

QQQ – Further Statistics 2 – Chapter 1 – Linear Regression (Version 2)

Total Marks: 24

(24 = Platinum, 22 = Gold, 19 = Silver, 16 = Bronze)

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.

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

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

(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 7 marks)

2. 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.

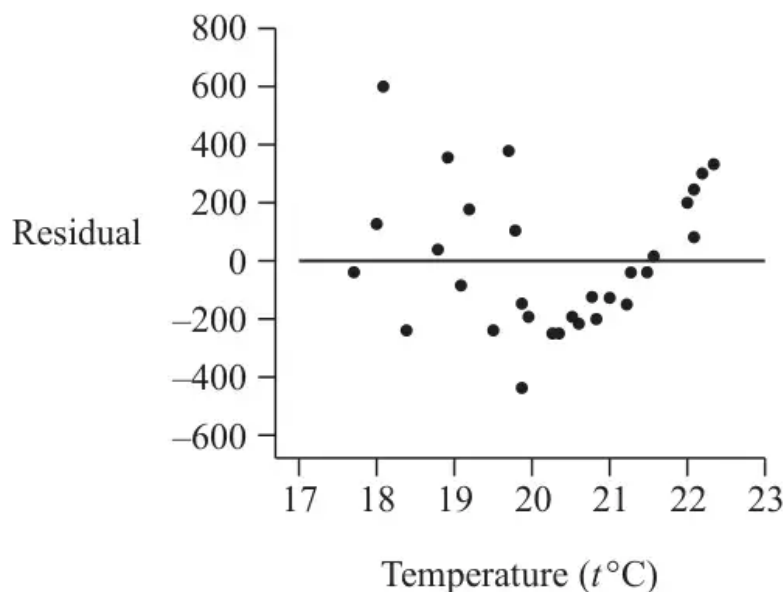
He uses 30 years of past data to produce the following summary statistics for a linear regression model, with tea yield as the dependent variable.

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

least squares regression line: gradient = 45.5 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.

$$\text{Residual Sum of Squares (RSS)} = 86\,754$$

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 7 marks)

Edexcel A2 June 2022 Q1

3. A researcher is investigating the relationship between elevation, x metres, and annual mean temperature, t °C.

From a random sample of 20 weather stations in Switzerland, the following results were obtained

$$S_{xx} = 8\,820\,655 \quad S_{tt} = 444.7 \quad \sum x = 28\,130 \quad \sum t = 94.62$$

The product moment correlation coefficient for these data is found to be -0.959

- (a) Interpret the value of this correlation coefficient. (1)
- (b) Show that the equation of the regression line of t on x can be written as

$$t = 14.3 - 0.00681x$$

(4)

The random variable W represents the elevations of the weather stations in kilometres.

- (c) Write down the equation of the regression line of t on w for these 20 weather stations in the form $t = a + bw$ (1)
- (d) Show that the residual sum of squares (RSS) for the model for t and x is 35.7 correct to one decimal place. (1)

One of the weather stations in the sample had a recorded elevation of 1100 metres and an annual mean temperature of 1.4 °C

- (e) (i) Calculate this weather station's contribution to the residual sum of squares. Give your answer as a percentage. (2)
- (ii) Comment on the data for this weather station in light of your answer to part (e)(i). (1)

(Total 10 marks)

Edexcel A2 October 2021 Q4

Solutions

1.

Mark scheme (a)

Scheme	Marks	AO
$r = \frac{1610}{\sqrt{314.55 \times 9026}}$	M1	1.1b
$= 0.9555 \dots$ awrt 0.956	A1	1.1b
	(2)	

Notes

M1: correct expression for r

A1: awrt **0.956**

Mark scheme (b)

Scheme	Marks	AO
$b = \frac{1610}{314.55}$	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)	

Notes

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**

Mark scheme (c)

Scheme	Marks	AO
Residual = $104 - (5.12(20) + 7.42) =$ awrt -5.8	B1	3.4

	(1)	
--	-----	--

Notes

B1: correct use of model awrt **-5.8**

Mark scheme (d)

Scheme	Marks	AO
<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

B1: correct contextual explanation mentioning positive residuals and underestimate o.e.

2.

Mark scheme (a)

Scheme	Marks	AO
$[20 \times 45.5 + 2080] = 2990$ [kg/ha]	B1	3.4
	(1)	

Notes

B1: cao

Mark scheme (b)

Scheme	Marks	AO
(A residual is the) difference between the observed value (oe) and the predicted value (oe) (of the dependent variable)	B1	1.2
	(1)	

Notes

B1: Correct definition. Allow equivalent wording.

Distance from regression line on its own is B0, but allow if vertical distance or y is referenced

Mark scheme (c)

Scheme	Marks	AO
$1666567 = 1774155(1 - r^2)$	M1	1.1b
$r = 0.246 \dots$ awrt 0.246	A1	1.1b
	(2)	

Notes

M1: Use of correct expression for r or r^2

Allow use of $S_{ty} = 45.5 \times 52.0 [= 2366]$

or RSS: $1774155 - \frac{(S_{ty})^2}{52} = 1666567 \rightarrow [S_{ty} = 2365.2 \dots]$

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)

Mark scheme (d)

Scheme	Marks	AO
(i) Since r is close to 0/weak correlation	B1	2.4
(ii) e.g. (For $t > 20 \dots$) the residuals do not appear randomly scattered about 0.	B1	3.5a
	(2)	

Notes

(i) B1: Correct explanation

(ii) 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.

Mark scheme (e)

Scheme	Marks	AO
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

B1: Correct assessment of the conclusion involving the units/size of the variables used to calculate the RSS

3.

Mark scheme (a)

Scheme	Marks	AO
As elevation increases, temperature decreases.	B1	3.4
	(1)	

Notes

B1: Correct contextual interpretation

Mark scheme (b)

Scheme	Marks	AO
$S_{xt} = -0.959\sqrt{8\,820\,655 \times 444.7} [= -60\,062.38727]$	M1	2.1
$b = \frac{'-60\,062 \dots'}{8\,820\,655} [= -0.006809 \dots]$	M1	1.1b
$a = \frac{94.62}{20} - 'b' \frac{28\,130}{20} [= 14.308 \dots]$	M1	1.1b
$t = 14.3 - 0.00681x^*$	A1cso*	2.2a
	(4)	

Notes

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.

Mark scheme (c)

Scheme	Marks	AO
$\left[w = \frac{x}{1000} \rightarrow \right] \quad t = 14.3 - 6.81w$	B1	3.3
	(1)	

Notes

B1: Correct model

Mark scheme (d)

Scheme	Marks	AO
$444.7(1 - (-0.959)^2)$ or $444.7 - \frac{(-60\,062 \dots)^2}{8\,820\,655}$ [= 35.7*]	B1cso*	1.1b
	(1)	

Notes

B1cso*: Either correct expression

Mark scheme (e)

Scheme	Marks	AO
(i) $(\text{residual})^2 =$ $[1.4 - (14.3 - 0.00681(1100))]^2$ [= 29.2 ...]	M1	3.4
$[29.2 \dots \div 35.7 \times 100\%]$ awrt 82%	A1	1.1b
	(2)	
(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

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