



RAFFLES INSTITUTION
H2 Mathematics 9758
2025 Year 6

Additional Practice Questions for Chapter S6: Correlation and Regression

- 1** Seven cities in a certain country are linked by rail to the capital city. The table below shows the distance of each city from the capital and the rail fare from the city to the capital.

City	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>
Distance, x km	124	44	76	148	16	180	104
Rail fare, \$ y	156	53	99	169	23	177	138

- (i) Give a sketch of the scatter diagram for the data as shown on your calculator. [2]
- (ii) Calculate the product moment correlation coefficient. [1]
You are given that the regression line of y on x has equation $y = 16.7 + 1.01x$, where the coefficients are given correct to 3 significant figures.
- (iii) Calculate the equation of the regression line of x on y , giving your answer in the form $x = a + by$. [1]
- (iv) Use the appropriate regression line to estimate
- (a) the rail fare from a city that is 28 km from the capital, [2]
- (b) the distance of a city from the capital if the rail fare is \$198. [2]
- (v) Comment briefly on the reliability of the estimates in part (iv). [2]
- 2** An engineering company makes cranes. The numbers, x , sold in each three-month period for two years, together with the profits, y thousand dollars, on the sale of these cranes are given in the following table.

x	15	17	13	21	16	22	14	18
y	290	350	270	430	340	410	300	360

- (i) Give a sketch of the scatter diagram for the data as shown in your calculator. [2]
- (ii) Find $\bar{x}$ and $\bar{y}$, and mark the point $(\bar{x}, \bar{y})$ on your scatter diagram. [2]
- (iii) Calculate the equation of the regression line of y on x , and draw this line on your scatter diagram. [2]
- (iv) Calculate the product moment correlation, and comment on its value in relation to your scatter diagram. [2]
- (v) For the next three-month period, the sales target is 20 cranes. Estimate the corresponding profit. [2]
- (vi) The company's sales director uses the regression line in (iii) to predict the profit if 40 cranes were to be sold in a three-month period. Comment on the validity of this prediction. [2]

- 3 Explain why it is advisable to draw a scatter diagram before interpreting a correlation coefficient calculated for a sample drawn from a bivariate distribution. [1]

Sketch a scatter diagram that shows an obvious relation between two variables but which yields a linear product-moment correlation coefficient that is close to zero. [1]

- 4 A medical officer wishes to investigate a patient's heart-beat rate h beats per minute (bpm) when he walks at different speeds s km/h. The data is shown below:

s (km/h)	1	1.5	2	2.5	3	3.5	4	4.5	5
h (bpm)	60	63	66	75	86	99	150	110	130

- (i) Sketch a scatter plot of the above data. [1]

- (ii) One of the values of h appears to be incorrect. Indicate the corresponding point on your diagram by labelling it P . [1]

Omit P for the remainder of this question.

- (iii) Calculate the product moment correlation coefficient for this set of data. Use the equation of an appropriate regression line to predict the value of s when $h = 100$, justifying your choice of regression line. [4]

It is suggested to use one of the following two models instead:

$$\text{Model (I): } h = a + bs^2,$$

$$\text{Model (II): } h = a + be^s,$$

where a and b are real constants.

- (iv) Determine which of the two models is a better choice, giving a reason for your answer. [2]

- (v) Suppose a new data pair $(\bar{s}, \bar{h})$ is added to the table above, where $\bar{s}$ and $\bar{h}$ are the patient's sample mean walking speed (in km/h) and his sample mean heart-beat rate (in bpm) respectively, based on the data above. Without any calculations, explain whether the equation of the regression line you have obtained in part (iii) would change. [2]

- 5 (a)** Sketch a scatter diagram that might be expected when x and y are related approximately by $y = px^2 + t$ in each of the cases **(i)** and **(ii)** below. In each case your diagram should include 6 points, approximately equally spaced with respect to x , and with all x -values positive.

(i) p and t are both positive.

(ii) p is negative and t is positive. [2]

- (b)** The ages in months (m) and prices in dollars (P) of a random sample of ten used cars of a certain model are given in the table.

m	11	20	28	36	40	47	58	62	68	75
P	112800	102600	76500	72000	72000	69000	65800	57000	50600	47600

It is thought that the price after m months can be modelled by one of the formulae

$$P = am + b, \quad P = c \ln m + d,$$

where a , b , c and d are constants.

- (i)** Find, correct to 4 decimal places, the value of the product moment correlation coefficient between

(A) m and P ,

(B) $\ln m$ and P . [2]

- (ii)** Explain which of $P = am + b$ and $P = c \ln m + d$ is the better model and find the equation of a suitable regression line for this model. [3]

- (iii)** Use the equation of your regression line to estimate the price of a car that is 50 months old. [1]

- 6** Research is being carried out into how the concentration of a drug in the bloodstream varies with time, measured from when the drug is given. Observations of successive times give the data shown in the following table.

Time (t minutes)	15	30	60	90	120	150	180	240	300
Concentration (x micrograms per litre)	82	65	43	37	22	19	12	6	2

It is given that the value of the product moment correlation coefficient for this data is -0.912 , correct to 3 decimal places. Draw a scatter diagram for the data. [1]

Calculate the equation of the regression line of x on t . [2]

Calculate the corresponding estimated value of x when $t = 300$, and comment on the suitability of the linear model. [2]

The variable y is defined by $y = \ln x$. For the variables y and t ,

- (i)** calculate the product moment correlation coefficient and comment on its value, [2]

- (ii)** calculate the equation of the appropriate regression line. [3]

Use a regression line to give the best estimate that you can of the time when the drug concentration is 15 micrograms per litre. [2]

- 7 Mr. Lee recorded the length of time, y minutes, taken to travel to work when leaving home x minutes after 6.00 am on seven selected mornings. The results are as follows:

x /min	0	10	20	30	40	50	60
y /min	16	27	28	39	43	48	51

- (i) Calculate the equations of the least squares regression lines of y on x and x on y . [2]
- (ii) Mr. Lee forgot to add an observation for a particular morning but remembered that the two regression lines remain unchanged on its inclusion. Find this missing observation. [1]

Mr. Lee needs to report to work by 8.00 am. He leaves home x minutes after 6.00 am and he arrives at his workplace z minutes before 8.00 am.

- (iii) Find z in terms of x . [1]
- (iv) Hence estimate, to the nearest minutes, the latest time Mr. Lee can leave home without arriving late at work. [2]

- 8 A website about electric motors gives information about the percentage efficiency of motors depending on their power, measured in horsepower. Xian has copied the following table for a particular type of electric motor, but he has copied one of the efficiency values wrongly.

Power, x	1	1.5	2	3	5	7.5	10	20	30	40	50
Efficiency, $y\%$	72.5	82.5	84.0	87.4	87.5	88.5	89.5	90.2	91.0	91.7	92.4

- (i) Plot a scatter diagram on graph paper for these values, labeling the axes, using a scale of 2 cm to represent 10% efficiency on the y -axis and an appropriate scale for the x -axis. On your diagram, circle the point that Xian has copied wrongly. [2]

For parts (ii), (iii) and (iv) of this question you should **exclude** the point for which Xian has copied the efficiency value wrongly.

- (ii) Explain from your scatter diagram why the relationship between x and y should not be modelled by an equation of the form $y = ax + b$. [1]
- (iii) Suppose that the relationship between x and y are modelled by an equation of the form $y = \frac{c}{x} + d$, where c and d are constants. State with a reason whether each of c and d is positive or negative. [2]
- (iv) Find the product moment correlation coefficient and the constants c and d for the model in part (iii). [3]
- (v) Use the model $y = \frac{c}{x} + d$ with the values of c and d found in part (iv), to estimate the efficiency value (y) that Xian has copied wrongly. Give two reasons why you would expect this estimate to be reliable. [3]

- 9** For 3 years a factory records the number of units it produces each quarter and the total cost of producing the units. The following table summarises these results.

Units of output (x) in 1000's	15	30	57	75	10	25	49	70	20	34	60	76
Total cost (y) in £1000's	33	49	71	99	34	40	61	82	40	51	84	93

- (i) Give a sketch of the scatter diagram for the data as displayed on your graphing calculator. [2]
- (ii) Calculate the equation of the regression line of y on x , draw this line on your diagram and calculate the product moment correlation coefficient. [3]

The selling price of each unit of output is £1.80.

- (iii) Use your graph to estimate the level of output (correct to 2 significant figures) at which the total cost and the total income balance. [2]
- (iv) Give a brief interpretation of what this value means. [1]

- 10** In an investigation of a shrub, research workers measured the average width y (in cm) and stem density x (number of stems per m^2) at ten sites with the following results:

x	4	5	6	9	11	14	15	19	21	22
y	0.8	2	0.65	0.6	0	0.65	0.55	0.35	0.40	0.38

- (i) Calculate the correlation coefficient for the data. [2]
- (ii) Give a sketch of the scatter diagram for the data, as shown on your calculator, and comment on your answer in (i). [3]
- (iii) Identify the two data pairs which should be removed, and calculate the correlation coefficient, and the regression line of y on x , for the revised data. [3]
- (iv) Use the regression line of y on x in (iii) to demonstrate that it is unwise to extrapolate beyond the range of the data. [2]
- (v) A new data pair (a, b) is now obtained at an eleventh site, where $4 \leq a \leq 22$. Using the revised data in (iii) and the new data pair (a, b) , the regression equation of y on x is calculated and found to be identical to that using only the revised data. Find a possible data pair (a, b) . [2]

- 11** In an experiment the following information was gathered about air pressure P , measured in inches of mercury, at different heights above sea-level h , measured in feet.

h	2000	5000	10000	15000	20000	25000	30000	35000	40000	45000
P	27.8	24.9	20.6	16.9	13.8	11.1	8.89	7.04	5.52	4.28

- (i) Draw a scatter diagram for these values, labelling the axes. [1]
- (ii) Find, correct to 4 decimal places, the product moment correlation coefficient between
 (a) h and P ,
 (b) $\ln h$ and P ,
 (c) $\sqrt{h}$ and P . [3]
- (iii) Using the most appropriate case from part (ii), find the equation which best models air pressure at different heights. [3]
- (iv) Given that 1 metre = 3.28 feet, re-write your equation from part (iii) so that it can be used to estimate the air pressure when the height is given in metres. [2]
- 12** (i) Sketch a scatter diagram that might be expected when x and y are related approximately as given in each of the cases (A), (B) and (C) below. In each case your diagram should include 6 points approximately equally spaced with respect to x , and with all x - and y -values positive. The letters a , b , c , d , e and f represent constants.

(A) $y = a + bx^2$, where a is positive and b is negative,

(B) $y = c + d \ln x$, where c is positive and d is negative,

(C) $y = e + \frac{f}{x}$, where e is positive and f is negative. [3]

A motoring website gives the following about the distance travelled, y km, by a certain type of car at different speeds, x kmh⁻¹, on a fixed amount of fuel.

Speed, x	88	96	104	112	120	128
Distance, y	148	147	144	138	126	107

- (ii) Draw the scatter diagram for these values, labelling the axes. [1]
- (iii) Explain which of the three cases in part (i) is the most appropriate for modelling these values, and calculate the product moment correlation coefficient for this case. [2]
- (iv) It is required to estimate the distance travelled at a speed of 110 kmh⁻¹. Use the case that you identified in part (iii) to find the equation of a suitable regression line, and use your equation to find the required estimate. [3]

- 13** Many electronic devices need a fan to keep them cool. In order to maximise the lifetime of such fans, the speed they run at is reduced when conditions allow. Running a fan at a lower speed reduces the power required. The following table gives details, for a particular type of fan, of the power required (P watts) at different fan speeds (R revolutions per minute).

Fan speed (R)	3600	4500	5400	6300	7200	8100	9000	9900
Power (P)	0.22	0.34	0.52	0.78	1.06	1.48	2.04	2.64

- Draw a scatter diagram of these data. Use your diagram to explain whether the relationship between P and R is likely to be well modelled by an equation of the form $P = aR + b$, where a and b are constants. [2]
- By calculating the relevant product moment correlation coefficients, determine whether the relationship between P and R is modelled better by $P = aR + b$ or by $P = aR^2 + b$. Explain how you decide which model is better, and state the equation in this case. [5]
- Use your equation to estimate the speed of the fan when the power is 0.9 watts. Explain whether your estimate is reliable. [2]
- Use your equation to estimate the power used when the speed of the fan is 3300 revolutions per minute. Explain whether your estimate is reliable. [2]
- Re-write your equation from part (ii) so that it can be used when the speed of the fan, R , is given in revolutions per second. [1]

- 14** A car is placed in a wind tunnel and the drag force F for different wind speeds v , in appropriate units is recorded. The results are shown in the table.

v	0	4	8	12	16	20	24	28	32	36
F	0	2.5	5.1	8.8	11.2	13.6	17.6	22.0	27.8	33.9

- Draw the scatter diagram for these values, labelling the axes clearly. [2]

It is thought that the drag force F can be modelled by one of the formulae

$$F = a + bv \quad \text{or} \quad F = c + dv^2$$

where a , b , c and d are constants.

- Find, correct to 4 decimal places, the value of the product moment correlation coefficient between
 - v and F ,
 - v^2 and F .
 [2]
- Use your answers to parts (i) and (ii) to explain which of $F = a + bv$ or $F = c + dv^2$ is the better model. [1]
- It is required to estimate the value of v for which $F = 26.0$. Find the equation of a suitable regression line, and use it to find the required estimate. Explain why neither the regression line of v on F nor the regression line of v^2 on F should be used. [4]