Regression (8 pages; 23/8/19)

(aka least squares regression, or line of best fit)

(1) Introduction

(1.1) Whereas correlation tests cannot establish a causal link between two variables; only that there is an 'association' (or connection) between them, when using a regression line to predict values it is generally assumed that there is a reasonably direct causal link.

In any event, there will be a correlation between the variables, but r need not be close to ± 1 for a regression line to be found. (However, the closer it is to ± 1, the better the predictions that can be made from the regression line.)

(1.2) Regression of y on x and regression of x on y.

The following situations can arise:

(a) 'random on non-random'

[Whilst the correlation coefficient is based on the assumption that both variables are random, this is not the case for the regression line.]

x is a 'controlled variable' (such as time - in which case it would obviously be labelled as t), and y is a random variable (eg a measurement taken at time t).

[It is customary to have the controlled variable on the x-axis.]

In this case, the regression line would be of y on x, and this can be used to predict the y value likely to be associated with a given x value.

The regression line of y on x could also be used to establish the likely x value, given that the y value is known (although there is no suggestion of the x value being *caused* by the y value). Note that we would not perform a regression of x on y (as the x value is not caused by the y value).

(b) 'random on random'

Both variables are random (eg heights and weights of individuals). It may be the case that one of the variables can be said to cause the other in some way (eg higher temperature giving rise to increased sales of ice-cream), or it may be that both variables depend on a third factor.

However, as neither variable is controlled, it is permissible to perform a regression of x on y, as well as a regression of y on x.

If we wish to predict the y value likely to be associated with a given x value, then we would use the regression line of y on x, and if the x value is to be predicted then we would use the regression line of x on y. As will be seen, the two lines are usually different, and so the two predictions will not be consistent.

(1.3) Terminology

Where one variable (eg y) depends on the other (eg x), then x may be described as the independent variable and y as the dependent variable (x may or may not be controlled). Alternatively, x may be described as the 'explanatory' (or 'predicting') variable and y as the 'response' (or 'explained') variable.

Examples

(a) x is the load on a spring and y is its extension

x is a controlled variable; x is the independent variable, whilst y is the dependent variable

(b) x is the height of a mother and y is the height of her daughter

x is the independent variable, and y is the dependent variable; x

would be said to be random, rather than controlled, as the heights of the mothers cannot be dictated; note that the dependence is less certain in this case than in (a), as the height of the father will have an effect as well (along with other factors - genetic and environmental)

(c) x is someone's wealth and y is their life span

In this case, the terms 'independent variable', and 'dependent variable' would not be used, as someone's wealth does not, in itself, determine their lifespan. There may well be an association however, making a regression line meaningful. Note that, despite the lack of a direct causal link, it makes sense to have the variables this way round.

(2) Equation of the regression line

(2.1) Derivations of the equation (for y on x)

Let the line be $y = a + bx$

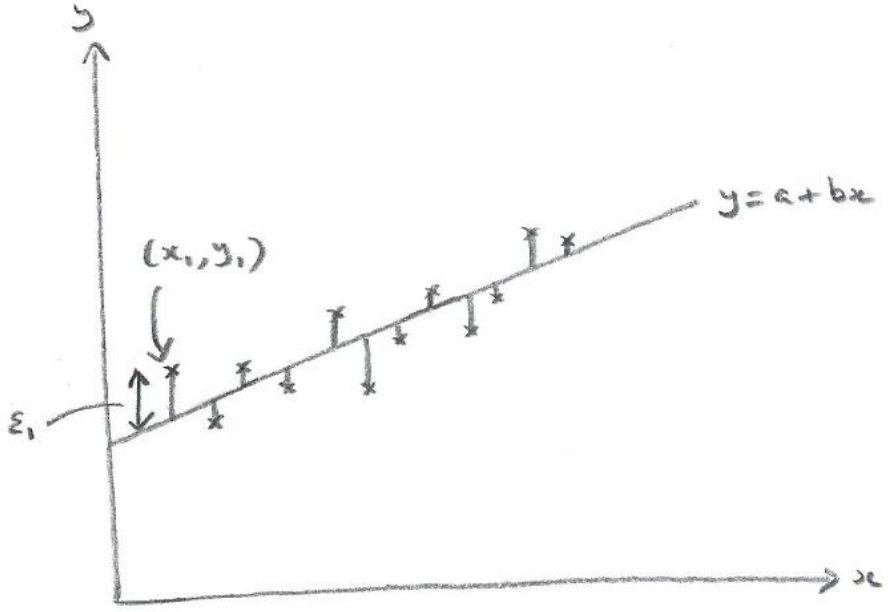
[Calculators sometimes use $y = ax + b$]

[b is referred to as the regression coefficient]

We want $(\overline{x}, \overline{y})$ to lie on the line, so that $\overline{y} = a + b\overline{x}$

For a given point (of data) (x_i, y_i),

$y_i = a + bx_i + \varepsilon_i$, where ε_i is the 'residual'



Let $T = \sum \varepsilon_i{}^2 = \sum (y_i - a - bx_i)^2$

The regression line is obtained by minimising T, which can be written as a quadratic in b. We then find the value of b that minimises T.

$T = \sum (y_i - [\overline{y} - b\overline{x}] - bx_i)^2$

$= \sum (y_i - \overline{y} - b(x_i - \overline{x}))^2$

$= b^2 \sum (x_i - \overline{x})^2 - 2b \sum (x_i - \overline{x})(y_i - \overline{y}) + \cdots$

(where the remaining term doesn't involve b)

Completing the square,

$$T = \{\sum (x_i - \overline{x})^2\} \left(b - \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}\right)^2 + \cdots$$

and to minimise T, we set $b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$

Notes

(i) Outliers should be removed, if this is considered appropriate.

Outliers could be determined by considering the sizes of residuals (ie initially performing the calculations including the suspected outliers, and then repeating the process with the outliers removed).

(ii) There is a link between r and b: $r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{{\sigma_{XY}}^2}{\sigma_X \sigma_Y}$ and

$b = \frac{S_{XY}}{S_{XX}} = \frac{{\sigma_{XY}}^2}{{\sigma_X}^2}$. So $b = r\frac{\sigma_Y}{\sigma_X}$

Exercise: Show that $\sum \varepsilon_i = 0$

Solution

$\sum \varepsilon_i = \sum(y_i - [a + bx_i])$

$= (\sum y_i) - na - b(\sum x_i)$

$= n\{\frac{\sum y_i}{n} - a - b\frac{\sum y_i}{n}\}$

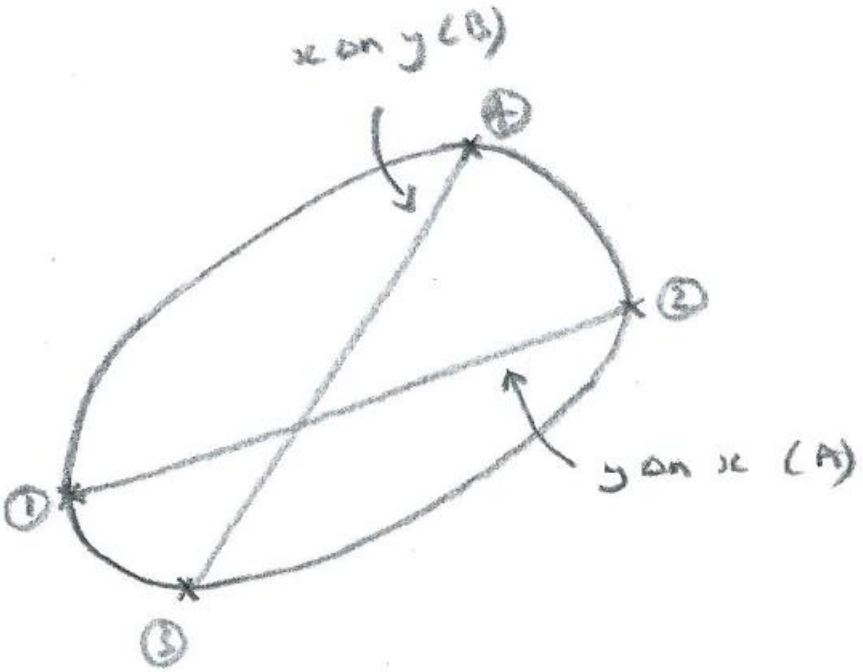
$= n\{\overline{y} - a - b\overline{x}\}$

$= 0,$ as $\overline{y} = a + b\overline{x}$

(2.2) Connection between the two lines of regression.

For the regression line of y on x, $\frac{y-\overline{y}}{x-\overline{x}} = \frac{S_{xy}}{S_{xx}} = b_A$, say;

and for the regression line of x on y, $\frac{x-\overline{x}}{y-\overline{y}} = \frac{S_{xy}}{S_{yy}} = b_B$, say.



[The data points have been omitted.]

Assuming that the scatter diagram shows an elliptical pattern, as in the diagram, then the regression lines of y on x (A) and x on y (B) will typically appear as shown. Note that the lines meet at $(\bar{x}, \bar{y})$.

For (A), the sum of the squares of the residual errors in y is being minimised, and it can be shown that the regression line will pass through the points (1) and (2), where $\frac{dx}{dy} = 0$. Similarly, the line for (B) will pass through (3) and (4), where $\frac{dy}{dx} = 0$.

The gradient of A is b_A, but the gradient of B is $\frac{1}{b_B}$ (as $\frac{x-\overline{x}}{y-\overline{y}} = b_B$, so that $\frac{y-\overline{y}}{x-\overline{x}} = \frac{1}{b_B}$).

$b_A b_B = \frac{S_{xy}}{S_{xx}} . \frac{S_{xy}}{S_{yy}} = r^2 \leq 1$,

so that $b_A \leq \frac{1}{b_B}$ (as in the diagram)

Also, when $r = \pm 1$, $b_A b_B = 1$, so that $b_A = \frac{1}{b_B}$ (and $\theta = 0$); ie the two regression lines are the same.

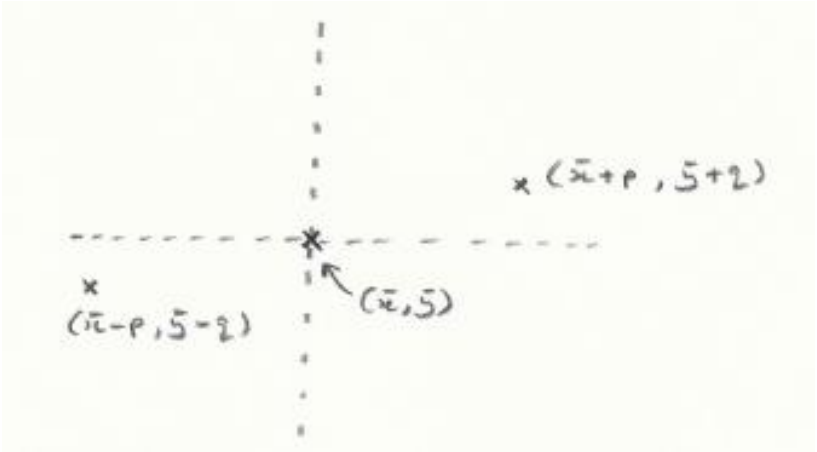
And $tan\theta = \frac{tan\theta_B - tan\theta_A}{1 + tan\theta_A tan\theta_B} = \frac{\frac{1}{b_B} - b_A}{1 + \frac{1}{b_B} b_A} = \frac{1 - b_A b_B}{b_B + b_A} = \frac{1 - r^2}{b_A + b_B}$

The angle between the two regression lines indicates the degree of dependence between the variables.

In the limiting case where $r = 0$ (so that the two variables are independent), the ellipse will become a circle, with $b_A = 0$ and $\frac{1}{b_B} = \infty$, so that the two lines are perpendicular.

(2.3) Gradient of the Regression Line

Consider the following example, involving just 3 points (referred to in "Correlation").



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

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

The formula for the gradient of the Regression line (usually given the letter b) can be justified for this example:

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

(3) Interpretation of a regression line

(3.1) The gradient can be used to say that "if x increases by ..., then y increases by ..."

(3.2) The predicted value of y, based on a particular value of x is commonly denoted: $\hat{y} = a + bx$

Whilst interpolation is generally safe, extrapolation is usually inadvisable. For example, it may be possible to demonstrate a negative correlation between a person's age and the time they take to run $100m$, between the ages of 10 and 20, but such a relation will not hold for older people.

(3.3) In order to assess the reliability of a predicted value, we can consider the size of the residuals in the region where the prediction is being made.

(4) Miscellaneous

(4.1) It may be convenient to 'code' the data, by a linear change of one or both of the variables (to give $x' = a + bx$ and/or

$y' = c + dy$)