Correlation 2A

1 $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \dfrac{100}{\sqrt{92 \times 112}} = \dfrac{100}{101.50862} = 0.985 \text{ (3 s.f.)}$

2 $S_{xx} = \sum x^2 - \dfrac{\left(\sum x\right)^2}{n} = 33845 - \dfrac{367 \times 367}{6} = 33845 - 22\,448.166\ldots = 11396.833\ldots$

$S_{yy} = \sum y^2 - \dfrac{\left(\sum y\right)^2}{n} = 12976 - \dfrac{270 \times 270}{6} = 12976 - 12150 = 826$

$S_{xy} = \sum xy - \dfrac{\sum x \sum y}{n} = 17135 - \dfrac{367 \times 270}{6} = 17135 - 16515 = 620$

$r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \dfrac{620}{\sqrt{11396.833 \times 826}} = \dfrac{620}{3068.189} = 0.202 \text{ (3 s.f.)}$

3 **a** $S_{aa} = \sum a^2 - \dfrac{\left(\sum a\right)^2}{n} = 1899 - \dfrac{115 \times 115}{7} = 9.7142\ldots = 9.71 \text{ (3 s.f.)}$

b $r = \dfrac{S_{ah}}{\sqrt{S_{aa}S_{hh}}} = \dfrac{72.1}{\sqrt{9.7142... \times 571.4}} = 0.96774\ldots = 0.968 \text{ (3 s.f.)}$

c There is positive correlation. The greater the age of the person, the taller the person.

4 **a** Calculating the summary statistics gives:

$\sum l = 26.8 \qquad \sum l^2 = 150.02 \qquad \sum t = 47.4 \qquad \sum t^2 = 399.58 \qquad \sum lt = 237.07$

$S_{ll} = 150.02 - \dfrac{26.8 \times 26.8}{6} = 150.02 - 119.7066\ldots = 30.3133\ldots = 30.3 \text{ (3 s.f.)}$

$S_{tt} = 399.58 - \dfrac{47.4 \times 47.4}{6} = 399.58 - 374.46 = 25.12$

$S_{lt} = 237.06 - \dfrac{26.8 \times 47.4}{6} = 237.07 - 211.72 = 25.35$

b $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \dfrac{25.35}{\sqrt{30.3133\ldots \times 25.12}} = \dfrac{25.35}{27.5947\ldots} = 0.91865\ldots = 0.919 \text{ (3 s.f.)}$

c The data in the scatter graph appear to be linear, and the correlation coefficient found in part **b** is close to 1. Therefore, a linear regression model is suitable to model the data.

5 a $S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 120\,123 - \frac{973 \times 973}{8} = 120\,123 - 118\,341.125 = 1781.875$

$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 33\,000 - \frac{490 \times 490}{8} = 33\,000 - 30\,012.5 = 2987.5$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 61\,595 - \frac{973 \times 490}{8} = 61\,595 - 59\,596.25 = 1998.75$

$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{1998.75}{\sqrt{1781.875 \times 2987.5}} = \frac{1998.75}{2307.2389} = 0.86629\ldots = 0.866$ (3 s.f.)

b The correlation is positive. The higher the IQ, the higher the mark gained in the general knowledge test. (Alternatively, the higher the mark gained in the intelligence test, the higher the IQ.)

6 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 0.973.

7 a This is the coded data set:

p	0	5	3	2	1
q	0	17	12	10	6

b Calculating summary statistics for the coded data gives:

$\sum p = 11 \quad \sum p^2 = 39 \quad \sum q = 45 \quad \sum q^2 = 569 \quad \sum pq = 147$

$S_{pp} = \sum p^2 - \frac{\left(\sum p\right)^2}{n} = 39 - \frac{11 \times 11}{5} = 14.8$

$S_{qq} = \sum q^2 - \frac{\left(\sum q\right)^2}{n} = 569 - \frac{45 \times 45}{5} = 164$

$S_{pq} = \sum pq - \frac{\sum p \sum q}{n} = 147 - \frac{11 \times 45}{5} = 48$

$r = \frac{S_{pq}}{\sqrt{S_{pp}S_{qq}}} = \frac{48}{\sqrt{14.8 \times 164}} = 0.97429\ldots = 0.974$ (3 s.f.)

c The coding is linear. The product moment correlation coefficient is independent of the linear coding, hence it is 0.974 (3 s.f.).

8 a This is the coded data set:

p	10	8	11	9	12
t	4	3	5	4	6

Calculating summary statistics for the coded data gives:

$\sum p = 50 \quad \sum p^2 = 510 \quad \sum t = 22 \quad \sum t^2 = 102 \quad \sum pt = 227$

$$S_{pp} = \sum p^2 - \frac{\left(\sum p\right)^2}{n} = 510 - \frac{50 \times 50}{5} = 10$$

$$S_{tt} = \sum t^2 - \frac{\left(\sum t\right)^2}{n} = 102 - \frac{22 \times 22}{5} = 5.2$$

$$S_{pt} = \sum pt - \frac{\sum p \sum t}{n} = 227 - \frac{50 \times 22}{5} = 7$$

b $r = \dfrac{S_{pt}}{\sqrt{S_{pp}S_{tt}}} = \dfrac{7}{\sqrt{10 \times 5.2}} = \dfrac{7}{7.2111\ldots} = 0.97072\ldots = 0.971$ (3 s.f.)

c The coding is linear. The product moment correlation coefficient is independent of the linear coding, hence it is 0.971 (3 s.f.).

9 a This is the coded data set:

x	15	37	5	0	45	27	20
y	30	13	34	43	20	14	0

Calculating summary statistics for the coded data gives:

$\sum x = 149 \quad \sum x^2 = 4773 \quad \sum y = 154 \quad \sum y^2 = 4670 \quad \sum xy = 2379$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 4773 - \frac{149 \times 149}{7} = 1601.4285\ldots = 1601 \text{ (4 s.f.)}$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 4670 - \frac{154 \times 154}{7} = 1282$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 2379 - \frac{149 \times 154}{7} = -899$$

b $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \dfrac{-899}{\sqrt{1601.4285 \times 1282}} = \dfrac{-899}{1432.84\ldots} = -0.62742\ldots = -0.627$ (3 s.f.)

c The shopkeeper is not correct. There is negative correlation, so as the newspaper sales go up the sweet sales go down.

10 a $S_{ff} = \sum f^2 - \frac{\left(\sum f\right)^2}{n} == \sum (10x)^2 - \frac{\left(\sum 10x\right)^2}{8} = 10^2\left(\sum x^2 - \frac{\left(\sum x\right)^2}{8}\right)$

$= 100S_{xx} = 100 \times 111.48 = 11148$

b $S_{gg} = \sum g^2 - \frac{\left(\sum g\right)^2}{n} = 74\,458.75 - \frac{\left(\sum 5(y+10)\right)^2}{n} = 74\,458.75 - \frac{\left(5\sum y + 50 \times n\right)^2}{n}$

$= 74\,458.75 - \frac{(5 \times 70.9 + 50 \times 8)^2}{8} = 3299.97$

$r = \frac{S_{fg}}{\sqrt{S_{ff}S_{gg}}} = \frac{5667.5}{\sqrt{11148 \times 3299.97}} = 0.934 \text{ (3 s.f.)}$

c The product moment correlation coefficient shows strong linear correlation. However, the scatter diagram suggests a non-linear fit.

11 a $S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 22.02 - \frac{12^2}{7} = 1.44857\ldots$

$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 1491.69 - \frac{97.7^2}{7} = 128.077\ldots$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 180.37 - \frac{12 \times 97.7}{7} = 12.8842\ldots$

$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{12.884\ldots}{\sqrt{1.4485\ldots \times 128.077\ldots}} = 0.946 \text{ (3 s.f.)}$

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

x	y	$y = -1.2905 + 8.8945x$	ε
1.1	6.2	8.49345	–2.29345
1.3	10.5	10.27235	0.22765
1.4	12	11.1618	0.8382
1.7	15	13.83015	1.16985
1.9	17	15.60905	1.39095
2.1	18	17.38795	0.61205
2.5	19	20.94575	–1.94575

c The linear model might not be a good model for this data, as the residuals do not appear to be randomly scattered about zero.