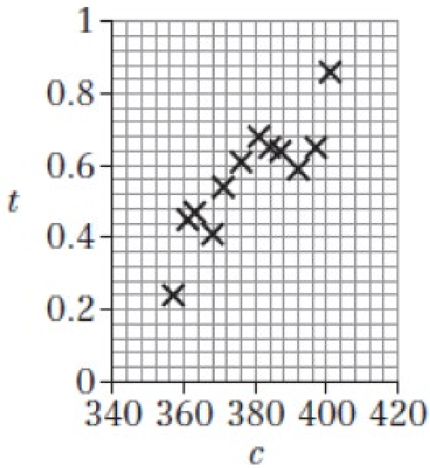
Chapter review 5

1 The data shows that the number of serious road accidents in a week strongly correlates with the number of fast food restaurants. However, it does not show whether the relationship is causal. Both variables could correlate with a third variable, e.g. the number of roads coming into a town.

2 a



b There is strong positive correlation.

c As mean CO_2 concentration in the atmosphere increased, mean temperatures also increased.

3 a There is strong positive correlation.

b If the number of items increases by 1, the time taken increases by approximately 2.64 minutes.

4 a $15.2 + 2 \times 11.4 = 38$

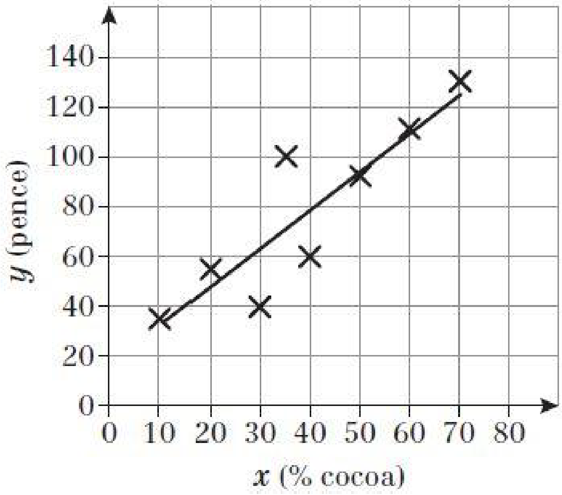
As $50 > 38$, $t = 50$ °C is an outlier.

b The outlier should be omitted, as it is very unlikely that the average temperature was 50 °C in a climate where people need to buy gloves, and so this data point is likely an anomaly.

c The equation of the regression line of t on g is $g = 99.6 - 5.2t$.

This means that for every increase in temperature of 1 °C, the shop sells 5.2 fewer pairs of pairs of gloves.

5 a and **b**



c Brand D is overpriced, since its price is much more than you would expect (the data point is far above the regression line).

d The regression equation should be used to predict a value for y given x, i.e. the price given the percentage of cocoa solids. So the student's method is a valid one.

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

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

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

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

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

7 a Calculating the summary statistics gives:

$\sum x^2 = 43\,622.85 \quad \sum x = 467.1 \quad \sum y = 7805 \quad \sum xy = 666\,045$

$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 43\,622.85 - \frac{467.1 \times 467.1}{8} = 16\,350.048\ldots = 16\,350 \text{ (5 s.f.)}$

$S_{xy} = 666\,045 - \frac{467.1 \times 7805}{8} = 210\,330.56\ldots = 210\,331 \text{ (6 s.f.)}$

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

$b = \frac{S_{xy}}{S_{xx}} = \frac{210\,330.56}{16\,350.048} = 12.8642\ldots = 12.86 \text{ (4 s.f.)}$

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

Equation is: $y = 224.5 + 12.86x$

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

7 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.

8 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: $\boldsymbol{y = 0.343 + 0.449x}$

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.)

9 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

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

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

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

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.)

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

Equation is: $y = 7.31 + 0.527x$ (parameters to 3 s.f.)

9 c $\frac{w}{400} = 7.3108\ldots + 0.52702\ldots(n-10)$ (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.

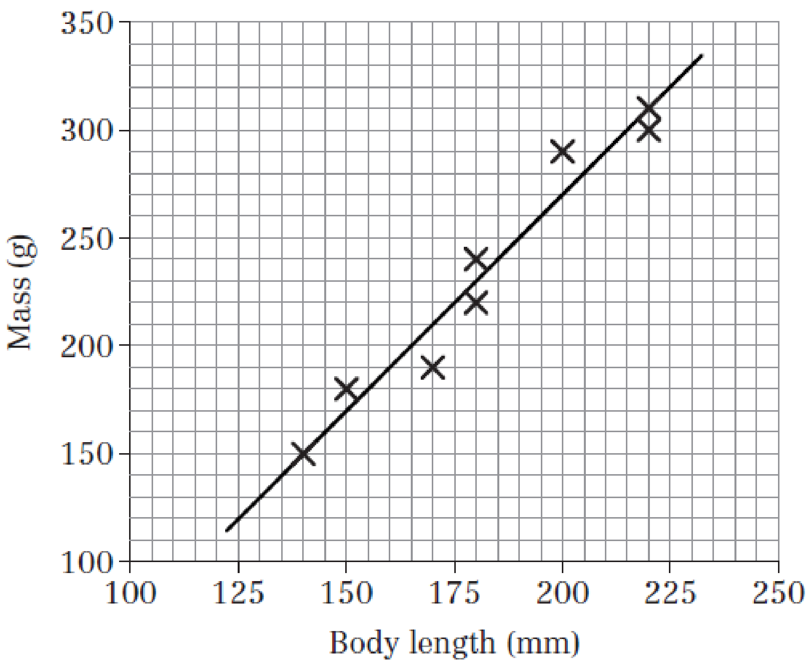
10 a The figure of 0.79 is the average amount of food consumed (in kg) in 1 week by 1 hen.

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 $= \frac{39.66}{10} \times 12 = €47.592 = €47.59$

11 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.

11 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

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

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

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

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

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

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

Equation is: $\boldsymbol{w = -12.7 + 1.98l}$

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

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

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

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

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

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

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

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

13 $\bar{x} = \frac{\sum x}{20} = 4.535 \Rightarrow \sum x = 4.535 \times 20 = 90.7$

$\bar{t} = \frac{\sum t}{20} = 15.15 \Rightarrow \sum t = 15.15 \times 20 = 303$

$$r = \frac{S_{xt}}{\sqrt{S_{xx}S_{tt}}} = \frac{\sum xt - \frac{\sum x \sum t}{n}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum t^2 - \frac{\left(\sum t\right)^2}{n}\right)}}$$

$$= \frac{1433.8 - \frac{(90.7)(303)}{20}}{\sqrt{\left(493.77 - \frac{90.7^2}{20}\right)\left(4897 - \frac{303^2}{20}\right)}} = 0.375 \text{ (3 s.f.)}$$

14 a $S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 465 - \frac{67 \times 67}{10} = 16.1$

$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 429 - \frac{65 \times 65}{10} = 6.5$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 434 - \frac{67 \times 65}{10} = -1.5$

$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-1.5}{\sqrt{16.1 \times 6.5}} = \frac{-1.5}{10.2298\ldots} = -0.1466\ldots = -0.147 \text{ (3 s.f.)}$

b The coding is linear, so the product moment correlation coefficient will be unaffected by the coding. So the product moment correlation coefficient between s and a is –0.147.

c This is a weak negative correlation that is close to 0. There is little evidence to suggest that students in the group who are good at science will also be good at art.

15 a $S_{jj} = \sum j^2 - \frac{\left(\sum j\right)^2}{n} = 52\,335 - \frac{979 \times 979}{20} = 4412.95$

$S_{pp} = \sum p^2 - \frac{\left(\sum p\right)^2}{n} = 32\,156 - \frac{735 \times 735}{20} = 5144.75$

$S_{jp} = \sum jp - \frac{\sum j \sum p}{n} = 39\,950 - \frac{979 \times 735}{20} = 3971.75$

b $r = \frac{S_{jp}}{\sqrt{S_{jj}S_{pp}}} = \frac{3971.75}{\sqrt{4412.95 \times 5144.75}} = \frac{3971.75}{4764.8215} = 0.8335\ldots = 0.834 \text{ (3 s.f.)}$

c There is a strong positive correlation between the amount of juice and the cost, as the product moment correlation coefficient is close to 1. So Nimer is correct.

16 a $S_{pp} = \sum p^2 - \frac{\left(\sum p\right)^2}{n} = \sum (x-10)^2 - \frac{\left(\sum (x-10)\right)^2}{n}$

$= \sum x^2 - 20\sum x + 100n - \frac{\left(\left(\sum x\right) - 10n\right)^2}{n}$

$= \sum x^2 - 20\sum x + 100n - \frac{\left(\sum x\right)^2 - 20n\sum x + 100n^2}{n}$

$= \sum x^2 - 20\sum x + 100n - \frac{\left(\sum x\right)^2}{n} + 20\sum x - 100n$

$= \sum x^2 - \frac{\left(\sum x\right)^2}{n} = S_{xx}$

b $S_{qq} = \sum q^2 - \frac{\left(\sum q\right)^2}{n} = 77.0375 - \frac{\left(\sum \frac{1}{20} y\right)^2}{n} = 77.0375 - \frac{\left(\sum y\right)^2}{400n}$

$= 77.0375 - \frac{491^2}{400 \times 8} = 1.69968\ldots = 1.70 \text{ (3 s.f.)}$

$r = \frac{S_{pq}}{\sqrt{S_{pp} S_{qq}}} = \frac{-11.625}{\sqrt{85.5 \times 1.69968\ldots}} = -0.964 \text{ (3.s.f)}.$

c The coding is linear, so the product moment correlation coefficient will be unaffected by the coding. So the product moment correlation coefficient between x and y is –0.964.

d The correlation coefficient suggests a strong negative linear correlation, but the scatter diagram shows a non-linear fit.

Challenge

a $\sum x = 104.5,\ \sum y = 113.6,\ \sum x^2 = 1954.1,\ \sum y^2 = 2100.6$

The regression line of x on y is of the form $x = a + by$ where

$$b = \frac{S_{xy}}{S_{yy}},\ S_{xy} = \sum xy - \frac{\sum x \sum y}{n},\ S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} \text{ and } n = 10$$

The gradient of the regression line of x on y is 0.8, therefore,

$$\frac{S_{xy}}{S_{yy}} = 0.8$$

$$\sum xy - \frac{\sum x \sum y}{n} = 0.8\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)$$

$$\sum xy = 0.8\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right) + \frac{\sum x \sum y}{n}$$

$$= 0.8\left(2100.6 - \frac{113.6^2}{10}\right) + \frac{104.5 \times 113.6}{10}$$

$$= 1835.203...$$

$$= 1835 \text{ (to the nearest whole number)}$$

b $y = 3.50 + 0.725x$

The regression line of y on x is of the form $y = a + bx$ where

$$a = \bar{y} - b\bar{x} \text{ and } b = \frac{S_{xy}}{S_{xx}}$$

$$\frac{S_{xy}}{S_{xx}} = 0.725$$

$$S_{xy} = 0.725S_{xx}$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$= 1954.1 - \frac{104.5^2}{10}$$

$$= 862.075$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

$$= 2100.6 - \frac{113.6^2}{10}$$

$$= 810.104$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$$= \frac{0.725S_{xx}}{\sqrt{S_{xx}S_{yy}}}$$

$$= \frac{0.725 \times 862.075}{\sqrt{862.075 \times 810.104}}$$

$$= 0.74789...$$

$$= 0.748 \text{ (3 s.f.)}$$