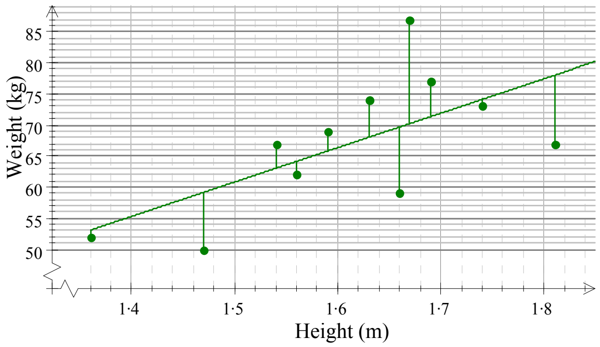
Regression: Finding the equation of the line of best fit

Objectives: To find the equation of the least squares regression line of y on x.

Background and general principle

The aim of regression is to calculate the equation of the line of best fit on a scatter graph.

Consider the scatter graph on the right. One possible line of best fit has been drawn on the diagram. Some of the points lie above the line and some lie below it.



The vertical distance each point is above or below the line has been added to the diagram. These distances are called *deviations* or *errors* – they are symbolised as $d_1, d_2, \ldots, d_n$.

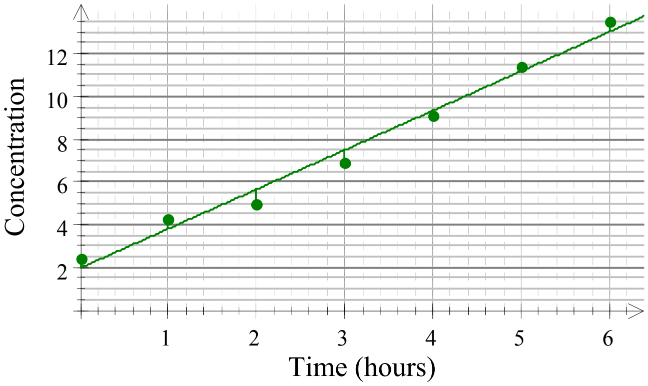
When drawing in a regression line, the aim is to make the line fit the points as closely as possible. We do this by making the **total of the squares of the deviations as small as possible**, i.e. we minimise $\sum d_i^2$.

If a line of best fit is found using this principle, it is called the **least-squares regression line**.

Example 1:
A patient is given a drip feed containing a particular chemical and its concentration in his blood is measured, in suitable units, at one hour intervals. The doctors believe that a linear relationship will exist between the variables.

Time, x (hours)	0	1	2	3	4	5	6
Concentration, y	2.4	4.3	5.0	6.9	9.1	11.4	13.5

We can plot these data on a scatter graph – time would be plotted on the horizontal axis (as it is the independent variable). Time is here referred to as a **controlled variable**, since the experimenter fixed the value of this variable in advance (measurements were taken every hour).



Concentration is the dependent variable as the concentration in the blood is likely to vary according to time.

The doctor may wish to estimate the concentration of the chemical in the blood after 3.5 hours.

She could do this by finding the equation of the line of best fit.

There is a formula which gives the equation of the line of best fit.

The equation of the line is $y = a + bx$

where $b = \frac{S_{xy}}{S_{xx}}$ and $a = \overline{y} - b\overline{x}$.

Note: $S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$ and $S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$.

Note 2: $\overline{x}$ and $\overline{y}$ are the mean values of *x* and *y* respectively.

This line is called the (least-squares) regression line of *y* on *x* (because the equation has been given with *y* the subject).
b is sometimes called the regression coefficient of *y* on *x*.

We can work out the equation for our example as follows:

$$\sum x = 0 + 1 + \ldots + 6 = 21 \quad \text{so} \quad \overline{x} = \frac{21}{7} = 3$$

$$\sum y = 2.4 + 4.3 + \ldots + 13.5 = 52.6 \quad \text{so} \quad \overline{y} = \frac{52.6}{7} = 7.514\ldots$$

$$\sum xy = (0 \times 2.4) + (1 \times 4.3) + \ldots + (6 \times 13.5) = 209.4$$

$$\sum x^2 = 0^2 + 1^2 + \ldots + 6^2 = 91 \quad \text{so} \quad \overline{x} = \frac{21}{7} = 3$$

These could all be found on a calculator (if you enter the data into a calculator).

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 209.4 - \frac{21 \times 52.6}{7} = 51.6$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 91 - \frac{(21)^2}{7} = 28$$

So, $b = \frac{S_{xy}}{S_{xx}} = \frac{51.6}{28} = 1.843$ and $a = \overline{y} - b\overline{x} = 7.514 - 1.843 \times 3 = 1.985$.

So the equation of the regression line is $y = 1.985 + 1.843x$.

To work out the concentration after 3.5 hours: y = 1.985 + 1.843 × 3.5 = 8.44 (3sf)

If you want to find how long it would be before the concentration reaches 8 units, we substitute y = 8 into the regression equation:

$$8 = 1.985 + 1.843x$$

Solving this we get: $x = 3.26$ hours

Note: It would not be sensible to predict the concentration after 8 hours from this equation – we don't know whether the relationship will continue to be linear. The process of trying to predict a value from outside the range of your data is called *extrapolation*.

Example 2:
The heights and weights of a sample of 11 students are:

Height (m) h	1.36	1.47	1.54	1.56	1.59	1.63	1.66	1.67	1.69	1.74	1.81
Weight (kg) w	52	50	67	62	69	74	59	87	77	73	67

$$[n = 11 \quad \sum h = 17.72 \quad \sum h^2 = 28.705 \quad \sum w = 737 \quad \sum w^2 = 50571 \quad \sum hw = 1196.1]$$

a) Calculate the regression line of *w* on *h*.
b) Use the regression line to estimate the weight of someone whose height is 1.6m.

Note: Both height and weight are referred to as **random** variables – their values could not have been predicted before the data were collected. If the sampling were repeated again, different values would be obtained for the heights and weights.

Solution:
a) We begin by finding the mean of each variable:

$$\bar{h} = \frac{\sum h}{n} = \frac{17.72}{11} = 1.6109...$$

$$\bar{w} = \frac{\sum w}{n} = \frac{737}{11} = 67$$

Next we find the sums of squares:

$$S_{hh} = \sum h^2 - \frac{\left(\sum h\right)^2}{n} = 28.705 - \frac{17.72^2}{11} = 0.1597$$

$$S_{ww} = \sum w^2 - \frac{\left(\sum w\right)^2}{n} = 50571 - \frac{737^2}{11} = 1192$$

$$S_{hw} = \sum hw - \frac{\sum h \sum w}{n} = 1196.1 - \frac{17.72 \times 737}{11} = 8.86$$

The equation of the regression line is:

$$w = a + bh$$

where

$$b = \frac{S_{hw}}{S_{hh}} = \frac{8.86}{0.1597} = 55.5$$

and

$$a = \bar{w} - b\bar{h} = 67 - 55.5 \times 1.6109 = -22.4$$

So the equation of the regression line of *w* on *h* is:

$$w = -22.4 + 55.5h$$

b) To find the weight for someone that is 1.6m high:

$$w = -22.4 + 55.5 \times 1.6 = 66.4 \text{ kg}$$

Two possible regression lines

When x and y are both **random variables**, there are two possible regression lines that can be calculated:

* the regression line of y on x;

* the regression line of x on y.

The regression line of y on x is the line that has already been met. Its equation is

$$y = a + bx$$

and it is used to find a value of y when we are given a value of x. This line minimises the vertical distances of each point from the line.

The regression line of x on y minimises the horizontal distances of each point from the line. It is used if you wish to work out a value of x when you are given a value of y. The equation of this regression line is

$$x = a' + b'y$$

where

$$b' = \frac{S_{xy}}{S_{yy}}$$

and

$$a' = \bar{x} - b'\bar{y}.$$

The regression lines will not in general be the same (unless the points lie on a perfect straight line). Both regression lines pass through the mean point $(\bar{x}, \bar{y})$.

Note: If x is a controlled variable, you always use the regression line of y on x (since the regression line of x on y doesn't have any statistical meaning in this case).

Example:

A psychologist wants to investigate the relationship between the IQ of a child and the IQ of their mother. She measures the IQ of a sample of 8 children and mothers:

Child's IQ, x	87	91	94	98	103	108	111	123
Mother's IQ, y	94	96	89	102	98	94	116	117

[$\sum x = 815 \quad \sum x^2 = 84013 \quad \sum y = 806 \quad \sum y^2 = 81962 \quad \sum xy = 82789$]

a) Work out the product moment correlation coefficient between x and y.

b) Calculate the regression line of y on x and the regression line of x on y. Write down a point that both lines pass through.

c) Use the appropriate regression line to estimate the IQ of a child born to a mother with an IQ of 100. Using your answer to part (a), explain how accurate you think this estimate is likely to be.

Solution:

a) Using the sums given in the question we find that:

$$S_{xy} = 82789 - \frac{815 \times 806}{8} = 677.75 \qquad S_{xx} = 84013 - \frac{815^2}{8} = 984.875$$

$$S_{yy} = 81962 - \frac{806^2}{8} = 757.5$$

So,

$$r = \frac{677.75}{\sqrt{984.875 \times 757.5}} = 0.785 \quad \text{(to 3 s.f.)}$$

b) For the regression line of *y* on *x*:

$$b = \frac{677.75}{984.875} = 0.688 \quad \text{(to 3 s.f.)}$$

$$\bar{x} = \frac{815}{8} = 101.875 \text{ and } \bar{y} = \frac{806}{8} = 100.75$$

$$a = \bar{y} - b\bar{x} = 100.75 - 0.688 \times 101.875 = 30.66$$

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

$y = 30.66 + 0.688x$.

For the regression line of *x* on *y*:

$$b' = \frac{677.75}{757.5} = 0.895 \quad \text{(to 3 s.f.)}$$

$$a' = \bar{x} - b'\bar{y} = 101.875 - 0.895 \times 100.75 = 11.70$$

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

$x = 11.70 + 0.895y$.

Both lines must pass through the mean point (101.875, 100.75).

c) Both variables are random variables.
To find *x* when $y = 100$, we use the regression line of *x* on *y*.

$x = 11.70 + 0.895 \times 100 = 101.2$

This accurate should be reasonably accurate since the product moment correlation coefficient shows fairly strong correlation.

Example 2:
The scores that 9 students obtained in their C1 and M1 mathematics examinations are as follows:

C1 mark, c%	82	51	68	45	30	55	64	77	28
M1 mark, m%	75	46	84	47	42	59	52	69	41

$$[\sum c = 500, \quad \sum c^2 = 30708, \quad \sum m = 515, \quad \sum m^2 = 31397, \quad \sum cm = 30617$$

a) Calculate the equation of the appropriate regression line in order to estimate the mark that a student scoring 53% in M1 might expect to obtain in C1.
b) Explain why it would not be sensible to use the same regression line to estimate the C1 mark that might be expected if a student scored 20% in M1.

Solution:
a) Both *c* and *m* are random variables. So we need to find the regression line of *c* on *m*:

$$S_{cm} = 30617 - \frac{500 \times 515}{9} = 2005.89 \qquad S_{mm} = 31397 - \frac{515^2}{9} = 1927.56$$

So, $b = \dfrac{2005.89}{1927.56} = 1.04$

$$\bar{c} = \frac{500}{9} = 55.56 \quad \text{and} \quad \bar{m} = \frac{515}{9} = 57.22$$

Therefore, $a = 55.56 - 1.04 \times 57.22 = -3.95$

So, the regression line of *c* on *m* is:

$c = -3.95 + 1.04m$

When $m = 53$, c = -3.95 + 1.04×53 = 51.17%

b) The value $m = 20$ lies outside the range of M1 marks seen in the table. The regression line calculated in part (a) cannot be assumed to still be valid outside the range of values given in the table.

Example 3:

A particular greenhouse plant is suspect able to a particular disease. An agricultural scientist wishes to see how the temperature of the greenhouse affects the prevalence of the disease. She designs an experiment in which she monitors the percentage of diseased leaves occurring at different temperatures:

Temperature, t °F	70	72	74	76	78	80
Percentage of diseased leaves, p	12.3	9.5	7.7	6.1	4.3	2.3

$$[\sum t = 450, \ \sum t^2 = 33820, \ \sum p = 42.2, \ \sum p^2 = 361.82, \ \sum tp = 3097.8]$$

The scientist wishes to estimate the temperature that she should set the greenhouse if she is aims for 5% of leaves being diseased. Calculate an appropriate regression line and use it to find the required temperature (giving your answer to the nearest whole number). Give a reason for your choice of regression line.

Solution:

In this situation, temperature is a controlled variable. Only the regression line of *p* on *t* makes sense here.

$$S_{tp} = 3097.8 - \frac{450 \times 42.2}{6} = -67.2 \qquad S_{pp} = 361.82 - \frac{42.2^2}{6} = 65.013$$

So, $b = \frac{-67.2}{65.013} = -1.0336$

$$\bar{t} = \frac{450}{6} = 75 \quad \text{and} \quad \bar{p} = \frac{42.2}{6} = 7.033$$

Therefore, $a = 7.033 - (-1.0336) \times 75 = 84.5$

So, the regression line of *t* on *p* is:

$p = 84.5 - 1.03t$

Therefore to find t, we solve

$5 = 84.5 - 1.03t$

i.e. $t = 77$°C