Derivation of the Equation for the Least Squares Regression Line for an Arbitrary Number of Points

Rustem Bilyalov

January 13, 2011

The Least Squares Regression Line for a set of n points of the form (x_i, y_i) where $i \in \{1, 2, 3, ...\}$ is of the form $y = mx + b$. The sum of the squares of the distances from the line to each individual point along a vertical line, $R(m, b)$, is expressed as

$$R(m,b) = (x_1 m + b - y_1)^2 + (x_2 m + b - y_2)^2 + \cdots + (x_n m + b - y_n)^2.$$

A partial derivative of R with respect to m is then

$$\frac{\partial R}{\partial m} = 2x_1(x_1 m + b - y_1) + 2x_2(x_2 m + b - y_2) + \cdots + 2x_n(x_n m + b - y_n),$$

$$\frac{\partial R}{\partial m} = 2m \sum_{i=1}^{n} x_i^2 + 2b \sum_{i=1}^{n} x_i - 2 \sum_{i=1}^{n} x_i y_i.$$

Similarly a partial derivative of R with respect to b is

$$\frac{\partial R}{\partial b} = 2(x_1 m + b - y_1) + 2(x_2 m + b - y_2) + \cdots + 2(x_n m + b - y_n),$$

$$\frac{\partial R}{\partial b} = 2m \sum_{i=1}^{n} x_i + 2nb - 2 \sum_{i=1}^{n} y_i.$$

Setting both derivatives equal to 0 sets up a system of equations with two variables:

$$2m \sum_{i=1}^{n} x_i^2 + 2b \sum_{i=1}^{n} x_i - 2 \sum_{i=1}^{n} x_i y_i = 0$$

$$2m \sum_{i=1}^{n} x_i + 2nb - 2 \sum_{i=1}^{n} y_i = 0$$

Simplifying yields

$$m \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$m \sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i$$

Using Cramer's rule,

$$m = \frac{\begin{vmatrix} \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i & n \end{vmatrix}}{\begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix}} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}.$$

and

$$b = \frac{\begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i \end{vmatrix}}{\begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix}} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}.$$

This means the equation for a Least Squares Regression Line for a set of n points of the form (x_i, y_i) where $i \in \{1, 2, 3, ...\}$ is:

$$y = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} x + \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}.$$