Correlation & Regression (4 pages; 12/9/13)

(1) Note that correlation tests cannot establish a causal link between two variables; only that there is an 'association' (or connection).

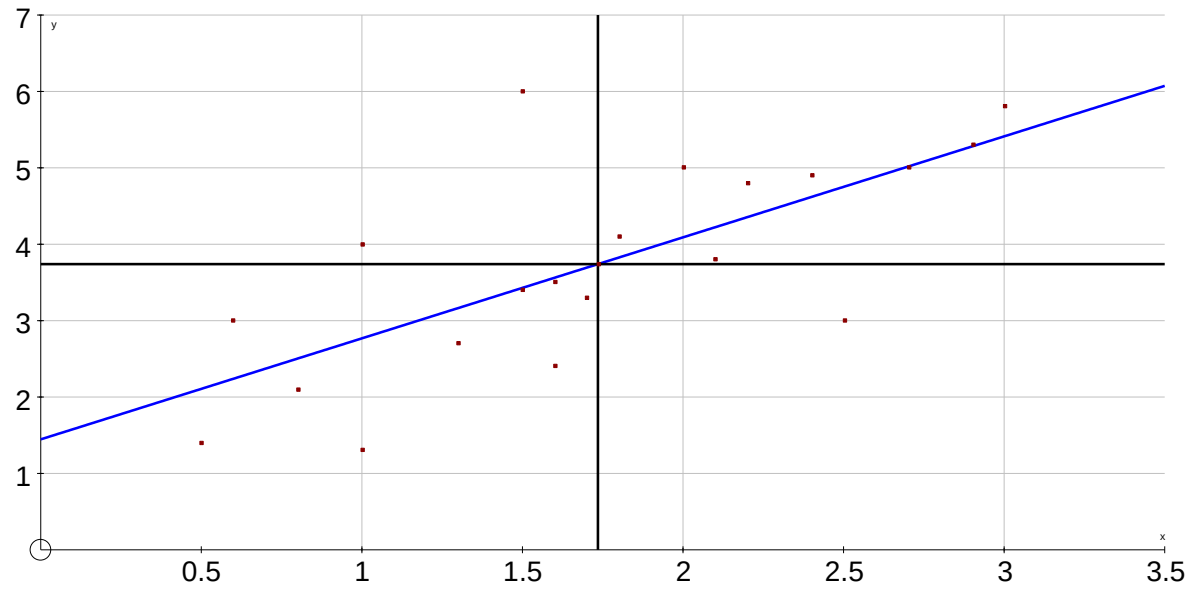
For example, it can probably be shown that there is a correlation between consumption of champagne and long life expectancy. This doesn't imply that champagne improves life expectancy (in fact the reverse may be true): the result has probably arisen because wealthy people (who can afford lots of champagne) are more likely to be in good health than poorer people (the latter category will include most drug addicts, for example).

However, when using a Regression line it is assumed that there is in fact a causal link. For a given value of the x variable, the line is used to predict the corresponding y value.

(2) Justification for the formula $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

where $S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y})$, $S_{xx} = \sum(x_i - \overline{x})^2$ & $S_{yy} = \sum(y_i - \overline{y})^2$

First of all, we can set up a new origin at $(\overline{x}, \overline{y})$. The point (x_i, y_i) then becomes $(x_i - \overline{x}, y_i - \overline{y})$ relative to this origin.



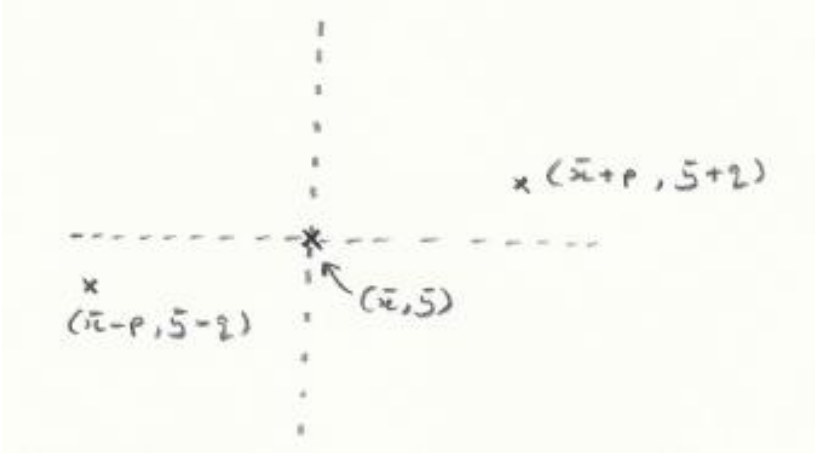
In the 1st and 3rd quadrants, $(x_i - \overline{x})(y_i - \overline{y})$ is positive, whilst it is negative in the 2nd and 4th quadrants.

Thus scatter diagrams where the points lie close to a line of best fit of positive gradient will have a high proportion of positive contributions to S_{xy}, leading to a large positive value of r.

Similarly, if the points tend to lie close to a line of best fit of negative gradient then there will be a high proportion of negative contributions to S_{xy}, leading to a large negative value of r.

The denominator has the effect of scaling down the value, depending on the distributions of the x and y values, to produce a value of r between -1 and 1.

It might be thought that the value of r should depend on the slope of the line of best fit, but this isn't in fact the case: for example, r can equal 1 whatever the gradient (provided it is positive). This can be seen from the following example, involving just 3 points.



$S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = (-p)(-q) + 0 + pq = 2pq$

$S_{xx} = \sum(x_i - \overline{x})^2 = (-p)^2 + 0 + p^2 = 2p^2$

$S_{yy} = \sum(y_i - \overline{y})^2 = (-q)^2 + 0 + q^2 = 2q^2$

so that $r = \frac{2pq}{\sqrt{(2p^2)(2q^2)}} = 1$

Also, if the points are $(\overline{x} - p, \overline{y} + q), (\overline{x}, \overline{y})$ & $(\overline{x} + p, \overline{y} - q)$ instead, then

$S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = (-p)q + 0 + p(-q) = -2pq$

$S_{xx} = \sum(x_i - \overline{x})^2 = (-p)^2 + 0 + p^2 = 2p^2$

$S_{yy} = \sum(y_i - \overline{y})^2 = (q)^2 + 0 + (-q)^2 = 2q^2$

so that $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-2pq}{\sqrt{(2p^2)(2q^2)}} = -1$

(3) Gradient of the Regression Line

The above example can also be used to justify the formula for the gradient of the Regression line (usually given the letter b):

b = $\frac{S_{xy}}{S_{xx}} = \frac{2pq}{2p^2} = \frac{q}{p}$, which is indeed the gradient of the line connecting the points in the diagram.