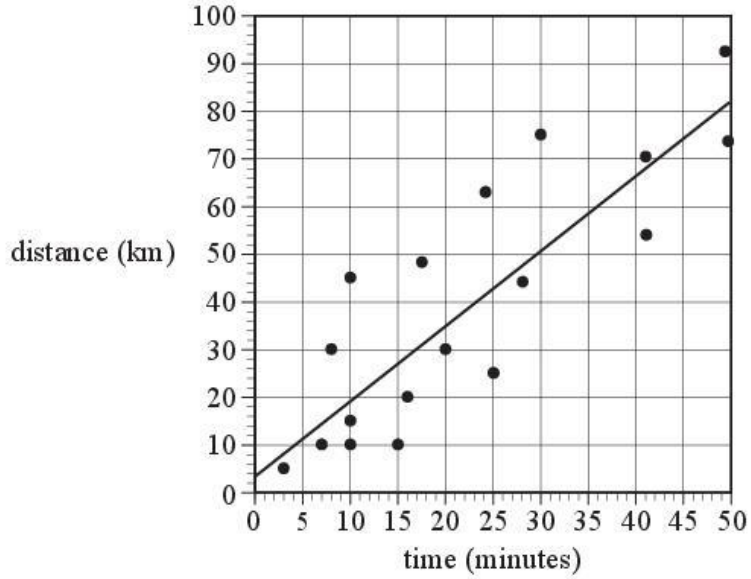
Question 1



A least squares regression line has been fitted to the scatterplot above to enable *distance*, in kilometres, to be predicted from *time*, in minutes.

The equation of this line is closest to

A. $distance = 3.5 + 1.6 \times time$

B. $time = 3.5 + 1.6 \times distance$

C. $distance = 1.6 + 3.5 \times time$

D. $time = 1.8 + 3.5 \times distance$

E. $distance = 3.5 + 1.8 \times time$

Question 2

For a set of bivariate data that involves the variables x and y:

$$r = -0.47, \qquad \bar{x} = 1.8, \qquad s_x = 1.2, \qquad \bar{y} = 7.2, \qquad s_y = 0.85$$

Given the information above, the least squares regression line predicting y from x is closest to

A. $y = 8.4 - 0.66x$

B. $y = 8.4 + 0.66x$

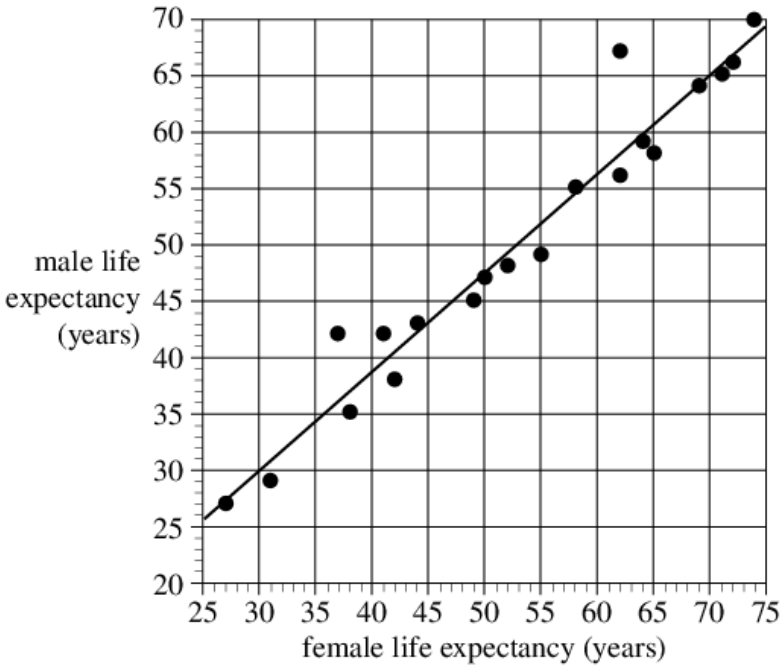
C. $y = 7.8 - 0.33x$

D. $y = 7.8 + 0.33x$

E. $y = 1.8 + 5.4x$

Question 3

The scatterplot below plots male life expectancy (*male*) against female life expectancy (*female*) in 1950 for a number of countries. A least squares regression line has been fitted to the scatterplot as shown.

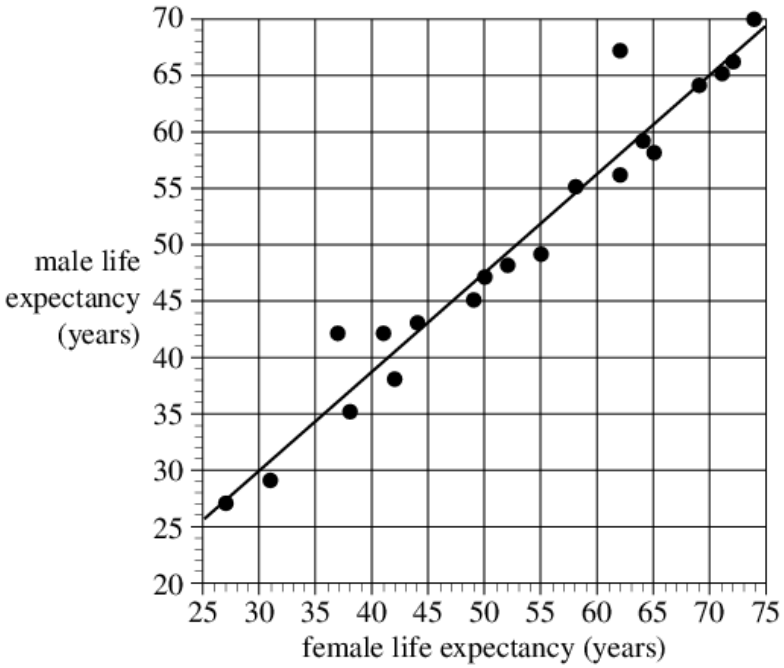


The slope of this least squares regression line is 0.88

a. Interpret the slope in terms of the variables *male* life expectancy and *female* life expectancy. 1 mark

Question 4

The scatterplot below plots male life expectancy (*male*) against female life expectancy (*female*) in 1950 for a number of countries. A least squares regression line has been fitted to the scatterplot as shown.



The slope of this least squares regression line is 0.88

The equation of this least squares regression line is

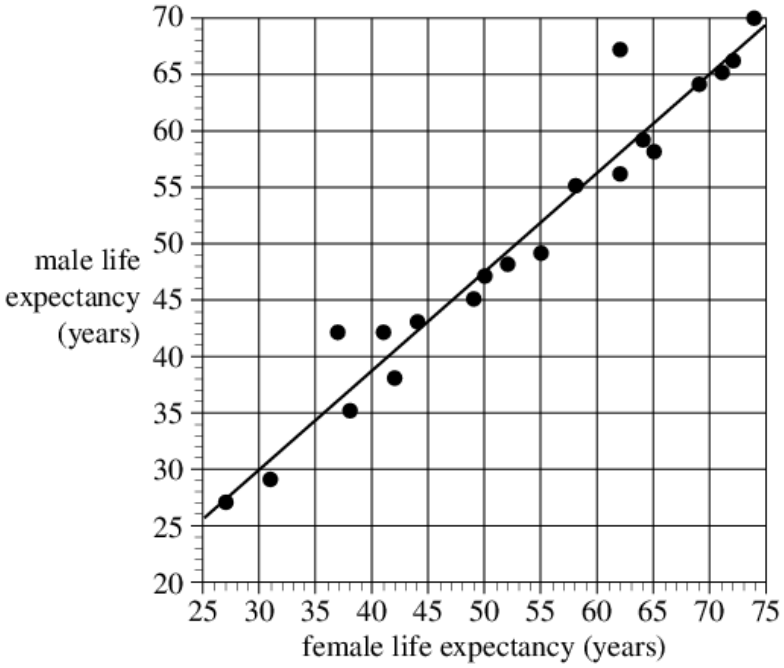
$$male = 3.6 + 0.88 \times female$$

b. In a particular country in 1950, *female* life expectancy was 35 years.

Use the equation to predict *male* life expectancy for that country. 1 mark

Question 5

The scatterplot below plots male life expectancy (*male*) against female life expectancy (*female*) in 1950 for a number of countries. A least squares regression line has been fitted to the scatterplot as shown.



The slope of this least squares regression line is 0.88

The equation of this least squares regression line is

$$male = 3.6 + 0.88 \times female$$

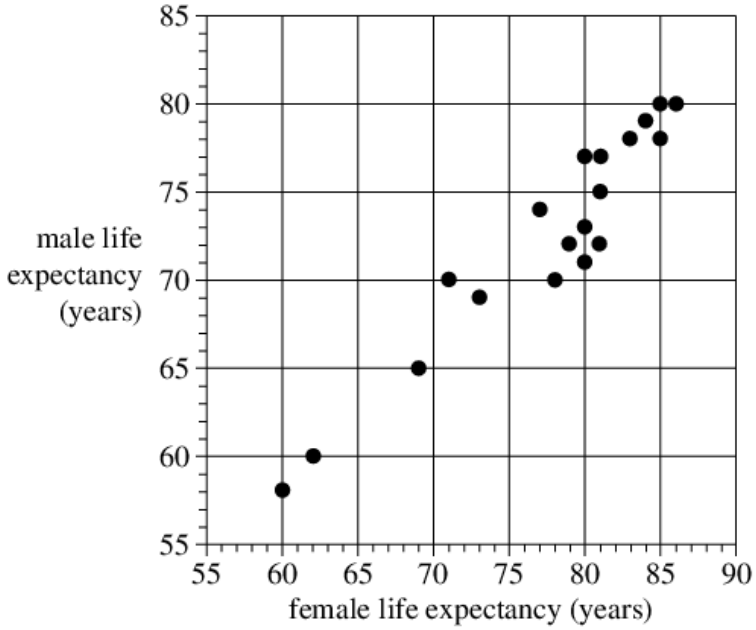
c. The coefficient of determination is 0.95

Interpret the coefficient of determination in terms of male life expectancy and female life expectancy. 1 mark

Question 6

The table below shows male life expectancy (*male*) and female life expectancy (*female*) for a number of countries in 2013. The scatterplot has been constructed from this data.

Life expectancy (in years) in 2013	
male	***female***
80	85
60	62
73	80
70	71
70	78
78	83
77	80
65	69
74	77
70	78
75	81
58	60
80	86
69	73
79	84
72	81
78	85
72	79
77	81
71	80



a. Use the scatterplot to describe the association between *male* life expectancy and *female* life expectancy in terms of strength, direction and form. 1 mark

b. Determine the equation of a least squares regression line that can be used to predict *male* life expectancy from *female* life expectancy for the year 2013.

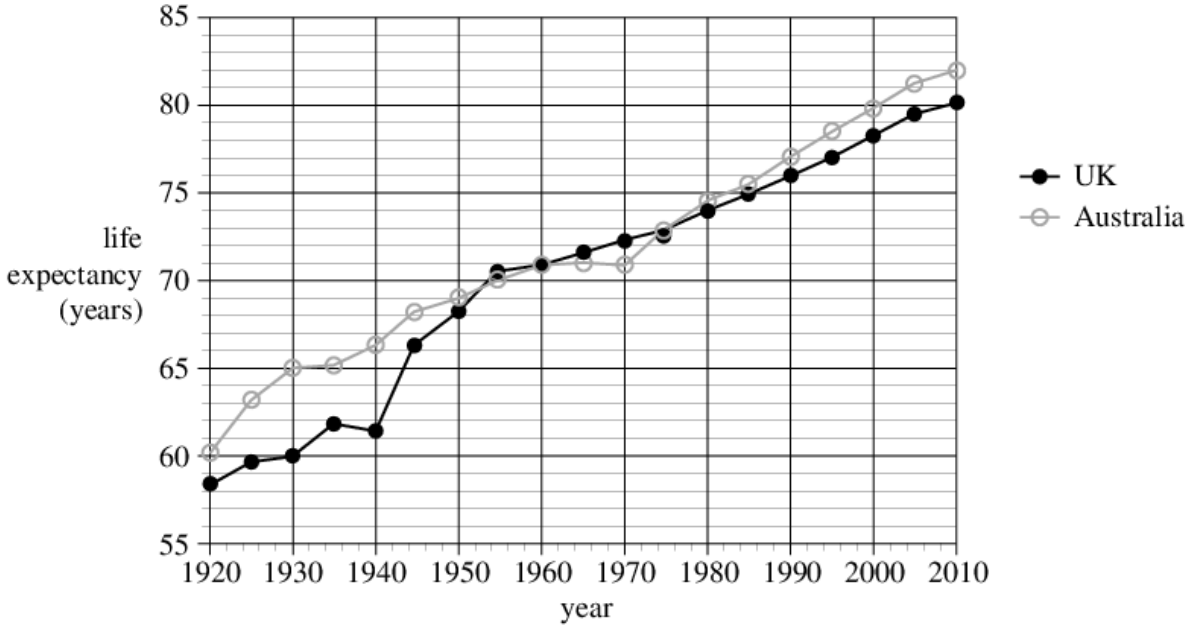
Complete the equation for the least squares regression line below by writing the intercept and slope in the boxes provided.

Write these values correct to two decimal places. 1 mark

male = [ ] + [ ] × *female*

Question 7

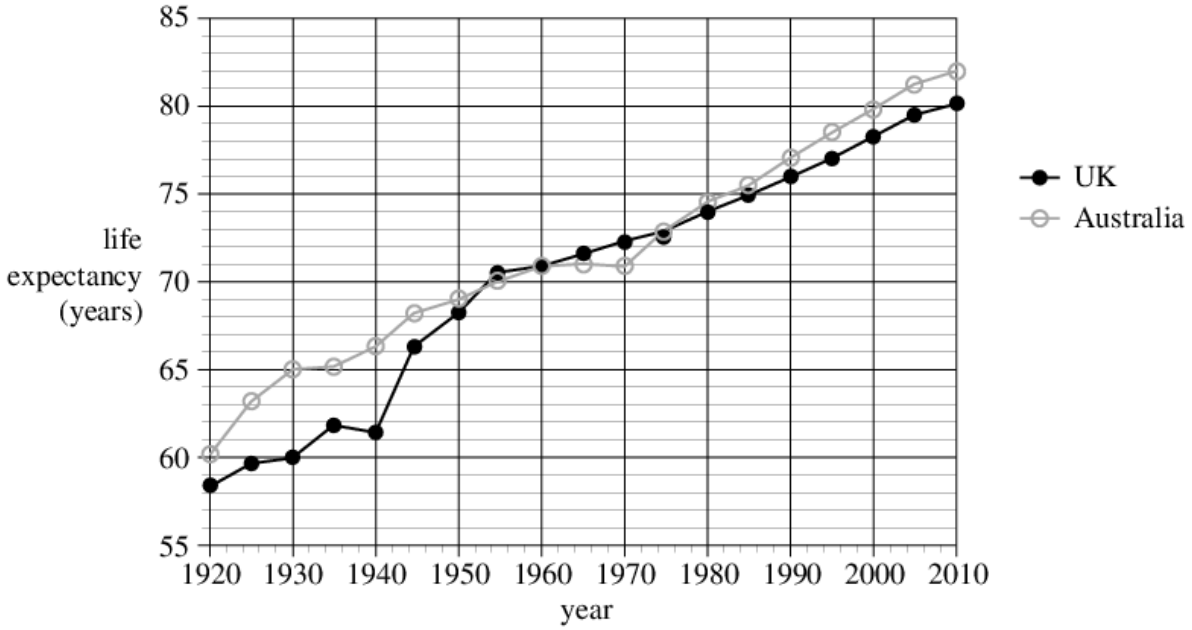
The time series plot below displays the *life expectancy*, in years, of people living in Australia and the United Kingdom (UK) for each *year* from 1920 to 2010.



a. By how much did *life expectancy* in Australia increase during the period 1920 to 2010? Write your answer correct to the nearest year. 1 mark

Question 8

The time series plot below displays the *life expectancy*, in years, of people living in Australia and the United Kingdom (UK) for each *year* from 1920 to 2010.

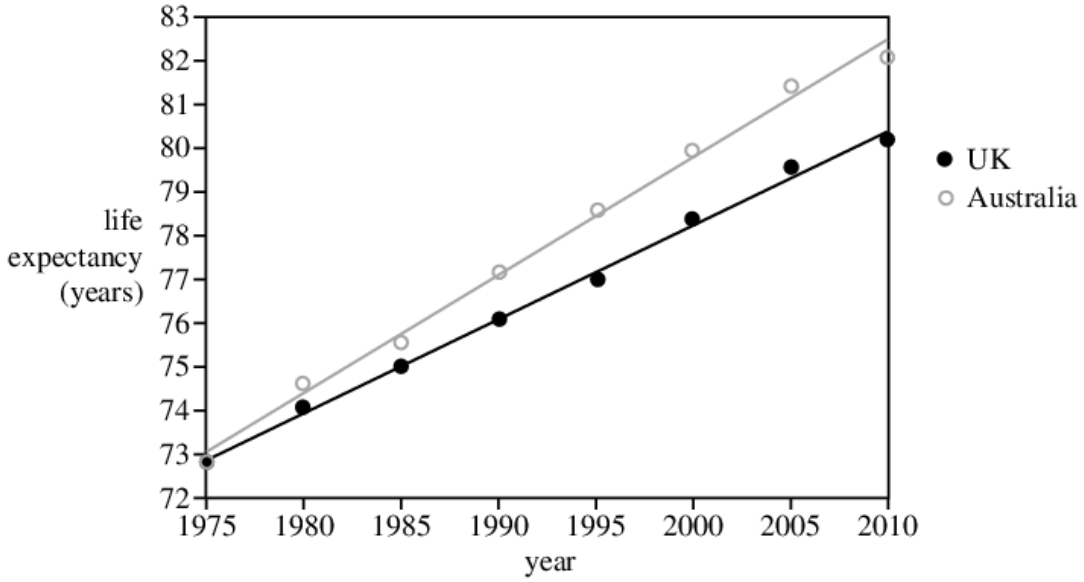


b. In 1975, the life expectancies in Australia and the UK were very similar.

From 1975, the gap between the life expectancies in the two countries increased, with people in Australia having a longer life expectancy than people in the UK.

To investigate the difference in life expectancies, least squares regression lines were fitted to the data for both Australia and the UK for the period 1975 to 2010.

The results are shown below.



The equations of the least squares regression lines are as follows.

Australia: $life\ expectancy = -451.7 + 0.2657 \times year$

UK: $life\ expectancy = -350.4 + 0.2143 \times year$

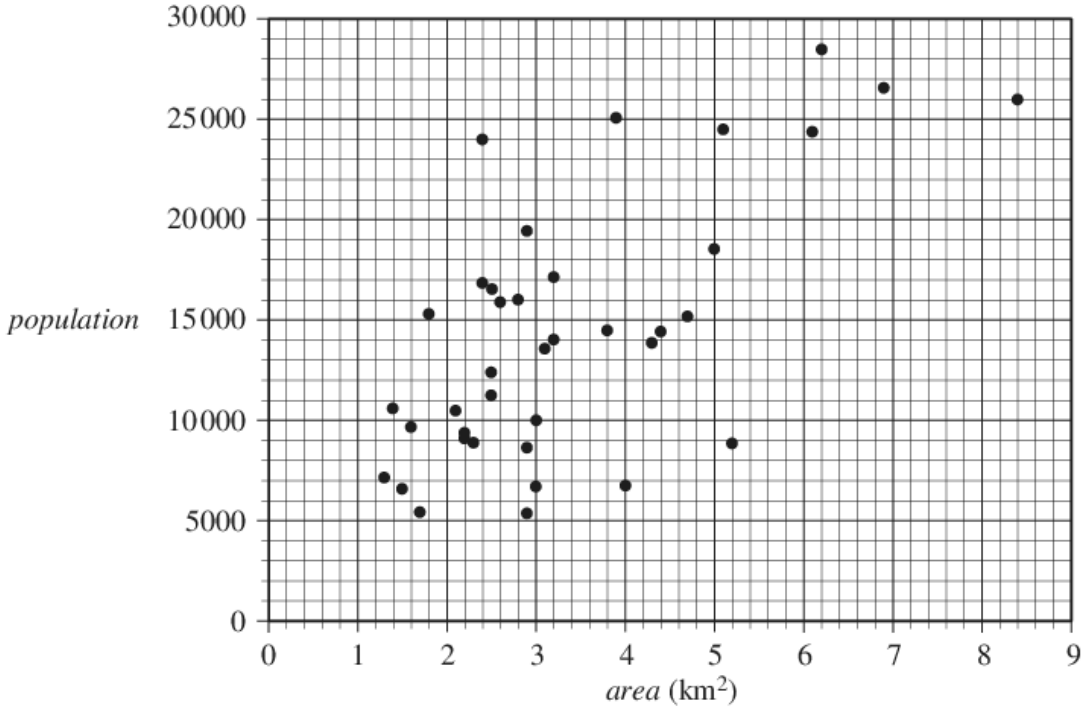
i. Use these equations to predict the difference between the life expectancies of Australia and the UK in 2030.

Give your answer correct to the nearest year. 2 marks

ii. Explain why this prediction may be of limited reliability. 1 mark

Question 9

The scatterplot below shows the *population* and *area* (in square kilometres) of a sample of inner suburbs of a large city.



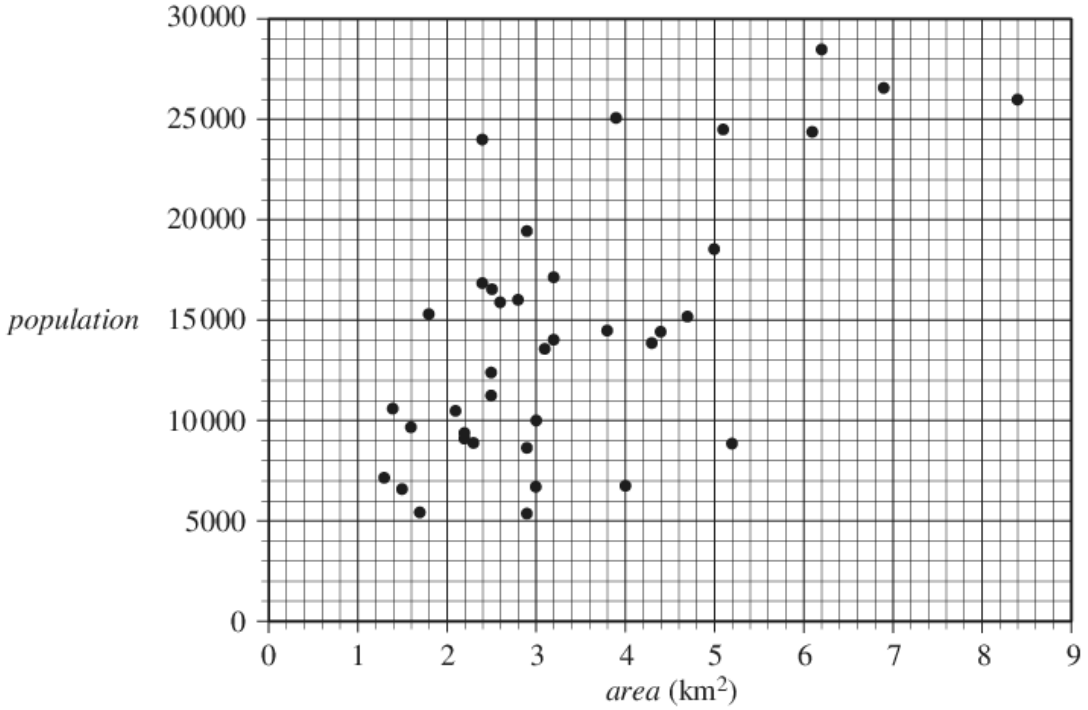
The equation of the least squares regression line for the data in the scatterplot is

$$population = 5330 + 2680 \times area$$

a. Write down the dependent variable. 1 mark

Question 10

The scatterplot below shows the *population* and *area* (in square kilometres) of a sample of inner suburbs of a large city.



The equation of the least squares regression line for the data in the scatterplot is

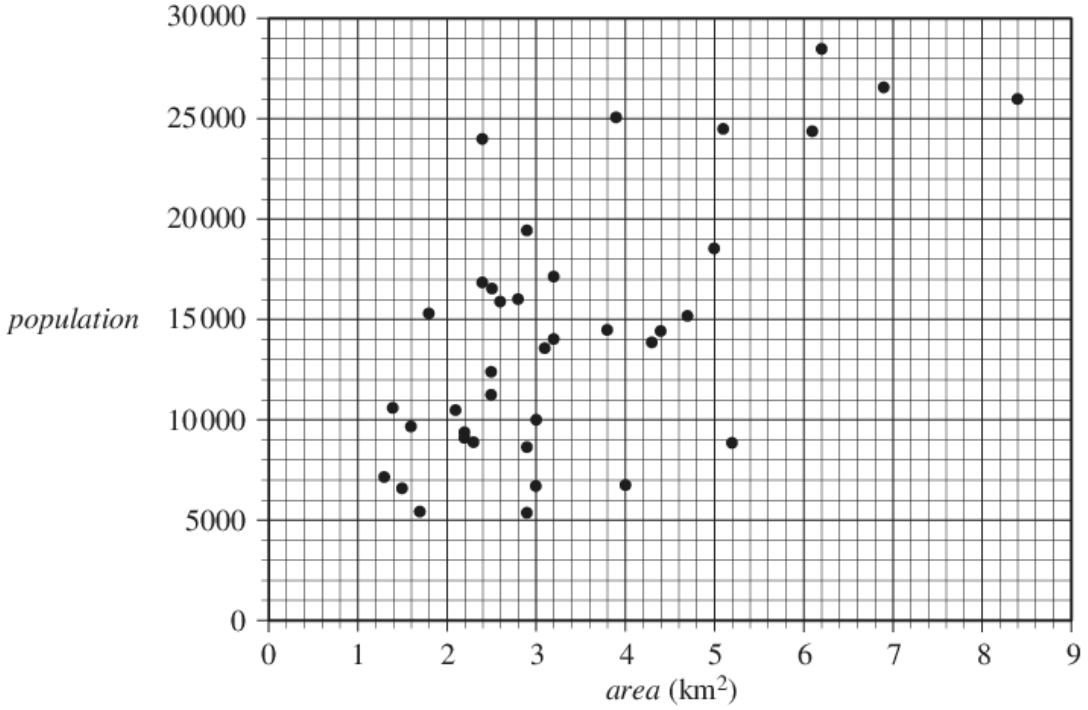
$$population = 5330 + 2680 \times area$$

b. Draw the least squares regression line on the **scatterplot above**. 1 mark

(*Answer on the scatterplot above.*)

Question 11

The scatterplot below shows the *population* and *area* (in square kilometres) of a sample of inner suburbs of a large city.



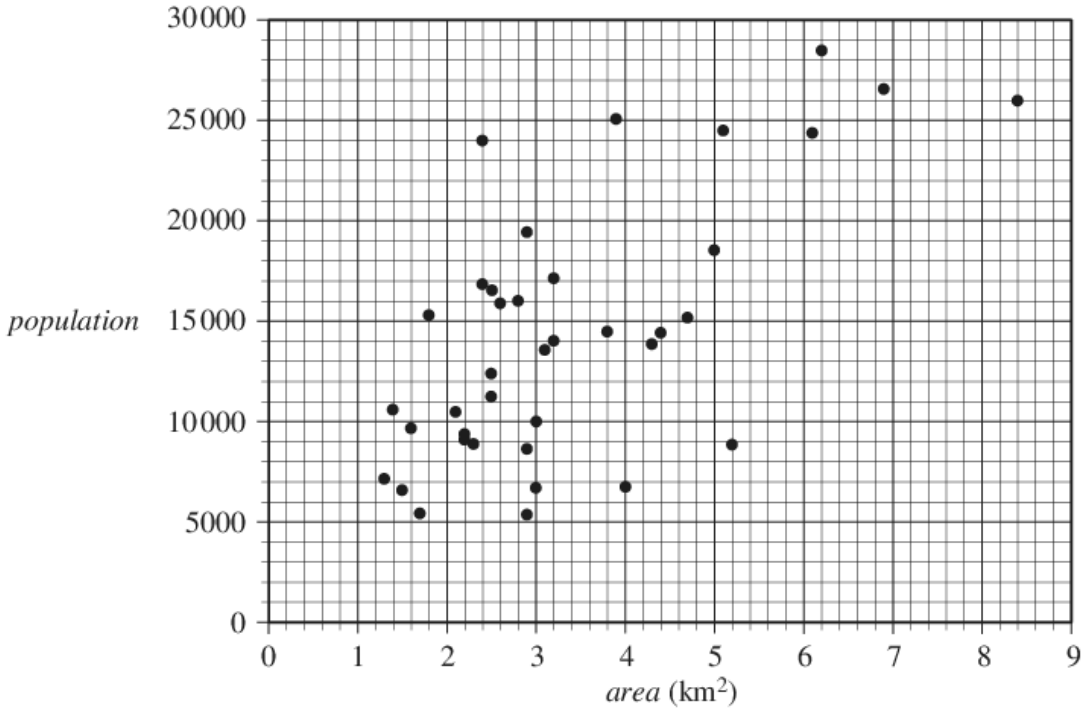
The equation of the least squares regression line for the data in the scatterplot is

$$population = 5330 + 2680 \times area$$

c. Interpret the slope of this least squares regression line in terms of the variables *area* and *population*. **2 marks**

Question 12

The scatterplot below shows the *population* and *area* (in square kilometres) of a sample of inner suburbs of a large city.



The equation of the least squares regression line for the data in the scatterplot is

$$population = 5330 + 2680 \times area$$

d. Wiston is an inner suburb. It has an area of 4 km^2 and a population of 6690.

The correlation coefficient, r, is equal to 0.668

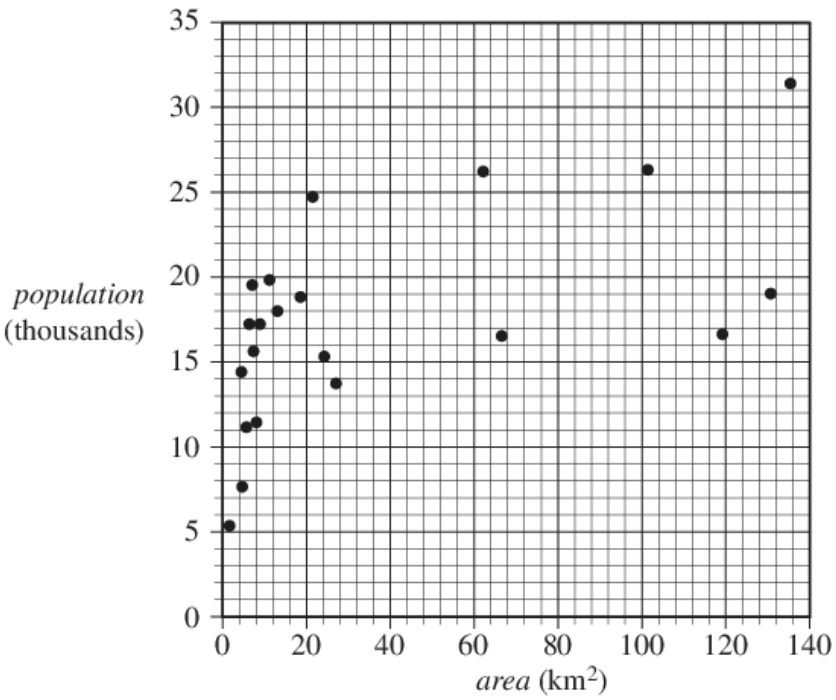
i. Calculate the residual when the least squares regression line is used to predict the population of Wiston from its area. 1 mark

ii. What percentage of the variation in the population of the suburbs is explained by the variation in area?

Write your answer, correct to one decimal place. 1 mark

Question 13

The scatterplot and table below show the *population*, in thousands, and the *area*, in square kilometres, for a sample of 21 outer suburbs of the same city.



Area (km²)	*Population* (thousands)
1.6	5.2
4.4	14.3
4.6	7.5
5.6	11.0
6.3	17.1
7.0	19.4
7.3	15.5
8.0	11.3
8.8	17.1
11.1	19.7
13.0	17.9
18.5	18.7
21.3	24.6
24.2	15.2
27.0	13.6
62.1	26.1
66.5	16.4
101.4	26.2
119.2	16.5
130.7	18.9
135.4	31.3

In the outer suburbs, the relationship between *population* and *area* is non-linear.

A **log** transformation can be applied to the variable *area* to linearise the scatterplot.

a. Apply the **log** transformation to the data and determine the equation of the least squares regression line that allows the population of an outer suburb to be predicted from the logarithm of its area.

Write the slope and intercept of this regression line in the boxes provided below.

Write your answers, correct to one decimal place. 1 mark

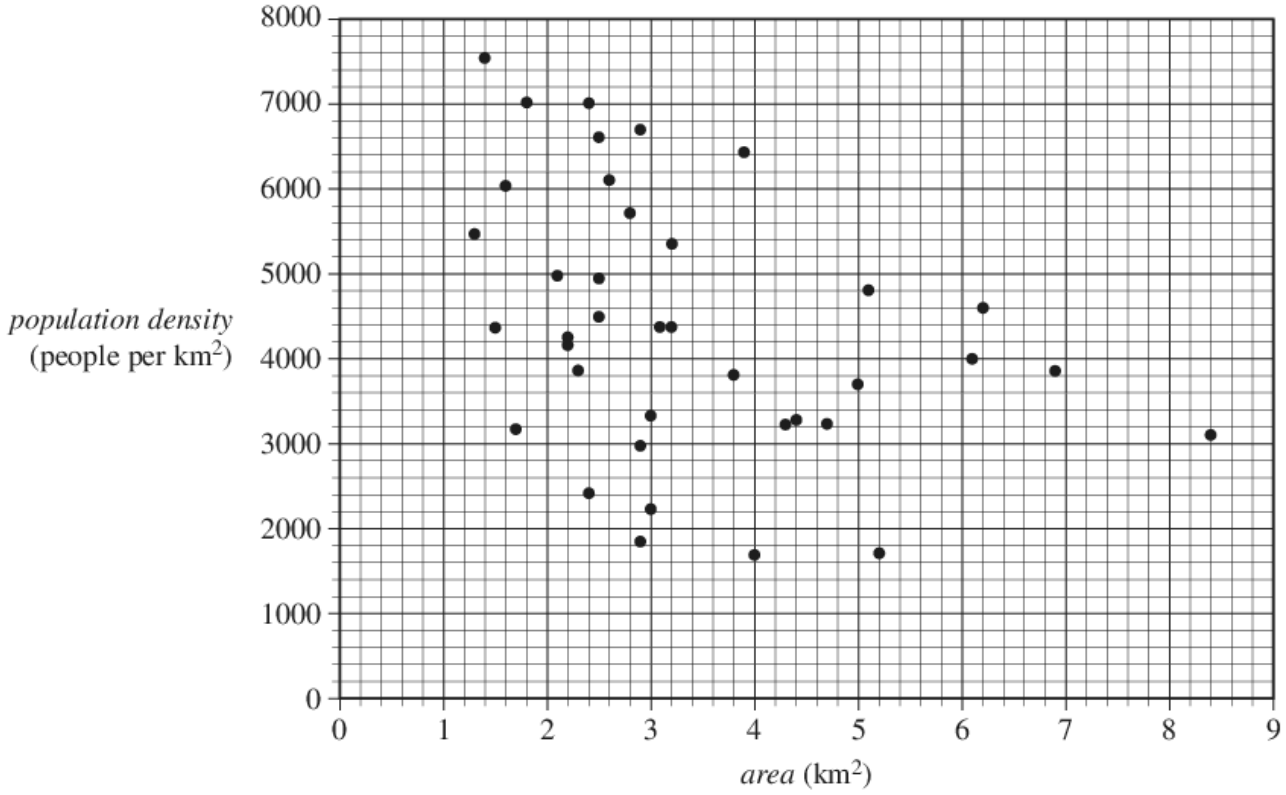
$$population = \boxed{} + \boxed{} \times \log_{10}(area)$$

b. Use this regression equation to predict the population of an outer suburb with an area of 90 km².

Write your answer, correct to the nearest one thousand people. 1 mark

Question 14

The scatterplot below shows the *population density*, in people per square kilometre, and the *area*, in square kilometres, of 38 inner suburbs of the same city.



For this scatterplot, $r^2 = 0.141$

a. Describe the association between the variables *population density* and *area* for these suburbs in terms of strength, direction and form. 1 mark

Question 15

For a city, the correlation coefficient between

- population density and distance from the centre of the city is $r = -0.563$
- house size and distance from the centre of the city is $r = 0.357$.

Given this information, which one of the following statements is true?

A. Around 31.7% of the variation observed in house size in the city can be explained by the variation in distance from the centre of the city.

B. Population density tends to increase as the distance from the centre of the city increases.

C. House sizes tend to be larger as the distance from the centre of the city decreases.

D. The slope of a least squares regression line relating population density to distance from the centre of the city is positive.

E. Population density is more strongly associated with distance from the centre of the city than is house size.

Question 16

The table below shows the hourly rate of pay earned by 10 employees in a company in 1990 and in 2010.

Employee	Hourly rate of pay ($)	
	1990	2010
Ben	9.53	17.02
Lani	9.15	16.71
Freya	8.88	15.10
Jill	8.60	15.93
David	7.67	14.40
Hong	7.96	13.32
Stuart	6.42	15.40
Mei Lien	11.86	19.79
Tim	14.64	23.38
Simon	15.31	25.11

The value of the correlation coefficient, r, for this set of data is closest to

A. 0.74

B. 0.86

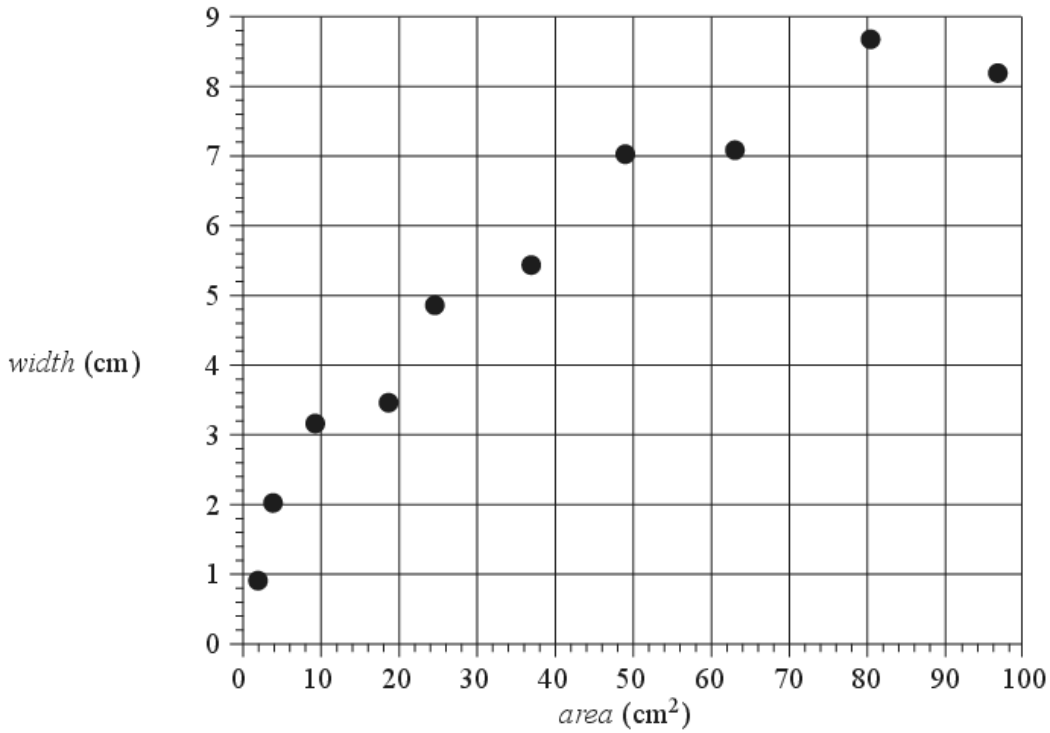
C. 0.92

D. 0.93

E. 0.96

Question 17

The data in the scatterplot below shows the *width*, in cm, and the surface *area*, in cm^2, of leaves sampled from 10 different trees. The scatterplot is non-linear.



To linearise the scatterplot, $(width)^2$ is plotted against *area* and a least squares regression line is then fitted to the linearised plot.

The equation of this least squares regression line is

$$(width)^2 = 1.8 + 0.8 \times area$$

Using this equation, a leaf with a surface area of 120 cm^2 is predicted to have a width, in cm, closest to

A. 9.2

B. 9.9

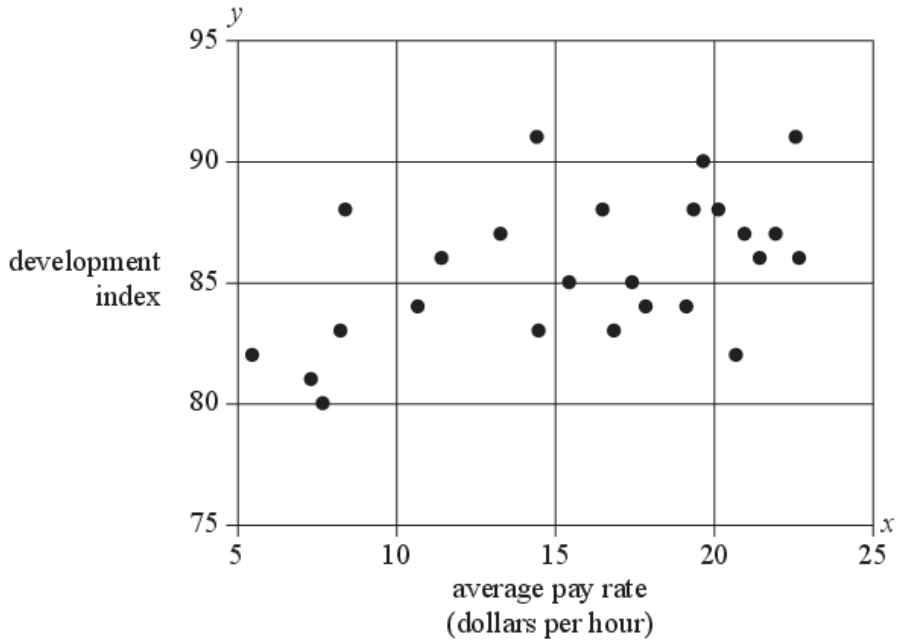
C. 10.6

D. 84.6

E. 97.8

Question 18

The development index and the average pay rate for workers, in dollars per hour, for a selection of 25 countries are displayed in the scatterplot below.



The table below contains the values of some statistics that have been calculated for this data.

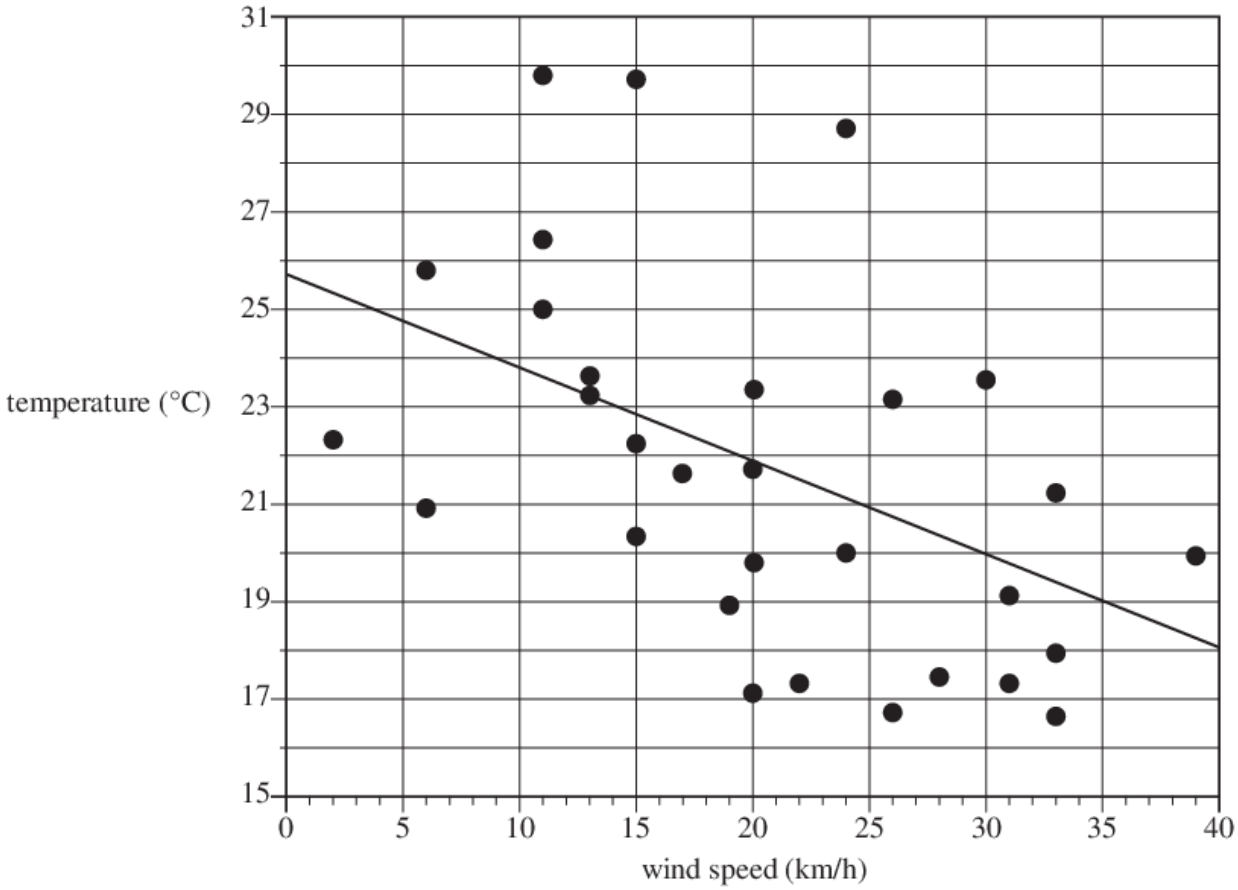
Statistic	Average pay rate (x)	Development index (y)
mean	$\bar{x} = 15.7$	$\bar{y} = 85.6$
standard deviation	$s_x = 5.37$	$s_y = 2.99$
correlation coefficient	$r = 0.488$	

b. Use the information in the table to show that the equation of the least squares regression line for a country's development index, y, in terms of its average pay rate, x, is given by 2 marks

$$y = 81.3 + 0.272x$$

Question 19

The maximum wind speed and maximum temperature were recorded each day for a month. The data is displayed in the scatterplot below and a least squares regression line has been fitted. The dependent variable is *temperature*. The independent variable is *wind speed*.



The equation of the least squares regression line is closest to

A. *temperature* = 25.7 – 0.191 × *wind speed*

B. *wind speed* = 25.7 – 0.191 × *temperature*

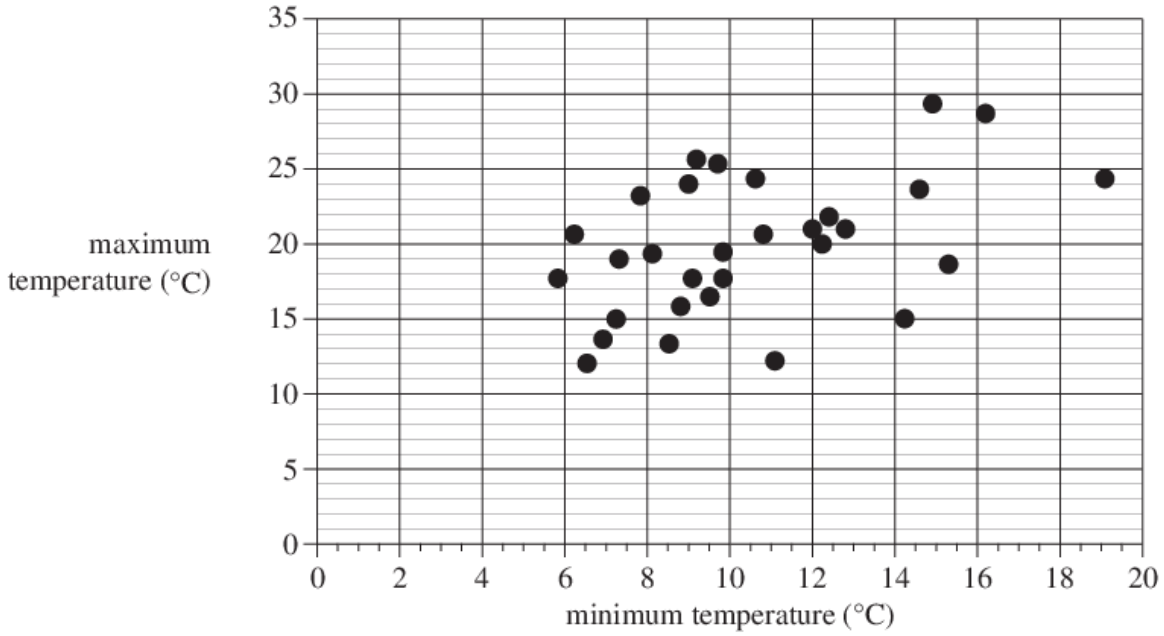
C. *temperature* = 0.191 + 25.7 × *wind speed*

D. *wind speed* = 25.7 + 0.191 × *temperature*

E. *temperature* = 25.7 + 0.191 × *wind speed*

Question 20

The maximum temperature and the minimum temperature at this weather station on each of the 30 days in November 2011 are displayed in the scatterplot below.



The correlation coefficient for this data set is $r = 0.630$.

The equation of the least squares regression line for this data set is

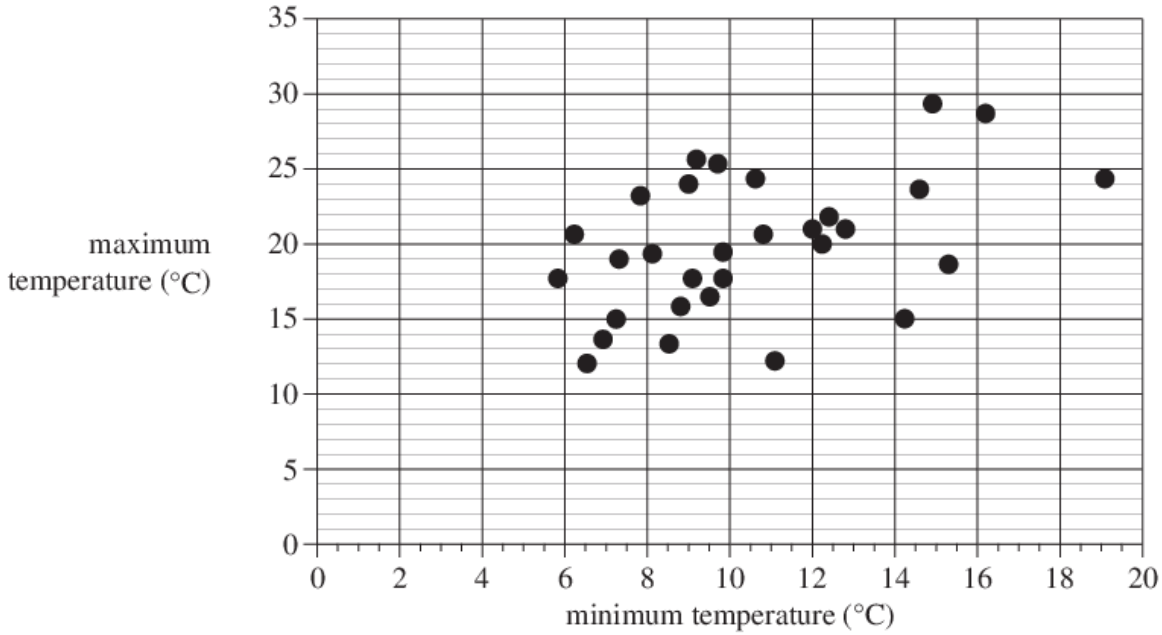
$$\textit{maximum temperature} = 13 + 0.67 \times \textit{minimum temperature}$$

a. **Draw** this least squares regression line on the scatterplot above.

1 mark

Question 21

The maximum temperature and the minimum temperature at this weather station on each of the 30 days in November 2011 are displayed in the scatterplot below.



The correlation coefficient for this data set is $r = 0.630$.

The equation of the least squares regression line for this data set is

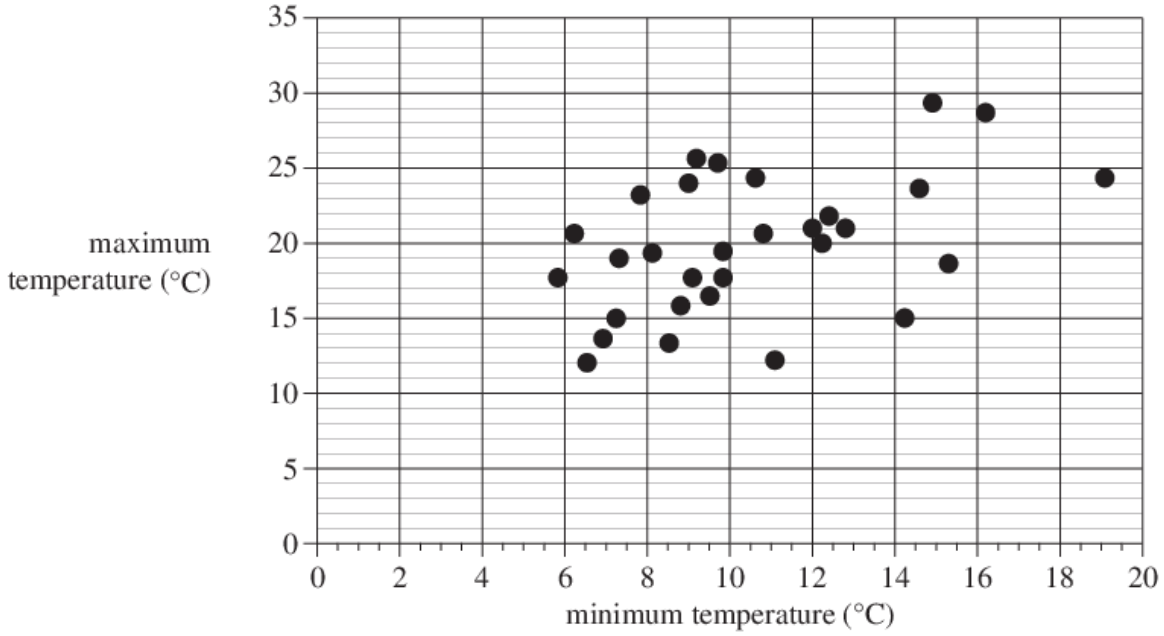
$$\textit{maximum temperature} = 13 + 0.67 \times \textit{minimum temperature}$$

b. Interpret the vertical intercept of the least squares regression line in terms of maximum temperature and minimum temperature.

1 mark

Question 22

The maximum temperature and the minimum temperature at this weather station on each of the 30 days in November 2011 are displayed in the scatterplot below.



The correlation coefficient for this data set is $r = 0.630$.

The equation of the least squares regression line for this data set is

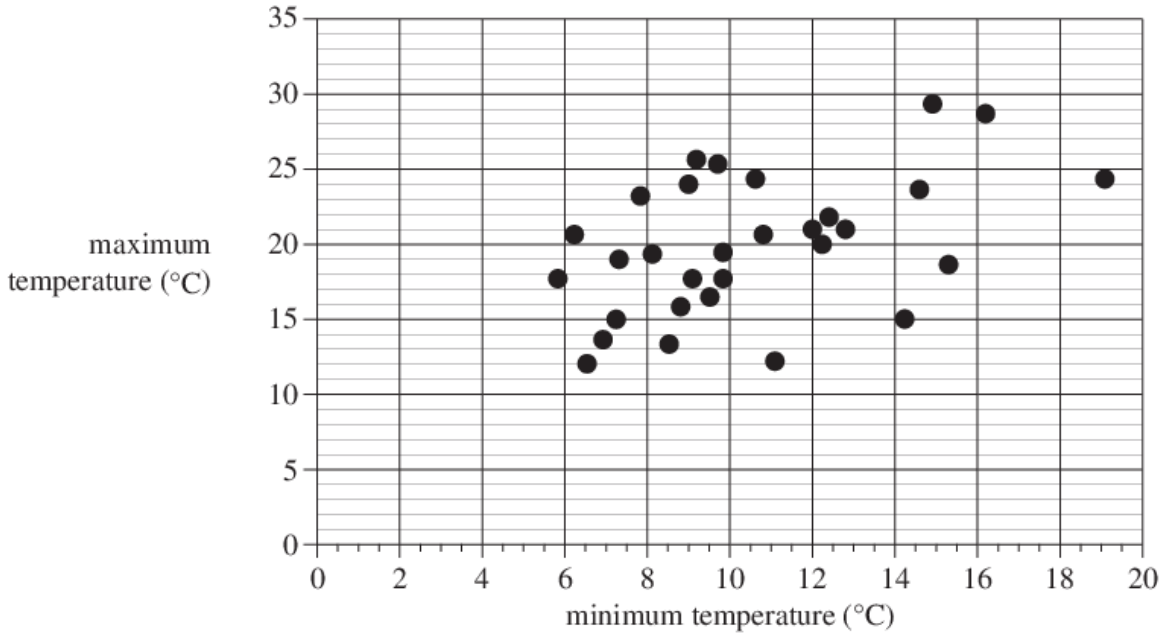
$$\textit{maximum temperature} = 13 + 0.67 \times \textit{minimum temperature}$$

c. Describe the relationship between the maximum temperature and the minimum temperature in terms of strength and direction.

1 mark

Question 23

The maximum temperature and the minimum temperature at this weather station on each of the 30 days in November 2011 are displayed in the scatterplot below.



The correlation coefficient for this data set is $r = 0.630$.

The equation of the least squares regression line for this data set is

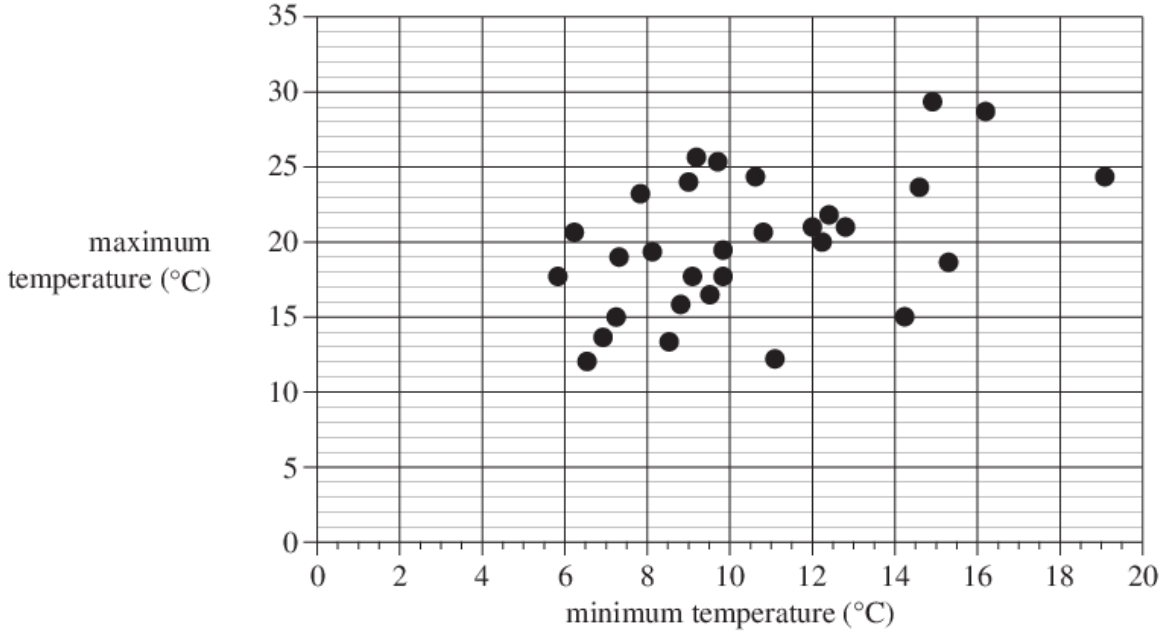
$$\textit{maximum temperature} = 13 + 0.67 \times \textit{minimum temperature}$$

d. Interpret the slope of the least squares regression line in terms of maximum temperature and minimum temperature.

1 mark

Question 24

The maximum temperature and the minimum temperature at this weather station on each of the 30 days in November 2011 are displayed in the scatterplot below.



The correlation coefficient for this data set is $r = 0.630$.

The equation of the least squares regression line for this data set is

$$\textit{maximum temperature} = 13 + 0.67 \times \textit{minimum temperature}$$

e. Determine the percentage of variation in the maximum temperature that may be explained by the variation in the minimum temperature.

Write your answer, correct to the nearest percentage.

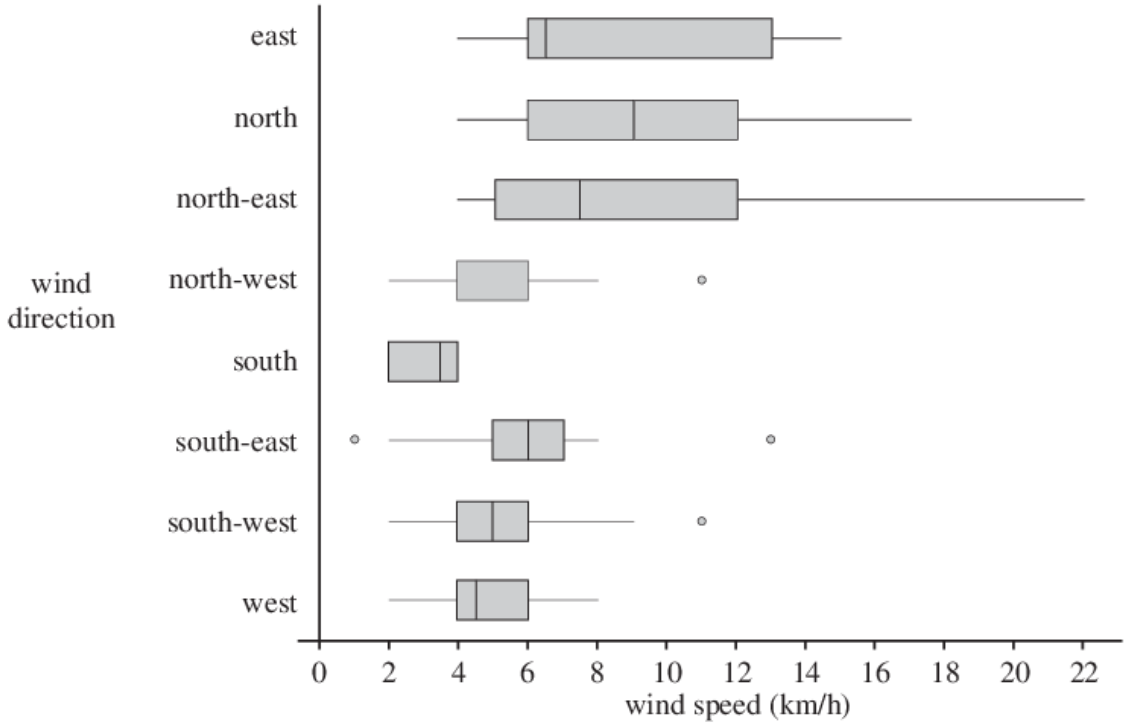
1 mark

Question 25

A weather station records the wind speed and the wind direction each day at 9.00 am.

The wind speed is recorded, correct to the nearest whole number.

The parallel boxplots below have been constructed from data that was collected on the 214 days from June to December in 2011.



a. Complete the following statements.

The wind direction with the lowest recorded wind speed was _____.

The wind direction with the largest range of recorded wind speeds was _____.

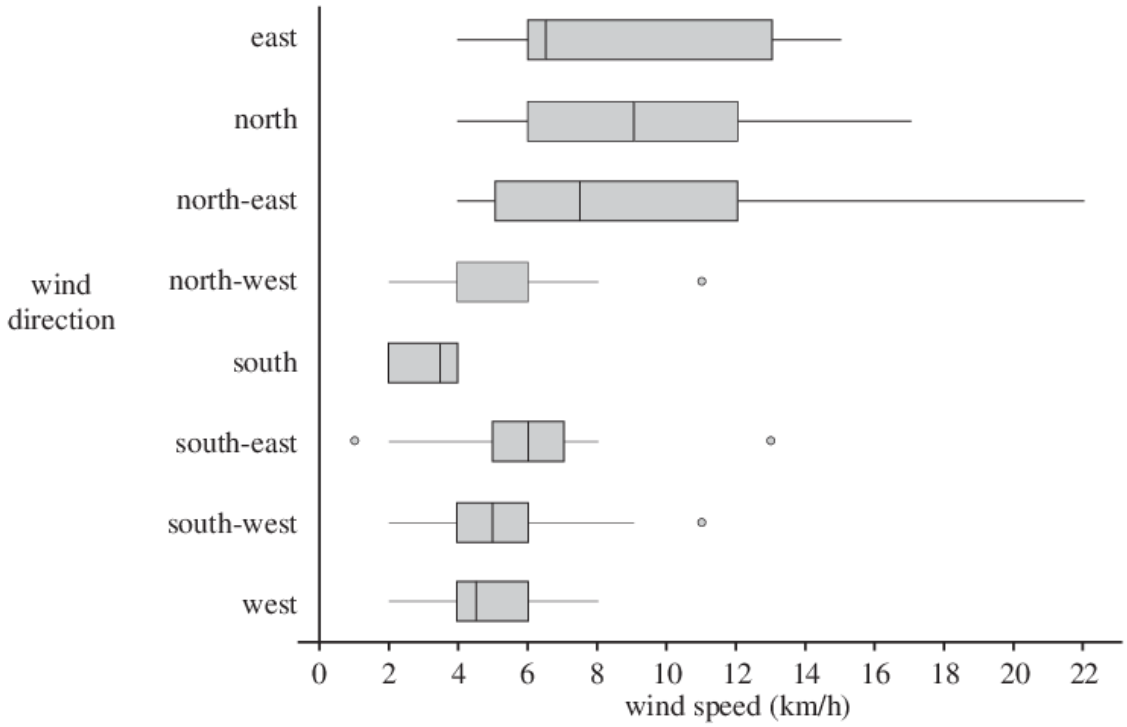
1 mark

Question 26

A weather station records the wind speed and the wind direction each day at 9.00 am.

The wind speed is recorded, correct to the nearest whole number.

The parallel boxplots below have been constructed from data that was collected on the 214 days from June to December in 2011.



b. The wind blew from the south on eight days.

Reading from the parallel boxplots above we know that, for these eight wind speeds, the

first quartile	$Q_1 = 2$ km/h
median	$M = 3.5$ km/h
third quartile	$Q_3 = 4$ km/h

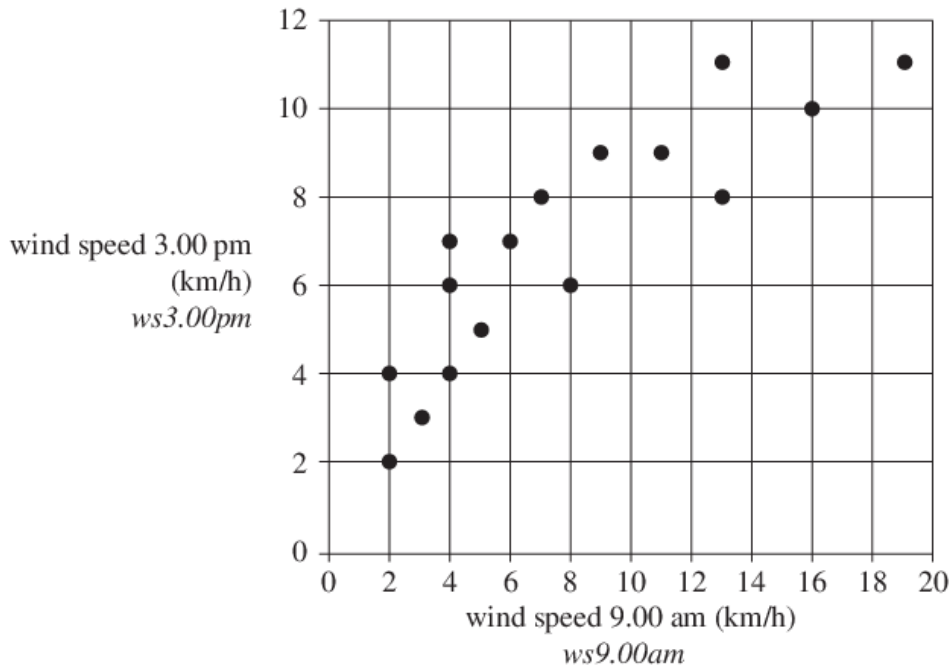
Given that the eight wind speeds were recorded to the nearest whole number, write down the eight wind speeds.

1 mark

Question 27

The wind speeds (in km/h) that were recorded at the weather station at 9.00 am and 3.00 pm respectively on 18 days in November are given in the table below. A scatterplot has been constructed from this data set.

Wind speed (km/h)	
9.00 am	**3.00 pm**
2	2
4	6
4	7
4	4
13	11
6	7
3	3
16	10
6	7
13	8
11	9
2	4
7	8
5	5
8	6
6	7
19	11
9	9



Let the wind speed at 9.00 am be represented by the variable *ws9.00am* and the wind speed at 3.00 pm be represented by the variable *ws3.00pm*.

The relationship between *ws9.00am* and *ws3.00pm* shown in the scatterplot above is nonlinear.

A **squared transformation** can be applied to the variable *ws3.00pm* to linearise the data in the scatterplot.

a. Apply the squared transformation to the variable *ws3.00pm* and determine the equation of the least squares regression line that allows $(ws3.00pm)^2$ to be predicted from *ws9.00am*.

In the boxes provided, write the coefficients for this equation, correct to one decimal place.

$$(ws3.00pm)^2 = \boxed{} + \boxed{} \times ws9.00am$$

2 marks

b. Use this equation to predict the wind speed at 3.00 pm on a day when the wind speed at 9.00 am is 24 km/h.

Write your answer, correct to the nearest whole number.

1 mark

Question 28

When blood pressure is measured, both the systolic (or maximum) pressure and the diastolic (or minimum) pressure are recorded.

Table 1 displays the blood pressure readings, in mmHg, that result from fifteen successive measurements of the same person's blood pressure.

Table 1

	Blood pressure	
Reading number	**systolic**	**diastolic**
1	121	73
2	126	75
3	141	73
4	125	73
5	122	67
6	126	74
7	129	70
8	130	72
9	125	69
10	121	65
11	118	66
12	134	77
13	125	70
14	127	64
15	119	69

Using systolic blood pressure (*systolic*) as the dependent variable, and diastolic blood pressure (*diastolic*) as the independent variable, a least squares regression line is fitted to the data in Table 1.

The equation of the least squares regression line is closest to

A. $systolic = 70.3 + 0.790 \times diastolic$

B. $diastolic = 70.3 + 0.790 \times systolic$

C. $systolic = 29.3 + 0.330 \times diastolic$

D. $diastolic = 0.330 + 29.3 \times systolic$

E. $systolic = 0.790 + 70.3 \times diastolic$

Question 29

When blood pressure is measured, both the systolic (or maximum) pressure and the diastolic (or minimum) pressure are recorded.

Table 1 displays the blood pressure readings, in mmHg, that result from fifteen successive measurements of the same person's blood pressure.

Table 1

	Blood pressure	
Reading number	**systolic**	**diastolic**
1	121	73
2	126	75
3	141	73
4	125	73
5	122	67
6	126	74
7	129	70
8	130	72
9	125	69
10	121	65
11	118	66
12	134	77
13	125	70
14	127	64
15	119	69

From the fifteen blood pressure measurements for this person, it can be concluded that the percentage of the variation in systolic blood pressure that is explained by the variation in diastolic blood pressure is closest to

A. 25.8%

B. 50.8%

C. 55.4%

D. 71.9%

E. 79.0%

Question 30

For a group of 15-year-old students who regularly played computer games, the correlation between the time spent playing computer games and fitness level was found to be $r = -0.56$.

On the basis of this information it can be concluded that

A. 56% of these students were not very fit.

B. these students would become fitter if they spent less time playing computer games.

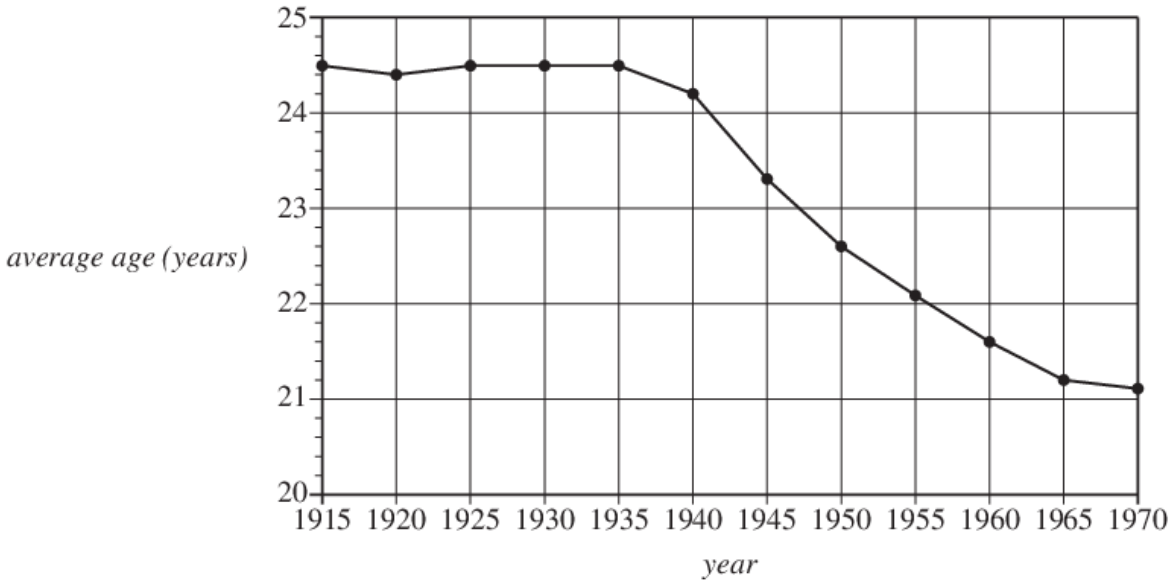
C. these students would become fitter if they spent more time playing computer games.

D. the students in the group who spent a short amount of time playing computer games tended to be fitter.

E. the students in the group who spent a large amount of time playing computer games tended to be fitter.

Question 31

The following time series plot shows the average age of women at first marriage in a particular country during the period 1915 to 1970.



a. Use this plot to describe, in general terms, the way in which the average age of women at first marriage in this country has changed during the period 1915 to 1970.

1 mark

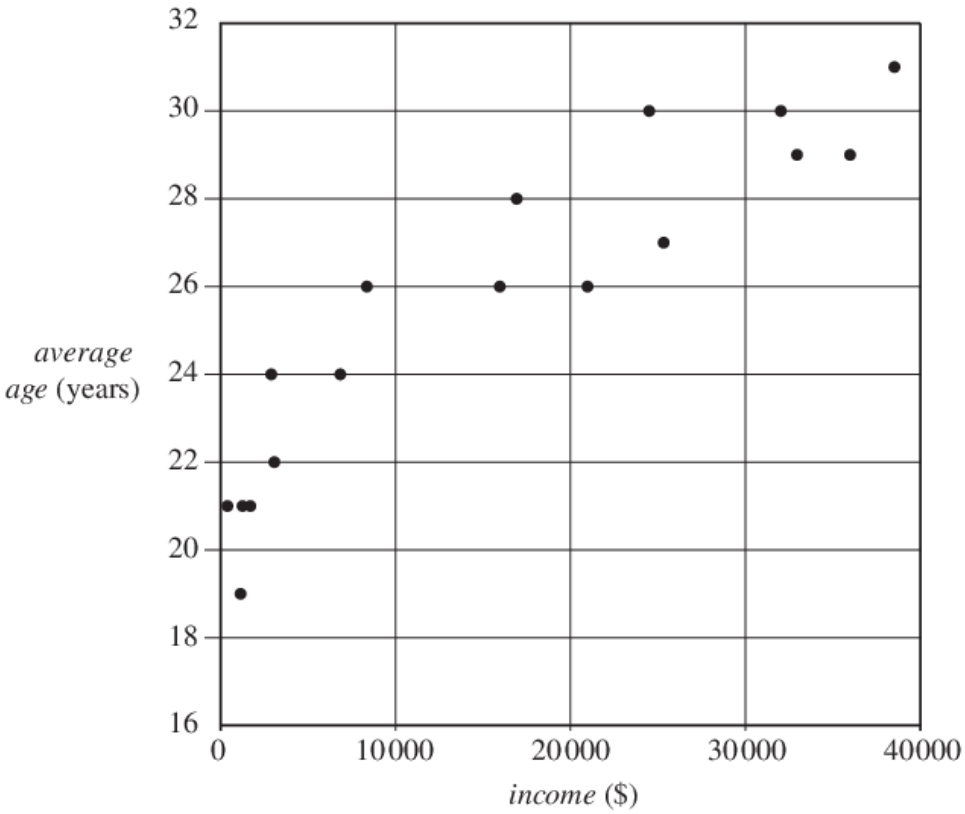
Question 32

The average age of women at first marriage in years (*average age*) and average yearly income in dollars per person (*income*) were recorded for a group of 17 countries.

The results are displayed in Table 2. A scatterplot of the data is also shown.

Table 2

average age (years)	*income* ($)
21	1750
22	3200
26	8600
26	16000
28	17000
26	21000
30	24500
30	32000
31	38500
29	33000
27	25500
29	36000
19	1300
21	600
24	3050
24	6900
21	1400



The relationship between *average age* and *income* is nonlinear.

A **log transformation** can be applied to the variable *income* and used to linearise the scatterplot.

a. Apply this log transformation to the data and determine the equation of the least squares regression line that allows *average age* to be predicted from log (*income*).

Write the coefficients for this equation, correct to two decimal places, in the spaces provided.

average age = [ ] + [ ] × log (*income*)

2 marks

b. Use this equation to predict the average age of women at first marriage in a country with an average yearly income of $20000 per person.

Write your answer correct to one decimal place.

1 mark

Question 33

For a set of bivariate data that involves the variables x and y, with y as the dependent variable

$$r = -0.644, \quad \bar{x} = 5.30, \quad \bar{y} = 5.60, \quad s_x = 3.06, \quad s_y = 3.20$$

The equation of the least squares regression line is closest to

A. $y = 9.2 - 0.7x$

B. $y = 9.2 + 0.7x$

C. $y = 2.0 - 0.6x$

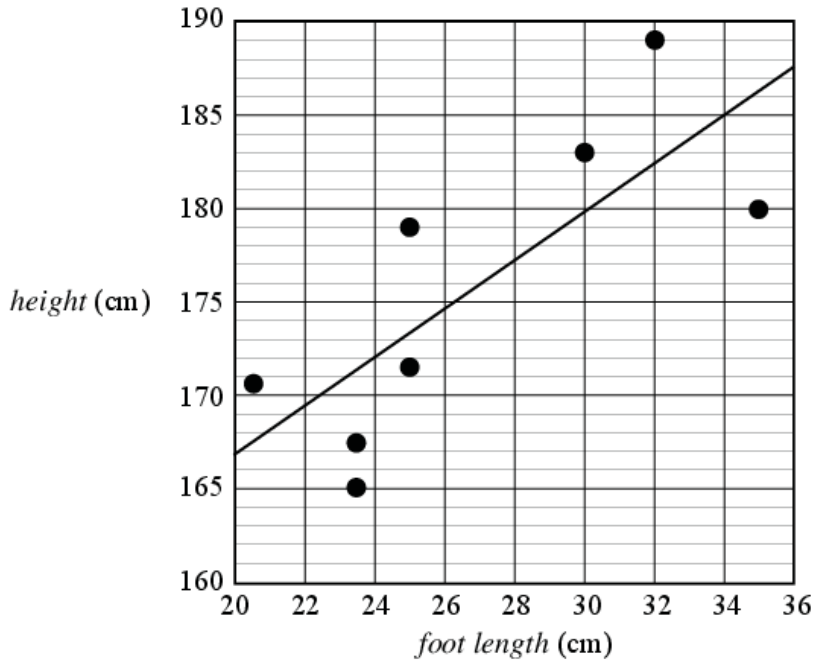
D. $y = 2.0 - 0.7x$

E. $y = 2.0 + 0.7x$

Question 34

The *height* (in cm) and *foot length* (in cm) for each of eight Year 12 students were recorded and displayed in the scatterplot below.

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



The independent variable is *foot length*.

The equation of the least squares regression line is closest to

A. *height* = –110 + 0.78 × *foot length*.

B. *height* = 141 + 1.3 × *foot length*.

C. *height* = 167 + 1.3 × *foot length*.

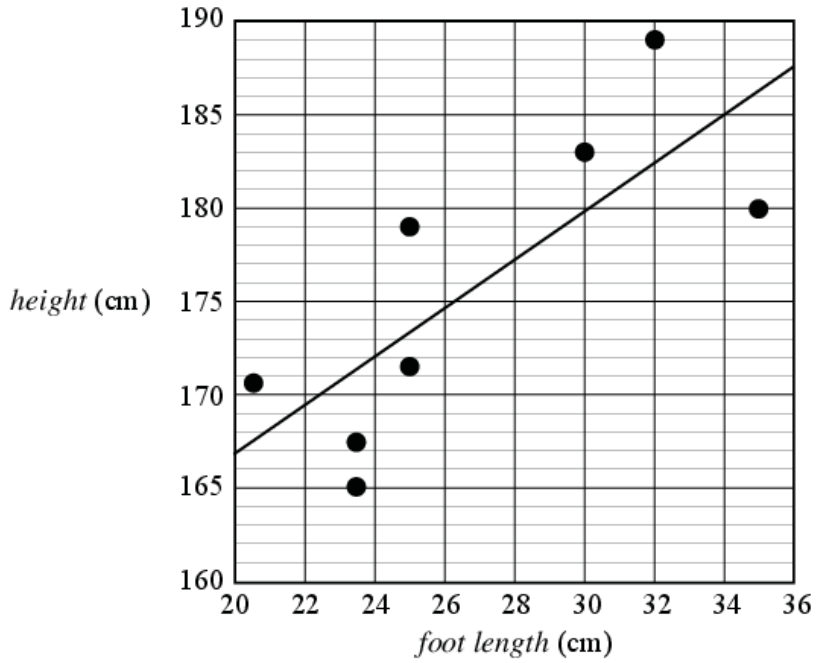
D. *height* = 167 + 0.67 × *foot length*.

E. *foot length* = 167 + 1.3 × *height*.

Question 35

The *height* (in cm) and *foot length* (in cm) for each of eight Year 12 students were recorded and displayed in the scatterplot below.

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



By inspection, the value of the product-moment correlation coefficient (r) for this data is closest to

A. 0.98

B. 0.78

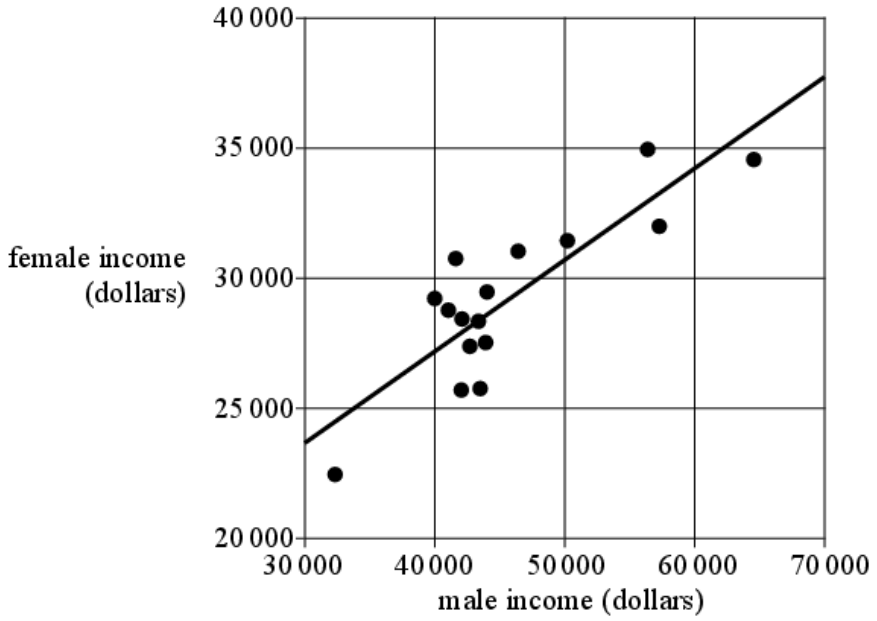
C. 0.23

D. –0.44

E. –0.67

Question 36

In the scatterplot below, average annual *female income*, in dollars, is plotted against average annual *male income*, in dollars, for 16 countries. A least squares regression line is fitted to the data.



The equation of the least squares regression line for predicting female income from male income is

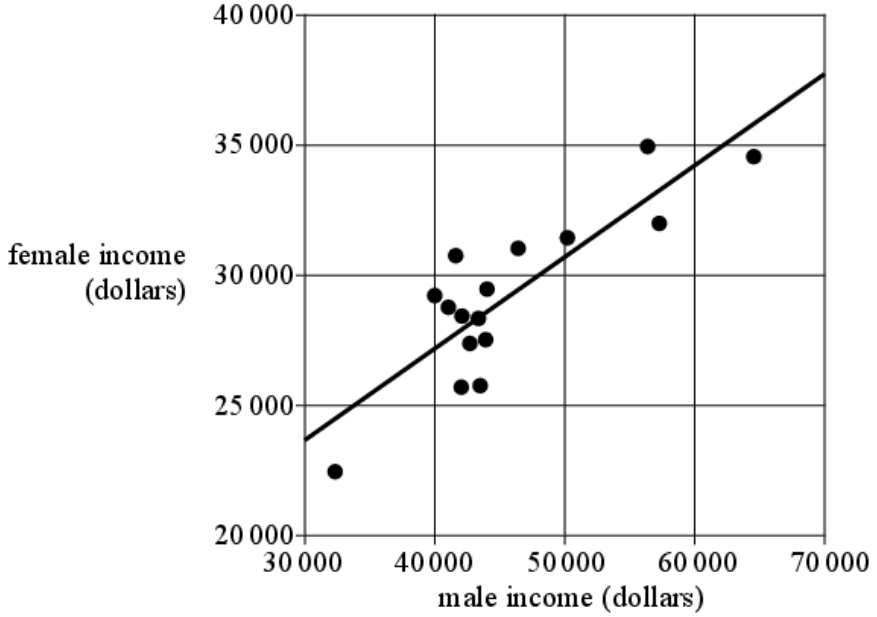
$$\textit{female income} = 13\,000 + 0.35 \times \textit{male income}$$

a. What is the independent variable?

1 mark

Question 37

In the scatterplot below, average annual *female income*, in dollars, is plotted against average annual *male income*, in dollars, for 16 countries. A least squares regression line is fitted to the data.



The equation of the least squares regression line for predicting female income from male income is

$$\textit{female income} = 13\,000 + 0.35 \times \textit{male income}$$

c. **i.** Use the least squares regression line equation to predict the average annual female income (in dollars) in a country where the average annual male income is $15 000.

ii. The prediction made in **part c. i.** is not likely to be reliable.
Explain why.

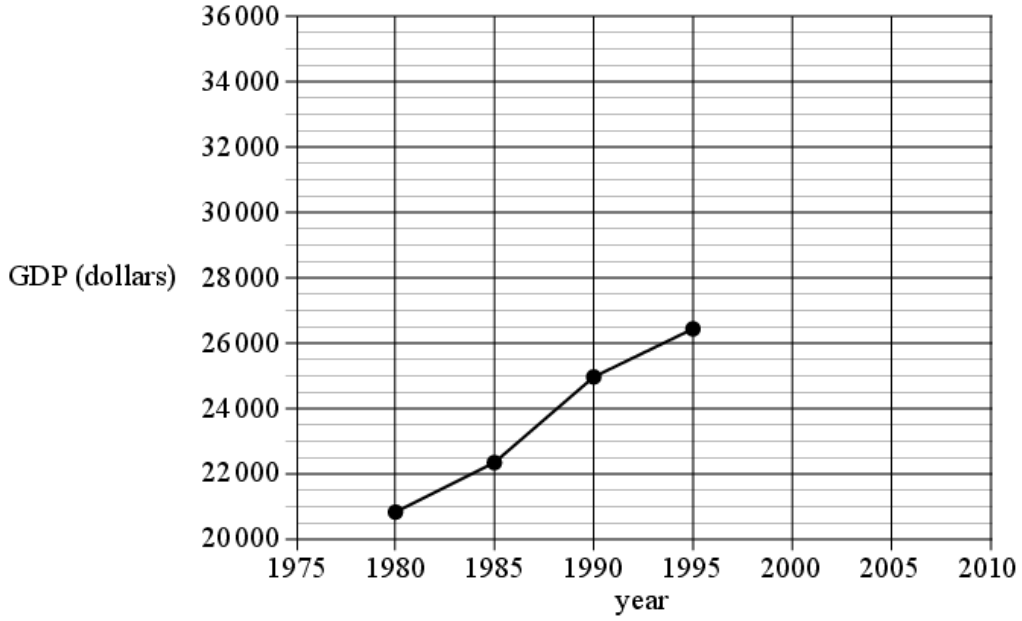
1 + 1 = 2 marks

Question 38

Table 2 shows the Australian gross domestic product (GDP) per person, in dollars, at five yearly intervals for the period 1980 to 2005.

Table 2

Year	1980	1985	1990	1995	2000	2005
GDP	20 900	22 300	25 000	26 400	30 900	33 800



In Table 3, the variable *year* has been rescaled using 1980 = 0, 1985 = 5 and so on. The new variable is *time*.

Table 3

Year	1980	1985	1990	1995	2000	2005
Time	0	5	10	15	20	25
GDP	20 900	22 300	25 000	26 400	30 900	33 800

c. Use the variables *time* and *GDP* to write down the equation of the least squares regression line that can be used to predict *GDP* from *time*. Take *time* as the independent variable.

2 marks

d. In the year 2007, the *GDP* was $34 900. Find the error in the prediction if the least squares regression line calculated in **part c.** is used to predict *GDP* in 2007.

2 marks