S1 Revision Notes: Regression

Section 1: Calculating the regression line of *y* on *x*

At GCSE you learnt to draw a line of best fit on a scatter graph.
Regression is the area of statistics that enables you to calculate the equation of the best fitting line.

The line of best fit is defined to be the line that minimises the sum of the squares of the deviations each point is from the line. Hence the name *least squares regression line*.

Recap:

- The regression line of *y* on *x* has equation

$$y = a + bx,$$

where

$$b = \frac{S_{xy}}{S_{xx}}$$

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

and

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

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}.$$

These formulae are in the formula book.

- This regression line passes through the mean point, $(\bar{x}, \bar{y})$.

- Your calculators have the facility to find the equation of the regression line for you – make sure you know how to use your calculator.

Example:
In addition to its full-time staff, a supermarket employs part-time sales staff on Saturdays. The manager experimented to see if there is a relationship between the takings and the number of part-time staff employed. He collected the following data on nine successive Saturdays.

Number of part-time staff employed, x	Takings, £'00, y
10	313
13	320
16	319
19	326
22	333
25	342
28	321
31	361
34	355

$$n = 9,\ \sum x = 198,\ \sum x^2,\ \sum xy = 66713,\ \sum y = 2990,\ \sum y^2 = 995646$$

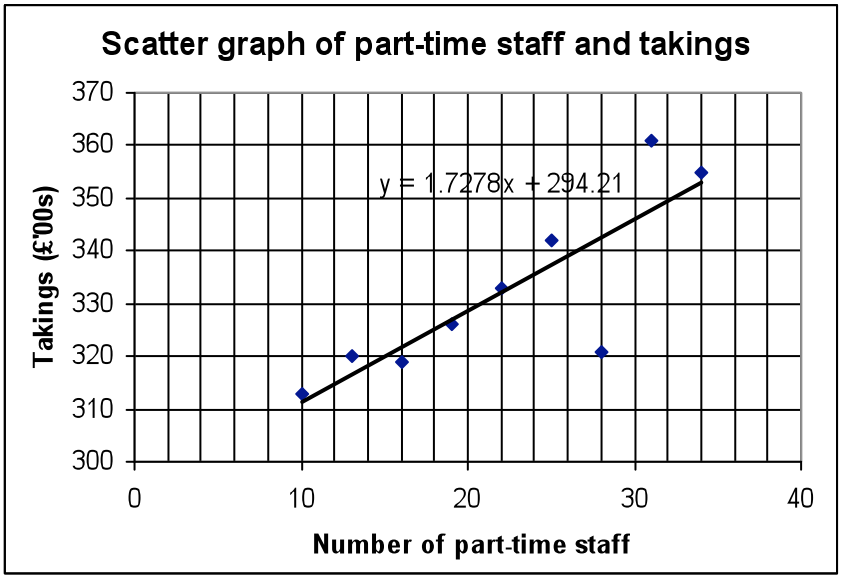
(i) Plot a scatter diagram of these data.

(ii) Calculate the equation of the regression line of takings on the number of part-time staff employed. Draw the line on your scatter diagram.

(iii) If the equation of the regression line is denoted $y = a + bx$, give an interpretation to b.

(iv) On one of the Saturdays, major road works blocked a nearby main road. Which Saturday do you think this was? Give a reason for your choice.

(i)



(ii) To find the equation of the regression line...

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 4896 - \frac{198^2}{9} = 540$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 66713 - \frac{198 \times 2990}{9} = 933.$$

So, $b = \frac{S_{xy}}{S_{xx}} = \frac{933}{540} = 1.7278$

Also, $\bar{x} = \frac{198}{9} = 22$ and $\bar{y} = \frac{2990}{9} = 332.222$.

So, $a = \bar{y} - b\bar{x} = 332.222 - 1.7278 \times 22 = 294.21$

Therefore the equation of the regression line is $y = 294 + 1.73x$

(iii) b is the gradient of the line. This says that the takings are increasing by about 1.73 hundred pounds (i.e. £173) for each additional member of part-time staff employed.

(The 294 represents the y-intercept, i.e. if no part-time staff were employed takings would be 294 hundred pounds, i.e. £29400)

(iv) There is an outlying point on the scatter graph where takings are not as high as you would have expected. This is the Saturday when 28 part-time staff were employed.

Section 2: Choosing the regression line to use

There are two possible regression lines: the regression line of y on x or the regression line of x on y. Sometimes you have to decide which regression line to use.

You use the following process to decide which regression line to find.

(1) Decide whether there are any **controlled variables**. Controlled variables arise in experiments that have been planned, i.e. the values of one of the variables has been fixed in advance. Contolled variables will usually increase in equal steps.

If x is a controlled variable then you should always use the regression line of y on x.

(2) If neither x or y is a controlled variable then you choose the regression line as follows:
If you wish to estimate a value of y given a value for x, find the regression line of y on x.
If you wish to estimate a value of x given a value for y, find the regression line of x on y.

Note: The regression line of x on y has equation $x = a' + b'y$
where

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

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

and

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

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}.$$

Note 2: The regression line of x on y also passes through the mean point.

Note 3: The two regression lines (x on y and y on x) will generally be different from each other. They are only the same if the points on the scatter graph are in a perfect straight line.

- **Example:**

The results of an experiment to determine how the percentage sand content of soil y varies with depth in centimetres below ground level d are given in the following table:

d	0	6	12	18	24	30	36	42	48
y	80.6	63.0	64.3	62.5	57.5	59.2	40.8	46.9	37.6

a) Calculate the appropriate regression line to estimate the depth below ground level which would correspond to soil with a 50% sand content. Explain why you chose to use this regression line.

b) Calculate the product-moment correlation coefficient. Use this to explain how reliable the estimate made in (a) is likely to be.

- **Solution**

a) You need to find the regression line of y on d since d is a controlled variable (we have an experiment and the soil composition has been measured at a series of depths that the experimenter has chosen – note how they increase in equal steps).

The summary values here are:

$$n = 9,\ \sum d = 216,\ \sum d^2 = 7344,\ \sum dy = 10674,\ \sum y = 512.4,\ \sum y^2 = 30595$$

$$S_{dd} = \sum d^2 - \frac{\left(\sum d\right)^2}{n} = 7344 - \frac{216^2}{9} = 2160$$

$$S_{dy} = \sum dy - \frac{\left(\sum d\right)\left(\sum y\right)}{n} = 10674 - \frac{216 \times 512.4}{9} = -1623.6$$

So, $$b = \frac{S_{dy}}{S_{dd}} = \frac{-1623.6}{2160} = -0.751667$$

As $\bar{d} = 24$ and $\bar{y} = 56.9333...$ we also find that $a = 74.973$

So the regression line is $y = 74.973 - 0.751667d$ (this is not the final answer so maintain accuracy by keeping lots of significant figures).

When $y = 50$: $50 = 74.973 - 0.751667d$

So, $0.751667d = 24.973$

$d = 33.2$ cm (3sf)

Therefore the depth is 33.2 cm

(b) To find the product moment correlation coefficient, we also need the value of S_{yy}:

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 30595 - \frac{512.4^2}{9} = 1422.36$$

The formula for PMCC is:

$$r = \frac{S_{dy}}{\sqrt{S_{dd}S_{yy}}}$$

We get: $r = -0.9262909 = -0.926$ (3sf)

This indicates strong negative correlation. As the points must lie close to a straight line, the estimate found in (a) is likely to be reliable.

Note: It would not be sensible to use our regression line to find the value of y corresponding to $d = 60$. This is because $d = 60$ is outside the range of values of data collected in the experiment. We don't know whether the same relationship would continue to hold.

- **Example**

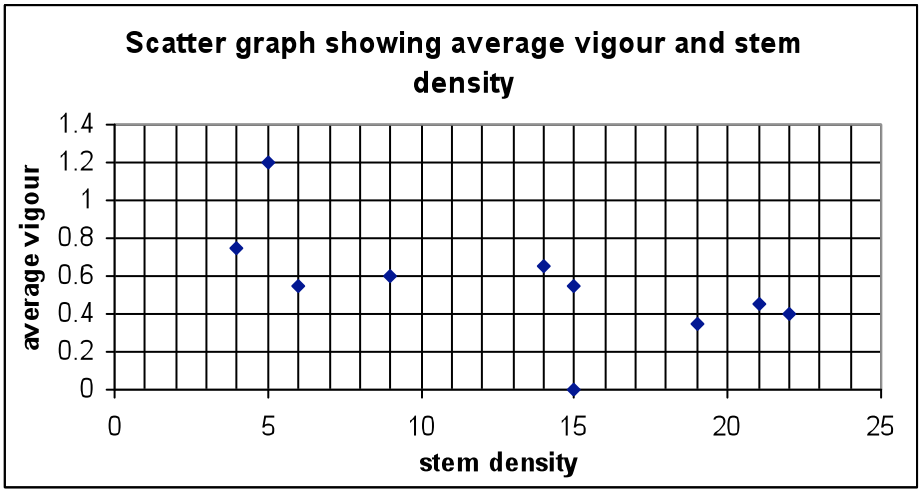
In an investigation of the genus *Tamarix* (a shrub able to withstand drought), research workers in Tunisia measured the average vigour, y, (defined as the average width in centimetres of the last two annual rings) and stem density, x, (defined as the number of stems per m²) at ten sites with the following results:

x	4	5	6	9	14	15	15	19	21	22
y	0.75	1.20	0.55	0.60	0.65	0.55	0	0.35	0.45	0.40

(i) Draw a scatter diagram for the data;
(ii) Calculate the appropriate regression line to estimate the stem density that would correspond to an average vigour value of 0.70.

- **Solution:**

(i)



(ii) Neither x nor y is a controlled variable (both variables are measured values – they weren't chosen by the experimenter).

Therefore since we have $y = 0.70$ and wish to estimate x, we need to find the regression line of x on y.

The summary values are:

$$n = 10,\ \sum x = 130,\ \sum x^2 = 2090,\ \sum xy = 59.95,\ \sum y = 5.5,\ \sum y^2 = 3.875$$

From these we can calculate:

$$S_{yy} = 3.875 - \frac{5.5^2}{10} = 0.85$$
$$S_{xy} = 59.95 - \frac{130 \times 5.5}{10} = -11.55$$

The formula for the regression line of x on y is:

$$x = a + by$$

where:

$$b = \frac{S_{xy}}{S_{yy}} = \frac{-11.55}{0.85} = -13.5882$$

As $\bar{x} = 13,\ \bar{y} = 0.55$, we get $a = \bar{x} - b\bar{y} = 13 - (-13.5882 \times 0.55) = 20.4735$

Therefore the regression line is $x = 20.4735 - 13.5882y$

When $y = 0.70$, $x = 20.4735 - 13.5882 \times 0.70 = 10.96$