Quantitative Methods for Business Decisions Sixth Edition
John Curwin & Roger Slater
ISBN: 9781844805747

Chapter 15 *Regression*

Derivation of the formula for Linear Regression

Equation of the straight line to predict y values is:

$$\hat{y} = a + bx$$

Thus for a value x_1, we have

$$\hat{y}_1 = a + bx_1$$

and

$$\begin{aligned} d_1 &= y_1 - \hat{y}_1 \\ &= y_1 - (a + bx_1) \\ &= y_1 - a - bx_1 \end{aligned}$$

where d is the difference between an observed y and its value predicted by the regression equation.

This vertical distance may be defined, and calculated as above, for each point on the scatter diagram. To minimize the prediction error, we may sum these vertical distances; however, if the observed y values deviate from the regression line in a random manner, then:

$$\sum d = 0$$

To overcome this problem, we may square each value of d, and then minimize the sum of these squares:

$$d_1 = y_1 - a - bx_1$$

$$d_1^2 = (y_1 - a - bx_1)^2$$

$$\begin{aligned} S &= \sum_{i=1}^{n} d_i^2 \\ &= \sum_{i=1}^{n} (y_i - a - bx_i)^2 \end{aligned}$$

To minimize S we need to differentiate the function with respect to a and b (since the values of x and y are fixed by the problem we are trying to solve). Thus:

$$\frac{\partial S}{\partial a} = -2\sum_{i=1}^{n}(y - a - bx_i) = 0$$

$$\frac{\partial S}{\partial b} = -2\sum_{i=1}^{n}x_i(y_i - a - bx_i) = 0$$

These are the **first-order conditions**, and if they are rearranged, we have the normal equations:

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

These equations could be used as a pair of simultaneous equations each time that we wanted to identify the values of a and b, but it is usually more convenient to find the general solution for a and b to give the following formulae.

Multiply equations 16.1 by $(\sum x)/n$:

$$\frac{\sum x \sum y}{n} = a\sum x + \frac{b}{n}(\sum x)^2$$

Subtracting equation 16.3 from equation 16.2:

$$\sum xy - \frac{\sum x \sum y}{n} = b\sum x^2 - \frac{b}{n}(\sum x)^2$$

$$= b\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$$

So

$$n\sum xy - \sum x \sum y = b\left[n\sum x^2 - (\sum x)^2\right]$$

or

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Rearranging equation 16.1 gives:

$$a = \frac{\sum y - b \sum x}{n}$$

or

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