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)

DO NOT WRITE IN THIS AREA DO NOT WRITE IN T

2. Two students, Olive and Shan, collect data on the weight, w grams, and the tail length, t cm, of 15 mice.

Olive summarised the data as follows

$S_{tt} = 5.3173$ $\quad \sum w^2 = 6089.12$ $\quad \sum tw = 2304.53$ $\quad \sum w = 297.8$ $\quad \sum t = 114.8$

(a) Calculate the value of S_{tw} and the value of S_{ww}

(3)

(b) Calculate the value of the product moment correlation coefficient between w and t

(2)

(c) Show that the equation of the regression line of w on t can be written as

$$w = -16.7 + 4.77t$$

(3)

(d) Give an interpretation of the gradient of the regression line.

(1)

(e) Explain why it would not be appropriate to use the regression line in part (c) to estimate the weight of a mouse with a tail length of 2 cm.

(2)

Shan decided to code the data using $x = t - 6$ and $y = \frac{w}{2} - 5$

(f) Write down the value of the product moment correlation coefficient between x and y

(1)

(g) Write down an equation of the regression line of y on x
You do not need to simplify your equation.

(1)

3: Statistical models are a cheap and quick way to make predictions about real-world situations.

(a) Give one other reason why statistical models are used.

(1)

Madison wants to develop a model to describe the relationship between the average daily temperature, t °C, and a household's daily gas consumption, d m^3, in winter.

Madison takes a random sample of 12 days in winter and codes the daily gas consumption so that $w = 11.5d$

These data are summarised as follows

$$S_{tt} = 26.43 \quad S_{tw} = -91.55 \quad \sum w = 339.25 \quad \sum t = 9.1 \quad \sum w^2 = 10\,036.45$$

(b) Show that $S_{ww} = 445.57$ to 2 decimal places.

(1)

(c) Find the value of S_{td} and the value of S_{dd}

(3)

(d) Find the product moment correlation coefficient between d and t

(2)

(e) Give an interpretation, in context, of your product moment correlation coefficient.

(1)

(f) Show that the equation of the regression line of w on t is

$$w = -3.46t + 30.9$$

where the values of the intercept and the gradient are given to 3 significant figures.

(3)

(g) Write down an equation of the regression line of d on t

(1)

(h) Using your equation in part (g)

(i) estimate the daily gas consumption in winter when the temperature is 2 °C

(ii) interpret the effect an increase of 1 °C in average daily temperature is expected to have on the daily gas consumption in winter.

(2)

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

4. A French test and a Spanish test were sat by 11 students.

The table below shows their marks.

Student	A	B	C	D	E	F	G	H	I	J	K
French mark (f)	24	30	32	32	36	36	40	44	50	60	68
Spanish mark (s)	16	90	24	28	32	36	38	44	48	48	68

Greg says that if these points were plotted on a scatter diagram, then the point (30, 90) would be an outlier because 90 is an outlier for the Spanish marks.

An outlier is defined as a value that is

$$\text{greater than } Q_3 + 1.5 \times (Q_3 - Q_1) \text{ or smaller than } Q_1 - 1.5 \times (Q_3 - Q_1)$$

(a) Show that 90 is an outlier for the Spanish marks.

(3)

Ignoring the point (30, 90), Greg calculated the following summary statistics.

$$\sum f = 422 \quad \sum s = 382 \quad S_{ff} = 1667.6 \quad S_{fs} = 1735.6$$

(b) Use these summary statistics to show that the equation of the least squares regression line of s on f for the remaining 10 students is

$$s = -5.72 + 1.04f$$

where the values of the intercept and gradient are given to 3 significant figures. You must show your working.

(3)

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

(1)

Two further students sat the French test but missed the Spanish test.

(d) Using the equation given in part (b), estimate

(i) a Spanish mark for the student who scored 55 marks in their French test,

(ii) a Spanish mark for the student who scored 18 marks in their French test.

(3)

(e) State, giving a reason, which of the two estimates found in part (d) would be the more reliable estimate.

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

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.

$$\sum c = 3660 \quad \sum h = 126 \quad \sum c^2 = 973228$$

$$\sum ch = 20023.4 \quad S_{cc} = 303448 \quad S_{ch} = -3034.6$$

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)

DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS

4. A biologist is studying bears. The biologist records the length, d cm, and the girth, g cm, of 8 bears. The biologist summarises the data as follows

$$\sum d = 1456.8 \quad \sum g = 713.2 \quad \sum dg = 141\,978.84 \quad \sum g^2 = 72\,675.98$$

$$S_{dd} = 16\,769.78$$

(a) Calculate the exact value of S_{dg} and the exact value of S_{gg}

(3)

(b) Calculate the value of the product moment correlation coefficient between d and g

(2)

(c) Show that the equation of the regression line of g on d can be written as

$$g = -42.3 + 0.722d$$

where the values of the intercept and gradient are given to 3 significant figures.

(3)

(d) Give an interpretation, in context, of the gradient of the regression line.

(1)

Using the equation of the regression line given in part (c)

(e) (i) estimate the girth of a bear with a length of 2.5 metres,

(ii) explain why an estimate for the girth of a bear with a length of 0.5 metres is not reliable.

(2)

Using the regression line from part (c), the biologist estimates that for each x cm increase in the length of a bear there will be a 17.3 cm increase in the girth.

(f) Find the value of x

(2)

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

6. As a part of a selection process, applicants for a television game show have to take a test, and then complete a task as quickly as possible. The test score, w, and the time taken to complete the task, t minutes, are recorded for each applicant.

The summary statistics below represent the data for a random sample of 30 applicants.

$$S_{wt} = -1648.83 \quad S_{ww} = 2396.97 \quad \sum w = 839 \quad \sum t = 635 \quad \sum t^2 = 14837$$

(a) Show that the product moment correlation coefficient for these data is -0.901 to 3 significant figures.

(2)

A scatter diagram of t against w is plotted for these data.

(b) State **two** features of the graph that you would expect to see, given the correlation coefficient in part (a)

(2)

(c) Calculate the equation of the regression line of t on w in the form

$$t = a + bw$$

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

(4)

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

(1)

The test score, w, was a score out of 50
The manager of the selection process now decides to double all the values of w to make them into percentages.
The manager then recalculates the product moment correlation coefficient and the equation of the regression line.

(e) State, for **each** of the following, whether the value would increase **or** decrease **or** stay the same as a result of applying this change.

(i) The product moment correlation coefficient

(ii) The magnitude of the gradient of the regression line

(iii) The t intercept of the regression line

(3)

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

Leave blank

5. A large company rents shops in different parts of the country. A random sample of 10 shops was taken and the floor area, x in $10\,m^2$, and the annual rent, y in thousands of dollars, were recorded.
The data are summarised by the following statistics

$$\sum x = 900 \quad \sum x^2 = 84\,818 \quad \sum y = 183 \quad \sum y^2 = 3434$$

and the regression line of y on x has equation $y = 6.066 + 0.136x$

(a) Use the regression line to estimate the annual rent in dollars for a shop with a floor area of $800\,m^2$

(2)

(b) Find S_{yy} and S_{xx}

(3)

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

(4)

An 11th shop is added to the sample. The floor area is $900\,m^2$ and the annual rent is 15000 dollars.

(d) Use the formula $S_{xy} = \sum(x - \bar{x})(y - \bar{y})$ to show that the value of S_{xy} for the 11 shops will be the same as it was for the original 10 shops.

(2)

(e) Find the new equation of the regression line of y on x for the 11 shops.

(3)

The company is considering renting a larger shop with area of $3000\,m^2$

(f) Comment on the suitability of using the new regression line to estimate the annual rent. Give a reason for your answer.

(1)

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