

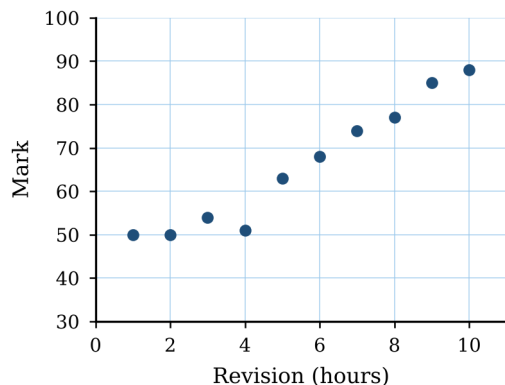
Scatter Diagrams

- 1 State whether you would expect positive, negative or no correlation between each pair of variables.

- (a) the height and the arm span of adults
- (b) the outside temperature and a household's heating bill
- (c) the shoe size of a student and the student's mark in a French test

[3]

- 2 The scatter diagram shows the hours of revision, x , and the test mark, y , of ten students. Describe the correlation.



[2]

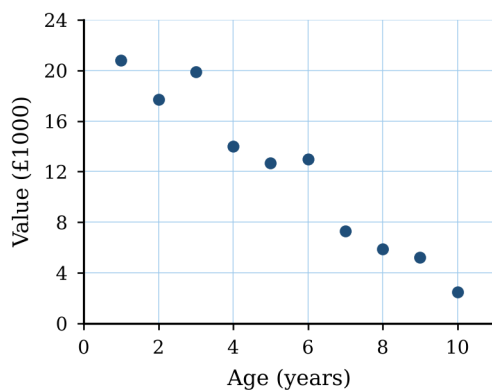
- 3 A scientist investigates how the mass of fertiliser (g) affects the yield (kg) of tomato plants. State which is the explanatory variable and which is the response variable.

[2]

- 4 Find the coordinates of the mean point $(\bar{x}, \bar{y})$ for the data (2, 5), (4, 9), (6, 10), (8, 14), (10, 17).

[2]

- 5 The scatter diagram shows the age in years, x , and the value in £1000, y , of ten cars. Describe the relationship shown and explain what it means in context.

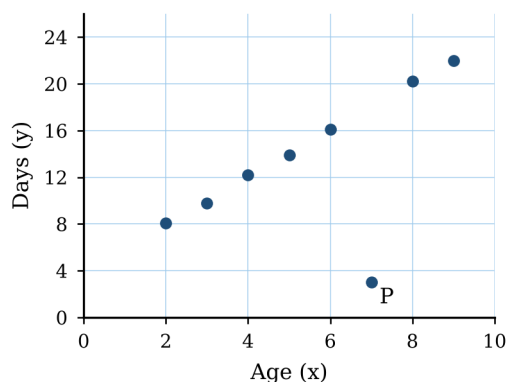


[2]

- 6 Sketch a scatter diagram that shows weak positive correlation between two variables x and y .

[2]

- 7 The scatter diagram shows the age (years), x , and the length of stay in hospital (days), y , of 8 patients. One point, P, does not follow the pattern.

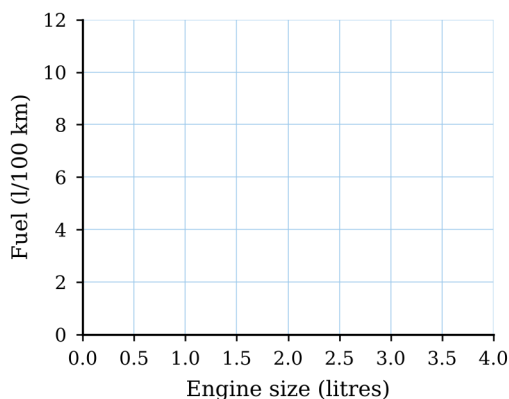


- (a) Write down the approximate coordinates of P. [1]
 (b) Suggest a reason for the anomaly and say what should be done with the point. [2]

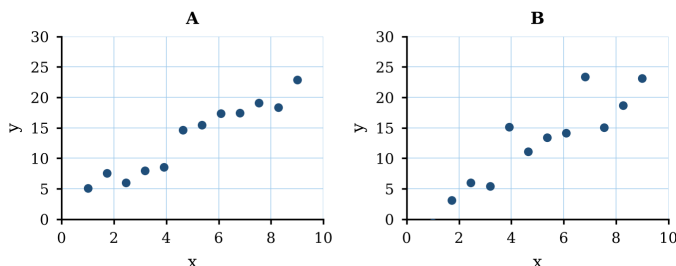
- 8 The table shows the engine size x (litres) and the fuel used y (litres per 100 km) of six cars.

x	1	1.4	1.6	2	2.4	3
y	5.1	6	6.4	7.2	8.1	9.2

- (a) Find the mean point $(\bar{x}, \bar{y})$. [2]
 (b) On the grid, plot the six points and the mean point, and draw a line of best fit through the mean point. [3]



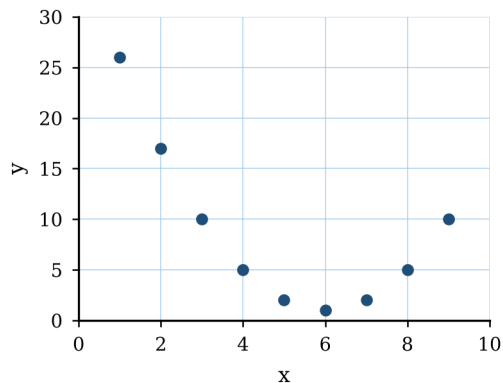
- 9 The scatter diagrams show two sets of data, A and B, each with 12 points.



- (a) Describe the correlation in each set. [2]
 (b) State which set would have the larger product moment correlation coefficient. [1]

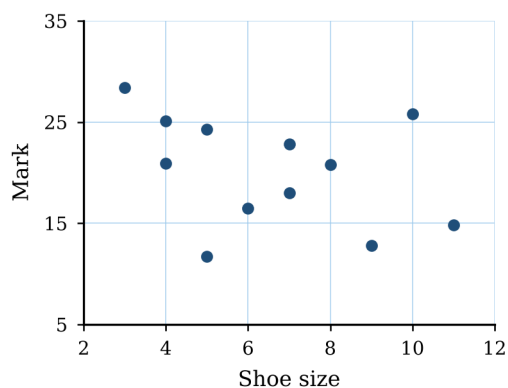
- 10 A newspaper reports that there is positive correlation between the monthly sales of ice cream and the number of people who drown. Explain why this does not show that ice cream causes drowning. [2]

- 11** The scatter diagram shows data that follow a curve. Describe the relationship and explain why a straight line is not a suitable model.



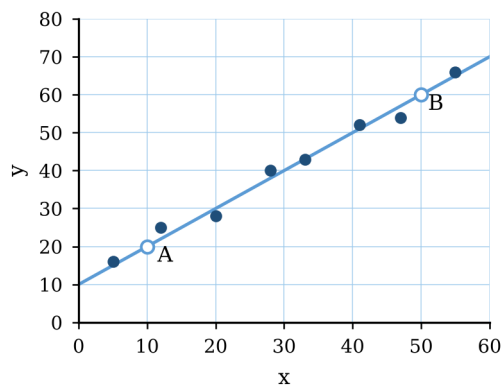
[2]

- 12** The scatter diagram shows the shoe size, x , and the mark, y , in a spelling test of 12 students. Describe the correlation and explain what it tells you about the two variables.



[2]

- 13** The line of best fit shown passes through A(10, 20) and B(50, 60). Find its equation in the form $y = a + bx$ and use it to estimate y when $x = 35$.



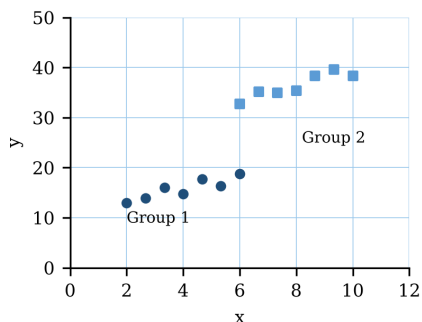
[4]

- 14** The table shows 9 pairs of values. A line of best fit with gradient 1.8 is drawn through the mean point.

x	3	5	6	8	9	11	12	14	15
y	14.9	14.4	15.8	15.1	22.2	25.2	26.5	31.8	33.8

- (a) Find the mean point. [2]
 (b) Find the equation of the line of best fit and use it to estimate y when $x = 20$. [3]

- 15** The scatter diagram shows data from two groups of people. Group 1 is shown as circles and Group 2 as squares.



- (a) Describe the correlation for the 14 points taken together. [1]
 (b) Describe the correlation within each group separately, and explain why the overall impression may be misleading. [3]

- 16** For the data in the table, calculate S_{xx} , S_{xy} and S_{yy} , and state whether the correlation is positive or negative.

x	2	4	5	7	8	10	11	13
y	14	15	19	20	24	25	29	30

[5]

- 17** Eight points have coordinates (1, 5), (2, 7), (3, 8), (4, 10), (5, 11), (6, 13), (7, 14), (8, 16). A ninth point P(9, 3) is added.

- (a) Find the mean point of the original eight points. [2]
 (b) Find the mean point of all nine points and describe the effect of P. [3]

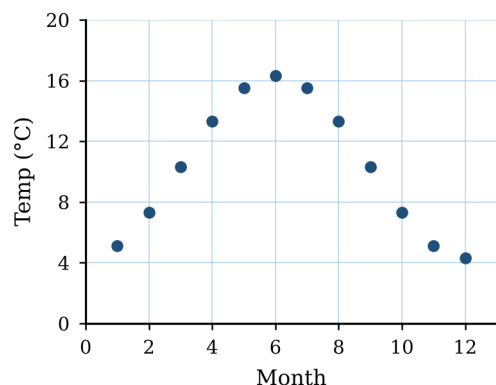
- 18** Data set A has pairs (x, y) with y close to 2x, and data set B has pairs with y close to 0.2x, for x = 1 to 10. A student says that A shows stronger correlation than B because its line is steeper. The product moment correlation coefficients are $r_A = 0.991$ and $r_B = 0.991$. Comment on the student's statement. [3]

- 19** The table shows seven pairs of values with $y = x^2$.

x	-3	-2	-1	0	1	2	3
y	9	4	1	0	1	4	9

- (a) Show that $S_{xy} = 0$. [3]
 (b) Hence explain why a product moment correlation coefficient of 0 does not mean that the variables are unrelated. [2]

- 20** The scatter diagram shows the average temperature ($^{\circ}\text{C}$) for each month of a year at one location, x being the month number and y the temperature.
A student uses the diagram to claim that temperature will keep rising next year at the same rate. Give two reasons the claim is unreliable.



[2]

- 21** The table shows data for 10 pairs of values (x, y) . One pair is suspected to be an error.

x	12	15	17	20	22	25	27	30	33	36
y	30	34	33	41	44	47	20	55	59	64

- (a) Identify the suspect pair and explain your choice. [2]
 (b) Calculate the product moment correlation coefficient for all ten pairs. [2]
 (c) Calculate r again without the suspect pair and comment on the change. [3]

- 22** A study records the pairs (x, y) for two groups. Group A: (1, 20), (2, 22), (3, 25), (4, 27), (5, 30). Group B: (6, 40), (7, 43), (8, 45), (9, 48), (10, 52).

- (a) Find the product moment correlation coefficient for each group, and for all 10 pairs. [3]
 (b) Explain why the overall correlation is not a good summary of the relationship within each group. [2]
 (c) Suggest how the data should be analysed. [1]

- 23** The data show (x, y) . The coded variables are $u = \frac{x}{10}$ and $v = \frac{y - 50}{4}$.

x	10	20	30	40	50	60
y	53	81	98	108	117	123

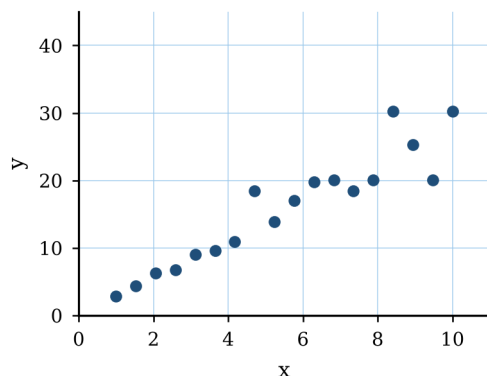
- (a) Find the product moment correlation coefficient between x and y . [2]
 (b) Write down the coefficient between u and v , giving a reason. [2]
 (c) State the coefficient between u and w , where $w = 20 - 3v$. [2]

- 24** The table shows the number y of bacteria (in thousands) in a culture at time x hours.

x	1	2	3	4	5	6	7	8
y	4.8	6.1	10	13.1	20.9	27.6	43.2	59

- (a) Calculate the product moment correlation coefficient between x and y . [1]
 (b) Calculate $\ln y$ for each value of x and find the correlation coefficient between x and $\ln y$. [3]
 (c) Suggest a type of model for y in terms of x and explain your choice. [2]

- 25** The scatter diagram shows the data for 18 pairs of values (x, y) . The spread of y about the general trend changes as x changes.



- (a) Describe the correlation and the pattern of the spread of the points. [2]
 (b) Explain why a prediction of y at $x = 10$ is less precise than at $x = 2$. [2]

- 26** The numbers of visitors (in thousands) to a museum in the years 2015 to 2024 are shown. Let $t = \text{year} - 2015$.

t	0	1	2	3	4	5	6	7	8	9
Visitors	42	45	47	51	54	31	55	60	63	66

- (a) Describe the general trend and identify the unusual year. [2]
 (b) Find the correlation coefficient between t and visitors, and then without the year 2020. [2]
 (c) Suggest a reason for the unusual year and say whether the point should be removed. [2]

- 27** The table shows 12 pairs of values of x and y .

x	1	2	3	4	5	6	7	8	9	10	11	12
y	5	11	9	16	14	20	17	24	21	27	25	31

- (a) Calculate the product moment correlation coefficient for all 12 pairs. [2]
 (b) A researcher only has the four pairs with $x = 5, 6, 7, 8$. Calculate r for these pairs. [2]
 (c) Comment on the effect on the correlation coefficient of using only a small range of x values, and on the reliability of conclusions from the small sample. [3]

- 28** For n pairs of values (x, y) , $S_{xy} = \Sigma(x - \bar{x})(y - \bar{y})$.

- (a) Show that $S_{xy} = \Sigma xy - n\bar{x}\bar{y}$. [3]
 (b) New variables are $u = ax + b$ and $v = cy + d$, where a and c are positive constants. Show that $S_{uv} = ac S_{xy}$. [3]
 (c) Hence show that the product moment correlation coefficient is unchanged by this coding. [3]

- 29** The table gives 10 pairs of values. The least squares line of y on x for all 10 pairs is $y = 6.1133 + 1.6606x$. A pair is flagged as an outlier if its residual (observed y minus the value given by the line) is greater than $2s$ in size, where $s = 2.1984$.

x	2	4	6	8	10	12	14	16	18	20
y	9.1	13.2	14.8	19.7	21	31.5	28.2	31.1	36	39.2

- (a) Calculate the residual for each pair. [3]
 (b) Identify any flagged pair. [2]
 (c) Recalculate the correlation coefficient without the flagged pair(s) and comment. [3]

- 30** The table shows the surface area y (cm^2) of a growing crystal at time x (hours). It is suggested that $y = ax^n$ or $y = ab^x$.

x	1	2	3	4	5	6	7	8
y	2.5	5.3	11.8	15.7	24.8	29.5	40.5	45.7

- (a) Show that the model $y = ax^n$ gives a linear relationship between $\ln y$ and $\ln x$, and state the gradient and intercept in terms of a and n . [3]
- (b) Calculate the correlation coefficients for (x, y) , $(x, \ln y)$ and $(\ln x, \ln y)$. [3]
- (c) Hence choose the better model, and find the values of a and n from the least squares line of $\ln y$ on $\ln x$, to 2 decimal places. [4]

Correlation

- 1** Describe the correlation indicated by each value of the product moment correlation coefficient: (a) 0.92 (b) -0.15 (c) -0.88 (d) 0.55. [4]
- 2** State the range of possible values of the product moment correlation coefficient r , and say what the values $r = 1$ and $r = -1$ mean. [3]
- 3** For a set of data, $S_{xx} = 40$, $S_{yy} = 90$ and $S_{xy} = 48$. Calculate the product moment correlation coefficient. [2]
- 4** A test is carried out to see whether there is positive correlation between two variables. State suitable null and alternative hypotheses in terms of the population correlation coefficient ρ . [2]
- 5** For 10 pairs of values, $r = 0.71$. The critical value for a one-tailed test at the 5% significance level is 0.5494. Test $H_0: \rho = 0$ against $H_1: \rho > 0$. [2]
- 6** The correlation coefficient between x and y is 0.6. Write down the correlation coefficient between (a) $x + 10$ and y , (b) $4x$ and y , (c) $-x$ and y . [3]
- 7** For 8 pairs of values, $n = 8$, $\Sigma x = 84$, $\Sigma y = 113$, $\Sigma x^2 = 1064$, $\Sigma y^2 = 1857$ and $\Sigma xy = 1397$. Calculate the product moment correlation coefficient. [4]
- 8** The table shows 8 pairs of values of x and y . Calculate the product moment correlation coefficient and carry out a test for positive correlation at the 5% level, given that the critical value is 0.6215.
- | x | 12 | 15 | 14 | 18 | 20 | 22 | 25 | 24 |
|-----|----|----|----|----|----|----|----|----|
| y | 40 | 46 | 41 | 55 | 52 | 63 | 66 | 68 |
- [4]
- 9** A sample of 12 pairs gives $r = -0.62$. Test, at the 5% significance level, whether there is correlation (either positive or negative) between the variables. The two-tailed critical value is 0.576. [3]
- 10** The product moment correlation coefficient between the height and the mass of 15 students is $r = 0.82$. Interpret this value in context. [2]
- 11** The correlation coefficient between daily temperature in $^{\circ}\text{C}$ and ice cream sales is 0.78. The temperatures are converted to $^{\circ}\text{F}$ using $F = 1.8C + 32$. State the new correlation coefficient and explain your answer. [2]

- 12** A student calculated $r = 0.58$ from 9 pairs of values and says "Since 0.58 is close to 1, the correlation is certainly significant." The critical value for a one-tailed test at the 5% level is 0.5822. Comment on her statement. [3]

- 13** The table shows the number x of absences and the exam mark y (%) for 6 students. Calculate the product moment correlation coefficient and interpret it in context.

x	2	4	5	7	9	10
y	31	28	27	22	20	15

[3]

- 14** The table shows the number x of hours of television watched per week and the mark y in a test, for 10 students. A teacher believes that watching more television is associated with lower marks.

x	20	24	27	31	33	36	40	44	47	52
y	62	58	55	57	49	45	44	38	40	31

- (a) Calculate the product moment correlation coefficient. [1]
 (b) Test the teacher's belief at the 1% significance level. The critical value is 0.7155. [4]
 (c) Does your result show that watching television causes lower marks? Explain. [1]

- 15** For 12 pairs of values the data are (5, 19), (7, 25), (8, 18), (11, 28), (12, 24), (14, 33), (15, 29), (17, 31), (19, 38), (20, 30), (22, 40), (24, 37). Test at the 10% level whether there is correlation between x and y . The two-tailed critical value is 0.4973. [5]

- 16** For 12 pairs, $\Sigma x = 144$, $\Sigma y = 180$, $\Sigma x^2 = 1930$, $\Sigma y^2 = 3000$ and $\Sigma xy = 2310$. Find r and test for positive correlation at the 1% level, given the critical value 0.6581. [6]

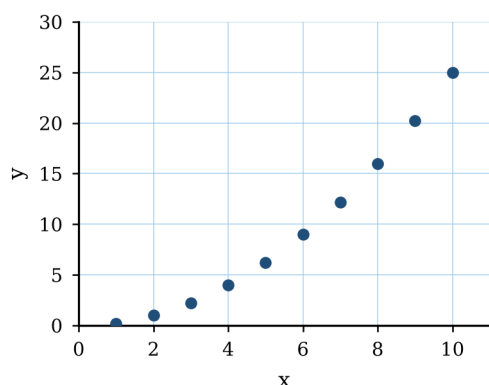
- 17** Seven pairs of values are (1, 3), (2, 5), (3, 4), (4, 7), (5, 6), (6, 9) and (7, 20). Critical values for a one-tailed 5% test are 0.6694 for $n = 7$ and 0.7293 for $n = 6$.
 (a) Calculate r for all seven pairs and for the first six pairs. [2]
 (b) Carry out a test for positive correlation in each case and comment on the influence of (7, 20). [3]

- 18** For 15 pairs of values, $\bar{x} = 12$, $\bar{y} = 18$ and $\Sigma xy = 3450$. The variance of the x values is 16 and the variance of the y values is 25.
 (a) Find the product moment correlation coefficient. [3]
 (b) Test for positive correlation at the 5% significance level. The critical value is 0.4409. [2]

- 19** Critical values for a one-tailed 5% test are 0.5494 for $n = 10$, 0.3783 for $n = 20$ and 0.2638 for $n = 40$.
 (a) Explain why the critical value decreases as the sample size increases. [2]
 (b) A sample gives $r = 0.5$. State whether this is significant at the 5% level for $n = 10$ and for $n = 20$, and comment. [3]

20

The scatter diagram shows 10 pairs (x, y) with $y = \frac{x^2}{4}$. A student calculates $r = 0.975$ and states that the relationship is therefore linear. Comment on this.



[3]

21

For 10 pairs (x, y) , the coded variables $u = x - 50$ and $v = \frac{y}{10}$ have $\Sigma u = -12$, $\Sigma v = 55$, $\Sigma u^2 = 150$, $\Sigma v^2 = 420$ and $\Sigma uv = -168$.

- Calculate the product moment correlation coefficient between u and v . [4]
- State the coefficient between x and y , with a reason. [2]
- Test for negative correlation at the 5% level; the critical value is 0.5494. [2]

22

A researcher tests 20 different pairs of unrelated variables for correlation, each at the 5% significance level, using a two-tailed test.

- State the probability that a single test wrongly rejects H_0 . [1]
- Assuming the tests are independent, find the probability that at least one test gives a significant result. [3]
- Explain what this shows about searching through many variables for correlation. [2]

23

The 10 pairs of values below have $y = \frac{x^2}{4}$. Test for positive correlation at the 1% level (critical value 0.7155) and comment on the validity of using the test.

x	1	2	3	4	5	6	7	8	9	10
y	0.2	1	2.2	4	6.2	9	12.2	16	20.2	25

[5]

24

For 10 pairs of values, $\bar{x} = 8$, $\bar{y} = 20$, $S_{xx} = 90$, $S_{yy} = 250$ and $S_{xy} = 135$. An eleventh pair (14, 32) is then added.

- Find the original value of r . [2]
- Show that for the original data $\Sigma x = 80$, $\Sigma y = 200$, $\Sigma x^2 = 730$, $\Sigma y^2 = 4250$ and $\Sigma xy = 1735$. [3]
- Find the new values of S_{xx} , S_{yy} and S_{xy} and hence the new value of r . [5]
- Comment on the effect of the new pair on the correlation. [1]

25

For a set of data, $r = 0.8$, $S_{xx} = 36$ and $S_{yy} = 100$.

- Find S_{xy} . [2]
- Find the gradient of the least squares line of y on x and of x on y . [2]
- Show that the product of the two gradients is r^2 . [2]

- 26** The six pairs (3, 7), (5, 11), (6, 10), (8, 15), (9, 14), (12, 19) have $r = 0.9695$. A seventh pair equal to the mean point is added.
- Find the mean point of the original data. [2]
 - Show that adding the mean point does not change S_{xx} , S_{yy} or S_{xy} , and state the new value of r . [3]
-
- 27** For n pairs of values, consider the expression $\sum[a(x - \bar{x}) - (y - \bar{y})]^2$, where a is a real constant.
- Show that the expression equals $a^2S_{xx} - 2aS_{xy} + S_{yy}$. [3]
 - Explain why the expression is never negative, and hence use the discriminant to show that $S_{xy}^2 \leq S_{xx}S_{yy}$. [4]
 - Deduce that $-1 \leq r \leq 1$. [2]
-
- 28** For a set of data, the regression line of y on x has gradient 0.9 and the regression line of x on y has gradient 0.5 when x is written as a function of y .
- Show that $r^2 = 0.45$ and find r , given that the correlation is positive. [3]
 - Find the ratio of the standard deviation of y to the standard deviation of x . [3]
 - Explain what would be true of the two lines if $r = 1$. [2]
-
- 29** A test for correlation can be carried out using the statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, which has a t -distribution with $n - 2$ degrees of freedom when $\rho = 0$.
- For $n = 12$ and a one-tailed 5% test, the critical t value with 10 degrees of freedom is 1.8125. Find the critical value of r by solving $t = 1.8125$, giving your answer to 4 decimal places. [4]
 - A sample of 12 pairs has $r = 0.55$. Find t and decide whether there is evidence of positive correlation at the 5% level. [3]
 - Show that this agrees with a comparison of $r = 0.55$ with the critical value in part (a). [1]
-
- 30** Three studies of the same relationship each use a one-tailed 5% test for positive correlation. The results and critical values are shown.
- | Study | n | r | Critical value |
|-------|-----|------|----------------|
| A | 10 | 0.45 | 0.5494 |
| B | 20 | 0.40 | 0.3783 |
| C | 40 | 0.32 | 0.2638 |
- State the conclusion of each study. [3]
 - Study A has the largest r but is not significant. Explain how this is possible. [2]
 - A newspaper reports Study C as showing "a strong link". Comment. [2]

Regression Lines

- 1** The regression line of y on x is $y = 12.4 + 3.5x$. Find the value of y when $x = 8$, and state the change in y when x increases by 2. [3]
-
- 2** For a set of data, $S_{xy} = 120$ and $S_{xx} = 40$, with $\bar{x} = 5$ and $\bar{y} = 20$. Find the equation of the regression line of y on x . [3]
-
- 3** A scientist wants to predict the mass of a plant from its height. Should she use the regression line of mass on height or of height on mass? Explain. [2]

- 4 Use your calculator to find the equation of the regression line of y on x for the data in the table, giving a and b to 3 significant figures.

x	1	2	3	4	5	6
y	4	7	8	12	13	16

[2]

- 5 A regression line is $y = 2 + 3x$. At $x = 7$ the observed value of y is 25. Find the residual.

[2]

- 6 The regression line of y on x is $y = 4 + 2x$ and the mean point is $(6, 16)$. Show that the line passes through the mean point.

[2]

- 7 For 10 pairs of values, $\Sigma x = 95$, $\Sigma y = 210$, $\Sigma x^2 = 1085$ and $\Sigma xy = 2285$. Find the equation of the regression line of y on x .

[5]

- 8 The table shows the temperature x ($^{\circ}\text{C}$) and the number y of cold drinks sold in a café on 8 days. Find the regression line of y on x and use it to estimate the sales when $x = 21$.

x	12	15	17	20	22	25	28	30
y	34	39	41	48	50	58	62	66

[4]

- 9 A regression line has gradient 2.5 and passes through the mean point $(8, 30)$. Find its equation in the form $y = a + bx$.

[3]

- 10 For a data set, $r = 0.8$, the standard deviation of x is 5 and the standard deviation of y is 12. Find the gradient of the regression line of y on x .

[2]

- 11 Data on a company's yearly profit y (£ million) are coded using $u = x - 2020$, where x is the year. The regression line of y on u is $y = 3.2 + 0.8u$. Find the line in terms of x and estimate y for the year 2026.

[3]

- 12 The residuals for the first four points of a regression with five points are 1.2, -0.8 , 0.5 and -2.1 . Find the fifth residual and explain your method.

[2]

- 13 The regression line of the cost C (£) of hiring a van for d days is $C = 35 + 48d$. Interpret the values 35 and 48 in context and estimate the cost of hiring the van for 3 days.

[3]

- 14 For 12 pairs, $\Sigma x = 120$, $\Sigma y = 426$, $\Sigma x^2 = 1518$, $\Sigma xy = 5020$.

(a) Find the regression line of y on x , giving a and b to 3 significant figures.

[4]

(b) Estimate y when $x = 10$ and find the residual if the observed value at $x = 10$ is 36.

[3]

- 15 For 10 pairs (x, y) , the coded variables $u = \frac{x}{10}$ and $v = y - 100$ give $\Sigma u = 25$, $\Sigma v = -30$, $\Sigma u^2 = 105$ and $\Sigma uv = -150$. Find the regression line of v on u and hence the regression line of y on x .

[6]

- 16 The table shows 7 pairs of values (x, y) .

x	4	6	7	9	11	13	14
y	9	11	16	15	20	22	26

(a) Find the regression line of y on x and the regression line of x on y .

[4]

(b) Show that both lines pass through the mean point.

[2]

- 17** The regression line of the mass M (kg) of a child on height h (cm) is $M = -48 + 0.52h$. The heights are now measured in metres, $H = \frac{h}{100}$. Find the regression line of M on H and the predicted mass for a child 1.40 m tall. [3]

- 18** A set of data has $\bar{x} = 10$, $\bar{y} = 34$, standard deviations $s_x = 2$ and $s_y = 6$ and correlation coefficient $r = 0.9$. Find the regression line of y on x and the regression line of x on y . [5]

- 19** Five points have coordinates (2, 6), (4, 7), (5, 11), (7, 12) and (8, 16). Complete the working below to find the regression line of y on x , showing the deviations from the mean point. [5]

- 20** The table shows 8 pairs of values. The regression line of y on x is $y = 8 + 2x$.

x	3	5	6	8	10	11	13	15
y	14	17	22	23	29	28	35	38

- (a) Calculate the residual for each pair. [3]
 (b) Identify the point with the largest residual and calculate Σe^2 for all eight points. [3]

- 21** The regression line of y on x is $y = 2 + 0.8x$ and the regression line of x on y is $x = 1 + 0.5y$.
 (a) Find the coordinates of the mean point. [3]
 (b) Find the product moment correlation coefficient. [3]

- 22** A model $y = bx$ (a line through the origin) is fitted to the data (1, 3), (2, 5), (3, 8), (4, 9), (5, 13) by minimising the sum of the squares of the vertical distances.
 (a) Show that the sum of squares $D = \Sigma(y - bx)^2$ is minimised when $b = \frac{\Sigma xy}{\Sigma x^2}$. [4]
 (b) Calculate b for these data. [2]

- 23** Seven points (10, 15), (20, 24), (30, 33), (40, 44), (50, 52), (60, 63), (70, 72) have regression line y on x . An eighth point (80, 40) is then added.
 (a) Find the original regression line. [2]
 (b) Find the new regression line and compare the predictions at $x = 55$. [4]

- 24** A taxi company's fares y (£) for journeys of x km are shown in the table.

x	2	5	8	12	15	20
y	7.1	9	10.8	13.4	15.4	19

- (a) Find the regression line of y on x . [3]
 (b) Interpret the values of a and b in context. [2]
 (c) Estimate the fare for a 10 km journey. [1]

- 25** The table shows the number y (hundreds) of users of an app after x months. The model $y = ab^x$ is proposed.

x	1	2	3	4	5	6
y	4.2	6.1	9.5	14	20.5	30.9

- (a) Show that $\ln y = \ln a + x \ln b$, and explain how a straight line can be used to find a and b . [2]
 (b) Find the regression line of $\ln y$ on x and hence a and b to 3 significant figures. [4]
 (c) Estimate y after 8 months. [1]

- 26** The power y (watts) of a wind turbine at wind speed x (m s^{-1}) is recorded. It is thought that $y = ax^n$.

x	1	2	3	4	5	6	8	10
y	3.1	8.6	15.9	24.5	34.9	46.1	71.8	100.5

- (a) Show that regressing $\ln y$ on $\ln x$ gives estimates of $\ln a$ and n . [2]
 (b) Calculate the regression line of $\ln y$ on $\ln x$ and find n and a . [4]

- 27** The least squares regression line $y = a + bx$ is found by minimising $D = \sum(y - a - bx)^2$ over a and b .

- (a) Show that $\frac{\partial D}{\partial a} = 0$ gives $a = \bar{y} - b\bar{x}$. [3]
 (b) Show that $\frac{\partial D}{\partial b} = 0$ then gives $b = \frac{S_{xy}}{S_{xx}}$. [4]
 (c) Hence show that the sum of the residuals is zero. [2]

- 28** For the data (1, 3), (2, 5), (3, 4), (4, 8), (5, 9), (6, 12), the regression line of y on x is $y = 0.7333 + 1.7429x$.

- (a) Calculate the six residuals and verify that their sum is zero. [4]
 (b) Show also that $\sum x \times (\text{residual}) = 0$ and explain why this must hold for a least squares line. [4]

- 29** The table shows five pairs of values (x, y) , where one value of y is missing and is represented by k . The regression line of y on x is $y = 4 + 3x$.

x	1	2	3	4	5
y	8	9	k	15	20

- (a) Find the value of k . [3]
 (b) Calculate the residual of each of the five points and show that they sum to zero. [3]
 (c) A different data set of six points has a least squares regression line. Five of its residuals are 1.2, -0.8 , 0.5, -2.1 and 1.0. Find the sixth residual and explain your method. [3]

- 30** Data pairs (x, y) are coded using $u = \frac{x}{10}$ and $v = \frac{y - 10}{2}$. The regression line of v on u is $v = p + qu$.

- (a) Show that the regression line of y on x is $y = 10 + 2p + \frac{q}{5}x$. [3]
 (b) For the data (10, 12), (20, 19), (30, 31), (40, 38), (50, 52), find p and q and hence the regression line of y on x . Check using the original data. [4]

Interpretation of Regression

- 1** A café's weekly profit £ P is modelled by $P = 450 + 12c$, where c is the number of customers. Interpret the value 12 in context. [2]
2 The regression line of the height h (cm) of a plant on its age t (days) is $h = 3.2 + 1.4t$. Interpret the value 3.2 in context. [2]
3 The percentage charge y of a phone battery after x hours of use is modelled by $y = 90 - 2.5x$. Interpret the gradient and find the charge after 8 hours. [3]
4 A scientist studies how the amount of rainfall (mm) affects crop yield (tonnes per hectare). State which is the explanatory variable and which is the response variable, and which regression line she should use to predict yield. [2]
5 A regression line is $y = 20 + 5x$, where x is the number of hours of tutoring and y is the mark. Find the predicted mark for 4 hours of tutoring. [2]

- 6** There is a strong positive correlation between the number of fire engines attending a fire and the damage caused. Does this mean that fire engines cause the damage? Explain. [2]
- 7** A regression line for the data is $y = 2.143 + 1.895x$. At $x = 14$, the observed value is 30. Find the residual and say whether the point lies above or below the line.
- | x | 5 | 8 | 11 | 14 | 17 | 20 |
|----------|----------|----------|-----------|-----------|-----------|-----------|
| y | 12 | 18 | 21 | 30 | 33 | 41 |
- [3]
- 8** The regression lines of exam mark y on revision hours x are $y = 32 + 4.5x$ for school A and $y = 40 + 3x$ for school B. Compare the two schools. [3]
- 9** The regression line of mass m (kg) on height h (cm) is $m = -60 + 0.75h$. Write the gradient in grams per millimetre. [2]
- 10** A student has the regression line of y on x , $y = 6.9 + 0.77x$, from the data in the table. To find x from a given y she rearranges the equation. Explain why this may not be appropriate and find the gradient of the proper line for predicting x .
- | x | 10 | 20 | 30 | 40 | 50 |
|----------|-----------|-----------|-----------|-----------|-----------|
| y | 15 | 22 | 28 | 41 | 44 |
- [4]
- 11** A model for the number of visitors y to a park on days with maximum temperature x °C is $y = -45 + 9.2x$. Find the predicted number of visitors when $x = 17$, giving your answer to the nearest whole number, and explain why the answer is rounded. [3]
- 12** The residuals from fitting a regression line to 10 data points are plotted against x . State what the pattern in the residuals suggests about the linear model.
-
- [2]
- 13** A regression line predicts a sales figure of 240 units, but the observed figure is 258. Find the residual and the percentage by which the observed value exceeds the prediction, to 3 significant figures. [3]
- 14** The table shows the price x (pence) of an item and the number y sold per week by a shop.
- | x | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 |
|----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| y | 310 | 285 | 262 | 240 | 214 | 195 | 170 | 148 |
- (a) Find the regression line of y on x . [3]
 (b) Interpret the gradient in context. [1]
 (c) Interpret the intercept and comment on whether it is meaningful. [2]
- 15** Daily temperatures in °C (c) and in °F (f) are related by $f = 1.8c + 32$. The regression line of the amount of electricity y (units) used on the temperature in °C is $y = 85 - 2.4c$. Find the regression line of y on f and interpret its gradient. [4]

- 16** For a group of children, $r = 0.7$, the standard deviation of age is 2.5 years and the standard deviation of height is 12 cm. Find the gradient of the regression line of height on age and interpret it. [3]

- 17** The heights h (cm) and masses m (kg) of 12 men are summarised by $\Sigma h = 2040$, $\Sigma m = 852$, $\Sigma h^2 = 348\,000$, $\Sigma m^2 = 61\,092$ and $\Sigma hm = 145\,560$.
A magazine claims that every extra 10 cm of height adds about 8 kg of mass.
- (a) Find S_{hh} and S_{hm} . [2]
- (b) Find the regression line of m on h . [3]
- (c) Use your line to comment on the magazine's claim. [2]

- 18** Two regression lines are found. Line 1 uses 8 pairs and has $r = 0.62$. Line 2 uses 40 pairs and has $r = 0.62$. Explain which line gives more reliable predictions and why. Critical values for a one-tailed 5% test are 0.6215 ($n = 8$) and 0.2638 ($n = 40$). [4]

- 19** A student says the line of best fit for the data should pass through the origin. The least squares line is $y = 1.84 + 0.744x$. The line through the origin that fits best is $y = 0.9566x$.

x	2	4	6	8	10	12
y	3.4	5.1	5.8	8.2	8.5	11.3

- (a) Calculate the sum of squared residuals for each line. [4]
- (b) Comment on the student's suggestion. [1]

- 20** The regression line of the time y (minutes) taken to complete a task on the number x of practice sessions is $y = 42 - 3.1x$, for x between 1 and 6.
- (a) Find the predicted time after 4 practice sessions. [1]
- (b) Find the least whole number of sessions for which the model predicts a time below 20 minutes. [3]
- (c) Comment on the use of the model for a very large number of sessions. [1]

- 21** The table shows the number of years x of experience and the salary y (£1000) of 10 employees.

x	1	2	3	4	5	6	8	9	10	12
y	22	24	25	29	28	31	35	36	40	42

- (a) Find the regression line of y on x and the product moment correlation coefficient. [3]
- (b) Test at the 5% level for positive correlation, given the critical value 0.5494. [2]
- (c) Interpret the gradient and the intercept in context. [2]
- (d) Explain why the regression line cannot be used to claim that experience causes higher salary. [1]

- 22** The data (1, 10), (2, 12), (3, 15), (4, 16), (5, 19), (6, 40) include an unusual final point.
- (a) Find the regression line of y on x with and without the final point. [4]
- (b) Describe the effect of the unusual point on the gradient and intercept, and on the prediction at $x = 3.5$. [3]

- 23** The value y (£1000) of an investment after x years is modelled by $y = ab^x$. The table shows the data.

x	1	2	3	4	5	6
y	5.2	6.1	7.4	8.9	10.6	12.9

- (a) Find a and b by regressing $\ln y$ on x . [4]
- (b) Interpret the values of a and b in context. [2]

- 24** A technician measures a voltage x and wants to estimate it from the current y ; another technician wants to estimate y from x . The table shows the data.
- | x | 10 | 15 | 20 | 25 | 30 | 35 |
|-----|----|----|----|----|----|----|
| y | 14 | 21 | 24 | 33 | 35 | 44 |
- (a) Find the regression line of y on x and the regression line of x on y . [4]
(b) Estimate x when $y = 30$ using the appropriate line, and compare with the value from rearranging the line of y on x . [3]
- 25** A linear model is fitted to the data $(x, y) = (1, 3.0), (2, 7.1), (3, 9.4), (4, 11.2), (5, 11.9), (6, 11.7), (7, 10.4), (8, 8.5), (9, 5.0)$. The regression line is $y = 7.464 + 0.245x$.
- (a) Calculate the nine residuals. [3]
(b) Describe the pattern of the residuals and explain what it shows about the model. [2]
- 26** Maths marks m have mean 58 and standard deviation 10. Physics marks p have mean 62 and standard deviation 12. The correlation coefficient is 0.72.
- (a) Find the regression line of p on m and the regression line of m on p . [4]
(b) A student scores 70 in maths. Predict the physics mark. Another scores 70 in physics. Predict the maths mark. [3]
(c) Explain why each predicted mark lies closer to the mean than the given mark. [2]
- 27** The profit y (£100) of a shop in month x , for $x = 1$ to 8, gives $\Sigma x = 36$, $\Sigma y = 244$, $\Sigma x^2 = 204$ and $\Sigma xy = 1224$.
- (a) Find the regression line of y on x . [4]
(b) Use the line to estimate the total profit for months 9 to 20 inclusive. [4]
(c) Show that the model predicts the total profit over the first N months to be $\frac{N(3N + 37)}{2}$, and find the least N for which this total exceeds 2000 (£100). [4]
(d) Comment on the reliability of your answer to part (c). [1]
- 28** Fathers' heights have mean 175 cm and standard deviation 6.5 cm. Sons' heights have mean 177 cm and standard deviation 6.5 cm. The correlation coefficient between a father's height and his son's height is 0.5.
- (a) Find the regression line of the son's height s on the father's height f . [3]
(b) A father is 188 cm tall (2 standard deviations above the mean). Predict his son's height and say how many standard deviations above the sons' mean this is. [3]
(c) Explain the phenomenon of regression towards the mean shown by your answer. [2]
- 29** A hot drink cools in a room at 20 °C. Its temperature T (°C) at time t (minutes) is recorded. It is modelled by $T - 20 = Ae^{-kt}$.
- | t | 0 | 5 | 10 | 15 | 20 | 25 |
|-----|----|------|------|------|------|------|
| T | 95 | 70.2 | 54.6 | 44.6 | 38.3 | 34.3 |
- (a) Show that $\ln(T - 20)$ is a linear function of t and state the gradient and intercept. [2]
(b) Regress $\ln(T - 20)$ on t and find estimates of A and k . [4]
(c) Interpret A , and find the time taken for the excess temperature $T - 20$ to halve. [3]
- 30** Twelve individuals fall into three groups of four. Group 1: $x = 1$ to 4, $y = 8.0, 9.1, 9.9, 11.2$. Group 2: $x = 5$ to 8, $y = 14.1, 14.8, 16.1, 17.2$. Group 3: $x = 9$ to 12, $y = 19.8, 21.3, 22.0, 23.1$.
- (a) Find the regression line of y on x for all 12 individuals. [3]
(b) Find the regression line for the three group means. [3]
(c) Find the gradient within each group and compare all the gradients. What does this show about using group averages to make claims about individuals? [4]

Extrapolation and Interpolation

1 Explain the difference between interpolation and extrapolation when using a regression line. [2]

2 A regression line is based on data with x values from 10 to 50. For each of $x = 35, 60, 8$ and 50 state whether using the line is interpolation or extrapolation. [4]

3 State which is usually more reliable, interpolation or extrapolation, and give a reason. [2]

4 A regression line $y = 5 + 2x$ is based on data with x between 10 and 40. Estimate y when $x = 25$ and say whether the estimate is likely to be reliable. [3]

5 Give one reason why extrapolation may give an unreliable prediction. [1]

6 The table shows the data used to find a regression line. State the range of x values for which predictions would be interpolation.

x	12	15	18	21	24	27
y	34	39	41	48	52	55

[1]

7 The table shows five pairs of values (x, y) , with one value of y missing. The regression line of y on x is $y = 14 + 1.1x$.

x	10	20	30	40	50
y	25	35	k	61	67

(a) Find the missing value k . [3]

(b) Estimate y when $x = 80$ and comment on the reliability of the estimate. [3]

8 A regression line for the weekly sales y (units) of a product against its price x (£) is $y = 500 - 85x$, using prices from £1.50 to £4.00. Find the predicted sales when the price is £6 and comment. [3]

9 The regression line of the height h (cm) of a child on age a (years) is $h = 80 + 6.2a$, using children aged from 5 to 12. Predict the height of a 30-year-old and comment. [3]

10 A study of 30 pairs of values finds $r = 0.97$. A student says: "Since the correlation is so strong, extrapolating far beyond the data will be reliable." Comment on this claim. [2]

11 The regression line of the number y (thousands) of members of a gym on t , the number of years after 2015, is $y = 18.143 + 3.143t$, using data for 2015 to 2020. Predict the number of members in 2040 and comment. [3]

12 A model predicts that the percentage y of students who pass a test is $y = 52 + 3.5x$, where x is the hours of revision ($1 \leq x \leq 8$). Find the predicted pass rate when $x = 15$ and comment. [3]

13 The table shows measured values of y at four values of x . By linear interpolation between the points (20, 27) and (30, 33), estimate y at $x = 25$. A regression line fitted to all four points is $y = 2 + 1.17x$. Compare its prediction at $x = 25$.

x	10	20	30	40
y	14	27	33	51

[3]

- 14** The table shows the length x (hours) that a machine ran and the amount y (litres) of fuel used.

x	4	6	8	10	12	14	16
y	11.2	15.1	19.6	23.4	27.9	31.5	36.1

- (a) Find the regression line of y on x . [3]
 (b) Predict y when $x = 9$, $x = 20$ and $x = 1$. State, with reasons, which predictions are reliable. [4]

- 15** The table shows the temperature x ($^{\circ}\text{C}$) at which a chemical reaction was run and the yield y (grams).

x	10	20	30	40	50
y	28	41	49	63	70

- (a) The scientist wants to know what temperature will give a yield of 56 g. Which regression line should be used? Find the temperature. [4]
 (b) Is your estimate in (a) reliable? Explain. [1]

- 16** The regression line for the speed y (km/h) of a car going uphill against the gradient x (%) of the road, based on data from $x = 8$ to $x = 16$, is $y = 98.6 - 4.75x$.

- (a) Find the gradient at which the model predicts the car will stop ($y = 0$). [2]
 (b) Comment on the reliability of this prediction. [2]

- 17** The sales y (£ million) of a company in the years 2015 to 2019 are 12.1, 13.4, 14.8, 16.0, 17.4. Let $t = \text{year} - 2015$. The regression line is $y = 12.1 + 1.32t$. Later the actual sales were £24.0 million in 2022 and £33.5 million in 2025.

- (a) Find the predicted sales in 2022 and 2025. [3]
 (b) Calculate the errors in the predictions and comment on how they change. [3]

- 18** The population y (thousands) of a town is recorded every 10 years, where t is the number of years since the first record. Both a linear model and an exponential model were fitted: linear $y = 9.44 + 0.301t$ and exponential $\ln y = 2.3033 + 0.0198t$.

t	0	10	20	30	40
y	10	12.2	14.9	18.1	22.1

- (a) Predict the population at $t = 80$ using each model. [4]
 (b) Explain why the two predictions are so different and why neither is certain. [2]

- 19** The stopping distance y (m) of a car at speed x (km/h) is recorded. A regression line for speeds from 20 to 70 km/h is $y = -26.2 + 1.36x$.

x	20	30	40	50	60	70
y	6	14	24	38	53	75

- (a) Predict the stopping distance at 45 km/h and at 120 km/h. [3]
 (b) Stopping distance is known to be roughly proportional to the square of the speed. Using $y = kx^2$ with the data point (70, 75), find k and the prediction at 120 km/h. Comment. [4]

- 20** The table shows five pairs of values (x , y), with one value of y shown as k . The regression line of y on x has gradient 2.4.

x	2	4	6	8	10
y	9	k	16	22	27

- (a) Use $S_{xy} = b S_{xx}$ to find the value of k . [4]
 (b) Find the regression line of y on x . [2]
 (c) Estimate y when $x = 7$ and when $x = 15$, commenting on the reliability of each. [3]

- 21** The table shows the number x of hours of sunshine in a week and the number y of visitors (hundreds) to a beach.

x	5	7	9	11	13	15	17	19
y	18	22	27	31	33	40	42	48

- (a) Find r and test for positive correlation at the 5% level, given the critical value 0.6215. [3]
 (b) Find the regression line and use it to predict y when $x = 12$ and when $x = 30$. [3]
 (c) Comment on the reliability of each prediction. [2]

- 22** A model for the battery level y (%) of a phone after x hours of use is $y = 55.667 - 3.429x$, fitted to data for 1 to 6 hours.

- (a) Find the number of hours at which the model predicts the battery will be fully empty ($y = 0$). [2]
 (b) Find the value of y when $x = 0$ and explain why it is not sensible. [2]
 (c) Suggest a valid range of x for the model and explain. [2]

- 23** The altitude x (in hundreds of metres) and the air temperature y ($^{\circ}\text{C}$) were recorded at 10 places, all between 200 m and 1000 m above sea level. The results are summarised by $\Sigma x = 60$, $\Sigma y = 120$, $\Sigma x^2 = 480$, $\Sigma y^2 = 1502.5$ and $\Sigma xy = 642$.

- (a) Find S_{xx} , S_{yy} and S_{xy} . [3]
 (b) Find r and test for negative correlation at the 1% level. The critical value is 0.7155. [3]
 (c) Find the regression line of y on x . [2]
 (d) Estimate the temperature at an altitude of 750 m and at sea level, and comment on the reliability of each. [3]

- 24** The exact values of the function $y = 3 + 0.6x + 0.03x^2$ are used to fit two regression lines. Line N uses $x = 10, 11, 12, 13, 14$. Line W uses $x = 4, 7, 10, 13, 16, 19$.

- (a) Find both lines. [3]
 (b) Find the error of each line when predicting y at $x = 30$. [3]
 (c) Comment on the effect of the range of the data used on the quality of extrapolation. [1]

- 25** The mean monthly temperature ($^{\circ}\text{C}$) at a location in months 1 to 6 (January to June) is 4, 4.9, 7.5, 11, 14.5, 17.1. A regression line fitted to these six values is $y = 0.053 + 2.794x$.

- (a) Use the line to predict the temperature in month 10 (October). [2]
 (b) The actual temperature in October is 11°C . Find the error and explain it. [3]

- 26** The table shows data for x and y . To estimate x from y , one student rearranges the line of y on x and another uses the line of x on y .

x	10	20	30	40	50
y	32	38	49	52	64

- (a) Find both lines. [3]
 (b) Find the two estimates of x when $y = 47, 67$ and 107 . [3]
 (c) Comment on how the two estimates differ as y moves away from the mean. [1]

- 27** The regression line for the data (10, 24), (15, 30), (20, 33), (25, 41), (30, 47), (35, 52), (40, 58) is $y = 12.1429 + 1.1429x$. Suppose the true gradient could lie within ± 0.15 of the fitted value b .

- (a) Show that the prediction at x can be written as $\bar{y} + b(x - \bar{x})$ and find $\bar{x}$ and $\bar{y}$. [3]
 (b) Hence find the greatest possible error in the predicted y at $x = 15, 30$ and 60 caused by the uncertainty in the gradient. [4]
 (c) Use this to explain why extrapolation is less reliable than interpolation. [2]

- 28** A report states: "Our regression line based on 5 towns (populations 20 000 to 40 000) shows that for every extra 1000 people, crime rises by 4.2 incidents per year ($r = 0.78$). So when the city of 2 million people reduces by a half, crime will fall by 4200 incidents. This proves that population growth causes crime."
- (a) Identify three flaws in the report's reasoning. [3]
 - (b) For $n = 5$ the critical value for a one-tailed 5% test is 0.8054. Is $r = 0.78$ significant? [2]
 - (c) Check the arithmetic of the final claim: by how much does the model predict crime changes when a population of 2 million halves? [2]
 - (d) Suggest how a better investigation could be designed. [2]
-
- 29** The percentage y of a population using a new service t years after its launch is recorded: (1, 8), (2, 14), (3, 23), (4, 33), (5, 45), (6, 56), (7, 66).
- (a) Find the regression line of y on t . [3]
 - (b) Find the time at which the model predicts $y = 100$, and say why the model cannot hold beyond this. [3]
 - (c) Suggest the likely shape of the true long-term relationship and a sensible restriction on the model. [2]
-
- 30** The number y (millions) of components on a chip t years after a base year is recorded for $t = 0, 2, 4, 6, 8, 10$ as 2.3, 5.4, 12.0, 28.5, 61.0, 140.0. A model $y = ab^t$ is fitted by regressing $\ln y$ on t .
- (a) Find the regression line of $\ln y$ on t and hence a and b . [4]
 - (b) Show that the model predicts y to double about every 1.692 years, and use it to predict y at $t = 20$. [3]
 - (c) Comment on the reliability of the prediction at $t = 20$ and on the linear regression of y on t for these data. [3]
-