



GuideMyGrades

— YOUR GOALS. OUR EXPERTISE. —

EDEXCEL A LEVEL FURTHER MATHEMATICS: QUESTIONS BY TOPIC

Correlation (3)

Further Statistics 2 · Topic 3

8 questions · 66 marks · Mark schemes included

Pearson Edexcel A Level Further Mathematics (9FM0), June 2019 - June 2025

 [guidemygrades.com](https://www.guidemygrades.com)

 @Guidemygrades

 @Guidemygrades

 @Guidemygrades

 info@guidemygrades.com

 Piccadilly Business Centre, Unit C, Aldow Enterprise Park
Blackett Street, Manchester, M12 6AE

HOW TO USE THIS BOOKLET

Questions are grouped by specification sub-topic and run from the lowest mark tariff to the highest within each one.

Click any row below to jump straight to that sub-topic or question.

On every question page, click ANSWER in the top-right corner to jump to its mark scheme.

Diagrams, graphs, tables and data are printed exactly as they appeared in the original papers.

QUESTION	SOURCE	MARKS	PAGE
3.1 PRODUCT MOMENT CORRELATION		3 Qs · 24 marks	3
1. Q1 <i>also Topic 1</i>	Option 4B: Further Statistics 2 June2022	7	4
2. Q1 <i>also Topic 1</i>	Option 4B: Further Statistics 2 June2023	7	8
3. Q2 <i>also Topic 1</i>	Option 4B: Further Statistics 2 June2019	10	10
3.2 SPEARMAN'S RANK CORRELATION		4 Qs · 32 marks	14
4. Q1	Option 4B: Further Statistics 2 November2021	7	15
5. Q2	Option 4B: Further Statistics 2 June2024	7	17
6. Q8	Option 4B: Further Statistics 2 June2019	11	19
7. Q2	Option 4B: Further Statistics 2 June2025	7	22
3.3 TESTING FOR ZERO CORRELATION		1 Qs · 10 marks	24
8. Q4 <i>also Topic 1</i>	Option 4B: Further Statistics 2 November2021	10	25



QUESTIONS

8 questions · 66 marks · 3 sub-topics



SUB-TOPIC 3.1

PRODUCT MOMENT CORRELATION

Sub-topic 3.1 · 3 questions · 24 marks · lowest tariff first



1. Kwame is investigating a possible relationship between average March temperature, $t^{\circ}\text{C}$, 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.

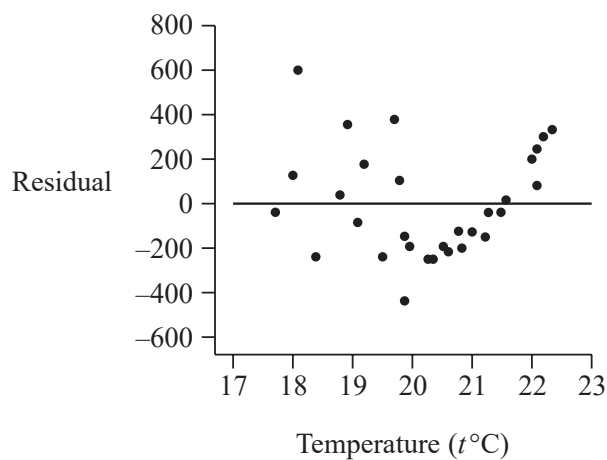
$$\text{Residual Sum of Squares (RSS)} = 1\,666\,567 \quad S_{tt} = 52.0 \quad S_{yy} = 1\,774\,155$$

$$\text{least squares regression line: gradient} = 45.5 \quad y\text{-intercept} = 2080$$

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

(1)

(2)

(2)

Question 1 continues on page 4



Question 1 continued

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 $\text{RSS} = 1\,666\,567$)

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

(1)



(Total for Question 1 is 7 marks)



1. Baako is investigating the times taken by children to run a 100 m race, x seconds, and a 500 m race, y seconds. For a sample of 20 children, Baako obtains the time taken by each child to run each race.

Here are Baako’s summary statistics.

$S_{xx} = 314.55$ $S_{yy} = 9026$ $S_{xy} = 1610$

$\bar{x} = 19.65$ $\bar{y} = 108$

(a) Calculate the product moment correlation coefficient between the times taken to run the 100 m race and the times taken to run the 500 m race. (2)

(b) Show that the equation of the regression line of y on x can be written as

$$y = 5.12x + 7.42$$

where the gradient and y intercept are given to 3 significant figures. (3)

The child who completed the 100 m race in 20 seconds took 104 seconds to complete the 500 m race.

(c) Find the residual for this child. (1)

The table below shows the signs of the residuals for the 20 children in order of finishing time for the 100 m race.

Sign of residual	+	+	+	+	−	−	+	−	−	−	−	−	−	−	−	+	+	+	+	+
------------------	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

(d) Explain what the signs of the residuals show about the model’s predictions of the 500 m race times for the children who are fastest and slowest over the 100 m race. (1)

(Total for Question 1 is 7 marks)

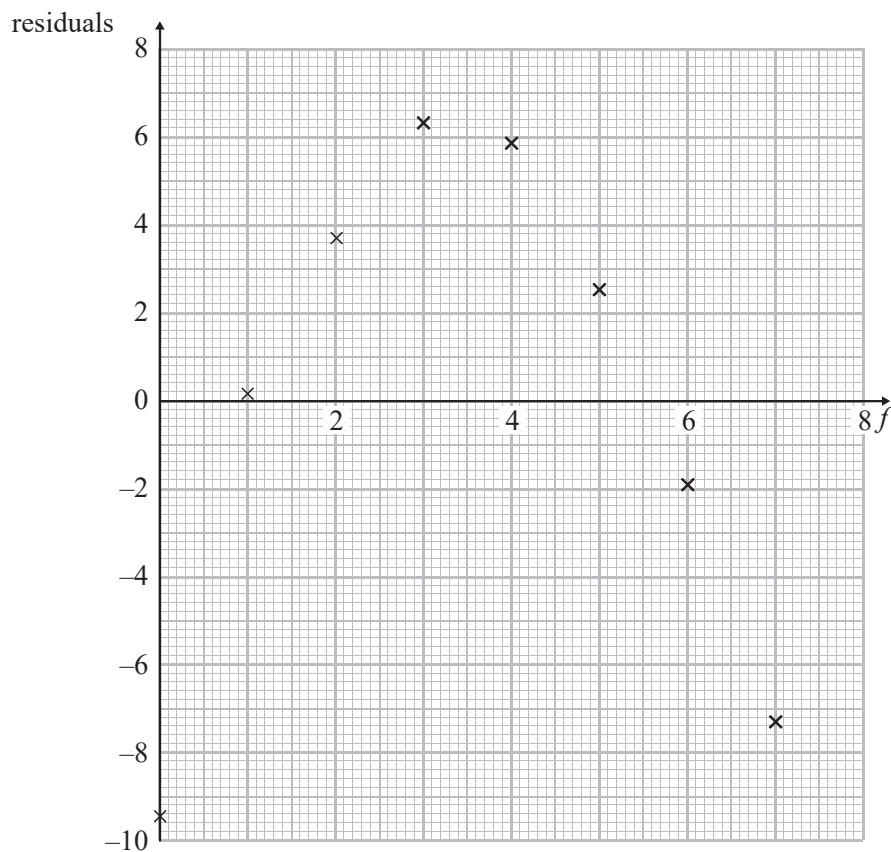


- 2 A large field of wheat is split into 8 plots of equal area. Each plot is treated with a different amount of fertiliser, f grams/m². The yield of wheat, w tonnes, from each plot is recorded. The results are summarised below.

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

- (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/m² (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 2 is 10 marks)



SUB-TOPIC 3.2

SPEARMAN'S RANK CORRELATION

Sub-topic 3.2 · 4 questions · 32 marks · lowest tariff first



1. 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 $r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$

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 1 is 7 marks)



2. An estate agent asks customers to rank 7 features of a house, A, B, C, D, E, F and G , in order of importance. The responses for two randomly selected customers are in the table below.

Rank	1	2	3	4	5	6	7
Customer 1	A	E	C	F	G	B	D
Customer 2	E	F	C	G	A	D	B

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

(4)

- (b) Stating your hypotheses and critical value clearly, test at the 5% level of significance, whether or not the two customers are generally in agreement.

(3)



(Total for Question 8 is 11 marks)

Question 2 continued

SUB-TOPIC 3.3

TESTING FOR ZERO CORRELATION

Sub-topic 3.3 · 1 question · 10 marks · lowest tariff first





Question 4 continued

(Total for Question 4 is 10 marks)



MARK SCHEMES

official mark schemes



Question	Scheme	Marks	AOs
1(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) (ii)	Since r is close to 0/weak correlation	B1	2.4
	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		
Question	Scheme	Marks	AOs



Question	Scheme	Marks	AOs
1(a)	$r = \frac{1610}{\sqrt{314.55 \times 9026}}$	M1	1.1b
	$= 0.9555\dots$ awrt <u>0.956</u>	A1	1.1b
		(2)	
(b)	$b = \frac{1610}{314.55}$	M1	2.1
	$a = \bar{y} - b\bar{x} = 108 - 'b' \times 19.65$	M1	1.1b
	$y = 5.12x + 7.42^*$	A1*	1.1b
		(3)	
(c)	Residual = $104 - (5.12(20) + 7.42) =$ awrt <u>-5.8</u>	B1	3.4
		(1)	
(d)	Positive residuals show that the model <u>underestimates</u> the 500m race times of the fastest and slowest children.	B1	3.5a
		(1)	
(7 marks)			
Notes			
(a)	M1: correct expression for r A1: awrt <u>0.956</u>		
(b)	M1: correct expression for gradient, or correct formula stated with b to more than 3.s.f M1: correct expression, or correct formula stated with a to more than 3.s.f follow through their b *A1: correct linear model with $a =$ awrt <u>7.42</u> and $b =$ awrt <u>5.12</u>		
(c)	B1: correct use of model awrt <u>-5.8</u>		
(d)	B1: correct contextual explanation mentioning positive residuals and underestimate o.e.		

2(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
	$a = \frac{303}{8} - 'b' \cdot \frac{28}{8} [= 15.41666...]$		M1	1.1b
	$w = 15.4 + 6.42 f$		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 $w = 15.4 + 6.42 f$ 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		

Question	Scheme	Marks	AOs
1(a)	$s = 4$	B1	1.1b
	$t = 4.5$	B1	1.1b
		(2)	
(b)	Because there are <u>tied ranks</u> .	B1	2.4
		(1)	
(c)	$H_0: \rho_s = 0 \quad H_1: \rho_s > 0$	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		
Question	Scheme	Marks	AOs

Question		Scheme							Marks	AOs
2(a)	Rank	A	B	C	D	E	F	G	M1	2.1
	Customer 1	1	6	3	7	2	4	5		
	Customer 2	5	7	3	6	1	2	4		
	$\sum d^2 = 16 + 1 + 1 + 1 + 4 + 1 [= 24]$								M1	1.1b
	$r_s = 1 - \frac{6 \times 24}{7(7^2 - 1)}$								dM1	1.1b
	$= 0.57142 \dots$ awrt 0.571								A1	1.1b
									(4)	
(b)	$H_0 : \rho_s = 0 \quad H_1 : \rho_s > 0$								B1	2.5
	5% critical value ($\pm$) 0.7143								B1	1.1a
	Insufficient evidence to reject H_0 . There is no evidence to suggest the customers are in agreement								B1	2.2b
									(3)	
Total 7										
Notes										
(a)	M1	For pairing up the ranks (at least 4 pairs correct) may be implied by a later calculation or awrt 0.571								
	M1	For an attempt at $\sum d^2$ condoning slips. Implied by 24 or awrt 0.571 (If $\sum d^2$ is not correct then look for attempts at the differences proceeding to a value for $\sum d^2$)								
	dM1	Dependent on the previous method mark for use of their $\sum d^2$ in a correct formula (which must be less than 112) Does not need to be evaluated. Implied by awrt 0.571								
	A1	awrt 0.571 correct answer with no incorrect working seen scores full marks Accept $\frac{4}{7}$								
(b)	B1	For both hypotheses in terms of ρ or ρ_s (do not allow r or r_s)								
	B1	Correct CV of ($\pm$) 0.7143 or better								
	B1	Correct conclusion in context based on their cv and their r_s such that $ r_s < CV $ suggesting there is <u>insufficient</u> evidence to suggest the <u>customers</u> are (generally) in <u>agreement</u> oe / <u>no evidence</u> to suggest a positive <u>correlation</u> between the <u>rankings (of features)</u> by the <u>customers</u> oe. Condone conclusions such as e.g. sufficient evidence to suggest the customers are not in agreement A comparison statement such as “not sig” is not needed but if seen must be correct. Do NOT award this mark if contradictory comments or working is seen e.g. “reject H_0 ” or comparison of their 0.517 with 0.05 or e.g. $0.571 > 0.7143$								
	SC	B0(for 2-tail) B1(for CV ± 0.7587) and B1 correct conclusion in context based on their cv and their r_s such that $ r_s < CV $ (as in main scheme)								

Question	Scheme	Marks	AOs
8(a)	$H_0: \rho_s = 0 \quad H_1: \rho_s > 0$	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 6 th and D is 8 th	A1	2.2a
	SC B7, C6, D8 with no reasons B1 marks as final A1 on open		
		(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)	
Notes: (11 marks)			
(a)B1: Both hypotheses correct written using the notation ρ			
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 6 th D 8 th			
(c)M1: Complete explanation why it decreases			
A1: using the information given to deduce that it decreases			

Qu	Scheme										Mark	AO
2(a)	Student	A	B	C	D	E	F	G	H	I		
	Rank 1	6	2	9	1	4	8	3	5	7	M1	1.1b
	Rank 2	8	3	6	1	4	9	2	7	5		
	$\sum d^2 = 4+1+9+0+0+1+1+4+4 (= 24)$										dM1	1.1b
	$r_s = 1 - \frac{6 \times "24"}{9 \times (9^2 - 1)}$										M1	1.1b
	$= \frac{4}{5} \text{ or } \underline{0.8}$										A1	1.1b
											(4)	
(b)	$H_0 : \rho = 0 \quad H_1 : \rho > 0$										B1	2.5
	cv one tail 5% = 0.6										M1	3.4
	(Significant result so there is evidence that) the <u>lecturers agree</u> about the rank order of the essays.										A1	2.2b
											(3)	
	(7 marks)											
	Notes											
(a)	1 st M1 for an attempt to rank marks (at least 5 correct in one row) Allow reverse ranking. May be implied by 24											
	2 nd dM1 dep on 1st M1 for attempt at calculating d^2 (does not need to be summed). (Allow 1 error squaring their differences) May be implied by 24											
	3 rd M1 indep of first two Ms for using correct formula with their $12 < \sum d^2 < 240$ May be implied by 0.8 o.e.											
	A1 for 0.8 o.e.											
(b)	B1 for both hypotheses correct in terms of ρ or ρ_s (condone if it appears as p). Do not allow hypotheses in words on their own.											
	M1 for selecting the correct critical value. Allow $\pm$ If they scored B0 for 2-tail test allow 0.7											
	A1 dep on $0.7 < "0.8" < 1$ for a correct conclusion in context stating <u>agreement</u> (or <u>agree</u>) between <u>lecturers</u> . Do not allow correlation for agreement (not contextual) Can award this mark independently of hypotheses. Do NOT award this mark if contradictory statements are seen e.g. not significant so lecturers agree. If a comparison is made then it must be correct. Note that M0A1 is not possible											

Question	Scheme	Marks	AOs
4(a)	As elevation increases, temperature decreases.	B1	3.4
		(1)	
(b)	$S_{xt} = -0.959\sqrt{8\,820\,655 \times 444.7} [= -60\,062.38727]$	M1	2.1
	$b = \frac{-60\,062...}{8\,820\,655} [= -0.006809...]$	M1	1.1b
	$a = \frac{94.62}{20} - 'b' \frac{28\,130}{20} [= 14.308...]$	M1	1.1b
	$t = 14.3 - 0.00681x$ *	A1cso*	2.2a
		(4)	
(c)	$[w = \frac{x}{1000} \rightarrow] \quad t = 14.3 - 6.81w$	B1	3.3
		(1)	
(d)	$444.7(1 - (-0.959)^2)$ or $444.7 - \frac{(-60\,062...)^2}{8\,820\,655} [= 35.7*]$	B1cso*	1.1b
		(1)	
(e)(i)	$(\text{residual})^2 = [1.4 - (14.3 - 0.00681(1100))]^2 [= 29.2...]$	M1	3.4
	$[29.2... \div 35.7 \times 100\%]$ awrt 82%	A1	1.1b
		(2)	
(e)(ii)	(As the point representing this data contributes to the majority of the RSS), the point is possibly an outlier and should be investigated.	B1	3.5a
		(1)	
(10 marks)			
Notes			
(a)	B1: Correct contextual interpretation		
(b)	M1: Using pmcc to find S_{xt} M1: Setting up linear model by attempting to find b Note: Allow M2 for $b = r \sqrt{\frac{S_{tt}}{S_{xx}}}$ M1: Setting up linear model by attempting to find a A1cso*: Correct model $t = 14.3 - 0.00681x$ with $a =$ awrt 14.3 and $b =$ awrt -0.00681 dependent upon all previous M marks.		
(c)	B1: Correct model		
(d)	B1cso*: Either correct expression		
(e)(i) and (ii)	M1: Using the model to evaluate the squared residual A1: awrt 82% B1: Evaluating the result obtained from the model to suggest that this point may be an outlier		