Edexcel past paper questions

Statistics 1

Correlation & Regression

Product-moment correlation coefficient

The product-moment correlation coefficient, *r*, measures how close the points on a scatter graph lie to a straight line (or more mathematically it measures the strength of the linear relationship between two variables).
The key points:

- ***r* always lies between -1 and 1**. (If you calculate a value of r that does not lie between -1 and 1 then you've made a mistake!!).
- The formula for calculating *r* is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

where

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}.$$

Note: These formulae are in the formula book. The PMCC value doesn't change with any linear coding. It is important that you interpret a value of the correlation coefficient in the context of the question.

Regression

Calculating the regression line of y on x

- The regression line of *y* on *x* has equation

$$y = a + bx,$$

where

$$b = \frac{S_{xy}}{S_{xx}}$$

$$a = \bar{y} - b\bar{x}\ ; \quad \bar{x} = \frac{\sum x_i}{n},\ \bar{y} = \frac{\sum y_i}{n}$$

and

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}.$$

These formulae are in the formula book.

An **independent (explanatory) variable** is one that is set independently of the other variable. (Plotted on the x axis).

A **dependent (response) variable** is one whose values are determined by the values of the independent variable. (Plotted on the y axis).

Interpolation is when you estimate the value of a dependent variable within the range of the data.

Extrapolation is when you estimate a value outside the range of the data. Values estimated by extrapolation can be unreliable.

1. A local authority is investigating the cost of reconditioning its incinerators. Data from 10 randomly chosen incinerators were collected. The variables monitored were the operating time x (in thousands of hours) since last reconditioning and the reconditioning cost y (in £1000). None of the incinerators had been used for more than 3000 hours since last reconditioning.

The data are summarised below,

$$\Sigma x = 25.0,\ \Sigma x^2 = 65.68,\ \Sigma y = 50.0,\ \Sigma y^2 = 260.48,\ \Sigma xy = 130.64.$$

(a) Find S_{xx}, S_{xy}, S_{yy}. **(3 marks)**

(b) Calculate the product moment correlation coefficient between x and y. **(3 marks)**

(c) Explain why this value might support the fitting of a linear regression model of the form $y = a + bx$. **(1 mark)**

(d) Find the values of a and b. **(4 marks)**

(e) Give an interpretation of a. **(1 mark)**

(f) Estimate

(i) the reconditioning cost for an operating time of 2400 hours,

(ii) the financial effect of an increase of 1500 hours in operating time. **(4 marks)**

(g) Suggest why the authority might be cautious about making a prediction of the reconditioning cost of an incinerator which had been operating for 4500 hours since its last reconditioning. **(2 marks)**

Jan 2001, Q6

2. On a particular day in summer 1993 at 0800 hours the height above sea level, x metres, and the temperature, y °C, were recorded in 10 Mediterranean towns. The following summary statistics were calculated from the results.

$$\Sigma x = 7300,\ \Sigma x^2 = 6\,599\,600,\ S_{xy} = -13\,060,\ S_{yy} = 140.9.$$

(a) Find S_{xx}. **(2)**

(b) Calculate, to 3 significant figures, the product moment correlation coefficient between x and y. **(2)**

(c) Give an interpretation of your coefficient. **(1)**

June 2001, Q2

3. A music teacher monitored the sight-reading ability of one of her pupils over a 10 week period. At the end of each week, the pupil was given a new piece to sight-read and the teacher noted the number of errors y. She also recorded the number of hours x that the pupil had practised each week. The data are shown in the table below.

x	12	15	7	11	1	8	4	6	9	3
y	8	4	13	8	18	12	15	14	12	16

(a) Plot these data on a scatter diagram.

(3)

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

(You may use $\Sigma x^2 = 746$, $\Sigma xy = 749$.)

(9)

(c) Give an interpretation of the slope and the intercept of your regression line.

(2)

(d) State whether or not you think the regression model is reasonable

(i) for the range of x-values given in the table,

(ii) for all possible x-values.

In each case justify your answer either by giving a reason for accepting the model or by suggesting an alternative model.

(2)

June 2001, Q7

4. A number of people were asked to guess the calorific content of 10 foods. The mean s of the guesses for each food and the true calorific content t are given in the table below.

Food	t	s
Packet of biscuits	170	420
1 potato	90	160
1 apple	80	110
Crisp breads	10	70
Chocolate bar	260	360
1 slice white bread	75	135
1 slice brown bread	60	115
Portion of beef curry	270	350
Portion of rice pudding	165	390
Half a pint of milk	160	200

[You may assume that $\Sigma t = 1340$, $\Sigma s = 2310$, $\Sigma ts = 396\,775$, $\Sigma t^2 = 246\,050$, $\Sigma s^2 = 694\,650$.]

(a) Draw a scatter diagram, indicating clearly which is the explanatory (independent) and which is the response (dependent) variable.

(3)

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

(7)

(c) State, with a reason, whether or not the value of the product moment correlation coefficient changes if all the guesses are 50 calories higher than the values in the table.

(2)

The mean of the guesses for the portion of rice pudding and for the packet of biscuits are outside the linear relation of the other eight foods.

(d) Find the equation of the regression line of s on t excluding the values for rice pudding and biscuits.

(3)

[You may now assume that $S_{ts} = 72\,587$, $S_{tt} = 63\,671.875$, $\bar{t} = 125.625$, $\bar{s} = 187.5$.]

(e) Draw the regression line on your scatter diagram.

(2)

(f) State, with a reason, what the effect would be on the regression line of including the values for a portion of rice pudding and a packet of biscuits.

(2)

Jan 2002, Q7

5. An ice cream seller believes that there is a relationship between the temperature on a summer day and the number of ice creams sold. Over a period of 10 days he records the temperature at 1p.m., t °C, and the number of ice creams sold, c, in the next hour. The data he collects is summarised in the table below.

t	c
13	24
22	55
17	35
20	45
10	20
15	30
19	39
12	19
18	36
23	54

[Use $\Sigma t^2 = 3025$, $\Sigma c^2 = 14245$, $\Sigma ct = 6526$.]

(a) Calculate the value of the product moment correlation coefficient between t and c.

(7)

(b) State whether or not your value supports the use of a regression equation to predict the number of ice creams sold. Give a reason for your answer. **(2)**

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

(2)

(d) Interpret the value of b. **(1)**

(e) Estimate the number of ice creams sold between 1 p.m. and 2 p.m. when the temperature at 1 p.m. is 16 °C.

(3)

(f) At 1 p.m. on a particular day, the highest temperature for 50 years was recorded. Give a reason why you should not use the regression equation to predict ice cream sales on that day.

(1)

May 2002, Q7

6. An agricultural researcher collected data, in appropriate units, on the annual rainfall x and the annual yield of wheat y at 8 randomly selected places.

The data were coded using $s = x - 6$ and $t = y - 20$ and the following summations were obtained.

$$\Sigma s = 48.5, \quad \Sigma t = 65.0, \quad \Sigma s^2 = 402.11, \quad \Sigma t^2 = 701.80, \quad \Sigma st = 523.23$$

(a) Find the equation of the regression line of t on s in the form $t = p + qs$.

(7)

(b) Find the equation of the regression line of y on x in the form $y = a + bx$, giving a and b to 3 decimal places.

(3)

The value of the product moment correlation coefficient between s and t is 0.943, to 3 decimal places.

(c) Write down the value of the product moment correlation coefficient between x and y. Give a justification for your answer.

(2)
Nov 2002, Q5

7. The chief executive of Rex cars wants to investigate the relationship between the number of new car sales and the amount of money spent on advertising. She collects data from company records on the number of new car sales, c, and the cost of advertising each year, p (£000). The data are shown in the table below.

Year	Number of new car sale, c	Cost of advertising (£000), p
1990	4240	120
1991	4380	126
1992	4420	132
1993	4440	134
1994	4430	137
1995	4520	144
1996	4590	148
1997	4660	150
1998	4700	153
1999	4790	158

(*a*) Using the coding $x = (p - 100)$ and $y = \frac{1}{10}(c - 4000)$, draw a scatter diagram to represent these data. Explain why x is the explanatory variable.

(5)

(*b*) Find the equation of the least squares regression line of y on x.

[Use $\Sigma x = 402$, $\Sigma y = 517$, $\Sigma x^2 = 17\,538$ and $\Sigma xy = 22\,611$.]

(7)

(*c*) Deduce the equation of the least squares regression line of c on p in the form $c = a + bp$.

(3)

(*d*) Interpret the value of a.

(2)

(*e*) Predict the number of extra new cars sales for an increase of £2000 in advertising budget. Comment on the validity of your answer.

(2)

Jan 2003, Q6

8. A company owns two petrol stations P and Q along a main road. Total daily sales in the same week for P (£p) and for Q (£q) are summarised in the table below.

	p	q
Monday	4760	5380
Tuesday	5395	4460
Wednesday	5840	4640
Thursday	4650	5450
Friday	5365	4340
Saturday	4990	5550
Sunday	4365	5840

When these data are coded using $x = \frac{p - 4365}{100}$ and $y = \frac{q - 4340}{100}$,

$\Sigma x = 48.1$, $\Sigma y = 52.8$, $\Sigma x^2 = 486.44$, $\Sigma y^2 = 613.22$ and $\Sigma xy = 204.95$.

(*a*) Calculate S_{xy}, S_{xx} and S_{yy}.

(4)

(*b*) Calculate, to 3 significant figures, the value of the product moment correlation coefficient between x and y.

(3)

(*c*) (i) Write down the value of the product moment correlation coefficient between p and q.

(ii) Give an interpretation of this value. **(3)**

June 2003, Q3

9. Eight students took tests in mathematics and physics. The marks for each student are given in the table below where *m* represents the mathematics mark and *p* the physics mark.

		Student							
		A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Mark	*m*	9	14	13	10	7	8	20	17
	p	11	23	21	15	19	10	31	26

A science teacher believes that students' marks in physics depend upon their mathematical ability. The teacher decides to investigate this relationship using the test marks.

(*a*) Write down which is the explanatory variable in this investigation.

(1)

(*b*) Draw a scatter diagram to illustrate these data.

(3)

(*c*) Showing your working, find the equation of the regression line of *p* on *m*.

(8)

(*d*) Draw the regression line on your scatter diagram.

(2)

A ninth student was absent for the physics test, but she sat the mathematics test and scored 15.

(*e*) Using this model, estimate the mark she would have scored in the physics test.

(2)

June 2003, Q7

10. A company wants to pay its employees according to their performance at work. The performance score x and the annual salary, y in £100s, for a random sample of 10 of its employees for last year were recorded. The results are shown in the table below.

x	15	40	27	39	27	15	20	30	19	24
y	216	384	234	399	226	132	175	316	187	196

[You may assume $\Sigma xy = 69\,798$, $\Sigma x^2 = 7\,266$]

(*a*) Draw a scatter diagram to represent these data. **(4)**

(b) Calculate exact values of S_{xy} and S_{xx}. **(4)**

(c) (i) Calculate the equation of the regression line of y on x, in the form $y = a + bx$.

Give the values of a and b to 3 significant figures.

(ii) Draw this line on your scatter diagram. **(5)**

(d) Interpret the gradient of the regression line. **(1)**

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

(e) Find the proposed annual salary for an employee who has a performance score of 35. **(2)**

Nov 2003, Q1

11. An office has the heating switched on at 7.00 a.m. each morning. On a particular day, the temperature of the office, t °C, was recorded m minutes after 7.00 a.m. The results are shown in the table below.

m	0	10	20	30	40	50
t	6.0	8.9	11.8	13.5	15.3	16.1

(*a*) Calculate the exact values of S_{mt} and S_{mm}. **(4)**

(*b*) Calculate the equation of the regression line of t on m in the form $t = a + bm$. **(3)**

(*c*) Use your equation to estimate the value of t at 7.35 a.m. **(2)**

(*d*) State, giving a reason, whether or not you would use the regression equation in (*b*) to estimate the temperature

(i) at 9.00 a.m. that day,

(ii) at 7.15 a.m. one month later. **(4)**

Jan 2004, Q1

12. A researcher thinks there is a link between a person's height and level of confidence. She measured the height h, to the nearest cm, of a random sample of 9 people. She also devised a test to measure the level of confidence c of each person. The data are shown in the table below.

h	179	169	187	166	162	193	161	177	168
c	569	561	579	561	540	598	542	565	573

[You may use $\Sigma h^2 = 272\,094$, $\Sigma c^2 = 2\,878\,966$, $\Sigma hc = 884\,484$]

(a) Draw a scatter diagram to illustrate these data. **(4)**
(b) Find exact values of S_{hc} S_{hh} and S_{cc}. **(4)**
(c) Calculate the value of the product moment correlation coefficient for these data. **(3)**
(d) Give an interpretation of your correlation coefficient. **(1)**
(e) Calculate the equation of the regression line of c on h in the form $c = a + bh$. **(3)**
(f) Estimate the level of confidence of a person of height 180 cm. **(2)**
(g) State the range of values of h for which estimates of c are reliable. **(1)**

June 2004, Q2

13. An experiment carried out by a student yielded pairs of (x, y) observations such that

$$\bar{x} = 36, \quad \bar{y} = 28.6, \quad S_{xx} = 4402, \quad S_{xy} = 3477.6$$

(a) Calculate the equation of the regression line of y on x in the form $y = a + bx$. Give your values of a and b to 2 decimal places. **(3)**

(b) Find the value of y when $x = 45$. **(1)**

Nov 2004

14. Students in Mr Brawn's exercise class have to do press-ups and sit-ups. The number of press-ups x and the number of sit-ups y done by a random sample of 8 students are summarised below.

$$\Sigma x = 272, \qquad \Sigma x^2 = 10\,164, \qquad \Sigma xy = 11\,222,$$
$$\Sigma y = 320, \qquad \Sigma y^2 = 13\,464.$$

(a) Evaluate S_{xx}, S_{yy} and S_{xy}. **(4)**
(b) Calculate, to 3 decimal places, the product moment correlation coefficient between x and y. **(3)**
(c) Give an interpretation of your coefficient. **(2)**
(d) Calculate the mean and the standard deviation of the number of press-ups done by these students. **(4)**

Nov 2004, Q2

15. The following table shows the height x, to the nearest cm, and the weight y, to the nearest kg, of a random sample of 12 students.

x	148	164	156	172	147	184	162	155	182	165	175	152
y	39	59	56	77	44	77	65	49	80	72	70	52

(a) On graph paper, draw a scatter diagram to represent these data.

(3)

(b) Write down, with a reason, whether the correlation coefficient between x and y is positive or negative.

(2)

The data in the table can be summarised as follows.

$$\Sigma x = 1962, \quad \Sigma y = 740, \quad \Sigma y^2 = 47\,746, \quad \Sigma xy = 122\,783, \quad S_{xx} = 1745.$$

(c) Find S_{xy}.

(2)

The equation of the regression line of y on x is $y = -106.331 + bx$.

(d) Find, to 3 decimal places, the value of b.

(2)

(e) Find, to 3 significant figures, the mean $\bar{y}$ and the standard deviation s of the weights of this sample of students.

(3)

(f) Find the values of $\bar{y} \pm 1.96s$.

(2)

(g) Comment on whether or not you think that the weights of these students could be modelled by a normal distribution.

(1)

Jan 2005, Q3

16. 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 $y = f - 100$.

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

(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)**

June 2005, Q3

17. 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.

(You may use $\Sigma x^2 = 1352$, $\Sigma y^2 = 53\,112$ and $\Sigma xy = 8354$.)

(7)

(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)

Jan 2006, Q3

18. 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)

May 2006, Q3

19. 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 $\Sigma tm = 2485$)

(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)

Jan 2007, Q1

20. 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)**

June 2007, Q1

21. 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. **(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)**

June 2007, Q3

22. 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)**

Jan 2008, Q1

23. 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)**

Jan 2008, Q4

24. 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)

May 2008, Q4

25. 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)**

Jan 2009, Q1

26. 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.

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

(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)**

May 2009, Q1

27. 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)**

May 2009, Q5

28. 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) 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)**

Jan 2010, Q6

29. 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:

$$S_{xx} = 1022500, \qquad S_{yy} = 130.9, \qquad S_{xy} = 8825.$$

(*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)**

May 2010, Q1

30. 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*) On graph paper, draw 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)**

May 2010, Q6

31. 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.

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

(*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)

Jan 2011, Q4

32. 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

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

(*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)

May 2011, Q1

33. 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 $\sum f = 186$ $\sum h = 1085$ $S_{ff} = 39.5$ $S_{hh} = 139.875$ $\sum fh = 25\,291$)

(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 part (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 part (b) should not be used to estimate the teacher's height. **(1)**

May 2011, Q7

34. 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)**

Jan 2012, Q5

35. 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.

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

(*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)**

May 2012, Q2

36. 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*) On graph paper, draw a scatter diagram 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 $(\bar{p}, \bar{t})$ 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)

May 2012, Q3

37. 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)

Jan 2013, Q3

38. A teacher asked a random sample of 10 students to record the number of hours of television, t, they watched in the week before their mock exam. She then calculated their grade, g, in their mock exam. The results are summarised as follows.

$$\sum t = 258 \quad \sum t^2 = 8702 \quad \sum g = 63.6 \quad S_{gg} = 7.864 \quad \sum gt = 1550.2$$

(a) Find S_{tt} and S_{gt}.

(3)

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

(2)

The teacher also recorded the number of hours of revision, v, these 10 students completed during the week before their mock exam. The correlation coefficient between t and v was –0.753.

(c) Describe, giving a reason, the nature of the correlation you would expect to find between v and g.

(2)

Jan 2013, Q1

39. 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 $\sum h = 7150$, $\sum t = 110$, $\sum h^2 = 7171500$, $\sum t^2 = 1716$, $\sum th = 64\,980$ 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)

May 2013, Q1

40. Sammy is studying the number of units of gas, g, and the number of units of electricity, e, used in her house each week. A random sample of 10 weeks use was recorded and the data for each week were coded so that $x = \frac{g-60}{4}$ and $y = \frac{e}{10}$. The results for the coded data are summarised below

$$\sum x = 48.0, \quad \sum y = 58.0, \quad S_{xx} = 312.1, \quad S_{yy} = 2.10, \quad S_{xy} = 18.35$$

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

Give the values of a and b correct to 3 significant figures.

(4)

(*b*) Hence find the equation of the regression line of e on g in the form $e = c + dg$.

Give the values of c and d correct to 2 significant figures.

(4)

(*c*) Use your regression equation to estimate the number of units of electricity used in a week when 100 units of gas were used.

(2)

May 2013_R, Q1

41. 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)**

May 2013_R, Q5

42. The table shows data on the number of visitors to the UK in a month, v (1000s), and the amount of money they spent, m (£ millions), for each of 8 months.

Number of visitors v (1000s)	2450	2480	2540	2420	2350	2290	2400	2460
Amount of money spent m (£ millions)	1370	1350	1400	1330	1270	1210	1330	1350

$S_{vv} = 42587.5$ $S_{vm} = 31512.5$ $S_{mm} = 25187.5$ $\Sigma v = 19390$ $\Sigma m = 10610$

(*a*) Find the product moment correlation coefficient between m and v. **(2)**
(*b*) Give a reason to support fitting a regression model of the form $m = a + bv$ to these data. **(1)**
(*c*) Find the value of b correct to 3 decimal places. **(2)**
(*d*) Find the equation of the regression line of m on v. **(2)**
(*e*) Interpret your value of b. **(2)**
(*f*) Use your answer to part (*d*) to estimate the amount of money spent when the number of visitors to the UK in a month is 2 500 000. **(2)**
(*g*) Comment on the reliability of your estimate in part (*f*). Give a reason for your answer. **(2)**

June 2014, Q3

43. A large company is analysing how much money it spends on paper in its offices every year. The number of employees, x, and the amount of money spent on paper, p (£ hundreds), in 8 randomly selected offices are given in the table below.

x	8	9	12	14	7	3	16	19
p (£ hundreds)	40.5	36.1	30.4	39.4	32.6	31.1	43.4	45.7

(You may use $\Sigma x^2 = 1160$ $\Sigma p = 299.2$ $\Sigma p^2 = 11\,422$ $\Sigma xp = 3449.5$)

(a) Show that $S_{pp} = 231.92$ and find the value of S_{xx} and the value of S_{xp}. **(5)**

(b) Calculate the product moment correlation coefficient between x and p. **(2)**

The equation of the regression line of p on x is given in the form $p = a + bx$.

(c) Show that, to 3 significant figures, $b = 0.824$ and find the value of a. **(4)**

(d) Estimate the amount of money spent on paper in an office with 10 employees. **(2)**

(e) Explain the effect each additional employee has on the amount of money spent on paper. **(1)**

Later the company realised it had made a mistake in adding up its costs, p. The true costs were actually half of the values recorded. The product moment correlation coefficient and the equation of the linear regression line are recalculated using this information.

(f) Write down the new value of

(i) the product moment correlation coefficient,

(ii) the gradient of the regression line. **(2)**

June 2014_R, Q3

44. Statistical models can provide a cheap and quick way to describe a real world situation.

(a) Give two other reasons why statistical models are used. **(2)**

A scientist wants to develop a model to describe the relationship between the average daily temperature, x °C, and her household's daily energy consumption, y kWh, in winter.

A random sample of the average daily temperature and her household's daily energy consumption are taken from 10 winter days and shown in the table.

x	–0.4	–0.2	0.3	0.8	1.1	1.4	1.8	2.1	2.5	2.6
y	28	30	26	25	26	27	26	24	22	21

[You may use $\sum x^2 = 24.76$ $\sum y = 255$ $\sum xy = 283.8$ $S_{xx} = 10.36$]

(b) Find S_{xy} for these data. **(3)**

(c) Find the equation of the 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) Give an interpretation of the value of a. **(1)**

(e) Estimate her household's daily energy consumption when the average daily temperature is 2°C. **(2)**

The scientist wants to use the linear regression model to predict her household's energy consumption in the summer.

(f) Discuss the reliability of using this model to predict her household's energy consumption in the summer. **(2)**

June 2015, Q4

45. An estate agent recorded the price per square metre, p £/m^2, for 7 two-bedroom houses.

He then coded the data using the coding $q = \frac{p - a}{b}$, where a and b are positive constants.

His results are shown in the table below.

p	1840	1848	1830	1824	1819	1834	1850
q	4.0	4.8	3.0	2.4	1.9	3.4	5.0

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

(2)

The estate agent also recorded the distance, d km, of each house from the nearest train station. The results are summarised below.

$$S_{dd} = 1.02 \qquad S_{qq} = 8.22 \qquad S_{dq} = -2.17$$

(b) Calculate the product moment correlation coefficient between d and q.

(2)

(c) Write down the value of the product moment correlation coefficient between d and p.

(1)

The estate agent records the price and size of 2 additional two-bedroom houses, H and J.

House	Price (£)	Size (m^2)
H	156 400	85
J	172 900	95

(d) Suggest which house is most likely to be closer to a train station. Justify your answer.

(3)

June 2015, Q2