# *Correlation and regression*

## *Introduction*

Here we will be plotting points to see if there is any linear relationship between two variables (*eg* pints of beer drunk per week and life expectancy). We will then seek to measure the strength of this relationship and draw a line to represent it and make predictions.

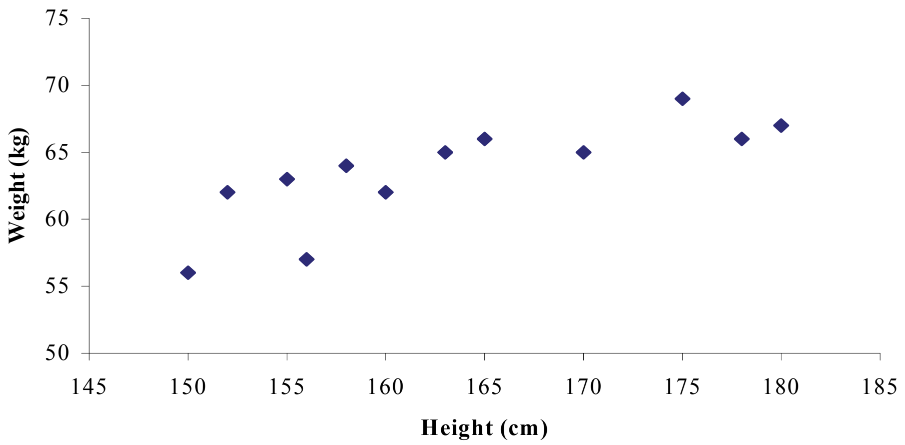
## *Scatterplots*

The table below shows the heights (in cm) and weights (in kg) of 12 individuals (A – L):

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Height (cm)** | 150 | 152 | 155 | 156 | 158 | 160 | 163 | 165 | 170 | 175 | 178 | 180 |
| **Weight (kg)** | 56 | 62 | 63 | 57 | 64 | 62 | 65 | 66 | 65 | 69 | 66 | 67 |

These data are called **bivariate data** as we have two variables (height and weight) for each individual.

A **scatterplot** (or scatter diagram) is simply a plot of our bivariate data, with one variable (*eg* height) plotted on the *x*-axis and the other variable (*eg* weight) plotted on the *y*-axis.



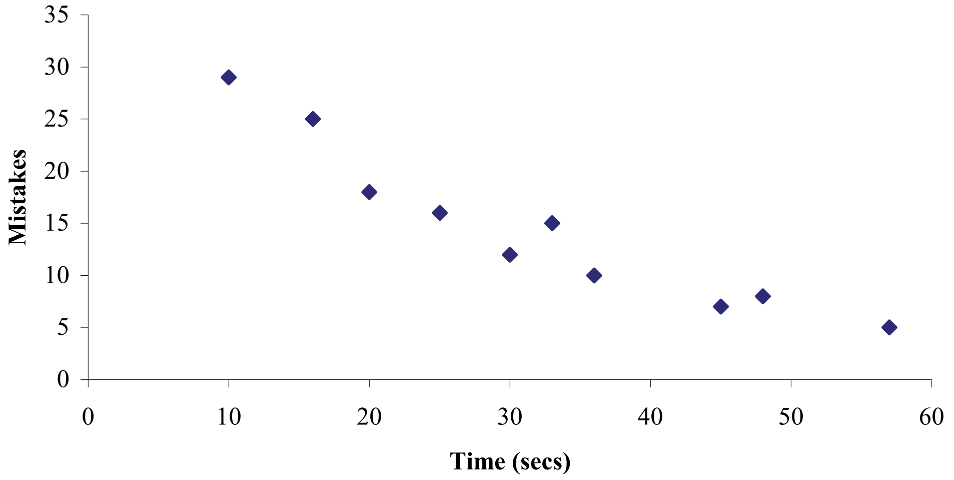
We use a scatterplot to see if there is any kind of relationship or connection between the two variables.

In this case we can see that there *is* a connection between height and weight. We have an *increasing* pattern in the points – generally the taller a person is, the more they weigh. Since this is not an exact relationship the points plotted are scattered, hence the name scatterplot.

Here we will only be looking to see if there is a *linear* relationship between the two variables.

***Question 1.1***

An experiment was carried out where a person had to draw a shape whilst looking in a mirror. The time taken to draw the shape (in seconds) and the number of mistakes were recorded. A scatterplot for 10 individual's results is shown:



Describe the relationship (if any) between the time taken and the number of mistakes made.

The variable on the *x*-axis is called the **explanatory variable** whereas the variable on the *y*-axis is called the **response variable**. These names arise from the use of scatterplots in experiments.

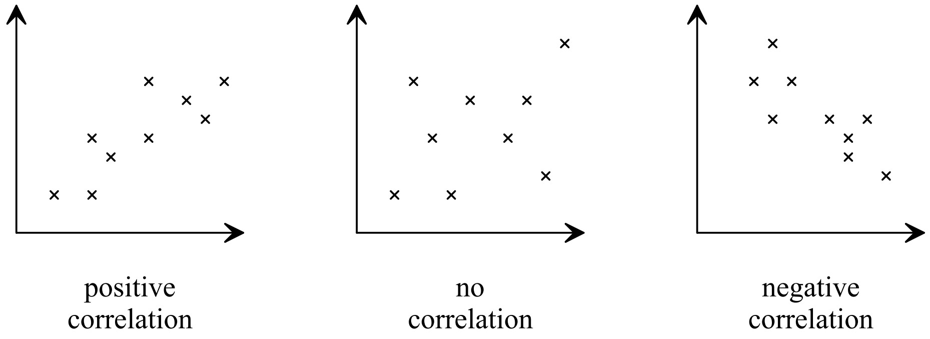
For example, suppose we measure the length of a spring when we hang weights on it. The weights we hang on the spring are chosen by the experimenter – so they can be 'explained', hence they are the *explanatory* variable. Once the weights are hung on we can then see the effect or response in the length of the spring, hence the length of the spring is the *response* variable.

## *Correlation*

The mathematical term for the relationship or connection between two variables (*eg* height and weight) is **correlation**.

### *Types of linear correlation*

There are three types of linear correlation between two variables.



**Positive correlation** - as one variable increases so does the other variable. This gives a positive sloping (*ie* upward) graph.

*For example, we saw that there was positive correlation between people's height and their weight. As height increases so does their weight.*

**Negative correlation** - as one variable increases, the other variable decreases. This gives a negative sloping (*ie* downward) graph.

*For example, we saw in* ***Error! Reference source not found.*** *that there was negative correlation between the time people took to complete the drawing and the number of mistakes made. As the time increases so the number of mistakes decrease.*

**No correlation** - there is no connection between the two variables.

*For example, there would be no correlation between the height of a person and the amount paid for their insurance.*

***Question 1.2***

For each pair of variables, state the type of correlation that would be shown:

(i) life expectancy and the number of cigarettes smoked per day

(ii) distance lived from work and the time taken to get there

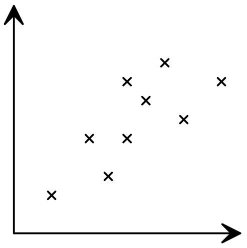
(iii) number of bedrooms and cost of home insurance

(iv) amount of 'no claims' discount and cost of car insurance

(v) number of exams passed in a sitting and length of hair.

## *Strength of linear correlation*

In addition to observing the type of correlation between two variables we can also look at how strong the connection or relationship is between them. The strength of the connection or correlation can be seen in how clear the pattern is:

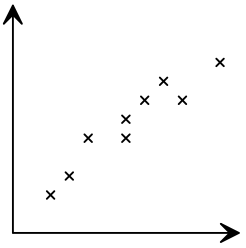


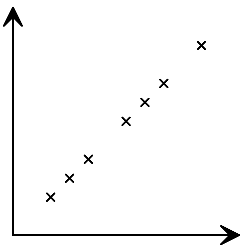
In this scatterplot we can see that there is positive correlation, but the pattern is not very clear (*ie* the points are quite scattered).

We say that there is **weak positive correlation**.

In this scatterplot there is again positive correlation, but this time the pattern is much clearer (*ie* the points are less scattered and more linear).

We say that there is **strong positive correlation**.





We now have the positive correlation with the pattern being a perfect straight line. There is an *exact* linear relationship between the two variables.

We say that there is **perfect positive correlation**.

***Question 1.3***

Sketch scatterplots showing:

(i) weak negative correlation

(ii) strong negative correlation

(iii) perfect negative correlation.

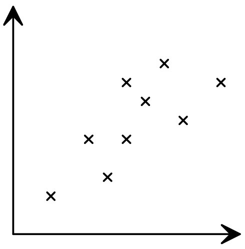
***Question 1.4***

Match each of these pairs of variables to one of the scatterplots shown below:

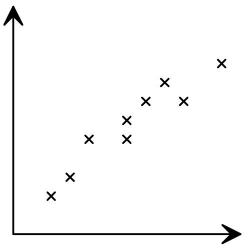
(i) pounds exchanged and dollars bought on a single day

(ii) height and shoe size of an individual

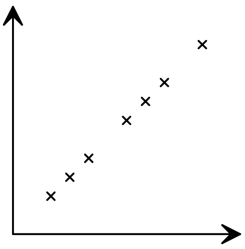
(iii) size of car engine and cost of car insurance.



Scatterplot A



Scatterplot B



Scatterplot C

## *Covariance*

We can now draw a scatterplot, state the type of linear correlation shown and describe how strong that correlation is. However, it would be nice to have a single *numerical* value to summarise the type and strength of the correlation. To do this we first need to define the sample covariance.

Recall that the sample variance of $x_1, \dots, x_n$ is:

$$\frac{1}{n-1}\sum(x_i-\overline{x})^2$$

Similarly, we could calculate the sample variance of the *y* values:

$$\frac{1}{n-1}\sum(y_i-\overline{y})^2$$

But what we want is some way to measure how *y* varies with *x* – this is called the sample **covariance** and is defined by:

$$\frac{1}{n-1}\sum(x_i-\overline{x})(y_i-\overline{y})$$

Taking the previous height and weight data:

| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Height (cm)** | 150 | 152 | 155 | 156 | 158 | 160 | 163 | 165 | 170 | 175 | 178 | 180 |
| **Weight (kg)** | 56 | 62 | 63 | 57 | 64 | 62 | 65 | 66 | 65 | 69 | 66 | 67 |

The mean of the heights is:

$$\overline{x}=\frac{150+\cdots+180}{12}=\frac{1,962}{12}=163.5 \text{ cm}$$

The mean of the weights is:

$$\overline{y}=\frac{56+\cdots+67}{12}=\frac{762}{12}=63.5 \text{ kg}$$

Now using the formula, we get a covariance of:

$$=\tfrac{1}{11}\left[(150-163.5)(56-63.5)+\cdots+(180-163.5)(67-63.5)\right]$$

$$=\tfrac{1}{11}\left[101.25+\cdots+57.75\right]$$

$$=\tfrac{1}{11}\times 344$$

$$=31\tfrac{3}{11}\text{ cmkg}$$

This is a positive result as there is positive correlation between these variables.

Now the formula we are using is fine for a small list of numbers, but is extremely tedious for this list of 12 numbers (and would be a real pain if the means were awkward numbers). So we are going to rearrange the formula into a nicer format.

Recall that the sample variance of $x_1,\ldots,x_n$ can be rewritten as:

$$\frac{1}{n-1}\left(\sum x_i^2-n\overline{x}^2\right)$$

Similarly, the sample variance of the *y* values can be rewritten as:

$$\frac{1}{n-1}\left(\sum y_i^2-n\overline{y}^2\right)$$

Using a similar method, we can rewrite the covariance as:

$$\frac{1}{n-1}\left(\sum x_i y_i-n\overline{x}\,\overline{y}\right)$$

The proof of this result can be found in Appendix A.

***Definition***

The **sample covariance** of $(x_1,y_1),\ldots,(x_n,y_n)$ is given by:

$$\frac{1}{n-1}\left(\sum x_i y_i-n\overline{x}\overline{y}\right)$$

***Question 1.5***

A garage is selling a particular make of used car. The table below shows the asking price and the mileage for five of these cars:

| | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **Mileage (000’s)** | 16 | 25 | 38 | 61 | 79 |
| **Price (£000’s)** | 7.8 | 5.2 | 4.3 | 2.5 | 1.7 |

(i) Show that the sample covariance for these data is $-59.175$.

(ii) What does this value tell us about the type of correlation shown?

Note how the covariance gives a positive answer for our positive correlation between height and weight and a negative answer for our negative correlation between price and mileage.

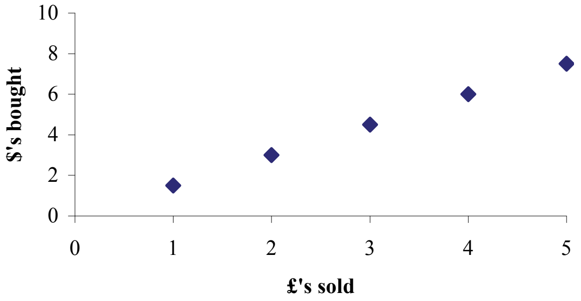
***Question 1.6***

What value will the covariance take if there is no correlation between two variables?

Appendix B explains why the covariance works to give us a positive answer for positive correlation and a negative answer for negative correlation.

## *The correlation coefficient*

The scatterplot below shows the number of dollars purchased in exchange for sterling.



$\overline{x} = £3 \quad \overline{y} = \$4.50$

$s_x = \sqrt{2.5} \simeq £1.58$

$s_y = \sqrt{5.625} \simeq \$2.37$

**covariance = £$3.75**

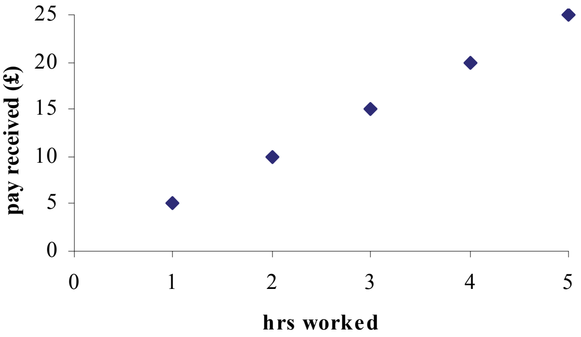
The next scatterplot shows the pay received for the number of hours worked.

$\overline{x} = 3\,\text{hrs} \quad \overline{y} = £15$

$s_x = \sqrt{2.5} \simeq 1.58\,\text{hrs}$

$s_y = \sqrt{62.5} \simeq £7.91$

**covariance = 12.5 £hrs**



Both of these scatterplots show perfect positive correlation. However, they have very different covariances. They also have different units for their covariances.

The difference in the size of the answers is to do with the spread of the results. This makes the covariance pretty unhelpful when comparing the strength of correlation for two separate cases, as does having different units.

What we need to do is to standardise the covariance so that any scatterplots with the same degree of correlation have the same value (regardless of the spread of the data). We also need to get rid of the units so we just have a number (*ie* a **coefficient**).

Recall that we standardised a normal distribution by subtracting the mean (which we have already done when calculating the covariance) and dividing by the standard deviation. Well here we will divide by *both* standard deviations. This gives us the **correlation coefficient**, *r*.

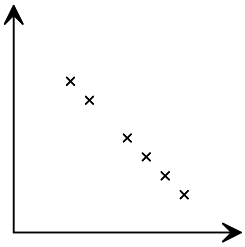
So the correlation coefficient, *r*, for the dollars purchased and sterling sold is:

$$r=\frac{\text{covariance of } X \text{ and } Y}{sd(X)\times sd(Y)}=\frac{3.75}{\sqrt{2.5}\sqrt{5.625}}=1$$

***Question 1.7***

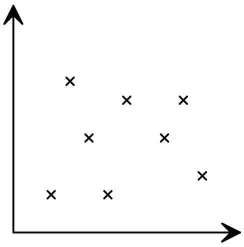
Calculate the correlation coefficient for the second scatterplot, pay received for the number of hours worked.

The value of the correlation coefficient ranges from $-1$ to 1 as follows:



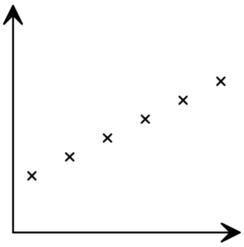
perfect negative correlation

$r = -1$



no correlation

$r = 0$

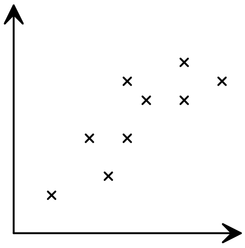


perfect positive correlation

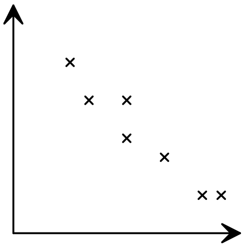
$r = +1$

***Question 1.8***

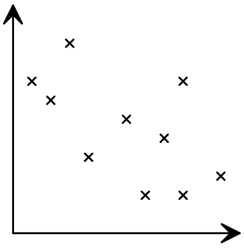
Give an approximate value for the correlation coefficient for each of these:



Scatterplot A



Scatterplot B



Scatterplot C

### *The formula*

The correlation coefficient was defined to be:

$$r=\frac{\text{covariance of } X \text{ and } Y}{sd(X)\times sd(Y)}$$

Using the definitions of the sample standard deviation and covariance, we get:

$$r=\frac{\frac{1}{n-1}\left(\sum x_i y_i-n\overline{xy}\right)}{\sqrt{\frac{1}{n-1}\left(\sum x_i^2-n\overline{x}^2\right)\frac{1}{n-1}\left(\sum y_i^2-n\overline{y}^2\right)}}$$

However the $\frac{1}{n-1}$’s all cancel so we get:

$$r=\frac{\sum x_i y_i-n\overline{xy}}{\sqrt{\left(\sum x_i^2-n\overline{x}^2\right)\left(\sum y_i^2-n\overline{y}^2\right)}}$$

Since we don’t need the $\frac{1}{n-1}$’s, it seems a bit pointless calculating them in the first place. We just need to calculate the **sum of squares**:

$$s_{xx}=\sum x_i^2-n\overline{x}^2$$

$$s_{yy}=\sum y_i^2-n\overline{y}^2$$

$$s_{xy}=\sum x_i y_i-n\overline{xy}$$

This gives us:

***Definition***

The sample **correlation coefficient** of $(x_1,y_1),\dots,(x_n,y_n)$ is given by:

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

### *Correlation and causation*

So far we have used scatterplots to see if there is any connection between two variables. We can then quantify the type and strength of that relationship using the correlation coefficient.

For example, gas and electricity bills from a particular household over the last few years have a correlation coefficient of 0.8. There is strong positive correlation between the amount charged on each of the bills. This means that as the gas bill increases so does the electricity bill for that household. BUT is the increase in the gas bill the *cause* of the increase in the electricity bill?

Clearly not. Both of them are due to the seasons – in summer we will use less gas for heating and less electricity for lights, whereas in winter we will use more of both. This is the difference between correlation (*ie* how they change together) and the *cause* of that correlation. Just because variables are correlated doesn't necessarily imply that one changing causes the other to change – there might some other variable causing them to change together.

***Question 1.9***

This table below compares the literacy rate (*x*) and life expectancy (*y*) of men in various countries:

**t●Algeria●Angola●Ireland●Bangladesh●Bolivia●Iran●●Literacy rate**

| | **Algeria** | **Angola** | **Ireland** | **Bangladesh** | **Bolivia** | **Iran** |
|---|---|---|---|---|---|---|
| **Literacy rate (%)** | 64 | 56 | 99 | 57 | 85 | 89 |
| **Life expectancy (yrs)** | 68 | 45 | 74 | 58 | 60 | 69 |

$\sum x = 450 \quad \sum x^2 = 35{,}428 \quad \sum y = 374 \quad \sum y^2 = 23{,}850 \quad \sum xy = 28{,}745$

(i) Show that the sample correlation coefficient for these data is 0.732.

(ii) What correlation is shown by the value in part (i)?

(iii) Is there cause and effect between literacy and life expectancy (*ie* does reading improve your life expectancy)? If so, explain how. If not, state the variable which is causing both of these to change together.

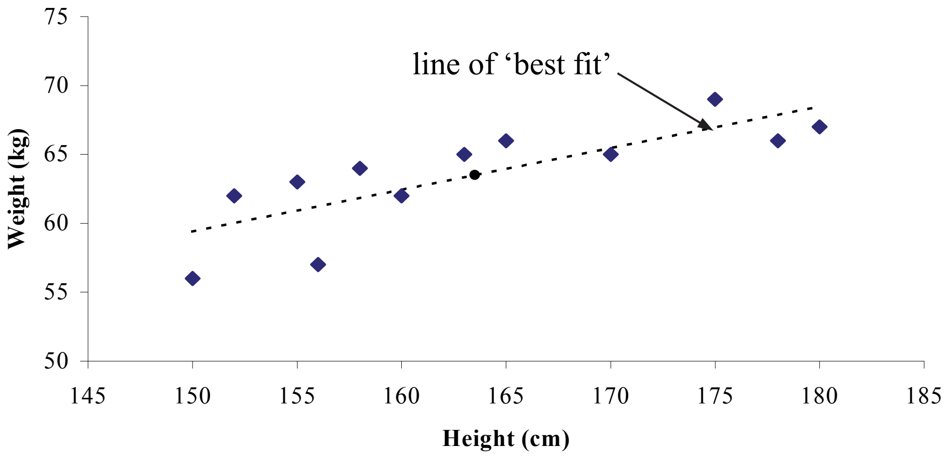
## *Regression*

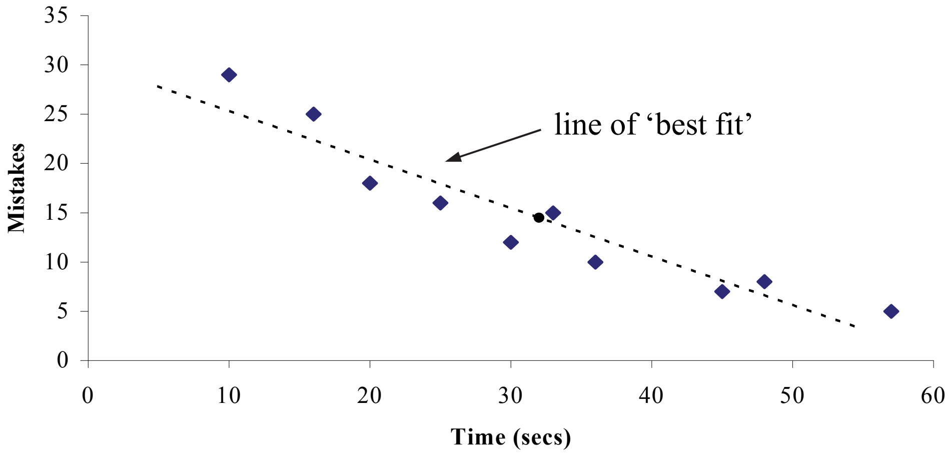
Once we have drawn a scatterplot and shown that there is a (strong) linear relationship between the two variables, we can attempt to represent that linear relationship by drawing a line.

If we have perfect correlation the line would pass through all the points. If we don't have a perfect relationship then the line just shows the general pattern of the points.

### *Line of 'best fit'*

The line of 'best fit' is drawn on the scatterplot by hand to show the relationship between the two variables. We would expect it to have the same slope as the pattern and to have roughly the same number of points on both sides of the line. We would also expect it to go through mean $(\bar{x}, \bar{y})$ of the co-ordinates. For example:

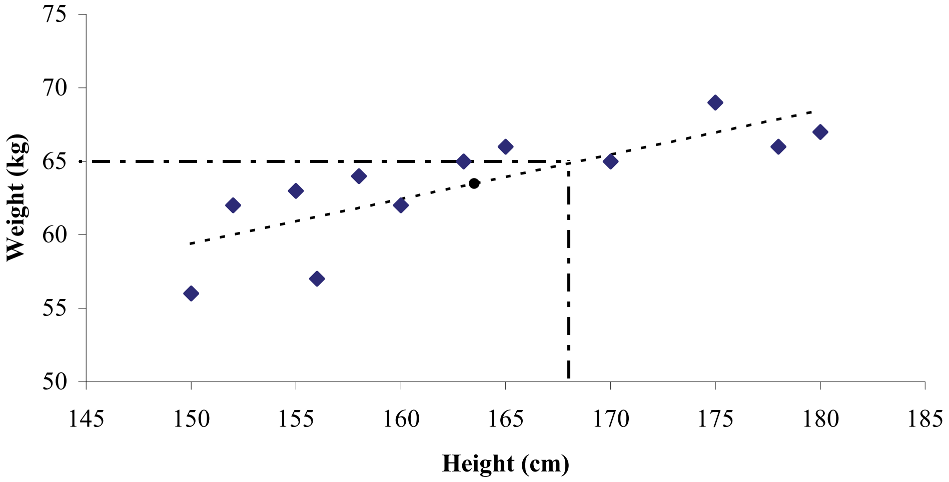




The method for drawing a line of 'best fit' by eye is:

- Plot the mean of the co-ordinates $(\overline{x}, \overline{y})$
- Place your ruler through the mean point and turn it so that it has the same slope as the pattern of points and has roughly the same number of points on either side
- Draw the line.

Once we have drawn the line of 'best fit' we can use it to read off predicted values. For example, we could estimate the height of someone who weighs 65 kg:



We can see, using the line of 'best fit', that someone who weighs 65kg would be expected to have a height of about 168 cm.

The accuracy of this estimate depends largely on how strong the correlation is.

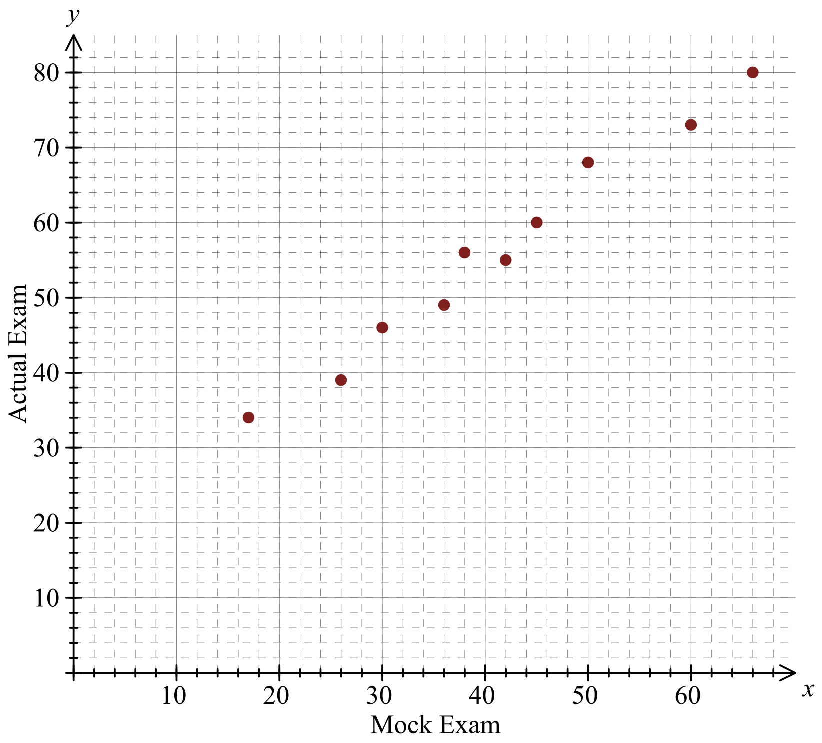
***Question 1.10***

The marks of 10 students in their mock exam and their actual exam are as follows:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Mock Exam (*x*)** | 36 | 50 | 17 | 42 | 38 | 66 | 30 | 60 | 26 | 45 |
| **Actual Exam (*y*)** | 49 | 68 | 34 | 55 | 56 | 80 | 46 | 73 | 39 | 60 |

$$\sum x = 410 \quad \sum x^2 = 18{,}850 \quad \sum y = 560 \quad \sum y^2 = 33{,}308 \quad \sum xy = 24{,}934$$

A scatterplot of these data is given below:



(i) Plot the mean of the co-ordinates $(\bar{x}, \bar{y})$.

(ii) Draw the line of 'best fit'.

(iii) Hence, estimate the final exam score for a student who obtained 56 in their mock exam.

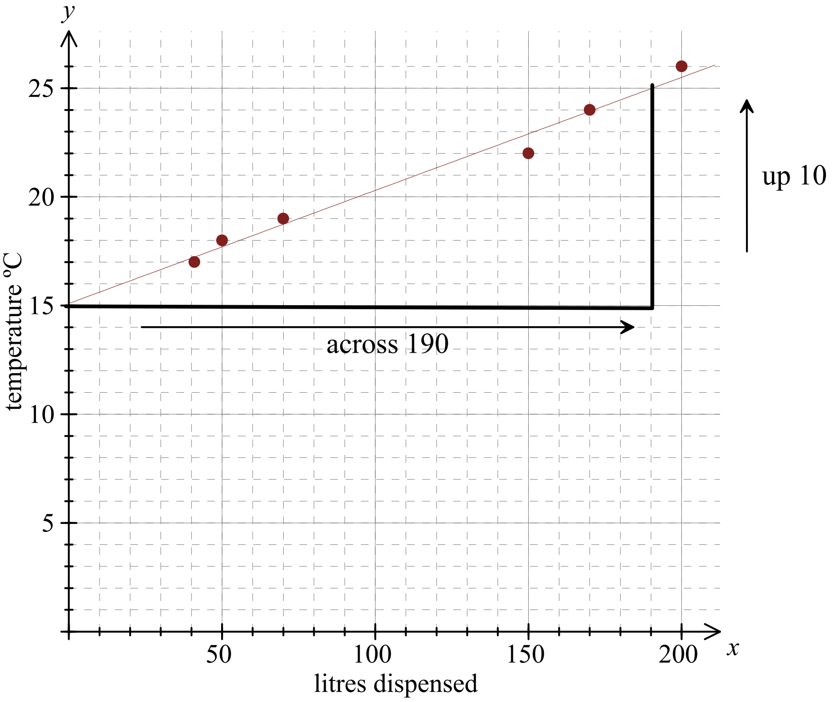
It is extremely useful to calculate the equation of our line of 'best fit'. This allows us to make predictions without referring to the graph.

The equation of a straight line is:

$$y = \alpha + \beta x$$

where $\beta$ is the gradient (*ie* how many units 'up' the line goes for every one unit 'across') and $\alpha$ is the *y*-intercept (*ie* where the line crosses the *y*-axis).

For example, a scatterplot of the number of litres of water dispensed per week from an office water machine against air temperature is shown below:



The *y*-intercept is where the graph crosses the *y*-axis:

$$\alpha = y\text{-intecept} = 15$$

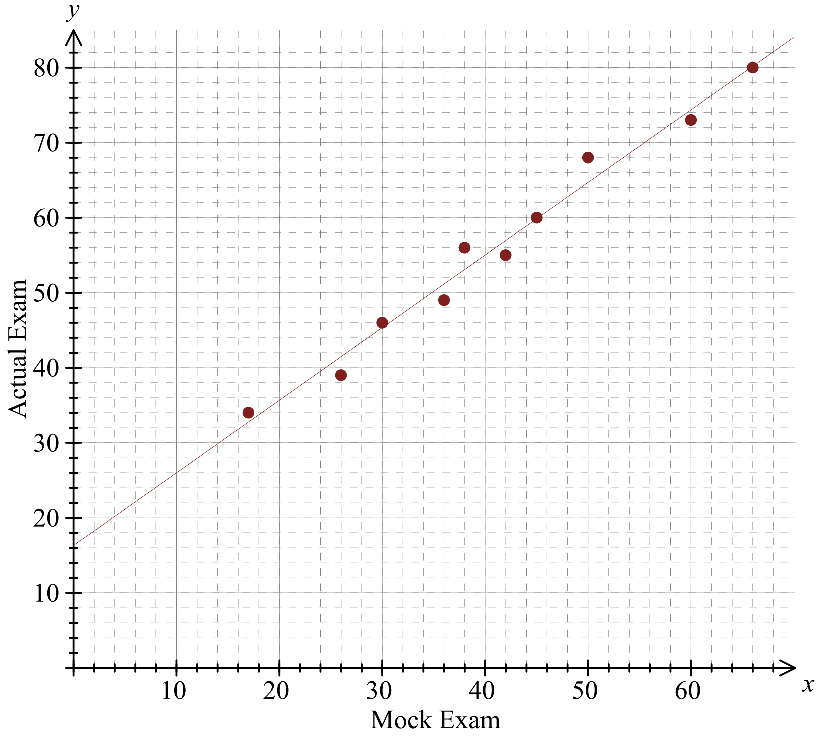
The gradient is the how many units 'up' the line goes for every one unit 'across':

$$\beta = \text{gradient} = \frac{\text{up}}{\text{across}} = \frac{10}{190} = 0.0526$$

So we get $y = 15 + 0.0526x$.

***Question 1.11***

Here is the line of 'best fit' for the mock and exam results from Question 1.10:



By finding the *y*-intercept and gradient, write down the equation of this line of 'best fit'.

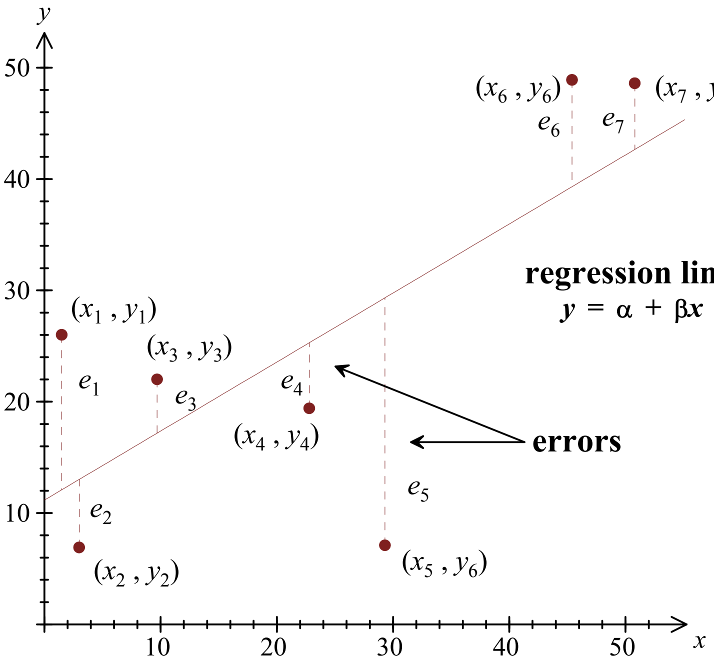
At school it is likely you would have used the notation $y = mx + c$ where *m* was the gradient and *c* was the *y*-intercept.

## *Regression line*

The line of 'best fit' drawn by eye is certainly not the most accurate way to obtain a line that represents the correlation. Nor is finding the equation of the line by looking at the graph an accurate method.

What we need is a way to *calculate* the *y*-intercept and the gradient just using the points themselves. The line of 'best fit' obtained using this method is called the **regression line** as we are working backwards (*ie* regressing) from the points to get the equation of the line.

The method involves considering how far 'out' each of the points is from our regression line:



If we had perfect correlation, all the points would lie on the regression line:

$$y_1 = \alpha + \beta x_1$$

$$y_2 = \alpha + \beta x_2$$

$$\ldots$$

But since we don't, the first point $(x_1, y_1)$ is a vertical distance of $e_1$ 'out' from the regression line, the second point $(x_2, y_2)$ is a vertical distance of $e_2$ 'out' from the regression line, and so on. So, we actually have:

$$y_1 = \alpha + \beta x_1 + e_1$$

$$y_2 = \alpha + \beta x_2 + e_2$$

$$\ldots$$

The $e_i$'s are called the errors or the **residuals** – these simply tell us how much our actual *y* value is 'out' from our *y* value on the regression line.

Now to make the regression line fit these points as closely as possible we would like to make all these errors as small as possible. Since we don't care whether the errors are positive or negative, we shall make their squares as small as possible:

$$\min \quad \sum e_i^2$$

So what we are going to do is to find the values of $\alpha$ and $\beta$ that make this sum of squares as small as possible. The values obtained are therefore called the **least squares estimates** of $\alpha$ and $\beta$.

Now to find the $\alpha$ and $\beta$ that minimises $\sum e_i^2$ we will differentiate it with respect to $\alpha$ and $\beta$. But first we need to get some $\alpha$'s and $\beta$'s in this equation. Rearranging our expressions above we get:

$$e_1 = y_1 - \alpha - \beta x_1$$

$$e_2 = y_2 - \alpha - \beta x_2$$

$$\ldots$$

So we want:

$$\min \quad \sum (y_i - \alpha - \beta x_i)^2$$

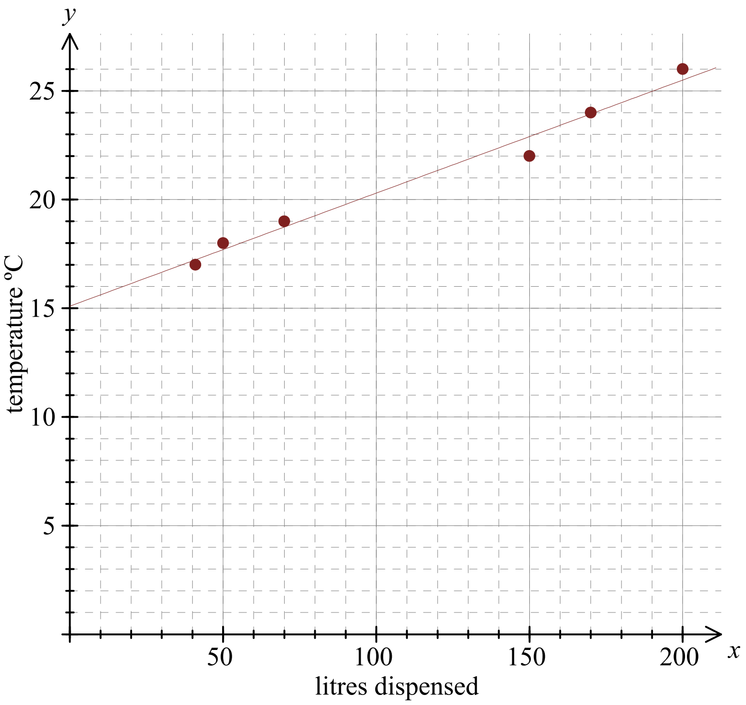
Differentiating this expression with respect to $\alpha$ and $\beta$, and setting equal to zero (to get the minimum), we get:

$$\hat{\alpha} = \overline{y} - \beta\overline{x}$$

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}}$$

The complete details of this proof are given in Appendix C. Note the little hats are a mathematical way of saying this is our *estimate* of the value (rather than the *true* value).

Let's calculate the least squares estimates of $\alpha$ and $\beta$ for our scatterplot of the litres of water dispensed against the temperature outside.



The values used in this graph were:

| **Litres dispensed** | 41 | 70 | 150 | 50 | 170 | 200 |
|---|---|---|---|---|---|---|
| **Temperature (ºC)** | 17 | 19 | 22 | 18 | 24 | 26 |

$$\sum x = 681 \quad \sum x^2 = 100,481 \quad \sum y = 126 \quad \sum y^2 = 2,710 \quad \sum xy = 15,507$$

Now:

$$\bar{x} = \frac{\sum x}{n} = \frac{681}{6} = 113.5 \text{ and } \quad \bar{y} = \frac{\sum y}{n} = \frac{126}{6} = 21$$

We also need the sum of squares:

$$s_{xx} = \sum x_i^2 - n\bar{x}^2 = 100,481 - 6 \times 113.5^2 = 23,187.5$$

$$s_{xy} = \sum x_i y_i - n\overline{xy} = 15,507 - 6 \times 113.5 \times 21 = 1,206$$

Hence:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{1,206}{23,187.5} = 0.0520$$

and:

$$\hat{\alpha} = \bar{y} - \beta\bar{x} = 21 - 0.0520 \times 113.5 = 15.1$$

So our fitted regression line is:

$$\hat{y} = 15.1 + 0.0520x$$

This is similar (but obviously more accurate) than our line of 'best fit' by eye which was $y = 15 + 0.0526x$ .

***Question 1.12***

The table below compares the GDP (billion US$) and child mortality rates (deaths of under 5's per 1,000 births) for various countries:

| | **Algeria** | **Angola** | **Ireland** | **Bangladesh** | **Bolivia** |
|---|---|---|---|---|---|
| **GDP (*x*)** | 45.9 | 7.4 | 72.7 | 32.8 | 8.1 |
| **Child mortality (*y*)** | 52 | 191 | 6 | 104 | 84 |

$\sum x = 166.9 \quad \sum x^2 = 8{,}588.31 \quad \sum y = 437 \quad \sum y^2 = 57{,}093 \quad \sum xy = 8{,}328$

(i) Show that the least squares estimates of $\alpha$ and $\beta$ are 156.6 and –2.074, respectively.

(ii) Write down the equation of the fitted regression line.

(iii) Use this line to estimate the child mortality for a country with a GDP of 60 billion US$.

## *Appendix A – rearranging the covariance formula*

The formula for the sample covariance is:

$$\frac{1}{n-1}\sum(x_i-\overline{x})(y_i-\overline{y})$$

Multiplying out the brackets and splitting up the sum:

$$=\frac{1}{n-1}\sum(x_iy_i-x_i\overline{y}-\overline{x}y_i+\overline{xy})$$

$$=\frac{1}{n-1}\left(\sum x_iy_i-\sum x_i\overline{y}-\sum\overline{x}y_i+\sum\overline{xy}\right)$$

We can take the $\overline{x}$ and $\overline{y}$ terms out of the sums, as they are constants (*ie* they don't depend on *i*):

$$=\frac{1}{n-1}\left(\sum x_iy_i-\overline{y}\sum x_i-\overline{x}\sum y_i+\overline{xy}\sum 1\right)$$

$$=\frac{1}{n-1}\left(\sum x_iy_i-\overline{y}\sum x_i-\overline{x}\sum y_i+n\overline{x}\,\overline{y}\right)$$

Now we use the fact that:

$$\overline{x}=\frac{1}{n}\sum x_i \quad\Rightarrow\quad \sum x_i=n\overline{x}$$

$$\overline{y}=\frac{1}{n}\sum y_i \quad\Rightarrow\quad \sum y_i=n\overline{y}$$

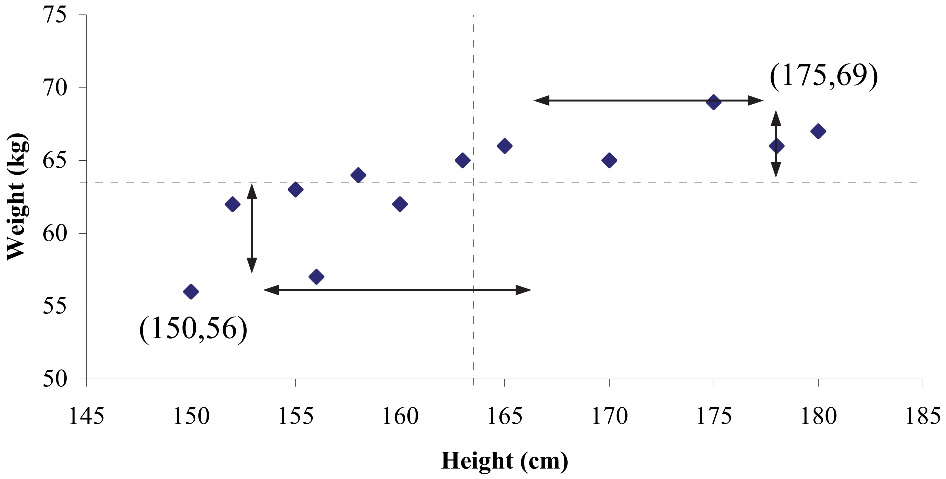
This gives:

$$=\frac{1}{n-1}\left(\sum x_iy_i-n\overline{x}\,\overline{y}-n\overline{x}\,\overline{y}+n\overline{x}\,\overline{y}\right)$$

$$=\frac{1}{n-1}\left(\sum x_iy_i-n\overline{x}\,\overline{y}\right)$$

## Appendix B – why does the covariance formula work?

To show why the covariance formula gives a positive result for variables with positive correlation we'll look at the height and weight example once more:



The mean height ( $\bar{x} = 163.5$ cm) and the mean weight ( $\bar{y} = 63.5$ kg) have been drawn on the scatterplot. From this we can see that for positive correlation, most of the points are in the top right and the bottom left quadrant.

To calculate the covariance we will multiply $x_i - \bar{x}$ and $y_i - \bar{y}$ for each point. From the diagram we can see that for points in the top right quadrant we have:

$$(x_i - \bar{x})(y_i - \bar{y}) = +ve \times +ve = +ve$$

*eg* $(175,69) \;\Rightarrow\; (x_i - \bar{x})(y_i - \bar{y}) = (175 - 163.5)(69 - 63.5) = 11.5 \times 5.5 = 63.25$

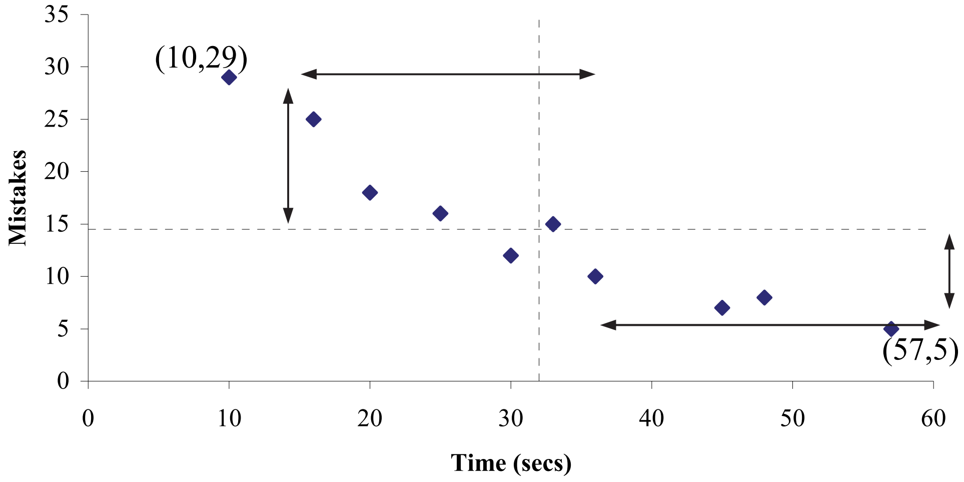
For points in the bottom left quadrant, $x_i - \bar{x}$ and $y_i - \bar{y}$ will both be negative. So:

$$(x_i - \bar{x})(y_i - \bar{y}) = -ve \times -ve = +ve$$

$$(150,56) \Rightarrow (x_i - \bar{x})(y_i - \bar{y}) = (150 - 163.5)(56 - 63.5) = -13.5 \times -7.5 = 101.25$$

So for positive correlation, the majority of points will give positive values of $(x_i - \bar{x})(y_i - \bar{y})$. Therefore, when we total these up we will get a positive covariance.

To show why the covariance formula gives a negative result for variables with negative correlation we'll look at the example of time taken and mistakes made whilst completing a drawing by looking in the mirror.



We can see that for negative correlation, most of the points are in the top left and bottom right quadrants.

From the diagram we can see that for points in the top left quadrant we have:

$$(x_i - \bar{x})(y_i - \bar{y}) = -ve \times +ve = -ve$$

*eg* $(10,29) \;\Rightarrow\; (x_i - \bar{x})(y_i - \bar{y}) = (10-32)(29-14.5) = -22 \times 14.5 = -319$

For points in the bottom right quadrant, we have:

$$(x_i - \bar{x})(y_i - \bar{y}) = +ve \times -ve = -ve$$

*eg* $(57,5) \Rightarrow (x_i - \bar{x})(y_i - \bar{y}) = (57-32)(5-14.5) = 25 \times -9.5 = -237.5$

So for negative correlation, the majority of points will give negative values of $(x_i - \bar{x})(y_i - \bar{y})$. Therefore, when we total these up we will get a negative covariance.

Finally, for variables with no correlation, the points will be scattered in all four quadrants. Therefore there will be a mixture of positive and negative values for $(x_i - \bar{x})(y_i - \bar{y})$. When we add these up they will cancel out to give zero or something near zero.

## *Appendix C – deriving the least squares estimates*

We want to find the values of $\alpha$ and $\beta$ that minimise:

$$\sum(y_i-\alpha-\beta x_i)^2$$

Let $S=\sum(y_i-\alpha-\beta x_i)^2$ . Now differentiating *S* with respect to $\alpha$ and setting the result equal to zero (to get the minimum):

$$\frac{\partial S}{\partial\alpha}=-2\sum(y_i-\alpha-\beta x_i)=0$$

*We have used the 'chain rule' – multiply by the power of the bracket, reduce the power of the bracket by 1 and then multiply by the derivative of the bracket.*

Dividing both sides by $-2$ and then splitting up the summation:

$$\begin{aligned}&\sum(y_i-\alpha-\beta x_i)=0\\ \Rightarrow\quad&\sum y_i-\sum\alpha-\sum\beta x_i=0\end{aligned}$$

Taking out the $\alpha$ and $\beta$ from the summations (as they don't depend on *i*):

$$\begin{aligned}&\sum y_i-\alpha\sum 1-\beta\sum x_i=0\\ \Rightarrow\quad&\sum y_i-n\alpha-\beta\sum x_i=0\end{aligned}$$

Rearranging:

$$\begin{aligned}&n\alpha=\sum y_i-\beta\sum x_i\\ \Rightarrow\quad&\alpha=\frac{\sum y_i}{n}-\beta\frac{\sum x_i}{n}\end{aligned}$$

Hence:

$$\hat{\alpha}=\overline{y}-\beta\overline{x}$$

Next we will find the value of $\beta$ that minimises:

$$S=\sum(y_i-\alpha-\beta x_i)^2$$

Differentiating by $\beta$ and setting the result equal to zero (to get the minimum):

$$\frac{\partial S}{\partial\beta}=-2\sum x_i(y_i-\alpha-\beta x_i)=0$$

Dividing both sides by $-2$ and then splitting up the summation:

$$\begin{aligned}&\sum x_i(y_i-\alpha-\beta x_i)=0\\ \Rightarrow\quad&\sum(x_iy_i-\alpha x_i-\beta x_i^2)=0\\ \Rightarrow\quad&\sum x_iy_i-\sum\alpha x_i-\sum\beta x_i^2=0\end{aligned}$$

Taking out the $\alpha$ and $\beta$ from the summations (as they don't depend on *i*):

$$\sum x_iy_i-\alpha\sum x_i-\beta\sum x_i^2=0$$

Now, earlier we found that $\alpha=\bar{y}-\beta\bar{x}$. Substituting this in, we get:

$$\begin{aligned}&\sum x_iy_i-(\bar{y}-\beta\bar{x})\sum x_i-\beta\sum x_i^2=0\\ \Rightarrow\quad&\sum x_iy_i-\bar{y}\sum x_i+\beta\bar{x}\sum x_i-\beta\sum x_i^2=0\end{aligned}$$

Rearranging:

$$\begin{aligned}&\sum x_iy_i-\bar{y}\sum x_i=\beta\left(\sum x_i^2-\bar{x}\sum x_i\right)\\ &\quad\Rightarrow\quad\beta=\frac{\sum x_iy_i-\bar{y}\sum x_i}{\sum x_i^2-\bar{x}\sum x_i}\end{aligned}$$

Now using the fact that $\sum x_i=n\bar{x}$ (since $\bar{x}=\dfrac{\sum x_i}{n}$), we get:

$$\hat{\beta}=\frac{\sum x_iy_i-n\bar{x}\bar{y}}{\sum x_i^2-n\bar{x}^2}=\frac{s_{xy}}{s_{xx}}$$

## *Extra practice questions*

***P1.1*** In a study into vehicle emissions eight vehicles were thoroughly tested and the following data on hydrocarbon emissions (grams/metre) and carbon monoxide emissions (grams/metre) were obtained:

| **Hydrocarbons (*x*)** | 0.83 | 0.72 | 0.65 | 0.57 | 0.55 | 0.51 | 0.43 | 0.37 |
|---|---|---|---|---|---|---|---|---|
| **Carbon Monoxide (*y*)** | 15.1 | 16.6 | 14.7 | 8.0 | 10.3 | 5.1 | 5.5 | 4.1 |

$$\sum x = 4.63, \quad \sum x^2 = 2.8391, \quad \sum y = 79.4, \quad \sum y^2 = 962.82, \quad \sum xy = 50.748$$

(i) Draw a scatterplot of these data and comment on the relationship between hydrocarbon and carbon monoxide emissions. [3]

(ii) Calculate the correlation coefficient *r*. [3]

(iii) Calculate the fitted regression line using carbon monoxide as the response variable. [2]

[Total 8]

***P1.2*** A random sample of 200 pairs of observations $(x, y)$ from a discrete bivariate distribution $(X, Y)$ is as follows:

the observation $(-2, 2)$ occurs 50 times

the observation $(0, 0)$ occurs 90 times

the observation $(2, -1)$ occurs 60 times.

Calculate the sample correlation coefficient for these data. [4]

***P1.3*** One of the conclusions of a 1980 study appearing in the journal *Advances in Cancer Research* was that '… none of the risk factors for cancer is probably more significant than diet and nutrition'.

The following data are from an investigation into the relationship between fat consumption *x* and prostrate cancer deaths *y*:

| **Country** | **Dietary fat**<br>***x (g/day)*** | **Death rate**<br>***y (per 100,000)*** |
|---|---|---|
| Philippines | 29 | 1.3 |
| Mexico | 57 | 4.5 |
| Colombia | 47 | 5.4 |
| Yugoslavia | 72 | 5.6 |
| Panama | 58 | 7.8 |
| Romania | 67 | 8.8 |
| Czechoslovakia | 96 | 9.1 |
| Spain | 97 | 10.1 |
| Finland | 112 | 11.7 |
| United Kingdom | 143 | 12.4 |
| Canada | 142 | 13.4 |
| France | 137 | 14.4 |
| Australia | 129 | 15.1 |
| United States | 147 | 16.3 |
| Sweden | 132 | 18.4 |

$$\sum x = 1465 \quad \sum x^2 = 165{,}561 \quad \sum y = 154.3 \quad \sum y^2 = 1{,}915.39 \quad \sum xy = 17{,}578.5$$

(i) Draw a scatterplot of these data and comment briefly on your findings. [3]

(ii) (a) Calculate the fitted regression line using death rate as the response variable and dietary fat as the explanatory variable.

(b) Comment briefly on the interpretation of the fitted slope coefficient. [4]

[Total 7]

***P1.4*** At the end of the skiing season the tourist board in a mountain region examines the records of ten ski resorts. For each one it obtains the total number ($y$, thousands) of visitor-days during the season as a measure of the resort's popularity, and the ski-lift capacity ($x$, thousands), being the maximum number of skiers that can be transported per hour. The resulting data are given in the following table:

| **Resort** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Lift Capacity *x*:** | 1.9 | 3.3 | 1.2 | 4.2 | 1.5 | 2.2 | 1.0 | 5.6 | 1.9 | 3.8 |
| **Visitor-days *y*:** | 15.1 | 22.6 | 9.2 | 37.5 | 8.9 | 21.2 | 5.8 | 41.0 | 9.2 | 32.4 |

$$\sum x = 26.6\,,\ \sum x^2 = 91.08\,,\ \sum y = 202.8\,,\ \sum y^2 = 5{,}603.12\,,\ \sum xy = 707.58$$

(i) Draw a scatterplot of $y$ against $x$ and comment briefly on any relationship between a resort's popularity and its ski-lift capacity. [2]

(ii) Calculate the correlation coefficient between $x$ and $y$ and comment briefly in the light of your comment in part (i). [3]

(iii) Calculate the fitted linear regression equation of $y$ on $x$. [2]

[Total 7]

***This page has been left blank so that you can keep the chapter summaries together for revision purposes.***

## *Summary*

### *Scatterplot*

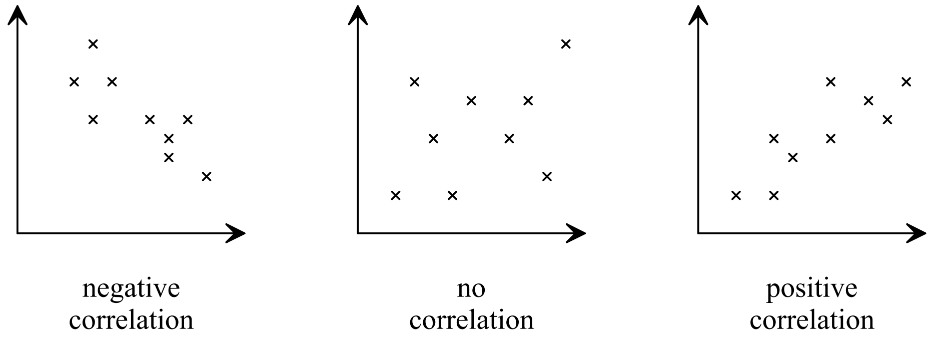
**Bivariate data** are data that have two variables (*eg* height and weight).

A **scatterplot** (or scatter diagram) is a plot of our bivariate data, with one variable (*eg* height) plotted on the *x*-axis (called the **explanatory variable**) and the other variable (*eg* weight) plotted on the *y*-axis (called the **response variable**).

A scatterplot is used to see if there is any relationship or connection (called **correlation**) between the two variables.

### *Correlation*

There are three types of linear correlation:



The clearer the pattern, the stronger the correlation. If there is an *exact* linear relationship we say that there is **perfect** linear correlation.

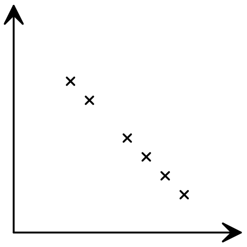
The sample **correlation coefficient**, *r*, measures the type and strength of the connection between two variables:

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

where:

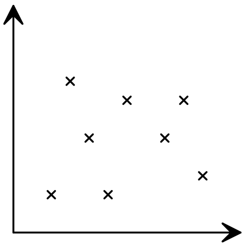
$$s_{xx} = \sum x_i^2 - n\bar{x}^2 \qquad s_{yy} = \sum y_i^2 - n\bar{y}^2 \qquad s_{xy} = \sum x_i y_i - n\overline{xy}$$

The value of the correlation coefficient ranges from $-1$ to 1:



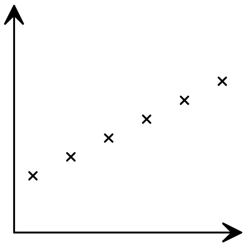
perfect negative correlation

$r = -1$



no correlation

$r = 0$



perfect positive correlation

$r = +1$

Correlation between two variables does not necessarily imply causation.

***Regression***

If there is a (strong) linear relationship between the two variables we can represent that linear relationship with a **regression line**:

$$y = \alpha + \beta x$$

where $\beta$ is the gradient (*ie* how many units 'up' the line goes for every one unit 'across') and $\alpha$ is the *y*-intercept (*ie* where the line crosses the *y*-axis).

The **least squares** estimates of $\alpha$ and $\beta$ are given by:

$$\hat{\alpha} = \overline{y} - \beta \overline{x}$$

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}}$$

They are calculated by minimising the sum of the squares of the **residuals** or errors (that is how much each *y* value is 'out' from the regression line).

## Solutions

### Solution 1.1

We have a *decreasing* pattern in the points – generally as the time taken increases, the number of mistakes made decreases.

However, it's not entirely clear whether this relationship is linear or not.

### Solution 1.2

(i) Negative correlation – as the number of cigarettes smoked increases, life expectancy would decrease.

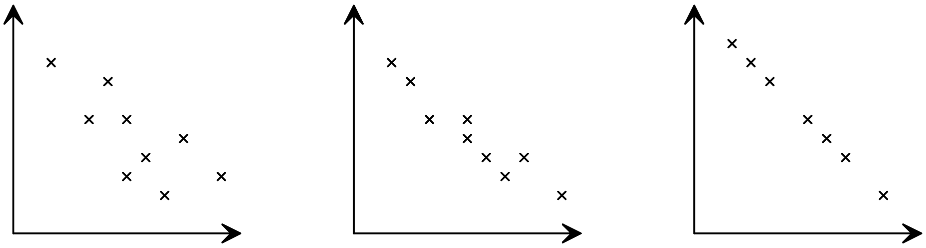
(ii) Positive correlation – the further one lives from work, the greater the time taken to get there.

(iii) Positive correlation. More bedrooms means the house is bigger and so the insurance charged on it will increase.

(iv) Negative correlation – as the amount of 'no claims' discount increases the cost of car insurance decreases.

(v) No correlation – there should be no connection between the number of exams passed in a sitting and the length of hair.

### Solution 1.3



weak negative correlation

strong negative correlation

perfect negative correlation

***Solution 1.4***

(i) There is an *exact* relationship between pounds exchanged and dollars bought (assuming that there are no fluctuations in the exchange rate over the day). This is Scatterplot C.

(ii) There will be a very weak relationship between height and shoe size (as there are many examples of small people with big feet and vice versa). This is Scatterplot A.

(iii) We would expect there to be a fairly strong relationship between engine size and the cost of car insurance. This is Scatterplot B.

***Solution 1.5***

(i) The formula for the sample covariance is:

$$\frac{1}{n-1}\left\{\sum x_i y_i - n\overline{xy}\right\}$$

Calculating the means:

$$\overline{x} = \frac{\sum x}{n} = \frac{16+25+38+61+79}{5} = \frac{219}{5} = 43.8$$

$$\overline{y} = \frac{\sum y}{n} = \frac{7.8+5.2+4.3+2.5+1.7}{5} = \frac{21.5}{5} = 4.3$$

Calculating $\sum x_i y_i$ :

$$(16\times 7.8)+(25\times 5.2)+(38\times 4.3)+(61\times 2.5)+(79\times 1.7) = 705$$

This gives us a covariance of:

$$\frac{1}{4}\left\{705-5\times 43.8\times 4.3\right\} = \frac{-236.7}{4} = -59.175$$

(ii) We have a negative value for the covariance which means that there is negative correlation between the variables. As the mileage of the car increases, the price will decrease.

***Solution 1.6***

The covariance will be zero or close to zero.

***Solution 1.7***

$$r=\frac{\text{covariance of } X \text{ and } Y}{sd(X)\times sd(Y)}=\frac{12.5}{\sqrt{2.5}\sqrt{62.5}}=1$$

There is perfect positive correlation between the pay received and the number of hours worked. Therefore the correlation coefficient is also 1.

***Solution 1.8***

Scatterplot A shows strong positive correlation, $r \simeq 0.8$ .

Scatterplot B shows very strong negative correlation, $r \simeq -0.95$ .

Scatterplot C shows weak negative correlation, $r \simeq -0.65$ .

*The value for Scatterplot C is probably more negative than you might think. The pattern is fairly clear and there are no outliers (ie points that are way off the general pattern).*

***Solution 1.9***

(i) First calculating the means:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{450}{6} = 75 \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{374}{6} = 62\tfrac{1}{3}$$

Next we calculate the sum of squares:

$$s_{xx} = \sum x_i^2 - n\bar{x}^2 = 35{,}428 - 6 \times 75^2 = 1{,}678$$

$$s_{yy} = \sum y_i^2 - n\bar{y}^2 = 23{,}850 - 6 \times \left(62\tfrac{1}{3}\right)^2 = 537\tfrac{1}{3}$$

$$s_{xy} = \sum x_i y_i - n\overline{xy} = 28{,}745 - 6 \times 75 \times 62\tfrac{1}{3} = 695$$

This gives:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{695}{\sqrt{1{,}678 \times 537\tfrac{1}{3}}} = 0.732$$

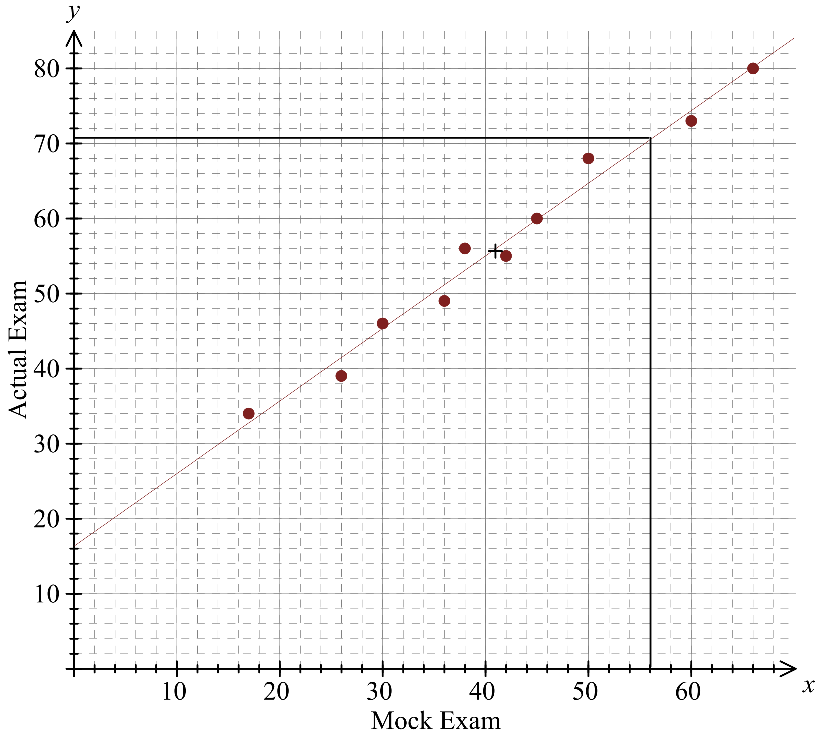
(ii) There is fairly strong positive correlation, *ie* as literacy rates increase so does life expectancy.

(iii) No. It is to do with how economically developed the country is. A less economically developed country has low literacy and low life expectancy whereas a more economically developed country will have higher literacy and higher life expectancy.

### *Solution 1.10*

(i) The mean of the co-ordinates $(\bar{x}, \bar{y})$ is $(41, 56)$.

(ii) A good line of 'best fit' is:



(iii) Reading off the graph, we get a value of about 71.

### *Solution 1.11*

The *y*-intercept is about 16.

Depending on the points you consider, the gradient will be between 0.93 to 1.

This will give a line of 'best fit' of, say, $y = 16 + 0.95x$.

*The true equation of the line is* $y = 16.3 + 0.968x$ *– our accuracy is limited by the plot of our graph and our ability to read from it.*

***Solution 1.12***

(i) First calculating the means:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{166.9}{5} = 33.38 \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{437}{5} = 87.4$$

Next we calculate the sum of squares:

$$s_{xx} = \sum x_i^2 - n\bar{x}^2 = 8{,}588.31 - 5 \times 33.38^2 = 3{,}017.188$$

$$s_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = 8{,}328 - 5 \times 33.38 \times 87.4 = -6{,}259.06$$

So we get:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{-6{,}259.06}{3{,}017.188} = -2.074$$

and:

$$\hat{\alpha} = \bar{y} - \beta\bar{x} = 87.4 - (-2.074) \times 33.38 = 156.6$$

(ii) The fitted regression line is:

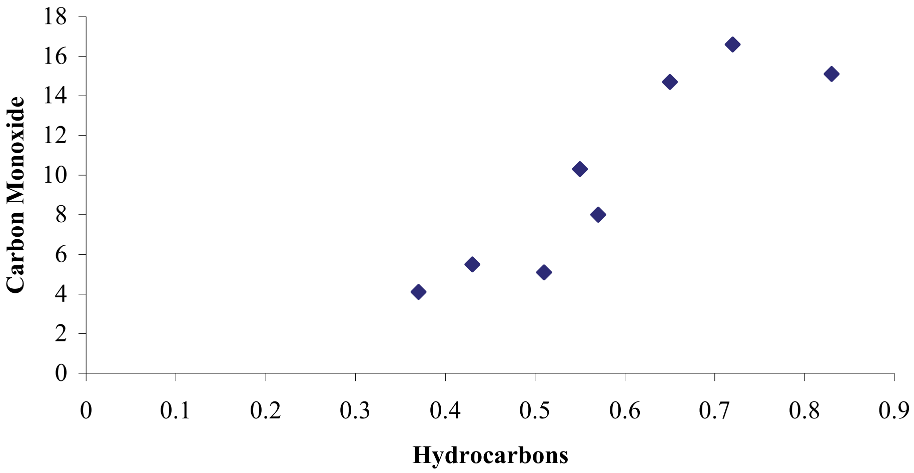
$$\hat{y} = 156.6 - 2.074x$$

(iii) Substituting $x = 60$ into our line of regression, we get:

$$\hat{y} = 156.6 - 2.074 \times 60 = 32.2$$

## *Solutions to extra practice questions*

***P1.1*** (i) The scatterplot for these data is shown below:



The scatterplot shows positive linear correlation, *ie* cars emitting more hydrocarbons also emit more carbon monoxide.

(ii) The sample correlation coefficient is given by:

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

So we require $s_{xx}, s_{yy}$ and $s_{xy}$ which in turn require $\bar{x}$ and $\bar{y}$.

$$\bar{x}=\frac{\sum x}{n}=\frac{4.63}{8}=0.57875 \quad \text{and} \quad \bar{y}=\frac{\sum y}{n}=\frac{79.4}{8}=9.925$$

$$s_{xx}=\sum x^2-n\bar{x}^2=2.8391-8\times 0.57875^2=0.1594875$$

$$s_{yy}=\sum y^2-n\bar{y}^2=962.82-8\times 9.925^2=174.775$$

$$s_{xy}=\sum xy-n\bar{x}\bar{y}=50.748-8\times 0.57875\times 9.925=4.79525$$

Hence:

$$r=\frac{4.79525}{\sqrt{0.1594875\times 174.775}}=0.908 \quad \text{(3 SF)}$$

(iii) We have:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{4.79525}{0.1594875} = 30.07$$

and:

$$\hat{\alpha} = \overline{y} - \beta\overline{x} = 9.925 - 30.07 \times 0.57875 = -7.48$$

Hence, the fitted regression line is:

$$\hat{y} = -7.48 + 30.07x$$

***P1.2*** The sample correlation coefficient is given by:

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

So we require $s_{xx}, s_{yy}$ and $s_{xy}$ which in turn require $\overline{x}$ and $\overline{y}$.

$$\overline{x} = \frac{\sum fx}{\sum f} = \frac{(50 \times -2) + (90 \times 0) + (60 \times 2)}{50 + 90 + 60} = \frac{20}{200} = 0.1$$

$$\overline{y} = \frac{\sum fy}{\sum f} = \frac{(50 \times 2) + (90 \times 0) + (60 \times -1)}{50 + 90 + 60} = \frac{40}{200} = 0.2$$

Now since there are only 3 different values for *x* and *y* it's probably slightly easier to use $s_{xx} = \sum (x - \overline{x})^2$, $s_{yy} = \sum (y - \overline{y})^2$ and $s_{xy} = \sum (x - \overline{x})(y - \overline{y})$. This gives:

$$s_{xx} = 50 \times (-2 - 0.1)^2 + 90 \times (0 - 0.1)^2 + 60 \times (2 - 0.1)^2 = 438$$

$$s_{yy} = 50 \times (2 - 0.2)^2 + 90 \times (0 - 0.2)^2 + 60 \times (-1 - 0.2)^2 = 252$$

$$\begin{aligned} s_{xy} &= 50 \times (-2 - 0.1)(2 - 0.2) + 90 \times (0 - 0.1)(0 - 0.2) + 60 \times (2 - 0.1)(-1 - 0.2) \\ &= -324 \end{aligned}$$

So the sample correlation coefficient is:

$$r = \frac{-324}{\sqrt{438 \times 252}} = -0.975$$

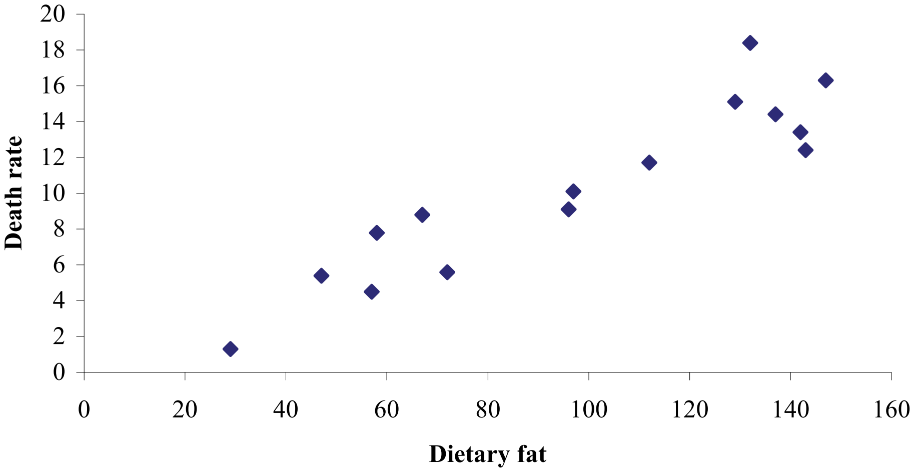
*Alternatively, using* $s_{xx} = \sum x^2 - n\bar{x}^2$*, etc gives:*

$$s_{xx} = \left[ 50 \times (-2)^2 + 90 \times 0^2 + 60 \times 2^2 \right] - 200 \times 0.1^2 = 438$$

$$s_{yy} = \left[ 50 \times 2^2 + 90 \times 0^2 + 60 \times (-1)^2 \right] - 200 \times 0.2^2 = 252$$

$$s_{xy} = \left[ 50 \times (-2 \times 2) + 90 \times (0 \times 0) + 60 \times (2 \times -1) \right] - 200 \times 0.1 \times 0.2 = -324$$

***P1.3*** (i) The scatterplot for these data is:



The scatterplot shows positive linear correlation, *ie* countries where the population eat more dietary fat have a greater death rate.

(ii) (a) The fitted regression line is $\hat{y}=\hat{\alpha}+\hat{\beta}x$ where:

$$\hat{\alpha}=\overline{y}-\beta\overline{x}$$

$$\hat{\beta}=\frac{s_{xy}}{s_{xx}}$$

So we require $\overline{x},\overline{y},s_{xx}$ and $s_{xy}$ :

$$\overline{x}=\frac{\sum x}{n}=\frac{1,465}{15}=97.\dot{6}$$

$$\overline{y}=\frac{\sum y}{n}=\frac{154.3}{15}=10.28\dot{6}$$

$$s_{xx}=\sum x^2-n\overline{x}^2=165,561-15\times 97.\dot{6}^2=22,479.\dot{3}$$

$$s_{xy}=\sum xy-n\overline{x}\,\overline{y}=17,578.5-15\times 97.\dot{6}\times 10.28\dot{6}=2,508.5\dot{3}$$

Hence:

$$\hat{\beta}=\frac{2,508.5\dot{3}}{22,478.\dot{3}}=0.11160$$

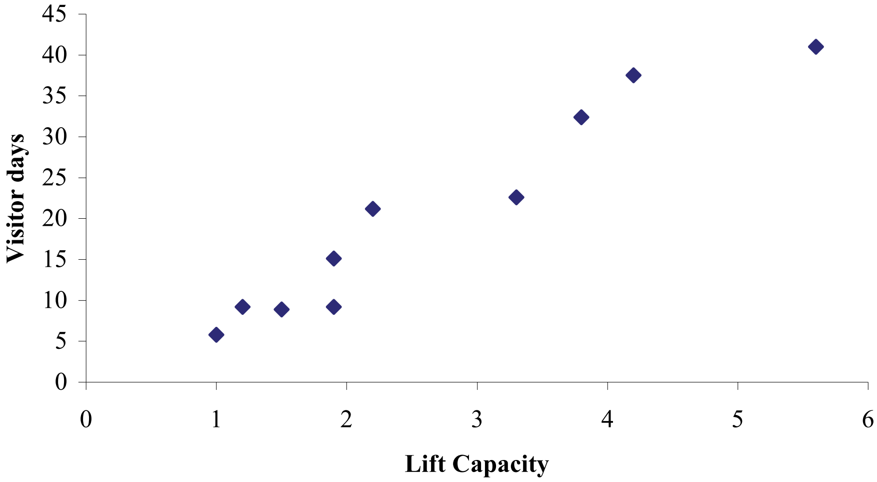
$$\hat{\alpha}=10.28\dot{6}-0.11160\times 97.\dot{6}=-0.61272$$

Therefore, our fitted regression line is:

$$\hat{y}=0.11160x-0.61272$$

(b) The slope parameter is the gradient of the regression line, so an increase in 1 g/day of fat increases the death rate by 0.11160 per 100,000.

***P1.4*** (i) The scatterplot for these data is:



The scatterplot shows positive linear correlation, *ie* resorts with a greater lift capacity are more popular.

(ii) The formula for the correlation coefficient is:

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

So we require $s_{xx}, s_{yy}$ and $s_{xy}$ which in turn require $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{\sum x}{n} = \frac{26.6}{10} = 2.66 \quad \text{and} \quad \bar{y} = \frac{\sum y}{n} = \frac{202.8}{10} = 20.28$$

$$s_{xx} = \sum x^2 - n\bar{x}^2 = 91.08 - 10 \times 2.66^2 = 20.324$$

$$s_{yy} = \sum y^2 - n\bar{y}^2 = 5{,}603.12 - 10 \times 20.28^2 = 1{,}490.336$$

$$s_{xy} = \sum xy - n\bar{x}\bar{y} = 707.58 - 10 \times 2.66 \times 20.28 = 168.132$$

Hence:

$$r = \frac{168.132}{\sqrt{20.324 \times 1490.336}} = 0.966$$

This indicates strong positive correlation, which backs up our theory of a linear relationship from (i).

(iii) The fitted regression line of *y* on *x* is $y = \alpha + \beta x$ where:

$$\hat{\alpha} = \overline{y} - \beta \overline{x}$$

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}}$$

Hence:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{168.132}{20.324} = 8.27$$

$$\hat{\alpha} = \overline{y} - \beta \overline{x} = 20.28 - 8.27 \times 2.66 = -1.73$$

So the equation of the regression line of *y* on *x* is:

$$\hat{y} = 8.27x - 1.73$$