Correlation

Questions

Q1.

Two students, Jim and Dora, collected data on the mean annual rainfall, w cm, and the annual yield of leeks, l tonnes per hectare, for 10 years.

Jim summarised the data as follows

$$S_{wl} = 42.786 \qquad S_{ww} = 9936.9 \qquad \sum l^2 = 26.2326 \qquad \sum l = 16.06$$

(a) Find the product moment correlation coefficient between l and w

(2)

Dora decided to code the data first using $s = w - 6$ and $t = l - 20$

(b) Write down the value of the product moment correlation coefficient between s and t.
Give a justification for your answer.

(1)

Dora calculates the equation of the regression line of t on s to be $t = 0.00431s - 18.87$

(c) Find the equation of the regression line of l on w in the form $l = a + bw$, giving the values of a and b to 3 significant figures.

(3)

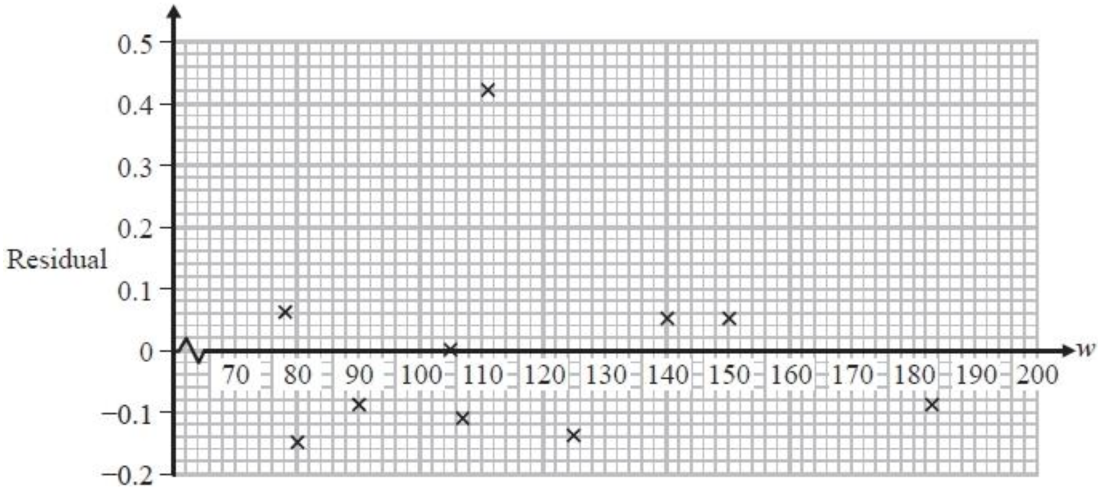
(d) Use your equation to estimate the yield of leeks when w is 100 cm.

(1)

(e) Calculate the residual sum of squares.

(2)

The graph shows the residual for each value of *l*



(f) (i) State whether this graph suggests that the use of a linear regression model is suitable for these data. Give a reason for your answer.

(ii) Other than collecting more data, suggest how to improve the fit of the model in part (c) to the data.

(2)

(Total for question = 11 marks)

Q2.

A large field of wheat is split into 8 plots of equal area. Each plot is treated with a different amount of fertiliser, *f* grams /m^2. The yield of wheat, *w* tonnes, from each plot is recorded. The results are summarised below.

$$\sum f = 28 \qquad \sum w = 303 \qquad \sum w^2 = 13\,447 \qquad S_{ff} = 42 \qquad S_{fw} = 269.5$$

(a) Calculate the product moment correlation coefficient between *f* and *w*

(2)

(b) Interpret the value of your product moment correlation coefficient.

(1)

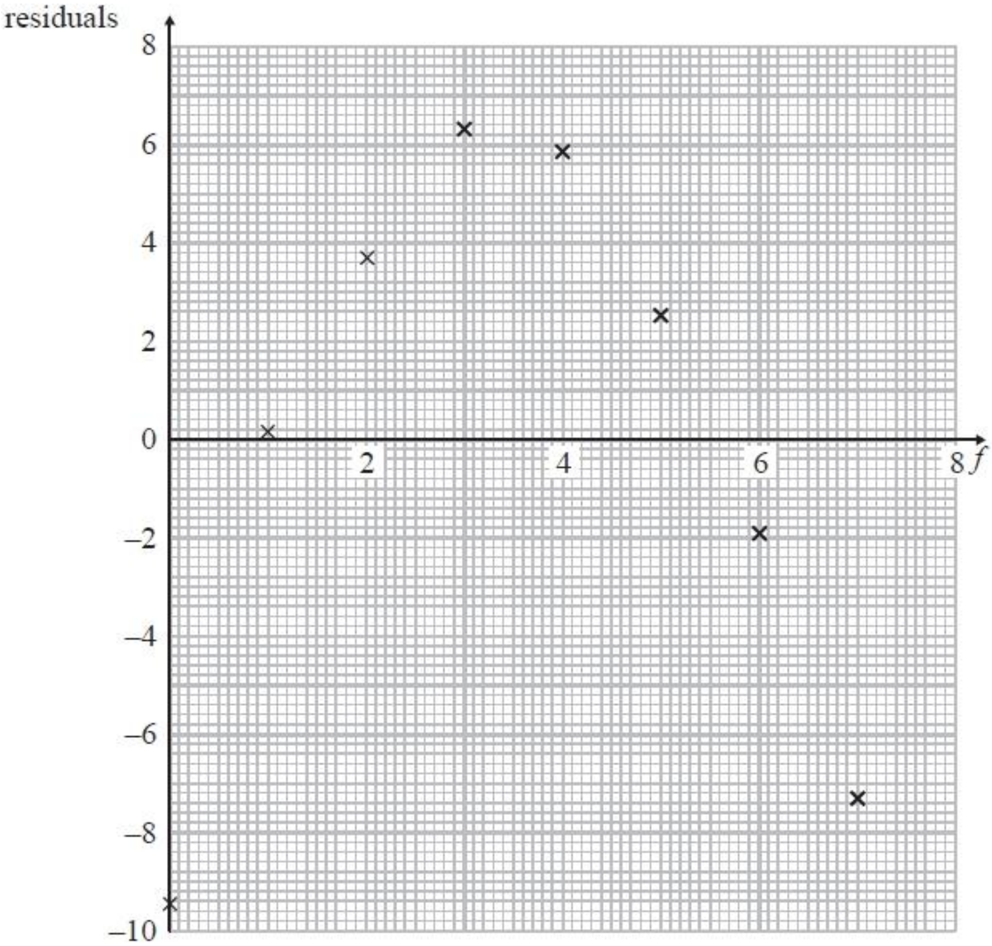
(c) Find the equation of the regression line of *w* on *f* in the form *w* = *a* + *bf*

(3)

(d) Using your equation, estimate the decrease in yield when the amount of fertiliser decreases by 0.5 grams /m^2

(1)

The residuals of the data recorded are calculated and plotted on the graph below.



(e) With reference to this graph, comment on the suitability of the model you found in part (c).

(2)

(f) Suggest how you might be able to refine your model.

(1)

(Total for question = 10 marks)

Q3.

Gabriela is investigating a particular type of fish, called bream. She wants to create a model to predict the weight, *w* grams, of bream based on their length, *x* cm.

For a sample of 27 bream, some summary statistics are given below.

$\bar{x} = 31.07$ $\quad\bar{w} = 628.59$ $\quad\sum w^2 = 11\ 386\ 134$

$S_{xw} = 13\ 082.3$ $\quad S_{xx} = 260.8$

(a) Find the value of the product moment correlation coefficient between *x* and *w*

(3)

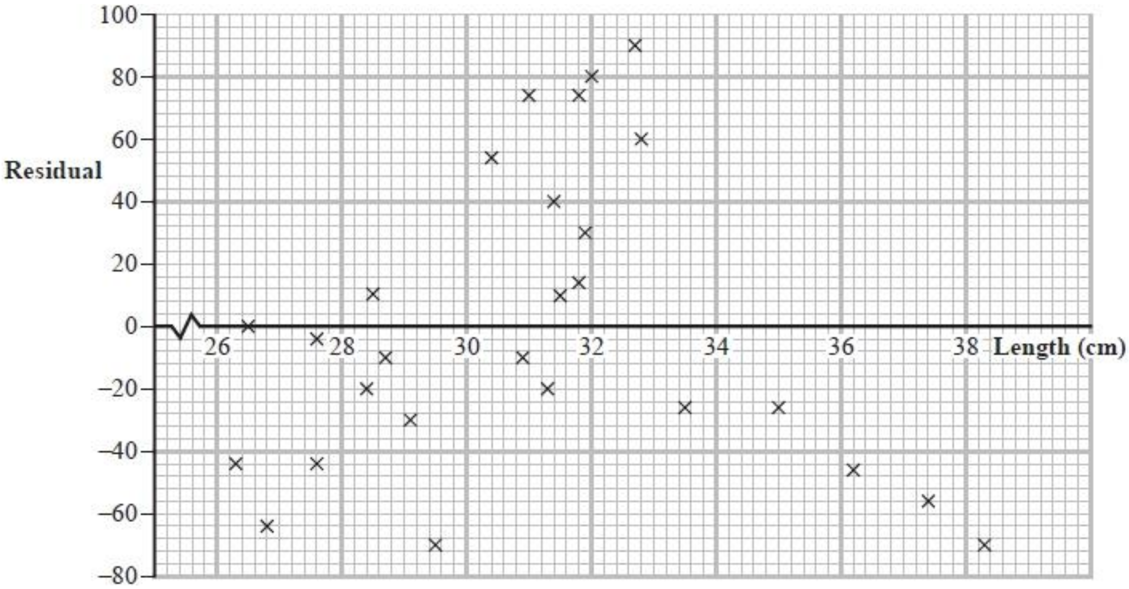
(b) Explain whether the answer to part (a) is consistent with a linear model for these data.

(1)

(c) Find the equation of the regression line of *w* on *x* in the form *w* = *a* + *bx*

(3)

A residual plot for these data is shown below.



One of the bream in the sample has a length of 32 cm.

(d) Find its weight.

(2)

(e) With reference to the residual plot, comment on the model for bream with lengths above 33 cm.

(1)

(Total for question = 10 marks)

Q4.

The scores achieved on a maths test, *m*, and the scores achieved on a physics test, p, by 16 students are summarised below.

$$\sum m = 392 \quad \sum p = 254 \quad \sum p^2 = 4748 \quad S_{mm} = 1846 \quad S_{mp} = 1115$$

(a) Find the product moment correlation coefficient between *m* and *p*

(2)

(b) Find the equation of the linear regression line of *p* on *m*

(3)

Figure 1 shows a plot of the residuals.

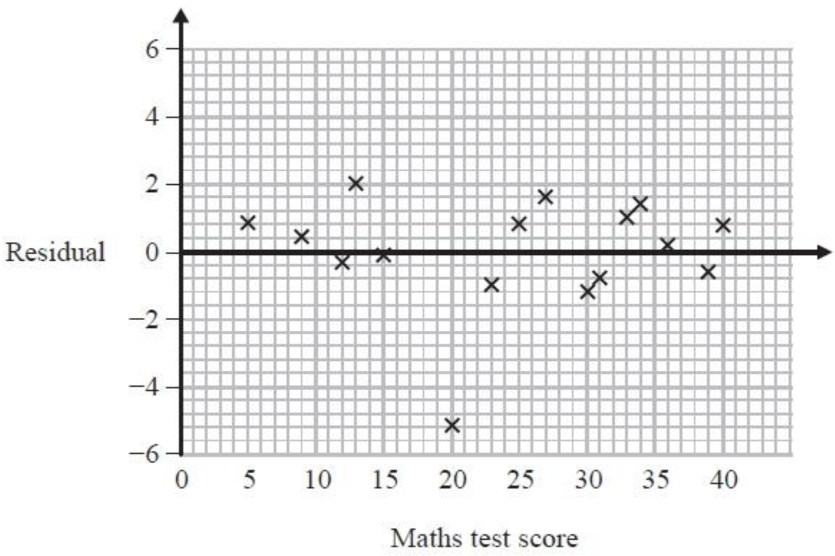


Figure 1

(c) Calculate the residual sum of squares (RSS).

(2)

For the person who scored 30 marks on the maths test,

(d) find the score on the physics test.

(2)

The data for the person who scored 20 on the maths test is removed from the data set.

(e) Suggest a reason why.

(1)

The product moment correlation coefficient between *m* and *p* is now recalculated for the remaining 15 students.

(f) Without carrying out any further calculations, suggest how you would expect this recalculated value to compare with your answer to part (a).

Give a reason for your answer.

(Total for question = 10 marks)

Q5.

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

A random sample of the average temperature and energy consumption are taken from 10 winter days and are summarised below.

$$\sum x = 12 \quad \sum x^2 = 24.76 \quad \sum y = 251 \quad \sum y^2 = 6341 \quad \sum xy = 284.8$$

$$S_{xx} = 10.36 \qquad S_{yy} = 40.9$$

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

(2)

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

(3)

(c) Use your equation to estimate the daily energy consumption when the average daily temperature is 2°C

(1)

(d) Calculate the residual sum of squares (RSS).

(2)

The table shows the residual for each value of *x*.

x	–0.4	–0.2	0.3	0.8	1.1	1.4	1.8	2.1	2.5	2.6
Residual	–0.63	–0.32	–0.52	–0.73	0.74	2.22	1.84	0.32	f	–1.88

(e) Find the value of *f*.

(2)

(f) By considering the signs of the residuals, explain whether or not the linear regression model is a suitable model for these data.

(1)

(Total for question = 11 marks)

Q6.

A clothes shop manager records the weekly sales figures, £ s, and the average weekly temperature, t °C, for 6 weeks during the summer. The sales figures were coded so that $w = \frac{s}{1000}$

The data are summarised as follows

$$S_{ww} = 50 \quad \sum wt = 784 \quad \sum t^2 = 2435 \quad \sum t = 119 \quad \sum w = 42$$

(a) Find S_{wt} and S_{tt}

(3)

(b) Write down the value of S_{ss} and the value of S_{st}

(2)

(c) Find the product moment correlation coefficient between s and t.

(2)

The manager of the clothes shop believes that a linear regression model may be appropriate to describe these data.

(d) State, giving a reason, whether or not your value of the correlation coefficient supports the manager's belief.

(1)

(e) Find the equation of the regression line of w on t, giving your answer in the form $w = a + bt$

(3)

(f) Hence find the equation of the regression line of s on t, giving your answer in the form $s = c + dt$, where c and d are correct to 3 significant figures.

(2)

(g) Using your equation in part (f), interpret the effect of a 1°C increase in average weekly temperature on weekly sales during the summer.

(1)

(Total for question = 14 marks)

Q7.

The table below shows the heights cleared, in metres, for each of 6 competitors in a high jump competition.

Competitor	A	B	C	D	E	F
Height (m)	2.05	1.93	2.02	1.96	1.81	2.02

These 6 competitors also took part in a long jump competition and finished in the following order, with C jumping the furthest.

C A F D B E

(a) Calculate Spearman's rank correlation coefficient for these data.

(4)

(b) Stating your hypotheses clearly, test at the 5% level of significance whether or not there is a positive correlation between results in the high jump and results in the long jump.

(4)

The product moment correlation coefficient between the height of the high jump and the length of the long jump for each competitor is found to be 0.678

(c) Use this value to test, at the 5% level of significance, for evidence of positive correlation between results in the high jump and results in the long jump.

(2)

(d) State the condition required for the test in part (c) to be valid.

(1)

(e) Explain what your conclusions in part (b) and part (c) suggest about the relationship between results in the high jump and results in the long jump.

(1)

(Total for question = 12 marks)

Q8.

In a gymnastics competition, two judges scored each of 8 competitors on the vault.

Competitor	A	B	C	D	E	F	G	H
Judge 1's scores	4.6	9.1	8.4	8.8	9.0	9.5	9.2	9.4
Judge 2's scores	7.8	8.8	8.6	8.5	9.1	9.6	9.0	9.3

(a) Calculate Spearman's rank correlation coefficient for these data.

(4)

(b) Stating your hypotheses clearly, test at the 1% level of significance, whether or not the two judges are generally in agreement.

(4)

(c) Give a reason to support the use of Spearman's rank correlation coefficient in this case.

(1)

The judges also scored the competitors on the beam.

Spearman's rank correlation coefficient for their ranks on the beam was found to be 0.952

(d) Compare the judges' ranks on the vault with their ranks on the beam.

(1)

(Total for question = 10 marks)

Q9.

Some students are investigating the strength of wire by suspending a weight at the end of the wire. They measure the diameter of the wire, d mm, and the weight, w grams, when the wire fails. Their results are given in the following table.

	These 14 points are plotted on page 13														Not yet plotted			
d	0.5	0.6	0.7	0.8	0.9	1.1	1.3	1.6	2	2.4	2.8	3.3	3.5	3.9	**4.5**	**4.6**	**4.8**	**5.4**
w	1.2	1.7	2.3	3.0	3.8	5.6	7.7	11.6	18	25.9	34.9	47.4	52.7	63.9	**81**	**83.6**	**89.9**	**109.4**

The first 14 points are plotted on the axes.

(a) On the axes complete the scatter diagram for these data.

(1)

(b) Use your calculator to write down the equation of the regression line of w on d.

(2)

(c) With reference to the scatter diagram, comment on the appropriateness of using this linear regression model to make predictions for w for different values of d between 0.5 and 5.4

(1)

The product moment correlation coefficient for these data is $r = 0.987$ (to 3 significant figures).

(d) Calculate the residual sum of squares (RSS) for this model.

(2)

Robert, one of the students, suggests that the model could be improved and intends to find the equation of the line of regression of w on u, where $u = d^2$
He finds the following statistics

$$S_{wu} = 5721.625 \qquad S_{uu} = 1482.619 \qquad \sum u = 157.57$$

(e) By considering the physical nature of the problem, give a reason to support Robert's suggestion.

(1)

(f) Find the equation of the regression line of w on u.

(3)

(g) Find the residual sum of squares (RSS) for Robert's model.

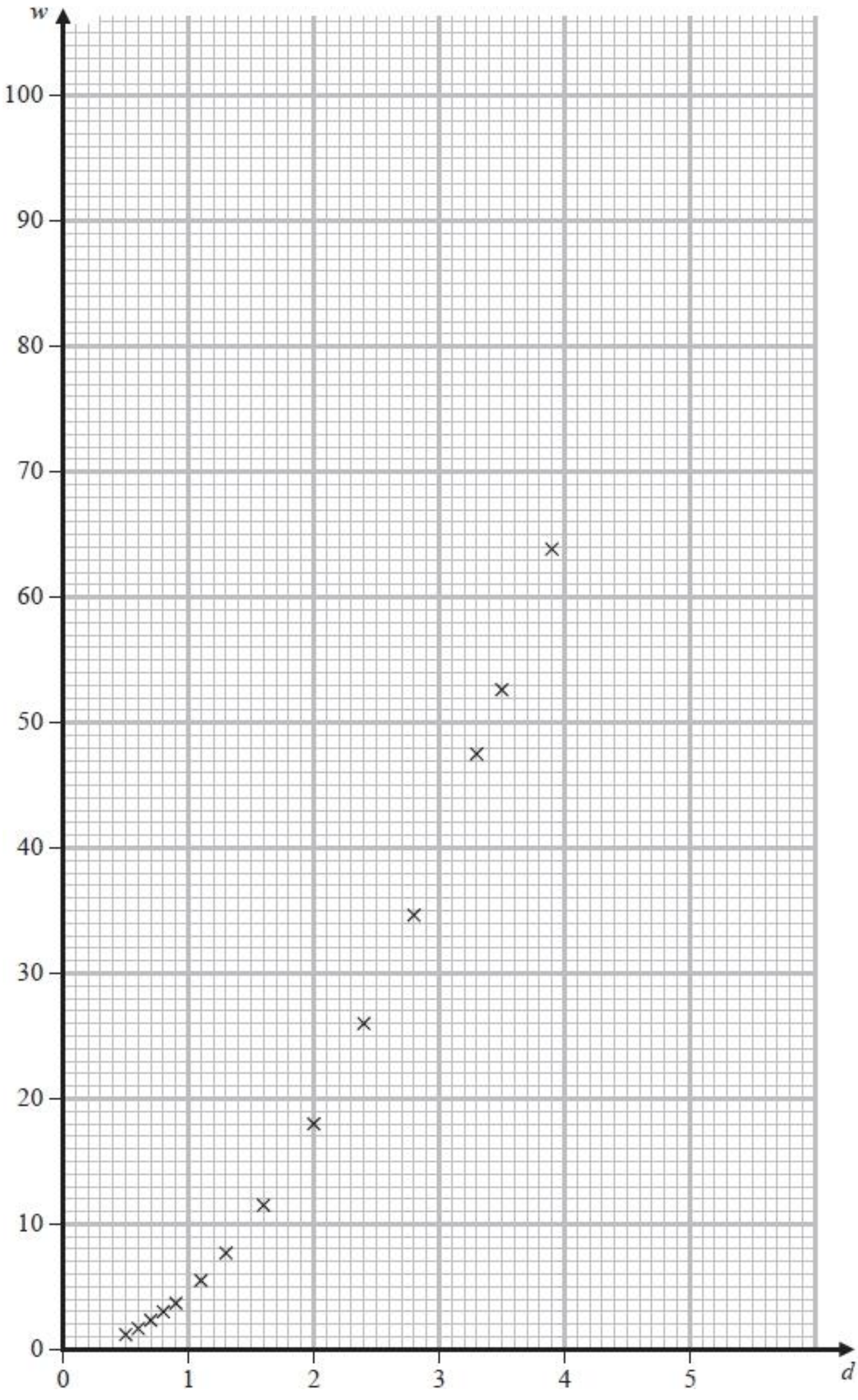
(2)

(h) State, giving a reason based on these calculations, which of these models better describes these data.

(1)

(i) Hence estimate the weight at which a piece of wire with diameter 3 mm will fail.

(1)



(Total for question = 14 marks)

Q10.

An estate agent in *Tornep* believes that houses further from the railway station are more expensive than those that are closer. She took a random sample of 22 three-bedroom houses in *Tornep* and calculated the product moment correlation coefficient between the house price and the distance from the station to be 0.3892

Stating your hypotheses clearly, use a 5% level of significance to test the estate agent's belief. State the critical region used in your test.

(Total for question = 3 marks)

Q11.

A junior judge is being trained by a senior judge to learn how to assess ice skaters. After the training, the judges each assess 6 ice skaters *A*, *B*, *C*, *D*, *E* and *F*. They each list them in order of preference with the best ice skater first. The results are shown in the table below.

Rank	1	2	3	4	5	6
Senior Judge	*A*	*B*	*D*	*C*	*F*	*E*
Junior Judge	*B*	*D*	*A*	*F*	*C*	*E*

(a) Calculate Spearman's rank correlation coefficient for these data.

(5)

(b) Test, at the 5% level of significance, whether or not there is evidence of a positive correlation between the rankings of the junior judge and the senior judge. State your hypotheses clearly.

(4)

(c) Comment on the effectiveness of the training delivered by the senior judge.

(1)

(Total for question = 10 marks)

Q12.

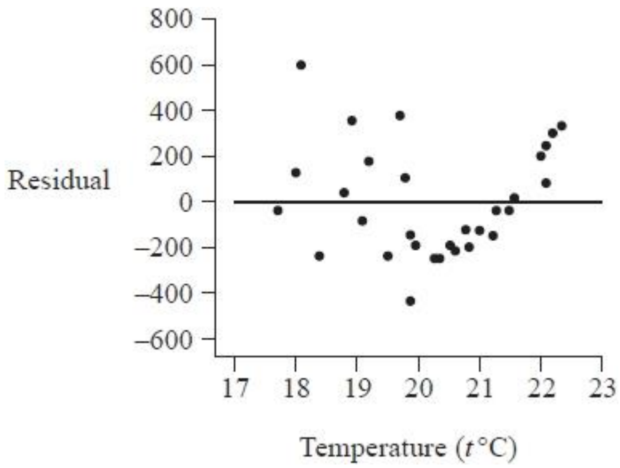
Kwame is investigating a possible relationship between average March temperature, t°C, and tea yield, y kg/hectare, for tea grown in a particular location.
He uses 30 years of past data to produce the following summary statistics for a linear regression model, with tea yield as the dependent variable.

Residual Sum of Squares (RSS) = 1 666 567 $\quad S_{tt} = 52.0 \quad S_{yy} = 1\ 774\ 155$
least squares regression line: gradient = 45.5 $\quad y$-intercept = 2080

(a) Use the regression model to predict the tea yield for an average March temperature of 20 °C

(1)

He also produces the following residual plot for the data.



(b) Explain what you understand by the term residual.

(1)

(c) Calculate the product moment correlation coefficient between t and y

(2)

(d) Explain why the linear model may not be a good fit for the data

(i) with reference to your answer to part (c)
(ii) with reference to the residual plot.

(2)

Kwame also collects data on total March rainfall, w mm, for each of these 30 years.

For a linear regression model of w on t the following summary statistic is found.

Residual Sum of Squares (RSS) = 86 754

Kwame concludes that since this model has a smaller RSS, there must be a stronger linear relationship between w and t than between y and t (where RSS = 1 666 567)

(e) State, giving a reason, whether or not you agree with the reasoning that led to Kwame's conclusion.

(1)

(Total for question = 7 marks)

Q13.

Mary, Jahil and Dawn are judging the cakes in a village show. They have 5 features to consider and each feature is awarded up to 5 points. The total score the judges gave each cake are given in the table below.

Cake	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Mary	19	17	23	10	21	15	12	8	14
Jahil	22	18	21	10	24	20	16	12	15
Dawn	9	11	6	18	9	15	13	20	13

(a) Calculate Spearman's rank correlation coefficient between Mary's scores and Jahil's scores.

(4)

(b) Calculate Spearman's rank correlation coefficient between Jahil's scores and Dawn's scores.

(3)

The judges discussed their interpretation of the points system and agreed that the first prize should go to cake *C*.

(c) Explain how different interpretations of the points system could give rise to the results in part (a) and part (b).

(2)

(Total for question = 9 marks)

Q14.

Abena and Meghan are both given the same list of 10 films.

Each of them ranks the 10 films from most favourite to least favourite.

For the differences, *d*, between their ranks for these 10 films, $\sum d^2 = 84$

(a) Calculate Spearman's rank correlation coefficient between Abena's ranks and Meghan's ranks.

(1)

A test is carried out at the 5% level of significance to see if there is agreement between their ranks for the films.

The hypotheses for the test are

$$H_0 : \rho_S = 0 \qquad\qquad H_1 : \rho_S > 0$$

(b) (i) Find the critical region for the test.

(ii) State the conclusion of the test.

(2)

An 11th film is added to the list. Abena and Meghan both agree that this film is their least favourite.

A new test is carried out at the 5% level of significance using the same hypotheses.

(c) Determine the conclusion of this test. You should state the test statistic and the critical value used.

(4)

(Total for question = 7 marks)

Q15.

Anisa is investigating the relationship between marks on a History test and marks on a Geography test. She collects information from 7 students. She wants to calculate the Spearman's rank correlation coefficient for the 7 students so she ranks their performance on each test.

Student	History mark	Geography mark	History rank	Geography rank
A	76	58	1	3
B	70	60	2	2
C	64	57	*s*	*t*
D	64	63	*s*	1
E	64	57	*s*	*t*
F	59	50	6	7
G	55	52	7	6

(a) Write down the value of *s* and the value of *t*

(2)

The full product moment correlation coefficient (pmcc) formula is used with the ranks to calculate the Spearman's rank correlation coefficient instead of $r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$ and the value obtained is 0.7106 to 4 significant figures.

(b) Explain why the full pmcc formula is used to carry out the calculation.

(1)

(c) Stating your hypotheses clearly, test whether or not there is evidence to suggest that the higher a student ranks in the History test, the higher the student ranks in the Geography test. Use a 5% level of significance.

(4)

(Total for question = 7 marks)

Q16.

Bara is investigating whether or not the two judges of a skating competition are in agreement. The two judges gave a score to each of the 8 skaters in the competition as shown in the table below.

	Skater							
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Judge 1	71	70	72	62	63	61	57	53
Judge 2	73	71	67	64	62	56	52	53

Bara decided to calculate Spearman's rank correlation coefficient for these data.

(a) Calculate Spearman's rank correlation coefficient between the ranks of the two judges.

(4)

(b) Test, at the 1% level of significance, whether or not the two judges are in agreement.

(4)

Judge 1 accidentally swapped the scores for skaters *D* and *E*. The score for skater *D* should be 63 and the score for skater *E* should be 62

(c) Without carrying out any further calculations, explain how Spearman's rank correlation coefficient will change. Give a reason for your answer.

(2)

(Total for question = 10 marks)

Q17.

Nine athletes, *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H* and *I*, competed in both the 100 m sprint and the long jump. After the two events the positions of each athlete were recorded and Spearman's rank correlation coefficient was calculated and found to be 0.85

(a) Stating your hypotheses clearly, test whether or not there is evidence to suggest that the higher an athlete's position is in the 100 m sprint, the higher their position is in the long jump. Use a 5% level of significance.

(4)

The piece of paper the positions were recorded on was mislaid. Although some of the athletes agreed their positions, there was some disagreement between athletes *B*, *C* and *D* over their long jump results.

The table shows the results that are agreed to be correct.

Athlete	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Position in 100 m sprint	4	6	7	9	2	8	3	1	5
Position in long jump	5				4	9	3	1	2

Given that there were no tied ranks,

(b) find the correct positions of athletes *B*, *C* and *D* in the long jump. You must show your working clearly and give reasons for your answers.

(5)

(c) Without recalculating the coefficient, explain how Spearman's rank correlation coefficient would change if athlete *H* was disqualified from both the 100 m sprint and the long jump.

(2)

(Total for question = 11 marks)

Q18.

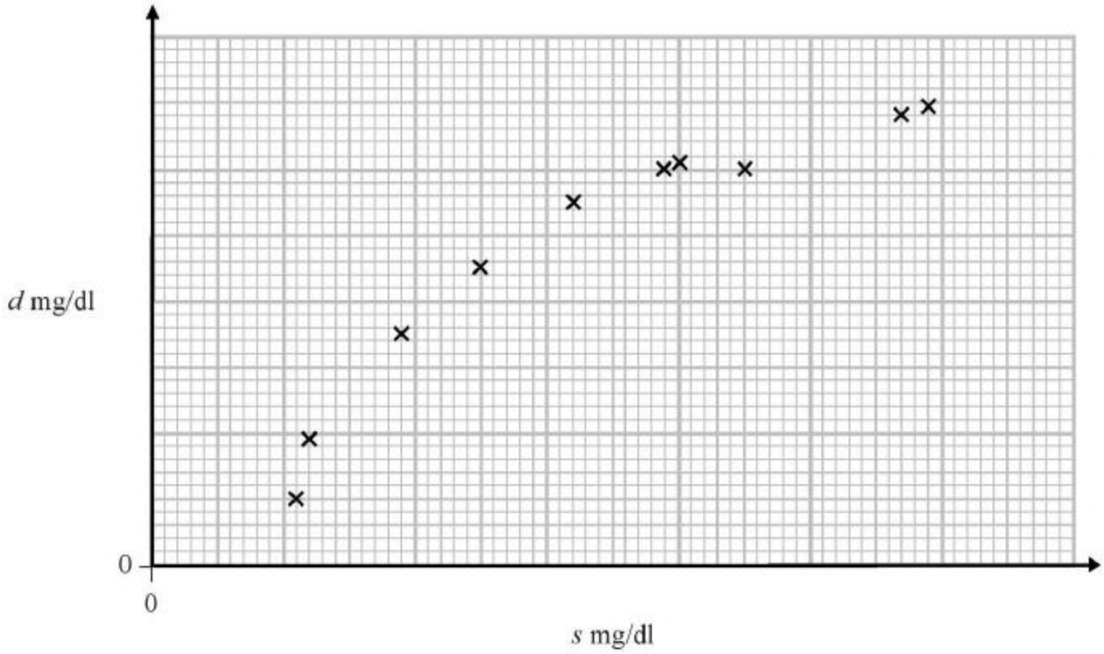
Over a period of time, researchers took 10 blood samples from one patient with a blood disease. For each sample, they measured the levels of serum magnesium, *s* mg/dl, in the blood and the corresponding level of the disease protein, *d* mg/dl. One of the researchers coded the data for each sample using $x = 10s$ and $y = 10(d - 9)$ but spilt ink over his work.

The following summary statistics and unfinished scatter diagram are the only remaining information.

$\sum d^2 = 1081.74$ $\quad S_{ds} = 59.524$

and

$\sum y = 64$ $\quad S_{xx} = 2658.9$



(a) Use the formula for S_{xx} to show that $S_{ss} = 26.589$

(3)

(b) Find the value of the product moment correlation coefficient between *s* and *d*.

(4)

(c) With reference to the unfinished scatter diagram, comment on your result in part (b).

(1)

(Total for question = 8 marks)

Q19.

A researcher claims that, at a river bend, the water gradually gets deeper as the distance from the inner bank increases. He measures the distance from the inner bank, *b* cm, and the depth of a river, *s* cm, at 7 positions. The results are shown in the table below.

Position	A	B	C	D	E	F	G
Distance from inner bank *b* cm	100	200	300	400	500	600	700
Depth *s* cm	60	75	85	76	110	120	104

The Spearman's rank correlation coefficient between *b* and *s* is $\frac{6}{7}$

(a) Stating your hypotheses clearly, test whether or not the data provides support for the researcher's claim. Use a 1% level of significance.

(4)

(b) Without re-calculating the correlation coefficient, explain how the Spearman's rank correlation coefficient would change if

(i) the depth for G is 109 instead of 104

(ii) an extra value H with distance from the inner bank of 800 cm and depth 130 cm is included.

(3)

The researcher decided to collect extra data and found that there were now many tied ranks.

(c) Describe how you would find the correlation with many tied ranks.

(2)

(Total for question = 9 marks)

Mark Scheme - Correlation

Q1.

Question	Scheme	Marks	AOs
(a)	$\left[S_{ll} = 26.2326 - \frac{16.06^2}{10} = 0.44024\right]$		
	$r = \frac{42.786}{\sqrt{9936.9 \times "0.44024"}}$	M1	1.1b
	$r = 0.64689...$ awrt 0.647	A1	1.1b
		(2)	
(b)	"0.647" coding has no effect on the pmcc	B1ft	1.1b
		(1)	
(c)	$l - 20 = 0.00431(w - 6) - 18.87$	M1	3.1a
	$l = 0.00431w + ...$	M1	1.1b
	$l = 0.00431w + 1.10414$	A1	1.1b
		(3)	
(d)	$l = 0.00431 \times 100 + 1.10 = 1.53$	B1ft	3.4
		(1)	
(e)	$\text{RSS} = "0.44024" - \frac{(42.786)^2}{9936.9}$ or $"0.44024"(1 - "0.647"^2)$	M1	1.1b
	RSS = 0.2560	A1	1.1b
		(2)	
(f)	(i) The points appear **randomly** scattered above and below zero giving us no reason to doubt the suitability of the linear model.	B1	3.5a
	(ii) There is a possible outlier that could be removed (and the regression line recalculated).	B1	3.5c
		(2)	
			(11 marks)

Notes
(a) **M1:** For a complete correct method to find r
A1: for awrt 0.647
(b) **B1ft:** stating their answer to part (a) and a correct reason
(c) **M1:** for use of a correct model. i.e. a correct expression for b
M1: for use of a correct model i.e. a correct expression (ft) for a
A1: for correct model $l = 0.00431w + 1.10$ with awrt 0.00431 and awrt 1.10
(d) **B1ft:** correct answer using their equation and $w = 100$ or using $t = 0.00431s - 18.87$ and $s = 94$ Allow awrt 1.53/1.54
(e) **M1:** for a correct expression for RSS
A1: awrt 0.256
(f) **B1:** For explaining why the model may be suitable. Allow randomly scattered around w (x) axis. Do not allow most residuals close to zero or not suitable as not randomly scattered.
B1: For explaining how the fit of the model might be improved.

Q2.

Question	Scheme	Marks	AOs
(a)	$S_{ww} = 13447 - \frac{303^2}{8} = 1970.875$		
	$r = \frac{269.5}{\sqrt{42 \times 1970.875}}$	M1	1.1b
	$r = 0.9367...$ awrt 0.937	A1	1.1b
		(2)	
(b)	As the amount of **fertiliser** increases the **yield** increases	B1	3.2a
		(1)	
(c)	$b = \frac{269.5}{42}$ $[= 6.41666...]$	M1	3.3
	$a = \frac{303}{8} - 'b'\frac{28}{8}$ $[= 15.41666...]$	M1	1.1b
	$w = 15.4 + 6.42f$	A1	1.1b
		(3)	
(d)	3.21 tonnes	B1ft	1.1b
		(1)	
(e)	The **residual** plot is close to an 'n' shape or the **residuals** appear not to be **randomly** scattered	M1	2.4
	The model in part(c) is unlikely to be suitable	A1	2.2b
		(2)	
(f)	Fit a curve rather than a line	B1	3.5c
		(1)	
			(10 marks)

Notes:		
(a)	**M1** **A1**	Complete correct method for finding r for awrt 0.937
(b)	**B1**	Correct contextual statement
(c)	**M1**	For use of a correct model ie a correct expression for b
	M1	For use of a correct model ie a correct expression (ft) for a
	A1	For a correct model $w = 15.4 + 6.42f$ with awrt 15.4 and awrt 6.42
(d)	**B1ft**	awrt 3.21 condone -3.21
(e)	**M1**	Explaining a reason for their conclusion eg there is **a pattern/trend** in the **residuals** Do not accept residuals not close to zero
	A1	concluding it is not valid oe
(f)	**B1**	A comment about not using a linear line eg use a quadratic model, logarithmic graph exponential

Q3.

Question	Scheme	Marks	AOs
(a)	$S_{ww} = 11386134 - 27(628.59)^2 [= 717\,748.5213]$	B1	1.1b
	$r = \dfrac{13\,082.3}{\sqrt{260.8 \times '717\,748.5213'}}$	M1	1.1b
	$r = 0.95618....$ awrt **0.956**	A1	1.1b
		(3)	
(b)	Since r is close to 1, data is consistent with a linear model.	B1	2.4
		(1)	
(c)	$b = \dfrac{13\,082.3}{260.8} [= 50.162...]$	M1	3.3
	$a = 628.59 - 'b'(31.07)$	M1	1.1b
	$w = -930 + 50.2x$	A1	1.1b
		(3)	
(d)	$w = -930 + 50.2(32) + 80$	M1	3.4
	$w = 756.4$	A1	1.1b
		(2)	
(e)	Negative residuals for all 5 observations with $x > 33$ suggests the model systematically overestimates weights for the longest bream.	B1	3.5a
		(1)	
			(10 marks)

Notes

(a)	**B1:** Correct expression for S_{ww} (implied by a correct answer) **M1:** Complete method to find r (Use of S_{ww} = 11 386 134 is M0) **A1:** awrt 0.956
(b)	**B1:** Correct explanation and conclusion
(c)	**M1:** Setting up linear model by finding gradient **M1:** Attempting y-intercept of linear model **A1:** Correct model with b = awrt 50.2 and a = awrt −930 (must use w and x)
(d)	**M1:** Using the model with the residual. Allow $\pm$ 80 **A1:** awrt 756 (allow awrt 755 from use of exact values)
(e)	**B1:** Evaluating the model for $x > 33$ (must reference both the residuals and the model) Negative correlation between residuals and length is B0.

Q4.

Question	Scheme	Marks	AOs
(a)	$[S_{pp} = 4748 - \frac{254^2}{16} = 715.75]$ $r = \frac{1115}{\sqrt{1846 \times ("715.75")}}$	M1	1.1b
	$r = 0.970014...$ awrt **0.970**	A1	1.1b
		(2)	
(b)	$b = \frac{1115}{1846}$ $[= 0.6040....]$	M1	3.3
	$a = \frac{254}{16} - "b"\frac{392}{16}$ $[=1.076...]$	M1	1.1b
	$p = 1.08 + 0.604m$	A1	1.1b
		(3)	
(c)	$\text{RSS} = "715.75" - \frac{1115^2}{1846}$ or $\text{RSS} = "715.75"(1 - "0.970"^2)$	M1	1.1b
	$\text{RSS} = 42.28033...$ awrt **42.3**	A1	1.1b
		(2)	
(d)	$p = 1.08 + 0.604(30) + \text{residual}$	M1	3.4
	$p = \underline{18}$	A1ft	1.1b
		(2)	
(e)	|residual| is large/may be an outlier	B1	3.5b
		(1)	
(f)	New r should be closer to 1 than part (a) since the remaining points are likely to be closer to the new regression line.	B1	2.2b
		(1)	
			(11 marks)

Notes	
(a)	M1 for a complete correct method for finding r A1 for awrt 0.970 (allow 0.97 from correct working)
(b)	1st M1 for use of a correct model i.e. a correct expression for b 2nd M1 for use of a correct model i.e. a correct (ft) expression for a A1 for correct model $p = 1.08 + 0.604m$ with awrt 1.08 and awrt 0.604 No fractions and must be in terms of p and m
(c)	M1 for a correct expression for RSS A1 for awrt 42.3
(d)	M1 for substitution of $m = 30$ into the regression equation and adding the residual A1ft for 18
(e)	B1 for identifying this point's residual is far from 0/it may be an outlier/anomaly/ does not fit the trend
(f)	B1 for closer to 1 than part (a) / increase oe and correct supporting reason about the relative strength of correlation (condone outlier removed)

Q5.

Question	Scheme	Marks	AOs
(a)	$r = \dfrac{284.4 - \frac{251(12)}{10}}{\sqrt{10.36 \times 40.9}}$	M1	1.1b
	$r = -0.79671...$ awrt **$\underline{-0.797}$**	A1	1.1b
		(2)	
(b)	$b = \dfrac{\text{'}-16.4\text{'}}{10.36}$	M1	3.3
	$a = \dfrac{251}{10} - \text{'}b\text{'}\dfrac{12}{10}$	M1	1.1b
	$y = 27.0 - 1.58x$	A1	1.1b
		(3)	
(c)	$y = [27.0 - 1.58(2)] = 23.84$ awrt **$\underline{23.8}$**	B1ft	3.4
		(1)	
(d)	$\text{RSS} = 40.9 - \dfrac{(-16.4)^2}{10.36}$	M1	1.1b
	$\text{RSS} = 14.938...$ awrt **$\underline{14.9}$**	A1	1.1b
		(2)	
(e)	$\sum \text{residuals} = 0 \rightarrow -0.63 + (-0.32) + \ldots + f + (-1.88) = 0$	M1	3.1a
	$f = \underline{\mathbf{-1.04}}$	A1	1.1b
		(2)	
(f)	The residuals should be randomly scattered above and below zero so linear model may not be appropriate	B1	3.5b
		(1)	
		(11 marks)	

Notes	
(a)	M1 for a complete correct method for finding r A1 for awrt -0.797
(b)	1st M1 for use of a correct model i.e. a correct expression for b (ft their S_{xy}) 2nd M1 for use of a correct model i.e. a correct (ft) expression for a A1 for $y = 27.0 - 1.58x$ [a correct answer here can imply both method marks]
(c)	B1ft for awrt 23.8 (evaluating their model found in part (b) with $x = 2$)
(d)	M1 for a correct expression for RSS A1 for awrt 14.9
(e)	M1 for use of $\sum \text{residuals} = 0$ [Use of regression equation needs correct sign] A1 for -1.04
(f)	B1 for identifying that the residuals are not randomly scattered above and below zero and concluding the linear regression model may not be appropriate.

Q6.

Question Number	Scheme	Marks
(a)	$[S_{wt}] = 784 - \frac{119 \times 42}{6} =, \ \underline{-49}$	A1
	$[S_{tt}] = 2435 - \frac{119^2}{6} =, \quad 74.8\dot{3}$ or $74\frac{5}{6}$ or $\frac{449}{6}$ (accept awrt $\underline{74.8}$)	M1 A1
		(3)
(b)	$S_{ss} = 5 \times 10^7$ or **50 000 000** (o.e.)	B1
	$S_{st} = \underline{\mathbf{-49\ 000}}$	B1ft
		(2)
(c)	$r = \frac{"-49"}{\sqrt{50 \times "74.8\dot{3}"}}$ or $\frac{"-49\ 000"}{\sqrt{"5 \times 10^7" \times "74.8\dot{3}"}} =, \ -0.80105... =$ awrt $\underline{\mathbf{-0.801}}$	M1, A1
		(2)
(d)	r is close to -1 or $\lvert r \rvert$ is close to 1 or "**strong**" (o.e.) [negative] **correlation** … so "**yes**" or **does** support the belief	B1ft
		(1)
(e)	$b = \frac{"-49"}{"74.8\dot{3}"} = [-0.6547...], \ a = \frac{42}{6} - b \times \frac{119}{6} = [19.9866...]$ or $a = 7 - b \times 19.8\dot{3}$	M1, M1
	So $\underline{\mathbf{w = 20.0 - 0.655t}}$	A1
		(3)
(f)	$\underline{\mathbf{s = 20\ 000 - 655t}}$ or $\underline{c = 20\ 000 \text{ and } d = -655}$	B1ft B1ft
		(2)
(g)	**Decrease** in sales of [£] 655 (ignore any minus sign)	B1ft
		(1)
		[14]

	Notes
(a)	M1 for a correct expression for S_{wt} or S_{tt} (May be implied by either correct answer) 1st A1 for $[S_{wt}] = -49$ 2nd A1 for $[S_{tt}] =$ awrt 74.8 **SC** If both values correct but clearly mislabelled award M1A0A1
(b)	2nd B1ft for multiplying their S_{wt} by 1000
(c)	M1 for a correct expression using their values provided S_{tt} and S_{ss} both > 0 A1 for awrt -0.801 (Correct ans. only M1A1, -0.80 with no working M1A0)
(d)	B1ft for a correct comment that uses their value of r as support, provided $0.5 \leqslant \lvert r \rvert \leqslant 1$ For $\lvert r \rvert < 0.5$ comment must be "does not support", because "weak" (o.e.) correlation. **NB** "points lie close to a straight line" is B0 unless supported by mention of their value of r
(e)	1st M1 for a correct expression for b or awrt -0.66 or -0.65 Ft their answers from (a) 2nd M1 for a correct expression for a ft their value for b A1 for a correct equation in w and t only with $a = 20$ or awrt 20.0 and $b =$ awrt -0.655 (No fractions) **If their a and b are given to more than 3 sf, accept answers in (f) to 3sf or better.**
(f)	1st B1 ft for correct c or "their 20.0" $\times 1000$ 2nd B1ft for correct d or their "-0.655" $\times 1000$ Values can be in an s, t eq'n or $c =, d =$ (Their a and b needn't be to 3 sf and ft their letter for t)
(g)	B1ft for stating clearly both decrease (o.e.) and [£] 655. Ft their d and allow "increase" if $d > 0$

Q7.

<table>
<tr><th>Question</th><th>Scheme</th><th>Marks</th><th>AOs</th></tr>
<tr><td>(a)</td><td><table><tr><td>Competitor</td><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td><td>F</td></tr><tr><td>High jump ranks</td><td>1</td><td>5</td><td>2.5</td><td>4</td><td>6</td><td>2.5</td></tr><tr><td>Long jump ranks</td><td>2</td><td>5</td><td>1</td><td>4</td><td>6</td><td>3</td></tr></table></td><td>M1
M1</td><td>1.1b
1.1b</td></tr>
<tr><td></td><td>$[\sum h^2 = 90.5 \quad \sum l^2 = 91 \quad \sum hl = 89]$</td><td></td><td></td></tr>
<tr><td></td><td>Use of pmcc $\quad r_s = \dfrac{89 - \frac{21\times21}{6}}{\sqrt{(90.5-\frac{21^2}{6})(91-\frac{21^2}{6})}}$</td><td>M1</td><td>1.1b</td></tr>
<tr><td></td><td>$r =$ awrt <u>0.899</u></td><td>A1</td><td>1.1b</td></tr>
<tr><td></td><td></td><td>(4)</td><td></td></tr>
<tr><td>(b)</td><td>H0: $\rho_s = 0$ H1: $\rho_s > 0$</td><td>B1</td><td>2.5</td></tr>
<tr><td></td><td>Critical value $\rho_s = 0.8286$</td><td>B1</td><td>1.1b</td></tr>
<tr><td></td><td>$r_s = 0.899$ lies in the critical region/reject H0</td><td>M1</td><td>2.1</td></tr>
<tr><td></td><td>There is <u>positive</u> rank correlation between <u>high jump</u> and <u>long jump</u> results.</td><td>A1ft</td><td>2.2b</td></tr>
<tr><td></td><td></td><td>(4)</td><td></td></tr>
<tr><td>(c)</td><td>[H0: $\rho = 0$ H1: $\rho > 0$]
$0.678 <$ Critical value $\rho = 0.7293$</td><td>M1</td><td>2.1</td></tr>
<tr><td></td><td>There is no evidence of (positive) correlation (between high jump and long jump).</td><td>A1</td><td>2.2b</td></tr>
<tr><td></td><td></td><td>(2)</td><td></td></tr>
<tr><td>(d)</td><td>The test in part (c) requires the data to come from a bivariate normal distribution.</td><td>B1</td><td>2.3</td></tr>
<tr><td></td><td></td><td>(1)</td><td></td></tr>
<tr><td>(e)</td><td>Although there is evidence of a positive correlation between the ranks, the data does not appear to fit a linear pattern.</td><td>B1</td><td>2.4</td></tr>
<tr><td></td><td></td><td>(1)</td><td></td></tr>
<tr><td colspan="4" align="right">(12 marks)</td></tr>
</table>

Notes	
(a)	1st M1 for an attempt to rank first row **using tied ranks** (at least 4 correct) 2nd M1 for an attempt to rank second row (at least 4 correct) 3rd M1 for use of pmcc with tied ranks A1 for awrt 0.899 SC: Use of Spearman with Σd^2 for their ranks may score M1M1M1A0
(b)	1st B1 both hypotheses stated in terms of ρ_s or ρ 2nd B1 for correct critical value M1 for comparing their '0.8286' with their '0.899' A1ft for a correct contextual conclusion (may ft their r_s)
(c)	M1 for comparing 0.7293 with 0.678 A1 for a correct conclusion
(d)	B1 for explaining the required condition for the pmcc test to be used.
(e)	B1 for comparing what each coefficient shows

Q8.

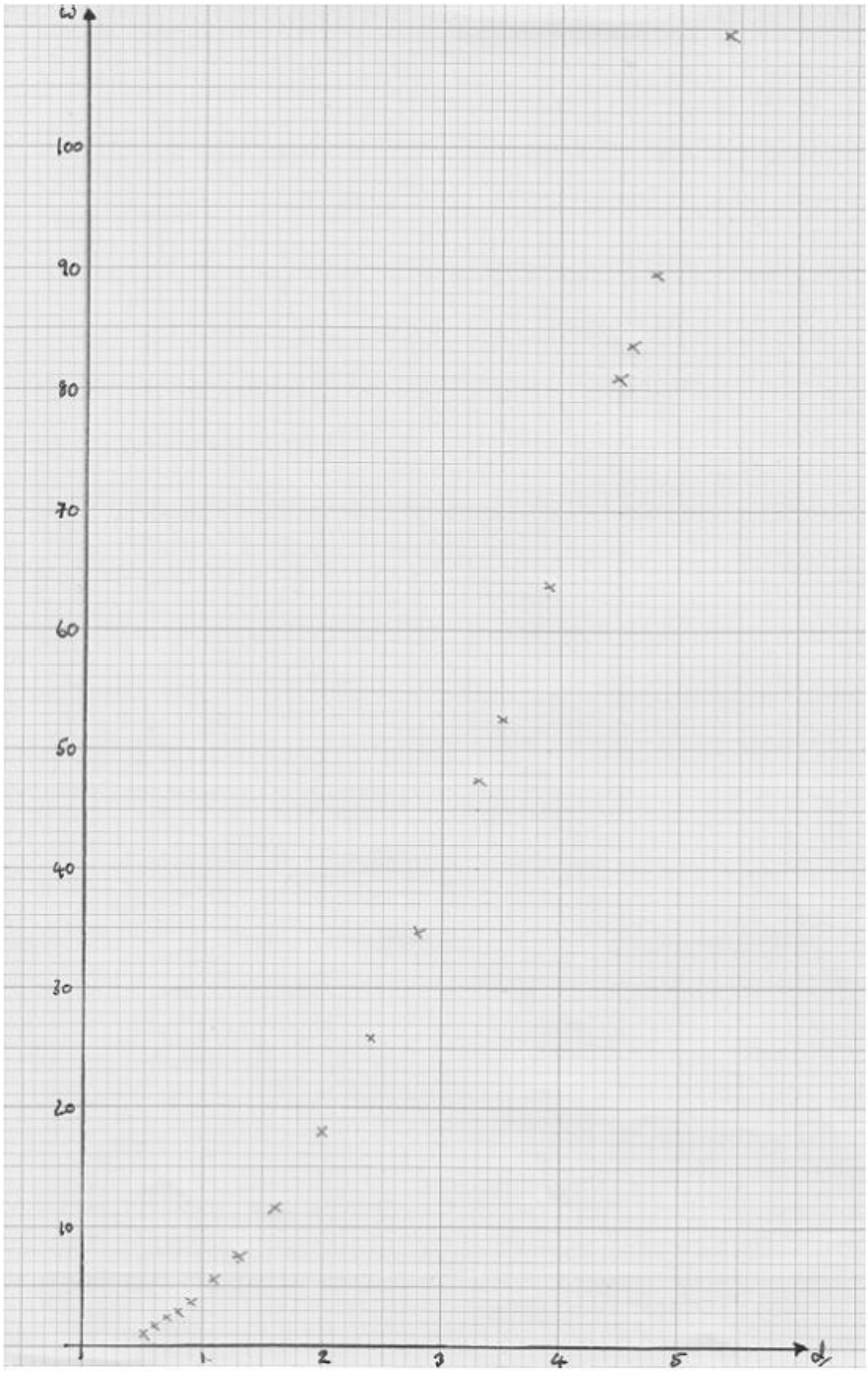
Question	Scheme	Marks	AOs
(a)	<table><tr><td>**Competitor**</td><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td><td>F</td><td>G</td><td>H</td></tr><tr><td>**Judge 1's ranks**</td><td>8</td><td>4</td><td>7</td><td>6</td><td>5</td><td>1</td><td>3</td><td>2</td></tr><tr><td>**Judge 2's ranks**</td><td>8</td><td>5</td><td>6</td><td>7</td><td>3</td><td>1</td><td>4</td><td>2</td></tr><tr><td>d^2</td><td>0</td><td>1</td><td>1</td><td>1</td><td>4</td><td>0</td><td>1</td><td>0</td></tr></table>	M1 M1	1.1b 1.1b
	$\sum d^2 = 8$		
	$r_s = 1 - \dfrac{6 \times 8}{8(64-1)}$	dM1	1.1b
	$r_s = 0.90476\ldots$ awrt <u>0.905</u>	A1	1.1b
		(4)	
(b)	H0: $\rho_s = 0$ H1: $\rho_s > 0$	B1	2.5
	Critical value $\rho_s = 0.8333$	B1	1.1b
	$r_s = 0.905$ lies in the critical region/reject H0	M1	2.1
	The two judges are in agreement.	A1	2.2b
		(4)	
(c)	e.g. The data is unlikely to be from a bivariate normal distribution (competitor A)/The emphasis here is on the ranks and not the individual scores.	B1	2.4
		(1)	
(d)	Both show positive correlation, but the judges agree more on the beam (since 0.952 is closer to 1)	B1	2.2b
		(1)	
			(10 marks)

Notes	
(a)	1st M1 for an attempt to rank at least one row (at least four correct) 2nd M1 for an attempt at d^2 row for their ranks 3rd M1 dependent on 1st M1 for use of $r_s = 1 - \dfrac{6 \times 8}{8(64-1)}$ with their $\sum d^2$ A1 for awrt 0.905
(b)	1st B1 both hypotheses stated in terms of ρ_s 2nd B1 for correct critical value M1 for comparing their '0.905' with their '0.8333' A1 for a correct contextual conclusion with no contradictions seen
(c)	B1 for a correct explanation to support the use of Spearman
(d)	B1 for a correct comparison of the correlation coefficients

Q9.

Qu	Answer	Marks	AO
(a)	Use overlay. All correct	B1 (1)	1.1b
(b)	Need to choose model of the form: $w = a + bd$ and have one of a or b correct to 2 sf	M1	3.3
	$\underline{w = 21.5d - 17.7}$	A1 (2)	1.1b
(c)	Not appropriate because eg the line is plotted and not close to the points or two lines with different gradients or overestimates values in the middle and underestimates the others or the points are more curved	B1 (1)	3.5a
(d)	$\left\{S_{ww} = \sum w^2 - \frac{\left(\sum w\right)^2}{18} = 45178.68 - \frac{643.6^2}{18}\right\} = 22166.404..$	M1	1.1b
	RSS = $S_{ww}\left(1-r^2\right) = 22166.404\ldots \times (1-0.987^2)$ = awrt $\underline{570}$ (g^2)	A1 (2)	1.1b
(e)	Thicker wire should be stronger and strength is proportional to area (i.e. d^2)	B1 (1)	2.4
(f)	$w = cu + f$ where $c = \frac{5721.625}{1482.619} = 3.85913\ldots$	M1	3.3
	$f\{=\bar{w} - c\bar{u}\} = \frac{"643.6"}{18} - "3.8591..." \times \frac{157.57}{18}$ $\{= 1.973\ldots\}$	M1	1.1b
	$\underline{w = 1.97 + 3.86u}$	A1 (3)	1.1b
(g)	RSS $= S_{ww} \times (1-r^2)$ or $S_{ww} - \frac{(S_{wu})^2}{S_{uu}}$, $= 85.8824\ldots$ awrt $\underline{85.9}$ (g^2)	M1, A1 (2)	1.1b (x2)
(h)	Robert's model is better since RSS is reduced	B1 (1)	2.4
(i)	Use Robert's model: w $\{= 3.859 \times 3^2 + 1.973\}$ = awrt $\underline{36.7}$	B1 (1)	3.4
		(14 marks)	

	Notes
(a)	1st B1 for fully correct scatter diagram
(b)	M1 for selecting the appropriate model and one coefficient correct to 2sf A1 for b = awrt 21.5 and a = awrt -17.7
(c)	B1 for comment suggesting not very good with a suitable reason.
(d)	M1 for calculation of S_{ww} or any other terms needed for their calculation A1 for RSS = 570.3299… i.e. awrt 570
(e)	B1 for a comment realising that strength is proportional to d^2 (area)
(f)	1st M1 for using correct expression for gradient 2nd M1 for correct expression for intercept A1 for correct line with coefficients awrt 3 sf
(g)	M1 for a correct expression (ft their S_{ww}) [NB r = awrt 0.998]
(h)	B1 for comment about reduced RSS (RSS needs to be lower but needn't be correct)


w
100
90
80
70
60
50
40
30
20
10
1
2
3
4
5
d

Q10.

Qu	Answer	Marks	AO
	$H_0: \rho = 0 \quad H_1: \rho > 0$	B1	2.5
	5% one-tail cv for r is: 0.3598	M1	1.1b
	Significant result so there is evidence to support the agent's belief	A1	2.2b
			(3 marks)
	Notes		
	B1 for both hypotheses correct in terms of ρ M1 for use of tables to find the cv of 0.3598 A1 for a correct conclusion in context mentioning " belief " or description of this		

Q11.

Question Number	Scheme	Notes	Marks
(a)	<table><tr><td></td><td>Senior Judge</td><td>Junior Judge</td><td>d</td><td>d^2</td></tr><tr><td>A</td><td>1</td><td>3</td><td>-2</td><td>4</td></tr><tr><td>B</td><td>2</td><td>1</td><td>1</td><td>1</td></tr><tr><td>C</td><td>4</td><td>5</td><td>-1</td><td>1</td></tr><tr><td>D</td><td>3</td><td>2</td><td>1</td><td>1</td></tr><tr><td>E</td><td>6</td><td>6</td><td>0</td><td>0</td></tr><tr><td>F</td><td>5</td><td>4</td><td>1</td><td>1</td></tr><tr><td></td><td></td><td></td><td></td><td>8</td></tr></table>	M1 for an attempt to rank judges lists (at least 4 correct for each judge) A1 for correct rankings for both (may be reversed). Can be implied by correct d^2 or r_s N.B. Table could be ordered in terms of Senior Judge.	M1A1
	$\sum d^2 = 8$ or 62	8 or 62 or correct d^2 row	A1
	$r_s = 1 - \frac{6 \times 8}{6 \times 35} = \frac{27}{35} = 0.771$	M1 for use of the correct formula, follow through their $\sum d^2$ (Dependent on 1st M1) If answer is not correct, a correct expression is required. A1 exact fraction or awrt $(\pm)0.771$	dM1A1
			(5)
(b)	$H_0: \rho = 0$ $H_1: \rho > 0$	Both hypotheses in terms of ρ or ρ_s . Hypotheses just in words e.g. "no correlation" score B0	B1
	Critical value ± 0.8286	Accept ± 0.8857 if 2-tailed H_1.	B1
	(0.771<0.8286) so insufficient evidence to reject H_0	Follow through their r_s and their c.v. if $\lvert cv \rvert < 1$ and $\lvert r_s \rvert < 1$	M1
	There is insufficient evidence to suggest a **positive** correlation between the judges.	A correct contextualised comment that includes "judges".	A1 ft
			(4)
(c)	(For positive correlation c.v.is 0.8286>0.771)		
	Training of junior judge was ineffective.	o.e. Follow through from their cv and r_s	B1 ft
			(1)
			Total 10

Q12.

Question	Scheme	Marks	AOs
(a)	$[20 \times 45.5 + 2080] = 2990$ [kg/ha]	B1	3.4
		(1)	
(b)	(A residual is the) difference between the observed value (oe) and the predicted value (oe) (of the dependent variable)	B1	1.2
		(1)	
(c)	$1666567 = 1774155(1 - r^2)$	M1	1.1b
	$r = 0.246...$ awrt 0.246	A1	1.1b
		(2)	
(d)(i)	Since r is close to 0/weak correlation	B1	2.4
(ii)	e.g. (For $t > 20$…) the residuals do not appear randomly scattered about 0.	B1	3.5a
		(2)	
(e)	Kwame's conclusion cannot be supported using RSS since the two values of RSS do not have the same units.	B1	2.3
		(1)	
			(7 marks)

Notes	
(a)	**B1:** cao
(b)	**B1:** Correct definition. Allow equivalent wording. Distance from regression line on its own is B0, but allow if vertical distance or y is referenced
(c)	**M1:** Use of correct expression for r or r^2 Allow use of $S_{ty} = 45.5 \times 52.0$ [=2366] or RSS: $1774155 - \frac{(S_{ty})^2}{52} = 1666567 \rightarrow [S_{ty} = 2365.2...]$ and then $r = \frac{\text{awrt } 2365 \text{ or awrt } 2366}{\sqrt{52.0 \times 1774155}}$ **A1:** awrt 0.246 (-0.246 or ± 0.246 scores M1A0)
(d)(i) (ii)	**B1:** Correct explanation **B1:** Correct evaluation of the fit of the model's residuals (e.g. variance either side of $t = 20$ does not appear to be the same) 'residuals not randomly scattered' on its own is B0.
(e)	**B1:** Correct assessment of the conclusion involving the units/size of the variables used to calculate the RSS

Q13.

Qu	Answer	Marks	AO
(a)	Cake: A B C D E F G H I Mary: 3 4 1 8 2 5 7 9 6 Jahil: 2 5 3 9 1 4 6 8 7	M1	1.1b
	$\sum d^2 = 1+1+4+1+1+1+1+1+1$ [= 12]	M1	1.1b
	$r_s = 1 - \frac{6\times\text{"12"}}{9\times 80}$	M1	1.1b
	$= \underline{0.9}$	A1	1.1b
		(4)	
(b)	Cake: A B C D E F G H I Jahil: 2 5 3 9 1 4 6 8 7 Dawn: 7.5 6 9 2 7.5 3 4.5 1 4.5	M1	1.1b
	$S_{JJ} = 60$ $S_{DD} = 59$ $S_{JD} = 176 - \frac{45^2}{9} = -49$ (o.e.)	M1	2.1
	$r_s = 0 - 0.823558\ldots$ awrt $\underline{-0.824}$	A1	1.1b
		(3)	
(c)	Mary and Jahil gave points for good features or high score is good Dawn gave points for poor features or low score is good	B1	2.4
	Both strong correlation, M&J positive, J&D negative so agree	B1	2.4
		(2)	
			(9 marks)

Notes

(a)	1st M1 an attempt to rank both – one row with at least 6 correct
	2nd dM1 (dep on an attempt at finding ranks) for an attempt to find $\sum d^2$ (some correct d values found and sum attempted)
	3rd M1 for using their $\sum d^2$ in formula for r_s with $n = 9$ (Independent of ranking)
	2nd A1 for 0.9 or exact fraction e.g. $\frac{9}{10}$
NB	No ranking leads to $\sum d^2 = 81$ and $r_s = 0.675$ and (a) M0M0M1A0 (b) M0M0A0
(b)	1st M1 for ranking Dawn's results and dealing with tied ranks 2nd M1 for selecting appropriate method to find r_s sight of 2 of these values or implied by ans A1 for using their calculator to evaluate r_s allow awrt -0.824
(c)	1st B1 for idea that M and J gave points for good features but D for bad features 2nd B1 for explaining that since both correlations are strong, one +, one − they agree

Q14.

Question	Scheme	Marks	AOs
(a)	$\left[r_s = 1 - \frac{6\times 84}{10(10^2-1)}\right]$ $= 0.4909...$ awrt **0.491**	B1	1.1b
		(1)	
(b)(i)	$(1>)r_s > 0.5636$	B1	1.1b
(ii)	(0.491 is not in the critical region) There is insufficient evidence of agreement between their film ranks.	B1ft	2.2b
		(2)	
(c)	new $r_s = 1 - \frac{6\times 84}{11(11^2-1)}$	M1	1.1b
	$= 0.61818...$ awrt **0.618**	A1	1.1b
	new critical value is 0.5364	B1	1.1b
	There is now sufficient evidence of agreement between their film ranks.	A1	2.2b
		(4)	
			(7 marks)

Notes	
(a)	**B1:** awrt 0.491 (allow $\frac{27}{55}$)
(b)(i) (ii)	**B1:** Correct critical region with 0.5636 or better. Condone use of ρ instead of r_s **B1ft:** Correct ft contextualised conclusion (must include film or ranks) based on their (a) and their CR Allow ft on their CV if a CR is not stated
(c)	**M1:** Use of formula with same $\sum d^2$ and 11 **A1:** awrt 0.618 (allow $\frac{34}{55}$) **B1:** 0.5364 or better **A1:** fully correct solution with awrt 0.618 and contextualised conclusion (must include film or ranks)

Q15.

Question	Scheme	Marks	AOs
(a)	$s = 4$	B1	1.1b
	$t = 4.5$	B1	1.1b
		(2)	
(b)	Because there are tied ranks.	B1	2.4
		(1)	
(c)	$\text{H}_0: \rho_s = 0 \quad \text{H}_1: \rho_s > 0$	B1	2.5
	CV = 0.7143	B1	1.1b
	$r_s = 0.7106$ does not lie in the critical region.	M1	2.1
	There is insufficient evidence to suggest that the higher the rank in the History test, the higher the rank in the Geography test (oe).	A1	2.2b
		(4)	
			(7 marks)

Notes	
(a)	**B1:** cao **B1:** cao
(b)	**B1:** Correct explanation
(c)	**B1:** Both hypotheses correct with correct notation (must use ρ_s or ρ) **B1:** Correct critical value 0.7143 or better **M1:** Drawing a correct inference using their CV and 0.7106 **A1:** Drawing a correct inference (condone "marks" instead of ranks) in context using their CV and 0.7106

Q16.

Question	Scheme	Marks	AOs
	(see ranking table below)	M1 dM1	1.1b 1.1b
	$\sum d^2 = 10$		
	$r_s = 1 - \frac{6 \times "10"}{8(64-1)}$	M1	1.1b
	$r_s = 0.8809...$ awrt 0.881	A1	1.1b
		(4)	
(b)	$H_0: \rho = 0 \quad H_1: \rho > 0$	B1	2.5
	Critical Value $\rho = 0.8333$	B1	1.1b
	$r_s = 0.8809$ lies in the critical region/reject H_0/significant	M1	2.1
	There is evidence that the two judges are in agreement.	A1cso	2.2b
		(4)	
(c)	The $\sum d^2$ will decrease since the new rankings given by Judge 1 are now the same as the rankings given by Judge 2 for Skater D and E whereas previously they were different	M1	2.4
	therefore Spearman's rank correlation coefficient will increase	A1	2.2a
		(2)	
		(10 marks)	

Skater / Judge	A	B	C	D	E	F	G	H
1	2	3	1	5	4	6	7	8
2	1	2	3	4	5	6	8	7
d^2	1	1	4	1	1	0	1	1

Notes

(a) **M1:** For an attempt to rank at least one row (at least 4 correct)

dM1: dep on previous M mark being awarded. For an attempt at d or d^2 row for their ranks.

M1: for use of $1 - \frac{6 \times "\text{their} \sum d^2"}{8(64-1)}$ Allow if not ranked $\sum d^2 = 85$

A1: awrt 0.881

(b) **B1:** Both hypotheses stated in terms of ρ

B1: for correct critical value. Allow even if 2 tail test (sign must match their r_s)

M1: for comparing their 0.881 with "their 0.8333"

A1cso: All previous marks awarded. For a correct contextual conclusion with no contradictions seen

(c) **M1:** For a correct explanation to support their answer given.

$\sum d^2$ decreases or d/d^2 decreases for D and E

and idea of same rankings eg d^2 will reduce by 2 . Do not allow d^2 will reduce by 1

A1: for a correct deduction from the information. Allow closer to 1.

Q17.

Question	Scheme	Marks	AOs
(a)	H0: $\rho_s = 0$ H1: $\rho_s > 0$	B1	2.5
	CV = 0.6	B1	1.1b
	$r_s = 0.85$ does lie in the critical region	M1	2.1
	There is evidence to suggest that there is a relationship between the position in the 100m sprint and the position in the long jump.	A1	2.2b
		(4)	
(b)	$1-\frac{6\sum d^2}{9(80)} = 0.85$	M1	3.1b
	$\sum d^2 = 18$	A1	1.1b
	$\sum d^2$ needed is '18' $- 15 = 3$	M1	1.1b
	Since $\sum d^2 = 3$ for the 3 missing places each place must contribute 1, therefore *B* must be in position 5 or 7. However, 5 has already been used so they must be position 7	A1	2.2a
	C is 6th and *D* is 8th	A1	2.2a
	SC B7, C6, D8 with no reasons B1 marks as final A1 on epen		
		(5)	
(c)	The $\sum d^2$ will not change but the value of *n* will decrease therefore	M1	2.4
	Spearman's rank correlation will decrease	A1	2.2a
		(2)	
		(11 marks)	

Notes:
(a)B1: Both hypotheses correct written using the notation r
B1: awrt 0.6
M1: Drawing a correct inference using their CV and the value of r_s
A1: Drawing a correct inference in context using their CV and the value of r_s
(b)M1: For realising they need to equate $1-\frac{6\sum d^2}{9(80)}$ to 0.85 to enable them to find the $\sum d^2$
A1: 18
M1: for $\sum d^2 = 3$
A1: For using the information in the question with the value for $\sum d^2$ to deduce that each must contribute 1 to the $\sum d^2$ and explain why *B* must be in position 7
A1: C 6th D 8th
(c)M1: Complete explanation why it decreases
A1: using the information given to deduce that it decreases

Q18.

Question	Scheme	Marks	AOs
(a)	$S_{xx} = \sum(10s)^2 - \frac{\left(\sum 10s\right)^2}{10}$	M1	2.1
	$2658.9 = 100\sum(s)^2 - \frac{100\left(\sum s\right)^2}{10}$	M1	1.1b
	$2658.9 = 100\ S_{ss}$		
	$S_{ss} = 26.589$ *	A1*cso	1.1b
		(3)	
(b)	$64 = \sum_1^{10} 10(d_i - 9)$	M1	3.1a
	$64 = 10\sum_1^{10} d_i - 900$		
	$\sum_1^{10} d_i = 96.4$	A1	1.1b
	$S_{dd} = 1081.74 - \frac{(\text{"96.4"})^2}{10}$	M1	1.1b
	$= 152.444$		
	$r = 0.935$	A1ft	1.1b
		(4)	
(c)	Linear correlation is significant but scatter diagram suggests a non-linear relationship between the level of serum magnesium, and the level of the disease protein	B1	3.5a
		(1)	
			(8 marks)

	Notes
(a)	**M1**: Attempting to use $S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{10}$ with $x = 10s$ **M1**: Substituting in 2658.9 and dealing with the 10 correctly **A1***: cso A complete solution with no errors leading to 26.589 only
(b)	**M1**: Realising that either $64 = \sum_1^{10} 10(d_i - 9)$ or $64 = 10\sum_1^{10} d_i - 900$ o.e. must be used. May be implied by seeing 96.4 **A1**: 96.4 only **M1**: Attempting to use $S_{dd} = \sum d^2 - \frac{\left(\sum d\right)^2}{10}$ may be implied by 0.935 **A1ft**: awrt 0.935 ft "their 96.4"
(c)	**B1**: A correct comment comparing their value of r and the scatter diagram in context

Q19.

Question	Scheme	Marks	AOs
(a)	$H_0 : \rho = 0, H_1 : \rho > 0$	B1	2.5
	Critical value at 1% level is 0.8929	B1	1.1b
	$r_s < 0.8929$ so not significant evidence to reject H_0 ,	M1	2.1
	The researcher's claim is not correct (at 1% level). or insufficient evidence for researcher's claim or there is insufficient evidence that water gets deeper further from inner bank. or no (positive) correlation between depth of water and distance from inner bank	A1ft	2.2b
		(4)	
(b)(i)	The **ranks will remain the same** therefore there will be **no change** to the spearman's rank correlation coefficient	**B1**	2.4
(ii)	Spearman's rank correlation coefficient will **increase** since	**B1**	2.2a
	The **ranks are the same** for both distance and depth therefore $d = 0$ however, ***n*** **has increased** **or** the new position follows the pattern that large b is assosciated with large s and so r_s will increase	**B1**	2.4
		(3)	
(c)	The mean of the tied ranks is given to each…	B1	2.4
	… then use PMCC	B1	2.4
		(2)	
			(9 marks)

Notes	
(a)	**B1:** Both hypotheses correct written using the notation ρ **B1:** awrt 0.893 **M1:** Drawing a correct inference using their answer to part(a) and their CV **A1ft:** Drawing a correct inference in context using their answer to part(a) and their CV
(b)(i)	**B1:** Stating **no change** and an explanation including **ranks remain unchanged** oe and **no change oe**
(b)(ii)	**B1:** Interpreted the outcome of adding a point as **increased** oe **B1:** Explaining why. Need to mention the **ranks are the same for both oe** and ***n*** **has increased oe**
(c)	**B1:** Explaining that The mean of the values for the tied ranks is given to both values **B1:** Explaining that the PMCC must be used