2014-2018 Chapter 5: Correlation and Regrression

Mr Faruk
Teacher of Mathematics
BSc/MSc/PGCE Mathematics
cieigcsesolutions@gmail.com

Leave blank

1. A price comparison website publishes data on the cost per month, £c, and the level of satisfaction, s, of a random sample of six internet service providers. A low value of s corresponds to a low level of satisfaction. The data are given in the table below.

Internet service provider	*A*	*B*	*C*	*D*	*E*	*F*
c	20	15	12	30	9	25
s	5	3	4	2	3	4

(You may use $\sum c = 111$, $\sum c^2 = 2375$, $\sum s = 21$, $\sum s^2 = 79$, $\sum cs = 380$, $S_{cc} = 321.5$)

(a) Calculate the value of S_{cs} and the value of S_{ss}

(3)

(b) Calculate the product moment correlation coefficient for these data.

(2)

Brad is not satisfied with his current internet service and decides to change his provider. He decides to pay a lot more for his new internet service.

(c) On the basis of your calculation in part (b), comment on Brad's decision. Give a reason for your answer.

(2)

Leave blank

Question 1 continued

Q1

(Total 7 marks)

Leave blank

3. Jean works for an insurance company. She randomly selects 8 people and records the price of their car insurance, £p, and the time, t years, since they passed their driving test. The data is shown in the table below.

t	10	13	17	18	22	24	25	27
p	720	650	430	490	500	390	280	300

(You may use $\bar{t} = 19.5$, $\bar{p} = 470$, $S_{tp} = -6080$, $S_{tt} = 254$, $S_{pp} = 169\,200$)

(a) On the graph below draw a scatter diagram for these data.

(2)

(b) Comment on the relationship between p and t.

(1)

(c) Find the equation of the regression line of p on t.

(4)

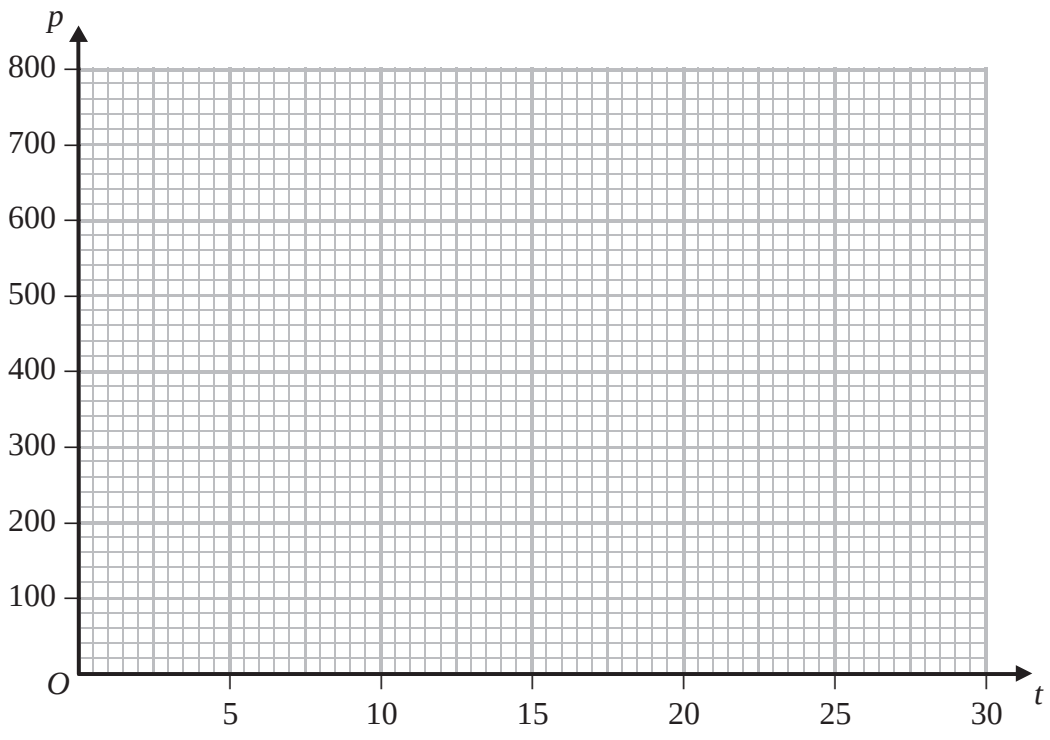
(d) Use your regression equation to estimate the price of car insurance for someone who passed their driving test 20 years ago.

(2)

Jack passed his test 39 years ago and decides to use Jean's data to predict the price of his car insurance.

(e) Comment on Jack's decision. Give a reason for your answer.

(2)



Leave blank

Question 3 continued

Leave blank

3. The table shows the price of a bottle of milk, m pence, and the price of a loaf of bread, b pence, for 8 different years.

m	29	29	35	39	41	43	44	46
b	75	83	91	121	120	126	119	126

(You may use $S_{bb} = 3083.875$ and $S_{mm} = 305.5$)

(a) Find the exact value of $\sum bm$

(1)

(b) Find S_{bm}

(3)

(c) Calculate the product moment correlation coefficient between b and m

(2)

(d) Interpret the value of the correlation coefficient.

(1)

A ninth year is added to the data set. In this year the price of the bottle of milk is 46 pence and the price of a loaf of bread is 175 pence.

(e) Without further calculation, state whether the value of the product moment correlation coefficient will increase, decrease or stay the same when all nine years are used. Give a reason for your answer.

(2)

Leave blank

Question 3 continued

Q3

(Total 9 marks)

Leave blank

5. The resting heart rate, h beats per minute (bpm), and average length of daily exercise, t minutes, of a random sample of 8 teachers are shown in the table below.

t	20	35	40	25	45	70	75	90
h	88	85	77	75	71	66	60	54

(a) State, with a reason, which variable is the response variable.

(2)

The equation of the least squares regression line of h on t is

$$h = 93.5 - 0.43t$$

(b) Give an interpretation of the gradient of this regression line.

(1)

(c) Find the value of $\bar{t}$ and the value of $\bar{h}$

(2)

(d) Show that the point $(\bar{t}, \bar{h})$ lies on the regression line.

(1)

(e) Estimate the resting heart rate of a teacher with an average length of daily exercise of 1 hour.

(1)

(f) Comment, giving a reason, on the reliability of the estimate in part (e).

(2)

The resting heart rate of teachers is assumed to be normally distributed with mean 73 bpm and standard deviation 8 bpm.

The middle 95% of resting heart rates of teachers lies between a and b

(g) Find the value of a and the value of b.

(4)

Leave blank

Question 5 continued

(Total 13 marks)

Q5

Leave blank

3. A publisher collects information about the amount spent on advertising, £ x, and the sales, y books, for some of her publications. She collects information for a random sample of 8 textbooks and codes the data using $v = \dfrac{x + 50}{200}$ and $s = \dfrac{y}{1000}$ to give

v	0.60	8.10	4.30	0.40	1.60	6.40	2.50	5.10
s	1.84	6.73	5.95	1.30	2.45	7.46	4.82	6.25

[You may use: $\sum v = 29$ $\sum s = 36.8$ $\sum s^2 = 209.72$ $\sum vs = 177.311$ $S_{vv} = 55.275$]

(a) Find S_{vs} and S_{ss}

(3)

(b) Calculate the product moment correlation coefficient for these data.

(2)

The publisher believes that a linear regression model may be appropriate to describe these data.

(c) State, giving a reason, whether or not your answer to part (b) supports the publisher's belief.

(1)

(d) Find the equation of the regression line of s on v, giving your answer in the form $s = a + bv$

(4)

(e) Hence find the equation of the regression line of y on x for the sample of textbooks, giving your answer in the form $y = c + dx$

(3)

The publisher calculated the regression line for a sample of novels and obtained the equation

$$y = 3100 + 1.2x$$

She wants to increase the sales of books by spending more money on advertising.

(f) State, giving your reasons, whether the publisher should spend more money on advertising textbooks or novels.

(2)

Leave blank

Question 3 continued

Leave blank

3. A scientist measured the salinity of water, x g/kg, and recorded the temperature at which the water froze, y °C, for 12 different water samples. The summary statistics are listed below.

$$\sum x = 504 \qquad \sum y = -27 \qquad \sum x^2 = 22842 \qquad \sum y^2 = 62.98$$

$$\sum xy = -1190.7 \qquad S_{xx} = 1674 \qquad S_{yy} = 2.23$$

(a) Find the mean and variance of the recorded temperatures.

(3)

Priya believes that the higher the salinity of water, the higher the temperature at which the water freezes.

(b) (i) Calculate the product moment correlation coefficient between x and y

(ii) State, with a reason, whether or not this value supports Priya's belief.

(4)

(c) Find the least squares regression line of y on x in the form $y = a + bx$
Give the value of a and the value of b to 3 significant figures.

(4)

(d) Estimate the temperature at which water freezes when the salinity is 32 g/kg

(1)

The coding $w = 1.8y + 32$ is used to convert the recorded temperatures from °C to °F

(e) Find an equation of the least squares regression line of w on x in the form $w = c + dx$

(2)

(f) Find

(i) the variance of the recorded temperatures when converted to °F

(ii) the product moment correlation coefficient between w and x

(3)

Leave blank

Question 3 continued

Leave blank

3. Martin is investigating the relationship between a person's daily caffeine consumption, c milligrams, and the amount of sleep they get, h hours, per night. He collected this information from 20 people and the results are summarised below.

$$\sum c = 3660 \qquad \sum h = 126 \qquad \sum c^2 = 973\,228$$

$$\sum ch = 20\,023.4 \qquad S_{cc} = 303\,448 \qquad S_{ch} = -3034.6$$

Martin calculates the product moment correlation coefficient for these data and obtains –0.833

(a) Give a reason why this value supports a linear relationship between c and h

(1)

The amount of sleep per night is the response variable.

(b) Explain what you understand by the term 'response variable'.

(1)

Martin says that for each additional 100 mg of caffeine consumed, the expected number of hours of sleep decreases by 1

(c) Determine, by calculation, whether or not the data support this statement.

(3)

(d) Use the data to calculate an estimate for the expected number of hours of sleep per night when no caffeine is consumed.

(3)

Question 3 continued

Leave blank

Leave blank

5. Franca is the manager of an accountancy firm. She is investigating the relationship between the salary, £x, and the length of commute, y minutes, for employees at the firm. She collected this information from 9 randomly selected employees.

The salary of each employee was then coded using $w = \dfrac{x - 20\,000}{1000}$

The table shows the values of w and y for the 9 employees.

w	6	8	8	–1	25	15	3	–2	19
y	45	50	35	65	25	40	50	75	20

(You may use $\sum w = 81$ $\sum y = 405$ $\sum wy = 2490$ $S_{ww} = 660$ $S_{yy} = 2500$)

(a) Calculate the salary of the employee with $w = -2$

(1)

(b) Show that, to 3 significant figures, the value of the product moment correlation coefficient between w and y is –0.899

(3)

(c) State, giving a reason, the value of the product moment correlation coefficient between x and y

(1)

The least squares regression line of y on w is $y = 60.75 - 1.75w$

(d) Find the equation of the least squares regression line of y on x giving your answer in the form $y = a + bx$

(3)

(e) Estimate the length of commute for an employee with a salary of £21 000

(2)

Franca uses the regression line to estimate the length of commute for employees with salaries between £25 000 and £40 000

(f) State, giving a reason, whether or not these estimates are reliable.

(2)

Question 5 continued

Leave blank

Leave blank

1. A medical researcher is studying the relationship between age (x years) and volume of blood (y ml) pumped by each contraction of the heart. The researcher obtained the following data from a random sample of 8 patients.

Age (x)	20	25	30	45	55	60	65	70
Volume (y)	74	76	77	72	68	67	64	62

[You may use $\sum x = 370$, $S_{xx} = 2587.5$, $\sum y = 560$, $\sum y^2 = 39\,418$, $S_{xy} = -710$]

(a) Calculate S_{yy}

(2)

(b) Calculate the product moment correlation coefficient for these data.

(2)

(c) Interpret your value of the correlation coefficient.

(1)

The researcher believes that a linear regression model may be appropriate to describe these data.

(d) State, giving a reason, whether or not your value of the correlation coefficient supports the researcher's belief.

(1)

(e) Find the equation of the regression line of y on x, giving your answer in the form $y = a + bx$

(4)

Jack is a 40-year-old patient.

(f) (i) Use your regression line to estimate the volume of blood pumped by each contraction of Jack's heart.

(ii) Comment, giving a reason, on the reliability of your estimate.

(2)

Leave blank

Question 1 continued

Leave blank

2. Paul believes there is a relationship between the value and the floor size of a house. He takes a random sample of 20 houses and records the value, £v, and the floor size, s m^2

The data were coded using $x = \dfrac{s - 50}{10}$ and $y = \dfrac{v}{100\,000}$ and the following statistics obtained.

$\sum x = 441.5, \quad \sum y = 59.8, \quad \sum x^2 = 11\,261.25, \quad \sum y^2 = 196.66, \quad \sum xy = 1474.1$

(a) Find the value of S_{xy} and the value of S_{xx}

(3)

(b) Find the equation of the least squares regression line of y on x in the form $y = a + bx$

(3)

The least squares regression line of v on s is $v = c + ds$

(c) Show that $d = 1020$ to 3 significant figures and find the value of c

(3)

(d) Estimate the value of a house of floor size 130 m^2

(2)

(e) Interpret the value d

(1)

Paul wants to increase the value of his house. He decides to add an extension to increase the floor size by 31 m^2

(f) Estimate the increase in the value of Paul's house after adding the extension.

(1)

Leave blank

Question 2 continued

Leave blank

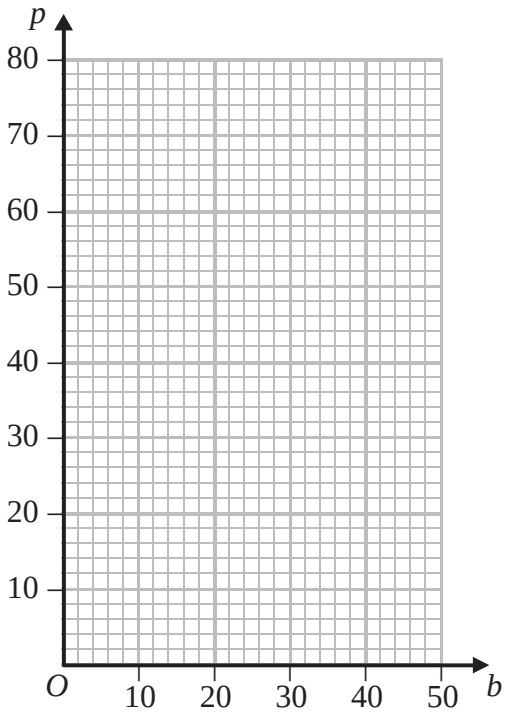
7. A doctor is investigating the correlation between blood protein, p, and body mass index, b.

He takes a random sample of 8 patients and the data are shown in the table below.

Patient	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
b	32	36	40	44	42	21	27	37
p	18	21	31	39	21	12	19	70

(a) Draw a scatter diagram of these data on the axes provided.

(2)



The doctor decides to leave out patient H from his calculations.

(b) Give a reason for the doctor's decision.

(1)

For the 7 patients A, B, C, D, E, F and G,

$$S_{bp} = 369, \quad S_{pp} = 490 \quad \text{and} \quad S_{bb} = 423\frac{5}{7}$$

(c) Find the product moment correlation coefficient, r, for these 7 patients.

(2)

(d) Without any further calculations, state how r would differ from your answer in part (c) if it was calculated for all 8 patients.

(1)

Leave blank

Question 7 continued

Q7

(Total 6 marks)

Leave blank

1. The percentage oil content, p, and the weight, w milligrams, of each of 10 randomly selected sunflower seeds were recorded. These data are summarised below.

$\sum w^2 = 41252$ $\sum wp = 27557.8$ $\sum w = 640$ $\sum p = 431$ $S_{pp} = 2.72$

(a) Find the value of S_{ww} and the value of S_{wp}

(3)

(b) Calculate the product moment correlation coefficient between p and w

(2)

(c) Give an interpretation of your product moment correlation coefficient.

(1)

The equation of the regression line of p on w is given in the form $p = a + bw$

(d) Find the equation of the regression line of p on w

(4)

(e) Hence estimate the percentage oil content of a sunflower seed which weighs 60 milligrams.

(2)

Question 1 continued

Leave blank

Leave blank

5. Tomas is studying the relationship between temperature and hours of sunshine in *Seapron*. He records the midday temperature, t °C, and the hours of sunshine, s hours, for a random sample of 9 days in October. He calculated the following statistics

$$\sum s = 15 \qquad \sum s^2 = 44.22 \qquad \sum t = 127 \qquad S_{tt} = 10.89$$

(a) Calculate S_{ss}

(2)

Tomas calculated the product moment correlation coefficient between s and t to be 0.832 correct to 3 decimal places.

(b) State, giving a reason, whether or not this correlation coefficient supports the use of a linear regression model to describe the relationship between midday temperature and hours of sunshine.

(1)

(c) State, giving a reason, why the hours of sunshine would be the explanatory variable in a linear regression model between midday temperature and hours of sunshine.

(1)

(d) Find S_{st}

(3)

(e) Calculate a suitable linear regression equation to model the relationship between midday temperature and hours of sunshine.

(4)

(f) Calculate the standard deviation of s

(1)

Tomas uses this model to estimate the midday temperature in *Seapron* for a day in October with 5 hours of sunshine.

(g) State the value of Tomas' estimate.

(1)

Given that the values of s are all within 2 standard deviations of the mean,

(h) comment, giving your reason, on the reliability of this estimate.

(2)

Leave blank

Question 5 continued

Leave blank

Question 5 continued

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Leave blank

1. A random sample of 10 cars of different makes and sizes is taken and the published miles per gallon, p, and the actual miles per gallon, m, are recorded. The data are coded using variables $x = \frac{p}{10}$ and $y = m - 25$

The results for the coded data are summarised below.

x	6.89	3.67	5.92	5.04	4.87	3.92	4.71	5.14	3.65	5.23
y	30	3	22	15	13	8	15	13.5	3	19

(You may use $\sum y^2 = 2628.25$ $\sum xy = 768.58$ $S_{xx} = 9.25924$ $S_{xy} = 74.664$)

(a) Show that $S_{yy} = 626.025$

(2)

(b) Find the product moment correlation coefficient between x and y.

(2)

(c) Give a reason to support fitting a regression model of the form $y = a + bx$ to these data.

(1)

(d) Find the equation of the regression line of y on x, giving your answer in the form $y = a + bx$.
Give the value of a and the value of b to 3 significant figures.

(3)

A car's published miles per gallon is 44

(e) Estimate the actual miles per gallon for this particular car.

(3)

(f) Comment on the reliability of your estimate in part (e). Give a reason for your answer.

(2)

Question 1 continued

Leave blank

Leave blank

4. A doctor is studying the scans of 30-week old foetuses. She takes a random sample of 8 scans and measures the length, f mm, of the leg bone called the femur. She obtains the following results.

52 53 56 57 57 59 60 62

(a) Show that $S_{ff} = 80$

(3)

The doctor also measures the head circumference, h mm, of each foetus and her results are summarised as

$$\sum h = 2209 \qquad \sum h^2 = 610\,463 \qquad S_{fh} = 182$$

(b) Find S_{hh}

(2)

(c) Calculate the product moment correlation coefficient between the length of the femur and the head circumference for these data.

(2)

The doctor believes that there is a linear relationship between the length of the femur and the head circumference of 30-week old foetuses.

(d) State, giving a reason, whether or not your calculation in part (c) supports the doctor's belief.

(1)

(e) Find an equation of the regression line of h on f.

(4)

The doctor plans in future to measure the femur length, f, and then use the regression line to estimate the corresponding head circumference, h.

A statistician points out that there will always be the chance of an error between the true head circumference and the estimated value of the head circumference.

Given that the error, E mm, has the normal distribution $N(0, 4^2)$

(f) find the probability that the estimate is within 3 mm of the true value.

(3)

Question 4 continued

Leave blank

Leave blank

Question 4 continued

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Leave blank

5. A company wants to pay its employees according to their performance at work. Last year's performance score x and annual salary y, in thousands of dollars, were recorded for a random sample of 10 employees of the company.

The performance scores were

15 24 32 39 41 18 16 22 34 42

(You may use $\sum x^2 = 9011$)

(a) Find the mean and the variance of these performance scores.

(3)

The corresponding y values for these 10 employees are summarised by

$$\sum y = 306.1 \quad \text{and} \quad S_{yy} = 546.3$$

(b) Find the mean and the variance of these y values.

(2)

The regression line of y on x based on this sample is

$$y = 12.0 + 0.659x$$

(c) Find the product moment correlation coefficient for these data.

(5)

(d) State, giving a reason, whether or not the value of the product moment correlation coefficient supports the use of a regression line to model the relationship between performance score and annual salary.

(1)

The company decides to use this regression model to determine future salaries.

(e) Find the proposed annual salary, in dollars, for an employee who has a performance score of 35

(2)

Question 5 continued

Leave blank

Leave blank

1. The heights above sea level (h hundred metres) and the temperatures (t °C) at 12 randomly selected places in France, at 7 am on July 31st, were recorded.
The data are summarised as follows

$$\sum h = 112 \quad \sum t = 136 \quad \sum t^2 = 1828 \quad S_{ht} = -236 \quad S_{hh} = 297$$

(a) Find the value of S_{tt}

(2)

(b) Calculate the product moment correlation coefficient for these data.

(2)

(c) Interpret the relationship between t and h.

(1)

(d) Find an equation of the regression line of t on h.

(3)

At 7 am on July 31st Yinka is on holiday in South Africa. He uses the regression equation to estimate the temperature when the height above sea level is 500 m.

(e) Find the estimated temperature Yinka calculates.

(2)

(f) Comment on the validity of your answer in part (e).

(1)

Question 1 continued

Leave blank