Linear Regression

Questions

Q1.

A random sample of 10 female pigs was taken. The number of piglets, x, born to each female pig and their average weight at birth, m kg, was recorded. The results were as follows:

Number of piglets, x	4	5	6	7	8	9	10	11	12	13
Average weight at birth, m kg	1.50	1.20	1.40	1.40	1.23	1.30	1.20	1.15	1.25	1.15

(You may use $S_{xx} = 82.5$ and $S_{mm} = 0.12756$ and $S_{xm} = -2.29$)

(a) Find the equation of the regression line of m on x in the form $m = a + bx$ as a model for these results.

(2)

(b) Show that the residual sum of squares (RSS) is 0.064 to 3 decimal places.

(2)

(c) Calculate the residual values.

(2)

(d) Write down the outlier.

(1)

(e) (i) Comment on the validity of ignoring this outlier .

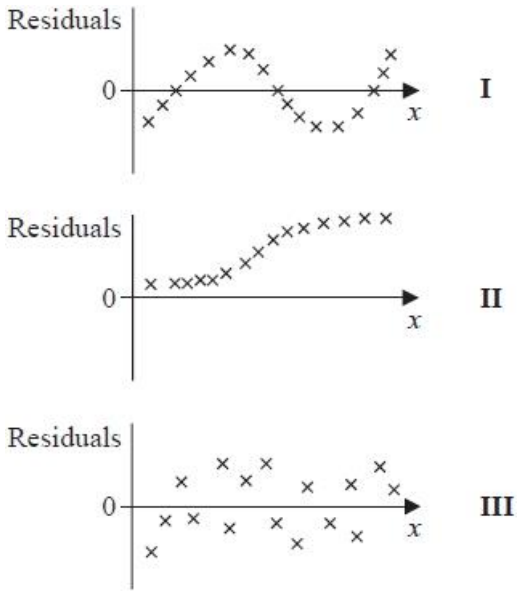
(ii) Ignoring the outlier, produce another model.
(iii) Use this model to estimate the average weight at birth if $x = 15$
(iv) Comment, giving a reason, on the reliability of your estimate.

(5)

(Total for question = 12 marks)

Q2.

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*.

(Total for question = 6 marks)

Q3.

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.

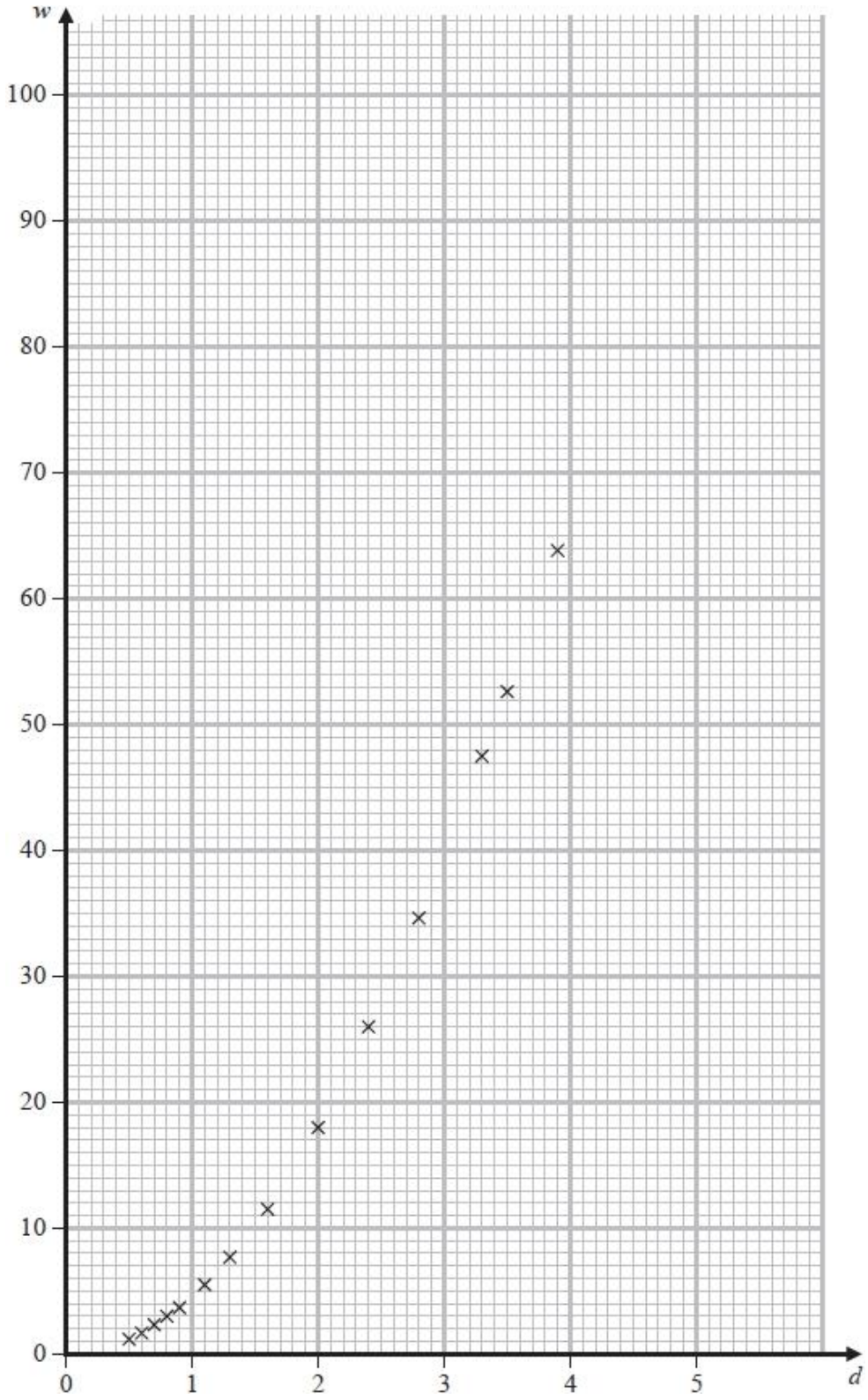
(2)

(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)

Mark Scheme – Linear Regression

Q1.

<table>
<tr><th>Question</th><th>Scheme</th><th>Marks</th><th>AOs</th></tr>
<tr><td rowspan="4">(a)</td><td>$\left[b=\frac{S_{xm}}{S_{xx}}=-0.0277576\right]$</td><td>M1</td><td>3.3</td></tr>
<tr><td>$[a=\bar{m}-b\bar{x}=1.278+0.0277576\times 8.5=1.5139]$</td><td></td><td></td></tr>
<tr><td>$m=1.5139-0.02775\ldots x$</td><td>A1</td><td>1.1b</td></tr>
<tr><td></td><td>(2)</td><td></td></tr>
<tr><td rowspan="3">(b)</td><td>$\text{RSS}=0.12756-\frac{(-2.29)^2}{82.5}$</td><td>M1</td><td>1.1b</td></tr>
<tr><td>$=0.06399$*</td><td>A1*</td><td>1.1b</td></tr>
<tr><td></td><td>(2)</td><td></td></tr>
<tr><td rowspan="2">(c)</td><td><table>
<tr><th>x</th><th>m</th><th>$m=a+bx$</th><th>ε</th></tr>
<tr><td>4</td><td>1.50</td><td>1.4029</td><td>+0.0971</td></tr>
<tr><td>5</td><td>1.20</td><td>1.3752</td><td>− 0.1752</td></tr>
<tr><td>6</td><td>1.40</td><td>1.3474</td><td>+0.0526</td></tr>
<tr><td>7</td><td>1.40</td><td>1.3196</td><td>+0.0804</td></tr>
<tr><td>8</td><td>1.23</td><td>1.2919</td><td>− 0.0619</td></tr>
<tr><td>9</td><td>1.30</td><td>1.2641</td><td>+0.0359</td></tr>
<tr><td>10</td><td>1.20</td><td>1.2364</td><td>− 0.0364</td></tr>
<tr><td>11</td><td>1.15</td><td>1.2086</td><td>− 0.0586</td></tr>
<tr><td>12</td><td>1.25</td><td>1.1808</td><td>+0.0692</td></tr>
<tr><td>13</td><td>1.15</td><td>1.1531</td><td>− 0.0031</td></tr>
</table></td><td>M1
A1</td><td>3.4
1.1b</td></tr>
<tr><td></td><td>(2)</td><td></td></tr>
<tr><td rowspan="2">(d)</td><td>The point (5, 1.2) is an outlier</td><td>B1ft</td><td>2.2b</td></tr>
<tr><td></td><td>(1)</td><td></td></tr>
<tr><td>(e)(i)</td><td>It is a valid piece of data so should be used
or It does not follow the pattern according to the residuals so may contain an error making the result invalid so should be removed</td><td>B1</td><td>2.4</td></tr>
<tr><td rowspan="2">(ii)</td><td>$a=\bar{m}-b\bar{x}=1.28667+0.03765\times 8.88889=1.6213$</td><td>M1</td><td>3.3</td></tr>
<tr><td>$m=1.6213-0.03765x$</td><td>A1</td><td>1.1b</td></tr>
<tr><td rowspan="2">(iii)</td><td>$m=1.6213-0.03765\times 15$</td><td></td><td></td></tr>
<tr><td>$=1.056$ or awrt 1.06</td><td>B1ft</td><td>3.4</td></tr>
<tr><td rowspan="2">(iv)</td><td>The model is only reliable if the values are limited to those in the given range so probably not reliable</td><td>B1</td><td>3.5b</td></tr>
<tr><td></td><td>(5)</td><td></td></tr>
<tr><td colspan="4">(12 marks)</td></tr>
</table>

Notes	
(a)	**M1:** Realsing the need to use $b = \frac{S_{xm}}{S_{xx}}$ and $a = \bar{m} - b\bar{x}$
	A1: m = awrt 1.51) – (awrt 0.0278) x. Award M1A1 for correct equation
(b)	**M1:** Using $S_{mm} - \frac{(S_{xm})^2}{S_{xx}}$
	A1*: awrt 0.064
(c)	**M1:** Using the model in part (a) i.e. m – ("1.5139" – "0.02775"x) implied by a correct value **A1:** All correct. Award M1A1 for a list of correct residuals
(d)	**B1:** Inferring from the residuals that the outlier is (5, 1.2) ft their residuals.
(e)(i)	**B1:** Explaining why the outlier should be removed or not.
(ii)	**M1:** Removing the outlier and refining the model by finding a new regression line. **A1:** $m = (\text{awrt } 1.62) - (\text{awrt } 0.0377)x$
(iii)	**B1ft:** using their model in e(i) with $x = 15$. awrt 1.06 or ft their e(ii)
(iv)	**B1:** Realising the limitations of the model by stating it is <u>not reliable</u> and giving the reason why ie <u>extrapolation/out of range</u> o.e.

Q2.

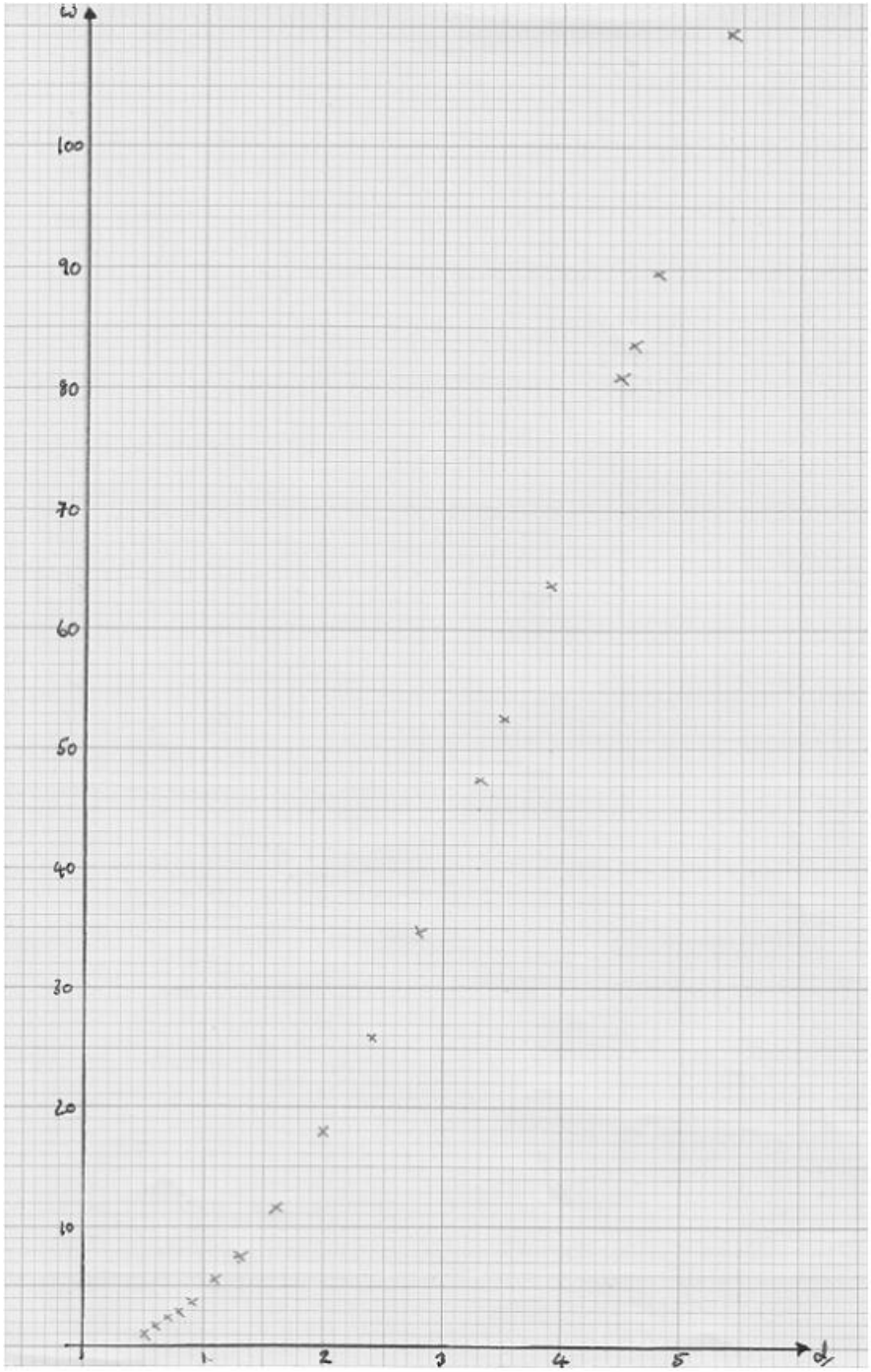
Qu	Scheme	Marks		AO
I	(Is feasible as a residual plot but) probably a non-linear relationship	B1		2.2b
	Since the residuals are not randomly scattered about zero	B1		2.4
II	Impossible as a residual plot	B1		2.2a
	Since the residuals do not sum to zero	B1		2.4
III	(Is feasible as a residual plot) and probably a linear relationship	B1		2.2b
	Since the points are randomly scattered about zero	B1		2.4
		(6)		
		(6 marks)		

	Notes
I	1st B1 for stating possibly non-linear (allow a suitable sketch) 2nd B1 for a suitable comment (e.g. follow a systematic pattern)
II	1st B1 for stating not feasible as a residual plot 2nd B1 for a correct reason
III	1st B1 for stating probably a linear relationship 2nd B1 for a suitable supporting reason

Q3.

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)	$\left\{ S_{ww} = \sum w^2 - \frac{(\sum w)^2}{18} = 45178.68 - \frac{643.6^2}{18} \right\} = 22166.404..$	M1	1.1b
	RSS $= S_{ww}(1-r^2) = 22166.404\ldots \times (1-0.987^2)$ = 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 $S_{ww} - \frac{(S_{uw})^2}{S_{uu}}$, $= 85.8824\ldots$ 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)


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