Linear regression Mixed exercise

1 a $S_{st} = \sum st - \dfrac{\sum s \sum t}{n} = 31\,185 - \dfrac{553 \times 549}{12} = 31\,185 - 25299.75 = 5885.25$

$b = \dfrac{S_{st}}{S_{ss}} = \dfrac{5885.25}{6193} = 0.95030\ldots = 0.950$ (3 s.f.)

$a = \bar{t} - b\bar{s} = 45.75 - (0.95030\ldots \times 46.0833) = 1.95672\ldots = 1.96$ (3 s.f.)

Hence equation of regression line of t on s is: $t = 1.96 + 0.95s$

b $t = 1.9567\ldots + (0.9503\ldots \times 50) = 49.4717 = 49.5$ (3 s.f.)

2 a $S_{xx} = \sum x^2 - \dfrac{\left(\sum x\right)^2}{n} = 97.73 - \dfrac{33.1 \times 33.1}{12} = 6.4291\ldots = 6.429$ (4 s.f.)

$S_{xy} = \sum xy - \dfrac{\sum x \sum y}{n} = 195.94 - \dfrac{33.1 \times 66.8}{12} = 195.94 - 184.26 = 11.68333\ldots = 11.68$ (4 s.f.)

b $\bar{x} = \dfrac{\sum x}{n} = \dfrac{33.1}{12} = 2.7583\ldots \qquad \bar{y} = \dfrac{\sum y}{n} = \dfrac{66.8}{12} = 5.5666\ldots$

$b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{11.6833\ldots}{6.4291\ldots} = 1.8172\ldots = 1.82$ (3 s.f.)

$a = \bar{y} - b\bar{x} = 5.5666\ldots - (1.8172\ldots \times 2.7583\ldots) = 0.55421\ldots = 0.554$ (3 s.f.)

Equation is: $y = 0.554 + 1.82x$

c Length of leaf $= 0.5542\ldots + (1.8172\ldots \times 3) = 6.0058\ldots = 6.01$cm (3 s.f.)

3 a Calculating the summary statistics gives:

$\sum x^2 = 43622.85 \qquad \sum x = 467.1 \qquad \sum y = 7805 \qquad \sum xy = 666045$

$S_{xx} = \sum x^2 - \dfrac{\left(\sum x\right)^2}{n} = 43622.85 - \dfrac{467.1 \times 467.1}{8} = 16350.048\ldots = 16350$ (5 s.f.)

$S_{xy} = 666045 - \dfrac{467.1 \times 7805}{8} = 210330.56\ldots = 210331$ (6 s.f.)

b $\bar{x} = \dfrac{\sum x}{n} = \dfrac{467.1}{8} = 58.3875 \qquad \bar{y} = \dfrac{\sum y}{n} = \dfrac{7805}{8} = 975.625$

$b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{210330.56}{16350.048} = 12.8642\ldots = 12.86$ (4 s.f.)

$a = \bar{y} - b\bar{x} = 975.625 - (12.8642\ldots \times 58.3875) = 224.5155\ldots = 224.5$ (4 s.f.)

Equation is: $y = 224.5 + 12.86x$

c Gross National Product $= 224.515\ldots + (12.8642\ldots \times 100) = 1510.93\ldots = 1511$ (4 s.f.)

3 d $3500 = 224.515\ldots + 12.864\ldots x$

$\Rightarrow$ Energy consumption $(x) = \dfrac{3500 - 224.515\ldots}{12.8642\ldots} = 255$ (3 s.f.)

e This answer is likely to be unreliable as it involves extrapolation. The value of 3500 is well outside the limits of the data set used.

4 a $S_{xy} = \sum xy - \dfrac{\sum x \sum y}{n} = 84.25 - \dfrac{25.5 \times 13.5}{6} = 84.25 - 57.375 = 26.875$

$\bar{x} = \dfrac{\sum x}{n} = \dfrac{25.5}{6} = 4.25 \qquad \bar{y} = \dfrac{\sum y}{n} = \dfrac{13.5}{6} = 2.25$

$b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{26.875}{59.88} = 0.44881\ldots = 0.449$ (3 s.f.)

$a = \bar{y} - b\bar{x} = 2.25 - (0.44881\ldots \times 4.25) = 0.3425\ldots = 0.343$ (3 s.f.)

Equation is: $y = 0.343 + 0.449x$

b $t - 2 = 0.3425\ldots + 0.4488\ldots\left(\dfrac{m}{2}\right)$

$\Rightarrow t = 2.3425\ldots + 0.2244\ldots m$

$\Rightarrow t = 2.34 + 0.224m$ (rounding the parameters to 3 s.f.)

c Tail length $= 2.3425\ldots + (0.2244\ldots \times 10) = 4.5865\ldots = 4.6\,\text{cm}$ (2 s.f.)

5 a $S_{xx} = \sum x^2 - \dfrac{\left(\sum x\right)^2}{n} = 747 - \dfrac{81 \times 81}{10} = 747 - 656.1 = 90.9$

$S_{xy} = \sum xy - \dfrac{\sum x \sum y}{n} = 1413 - \dfrac{81 \times 151}{10} = 1413 - 1223.1 = 189.9 = 190$ (3 s.f.)

b $b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{189.9}{90.9} = 2.0891\ldots = 2.09$ (3 s.f.)

$\bar{x} = \dfrac{\sum x}{n} = \dfrac{81}{10} = 8.1 \qquad \bar{y} = \dfrac{\sum y}{n} = \dfrac{151}{10} = 15.1$

$a = \bar{y} - b\bar{x} = 15.1 - (2.0891\ldots \times 8.1) = -1.8217\ldots = -1.82$ (3 s.f.)

Equation is: $y = -1.82 + 2.09x$

c $\dfrac{p - 50}{2} = -1.8217\ldots + 2.0891\ldots\left(\dfrac{r - 10}{2}\right)$

$\Rightarrow p - 50 = 3.6434\ldots + 2.0891\ldots r - 20.891$

$\Rightarrow p = 25.465\ldots + 2.0891\ldots r$

$\Rightarrow p = 25.5 + 2.09r$ (rounding paraemeters to 3 s.f.)

d Pulse rate $= 25.5 + (2.09 \times 22) = 71$ beats per minute (2 s.f.)

5 e The value of 22 breaths per minute equates to $x = 6$ and is within range of the data set used. So the answer to part **d** is reasonably reliable since it involves interpolation.

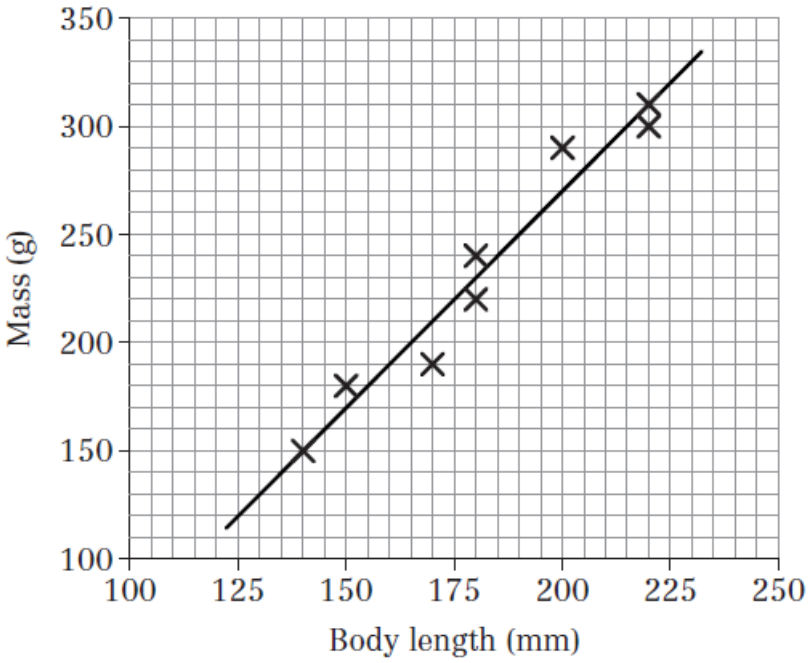
6 a The figure of 0.79 is the average amount of food consumed (in kg) in 1 week by 1 hen.

b $y = 0.16 + 0.79 \times 30 = 23.86 = 23.9\,\text{kg}$ (3 s.f.)

c Food needed $= 0.16 + 0.79 \times 50 = 39.66\,\text{kg}$

Cost of feed $= \dfrac{39.66}{10} \times 12 = £47.592 = £47.59$

7 a This is a scatter diagram of the data. (The diagram also shows the regression line, found in part **e**.)



b There appears to be a linear relationship between body length and body mass.

7 c Calculating the summary statistics for l and w gives:

l	14	15	17	18	18	20	22	22
w	15	18	19	22	24	29	30	31

$\sum l^2 = 2726 \quad \sum l = 146 \quad \sum w = 188 \quad \sum lw = 3553$

$\bar{l} = \frac{\sum l}{n} = \frac{146}{8} = 18.25 \qquad \bar{w} = \frac{\sum w}{n} = \frac{188}{8} = 23.5$

$S_{ll} = \sum l^2 - \frac{\left(\sum l\right)^2}{n} = 2726 - \frac{146 \times 146}{8} = 2726 - 2664.5 = 61.5$

$S_{lw} = \sum lw - \frac{\sum l \sum w}{n} = 3553 - \frac{146 \times 188}{8} = 3553 - 3431 = 122$

$b = \frac{S_{lw}}{S_{ll}} = \frac{122}{61.5} = 1.9837\ldots = 1.98 \text{ (3 s.f.)}$

$a = \bar{w} - b\bar{l} = 23.5 - (1.9837\ldots \times 18.25) = 23.5 - 36.2032\ldots = -12.7032\ldots = -12.7 \text{ (3 s.f.)}$

Equation is: $w = -12.7 + 1.98l$

d $\frac{y}{10} = -12.7 + \left(1.98 \times \frac{x}{10}\right) \Rightarrow y = -127 + 1.98x$ (multiply through by 10)

e See diagram for part **a**.

f Mass $= -127.0\ldots + 1.983\ldots \times 210 = 289.43\ldots = 290$ grams (2 s.f.)

This is reliable since it involves interpolation. The mass of 210 is within the range of the data.

g Voles B and C are both underweight so were probably removed from the river. Vole A is slightly overweight so was probably left in the river.

8 a Calculating the summary statistics for x and y gives:

x	0	3	12	5	14	6	9
y	7	9	15	9	13	11	13

$\sum x = 49 \quad \sum x^2 = 491 \quad \sum y = 77 \quad \sum xy = 617$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 617 - \frac{49 \times 77}{7} = 617 - 539 = 78$

$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 491 - \frac{49^2}{7} = 491 - 343 = 148$

8 b $\bar{x} = \dfrac{\sum x}{7} = \dfrac{49}{7} = 7 \qquad \bar{y} = \dfrac{\sum y}{7} = \dfrac{77}{7} = 11$

$b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{78}{148} = 0.52702\ldots = 0.5270$ (4 s.f.)

$a = \bar{y} - b\bar{x} = 11 - (0.52702\ldots \times 7) = 7.3108\ldots = 7.311$ (4 s.f.)

Equation is: $y = 7.311 + 0.5270x$ (parameters to 4 s.f.)

c $\dfrac{w}{400} = 7.3108\ldots + 0.52702\ldots(n-10)$ (multiply by 400)

$\Rightarrow w = 816.24\ldots + 210.808n$

$\Rightarrow w = 816.2 + 210.8n$ (parameters to 4 s.f.)

d $w = 816.24\ldots + 210.808\ldots \times 20 = 5032\,\text{kg}$

e This is far outside the range of values. This is extrapolation.

f This table sets out the residuals for each coded data point (x, y):

x	y	$y = 7.311 + 0.527x$	ε
0	7	7.311	–0.311
3	9	8.892	0.108
12	15	13.635	1.365
5	9	9.946	–0.946
14	13	14.689	–1.689
6	11	10.473	0.527
9	13	12.054	0.946

g This table sets out the residuals for each coded data point (n, w):

n	w	$w = 816.2 + 210.8n$	ε
10	2800	2924.2	–124.2
13	3600	3556.6	43.4
22	6000	5453.8	546.2
15	3600	3978.2	–378.2
24	5200	5875.4	–675.4
16	4400	4189	211
19	5200	4821.4	378.6

h They are related by the same code as that used for y in terms of w.

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

t	T	$T = 80.445 - 4.289t$	ε
2	72	71.867	0.133
3	68	67.578	0.422
5	59	59	0
6	54	54.711	–0.711
7	50	50.422	–0.422
9	42	41.844	0.156
10	38	37.555	0.445

b No: the residuals are not randomly scattered about zero.

c $\text{RSS} = S_{TT} - \dfrac{(S_{tT})^2}{S_{tt}} = 957.43 - \dfrac{(-223)^2}{52} = 1.10$ (3 s.f.)

d The second reaction is most likely to have a linear fit since the RSS is lower.

10 a $S_{ss} = \sum s^2 - \dfrac{\left(\sum s\right)^2}{n} = 395.76 - \dfrac{53.4^2}{8} = 39.315$

$S_{sf} = \sum sf - \dfrac{\sum s \sum f}{n} = 171.66 - \dfrac{53.4 \times 29.9}{8} = -27.9225$

b $b = \dfrac{S_{sf}}{S_{ss}} = \dfrac{-27.9225}{39.315} = -0.71022\ldots = -0.710$ (3 s.f.)

$\bar{s} = \dfrac{\sum s}{n} = \dfrac{53.4}{8} = 6.675$ $\qquad$ $\bar{f} = \dfrac{\sum f}{n} = \dfrac{29.9}{8} = 3.7375$

$a = \bar{f} - b\bar{s} = 3.7375 - \dfrac{(-27.9225)}{39.315} \times 6.675 = 8.4782\ldots = 8.48$ (3 s.f.)

Hence the equation of the regression line of f on s is: $f = 8.48 - 0.710s$

c $f = 8.48 - 0.710 \times 7.5 = 3.2\,\text{mm}$ (2 s.f.)

d $S_{ff} = \sum f^2 - \dfrac{\left(\sum f\right)^2}{n} = 131.93 - \dfrac{29.9^2}{8} = 20.17875$

$\text{RSS} = S_{ff} - \dfrac{(S_{sf})^2}{Sss} = 20.17875 - \dfrac{(-27.9225)^2}{39.315} = 0.347$ (3 s.f.)

e $\sum \varepsilon = 0 \Rightarrow -0.177 - 0.196 + 0.256 + 0.124 + x - 0.129 - 0.216 - 0.032 = 0 \Rightarrow x = 0.37$

f A linear regression model is not suitable: the residuals go negative, positive, negative so they are not randomly scattered about zero.

11 a $b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{75}{60} = 1.25$

$\bar{x} = \dfrac{2+3+4+5+6+7+8+9+10}{9} = 6$ $\qquad$ $\bar{y} = \dfrac{4+5+7+8+9+11+12+11+15}{9} = \dfrac{82}{9}$

$a = \bar{y} - b\bar{x} = \dfrac{82}{9} - 1.25 \times 6 = 1.6111\ldots = 1.61$ (3 s.f.)

Hence the equation of the regression line of y on x is: $y = 1.61 + 1.25x$

b $\text{RSS} = S_{yy} - \dfrac{(S_{xy})^2}{S_{xx}} = 98.89 - \dfrac{75^2}{60} = 5.14$

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

x	y	$y = 1.61 + 1.25x$	ε
2	4	4.11	–0.11
3	5	5.36	–0.36
4	7	6.61	0.39
5	8	7.86	0.14
6	9	9.11	–0.11
7	11	10.36	0.64
8	12	11.61	0.39
9	11	12.86	–1.86
10	15	14.11	0.89

d The outlier is the wombat that is nine days old, i.e. data point (9, 11).

e i There is no correct answer.

It could be a recording error – so ignore this data point.

It could be a valid data point; the wombat could have been a 'runt' or underweight due to illness, for example – so include this reading.

ii Removing the data point gives the following summary statistics:

$S_{xx} = 49.875$ $\quad$ $S_{xy} = 68.625$ $\quad$ $\bar{x} = \dfrac{45}{8} = 5.625$ $\quad$ $\bar{y} = \dfrac{71}{8} = 8.875$

$b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{68.625}{49.875} = 1.37593\ldots = 1.38$ (3 s.f.)

$a = \bar{y} - b\bar{x} = 8.875 - \dfrac{68.625}{49.875} \times 5.625 = 1.13533\ldots = 1.14$ (3 s.f.)

So the equation of the regression line is now: $y = 1.14 + 1.38x$

iii Mass of wombat $= 1.135\ldots + 1.376\ldots \times 20 = 28.65\ldots = 28.7$ g (3 s.f.)

iv The estimate is unreliable since 20 days lies (well) outside the range of the data.

12 a $S_{tt} = \sum t^2 - \frac{\left(\sum t\right)^2}{n} = 42.33 - \frac{17.7^2}{8} = 3.16875$

$S_{ts} = \sum ts - \frac{\sum t \sum s}{n} = 42.16 - \frac{17.7 \times 17.5}{8} = 3.44125$

$b = \frac{S_{ts}}{S_{tt}} = \frac{3.44125}{3.16875} = 1.0859\ldots = 1.09$ (3 s.f.)

$\bar{t} = \frac{\sum t}{n} = \frac{17.7}{8} = 2.2125$ $\bar{s} = \frac{\sum n}{n} = \frac{17.5}{8} = 2.1875$

$a = \bar{s} + b\bar{t} = 2.1875 - \frac{3.44125}{3.16875} \times 2.2125 = -0.21526\ldots = -0.215$ (3 s.f.)

Hence the equation of the regression line of s on t is: $s = -0.215 + 1.09t$

b Predicted number of employees (s) $= (-0.215 + 1.09 \times 2.3) \times 100 = 229$ (to nearest whole number)

c $\sum \varepsilon = 0 \Rightarrow 0.0121 - 0.0137 - 0.0395 + 0.1347 - 0.0997 + p + 0.0745 + 0.0487 = 0 \Rightarrow p = -0.1171$

So $p = 0.117$ (3 s.f.)

d Linear model is suitable since the residuals are randomly scattered about zero.

e $S_{ss} = \sum s^2 - \frac{\left(\sum s\right)^2}{n} = 42.07 - \frac{17.5^2}{8} = 3.78875$

$\text{RSS} = S_{ss} - \frac{\left(S_{st}\right)^2}{S_{tt}} = 3.78875 - \frac{(3.44125)^2}{3.16875} = 0.0516$ (3 s.f.)

f The UK sample is likely to have a better linear fit since the RSS is much smaller.

Challenge

a $\varepsilon_i = y_i - (a + bx_i)$ where $i = 1, 2, \ldots, n$

$\sum \varepsilon = \sum (y_i - a - bx_i)$

Given that $a = \bar{y} - b\bar{x}$, this can be written as $\sum \varepsilon = \sum (y_i - \bar{y} + b\bar{x} - bx_i)$

$\sum (y_i - \bar{y} + b\bar{x} - bx_i) = \sum y_i - \bar{y}\sum 1 + b\bar{x}\sum 1 - b\sum x_i = \sum y_i - n\bar{y} + nb\bar{x} - b\sum x_i$

Now $\bar{y} = \frac{\sum y_i}{n}$ and $\bar{x} = \frac{\sum x_i}{n}$ so

$\sum \varepsilon = \sum y_i - n\frac{\sum y_i}{n} + nb\frac{\sum x_i}{n} - b\sum x_i = \sum y_i - \sum y_i + b\sum x_i - b\sum x_i = 0$

b A linear regression line can be calculated for any set of data, following a linear trend or otherwise, and since by definition the sum of the residuals is equal to zero, this tells you nothing about the actual data itself and so does not guarantee a linear fit. There may be outliers and there is no guarantee that the RSS is close to 0.