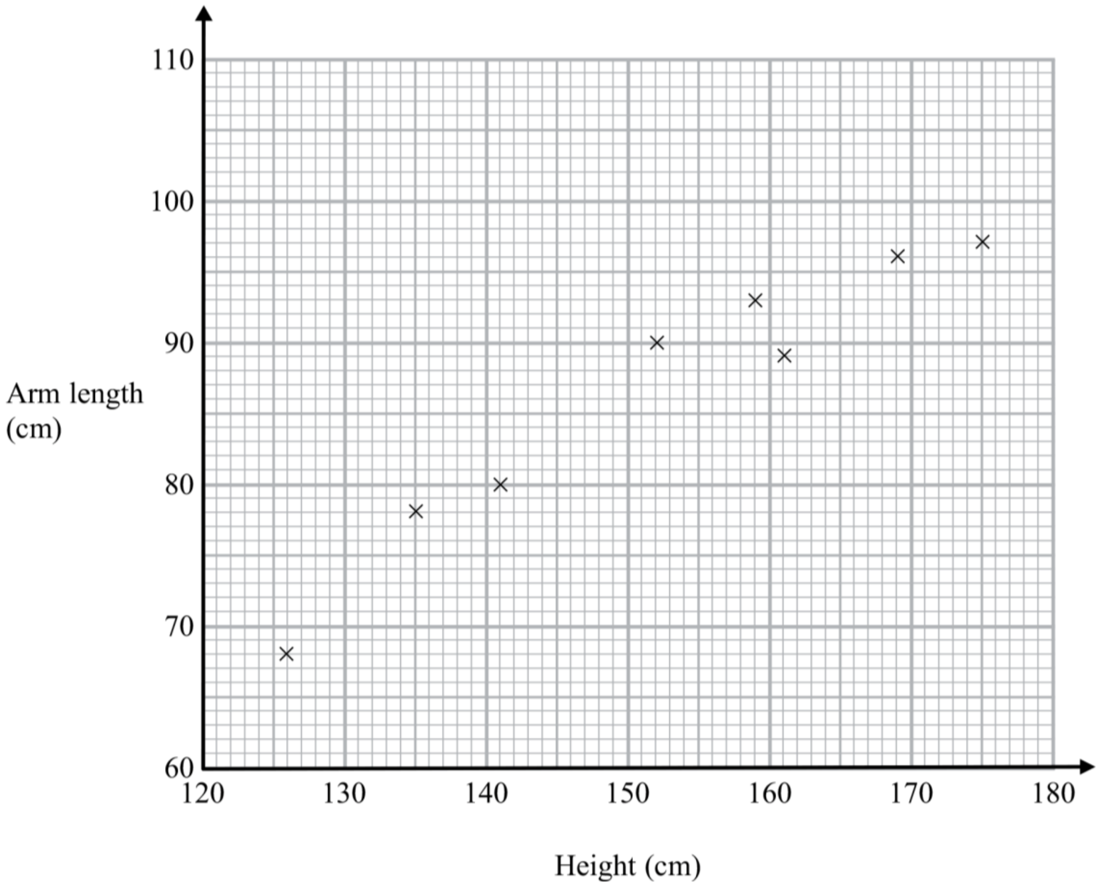
Scatter Graphs (and Least Squares Regression Lines)

The scatter graph shows information about the height and the arm length of some students in Year 11.



i. On the scattergraph, plot the remaining information from the table below.

Height (cm)	126	132	135	141	152	159	161	166	169	175
Arm length (cm)	68	75	78	80	90	93	89	93	96	97

ii. What type of correlation does this scatter graph show?

iii. Draw in a line of best fit

A different student in Year 11 has a height of 148 cm.

iv. Estimate the arm length of this student.

(Exam question adapted from Edexcel higher tier paper 1, November 2012)

x_i	126	132	135	141	152	159	161	166	169	175
x_i^2										
y_i	68	75	78	80	90	93	89	93	96	97
y_i^2										
$x_i\, y_i$										

Find:

a) $\bar{x}$

b) $\bar{y}$

c) $\sum x_i$

d) $\sum y_i$

e) $\sum {x_i}^2$

f) $\sum {y_i}^2$

g) $\sum x_i y_i$

h) $\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}$

i) $\sum {y_i}^2 - \frac{(\sum y_i)^2}{n}$

j) $\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$

Correlation and regression

For a set of n pairs of values (x_i, y_i)

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

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

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

The product moment correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{\sum (x_i - \bar{x})^2\right\}\left\{\sum (y_i - \bar{y})^2\right\}}} = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)}}$$

Spearman's rank correlation coefficient is the product moment correlation coefficient between ranks

When there are no tied ranks it may be calculated using $r_s = 1 - \dfrac{6\sum d_i^2}{n(n^2 - 1)}$

The regression coefficient of y on x is $b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

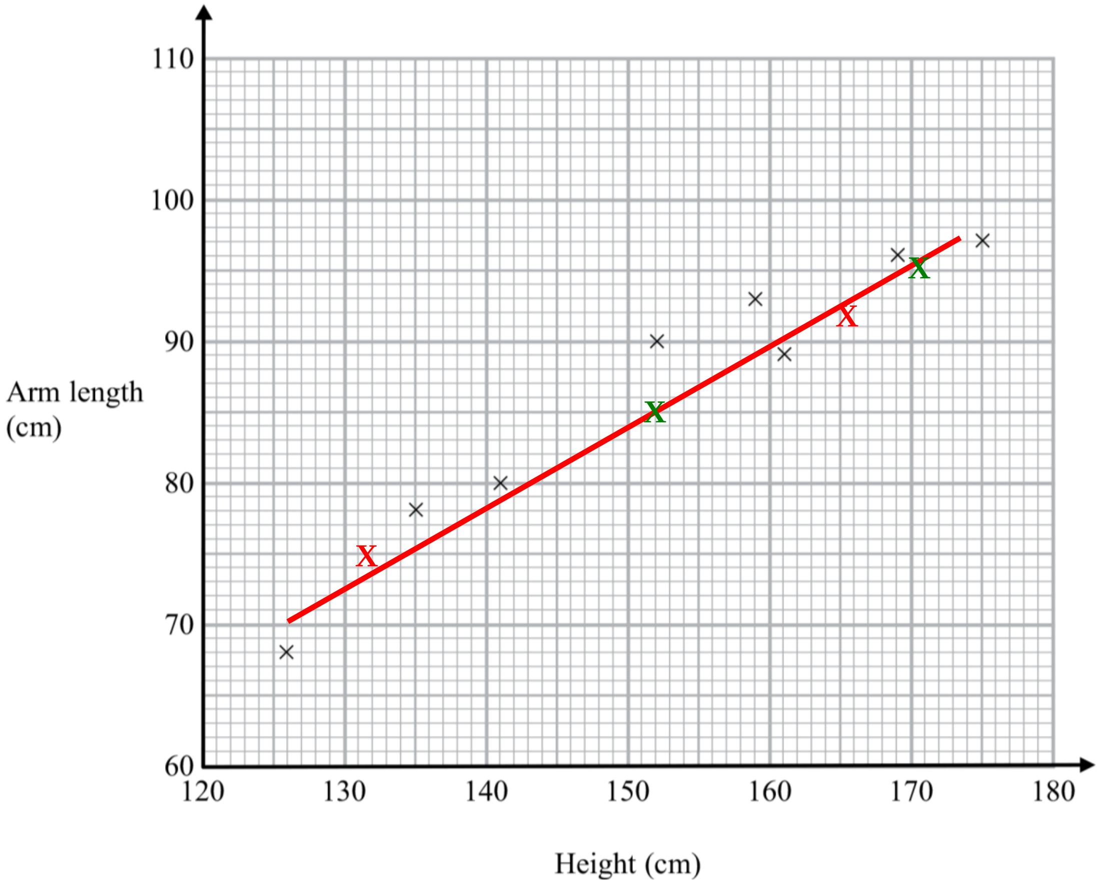
Least squares regression line of y on x is $y = a + bx$, where $a = \bar{y} - b\bar{x}$

Use the values that you found in parts (a)-(j) together with the excerpt of the formula book above to find;

1. The product moment correlation coefficient, r, of x and y.
2. The regression coefficient, b, of y on x.
3. The equation of the least squares regression line of y on x, in the form $y = a + bx$.
4. Use your values of $\bar{x}$ and $\bar{y}$ to plot the coordinate $(\bar{x}, \bar{y})$ on the scatter graph.
5. Use your equation of the least squares regression line to find a corresponding y value for $x = 170$ and plot this coordinate on the scatter graph. Draw in the least squares regression line of y on x. Compare this with your original line of best fit.

Scatter Graphs (and Least Squares Regression Lines) - Answers

The scatter graph shows information about the height and the arm length of some students in Year 11.



i. On the scattergraph, plot the remaining information from the table below.

Height (cm)	126	132	135	141	152	159	161	166	169	175
Arm length (cm)	68	75	78	80	90	93	89	93	96	97

ii. What type of correlation does this scatter graph show? (strong) positive

iii. Draw in a line of best fit

A different student in Year 11 has a height of 148 cm.

iv. Estimate the arm length of this student. 84 cm

(Exam question adapted from Edexcel higher tier paper 1, November 2012)

x_i	126	132	135	141	152	159	161	166	169	175
x_i^2										
y_i	68	75	78	80	90	93	89	93	96	97
y_i^2										
$x_i\, y_i$										

Find:

a) $\bar{x} = 151.6$

b) $\bar{y} = 85.9$

c) $\sum x_i = 1516$

d) $\sum y_i = 859$

e) $\sum {x_i}^2 = 232454$

f) $\sum {y_i}^2 = 74677$

g) $\sum x_i y_i = 131711$

h) $\sum {x_i}^2 - \frac{(\sum x_i)^2}{n} = 2628.4$

i) $\sum {y_i}^2 - \frac{(\sum y_i)^2}{n} = 888.9$

j) $\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 1486.6$

Correlation and regression

For a set of n pairs of values (x_i, y_i)

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

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

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

The product moment correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{\sum (x_i - \bar{x})^2\right\}\left\{\sum (y_i - \bar{y})^2\right\}}} = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)}}$$

Spearman's rank correlation coefficient is the product moment correlation coefficient between ranks

When there are no tied ranks it may be calculated using $r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$

The regression coefficient of y on x is $b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

Least squares regression line of y on x is $y = a + bx$, where $a = \bar{y} - b\bar{x}$

Use the values that you found in parts (a)-(j) together with the excerpt of the formula book above to find;

1. The product moment correlation coefficient, r, of x and y. 0.973
2. The regression coefficient, b, of y on x. 0.566
3. The equation of the least squares regression line of y on x, in the form $y = a + bx$. $y = 0.156 + 0.566x$
4. Use your values of $\bar{x}$ and $\bar{y}$ to plot the coordinate $(\bar{x}, \bar{y})$ on the scatter graph. $(151.6,\ 85.9)$
5. Use your equation of the least squares regression line to find a corresponding y value for $x = 170$ and plot this coordinate on the scatter graph. Draw in the least squares regression line of y on x. Compare this with your original line of best fit. $y = 0.156 + 0.566 \times 170 = 96.376$