Save
My
Exams

A Level · Edexcel · Maths

3 hours

24 questions

2.4 Correlation & Regression

2.4.1 Correlation & Regression

Easy (8 questions)	/42
Medium (5 questions)	/33
Hard (5 questions)	/36
Very Hard (6 questions)	/44
Total Marks	**/155**

Scan here to return to the course

or visit savemyexams.com

Easy Questions

1 (a) For each of the following four scatter graphs, identify the type and strength of any linear correlation shown.

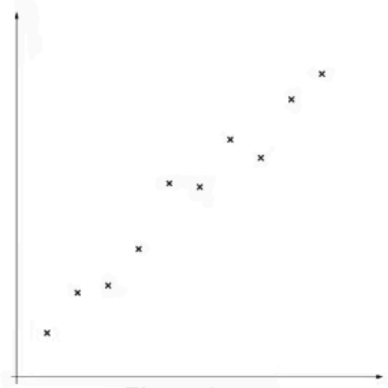


Figure 1

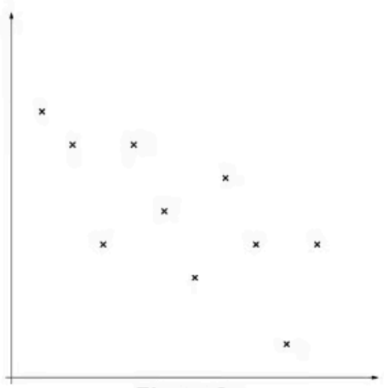


Figure 2

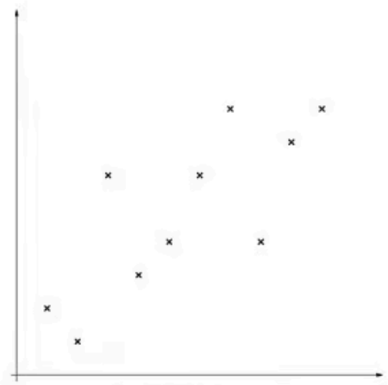


Figure 3

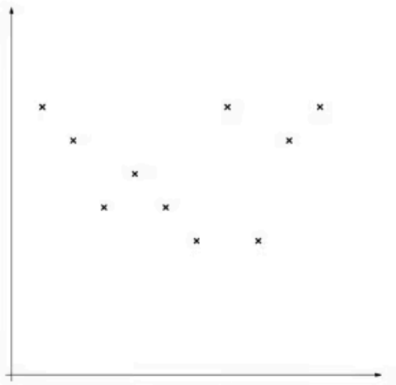


Figure 4

(4 marks)

(b) Sketch a scatter graph to show a perfect negative linear correlation between two variables.

(1 mark)

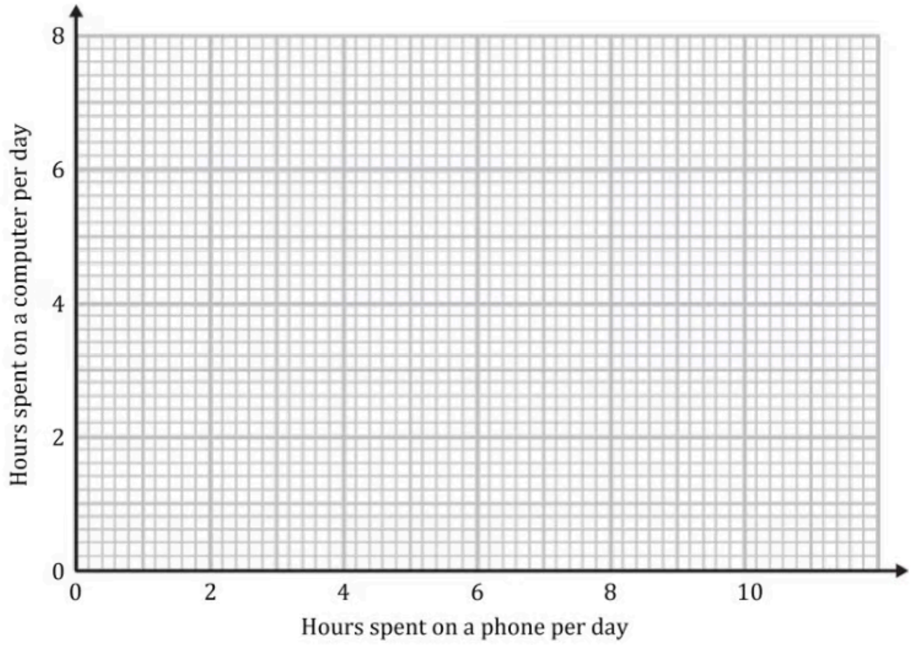
2 A teacher is interested in the relationship between the number of hours her students spend on a phone per day and the number of hours they spend on a computer. She takes a sample of nine students and records the results in the table below.

Hours spent on a phone per day	7.6	7	8.9	3	3	7.5	2.1	1.3	5.8
Hours spent on a computer per day	1.7	1.1	0.7	5.8	5.2	1.7	6.9	7.1	3.3

(i) Plot a scatter diagram of this data on the axes below.

(ii) Describe the linear correlation shown in your diagram.

(iii) Interpret the correlation in the context of the question.



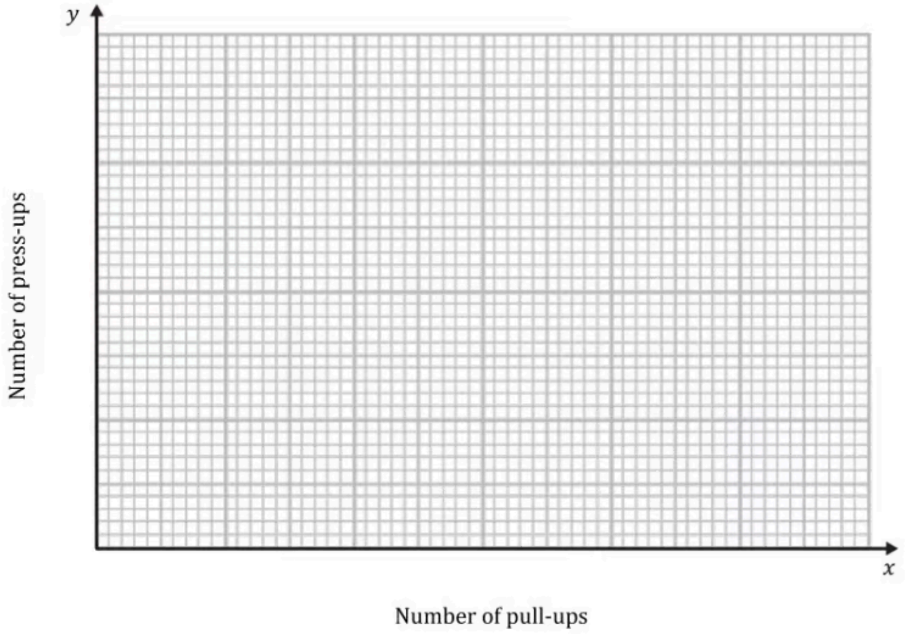
(5 marks)

3 (a) The table below shows data for a sample of 8 people comparing the maximum number of pull-ups they are able to complete, x, with the maximum number of press-ups, y.

Number of pull-ups (x)	5	10	8	3	6	8	1	4
Number of press-ups (y)	24	34	36	18	30	35	11	19

(i) Plot a scatter diagram on the axes below.

(ii) Describe the type of correlation shown in your scatter diagram.



(4 marks)

(b) The equation of the regression line of y on x is y=3x + 9.

(i) Add this regression line to your scatter diagram.

(ii) Explain the purpose of regression lines and how they may be used.

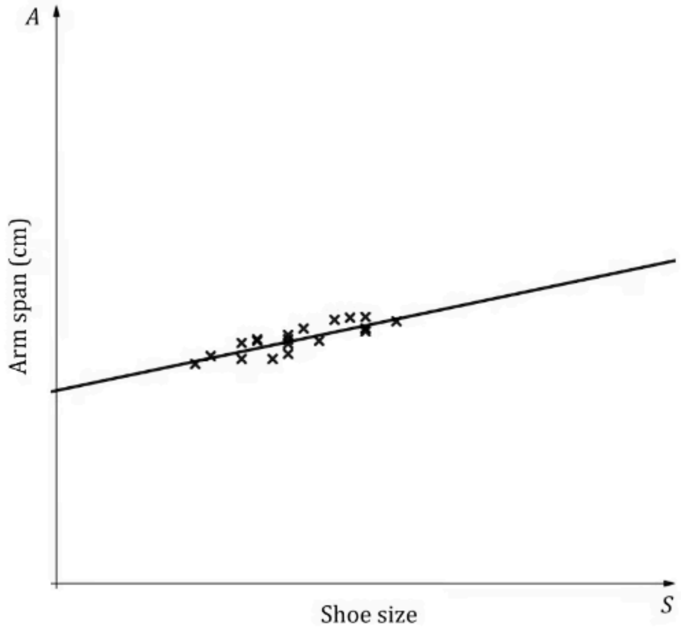
(4 marks)

4 (a) A class is asked to collect a sample of bivariate data. They collect data on the shoe size, *S*, and the arm span, *A* cm, of 20 randomly selected boys from the class.

Explain what is meant by the term 'bivariate data'.

(1 mark)

(b) The class plot the data in a scatter diagram and find the equation of the regression line of *A* on *S* to be *A*=4.5 *S* + 133. These are both plotted in the diagram below.



(i) Interpret the value 4.5 in the context of the question.

(ii) Interpret the value 133 in the context of the question.

(iii) Explain how the sign of the coefficient of S in the equation is related to the correlation shown in the scatter diagram.

(3 marks)

5 (a) The following table shows data comparing the length of time a cake was baked for, t minutes, with the mass of the cake once it has cooled, m grams. Each cake in the sample weighed the same before being baked.

t	37	35	36	31	30	28	36
m	825	868	812	943	947	997	837

State which variable is the explanatory (independent) variable and which is the response (dependent) variable.

(1 mark)

(b) The equation for the regression line of m on t is $m = 1531 - 19t$.

(i) Use the regression line to estimate the mass of a cake if it is baked for 32 minutes.

(ii) Comment on the validity of your estimate in part (b)(i).

(2 marks)

(c) (i) Use the regression line to estimate the mass of a cake if it is baked for 80 minutes.

(ii) Comment on the validity of your estimate in part (c)(i).

(2 marks)

6 (a) Isla is investigating whether the number of deep-fried chocolate bars a person eats has an impact on his or her level of fitness. She takes a sample of 10 people and records how many deep-fried chocolate bars they eat during a month, c, and then times how long it takes them to complete a 100-metre sprint, t seconds, at the end of the month.

She plotted the data in a scatter diagram and found the equation of the regression line of t on c to be $t = 5c + 12$.

Find an estimate for the 100-metre sprint time for a person if they eat:

(i) 2 deep-fried chocolate bars in a month,

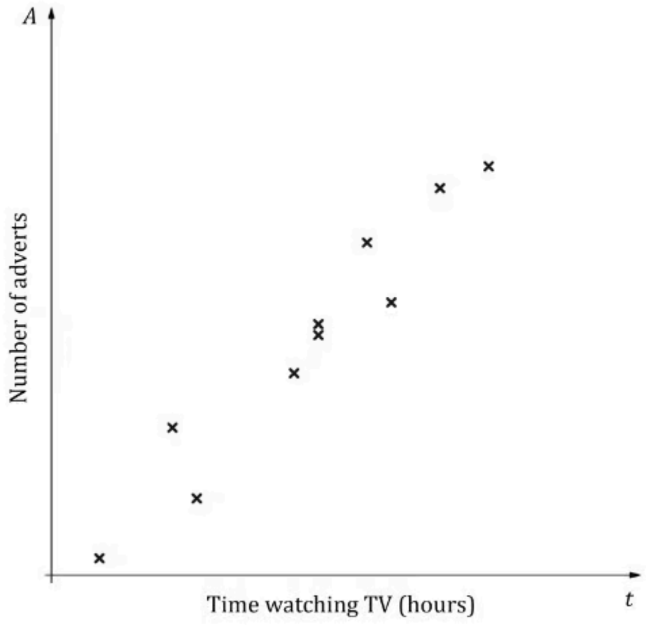
(ii) 54 deep-fried chocolate bars in a **year**.

(3 marks)

(b) Describe the type of linear correlation you would expect to see on Isla's scatter diagram and state which value in the regression equation tells you this.

(2 marks)

7 (a) Terrence has collected data comparing how many adverts, A, he sees whilst watching TV for different lengths of time, t hours. With this data, Terrence plotted the scatter diagram shown below.



(i) Describe the linear correlation shown in this scatter diagram.

(ii) What does the correlation suggest about the relationship between the number of adverts Terrence sees and the length of time he watches TV?

(2 marks)

(b) State, with a reason, whether each of the following equations would be appropriate for the equation of the regression line of *A* on *t*:

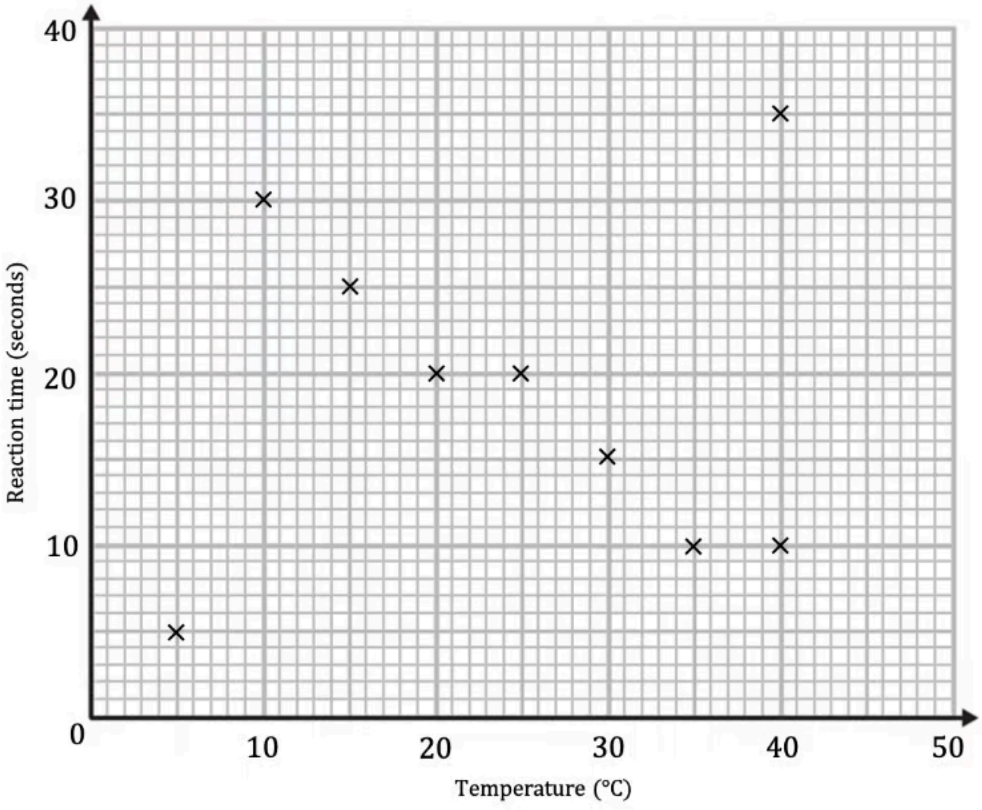
(i) A=18t+5,

(ii) t=18A+5,

(iii) A=-18t+5.

(3 marks)

8 (a) Two liquids are mixed and heated to a particular temperature. The time, in seconds, it takes the two liquids to react is recorded. The scatter diagram below shows the results.



(i) Identify the two outliers shown on the scatter diagram.

(ii) Clean the data by removing these outliers and find the mean reaction time.

(3 marks)

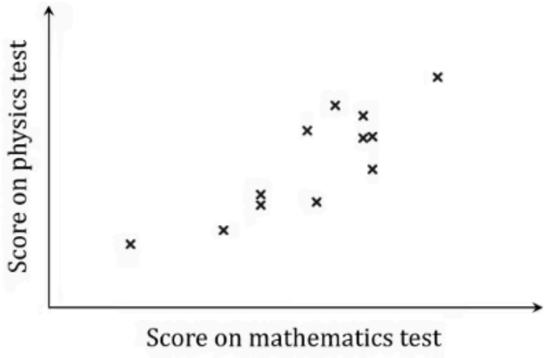
(b) (i) Describe the correlation shown by the scatter diagram.

(ii) A student says that if the mixture is heated to 60 °C the two liquids will react almost instantly. Explain why the student may be incorrect.

(2 marks)

Medium Questions

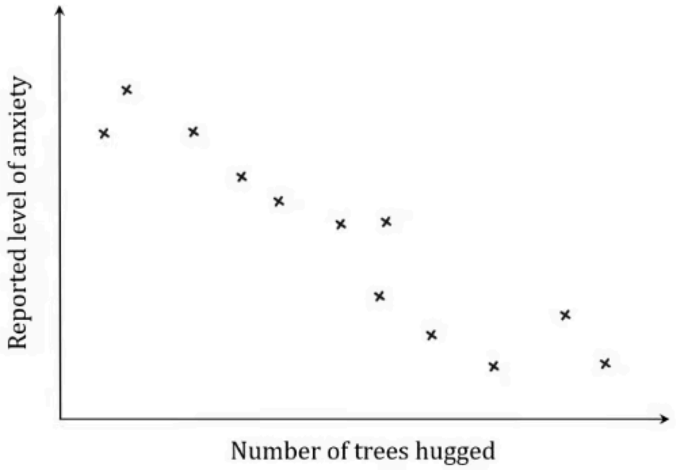
1 (a) A teacher collected the maths and physics test scores of a number of students and drew a scatter diagram to represent this data.



Describe the correlation shown by the scatter diagram, and interpret the correlation in context.

(2 marks)

(b) An alternative therapist collected data on his clients' reported levels of anxiety as well as the number of trees they had hugged in the course of therapy. He drew a scatter diagram to represent this data.



Describe the correlation shown by the scatter diagram, and interpret the correlation in context.

(2 marks)

2 (a) The table below shows data from the United States regarding annual per capita cheese consumption (in pounds) and the divorce rate (number of divorces per 1000 people) for ten years between 2000 and 2018:

Year	2000	2002	2004	2006	2008	2010	2012	2014	2016	2018
Cheese consumption (pounds)	32.1	32.8	33.6	34.8	34.5	35	35.5	36.2	38.5	40
Divorce rate (number per 1000 people)	4	3.9	3.7	3.7	3.5	3.6	3.4	3.2	3.0	2.9

Draw a scatter diagram to represent this data, with per capita cheese consumption on the horizontal axis and divorce rate on the vertical axis.

(3 marks)

(b) (i) Describe the correlation between per capita cheese consumption and divorce rate.

(ii) Do you think there is a causal relationship between per capita cheese consumption and divorce rate in the United States?
Explain your reasoning.

(2 marks)

3 (a) Myfanwy has been applying different voltages (v, measured in volts) to an electrical circuit in her lab and recording the resulting currents (i, measured in amps). The smallest voltage she applied was 0.5 volts, and the largest voltage she applied was 120 volts.

She found the equation of the regression line of i on v to be $i = 0.056+0.332v$.

(i) Interpret the value 0.332 in this context.

(ii) Use the equation to predict the current for a voltage of 70 volts.

(2 marks)

(b) Explain why it would not be sensible to use the regression equation to work out:

(i) the current resulting from a voltage of 2000 volts

(ii) the voltage corresponding to a current of 20 amps.

(2 marks)

(c) Myfanwy's lab partner suggests that the value 0.056 in the regression equation represents the current in the circuit when the voltage applied is zero. Explain why he might suggest this, but also suggest a reason why his interpretation is most likely incorrect.

(2 marks)

4 (a) The following table shows the height, h cm, and weight, w kg, for each of eleven students at a sixth form college.

h	167	182	176	173	17	174	177	178	172	170	169
w	51	62	69	65	65	56	64	62	51	55	58

The following statistics were calculated for the data on height:

mean=159.5 cm,   standard deviation=45.3 cm

An outlier is an observation which lies more than ±2 standard deviations from the mean.

(i) Show that h=17 is an outlier.

(ii) Explain why this outlier should be omitted from the data.

(2 marks)

(b) With the outlier data excluded, the equation of the regression line of w on h is $w = -87.6 + 0.845h$.

(i) Exclude the outlier data from the recorded measurements and draw a scatter diagram to represent the data for the remaining ten students.

(ii) Draw the regression line on your diagram.

(5 marks)

(c) Based on your diagram, along with the regression equation, to what extent would you say that a person's height may be used as an accurate predictor of his or her weight?

(2 marks)

5 (a) The table below shows data from the large data set on the daily mean pressure, p (hPa), and daily total sunshine, s (hrs), in Camborne for a random sample of 12 days in 2015.

p	1007	1023	1011	1022	1011	1019	1017	1016	1022	997	1030	1023
s	0	6.3	2.4	6.2	1.7	8.4	1.9	6.7	7.7	2.3	10.3	4.1

The equation of the regression line of s on p is $s = -270.5 + 0.271\, p$.

Give an interpretation of the value of the gradient of the regression line.

(1 mark)

(b) Use your knowledge of the large data set to explain whether there is likely to be a causal relationship between daily mean pressure and daily total sunshine.

(2 marks)

(c) Explain why it would not be reliable to use this regression equation to predict:

(i) the daily total sunshine on a day with a mean daily pressure of 980 hPa

(ii) the mean daily pressure on a day with 5.6 hours of total sunshine.

(2 marks)

(d) Use the regression equation to predict the daily total sunshine for a day with a daily mean pressure of 1017 hPa. How does this compare with the actual data from the 1017 hPa day in the table?

(2 marks)

(e) The table includes data for the 1st and 15th days of each of the months covered by the large data set for 2015.

Suggest another factor that may have had an effect on the daily total sunshine data recorded in the table. Explain how taking account of that factor might allow a better model to be produced describing the relationship between daily mean pressure and daily total sunshine.

(2 marks)

Hard Questions

1 (a) Ella measures how the extension, x mm, of a thin piece of metal wire varies with the force applied to it, F kN. She records her results in the table below.

F	15	32	49	76	99	106	112	124	132
x	0.2	0.4	0.6	0.9	1.4	1.5	1.6	1.8	1.8

Ella calculates the regression line of F on x to be $F = 0.004 - 69.3\,x$.

Explain why this equation must be wrong.

(1 mark)

(b) The correct equation for the regression line of F on x is $F = 6.16 + 67.6x$.

Interpret the value of 67.6 in this context.

(1 mark)

(c) Using the correct regression line, Ella estimates that if she applies a force of 1000 kN then the wire will show an extension of 14.7 mm.

Give two reasons why Ella's estimate may not be accurate.

(2 marks)

2 (a) The table below shows a comparison of the average house price, *H* (£100 000), and the average yearly income, *I* (£10 000), for different areas around the UK in 2021.

Area	***H***	***I***
Conwy	155.1	26.4
Perth and Kinross	181.3	27.9
Richmondshire	190.3	25.1
Monmouthshire	232.6	31.4
Trafford	260.2	32.0
Gwynedd	148.5	23.6
Basingstoke and Dean	297.7	33.7
Daventry	259.2	29.5

(i) Plot a scatter diagram of *I* against *H*, and

(ii) describe the correlation shown.

(4 marks)

(b) The equation of the regression line of *I* on *H* is calculated to be $I = 0.06H + 15.92$.
A particularly unscrupulous politician uses this to claim that if you want a salary of £35 000, all you need to do is buy a house that costs £583 000.

Comment on the validity of the politician's claim.

(2 marks)

3 (a) Two researchers, Alwyn and Beth, are working on a project collecting data about the self-reported happiness of students on a scale from 0 to 10, H, and the number of exams sat by those students, n. After collecting data from 1000 students, they construct a scatter diagram and find the equation of the regression line of H on n to be $H = 7.63 - 0.82n$.

Explain what correlation the data is likely to show in the scatter diagram.

(2 marks)

(b) What information about the original data set would need to be checked before using the regression line equation to estimate the self-reported happiness of a student sitting 8 exams?

(1 mark)

(c) After calculating the equation of the line of regression, Alwyn accidentally deletes all the data collected about the self-reported happiness scores. Alwyn says it's not a problem since he can use the regression line and the number of exams sat to recalculate all the values. Beth says that Alwyn is wrong and the original data is lost forever.

Explain which researcher is correct.

(2 marks)

4 (a) A consultant is trying to improve the efficiency of how a factory making chewing gum operates. To help them do this, they collect many types of data about the factory workers. One such type of data is the number of chewing gum packets made per shift. The list below shows the number of chewing gum packets made by a particular worker (Worker 1) during the last 10 shifts worked.

392 414 536 474 212 396 427 545 459 234

Calculate the mean number of chewing gum packets made per shift by Worker 1 to the nearest whole number of packets.

(1 mark)

(b) The table below shows the mean number of chewing gum packets, *N*, made by various workers along with how many hours of training, *T* hours, they have received.

Worker	1	2	3	4	5	6	7	8	9
N		512	499	359	393	432	456	520	475
T	18	24	22.5	15	16	20	21	22	21

(i) Including your answer from (a), plot a scatter diagram of the data in the table above.

(ii) Given that the equation of the regression line of *N* on *T* is $N = 18T+95$, add the regression line to your scatter diagram.

(5 marks)

(c) The consultant then goes on to collect even more data on other factory workers and records some of it in the table below.

Worker	10	11	12	13	14	15	16	17	18
N	600	598	584	602	593	585	591	601	605
T	29	28.5	32	29	34.5	30.5	37	31	30

Without adding this new data to your scatter diagram, what advice could the consultant give to the factory to improve the efficiency of their workers?

(3 marks)

5 (a) The table below shows data from the large data set on the daily mean temperature in Heathrow, T_H℃, and the daily mean temperature in Beijing, T_B℃, on the same day for a random sample of 10 days during May 2015.

T_H	11.9	11.0	16.0	13.1	15.3	13.5	11.7	14.8	9.8	14.5
T_B	23.9	20.0	26.3	24.6	25.3	28.0	19.6	25.6	17.5	27.5

(i) Plot a scatter diagram of T_B against T_H, and

(ii) explain the correlation shown in this context.

(4 marks)

(b) The equation for the regression line of T_B on T_H is $T_B = 4.55+1.46T_H$.

The table below shows T_H and T_B on the same day for a random sample of 3 other days during May 2015.

$\mathbf{T_H}$	14.3	11.0	12.4
$\mathbf{T_B}$	21.1	22.8	24.0

Considering this second sample, use the regression line equation and the values of T_H to predict values of T_B and find the average percentage difference of these estimated values of T_B from the true values of T_B. Hence, comment on how accurately this regression line equation can predict values.

(6 marks)

(c) Using your knowledge of the large data set, explain whether there is likely to be a causal relationship between T_H and T_B.

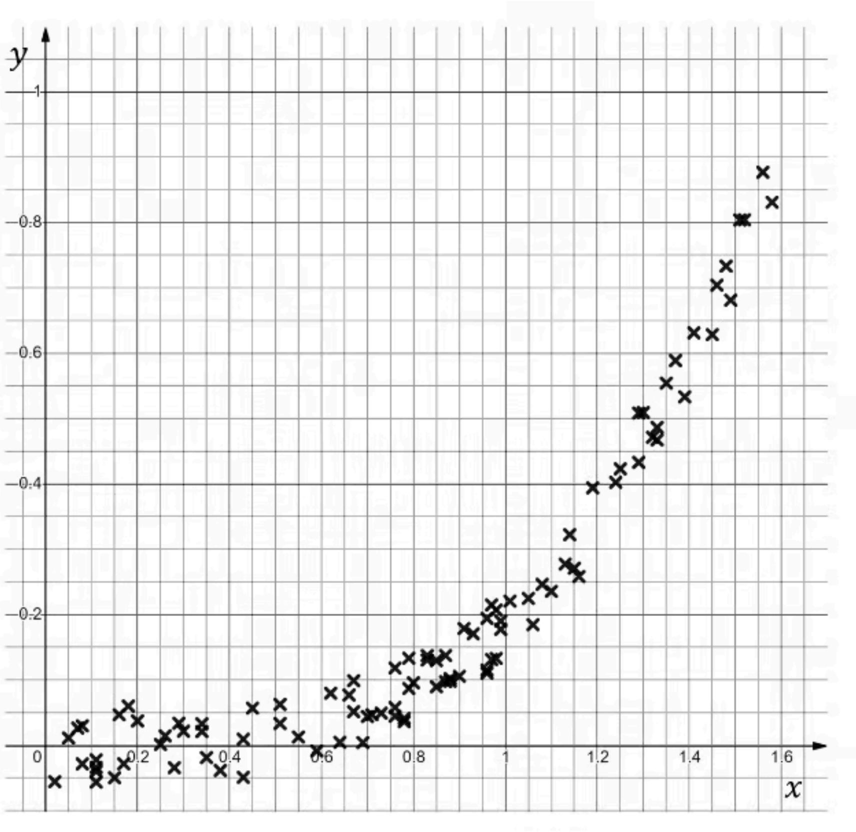
(1 mark)

(d) A researcher claims that this correlation between T_H and T_B is a coincidence. How could you use data from the large data set to check this claim?

(1 mark)

Very Hard Questions

1 Four statisticians are arguing over which line best highlights the trend of the set of data shown in the scatter diagram below.



The first statistician draws, by eye, a line of best fit and claims its equation is $y = -0.05 + 0.17x$. The second draws, again by eye, a different line of best fit and claims its equation is $y = -1.08 + 1.3x$. The third calculates the equation of the regression line of y on x claims it is $y = 0.18 + 0.11x$. The fourth statistician claims that all three of the other statisticians are definitely wrong and that there is no line of best fit.

By adding each of these lines to the scatter diagram, comment on the claims of each of the statisticians.

(5 marks)

2 (a) Paige takes a sample of 9 cities throughout the UK to compare the percentage of people living in a city who identify as vegan, *V* %, and the percentage of restaurants offering vegan options in that same city, *R* %.

The regression line of *R* on *V* is calculated, and it is used to predict values of *R* for *V*=1.35 and *V*=1.03, the values returned are *R*=70.73 and *R*=50.314 respectively.

Find the equation of the regression line of *R* on *V*.

(4 marks)

(b) In one of the cities, 1.16% of people were vegan and 55.9% of restaurants offered vegan options.

Use the equation of the regression line of *R* on *V* to estimate the percentage of restaurants offering vegan options in a city in which 1.16% of people are vegan. Give your estimated value of *R* to 3 significant figures. Compare this to the information above.

(2 marks)

(c) Paige discovers that in one city every restaurant offers vegan options. Paige suggests that the equation of the regression line of *R* on *V* can be used to find the percentage of people in this city who identify as vegan. Explain why Paige is likely wrong.

(2 marks)

3 (a) A ride sharing app collected data on the time, t minutes, taken to complete a journey of distance, d miles. Data from a random sample of 8 journeys is detailed in the table below.

d	3.9	6.6	8.5	1.3	1.7	3.7	7.4	6.1
t	25	36	39	6	8	19	38	32

By plotting a scatter diagram of t on d for this data, explain whether or not it is appropriate to use a linear regression model on this data.

(5 marks)

(b) Using a new random sample of thousands of journeys, the ride sharing app calculated the regression line of time on distance to be $t = -1.8 + 5.9d$.

The app uses this regression equation to predict that a journey of distance 7 km would take 39.5 minutes. Explain why this is incorrect.

(1 mark)

(c) The regression equation predicts that for journeys less than 0.3 miles the time taken will be less than zero minutes. What is the most likely reason that the regression equation gives this false prediction?

(1 mark)

4 (a) A maths teacher randomly selects 10 students from a class of 30 to answer a survey. The survey asks students how many practice questions they completed when revising for a recent test, Q, and their percentage score in that test, S %. Summary statistics for Q are shown below

$$\overline{Q}=21 \qquad \text{Range of Q=20}$$

The equation of the regression line of S on Q is $S = 34 + 2Q$.

Explain which variable is the response variable.

(2 marks)

(b) Use the regression equation to find an estimate for the mean value and range of S. State any assumptions that are needed.

(6 marks)

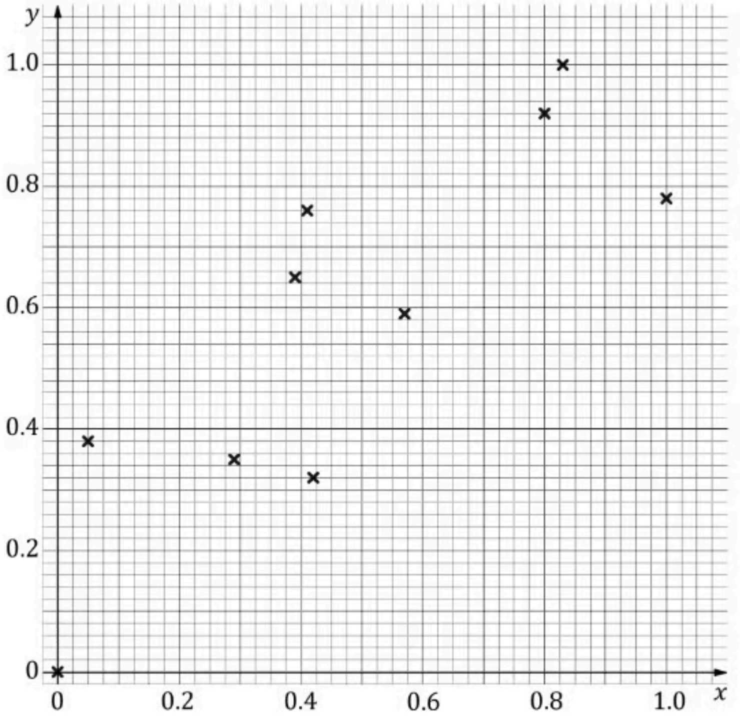
(c) Comment on the reliability of using the regression equation to:

(i) estimate the scores of the other students in the maths class,

(ii) estimate the scores of this cohort of students in a science class.

(2 marks)

5 (a) An owner of a beach resort is comparing parasol sales, £p, and sun cream sales, £s, at the resort over a period of eleven days. The data is standardised by coding the variables using $x = \frac{s - 153}{103}$ and $y = \frac{p - 32}{37}$. The values for the first ten days are plotted on the scatter diagram below.



(i) On the eleventh day, the resort sold £246 worth of sun cream and £69 worth of parasols. Use this information to complete the scatter diagram.

(ii) The equation for the regression line of y on x is $y = 0.19+0.83x$. Add the regression line to the scatter diagram.

(4 marks)

(b) (i) Show that by using the regression line of y on x and the coding equations above, the regression line of p on s can be written in the form $p = a + bs$, where a and b are constants to be found to 3 significant figures.

(ii) Hence, or otherwise, find an estimate for the amount of parasol sales on a day where there are £170 of sun cream sales.

(5 marks)

6 (a) An ice cream shop owner in Camborne is trying to use data from the large data set alongside their own past sales data to help them estimate future sales. The mean daily temperature per month, T ℃, is shown with the mean daily number of ice creams sold per month, I, from 2015 in the table below.

Month	May	June	July	August	September	October
T	11.2	13.8	15.7	15.4	13.6	12.2
I	57	132	259	227	133	101

The equation for the regression line of I on T is $I = -429.5 + 42.5T$.

Find an estimate for the expected total number of ice creams sold in the month of July if the average daily temperature for that month is 14.9° C.

(2 marks)

(b) Suggest which other data from the large data set could be used to improve this model.

(1 mark)

(c) The ice cream shop owner claims that there is a causal link between I and T, and so if the shop sells more ice cream, the month will be hotter.

Comment on this claim.

(2 marks)