Linear regression 1C

1 This table sets out the residuals for each data point:

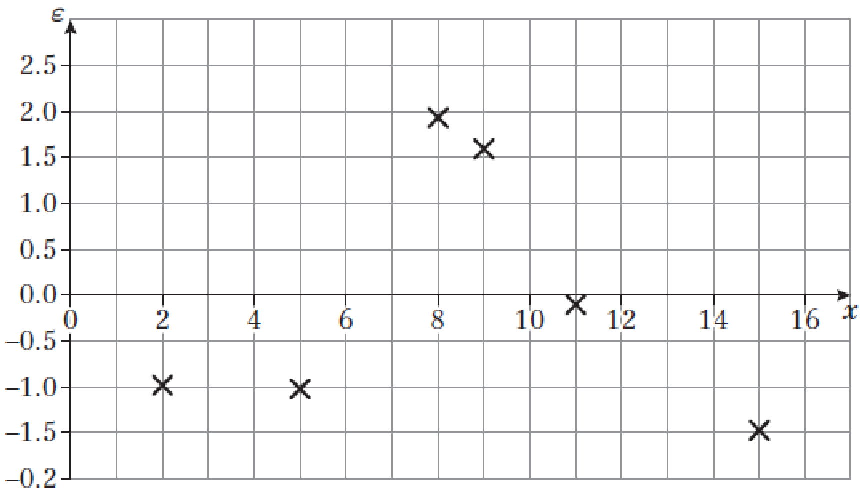
x	y	$y = -3.633 + 14.33x$	ε
1.1	12.2	12.13	0.07
1.3	14.5	14.996	–0.496
1.4	16.9	16.429	0.471
1.7	p	20.728	$p - 20.728$
1.9	23.5	23.594	–0.094

$$\sum \varepsilon = 0 \Rightarrow 0.07 - 0.496 + 0.471 + (p - 20.728) - 0.094 = 0$$
$$\Rightarrow p = 20.777 = 20.8 \text{ (3 s.f.)}$$

2 **a** This table sets out the residuals for each data point:

x	m	$m = 114.3 - 1.655x$	ε
2	110	110.99	–0.99
5	105	106.025	–1.025
8	103	101.06	1.94
9	101	99.405	1.595
11	96	96.095	–0.095
15	88	89.475	–1.475

b



c A linear model is suitable since the residuals appear to be randomly scattered about zero.

3 a This table sets out the residuals for each data point:

t	p	$p = 51.04 + 5.308t$	ε
5.1	79	78.121	0.8892
5.7	81	81.307	–0.2956
6.3	85	84.493	0.5196
6.4	86	85.024	0.9888
7.1	89	88.741	0.2732
7.2	84	89.272	–5.2576
8.0	95	93.52	1.496
8.3	96	95.113	0.9036
8.7	98	97.237	0.7804
9.1	99	99.361	–0.3428

The outlier is the point (7.2, 84) since the size of the residual is far greater than for the other points.

b There is no correct answer here.

Possible reason for including this outlier in the data: Sarah might have just had a bad day and it is a legitimate percentage.

Possible answer for excluding this outlier from the data: the data point could have been incorrectly recorded.

c Use of calculator with the remaining data points gives a p on t regression line equation of:

$p = 51.592\ldots + 5.3116\ldots t$

$\Rightarrow p = 51.6 + 53.1t$ (parameters to 3 s.f.)

d $p = 51.592\ldots + 5.3116 \times 7.8 = 93.022\ldots = 93.0\%$ (3 s.f.)

4 a This table sets out the residuals for each data point:

x	y	$y = 15.7 - 2.02x$	ε
1.2	13.1	13.276	–0.176
1.7	12.5	12.266	0.234
2.4	10.9	10.852	0.048
3.1	9.4	9.438	–0.038
3.8	7.9	8.024	–0.124
4.2	a	7.216	$a - 7.216$
5.1	5.8	5.398	0.402

$\sum \varepsilon = 0 \Rightarrow -0.176 + 0.234 + 0.478 - 0.038 - 0.124 + (a - 7.216) + 0.402 = 0$

$\Rightarrow a = 6.87$

4 b A linear model is suitable since the residuals are randomly distributed about zero.

5 a The residual sum of squares measures the reasonableness of linear fit.

b $S_{aa} = \sum a^2 - \frac{\left(\sum a\right)^2}{n} = 7720 - \frac{236^2}{8} = 758$

$S_{tt} = \sum t^2 - \frac{\left(\sum t\right)^2}{n} = 4821 - \frac{193^2}{8} = 164.875$

$S_{at} = \sum at - \frac{\sum a \sum t}{n} = 6046 - \frac{236 \times 193}{8} = 352.5$

$\text{RSS} = S_{tt} - \frac{\left(S_{at}\right)^2}{S_{aa}} = 164.875 - \frac{352.5^2}{758} = 0.949 \text{ (3 s.f.)}$

c The obstacle course data is more likely to have a linear fit since the RSS is lower.

6 a $\text{RSS} = S_{dd} - \frac{\left(S_{hd}\right)^2}{S_{hh}} = 1.949 - \frac{23.13^2}{289.4} = 0.100 \text{ (3 s.f.)}$

b The sample from October is more likely to have a linear fit since the RSS is lower.

7 a $S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 33.56 - \frac{12.8^2}{6} = 6.25333\ldots$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 120.03 - \frac{12.8 \times 65.4}{6} = -19.49$

$b = \frac{S_{xy}}{S_{xx}} = \frac{-19.49}{6.25333} = -3.1167\ldots = -3.117 \text{ (4 s.f.)}$

$\bar{x} = \frac{\sum x}{n} = \frac{12.8}{6} = \frac{32}{15}$ $\qquad \bar{y} = \frac{\sum y}{n} = \frac{65.4}{6} = 10.9$

$a = \bar{y} - b\bar{x} = 10.9 + 6.25333\ldots \times \frac{32}{15} = 17.5490\ldots = 17.55 \text{ (4 s.f.)}$

Hence the equation of the regression line of y on x is: $y = 17.55 - 3.117x$

b The value of a (17.55) means that the price of a brand new car before depreciation is £17 550).

c $y = 17.5490\ldots - 3.1167 \times 2 = 11.3156$

So an estimate for the value of a 2-year old car is £11 316 (to the nearest pound).

7 d This table sets out the residuals for each data point:

x	y	$y = 17.55 - 3.117x$	ε
0.7	15.4	15.3681	0.0319
1.3	13.5	13.4979	0.0021
1.8	12.1	11.9394	0.1606
2.3	10.1	10.3809	–0.2809
2.9	8.5	8.5107	–0.0107
3.8	5.8	5.7054	0.0946

e A linear model is suitable since the residuals are close to zero and scattered about zero.

f $S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 773.72 - \frac{65.4^2}{6} = 60.86$

$\text{RSS} = S_{yy} - \frac{\left(S_{xy}\right)^2}{S_{xx}} = 60.86 - \frac{-19.49^2}{6.25333} = 0.1148 \text{ (4 s.f.)}$

g The first sample is more likely to have a linear fit since the RSS is lower.

Challenge

Using equation of regression line with $\bar{y}$ and $\bar{x}$:

$$\frac{9+p+q}{3} = 2 + 4\left(\frac{1+5+7}{3}\right) \Rightarrow p + q = 49 \qquad (1)$$

Using $b = \frac{S_{xy}}{S_{xx}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$:

$$4 = \frac{9 + 5p + 7q - \frac{13(9+p+q)}{3}}{75 - \frac{13^2}{3}} = \frac{27 + 15p + 21q - 117 - 13p - 13q}{56}$$

$$\Rightarrow p + 4q = 157 \qquad (2)$$

Subtracting equation (1) from equation (2) gives:

$3q = 108 \Rightarrow q = 36 \Rightarrow p = 13$