



GuideMyGrades

— YOUR GOALS. OUR EXPERTISE. —

EDEXCEL A LEVEL FURTHER MATHEMATICS: QUESTIONS BY TOPIC

Linear Regression (1)

Further Statistics 2 · Topic 1

7 questions · 59 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
1.1 LEAST SQUARES REGRESSION		6 Qs · 53 marks	3
1. Q1 <i>also Topic 3</i>	Option 4B: Further Statistics 2 June2022	7	4
2. Q1 <i>also Topic 3</i>	Option 4B: Further Statistics 2 June2023	7	8
3. Q1	Option 4B: Further Statistics 2 June2024	9	10
4. Q2 <i>also Topic 3</i>	Option 4B: Further Statistics 2 June2019	10	14
5. Q4 <i>also Topic 3</i>	Option 4B: Further Statistics 2 November2021	10	18
6. Q1	Option 4B: Further Statistics 2 June2025	10	22
1.2 RESIDUALS		1 Qs · 6 marks	26
7. Q3	Option 4B: Further Statistics 2 October2020	6	27



QUESTIONS

7 questions · 59 marks · 2 sub-topics



SUB-TOPIC 1.1

LEAST SQUARES REGRESSION

Sub-topic 1.1 · 6 questions · 53 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.

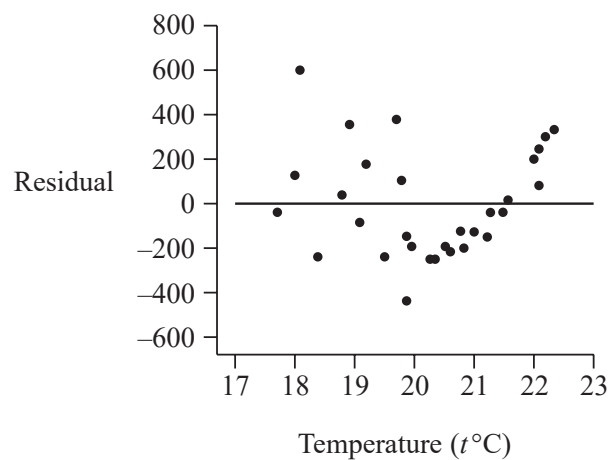
$$\text{Residual Sum of Squares (RSS)} = 1\,666\,567 \quad S_{tt} = 52.0 \quad S_{yy} = 1\,774\,155$$

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

- (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.

(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)

Question 1 continues on page 4



(Total for Question 1 is 7 marks)



(Total for Question 1 is 7 marks)





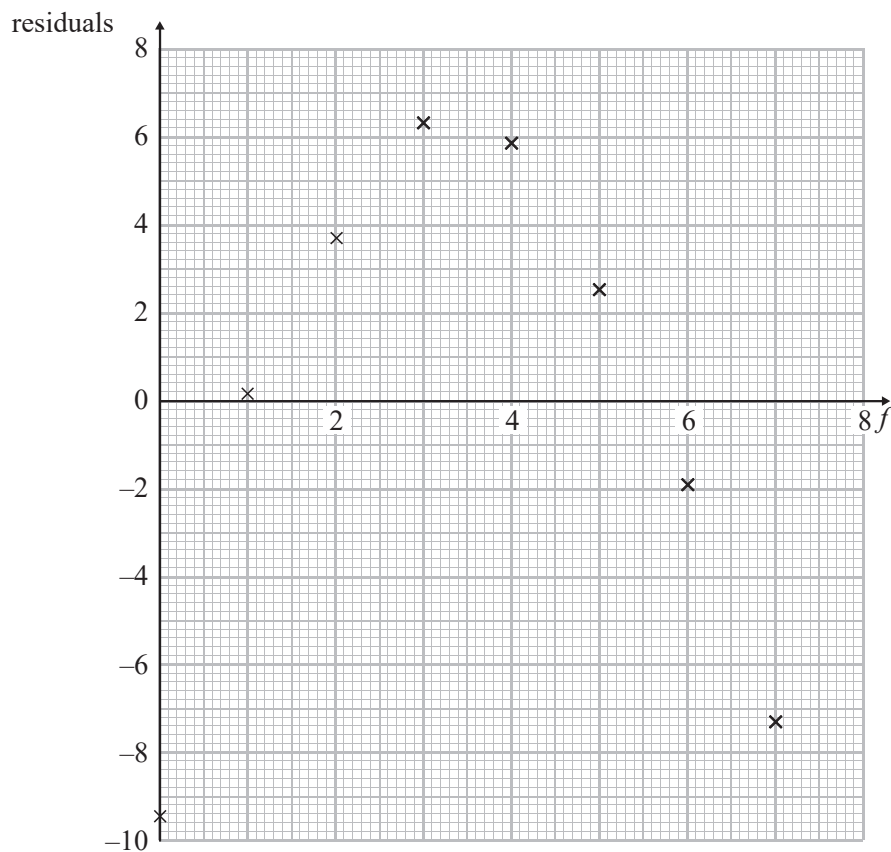


- 2 A large field of wheat is split into 8 plots of equal area. Each plot is treated with a different amount of fertiliser, f grams/m². The yield of wheat, w tonnes, from each plot is recorded. The results are summarised below.

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

- (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)

Question 2 continued

Lined area for writing the answer to Question 2 continued.





Question 2 continued

(Total for Question 2 is 10 marks)







Question 4 continued

(Total for Question 4 is 10 marks)



1. A medical researcher is exploring the relationship between the weight, x kg, and the head circumference, y cm, of newborn babies. A random sample of 15 newborn babies is taken and the data are summarised by the following statistics

$$\sum x = 50.46 \quad \sum x^2 = 171.828 \quad \sum y = 518.9 \quad \sum y^2 = 18004.47 \quad S_{xy} = 7.8284$$

- (a) Show that, to 3 decimal places, $S_{xx} = 2.081$ and $S_{yy} = 53.989$ (3)

- (b) Find the equation of the regression line of y on x , giving your answer in the form $y = a + bx$ (3)

- (c) Find the residual sum of squares (RSS) for these data. (1)

One of these 15 babies had a birth weight of 3.26 kg and a head circumference of 36.8 cm

- (d) Find the residual for this baby. (2)

The researcher claims that this baby could be an outlier.

- (e) Using your answers to part (c) and part (d), give a reason that might support this claim. (1)

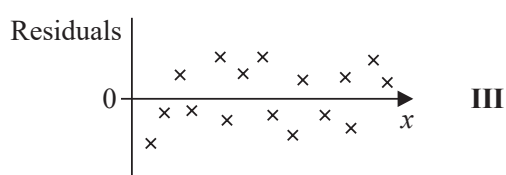
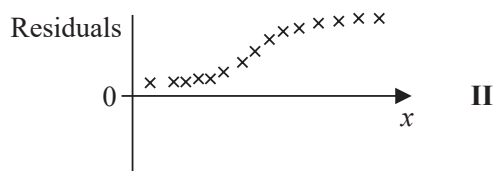
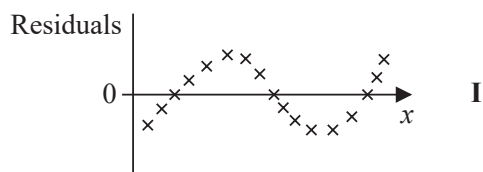
SUB-TOPIC 1.2

RESIDUALS

Sub-topic 1.2 · 1 question · 6 marks · lowest tariff first



3 Below are 3 sketches from some students of the residuals from their linear regressions of y on x .



For each sketch you should state, giving your reason,

- (i) whether or not the sketch is feasible

and if it is feasible

- (ii) whether or not the sketch suggests a linear or a non-linear relationship between y and x .

(6)

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 0.956	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 -5.8	B1	3.4
		(1)	
(d)	<u>Positive residuals</u> 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 0.956		
(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 7.42 and $b =$ awrt 5.12		
(c)	B1: correct use of model awrt -5.8		
(d)	B1: correct contextual explanation mentioning positive residuals and underestimate o.e.		

Question	Scheme	Marks	AOs
1(a)	$b = \frac{-338.83}{4.52} [= -74.96...]$	M1	3.3
	$a = \frac{626}{10} - "b" \frac{22.47}{10} [= 231.04.....]$	M1	1.1b
	$t - 40 = "231.04..." + (" - "74.96...") \sqrt{h}$	dM1	3.1a
	$t = 271.04... - 74.96... \sqrt{h}$	A1	1.1b
		(4)	
(b)	Residual = $47 - ("271.04..." - "74.96..." \times \sqrt{9})$	M1	3.4
	= 0.8466...	A1	1.1b
		(2)	
(c)	$RSS = \left[64678 - \frac{626^2}{10} \right] - \frac{(-338.83)^2}{4.52} \left(= 25490.4 - \frac{(-338.83)^2}{4.52} \right)$	M1	1.1b
	= 90.89...[s ²]	A1	1.1b
		(2)	
(d)	Student A's model as the sum of squares of the residuals is lower	B1	2.4
		(1)	
(9 marks)			
Notes:			
(a)	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	
	dM1	dep on both previous method marks for proceeding from an equation of the form $v = "a" + "b" w$ to a correct un-simplified model in terms of h and t ft their a and b	
	A1	For a correct model $t = 271.04... - 74.95... \sqrt{h}$ with awrt 271 and awrt 75	
(b)	M1	For a correct method to find the residual. States $h = 9$ and a correct expression to find the residual for their model of the form $t = a + b\sqrt{h}$ or 9 substituted into a correct expression. Must be subtracting the correct way round. Alternatively, may code the data to v and w and attempt $47 - ("231.04..." - "74.96..." \times 3)$	
	A1	awrt 0.85 Allow answers in range awrt 0.84 to awrt 0.86 if working shown following a correct model in (a). If no working is shown then look for awrt 0.847	
(c)	M1	For a correct expression for RSS. 25490.4 may be seen as $\frac{127452}{5}$ May also use $RSS = S_w(1 - r^2) = 25490.4 \left(1 - \frac{(-338.83)^2}{4.52 \times 25490.4} \right)$	
	A1	awrt 90.9	
(d)	B1	Explaining a reason for their conclusion that A is a more suitable model provided their positive RSS found in (c) is less than 980. e.g. RSS is smaller so model A Condone references to the model being more accurate oe. Must be a comparison with B or implied so do not accept statements such as "A because it has a small RSS"	



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
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		

Qu	Scheme	Mark	AO
1(a)	$S_{xx} = 171.828 - \frac{50.46^2}{15} \quad \text{or} \quad S_{yy} = 18\,004.47 - \frac{518.9^2}{15}$ $= 2.08056 = \underline{2.081}^*(3dp) \quad ; = 53.98933... = \underline{53.989}^*(3dp)$	M1 A1*; A1*	1.1b 1.1b(x2)
		(3)	
(b)	$b = \left[\frac{S_{xy}}{S_{xx}} \right] = \frac{7.8284}{2.081} = (3.76...)$	M1	1.1b
	$a = \frac{518.9}{15} - b \frac{50.46}{15} = 34.593... - "3.76..." \times 3.364 (= 21.9...)$	M1	1.1b
	Equation of line is: $y = \underline{21.9 + 3.76x}$	A1	1.1b
		(3)	
(c)	$[RSS = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}] = 24.5397... = \text{awrt } \underline{24.5}$	B1	1.1b
		(1)	
(d)	$\hat{y} = "21.9" + "3.76" \times 3.26 (= 34.1576)$	M1	3.4
	$[\text{Residual} = y - \hat{y} = 36.8 - "34.15..." = 2.6424] = \text{awrt } \underline{2.6}$	A1	1.1b
		(2)	
(e)	"2.64..." ² ≈ 7 which is a high proportion of RSS (so this value may be an outlier)	B1	2.4
		(1)	
(10 marks)			
Notes			
(a)	M1 for either correct expression		
*	1 st A1 cso for 2.081 or better following a correct expression		
*	2 nd A1 cso for 53.989 or better following a correct expression		
(b)	1 st M1 for a correct numerical expression for gradient (or ft their $S_{xx} \neq 2.081$, $S_{xy} \neq 7.8284$) 2 nd M1 for a correct expression for intercept (ft their b). May be implied by awrt 21.9 A1 for equation with $a = \text{awrt } 21.9$ and $b = \text{awrt } 3.76$ Note A correct equation scores M1M1A1		
(c)	B1 for awrt 24.5		
(d)	M1 for an expression for $\hat{y}$ or the residual (ft their regression line equation). Implied by awrt 2.6 A1 for awrt 2.6 (exact figures give 2.5979...)		
(e)	B1 Dependent on $0.25 < \frac{"2.6"{}^2}{\text{"24.5"}} < 1$ for a suitable explanation based on their (d) ² compared with their (c) or other suitable calculations which enable a comparison to be made. e.g. this residual is 27.5% of the total which is a large proportion". References to size e.g. "large" requires quoting the numerical value as well.		



Qu	Scheme	Marks	AO
3	I (Is feasible as a residual plot but) probably a non-linear relationship Since the residuals are not randomly scattered about zero	B1 B1	2.2b 2.4
	II Impossible as a residual plot Since the residuals do not sum to zero	B1 B1	2.2a 2.4
	III (Is feasible as a residual plot) and probably a linear relationship Since the points are randomly scattered about zero	B1 B1 (6)	2.2b 2.4
		(6 marks)	
	Notes		
I	1 st B1 for stating possibly non-linear (allow a suitable sketch) 2 nd B1 for a suitable comment (e.g. follow a systematic pattern)		
II	1 st B1 for stating not feasible as a residual plot 2 nd B1 for a correct reason		
III	1 st B1 for stating probably a linear relationship 2 nd B1 for a suitable supporting reason		