Questions

Q1.

A company is introducing a job evaluation scheme. Points (x) will be awarded to each job based on the qualifications and skills needed and the level of responsibility. Pay (£y) will then be allocated to each job according to the number of points awarded.

Before the scheme is introduced, a random sample of 8 employees was taken and the linear regression equation of pay on points was $y = 4.5x - 47$

(a) Describe the correlation between points and pay.

(1)

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

(1)

(c) Explain why this model might not be appropriate for all jobs in the company.

(1)

(Total for question = 3 marks)

Q2.

A sixth form college has 84 students in Year 12 and 56 students in Year 13

The head teacher selects a stratified sample of 40 students, stratified by year group.

(a) Describe how this sample could be taken.

(3)

The head teacher is investigating the relationship between the amount of sleep, *s* hours, that each student had the night before they took an aptitude test and their performance in the test, *p* marks.
For the sample of 40 students, he finds the equation of the regression line of *p* on *s* to be

$$p = 26.1 + 5.60s$$

(b) With reference to this equation, describe the effect that an extra 0.5 hours of sleep may have, on average, on a student's performance in the aptitude test.

(1)

(c) Describe one limitation of this regression model.

(1)

(Total for question = 5 marks)

Q3.

Sara was studying the relationship between rainfall, *r* mm, and humidity, *h* %, in the UK. She takes a random sample of 11 days from May 1987 for Leuchars from the large data set.

She obtained the following results.

h	93	86	95	97	86	94	97	97	87	97	86
r	1.1	0.3	3.7	20.6	0	0	2.4	1.1	0.1	0.9	0.1

Sara examined the rainfall figures and found

$$Q_1 = 0.1 \qquad Q_2 = 0.9 \qquad Q_3 = 2.4$$

A value that is more than 1.5 times the interquartile range (IQR) above Q_3 is called an outlier.

(a) Show that *r* = 20.6 is an outlier.

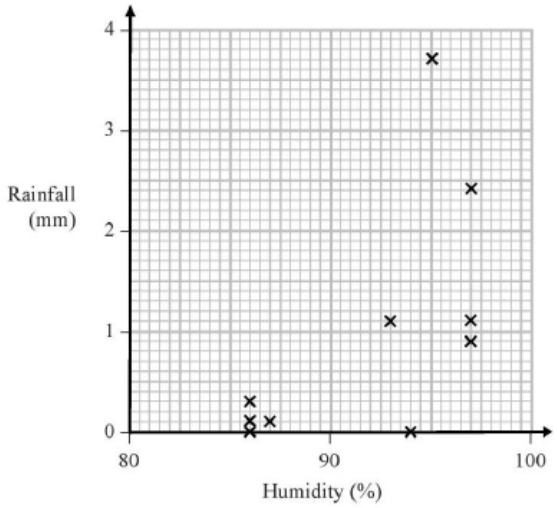
(1)

(b) Give a reason why Sara might

(i) include
(ii) exclude
this day's reading.

(2)

Sara decided to exclude this day's reading and drew the following scatter diagram for the remaining 10 days' values of *r* and *h*.



(c) Give an interpretation of the correlation between rainfall and humidity.

(1)

The equation of the regression line of *r* on *h* for these 10 days is *r* = –12.8 + 0.15*h*

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

(1)

(e) (i) Comment on the suitability of Sara's sampling method for this study.

(ii) Suggest how Sara could make better use of the large data set for her study.

(2)

(Total for question = 7 marks)

Q4.

Jerry is studying visibility for Camborne using the large data set June 1987.

The table below contains two extracts from the large data set.

It shows the daily maximum relative humidity and the daily mean visibility.

Date	Daily Maximum Relative Humidity	Daily Mean Visibility
Units	%	
10/06/1987	90	5300
28/06/1987	100	0

(The units for Daily Mean Visibility are deliberately omitted.)

Given that daily mean visibility is given to the nearest 100,

(a) write down the range of distances in metres that corresponds to the recorded value 0 for the daily mean visibility.

(1)

Jerry drew the following scatter diagram, Figure 2, and calculated some statistics using the June 1987 data for Camborne from the large data set.

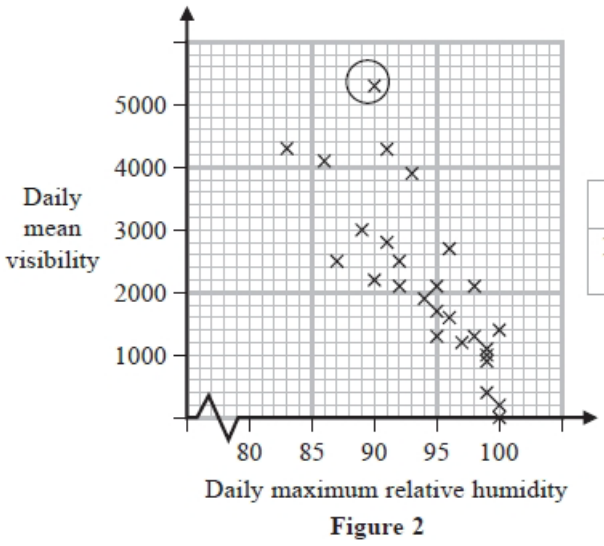


Figure 2

	Q_1	IQR
Daily mean visibility	1100	1600
Daily maximum relative humidity (%)	92	8

Jerry defines an outlier as a value that is more than 1.5 times the interquartile range above Q_3 or more than 1.5 times the interquartile range below Q_1.

(b) Show that the point circled on the scatter diagram is an outlier for visibility.

(2)

(c) Interpret the correlation between the daily mean visibility and the daily maximum relative humidity.

(1)

Jerry drew the following scatter diagram, Figure 3, using the June 1987 data for Camborne from the large data set, but forgot to label the *x*–axis.

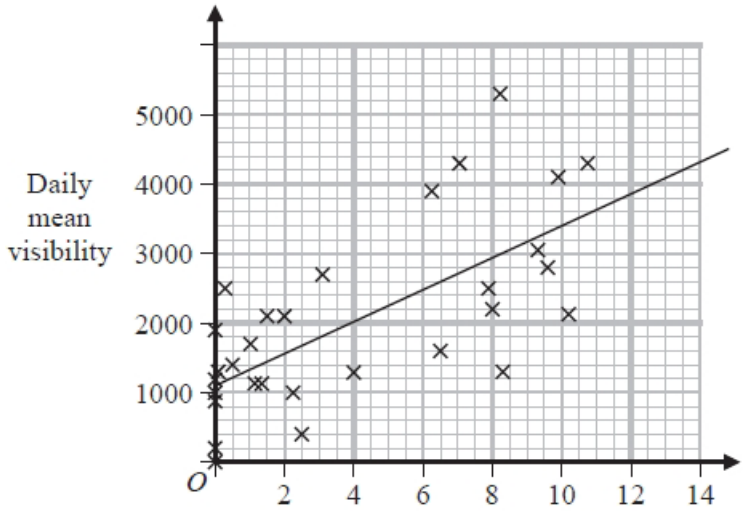


Figure 3

(d) Using your knowledge of the large data set, suggest which variable the *x*-axis on this scatter diagram represents.

(1)

(Total for question = 5 marks)

Mark Scheme

Q1.

Qu	Scheme	Marks	AO
(a)	Positive (correlation)	B1 (1)	1.2
(b)	Every extra point gives £4.5(0) more on pay (o.e.)	B1 (1)	3.4
(c)	e.g. For points < 11 it would give pay < 0 which is ridiculous	B1 (1)	2.4
		(3 marks)	
	Notes		
(a)	B1 for "positive". Allow an interpretation e.g. "as points increase pay increases" is B1 Read whole answer: contradictory comments such as "positive correlation, as points increase pay decreases" scores B0		
(b)	B1 for any correct comment conveying idea of £s per point and including a correct value; must have idea of rate. Can condone missing £ sign. Accept 4.5 e.g. "every 10 points earns an extra (or increase) of £45" is B1 BUT "every point earns £4.5(0)" is B0 *doesn't have idea of rate*		
(c)	B1 for a suitable comment mentioning "points" or "pay" (o.e. e.g. "amount") or commenting on "small sample" or "range of points" used to find line The following examples would score B1 Can say that n points (for $n < 10.4$) would give negative pay so not suitable Any comment suggesting that some jobs would end up with negative pay Don't know the range of points used to find the regression line A small sample of size 8 may not be representative to cover all jobs B0 for a focus on "qualifications" or "hours" worked only The following examples would score B0 Some jobs require no (or low) skills or qualifications (*need negative pay*)		

Q2.

Question	Scheme	Marks	AOs
(a)	Label **each** year group	B1	1.1b
	Use random numbers to select a ...	B1	1.1b
	Simple random sample of 24 Year 12s and 16 Year 13s.	B1	1.1b
		(3)	
(b)	Increase by 2.8 marks	B1	3.4
		(1)	
(c)	e.g. 'the best performance is predicted for the students who never wake up'	B1	3.5b
		(1)	
			(5 marks)

Notes	
(a)	**B1:** for a suitable numbered/labelled/ordered(o.e.) list/database/register(o.e.) for **each** year group. Condone poor numbering but if just one list, then the Year 12s must be distinguishable from the Year 13s
	B1: for use of random numbers/sample/selection to choose students
	B1: for 24 Year 12s, and 16 Year 13s
Note:	A description of a systematic sample: only allow access to the first mark and therefore may score maximum B1B0B0
(b)	**B1:** Using the gradient of the regression equation must include increase(o.e.) and 2.8 'Increase by approximately 3 marks' is B0 but isw if 2.8 is seen $5.6 \div 2$ is not sufficient
(c)	**B1:** for any suitable limitation of the model e.g. the idea that the longer you sleep the better performance in the test or only valid between 0 and 24 hours (within range of the data) or only applicable to the amount of sleep the night before the test or only takes sleep into consideration/does not include other variables (factors) or cannot score below 26.1 marks on the test or the model might not be linear over the entire range or the model might predict more than the maximum mark B0: e.g. might not be correlation between s and p or individual student performance may vary

Q3.

Question	Scheme	Marks	AOs
(a)	IQR = 2.3 and 20.6 ≫ 2.4 + 1.5×2.3 (= 5.85) (Compare correct values)	B1	1.1b
		(1)	
(b)(i)	e.g. it is a piece of data and we should consider all the data (o.e.)	B1	2.4
(ii)	e.g. it is an extreme value and could unduly influence the analysis or it could be a mistake	B1	2.4
		(2)	
(c)	e.g. "as humidity increases rainfall increases"	B1	2.2b
		(1)	
(d)	e.g. a 10% increase in humidity gives rise to a 1.5 mm increase in rainfall or represents 0.15mm of rainfall per percentage of humidity	B1	3.4
		(1)	
(e)(i)	Not a good method since only uses 11 days from one location in one month.	B1	2.4
(ii)	e.g. She should use data from more of the UK locations and more of the months or using a spreadsheet or computer package she could use all of the available UK data	B1	2.4
		(2)	
		(7 marks)	

Part	Notes
(a)	B1 for sight of the correct calculation and suitable comparison with 20.6
(b)(i)	B1 for a suitable reason for including the data point
(ii)	B1 for a suitable reason for excluding the data point
(c)	B1 for a suitable interpretation of positive correlation mentioning humidity and rainfall
(d)	B1 for a suitable description of the rate: rainfall per percentage of humidity including reference to values.
(e)(i)	B1 for a comment that supports the idea that her sampling method was not a good one
(ii)	B1 for some sensible suggestions that would give a better representation of the data across the UK. Must show some awareness of the fact that LDS has different locations and more months of data available but must be clear they are NOT using any overseas locations. NB B0 for a comment that says use more than one location without specifying that only UK locations are required.

Q4.

Question	Scheme	Marks	AOs
(a)	0 to 500 m	B1	1.2
		(1)	
(b)	$1100+1600+1.5\times1600$ [= 5100]	M1	2.1
	5300 > 5100 therefore outlier	A1	1.1b
		(2)	
(c)	As the humidity increases the mean visibility decreases	B1	2.4
		(1)	
(d)	(Hours of) sunshine	B1	2.2b
		(1)	
			(5 marks)

Notes		
(a)	B1:	For realising it is the maximum distance and distance given with correct units. Allow 0 to 50dm or < 500m or < 50dm
(b)	M1:	Attempt to find Q_3 and the upper limit
	A1:	5100, if a value for the point is stated it must be above 5100 otherwise it is A0. For a statement comparing and conclusion it is an outlier or it is above Q_3 +1.5IQR. Allow accept the point circled is greater than 5100 oe
(c)	B1:	For a suitable interpretation of a negative correlation mentioning humidity and visibility
(d)	B1:	A correct deduction that the unlabelled variable is the hours of sunshine. Condone missing hours. Do not allow if more than one variable given. Must be quantative variable Not cloud cover since values bigger than 8 Not wind speed since values not integers Not daily mean temperature since mean temperature near to zero are unlikely in June