Edexcel (U.K.) Pre 2017

Questions By Topic

S1 Chap06–7 Correlation and Regression

Compiled By: Dr Yu

Editors: Betül, Signal, Vivian

www.CasperYC.club

Last updated: February 7, 2026

DrYuFromShanghai@QQ.com

Leave blank

1. The scatter diagrams below were drawn by a student.

Diagram *A*

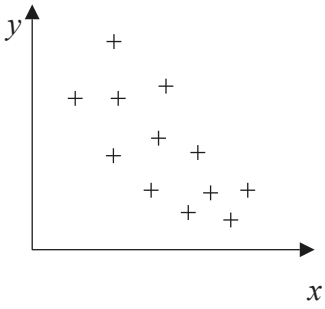


Diagram *B*

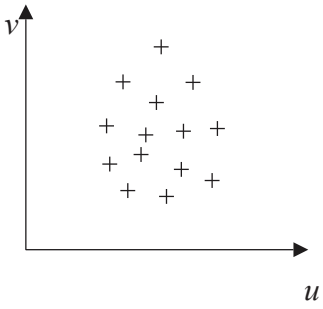
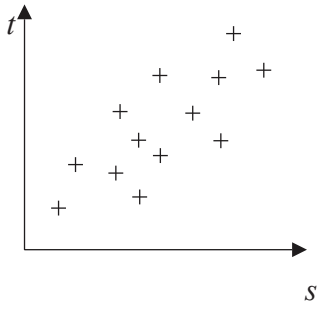


Diagram *C*



The student calculated the value of the product moment correlation coefficient for each of the sets of data.

The values were

$$0.68 \qquad -0.79 \qquad 0.08$$

Write down, with a reason, which value corresponds to which scatter diagram.

(6)

Leave blank

3. A long distance lorry driver recorded the distance travelled, m miles, and the amount of fuel used, f litres, each day. Summarised below are data from the driver's records for a random sample of 8 days.

The data are coded such that $x = m - 250$ and $y = f - 100$.

$$\Sigma x = 130 \quad \Sigma y = 48 \quad \Sigma xy = 8880 \quad S_{xx} = 20\,487.5$$

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

(6)

(b) Hence find the equation of the regression line of f on m.

(3)

(c) Predict the amount of fuel used on a journey of 235 miles.

(1)

Leave blank

3. 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)

Leave blank

1. 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 $\Sigma t^2 = 5478 \quad \Sigma m^2 = 2101 \quad \Sigma tm = 2485$)

(6)

(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)

Leave blank

1. 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

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

(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)

Leave blank

3. A student is investigating the relationship between the price (y pence) of 100g of chocolate and the percentage (x%) of 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) On the graph paper on page 9 draw a scatter diagram to represent these data.

(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 scatter 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)

Leave blank

1. 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)

Leave blank

4. 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)

Leave blank

4. 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.

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

(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)

Leave blank

1. 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)

Leave blank

1. 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.

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

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

(3)

Given that $S_{tt} = 21\,760$,

(b) calculate the product moment correlation coefficient.

(2)

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

(1)

Leave blank

5. 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 $S_{ll} = 33.381$ $S_{wl} = 59.99$ $S_{ww} = 120.1$)

(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)

Leave blank

6. The blood pressures, p mmHg, and the ages, t years, of 7 hospital patients are shown in the table below.

Patient	A	B	C	D	E	F	G
t	42	74	48	35	56	26	60
p	98	130	120	88	182	80	135

$$[\sum t = 341, \sum p = 833, \sum t^2 = 18\,181, \sum p^2 = 106\,397, \sum tp = 42\,948]$$

(a) Find S_{pp}, S_{tp} and S_{tt} for these data.

(4)

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

(3)

(c) Interpret the correlation coefficient.

(1)

(d) On the graph paper on page 17, draw the scatter diagram of blood pressure against age for these 7 patients.

(2)

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

(4)

(f) Plot your regression line on your scatter diagram.

(2)

(g) Use your regression line to estimate the blood pressure of a 40 year old patient.

(2)

Leave blank

1. Gary compared the total attendance, x, at home matches and the total number of goals, y, scored at home during a season for each of 12 football teams playing in a league. He correctly calculated:

$$S_{xx} = 1022\,500 \quad S_{yy} = 130.9 \quad S_{xy} = 8825$$

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

(2)

(b) Interpret the value of the correlation coefficient.

(1)

Helen was given the same data to analyse. In view of the large numbers involved she decided to divide the attendance figures by 100. She then calculated the product moment correlation coefficient between $\frac{x}{100}$ and y.

(c) Write down the value Helen should have obtained.

(1)

Leave blank

6. A travel agent sells flights to different destinations from *Beerow* airport. The distance d, measured in 100 km, of the destination from the airport and the fare £f are recorded for a random sample of 6 destinations.

Destination	*A*	*B*	*C*	*D*	*E*	*F*
d	2.2	4.0	6.0	2.5	8.0	5.0
f	18	20	25	23	32	28

[You may use $\sum d^2 = 152.09$ $\sum f^2 = 3686$ $\sum fd = 723.1$]

(a) Using the axes below, complete a scatter diagram to illustrate this information.

(2)

(b) Explain why a linear regression model may be appropriate to describe the relationship between f and d.

(1)

(c) Calculate S_{dd} and S_{fd}

(4)

(d) Calculate the equation of the regression line of f on d giving your answer in the form $f = a + bd$.

(4)

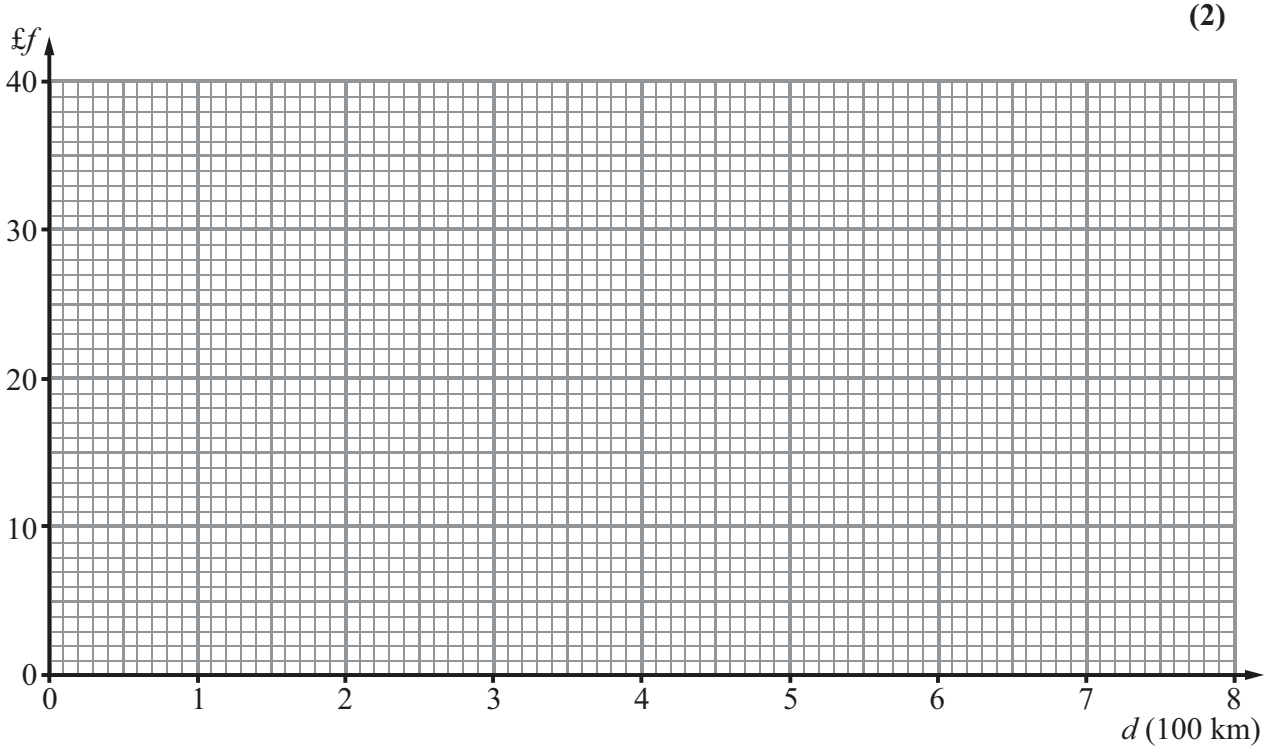
(e) Give an interpretation of the value of b.

(1)

Jane is planning her holiday and wishes to fly from *Beerow* airport to a destination t km away. A rival travel agent charges 5p per km.

(f) Find the range of values of t for which the first travel agent is cheaper than the rival.

(2)



Leave blank

1. A random sample of 50 salmon was caught by a scientist. He recorded the length l cm and weight w kg of each salmon.

The following summary statistics were calculated from these data.

$\sum l = 4027$ $\sum l^2 = 327\,754.5$ $\sum w = 357.1$ $\sum lw = 29\,330.5$ $S_{ww} = 289.6$

(a) Find S_{ll} and S_{lw}

(3)

(b) Calculate, to 3 significant figures, the product moment correlation coefficient between l and w.

(2)

(c) Give an interpretation of your coefficient.

(1)

Leave blank

4. A farmer collected data on the annual rainfall, x cm, and the annual yield of peas, p tonnes per acre.

The data for annual rainfall was coded using $v = \frac{x-5}{10}$ and the following statistics were found.

$$S_{vv} = 5.753 \qquad S_{pv} = 1.688 \qquad S_{pp} = 1.168 \qquad \overline{p} = 3.22 \qquad \overline{v} = 4.42$$

(a) Find the equation of the regression line of p on v in the form $p = a + bv$.

(4)

(b) Using your regression line estimate the annual yield of peas per acre when the annual rainfall is 85 cm.

(2)

Leave blank

1. On a particular day the height above sea level, x metres, and the mid-day temperature, y°C, were recorded in 8 north European towns. These data are summarised below

$$S_{xx} = 3\ 535\ 237.5 \quad \sum y = 181 \quad \sum y^2 = 4305 \quad S_{xy} = -23\ 726.25$$

(a) Find S_{yy}

(2)

(b) Calculate, to 3 significant figures, the product moment correlation coefficient for these data.

(2)

(c) Give an interpretation of your coefficient.

(1)

A student thought that the calculations would be simpler if the height above sea level, h, was measured in kilometres and used the variable $h = \frac{x}{1000}$ instead of x.

(d) Write down the value of S_{hh}

(1)

(e) Write down the value of the correlation coefficient between h and y.

(1)

Leave blank

7. A teacher took a random sample of 8 children from a class. For each child the teacher recorded the length of their left foot, f cm, and their height, h cm. The results are given in the table below.

f	23	26	23	22	27	24	20	21
h	135	144	134	136	140	134	130	132

(You may use $\sum f = 186$ $\sum h = 1085$ $S_{ff} = 39.5$ $S_{hh} = 139.875$ $\sum fh = 25291$)

(a) Calculate S_{fh}

(2)

(b) Find the equation of the regression line of h on f in the form $h = a + bf$.
Give the value of a and the value of b correct to 3 significant figures.

(5)

(c) Use your equation to estimate the height of a child with a left foot length of 25 cm.

(2)

(d) Comment on the reliability of your estimate in (c), giving a reason for your answer.

(2)

The left foot length of the teacher is 25 cm.

(e) Give a reason why the equation in (b) should not be used to estimate the teacher's height.

(1)

Leave blank

5. The age, t years, and weight, w grams, of each of 10 coins were recorded. These data are summarised below.

$$\sum t^2 = 2688 \quad \sum tw = 1760.62 \quad \sum t = 158 \quad \sum w = 111.75 \quad S_{ww} = 0.16$$

(a) Find S_{tt} and S_{tw} for these data.

(3)

(b) Calculate, to 3 significant figures, the product moment correlation coefficient between t and w.

(2)

(c) Find the equation of the regression line of w on t in the form $w = a + bt$

(4)

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

(2)

(e) Using this model, estimate
(i) the weight of a coin which is 5 years old,
(ii) the effect of an increase of 4 years in age on the weight of a coin.

(2)

It was discovered that a coin in the original sample, which was 5 years old and weighed 20 grams, was a fake.

(f) State, without any further calculations, whether the exclusion of this coin would increase or decrease the value of the product moment correlation coefficient. Give a reason for your answer.

(2)

Leave blank

2. A bank reviews its customer records at the end of each month to find out how many customers have become unemployed, u, and how many have had their house repossessed, h, during that month. The bank codes the data using variables $x = \frac{u-100}{3}$ and $y = \frac{h-20}{7}$ The results for the 12 months of 2009 are summarised below.

$$\sum x = 477 \quad S_{xx} = 5606.25 \quad \sum y = 480 \quad S_{yy} = 4244 \quad \sum xy = 23\ 070$$

(a) Calculate the value of the product moment correlation coefficient for x and y.

(3)

(b) Write down the product moment correlation coefficient for u and h.

(1)

The bank claims that an increase in unemployment among its customers is associated with an increase in house repossessions.

(c) State, with a reason, whether or not the bank's claim is supported by these data.

(2)

Leave blank

3. A scientist is researching whether or not birds of prey exposed to pollutants lay eggs with thinner shells. He collects a random sample of egg shells from each of 6 different nests and tests for pollutant level, p, and measures the thinning of the shell, t. The results are shown in the table below.

p	3	8	30	25	15	12
t	1	3	9	10	5	6

[You may use $\sum p^2 = 1967$ and $\sum pt = 694$]

(a) Draw a scatter diagram on the axes on page 7 to represent these data.

(2)

(b) Explain why a linear regression model may be appropriate to describe the relationship between p and t.

(1)

(c) Calculate the value of S_{pt} and the value of S_{pp}.

(4)

(d) Find the equation of the regression line of t on p, giving your answer in the form $t = a + bp$.

(4)

(e) Plot the point $(\overline{p}, \overline{t})$ and draw the regression line on your scatter diagram.

(2)

The scientist reviews similar studies and finds that pollutant levels above 16 are likely to result in the death of a chick soon after hatching.

(f) Estimate the minimum thinning of the shell that is likely to result in the death of a chick.

(2)

Leave blank

3. A biologist is comparing the intervals (m seconds) between the mating calls of a certain species of tree frog and the surrounding temperature (t °C). The following results were obtained.

t °C	8	13	14	15	15	20	25	30
m secs	6.5	4.5	6	5	4	3	2	1

(You may use $\sum tm = 469.5$, $S_{tt} = 354$, $S_{mm} = 25.5$)

(a) Show that $S_{tm} = -90.5$

(4)

(b) Find the equation of the regression line of m on t giving your answer in the form $m = a + bt$.

(4)

(c) Use your regression line to estimate the time interval between mating calls when the surrounding temperature is 10 °C.

(1)

(d) Comment on the reliability of this estimate, giving a reason for your answer.

(1)

Leave blank

1. A meteorologist believes that there is a relationship between the height above sea level, h m, and the air temperature, t °C. Data is collected at the same time from 9 different places on the same mountain. The data is summarised in the table below.

h	1400	1100	260	840	900	550	1230	100	770
t	3	10	20	9	10	13	5	24	16

[You may assume that $\sum h = 7150$, $\sum t = 110$, $\sum h^2 = 7\,171\,500$, $\sum t^2 = 1716$, $\sum th = 64\,980$ and $S_{tt} = 371.56$]

(a) Calculate S_{th} and S_{hh}. Give your answers to 3 significant figures.

(3)

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

(2)

(c) State whether or not your value supports the use of a regression equation to predict the air temperature at different heights on this mountain. Give a reason for your answer.

(1)

(d) Find the equation of the regression line of t on h giving your answer in the form $t = a + bh$.

(4)

(e) Interpret the value of b.

(1)

(f) Estimate the difference in air temperature between a height of 500 m and a height of 1000 m.

(2)

Leave blank

3. An agriculturalist is studying the yields, y kg, from tomato plants. The data from a random sample of 70 tomato plants are summarised below.

Yield (y kg)	**Frequency** (f)	**Yield midpoint** (x kg)
$0 \leqslant y < 5$	16	2.5
$5 \leqslant y < 10$	24	7.5
$10 \leqslant y < 15$	14	12.5
$15 \leqslant y < 25$	12	20
$25 \leqslant y < 35$	4	30

(You may use $\sum fx = 755$ and $\sum fx^2 = 12037.5$)

A histogram has been drawn to represent these data.

The bar representing the yield $5 \leqslant y < 10$ has a width of 1.5 cm and a height of 8 cm.

(a) Calculate the width and the height of the bar representing the yield $15 \leqslant y < 25$

(3)

(b) Use linear interpolation to estimate the median yield of the tomato plants.

(2)

(c) Estimate the mean and the standard deviation of the yields of the tomato plants.

(4)

(d) Describe, giving a reason, the skewness of the data.

(2)

(e) Estimate the number of tomato plants in the sample that have a yield of more than 1 standard deviation above the mean.

(2)

Leave blank

5. A researcher believes that parents with a short family name tended to give their children a long first name. A random sample of 10 children was selected and the number of letters in their family name, x, and the number of letters in their first name, y, were recorded.

The data are summarised as:

$$\sum x = 60, \quad \sum y = 61, \quad \sum y^2 = 393, \quad \sum xy = 382, \quad S_{xx} = 28$$

(a) Find S_{yy} and S_{xy}

(3)

(b) Calculate the product moment correlation coefficient, r, between x and y.

(2)

(c) State, giving a reason, whether or not these data support the researcher's belief.

(2)

The researcher decides to add a child with family name "Turner" to the sample.

(d) Using the definition $S_{xx} = \sum (x - \bar{x})^2$, state the new value of S_{xx} giving a reason for your answer.

(2)

Given that the addition of the child with family name "Turner" to the sample leads to an increase in S_{yy}

(e) use the definition $S_{xy} = \sum (x - \bar{x})(y - \bar{y})$ to determine whether or not the value of r will increase, decrease or stay the same. Give a reason for your answer.

(2)