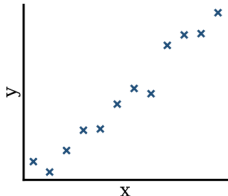
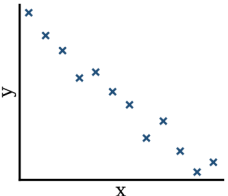


Product Moment Correlation Coefficient

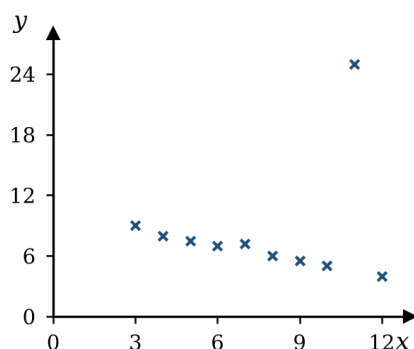
- 1 For a set of bivariate data, $S_{xx} = 40$, $S_{yy} = 90$ and $S_{xy} = 48$.
Calculate the product moment correlation coefficient. [2]
- 2 Eight pairs of values (x, y) give $n = 8$, $\Sigma x = 52$, $\Sigma y = 80$, $\Sigma x^2 = 380$, $\Sigma y^2 = 920$ and $\Sigma xy = 588$.
Calculate the product moment correlation coefficient, correct to 3 significant figures. [3]
- 3 Describe the correlation, in words, indicated by each of the following values of the product moment correlation coefficient: (i) $r = -0.92$, (ii) $r = 0.05$, (iii) $r = 0.6$. [3]
- 4 The scatter diagrams A and B show two sets of bivariate data.
State which diagram is more likely to have a correlation coefficient close to 0.95 and which close to -0.95 . Give a reason.
- A



B


- [2]
- 5 State the range of possible values of the product moment correlation coefficient, and describe the data when $r = 1$ and when $r = -1$. [2]
- 6 Find the product moment correlation coefficient for the data $(1, 2)$, $(2, 3)$, $(3, 5)$, $(4, 6)$. [3]
- 7 Six pairs of readings (x, y) are coded using $u = x - 10$ and $w = y - 50$, giving $\Sigma u = 26$, $\Sigma w = 49$, $\Sigma u^2 = 146$, $\Sigma w^2 = 579$ and $\Sigma uw = 289$.
Find the product moment correlation coefficient between x and y . [4]
- 8 The product moment correlation coefficient between two variables x and y is 0.8.
Write down the product moment correlation coefficient between (a) x and $3y - 2$, (b) x and $-2y$, (c) $5x + 1$ and $4y$. [3]
- 9 For a sample, the variance of x is 16, the variance of y is 25 and the covariance of x and y is 12.
Find the product moment correlation coefficient. [3]
- 10 The table shows six pairs of observations of two variables x and y .
- | | | | | | | |
|----------|----|----|----|----|----|----|
| x | 2 | 4 | 5 | 7 | 8 | 11 |
| y | 30 | 26 | 25 | 20 | 19 | 12 |
- (a) Calculate the product moment correlation coefficient. [4]
(b) Describe the relationship between x and y . [1]
- 11 For a sample of data, $S_{xx} = 25$, $S_{yy} = 36$ and the product moment correlation coefficient is $r = 0.8$.
Find S_{xy} , and hence find the gradient of the regression line of y on x . [4]

- 12** The scatter diagram shows ten observations (x, y) of two variables.



- (a) Identify the point that is likely to be an outlier, and say how it affects the product moment correlation coefficient. [2]
- (b) Without calculation, explain whether r for the nine other points would be closer to -1 than r for all ten points. [2]
-
- 13** A teacher finds that the product moment correlation coefficient between the number of hours spent revising and the percentage mark in a test is 0.78 for a class of students. Interpret this value in context, and explain why it does not prove that more revision causes higher marks. [3]
-
- 14** For ten pairs of observations, $\Sigma x = 60$, $\Sigma y = 129$, $\Sigma x^2 = 420$, $\Sigma y^2 = 1849$ and $\Sigma xy = 874$. Calculate the product moment correlation coefficient, and state what your value suggests about the data. [5]
-
- 15** For ten pairs of values, $\Sigma x = 113$, $\Sigma y = 138$, $\Sigma x^2 = 1801$, $\Sigma y^2 = 2166$ and $\Sigma xy = 1661$. The pair (30, 10) is thought to be a recording error and is removed.
- (a) Find the product moment correlation coefficient for all ten pairs. [2]
- (b) Find the product moment correlation coefficient for the remaining nine pairs. [4]
-
- 16** Show that, for constants a , b , c , d with $b > 0$ and $d > 0$, the product moment correlation coefficient between $u = a + bx$ and $w = c + dy$ equals that between x and y . [4]
-
- 17** Five pairs of values are (1, 3), (2, 5), (3, k), (4, 9) and (5, 11). The product moment correlation coefficient for these data is exactly 1. Find the value of k , giving a reason. [3]
-
- 18** Five pairs of observations are (−2, 4), (−1, 1), (0, 0), (1, 1) and (2, 4), so that $y = x^2$ in every case. Calculate the product moment correlation coefficient and explain why this does not mean that x and y are unrelated. [5]
-
- 19** The table shows the mean daily number of visitors (in hundreds) to a museum in six years. Using the coding $u = \text{year} - 2000$ and $w = 10(\text{visitors} - 40)$, find the product moment correlation coefficient between year and number of visitors.
- | Year | 2002 | 2004 | 2006 | 2008 | 2010 | 2012 |
|----------|------|------|------|------|------|------|
| Visitors | 41.2 | 43.5 | 44.1 | 47.3 | 48 | 51.1 |
- [6]
-
- 20** For a set of data, $S_{xx} = 16$, the regression line of y on x has gradient 1.5, and the product moment correlation coefficient is 0.9. Find S_{xy} and S_{yy} . [5]

- 21** Two groups of observations are recorded. Group A has summary statistics $n = 6$, $\Sigma x = 21$, $\Sigma y = 22$, $\Sigma x^2 = 91$, $\Sigma y^2 = 88$, $\Sigma xy = 88$. Group B has $n = 6$, $\Sigma x = 51$, $\Sigma y = 64$, $\Sigma x^2 = 451$, $\Sigma y^2 = 690$, $\Sigma xy = 555$.
- Calculate the product moment correlation coefficient for each group separately. [3]
 - Combine the two groups and calculate the product moment correlation coefficient for all 12 observations. [4]
 - Comment on the values obtained. [1]
-
- 22** A student has the summary statistics $n = 6$, $\Sigma x = 30$, $\Sigma x^2 = 190$, $\Sigma y = 42$, $\Sigma y^2 = 340$ and $\Sigma xy = 258$, and writes:
- " $S_{xx} = 190 - 30^2 = -710$, $S_{yy} = 340 - 42^2/6 = 46$, $S_{xy} = 258 - 30 \times 42/6 = 48$, so $r = 48/(-710 \times 46) \approx -0.0015$."
- Identify two errors in the student's working. [2]
 - Calculate the correct value of the product moment correlation coefficient. [3]
 - Explain how the student could tell that the answer -0.0015 is unlikely to be correct, given that $S_{xy} > 0$. [1]
-
- 23** Five pairs of observations are (2, 4), (3, 7), (5, 8), (6, 12), (9, 14). A sixth observation is then added at $(\bar{x}, \bar{y})$, the point whose coordinates are the means of the original five x-values and y-values.
- Calculate the product moment correlation coefficient before and after the sixth pair is added, and explain your result. [6]
-
- 24** For a sample of data, the standard deviation of x is 2.5 and the standard deviation of y is 4. The regression line of y on x has gradient 1.2.
- Show that the gradient of this regression line can be written as $b = r \frac{s_y}{s_x}$, where r is the product moment correlation coefficient, and hence find r . [5]
-
- 25** Standardised values are defined by $z_x = \frac{x - \bar{x}}{s_x}$ and $z_y = \frac{y - \bar{y}}{s_y}$, where s_x and s_y are the standard deviations (with divisor n).
- Show that the product moment correlation coefficient is $r = \frac{1}{n} \Sigma z_x z_y$. [4]
 - Hence show that r is unchanged if the variables are measured in different units. [2]
-
- 26** For a sample of 15 pairs of observations, $\Sigma(x - \bar{x})^2 = 135$, $\Sigma(y - \bar{y})^2 = 60$ and $\Sigma(x - \bar{x})(y - \bar{y}) = -54$.
- Find the product moment correlation coefficient between x and y . [2]
 - Find the product moment correlation coefficient between $p = 4 - \frac{x}{3}$ and $q = 2y + 1$. [2]
 - Find the standard deviation of x and of y , and the gradient of the regression line of y on x . [3]
-
- 27** For n pairs of observations (x_i, y_i) , write $S_{xx} = \Sigma(x_i - \bar{x})^2$, $S_{yy} = \Sigma(y_i - \bar{y})^2$ and $S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y})$.
- By considering $\Sigma[\lambda(x_i - \bar{x}) + (y_i - \bar{y})]^2 \geq 0$ for all real λ , show that $S_{xy}^2 \leq S_{xx}S_{yy}$, and deduce that $-1 \leq r \leq 1$. [6]
 - Show that $r = \pm 1$ only if every point lies on a straight line. [3]
-
- 28** Two variables x and y have $S_{xx} = S_{yy}$ (that is, equal variances) in a sample of n pairs.
- Show that the variables $u = x + y$ and $w = x - y$ are uncorrelated, whatever the correlation between x and y . [5]
 - The values of x are 1, 3, 4, 7, 10 and the values of y are 4, 10, 1, 7, 3 in the same order. Verify the result in part (a) numerically, and find the correlation between x and y . [3]

- 29** A variable y is measured with error. The recorded values are $y' = y + e$, where e is the measurement error. In a sample of n pairs, $S_{ee} = 36$ and $S_{ye} = S_{xe} = 0$ (the errors are uncorrelated with both x and y in the sample). Also $S_{yy} = 64$ and the correlation between x and y is 0.8.

Show that the correlation between x and y' is $r' = r\sqrt{\frac{S_{yy}}{S_{yy} + S_{ee}}}$, and calculate r' . Explain the effect of measurement error.

[7]

- 30** The pairs (1, 3), (2, 5), (3, a), (4, 8) and (5, 9) are observed, where a is a constant.

(a) Show that $S_{xx} = 10$ and find expressions in terms of a for S_{xy} and S_{yy} .

[4]

(b) Find the range of values of a for which the product moment correlation coefficient is greater than 0.9, giving the end points to 3 significant figures.

[5]

Rank Correlation

- 1** Five students score 62, 48, 75, 53 and 69 in a test. Give each score a rank, where rank 1 is the highest score.

[2]

- 2** Eight pairs of ranks give $\sum d^2 = 30$, where d is the difference between the ranks in each pair. Calculate Spearman's rank correlation coefficient.

[2]

- 3** Rank the values 7, 3, 7, 9, 3, giving rank 1 to the smallest and sharing ranks where values are equal.

[2]

- 4** Interpret each of the following values of Spearman's rank correlation coefficient: (i) $r_s = 1$, (ii) $r_s = -0.8$, (iii) $r_s = 0.02$.

[3]

- 5** Two judges rank five dancers. The table shows the ranks given. Calculate Spearman's rank correlation coefficient.

Judge A	1	2	3	4	5
Judge B	2	1	4	3	5

[3]

- 6** Two judges rank seven entries in exactly opposite orders (rank 1 and rank 7 are swapped, rank 2 and 6, and so on). Show that Spearman's rank correlation coefficient is -1 .

[3]

- 7** Eight students sit two tests. Their marks are shown in the table. Calculate Spearman's rank correlation coefficient between the two sets of marks, giving rank 1 to the highest mark.

Test 1	56	72	45	80	63	91	68	59
Test 2	60	70	52	75	58	85	72	66

[5]

- 8** Two judges rank six contestants A to F as shown.
Calculate Spearman's rank correlation coefficient, and comment on the level of agreement.

Contestant	A	B	C	D	E	F
Judge 1	2	1	4	3	6	5
Judge 2	1	3	2	5	4	6

[4]

- 9** Six pairs of values satisfy $y = x^3$: (1, 1), (2, 8), (3, 27), (4, 64), (5, 125), (6, 216).
(a) Find Spearman's rank correlation coefficient for these data. [2]
(b) The product moment correlation coefficient for these data is 0.938 to 3 significant figures. Explain why the two coefficients differ. [3]

- 10** Six pairs of observations are (8, 14), (12, 18), (12, 16), (15, 22), (17, 21), (20, 25).
Rank the data (smallest = 1), allowing for the tie, and use the formula $r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$ to calculate an approximate value of Spearman's coefficient. [5]

- 11** For n pairs of ranks, $\sum d^2 = 12$ and Spearman's rank correlation coefficient is $\frac{11}{14}$.
Find n . [4]

- 12** Six wines A to F are ranked by two tasters. Taster 1 ranks them A, B, C, D, E, F as 3, 1, 5, 2, 6, 4. Taster 2 ranks A, B, C, D, F as 2, 1, 4, 3, 6 but the rank for E is missing.
Find the missing rank and calculate Spearman's rank correlation coefficient. [5]

- 13** For each of the following situations, state whether Spearman's rank correlation coefficient or the product moment correlation coefficient is more appropriate, giving a reason.
(i) Two judges' rankings of ten singers. (ii) The heights and masses of 40 adults, where the scatter diagram is an elliptical cloud. (iii) Data with a clearly curved but always increasing relationship. [4]

- 14** Eight pairs of observations are shown in the table. Some values are repeated.
Rank each variable (smallest = 1) and calculate Spearman's rank correlation coefficient by finding the product moment correlation coefficient of the ranks.

x	14	22	22	9	31	17	26	22
y	3.1	4.5	3.8	2.2	5.9	3.8	5.1	4.9

[7]

- 15** Three judges A, B and C each rank six entries in a competition. The table shows the ranks awarded to the entries.
Calculate Spearman's rank correlation coefficient for each pair of judges, and state which two judges agree most closely.

Entry	1	2	3	4	5	6
Judge A	1	2	3	4	5	6
Judge B	2	1	3	5	4	6
Judge C	3	2	1	6	5	4

[7]

- 16** Seven pairs of values are (4, 10), (6, 7), (6, 12), (9, 14), (11, 20), (11, 15), (14, 22). Calculate Spearman's rank correlation coefficient by finding the product moment correlation coefficient between the ranks. [6]
-
- 17** A company ranks 8 job applicants by interview score and by aptitude test score. Spearman's rank correlation coefficient is found to be 0.83. Interpret this in context and state one limitation of using Spearman's coefficient here. [4]
-
- 18** Two examiners each rank eight scripts. Examiner A ranks the scripts 1, 2, 3, 4, 5, 6, 7, 8 and examiner B ranks them 2, 1, 3, 4, 5, 6, 8, 7. Calculate Spearman's rank correlation coefficient, and explain why the value would be unchanged if the two swapped pairs of adjacent scripts were in different places in the ranking. [5]
-
- 19** Five competitors are ranked 1 to 5 by two judges. Show that the largest possible value of $\sum d^2$ is 40, and hence that the smallest value of Spearman's rank correlation coefficient is -1 . [5]
-
- 20** Eight pairs of ranks give Spearman's rank correlation coefficient 0.5. For seven of the pairs the differences d between the ranks are $-1, -1, 0, 1, 1, 2, 3$. Find the difference d for the eighth pair, using the fact that the sum of the differences is zero. [5]
-
- 21** The table shows eight pairs of observations, one of which has an extreme x -value.
- | | | | | | | | | |
|----------|----|----|----|----|----|----|----|-----|
| x | 3 | 5 | 6 | 8 | 9 | 12 | 14 | 220 |
| y | 12 | 10 | 15 | 17 | 16 | 20 | 21 | 19 |
- (a) Calculate the product moment correlation coefficient. [3]
 (b) Calculate Spearman's rank correlation coefficient. [3]
 (c) Explain why the two coefficients are so different, and which is more suitable. [2]
-
- 22** Eight values of a variable are 5, 8, 8, 8, 12, 15, 15, 20.
 (a) Find the ranks (smallest = 1), sharing ranks for tied values. [3]
 (b) Show that the sum of the ranks is the same as it would be with no ties, and find the value of $\sum r^2$. [3]
 (c) Hence find $S_{rr} = \sum r^2 - \frac{(\sum r)^2}{8}$ and compare it with its value, 42, when there are no ties. [2]
-
- 23** Six students' maths marks and physics marks are listed from highest to lowest. For maths the student gives rank 1 to the highest mark, but for physics the student gives rank 1 to the lowest mark. The student's ranks are: maths 1, 2, 3, 4, 5, 6 (students P to U) and physics 5, 6, 4, 2, 3, 1. The student calculates $\sum d^2 = 66$ and $r_s = -0.886$, and concludes that students who do well in maths tend to do badly in physics.
 (a) Explain the mistake in the way the ranks have been assigned. [2]
 (b) Convert the physics ranks to the same convention as maths and find the correct value of r_s . [4]
-
- 24** A coach predicts the finishing order of ten runners as 1, 2, 3, ..., 10. The actual finishing positions of runners 1 to 10 are 1, 3, 2, 4, 6, 5, 7, 8, 10, 9.
 (a) Calculate Spearman's rank correlation coefficient between the predicted and actual positions. [3]
 (b) Show that swapping the positions of two runners whose actual positions are adjacent increases $\sum d^2$ by 2 (starting from a perfect prediction), and use this to explain the value of $\sum d^2$ found in part (a). [4]
-

- 25** Twelve pairs of observations are shown in the table. Several values are tied. Calculate Spearman's rank correlation coefficient between x and y by finding the product moment correlation coefficient of the ranks.

x	34	41	41	28	55	47	41	62	50	39	28	58
y	6.1	7.4	6.8	5.2	8.9	8.1	7.4	9	8.1	6.8	5.5	8.5

[7]

- 26** Eight students are ranked on two tests. A student records the differences in ranks as $d = -1, 1, -1, 1, 0, -3, 1, 1$.

(a) Show that these differences cannot be correct.

[2]

(b) The sixth difference was in fact -2 . Find the corrected $\sum d$ and calculate Spearman's rank correlation coefficient.

[4]

- 27** Let r_x and r_y be the ranks $1, 2, \dots, n$ assigned to n pairs of observations, with no ties, and let $d = r_x - r_y$. You may use $\sum i = \frac{n(n+1)}{2}$ and $\sum i^2 = \frac{n(n+1)(2n+1)}{6}$.

Show that the product moment correlation coefficient of the ranks is $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$.

[9]

- 28** Using $\sum_{i=1}^n i = \frac{n(n+1)}{2}$ and $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$, show that the largest possible value of $\sum d^2$ for n pairs of ranks without ties is $\frac{n(n^2-1)}{3}$, and deduce that $r_s \geq -1$ with equality only when the rankings are reversed.

State the value of $\sum d^2$ for $n = 6$, and check your formula.

[8]

- 29** Six candidates are ranked on two measures. The x -ranks are $1, 2.5, 2.5, 4, 5, 6$ (a tie for second place) and the y -ranks are $2, 1, 4, 3, 6, 5$.

(a) Show that $\sum r_x^2$ is smaller by $\frac{1}{2}$ than the value 91 it would take if there were no tie, and

$$\text{hence that } S_{xx} = \frac{n(n^2-1)}{12} - \frac{1}{2} = 17.$$

[3]

(b) Calculate Spearman's rank correlation coefficient using the product moment formula on the ranks.

[4]

(c) Calculate the value given by the simple formula $1 - \frac{6\sum d^2}{n(n^2-1)}$ and comment on the difference.

[2]

- 30** Four items are ranked 1 to 4 by one judge. A second judge, who cannot discriminate between the items, ranks them 1 to 4 at random, so that each of the $4! = 24$ possible orderings is equally likely.

(a) Show that for $n = 4$, $r_s = 1 - \frac{\sum d^2}{10}$.

[1]

(b) List the orderings of $1, 2, 3, 4$ that give $\sum d^2 = 0, 2$ and 4 , and find the probability that $r_s = 1$ and the probability that $r_s \geq 0.8$.

[6]

(c) Explain why, with only four items, a one-tailed test at the 5% significance level can only reject the hypothesis of no association when $r_s = 1$.

[2]

Linear Regression

- 1** A regression line of y on x is $y = 2.4 + 0.7x$. Estimate the value of y when $x = 10$.

[2]

- 2 For a set of data, $n = 10$, $S_{xy} = 60$, $S_{xx} = 40$, $\bar{x} = 5$ and $\bar{y} = 12$.
Find the equation of the regression line of y on x . [3]
- 3 A printing firm's costs are modelled by $C = 12 + 3.5n$, where C is the total cost in pounds and n is the number of posters printed.
Interpret the values 12 and 3.5 in this model. [3]
- 4 For each pair of variables, state which should be the explanatory (independent) variable and which the response (dependent) variable in a regression: (i) the mass of fertiliser applied to a field and the yield of wheat, (ii) the age of a car and its resale price. [2]
- 5 A regression line of y on x is found using data with x -values between 10 and 40.
The line is used to estimate y when $x = 25$ and when $x = 60$. Comment on the reliability of each estimate. [3]
- 6 A regression line is $y = 3 + 1.1x$. When $x = 10$ the observed value of y is 14.2.
Calculate the residual for this observation, and say whether the observed value is above or below the line. [2]
- 7 Find the equation of the regression line of y on x for the data (1, 3), (2, 5), (3, 6), (4, 9), (5, 10). [5]
- 8 For eight pairs of values, $\Sigma x = 76$, $\Sigma y = 212$, $\Sigma x^2 = 820$, $\Sigma y^2 = 6128$, $\Sigma xy = 2237$.
Find the equation of the regression line of y on x , giving the coefficients to 3 significant figures. [5]
- 9 For six pairs of values, $S_{xx} = 50.8$, $S_{yy} = 122$ and $S_{xy} = -78.5$, with $\bar{x} = 7.17$ and $\bar{y} = 13.5$.
Find the equation of the regression line of x on y , and use it to estimate x when $y = 10$. [5]
- 10 The regression line of y on x is $y = 2 + 0.5x$ and the regression line of x on y is $x = 1 + 0.8y$.
(a) Find the means $\bar{x}$ and $\bar{y}$. [3]
(b) Find the product moment correlation coefficient. [3]
- 11 The table shows the number of members, y (in hundreds), of a society in six years. The data are coded using $u = \text{year} - 2000$.
 $\Sigma u = 36$, $\Sigma y = 155$, $\Sigma u^2 = 286$, $\Sigma uy = 1037$. Find the regression line of y on u , and hence y on the year x .
- | Year | 2001 | 2003 | 2005 | 2007 | 2009 | 2011 |
|---------|------|------|------|------|------|------|
| Members | 18 | 22 | 23 | 28 | 31 | 33 |
- [5]
- 12 Two regression lines are fitted to different data. Line A has $r = 0.4$ and line B has $r = 0.95$ for the same type of variables, each based on 12 observations.
Explain which line is likely to give the more reliable predictions within the range of the data. [4]
- 13 A regression line of y on x has gradient 2.5 and passes through the point $(\bar{x}, \bar{y}) = (6, 20)$.
Find the equation of the line and estimate the value of y when $x = 8$. [3]

- 14** The table shows eight pairs of values of x and y .

x	3.2	4.1	5	5.8	6.7	7.5	8.4	9
y	12.5	15	17.9	19.5	23.4	25.6	28	29.1

- (a) Calculate the product moment correlation coefficient. [3]
 (b) Find the regression line of y on x and use it to estimate y when $x = 6$. [3]
 (c) Comment on the reliability of the estimate in part (b). [1]

- 15** For the data (2, 7), (4, 10), (6, 16), (8, 17), (10, 25), (12, 24), the regression line of y on x is $y = 3.40 + 1.87x$.

- (a) Calculate the residual for each point. [3]
 (b) Show that the residuals sum to zero (to rounding), and state which point is furthest from the line. [2]
 (c) Explain why residuals that show a systematic pattern would suggest a linear model is unsuitable. [1]

- 16** In an experiment, a chemist sets the temperature T ($^{\circ}\text{C}$) of a reaction mixture at six values and measures the mass M (g) of product formed: (10, 4.1), (20, 5.6), (30, 6.8), (40, 8.7), (50, 9.9), (60, 11.5).

- (a) State which regression line (M on T , or T on M) should be used to estimate M at $T = 35$, and give a reason. [2]
 (b) Find the equation of the regression line of M on T , and estimate M when $T = 35$. [4]

- 17** The regression line of y on x is $y = 3 + 2x$, where x is a distance in kilometres and y is a cost in pounds.

Find the regression line of cost y on distance m in miles, given that 1 mile = 1.6 km, and of cost in pence ($p = 100y$) on distance in kilometres.

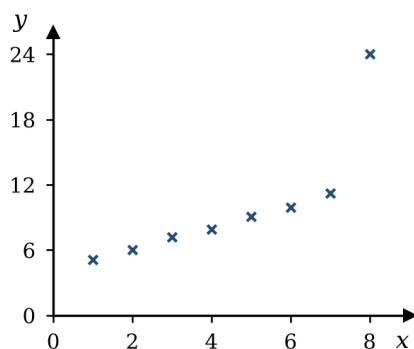
[5]

- 18** For a sample, $\bar{x} = 10$, $\bar{y} = 30$, the standard deviations are $s_x = 2$ and $s_y = 5$, and the product moment correlation coefficient is 0.8.

Find the equation of the regression line of y on x , and of x on y .

[5]

- 19** The scatter diagram shows eight observations of (x, y) . The last observation, (8, 24.0), is suspected to be an error. The data are (1, 5.1), (2, 6.0), (3, 7.2), (4, 7.9), (5, 9.1), (6, 9.9), (7, 11.2), (8, 24.0).



- (a) Find the regression line of y on x using all eight points. [3]
 (b) Find the regression line using only the first seven points, and compare the two lines. [4]

- 20** For the data (5, 14), (8, 15), (11, 22), (14, 21), (17, 28), (20, 30), a researcher wants to estimate the value of x for which $y = 25$.

- (a) Find the regression line of x on y and use it to estimate x when $y = 25$. [4]
 (b) The researcher instead rearranges the regression line of y on x , $y = 7.62 + 1.12x$. Find the estimate this gives and explain which method is better. [2]

- 21** Observations (x_i, y_i) have regression line $y = a + bx$, where $b = \frac{S_{xy}}{S_{xx}}$ and $a = \bar{y} - b\bar{x}$. The residuals are $e_i = y_i - a - bx_i$.
Show that $\sum e_i = 0$ and that $\sum x_i e_i = 0$. [6]

- 22** The table shows the annual sales, in thousands of units, of a product over ten years. Let $t = \text{year} - 2014$.

Year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Sales	41	44	46	51	53	58	59	64	67	70

- (a) Find the regression line of sales s on t , giving the coefficients to 3 significant figures. [4]
(b) Interpret the gradient in context, and use the line to estimate the sales in 2017 and calculate the residual for 2017. [3]
(c) A manager uses the line to forecast sales for 2032. Calculate the forecast and give one reason why it may be unreliable. [1]

- 23** A student has data with $n = 6$, $S_{xx} = 42.0$, $S_{yy} = 212$, $S_{xy} = 94.0$, $\bar{x} = 6.00$ and $\bar{y} = 18.0$. They calculate the regression line of y on x as follows:

“ $b = S_{xx}/S_{xy} = 0.447$, $a = \bar{x} - b\bar{y}$, so $y = a + bx$.”

- (a) Identify the two errors. [2]
(b) Find the correct regression line of y on x . [3]
(c) Use the line to estimate y when $x = 6$. [1]

- 24** The regression line of y on x is $y = 0.6x + 4$ and the regression line of x on y is $x = 0.9y - 1$.

- (a) Find the means $\bar{x}$ and $\bar{y}$ of the data. [3]
(b) Find the product moment correlation coefficient, and state why it is positive. [3]
(c) Find the ratio $s_y : s_x$ of the standard deviations. [2]

- 25** Variables x and y are standardised to $z_x = \frac{x - \bar{x}}{s_x}$ and $z_y = \frac{y - \bar{y}}{s_y}$.

Show that the regression line of z_y on z_x is $z_y = r z_x$, where r is the product moment correlation coefficient. [6]

- 26** Students' scores on two tests, each with a mean of 60 and a standard deviation of 10, have a product moment correlation coefficient of 0.6. A student scores 75 on test 1.

- (a) Find the regression line of test 2 score (y) on test 1 score (x), and predict the student's score in test 2. [4]
(b) The student scored 1.5 standard deviations above the mean on test 1. By how many standard deviations above the mean is the predicted score on test 2? Explain what this shows. [2]

- 27** The regression line of y on x has gradient $b_{yx} = \frac{S_{xy}}{S_{xx}}$ and the regression line of x on y is $x = a' + b_{xy}y$ with

$$b_{xy} = \frac{S_{xy}}{S_{yy}}.$$

- (a) Show that $b_{yx}b_{xy} = r^2$, and that both gradients have the same sign as r . [3]
(b) Plotted with y vertical, the line of x on y has gradient $\frac{1}{b_{xy}}$. Show that when $r > 0$ the line of y on x is the less steep of the two, and that the lines coincide only when $r = \pm 1$. [3]
(c) A student claims that for some data the regression line of y on x is $y = 1.5x + 1$ and that of x on y is $x = 0.8y + 2$. Show that this is impossible. [2]

- 28** Two samples of paired data are collected. Sample A has $n = 8$, $\Sigma x = 64$, $\Sigma y = 162$, $\Sigma x^2 = 604$, $\Sigma xy = 1478$. Sample B has $n = 7$, $\Sigma x = 88$, $\Sigma y = 231$, $\Sigma x^2 = 1244$, $\Sigma xy = 3146$.
- (a) Find the regression line of y on x for sample A alone and for sample B alone. [4]
- (b) Combine the samples (15 pairs) and find the regression line of y on x for the combined data. Is it the average of the two separate lines? [5]
-
- 29** For the least squares regression line $\hat{y} = a + bx$, with fitted values $\hat{y}_i$ and residuals $e_i = y_i - \hat{y}_i$, you may assume $\Sigma e_i = 0$ and $\Sigma(x_i - \bar{x})e_i = 0$.
- (a) Show that $\Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma e_i^2$, that is $S_{yy} = b^2 S_{xx} + \Sigma e_i^2$. [4]
- (b) Hence show that $\Sigma e_i^2 = S_{yy}(1 - r^2)$. [2]
- (c) For the data (1, 4), (3, 5), (4, 9), (6, 8), (8, 13), (10, 14), calculate r^2 and Σe_i^2 , and say what proportion of the variation in y is explained by the line. [2]
-
- 30** The true values x of an explanatory variable are measured with error, so the recorded values are $w = x + v$, where v is the error. In a sample, $S_{xx} = 40$, $S_{xy} = 30$, $S_{vv} = 10$, and the errors are uncorrelated in the sample with both x and y ($S_{xv} = S_{yv} = 0$).
- (a) Find the gradient b of the regression line of y on x , and show that the gradient b' of the regression line of y on w is $b' = b \frac{S_{xx}}{S_{xx} + S_{vv}}$. [5]
- (b) Calculate b' and state the effect of measurement error in the explanatory variable on the fitted gradient. [2]
- (c) Both samples have $\bar{w} = \bar{x} = 20$ and $\bar{y} = 50$. Find the two regression lines and the predicted value of y at $x = 30$ from each. Explain the difference. [2]

Least Squares Regression

- 1** State what quantity is minimised in the method of least squares when finding the regression line of y on x , and write it in terms of the observations (x_i, y_i) and the line $y = a + bx$. [2]
-
- 2** Calculate the sum of the squares of the residuals for the line $y = 1 + x$ and the points (1, 2), (2, 3), (3, 5), (4, 5). [3]
-
- 3** For a set of data, $S_{xy} = 34$, $S_{xx} = 17$, $\bar{x} = 3$ and $\bar{y} = 10$. Find the equation of the least squares regression line of y on x . [3]
-
- 4** Give two reasons why the residuals are squared before they are added when fitting a regression line by least squares. [2]
-
- 5** The least squares regression line for the points (1, 3), (2, 5), (3, 6) is $y = \frac{5}{3} + \frac{3}{2}x$. Calculate the three residuals and show that they sum to zero. [3]
-
- 6** Two lines are proposed for the points (1, 2), (2, 5), (3, 6), (4, 9): line A is $y = 2x$ and line B is $y = 1 + 1.5x$. Calculate $\Sigma(y - \hat{y})^2$ for each line and state which fits better by the least squares criterion. [3]
-
- 7** The normal equations for the least squares line $y = a + bx$ are $\Sigma y = na + b\Sigma x$ and $\Sigma xy = a\Sigma x + b\Sigma x^2$. For five points, $\Sigma x = 15$, $\Sigma y = 40$, $\Sigma x^2 = 55$ and $\Sigma xy = 146$. Solve the normal equations to find a and b . [5]

- 8** The sum of squared residuals is $S = \sum (y_i - a - bx_i)^2$.
By finding $\frac{\partial S}{\partial a}$ and setting it equal to zero, show that the least squares line passes through $(\bar{x}, \bar{y})$. [4]
- 9** For a set of data, $S_{xx} = 20$, $S_{yy} = 45$ and $S_{xy} = 24$. Calculate the sum of squares of the residuals for the least squares regression line of y on x , using $\sum e^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$. [4]
- 10** A line through the origin, $y = bx$, is fitted to the points (1, 2.2), (2, 3.9), (3, 6.3), (4, 7.8), (5, 10.4) by minimising $S = \sum (y_i - bx_i)^2$.
Show that $b = \frac{\sum xy}{\sum x^2}$ and calculate it. [5]
- 11** A line with fixed intercept 2, $y = 2 + bx$, is fitted to the points (0, 2.9), (1, 4.8), (2, 7.1), (3, 8.9), (4, 11.2) by least squares.
Show that the value of b minimising $\sum (y_i - 2 - bx_i)^2$ is $b = \frac{\sum x(y - 2)}{\sum x^2}$, and calculate it. [5]
- 12** The least squares line for the points (2, 5), (4, 9), (6, 11), (8, 16) is $y = 1.5 + 1.75x$.
Calculate the residuals and verify that $\sum e = 0$ and $\sum xe = 0$. [5]
- 13** For a data set, the product moment correlation coefficient is 0.9 and $S_{yy} = 50$. Find the sum of the squares of the residuals about the regression line of y on x . [4]
- 14** Five observations of (x, y) are (2, 5), (4, 9), (6, 10), (8, 15), (10, 16). A line $y = a + bx$ is fitted by least squares.
(a) Write down $S = \sum (y_i - a - bx_i)^2$, find $\frac{\partial S}{\partial a}$ and $\frac{\partial S}{\partial b}$, and hence show that the normal equations are $\sum y = na + b\sum x$ and $\sum xy = a\sum x + b\sum x^2$. [4]
(b) Use the normal equations to find a and b for these data. [3]
- 15** By expanding $S = \sum (y_i - a - bx_i)^2$, show that for a fixed value of b , S is minimised when $a = \bar{y} - b\bar{x}$, and that the minimum value is $S = S_{yy} - 2bS_{xy} + b^2S_{xx}$. [6]
- 16** The least squares line of x on y minimises the sum of the squares of the horizontal distances, $\sum (x_i - c - dy_i)^2$.
For the points (1, 3.0), (2, 4.1), (4, 6.2), (5, 6.9), (7, 9.8), (9, 11.0), find the equation of this line and compare it with the regression line of y on x . [6]
- 17** The residuals about the least squares line $y = a + bx$ are $e_i = y_i - a - bx_i$.
Use the normal equation $\sum xy = a\sum x + b\sum x^2$ and $\sum y = na + b\sum x$ to show that $\sum x_i e_i = 0$, and explain what this means. [5]
- 18** A line through the origin $y = bx$ is fitted to the points (1, 2.1), (2, 3.8), (3, 6.3), (4, 7.9). The points have weights 1, 2, 2, 1 respectively, so the quantity minimised is $S = \sum w_i (y_i - bx_i)^2$.
Show that $b = \frac{\sum w_i xy}{\sum w_i x^2}$ and calculate it. [6]

- 19** The table gives a company's profit y (in £ thousands) for five years x . Let $u = \frac{x - 2002}{2}$.

Find the least squares line of y on u , and hence y on x .

x	1998	2000	2002	2004	2006
y	14	16.5	19	21	24

[6]

- 20** Show that the minimum value of $S = \sum(y_i - a - bx_i)^2$ is $S_{yy} - \frac{S_{xy}^2}{S_{xx}}$, and deduce that this equals $S_{yy}(1 - r^2)$.

[7]

- 21** The model $y = a + bx^2$ is to be fitted by least squares to the data (1, 3.4), (2, 5.1), (3, 9.2), (4, 14.8), (5, 22.5), (6, 31.2).

(a) By minimising $S = \sum(y_i - a - bx_i^2)^2$, show that the normal equations are $\sum y = na + b\sum x^2$ and $\sum x^2 y = a\sum x^2 + b\sum x^4$.

[3]

(b) Hence find a and b .

[4]

- 22** A line is required that passes through the fixed point (2, 5) and fits the points (0, 1), (1, 3), (3, 6), (4, 7), (6, 11) as closely as possible in the least squares sense.

Show that the gradient is $b = \frac{\sum(x - 2)(y - 5)}{\sum(x - 2)^2}$ and calculate it, and hence write down the line.

[7]

- 23** A student tries to find the least squares line $y = a + bx$ by minimising $T = \sum(y_i - a - bx_i)$ (without squaring), writing:

" $\frac{\partial T}{\partial a} = -n = 0$ and $\frac{\partial T}{\partial b} = -\sum x_i = 0$, so no solution exists."

(a) Explain what is wrong with minimising T .

[2]

(b) Show correctly that minimising $S = \sum(y_i - a - bx_i)^2$ gives $b = \frac{S_{xy}}{S_{xx}}$, for the data (1, 3), (2, 7), (3, 8), (4, 12), (5, 13).

[4]

- 24** A line of gradient b is drawn through the mean point $(\bar{x}, \bar{y})$ of a data set, so $y = \bar{y} + b(x - \bar{x})$.

(a) Show that the sum of squared residuals is $S(b) = S_{yy} - 2bS_{xy} + b^2S_{xx}$.

[3]

(b) For the points (2, 4), (3, 6), (5, 7), (6, 11), (9, 12), calculate $S(b)$ for $b = 1$ and $b = 1.2$, and compare with the least squares value of b .

[4]

- 25** For data (x_i, y_i) , define $u_i = x_i - \bar{x}$, so that $\sum u_i = 0$. A line $y = \alpha + \beta u$ is fitted by least squares.

(a) Show that the normal equations for α and β are $\sum y = n\alpha$ and $\sum uy = \beta \sum u^2$, and hence that $\alpha = \bar{y}$ and $\beta = \frac{\sum uy}{\sum u^2}$.

[4]

(b) Use this to find the regression line of y on x for the data (4, 5), (6, 8), (7, 8), (9, 12), (12, 15), and give the line in terms of x .

[3]

- 26** For observations $y_1, \dots, y_n$, let $S(c) = \sum(y_i - c)^2$.

(a) Show that $S(c)$ is minimised when $c = \bar{y}$, and that the minimum value is S_{yy} .

[3]

(b) The least squares line with $b = 0$ is $y = \bar{y}$. Explain why $r^2 = 1 - \frac{\sum e^2}{S_{yy}}$ can be interpreted as the proportion of the variation in y explained by the regression on x , and calculate r^2 for data with $S_{yy} = 30$ and $\sum e^2 = 4.8$.

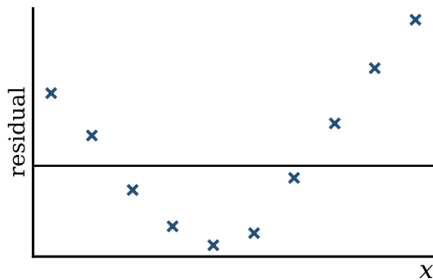
[3]

- 27** Let $a^* = \bar{y} - b^*\bar{x}$ and $b^* = \frac{S_{xy}}{S_{xx}}$ be the least squares estimates.
Show that for any a and b , $S(a, b) = \sum (y_i - a - bx_i)^2 = S_{\min} + n(a - a^* + (b - b^*)\bar{x})^2 + S_{xx}(b - b^*)^2$, where $S_{\min} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$, and deduce that the least squares solution is the unique minimiser. [9]
- 28** To reduce the effect of noise, a modified criterion is used for fitting a line through the mean point $\bar{y} - \bar{y} = b(x - \bar{x})$: minimise $S_\lambda(b) = \sum [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 + \lambda b^2$, where $\lambda > 0$ is a constant.
- (a) Show that the minimising value of b is $b_\lambda = \frac{S_{xy}}{S_{xx} + \lambda}$, and verify that it is a minimum. [4]
- (b) For the data (1, 2.0), (2, 3.1), (3, 5.2), (4, 5.9), (5, 8.1) find b_λ when $\lambda = 5$ and compare it with the ordinary least squares gradient. [3]
- (c) State what happens to b_λ as $\lambda \rightarrow 0$ and as $\lambda \rightarrow \infty$. [1]
- 29** Instead of vertical distances, a line through $(\bar{x}, \bar{y})$ is chosen to minimise the sum of the squares of the perpendicular distances from the points to the line. For data with $S_{xx} = 9$, $S_{yy} = 6$ and $S_{xy} = 2$, the line has direction vector $(\cos \theta, \sin \theta)$, and the sum of squared perpendicular distances is $D(\theta) = S_{xx}\sin^2\theta - 2S_{xy}\sin\theta\cos\theta + S_{yy}\cos^2\theta$.
- (a) Show that D is stationary when $\tan 2\theta = \frac{2S_{xy}}{S_{xx} - S_{yy}}$, and find the gradient $m = \tan \theta$ of the best line. [5]
- (b) Calculate the minimum value of D , and compare it with the sum of squared vertical residuals about the ordinary regression line of y on x . [3]
- (c) Find the gradient of the ordinary regression line of y on x and of x on y (as a gradient of y against x), and show that m lies between them. [2]
- 30** A line $y = a + bx$ fitted to n points has residuals $e_i = y_i - a - bx_i$ satisfying $\sum e_i = 0$ and $\sum x_i e_i = 0$, where the x_i are not all equal.
- (a) Show that a and b must be the least squares values $a = \bar{y} - b\bar{x}$ and $b = \frac{S_{xy}}{S_{xx}}$. [4]
- (b) Hence explain why the least squares line is the only line whose residuals have both zero sum and zero correlation with x . [1]
- (c) For the points (1, 4), (2, 5), (4, 9), (5, 8), the line $y = 2.5 + 1.5x$ has residuals with sum -2 and $\sum xe = -9$. Decide whether it is the least squares line, and if not find the least squares line. [2]

Regression Modelling

- 1** The relationship between two variables is modelled by $y = 5x^3$. By taking natural logarithms, state what should be plotted on each axis to obtain a straight line, and give the gradient and the intercept of that line. [3]
- 2** Data are modelled by $\log_{10}y = 0.5 + 1.2 \log_{10}x$. Express y in terms of x in the form $y = ax^n$. [3]

- 3 A straight line was fitted to some data. The residuals are plotted against x in the diagram. State what the pattern of the residuals suggests about the suitability of a linear model.



[2]

- 4 Data are modelled by $y = ab^x$. A regression of $\ln y$ on x gives $\ln y = 1.2 + 0.3x$. Find the values of a and b to 3 significant figures.

[3]

- 5 A population is modelled by $P = 2.5 \times 1.4^t$, where t is in years. Find the predicted population when $t = 6$, to 3 significant figures.

[2]

- 6 The values of y for $x = 1, 2, 3, 4$ are 3, 6, 12, 24. Show that $\ln y$ plotted against x gives a straight line and find its gradient.

[3]

- 7 It is thought that $y = ax^b$. Six pairs of readings (x, y) are (1, 3.1), (2, 8.6), (3, 15.8), (4, 24.2), (5, 33.5), (6, 44.4). Let $X = \ln x$ and $Y = \ln y$. Find the regression line of Y on X and hence estimate a and b .

x	1	2	3	4	5	6
y	3.1	8.6	15.8	24.2	33.5	44.4

[6]

- 8 It is suggested that $y = ab^x$. The table shows the values of y for $x = 0, 1, \dots, 5$: 4.1, 6.0, 9.2, 13.1, 19.8, 29.5.

x	0	1	2	3	4	5
y	4.1	6	9.2	13.1	19.8	29.5

- (a) Find the regression line of $\ln y$ on x and hence estimate a and b .
 (b) Estimate y when $x = 6$.

[4]

[2]

- 9 It is thought that $y = a + \frac{b}{x}$. Five readings are (1, 11.2), (2, 7.1), (4, 4.6), (5, 4.1), (8, 3.2). Let $u = \frac{1}{x}$.

Find the regression line of y on u and hence estimate a and b .

[5]

- 10 It is thought that $y = a + bx^2$. Five readings (x, y) are (1, 3.2), (2, 5.9), (3, 11.0), (4, 18.2), (5, 27.1).

Use the substitution $w = x^2$ to estimate a and b .

[5]

- 11 A town's population is modelled by $P = P_0 e^{kt}$ with $k = 0.03$, where t is the time in years. Find the percentage increase in the population each year (to 3 significant figures) and the time taken for the population to double.

[4]

- 12 A regression model is fitted to data with $x = 1, 2, \dots, 8$. The residuals are 0.31, -0.22 , 0.12, -0.35 , 0.28, -0.18 , 0.05, -0.02 .

Comment on whether the residuals suggest the model is appropriate.

[4]

- 13** Data are fitted by two models: a linear model $y = a + bx$, for which $r = 0.93$, and the model $y = ab^x$ (using a linear regression of $\ln y$ on x), for which $r = 0.998$. State, with a reason, which model is preferable, and describe how the preferred model could be written in a form suitable for prediction. [4]

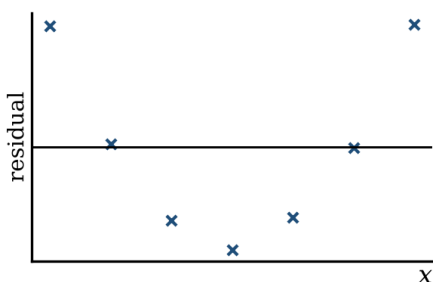
- 14** The period T seconds of a pendulum of length L metres is thought to satisfy $T = kL^n$. Experimental results are shown.
By taking logarithms, estimate k and n , and comment on your value of n .

L (m)	0.25	0.5	1	1.5	2
T (s)	1.01	1.42	2	2.46	2.83

[7]

- 15** The count rate N of a radioactive sample is recorded at times t minutes: (0, 980), (10, 640), (20, 410), (30, 270), (40, 175). The model is $N = N_0 e^{-kt}$.
(a) Use a regression of $\ln N$ on t to estimate N_0 and k . [4]
(b) Hence estimate the half-life of the sample. [2]

- 16** Seven observations are (1, 2.1), (2, 4.8), (3, 9.3), (4, 15.9), (5, 25.2), (6, 36.1), (7, 49.4). A student fits $y = a + bx$.



- (a) Show that the regression line is $y = -11.1 + 7.87x$, and calculate the residuals. [4]
(b) Describe the pattern in the residuals and suggest a better model. [2]
- 17** The data (1, 5.2), (2, 7.9), (3, 12.8), (4, 20.1), (5, 31.0), (6, 48.8) are modelled in two ways: model A is $y = a + bx$ and model B is $y = pq^x$.
(a) Calculate the product moment correlation coefficient between x and y , and between x and $\ln y$. [4]
(b) Hence state which model is better and estimate p and q for it. [1]

- 18** The number of users (in millions) of a service in six years is: 2016: 1.2, 2017: 1.9, 2018: 3.1, 2019: 4.9, 2020: 8.2, 2021: 12.9. Let $t = \text{year} - 2016$.

Year	2016	2017	2018	2019	2020	2021
Users	1.2	1.9	3.1	4.9	8.2	12.9

- (a) Fit the model $u = ab^t$, by regressing $\ln u$ on t , giving a and b to 3 significant figures. [4]
(b) Use the model to predict the number of users in 2024, and explain why this prediction should be treated with caution. [2]
- 19** A regression of $\log_{10} y$ on x gives $\log_{10} y = 0.3 + 0.8x$.
Express y in the form $y = ab^x$, giving a and b to 3 significant figures, and find the value of x for which $y = 100$. [5]

- 20** It is thought that $y = \frac{1}{c + dx}$. Six readings (x, y) are (1, 0.50), (2, 0.33), (3, 0.26), (4, 0.20), (5, 0.17), (6, 0.15).
- (a) Show that this model leads to a linear relationship between $z = \frac{1}{y}$ and x. [1]
- (b) Find the regression line of z on x, hence estimates of c and d, and use the model to predict y when x = 8. [6]
-
- 21** The model $y = a + bx + cx^2$ is fitted by least squares to the data (−2, 7.1), (−1, 3.9), (0, 2.2), (1, 3.1), (2, 6.8).
- (a) By minimising $S = \sum(y - a - bx - cx^2)^2$ with respect to a, b and c, show that the normal equations are $\sum y = na + b\sum x + c\sum x^2$, $\sum xy = a\sum x + b\sum x^2 + c\sum x^3$ and $\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$. [4]
- (b) Calculate the three sums needed and solve the equations to find a, b and c. [5]
-
- 22** A student wants to fit $y = ax^b$ and plots y against ln x, obtaining the regression line $y = 1.8 + 2.4 \ln x$. The student concludes that “a = 1.8 and b = 2.4”.
- (a) Explain why the student's conclusion is wrong. [2]
- (b) The correct regression of ln y on ln x is $\ln y = 1.8 + 2.4 \ln x$. Find a and b to 3 significant figures and the predicted value of y when x = 3. [4]
-
- 23** The data (1, 2.6), (2, 4.1), (3, 7.0), (4, 11.4), (5, 18.5) are fitted by model A, $y = a + bx$, and by model B, $y = pq^x$ (found by regressing ln y on x).
- (a) Find the equations of the two fitted models. [5]
- (b) Calculate the sum of the squares of the residuals $\sum(y - \hat{y})^2$ (on the original y scale) for each model, and decide which model fits better. [3]
-
- 24** The value V (in £ thousands) of an investment t years after it was made is recorded: (0, 120), (1, 138), (2, 158), (3, 183), (4, 209), (5, 241), (6, 276).
- (a) Show that the model $V = ab^t$ is suitable by calculating the correlation coefficient between ln V and t. [3]
- (b) Find a and b, and interpret b in context. [3]
- (c) Calculate the residual for t = 3 on the original scale, and predict V when t = 10, commenting on the reliability of the prediction. [2]
-
- 25** A power law $y = ax^b$ passes through the points (2, 6) and (8, 48).
- (a) Find a and b exactly. [4]
- (b) Test whether the point (4, 17) lies on the curve, and describe what this shows if it does not. [2]
-
- 26** The spread of a rumour is modelled by $y = \frac{100}{1 + e^{c-kt}}$, where y is the percentage of a school that has heard it t days after it started. Readings (t, y) are (0, 12), (1, 21), (2, 34), (3, 51), (4, 68), (5, 81).
- (a) Show that $\ln\left(\frac{100}{y} - 1\right) = c - kt$. [2]
- (b) Use a regression of $Z = \ln\left(\frac{100}{y} - 1\right)$ on t to estimate c and k. [4]
- (c) Estimate the day on which half the school has heard the rumour. [2]

- 27** It is thought that $y = ax^b\varepsilon$, where ε is a multiplicative random error close to 1. Six observations (x, y) are (1, 3.4), (2, 10.2), (3, 19.6), (4, 33.8), (5, 50.9), (6, 69.5).
- Show that taking natural logarithms turns the multiplicative error into an additive error, and explain why this makes a linear regression of $\ln y$ on $\ln x$ appropriate. [2]
 - Fit the model, giving a and b to 3 significant figures. [4]
 - Calculate the residuals on the log scale, $\ln y - \ln \hat{y}$, for each point, and comment. [3]
 - Predict y at $x = 8$ and explain why a prediction on the original scale made by $e^{\ln \hat{y}}$ may slightly underestimate the mean value of y . [3]
-
- 28** The values of y at $x = 0, 1, 2, \dots, 7$ are 3.0, 4.4, 6.5, 9.4, 13.8, 20.3, 30.1, 44.0.
- Calculate the ratios of consecutive values y_{x+1}/y_x and explain why they suggest the model $y = ab^x$. [3]
 - Fit the model using a regression of $\ln y$ on x , and give the percentage growth per unit of x . [4]
 - Fit $y = a + bx$ to the data instead. Calculate the residuals at $x = 0, 3$ and 7 and explain what their signs show. [2]
-
- 29** A cup of coffee cools in a room at temperature 20°C . According to Newton's law of cooling, its temperature $T^\circ\text{C}$ after t minutes is $T = 20 + Ae^{-kt}$. Measurements (t, T) are (0, 95), (5, 71.5), (10, 55.3), (15, 44.0), (20, 36.0), (25, 30.5).
- Show that $\ln(T - 20)$ is a linear function of t . [1]
 - Use a regression of $\ln(T - 20)$ on t to estimate A and k . [5]
 - Predict the temperature of the coffee after 40 minutes. [2]
 - A student wrongly assumes the room temperature is 22°C and regresses $\ln(T - 22)$ on t . Find the gradient they obtain and comment on the effect of the incorrect assumption on the linearity. [2]
-
- 30** A drug concentration y at time x hours after a dose is modelled by $y = axe^{bx}$. Readings (x, y) are (0.5, 1.62), (1, 2.69), (1.5, 3.27), (2, 3.69), (3, 3.62), (4, 3.10).
- Show that $\ln(\frac{y}{x})$ is a linear function of x and state the gradient and intercept in terms of a and b . [2]
 - Use a regression to estimate a and b . [4]
 - Show that the model has a maximum at $x = -\frac{1}{b}$, and use your values to find the time of peak concentration and the peak concentration. [2]

Statistical Inference for Correlation

- 1** A researcher wants to test whether there is positive correlation between the age of a car and its mileage. Write down suitable null and alternative hypotheses in terms of the population product moment correlation coefficient ρ . [2]
-
- 2** A test for positive correlation is carried out at the 5% significance level on 8 pairs of observations. The sample product moment correlation coefficient is $r = 0.81$ and the critical value is 0.6215.
State, with a reason, whether $H_0: \rho = 0$ is rejected. [3]
-
- 3** The marks of 10 students in a mathematics test and a physics test have product moment correlation coefficient $r = 0.62$. The critical value for a one-tailed test at the 5% level is 0.5494, for $H_0: \rho = 0$ against $H_1: \rho > 0$.
Write down the conclusion of the test in context. [3]

4 State whether a one-tailed or a two-tailed test is appropriate in each case, and give the alternative hypothesis. (i) "Is there any correlation between height and shoe size?" (ii) "Does taller height tend to mean larger shoe size?" (iii) "Do cars with more mileage tend to be cheaper?" [3]

5 In a hypothesis test, the significance level is 5%. Explain what this means. [2]

6 A sample of 5 pairs of observations has $r = 0.5$. The critical value for a one-tailed test at the 5% level with $n = 5$ is 0.8054. Carry out the test of $H_0: \rho = 0$ against $H_1: \rho > 0$ and comment on the effect of the small sample. [3]

7 For 10 pairs of observations $r = 0.68$. Test, at the 5% significance level, whether there is evidence of positive correlation, given that the critical value is 0.5494. [4]

8 A sample of 12 pairs gives $r = -0.58$. Test at the 5% level whether there is any correlation, given that the two-tailed critical value for $n = 12$ is 0.5760. [4]

9 Nine pairs of observations give $n = 9$, $\Sigma x = 84$, $\Sigma y = 166$, $\Sigma x^2 = 932$, $\Sigma y^2 = 3324$ and $\Sigma xy = 1738$. Calculate r and test, at the 5% level, for positive correlation, given that the critical value is 0.5822. [6]

10 Two judges rank 8 contestants and Spearman's rank correlation coefficient is $r_s = 0.77$. The critical value for a two-tailed test at the 5% level for $n = 8$ is 0.7381. Test whether there is any association between the judges' rankings. [4]

11 A study finds a significant positive correlation between the number of ice creams sold at a seaside resort and the number of sunburn cases treated each day. Does this show that ice cream causes sunburn? Explain. [3]

12 A sample of 15 pairs gives $r = -0.62$. Test, at the 1% significance level, for negative correlation, given that the one-tailed critical value is -0.5923 . [4]

13 A statistician carries out a hypothesis test for correlation and obtains a p-value of 0.032. State the conclusion at the 5% significance level and at the 1% significance level, and explain what the p-value means. [4]

14 The table shows the number of hours of sleep deprivation (x) and the number of errors (y) made by six volunteers in a task. Calculate the product moment correlation coefficient and test, at the 5% significance level, whether there is positive correlation. The critical value is 0.7293.

Hours	12	15	18	21	24	27
Errors	31	38	36	47	50	49

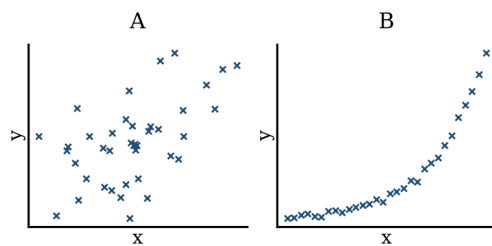
[7]

15 Seven students record their weekly hours of exercise (x) and their resting heart rate scores (y) as shown. The scatter diagram is not elliptical, so a Spearman rank test is used. Calculate Spearman's rank correlation coefficient and test, at the 5% level, for positive association. The critical value for $n = 7$ is 0.7143.

x	14	22	9	31	17	26	12
y	5.1	7.4	3.9	6.2	5.6	8.8	4

[7]

- 16** Scatter diagrams A and B show two samples of data, each with a significant positive correlation.



- (a) The PMCC hypothesis test assumes the data come from a bivariate normal distribution. State which diagram it is appropriate for, with a reason. [2]
- (b) State, with a reason, which test should be used for diagram B. [2]
- (c) Explain why a scatter diagram should be drawn before carrying out a PMCC test. [1]
-
- 17** For 12 pairs of observations from a bivariate normal population, $r = 0.62$. Under $H_0: \rho = 0$, the statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ has a t-distribution with $n - 2$ degrees of freedom. Test for correlation (two-tailed) at the 5% level, given that the critical value of t_{10} is 2.2281. [6]
-
- 18** For $n = 20$ pairs of observations, the critical value of the t-distribution with 18 degrees of freedom for a two-tailed test at the 5% level is 2.1009. Using $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, find the smallest value of $|r|$ that is significant at the 5% level. [6]
-
- 19** In a study with 10 pairs of observations the sample correlation coefficient is $r = 0.40$. Another study with 30 pairs also gives $r = 0.40$. The critical values for a one-tailed test at the 5% level are 0.5494 for $n = 10$ and 0.3061 for $n = 30$. Carry out the test in each case and comment on the difference in the conclusions. [5]
-
- 20** Sample A (10 pairs) has $r = 0.55$ and sample B (25 pairs) has $r = 0.40$. Two-tailed critical values at the 5% level are 0.6319 for $n = 10$ and 0.3961 for $n = 25$. Test each sample for correlation at the 5% level and explain why the sample with the smaller r gives the stronger evidence. [6]
-
- 21** Under $H_0: \rho = 0$, $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ has a t-distribution with $n - 2$ degrees of freedom.
- (a) Show that $r = \frac{t}{\sqrt{n-2+t^2}}$. [3]
- (b) For $n = 8$, the critical value of t_6 for a one-tailed test at the 5% level is 1.9432. Find the corresponding critical value of r . [2]
- (c) Hence show that, as n increases, the critical value of r decreases. [2]
-
- 22** A student tests for any correlation between two variables using 12 pairs of observations, obtaining $r = -0.55$. The student writes: “ $H_0: \rho = 0$, $H_1: \rho < 0$. Critical value (one-tailed, 5%) is 0.4973. Since -0.55 is less than 0.4973, reject H_0 . This proves that the variables are negatively correlated.”
- (a) Identify three errors in the student's solution. [3]
- (b) Carry out the test correctly and state the conclusion. [3]

- 23** A farmer records the mean daily rainfall x (mm) in a month and the mass of hay y (tonnes per field) produced in eight fields: (4.1, 32), (5.6, 41), (6.0, 38), (7.2, 49), (8.5, 55), (9.1, 52), (10.4, 63), (11.8, 66). The farmer believes that more rain leads to more hay. Test this belief at the 1% significance level, assuming bivariate normality. The one-tailed critical value for $n = 8$ is 0.7887.

Rain	4.1	5.6	6	7.2	8.5	9.1	10.4	11.8
Hay	32	41	38	49	55	52	63	66

[8]

- 24** A study of 25 pairs of observations gives $r = 0.45$. The two-tailed critical value at the 5% level for $n = 25$ is 0.3961.

(a) Test whether there is correlation at the 5% level.

[3]

(b) Calculate r^2 and explain why the correlation, although statistically significant, may not be of practical use for predictions.

[3]

- 25** A sample of 9 pairs has $r = 0.72$. The one-tailed critical values of the t -distribution with 7 degrees of freedom are 1.8946 (5%), 2.9980 (1%).

Calculate the test statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ for $H_0: \rho = 0$ against $H_1: \rho > 0$, and decide the outcome of the test at the 5% and at the 1% levels.

[7]

- 26** A researcher tests 20 independent pairs of unrelated variables for correlation, each test at the 5% significance level.

(a) Find the probability that none of the tests gives a significant result, assuming all null hypotheses are true.

[2]

(b) Hence find the probability that at least one test is significant, and explain what this shows about searching for correlations among many variables.

[4]

- 27** For n pairs (x_i, y_i) , the regression line of y on x has gradient $b = \frac{S_{xy}}{S_{xx}}$. Under the usual

assumptions, the test statistic for $H_0: \text{slope} = 0$ is $t = \frac{b}{\sqrt{s^2/S_{xx}}}$, where $s^2 = \frac{\sum e^2}{n-2}$ and $\sum e^2 =$

$S_{yy}(1 - r^2)$, with $n - 2$ degrees of freedom.

Show that this statistic equals the statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ used to test $H_0: \rho = 0$, so that the two tests are equivalent.

[9]

- 28** For a sample of n pairs, the one-tailed 5% critical values of the product moment correlation coefficient are: $n = 8$: 0.6215, $n = 9$: 0.5822, $n = 10$: 0.5494, $n = 11$: 0.5214, $n = 12$: 0.4973, $n = 13$: 0.4762.

A researcher expects a true correlation of about 0.5. Find the smallest sample size for which a sample correlation of exactly 0.5 would lead to rejection of $H_0: \rho = 0$ in favour of $H_1: \rho > 0$ at the 5% level, and verify your answer by calculating the t -statistic for that sample size.

[7]

- 29** Ten pairs of observations (x, y) are shown in the table. The last pair, (19.0, 2.0), is suspected to be an error.

x	9.5	15	2	7.4	4.6	3.7	5.4	8.2	9.1	19
y	9.3	26.4	4.3	15.8	7.9	14.8	8.1	15.4	17.6	2

(a) Test at the 5% level for correlation using all ten pairs. The two-tailed critical value for $n = 10$ is 0.6319.

[4]

(b) Repeat the test without the suspect pair. The two-tailed critical value for $n = 9$ is 0.6664.

[4]

(c) Discuss what should be done about the suspect pair.

[1]

- 30** A school records the number of hours x that ten students spent on social media in a week and their test score y (out of 100). The values are shown.
- In each part, test at the 5% level (two-tailed). The critical values for $n = 10$ are 0.6319 for the PMCC and 0.6485 for Spearman's coefficient.

Hours x	17	59	26	20	11	53	29	27	25	31
Score y	13	70	45	22	43	69	50	49	35	36

[9]