

Edexcel GCE

Core Mathematics S1

Correlation and Regression

Materials required for examination
Mathematical Formulae (Green)

Items included with question papers
Nil

Advice to Candidates

You must ensure that your answers to parts of questions are clearly labelled.
You must show sufficient working to make your methods clear to the Examiner. Answers without working may gain no credit.

Jan 2006 Q3

11. A manufacturer stores drums of chemicals. During storage, evaporation takes place. A random sample of 10 drums was taken and the time in storage, x weeks, and the evaporation loss, y ml, are shown in the table below.

x	3	5	6	8	10	12	13	15	16	18
y	36	50	53	61	69	79	82	90	88	96

- (a) On graph paper, draw a scatter diagram to represent these data. (3)
- (b) Give a reason to support fitting a regression model of the form $y = a + bx$ to these data. (1)
- (c) Find, to 2 decimal places, the value of a and the value of b . (7)
- (You may use $\Sigma x^2 = 1352$, $\Sigma y^2 = 53112$ and $\Sigma xy = 8354$.)
- (d) Give an interpretation of the value of b . (1)
- (e) Using your model, predict the amount of evaporation that would take place after
- (i) 19 weeks,
- (ii) 35 weeks. (2)
- (f) Comment, with a reason, on the reliability of each of your predictions. (4)

a) DRAWN.

b) points close to a straight line

$$c) b = \frac{891.6}{228.4} = 3.903677... = \underline{\underline{3.90}}$$

$$a = \frac{704}{10} - b \left(\frac{106}{10} \right) = 29.021015 = \underline{\underline{29.02}}$$

d) For every extra week 3.90ml of chemical evaporates

e i) 103.12

ii) 165.52

f i \rightarrow close to range of x (reasonably reliable)

ii \rightarrow outside of range of x (could be unreliable)

May 2006 Q3

1. A metallurgist measured the length, l mm, of a copper rod at various temperatures, t °C, and recorded the following results.

t	l
20.4	2461.12
27.3	2461.41
32.1	2461.73
39.0	2461.88
42.9	2462.03
49.7	2462.37
58.3	2462.69
67.4	2463.05

The results were then coded such that $x = t$ and $y = l - 2460.00$.

- (a) Calculate S_{xy} and S_{xx} .

(You may use $\Sigma x^2 = 15965.01$ and $\Sigma xy = 757.467$)

(5)

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

(5)

- (c) Estimate the length of the rod at 40 °C.

(3)

- (d) Find the equation of the regression line of l on t .

(2)

- (e) Estimate the length of the rod at 90 °C.

(1)

- (f) Comment on the reliability of your estimate in part (e).

(2)

a) $S_{xy} = 71.4685$
 $S_{xx} = 1760.45875$

b) $b = 0.04059652$
 $a = 0.324364$
 $y = 0.324 + 0.0406x$

c) $x = 40$ $y = 1.948$ $l = \underline{\underline{2461.948}}$

d) $l - 2460 = 0.324 + 0.0406t$
 $l = 2460.324 + 0.0406t$

e) $t = 90$ $l = 2463.978$

f) outside of range. ~~could be~~ unreliable likely to be.

Jan 2007 Q1

2. As part of a statistics project, Gill collected data relating to the length of time, to the nearest minute, spent by shoppers in a supermarket and the amount of money they spent. Her data for a random sample of 10 shoppers are summarised in the table below, where t represents time and m the amount spent over £20.

t (minutes)	m
15	-3
23	17
5	-19
16	4
30	12
6	-9
32	27
23	6
35	20
27	6

- (a) Write down the actual amount spent by the shopper who was in the supermarket for 15 minutes. (1)

- (b) Calculate S_{tt} , S_{mm} and S_{tm} .

(You may use $\Sigma t^2 = 5478$, $\Sigma m^2 = 2101$, and $\Sigma tm = 2485$)

- (c) Calculate the value of the product moment correlation coefficient between t and m . (3)
- (d) Write down the value of the product moment correlation coefficient between t and the actual amount spent. Give a reason to justify your value. (2)

On another day Gill collected similar data. For these data the product moment correlation coefficient was 0.178.

- (e) Give an interpretation to both of these coefficients. (2)
- (f) Suggest a practical reason why these two values are so different. (1)

a) £17

b) $S_{tt} = 983.6$
 $S_{tm} = 1191.8$ $S_{mm} = 1728.9$

c) $r = 0.913922...$

d) $0.913922...$ Coding does not affect PMCC

e) $0.914 \rightarrow$ longer spent shopping more money spent

$0.178 \rightarrow$ not people spent different amounts for same time

f) short time spent buying single expensive item/ lots of cheap items taking a long time

3. A young family were looking for a new 3 bedroom semi-detached house. A local survey recorded the price x , in £1000, and the distance y , in miles, from the station of such houses. The following summary statistics were provided

$$S_{xx} = 113\,573, S_{yy} = 8.657, S_{xy} = -808.917$$

- (a) Use these values to calculate the product moment correlation coefficient.

(2)

- (b) Give an interpretation of your answer to part (a).

(1)

Another family asked for the distances to be measured in km rather than miles.

- (c) State the value of the product moment correlation coefficient in this case.

(1)

$$a/ \quad r = \frac{-808.917}{\sqrt{113573 \times 8.657}} = -0.81579$$

b/ Houses are cheaper further away from the station

$$c/ \quad -0.816$$

June 2007 Q3

4. A student is investigating the relationship between the price (y pence) of 100g of chocolate and the percentage ($x\%$) of the cocoa solids in the chocolate.
The following data is obtained

Chocolate brand	A	B	C	D	E	F	G	H
x (% cocoa)	10	20	30	35	40	50	60	70
y (pence)	35	55	40	100	60	90	110	130

(You may use: $\sum x = 315$, $\sum x^2 = 15\,225$, $\sum y = 620$, $\sum y^2 = 56\,550$, $\sum xy = 28\,750$)

- (a) Draw a scatter diagram to represent these data.

DRAWN

(2)

- (b) Show that $S_{xy} = 4337.5$ and find S_{xx} .

(3)

The student believes that a linear relationship of the form $y = a + bx$ could be used to describe these data.

- (c) Use linear regression to find the value of a and the value of b , giving your answers to 1 decimal place.

(4)

- (d) Draw the regression line on your diagram.

(2)

The student believes that one brand of chocolate is overpriced.

- (e) Use the scatter diagram to

(i) state which brand is overpriced,

(ii) suggest a fair price for this brand.

Give reasons for both your answers.

(4)

a/ DRAWN

$$b) S_{xy} = 28750 - \frac{315 \times 620}{8} = \underline{\underline{4337.5}}$$

$$S_{xx} = 2821.875$$

$$c/ b = 1.537$$

$$a = 16.97 \dots$$

d/ DRAWN

e/ BRAND D (above the line)

$$ii/ y = 17 + 35(1.5) \\ = 69.5$$

Jan 2008 Q1

5. A personnel manager wants to find out if a test carried out during an employee's interview and a skills assessment at the end of basic training is a guide to performance after working for the company for one year.

The table below shows the results of the interview test of 10 employees and their performance after one year.

Employee	A	B	C	D	E	F	G	H	I	J
Interview test, $x\%$	65	71	79	77	85	78	85	90	81	62
Performance after one year, $y\%$	65	74	82	64	87	78	61	65	79	69

[You may use $\sum x^2 = 60\,475$, $\sum y^2 = 53\,122$, $\sum xy = 56\,076$]

- (a) Showing your working clearly, calculate the product moment correlation coefficient between the interview test and the performance after one year.

(5)

The product moment correlation coefficient between the skills assessment and the performance after one year is -0.156 to 3 significant figures.

- (b) Use your answer to part (a) to comment on whether or not the interview test and skills assessment are a guide to the performance after one year. Give clear reasons for your answers.

(2)

a/ $\sum x = 773$ $\sum y = 724$

S_{xy} S_{xx} S_{yy}

$r = 0.155357$

b/ Both weak correlation

Interview test is slightly better as it has a positive correlation

Jan 2008 Q4

6. A second hand car dealer has 10 cars for sale. She decides to investigate the link between the age of the cars, x years, and the mileage, y thousand miles. The data collected from the cars are shown in the table below.

Age, x (years)	2	2.5	3	4	4.5	4.5	5	3	6	6.5
Mileage, y (thousands)	22	34	33	37	40	45	49	30	58	58

[You may assume that $\sum x = 41$, $\sum y = 406$, $\sum x^2 = 188$, $\sum xy = 1818.5$]

- (a) Find S_{xx} and S_{xy} . (3)
- (b) Find the equation of the least squares regression line in the form $y = a + bx$. Give the values of a and b to 2 decimal places. (4)
- (c) Give a practical interpretation of the slope b . (1)
- (d) Using your answer to part (b), find the mileage predicted by the regression line for a 5 year old car. (2)

a/ $S_{xy} = 153.9$

$S_{xx} = 19.9$

b/ $b = \frac{153.9}{19.9} = 7.733668...$

$a = 40.6 - b(4.1) = 8.89796...$

$y = 8.89 + 7.73x$

c/ average car travels 7700 miles per year

d/ $x = 5$ $y = 8.89 + 7.73(5) = 47.5 - 47.6$

mileage predicted 47500 - 47600

May 2008 Q4

7. Crickets make a noise. The pitch, v kHz, of the noise made by a cricket was recorded at 15 different temperatures, t °C. These data are summarised below.

$$\sum t^2 = 10\,922.81, \quad \sum v^2 = 42.3356, \quad \sum tv = 677.971, \quad \sum t = 401.3, \quad \sum v = 25.08$$

- (a) Find S_{tt} , S_{vv} and S_{tv} for these data. (4)
- (b) Find the product moment correlation coefficient between t and v . (3)
- (c) State, with a reason, which variable is the explanatory variable. (2)
- (d) Give a reason to support fitting a regression model of the form $v = a + bt$ to these data. (1)
- (e) Find the value of a and the value of b . Give your answers to 3 significant figures. (4)
- (f) Using this model, predict the pitch of the noise at 19 °C. (1)

a/ $S_{tt} = 186.6973$
 $S_{vv} = 0.40184$
 $S_{tv} = 6.9974$

b/ $r = 0.807867$

c/ t is the explanatory variable
temperature affects noise / temp. can be controlled

d/ r is close to 1

e/ $b = \frac{6.9974}{186.6973} = 0.03748$

$$a = \frac{25.08}{15} - b \left(\frac{401.3}{15} \right) = 0.6692874$$

f/ $t = 19 \quad v = 0.6692874 + 0.03748 \times 19$
 $= \underline{1.381406}$

Jan 2009 Q1

8. A teacher is monitoring the progress of students using a computer based revision course. The improvement in performance, y marks, is recorded for each student along with the time, x hours, that the student spent using the revision course. The results for a random sample of 10 students are recorded below.

x hours	1.0	3.5	4.0	1.5	1.3	0.5	1.8	2.5	2.3	3.0
y marks	5	30	27	10	-3	-5	7	15	-10	20

[You may use $\sum x = 21.4$, $\sum y = 96$, $\sum x^2 = 57.22$, $\sum xy = 313.7$]

(a) Calculate S_{xx} and S_{xy} .

(3)

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

(4)

(c) Give an interpretation of the gradient of your regression line.

(1)

Rosemary spends 3.3 hours using the revision course.

(d) Predict her improvement in marks.

(2)

Lee spends 8 hours using the revision course claiming that this should give him an improvement in performance of over 60 marks.

(e) Comment on Lee's claim.

(1)

a/ $S_{xx} = 11.424$
 $S_{xy} = 108.26$

b/ $b = 9.4765...$
 $a = -10.679...$

$y = -10.7 + 9.48x$

c/ every extra hour spent using the programme produce 9.5 mark of improvement

d/ $y = -10.7 + 9.48 \times 3.3 = \underline{20.6}$

e/ Model is not valid 8hrs is outside of the range

May 2009 Q1

9. The volume of a sample of gas is kept constant. The gas is heated and the pressure, p , is measured at 10 different temperatures, t . The results are summarised below.

$$\Sigma p = 445 \quad \Sigma p^2 = 38\,125 \quad \Sigma t = 240 \quad \Sigma t^2 = 27\,520 \quad \Sigma pt = 26\,830$$

(a) Find S_{pp} and S_{pt} .

(3)

Given that $S_n = 21\,760$,

(b) calculate the product moment correlation coefficient.

(2)

(c) Give an interpretation of your answer to part (b).

(1)

a/ $S_{pp} = 18322.5$

$$S_{pt} = 16150$$

b/ $r = 0.8088\dots$

c/ As temperature increases price increases

May 2009 Q5

10. The weight, w grams, and the length, l mm, of 10 randomly selected newborn turtles are given in the table below.

l	49.0	52.0	53.0	54.5	54.1	53.4	50.0	51.6	49.5	51.2
w	29	32	34	39	38	35	30	31	29	30

(You may use $S_l = 33.381$ $S_w = 59.99$ $S_{wl} = 120.1$)

- (a) Find the equation of the regression line of w on l in the form $w = a + bl$. (5)
- (b) Use your regression line to estimate the weight of a newborn turtle of length 60 mm. (2)
- (c) Comment on the reliability of your estimate giving a reason for your answer. (2)

$$a) \quad b = \frac{59.99}{33.381} = 1.79713$$

$$a = 32.7 - 1.79713 \dots \times 51.83 \\ = -60.44525$$

$$w = -60.445251 \dots + 1.79713 \dots l$$

$$b) \quad w = "-60.445251" + "1.79713" \times 60 \\ = \cancel{40} 47.3825 \dots$$

c) extrapolating (outside of range) so (may be) unreliable