.5
B 0.5
C -1
D 1

5 Using the equation of a line of best fit to predict the value of a variable within the domain of the data set is called:

A extrapolation **B** causality **C** interpolation **D** correlation

ISBN 9780170413220

6 Make a scatterplot of this table of bivariate data, then:

x	1	2	3	4	5	6	7	8	9
y	17	21	24	29	36	43	44	52	58

a draw a line of best fit and find its equation

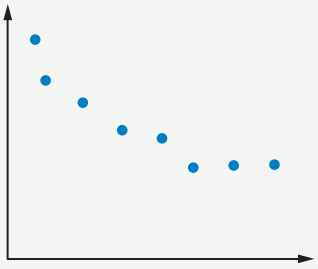
b draw a least-squares regression line and find its equation.

7 For each scatterplot, state whether it has:

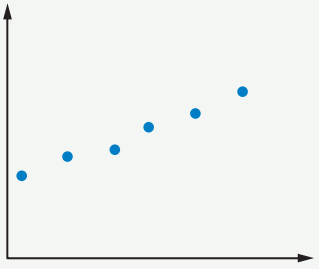
A a positive linear correlation **B** a negative linear correlation

C a non-linear correlation **D** little or no correlation

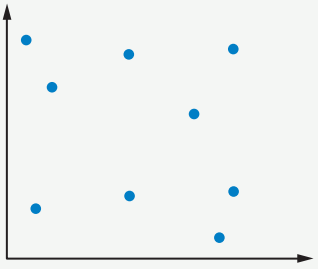
a



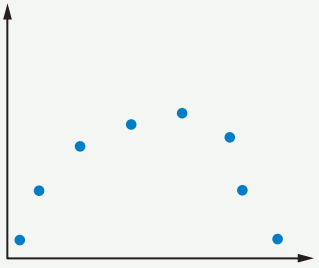
b



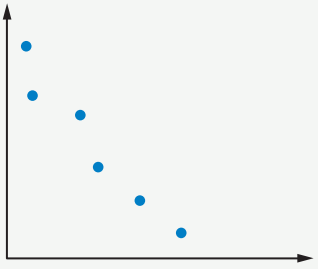
c



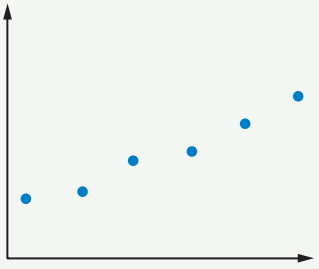
d



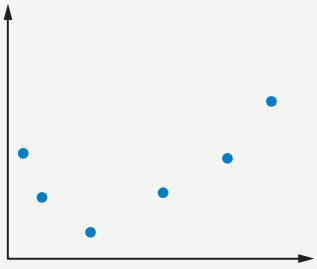
e



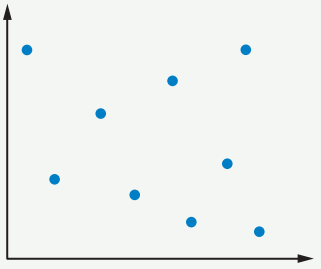
f



g



h



ISBN 9780170413220

8 A group of students was surveyed to find whether there was a correlation between a student's music and maths assessment marks. The results are in the table below.

Music	79	58	91	93	65	43	39	64	82	51
Maths	62	63	82	79	73	57	29	52	76	40

a Find the correlation coefficient.

b Find the equation of the least-squares regression line.

c Using the equation, find the maths assessment mark for a student who scores 60 in music.

d Find the music assessment mark for a student who scores 70 in maths.

9 Determine whether each pair of variables are likely to have a causal relationship.

a Height and amount of food eaten

b Number of years playing sport and weight

c Height and arm length

d Number of chickens and number of eggs

e Size of bookshelves and number of books

ISBN 9780170413220

8. CHALLENGE EXERCISE

1 The equation of the least-squares regression line is given by $y = mx + c$ where $m = r\frac{s_y}{s_x}$ and $c = \overline{y} - m\overline{x}$, where r = correlation coefficient, $\overline{x}$ and $\overline{y}$ are the sample means and s_x and s_y are the sample standard deviations. Evaluate r and $\overline{y}$ given the equation of the least-squares regression line is $y = 2x + 4$, $\overline{x} = 1.2$, $s_x = 1.8$ and $s_y = 4.5$.

2 Sketch a scatterplot that shows a linear correlation of:

a –1 **b** approximately 0.8 **c** 0

d approximately –0.2 **e** 1

3 The table below shows the results of an experiment into the volume of water evaporating from a body of water at different temperatures.

T (°C)	10	15	20	25	30	35	40	45
V (L)	0.5	1.3	2.9	5.8	10.3	15.7	39.8	76.1

a Draw a scatterplot to show this data.

b Why would a least-squares regression line not give a good approximation for this data?

c Use technology or otherwise to find an equation that might approximately model this data.

4 The table shows heights and shoe sizes of several males.

Shoe size	4	5	6	7	8	9	10	11
Height (m)	1.54	1.65	1.68	1.73	1.59	1.82	1.89	1.95

a Draw a scatterplot to show this data.

b Find the correlation coefficient.

c Find the equation of the least-squares regression line.

d The male in the sample with shoe size 8 is found to be an outlier for this data. For the sample without this outlier, find:

i the correlation coefficient

ii the equation of the least-squares regression line

e Is this equation a good model for shoe sizes and height?

ISBN 9780170413220