Linear regression 1A

1 $b = \frac{S_{xy}}{S_{xx}} = \frac{90}{15} = 6$

$a = \bar{y} - b\bar{x} = 15 - (6 \times 3) = -3$

2 $b = \frac{S_{xy}}{S_{xx}} = \frac{165}{30} = 5.5$

$a = \bar{y} - b\bar{x} = 8 - (5.5 \times 4) = 8 - 22 = -14$

Equation is: $y = -14 + 5.5x$

3 $b = \frac{S_{xy}}{S_{xx}} = \frac{80}{40} = 2$ $b = \frac{80}{40} = 2$

$a = \bar{y} - b\bar{x} = 12 - (2 \times 6) = 0$

Equation is: $y = 2x$

4 a $\bar{x} = \frac{\sum x}{n} = \frac{10}{4} = 2.5$

$\bar{y} = \frac{\sum y}{n} = \frac{48}{4} = 12$

$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 30 - \frac{10 \times 10}{4} = 30 - 25 = 5$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 140 - \frac{10 \times 48}{4} = 140 - 120 = 20$

b $b = \frac{S_{xy}}{S_{xx}} = \frac{20}{5} = 4$

$a = \bar{y} - b\bar{x} = 12 - (4 \times 2.5) = 12 - 10 = 2$

Equation is: $y = 2 + 4x$

5 a Calculating the summary statistics gives:

$\sum x = 29$ $\sum x^2 = 209$ $\sum y = 48$ $\sum xy = 348$

$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 209 - \frac{29^2}{5} = \frac{1045 - 841}{5} = \frac{204}{5} = 40.8$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 348 - \frac{29 \times 48}{5} = \frac{1740 - 1392}{5} = \frac{348}{5} = 69.6$

5 b $\bar{x} = \dfrac{\sum x}{n} = \dfrac{29}{5} = 5.8 \qquad \bar{y} = \dfrac{\sum y}{n} = \dfrac{48}{5} = 9.6$

$b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{69.6}{40.8} = 1.70588\ldots = 1.71 \text{ (3 s.f.)}$

$a = \bar{y} - b\bar{x} = 9.6 - (1.7059 \times 5.8) = -0.294 \text{ (3 s.f.)}$

Equation is: $y = -0.294 + 1.71x$

6 a $y = -59 + (57 \times 6) = 283$

b The value of 57 is the gradient of the regression line. For every unit increase in someone's dexterity score, that person's productivity rises by 57.

c i It may be a little unreliable to use the equation to work out the productivity of someone with dexterity of 2, since 2 lies just outside the values in the table. It would involve extrapolation.

ii It may be very unreliable to use the equation to work out the productivity of someone with dexterity of 14, since 14 lies well outside the range of the values in the table. It would involve extrapolation.

7 $S_{hh} = \sum h^2 - \dfrac{\left(\sum h\right)^2}{n} = 45.04 - \dfrac{22.09 \times 22.09}{12} = 45.04 - 40.6640\ldots = 4.37599\ldots = 4.376 \text{ (4 s.f.)}$

$S_{hg} = \sum hg - \dfrac{\sum h \sum g}{n} = 97.778 - \dfrac{22.09 \times 49.7}{12} = 97.778 - 91.48941\ldots = 6.28858\ldots = 6.286 \text{ (4 s.f.)}$

$\bar{h} = \dfrac{\sum h}{n} = \dfrac{22.09}{12} = 1.841 \text{ (4 s.f.)}$

$\bar{g} = \dfrac{\sum g}{n} = \dfrac{49.7}{12} = 4.141 \text{ (4 s.f.)}$

$b = \dfrac{S_{hg}}{S_{hh}} = \dfrac{6.286}{4.376} = 1.43647\ldots = 1.44 \text{ (3 s.f.)}$

$a = \bar{g} - b\bar{h} = 4.141 - (1.436 \times 1.841) = 1.4973\ldots = 1.50 \text{ (3 s.f.)}$

So the equation is: $g = 1.50 + 1.44h$

8 a $S_{wp} = \sum wp - \dfrac{\sum w \sum p}{n} = 6797 - \dfrac{186 \times 397}{10} = -587.2$

$S_{ww} = \sum w^2 - \dfrac{\left(\sum w\right)^2}{n} = 3886 - \dfrac{186^2}{10} = 3886 - 3459.6 = 426.4$

$b = \dfrac{S_{wp}}{S_{ww}} = -\dfrac{587.2}{426.4} = -1.37711\ldots = -1.38$ (3 s.f.)

$\overline{p} = \dfrac{\sum p}{n} = \dfrac{397}{10} = 39.7$

$\overline{w} = \dfrac{\sum w}{n} = \dfrac{186}{10} = 18.6$

$a = \overline{p} - b\overline{w} = 39.7 + \dfrac{587.2}{426.4} \times 18.6 = 65.31425\ldots = 65.3$ (3 s.f.)

Hence equation of regression line of p on w is: $p = 65.3 - 1.38w$

b $p = (65.3142\ldots) - (1.3771\ldots)w \Rightarrow w = \dfrac{65.3142}{1.3771} - \dfrac{1}{1.3771}p$

This gives (to 3 s.f.) the equation: $w = 47.4 - 0.726p$

c The w on p regression line is calculated using different summary statistics rather than just the reciprocal of the summary statistics used for the p on w regression line.

For example, the gradient of the w on p regression line is obtained from $d = \dfrac{S_{wp}}{S_{pp}}$

But the gradient in the equation is part **b** is obtained from $\dfrac{1}{b} = \dfrac{S_{ww}}{S_{wp}}$

d i The p on w regression line, since a p value is required for a given w value.

ii The w on p regression line, since a w value is required for a given p value.

9 a Calculating the summary statistics gives:

$\sum x = 1650 \quad \sum x^2 = 258500 \quad \sum y = 374 \quad \sum y^2 = 14712 \quad \sum xy = 52870$

$S_{xy} = \sum xy - \dfrac{\sum x \sum y}{n} = 52870 - \dfrac{1650 \times 374}{11} = -3230$

$S_{xx} = \sum x^2 - \dfrac{\left(\sum x\right)^2}{n} = 258500 - \dfrac{(1650)^2}{11} = 11000$

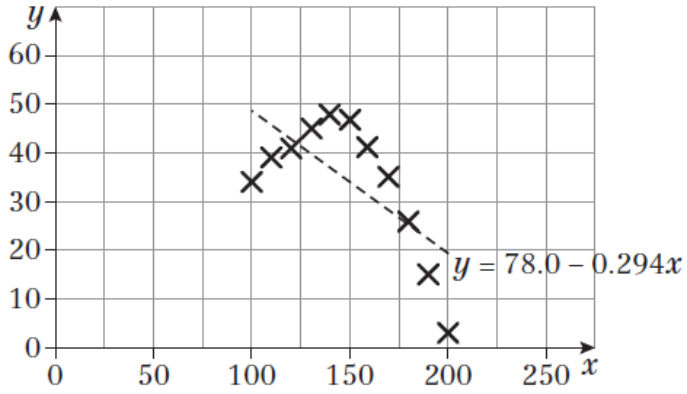
$b = \dfrac{S_{xy}}{S_{xx}} = -\dfrac{3230}{11000} = -0.29363\ldots = -0.294$ (3 s.f.) (3 s.f.)

$\overline{x} = \dfrac{1650}{11} = 150 \qquad \overline{y} = \dfrac{374}{11} = 34$

$a = \overline{y} - b\overline{x} = 34 + \dfrac{3230}{11000} \times 150 = 78.0454\ldots = 78.0$ (3 s.f.)

Hence equation of regression line of y on x is: $y = 78.0 - 0.294x$

9 b



c The model is not valid since the data does not follow a linear pattern.

10 a There are 12 data points, so:

$$S_{nn} = \sum n^2 - \frac{\left(\sum n\right)^2}{12} = 30\,786 - \frac{540 \times 540}{12} = 30\,786 - 24\,300 = 6486$$

$$S_{np} = \sum np - \frac{\sum n \sum p}{12} = 41\,444 - \frac{540 \times 780}{12} = 41\,444 - 35\,100 = 6344$$

b $\overline{n} = \frac{\sum n}{12} = \frac{540}{12} = 45 \quad \overline{p} = \frac{\sum p}{12} = \frac{780}{12} = 65$

$$b = \frac{S_{np}}{S_{nn}} = \frac{6344}{6486} = 0.97810\ldots = 0.978 \text{ (3 s.f.)}$$

$$a = \overline{p} - b\overline{n} = 65 - (0.9781 \times 45) = 20.98519\ldots = 21.0 \text{ (3 s.f.)}$$

So the equation is: $p = 21.0 + 0.978n$

c $p = 21.98\ldots + 0.9781\ldots \times 40 = 60.10$ (3 s.f.)
Hence the production costs of 40 000 items is £60 100 (3 s.f.).

d This estimate is reliable since 40 000 items lies within the range of the data.

11 a There are 10 data points, so:

$$S_{nn} = \sum n^2 - \frac{\left(\sum n\right)^2}{10} = 1844 - \frac{(112)^2}{10} = 589.6$$

$$S_{np} = \sum np - \frac{\sum n \sum p}{10} = 6850 - \frac{112 \times 480}{10} = 1474$$

b $\overline{n} = \frac{\sum n}{10} = \frac{112}{10} = 11.2 \quad \overline{p} = \frac{\sum p}{10} = \frac{480}{10} = 48$

$$b = \frac{S_{np}}{S_{nn}} = \frac{1474}{589.6} = 2.5$$

$$a = \overline{p} - b\overline{n} = 48 - 2.5 \times 11.2 = 20$$

Hence the equation of the regression line of p on n is $p = 20 + 2.5n$

11 c The increase in cost, in pounds, for every 100 leaflets printed.

d $20 + 2.5t < 5t \Rightarrow 2.5t > 20 \Rightarrow t > 8$

The first printing company is cheaper than the rival when printing more than 800 leaflets.

12 a Calculator gives the equation of the regression line as $y = -0.07 + 1.45x$.

b The value b (1.45) is the additional protection given (in number of years) for each additional coat of paint.

c The model would be unsuitable because 7 years lies outside the range of the data. The equation would also give a non-integer solution, but it is only possible to apply a whole number of coats.

d $y = -0.07 + 1.45 \times 7 = 10.1$ years (3 s.f.)

e i Calculator gives the new equation of the regression line as $y = 0.478 + 1.25x$ (3 s.f.).

ii $y = 0.478 + 1.25 \times 7 = 9.23$ years (2 s.f.)

iii The original prediction in part **d** was extrapolated. This result uses interpolation. More data generally gives a more accurate regression model.