Stephen Epstein

Derivation of the Linear Least Square Regression Line

Problem Statement

Linear Least Square Regression is a method of fitting an affine line to set of data points. This method is used throughout many disciplines including statistic, engineering, and science. The derivation of the formula for the Linear Least Square Regression Line is a classic optimization problem. Although used throughout many statistics books the derivation of the Linear Least Square Regression Line is often omitted. I will derive the formula for the Linear Least Square Regression Line and thus fill in the void left by many textbooks.

Procedure:

1. I will find the critical point for the sum of square errors formula
2. I will conjecture that this critical point must be the minimum point

Known:

The Linear Least Square Regression is the line that minimizes the sum of the square of the errors between the y-component of the fitted line and the y-component of the data points:

$$E(m,b) = \sum_{i=1}^{N} \left(y_i - (mx_i + b)\right)^2 \tag{1}$$

Solution:

Find the critical points

A point, $\mathbf{x_o}$, is a critical point of (1) if:

$$\nabla E(\mathbf{x_o}) = \mathbf{0} \tag{2}$$

Applying (2) to (1) yields:

$$\begin{pmatrix} \sum_{i=1}^{N} -2x_i(y_i - (mx_i + b)) \\ \sum_{i=1}^{N} -2(y_i - (mx_i + b)) \end{pmatrix} = \mathbf{0} \tag{3}$$

Transforming (3) into matrix form:

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} 1 \end{pmatrix} \begin{pmatrix} \mathrm{m} \\ \mathrm{b} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix} \tag{4}$$

Using Cramer's Rule to solve for m and b:

$$m = \frac{N \sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \tag{5}$$

$$b = \frac{\left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} y_i\right) - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} x_i y_i\right)}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \tag{6}$$

Using the following definitions of the mean of a set of data:

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{7}$$

$$\overline{y} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{8}$$

Substituting (7) and (8) into (5) and (6) reveals a more readable version:

$$m = \frac{\sum_{i=1}^{N} x_i y_i - \overline{x}\,\overline{y}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2} \tag{9}$$

$$b = \frac{\overline{y}\left(\sum_{i=1}^{N} x_i^2\right) - \overline{x}\left(\sum_{i=1}^{N} x_i y_i\right)}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2} \tag{10}$$

Finding the minimum of the sum of the square of the errors

The critical point given by (9) and (10) must the minimum of the sum of square errors. This can be seen by the fact that there is no maximizer of equation (1) and there is a minimizer. For a point to be a minimizer it must satisfy (1) and thus (9) and (10) minimize equation (1). This could analytically be proved by taking the Hessian of equation (1) and evaluating it at the point given by (9) and (10). This would show that the Hessian is positive definite and thus the point given by (9) and (10) is a minimizer.

The Linear Least Square Regression line

The Linear Least Square Regression line is simply the affine line where the slope (m) is given by (9) and the offset (b) is given by (10).

$$y = mx + b \tag{11}$$

Comments

By examination of equation (1) we notice that the error function is affected by the error squared. This avoids the problem of negative errors; however it leads to an un-proportional weighting of errors. For instance if a point is two units away from the LSQR line than its error with be 4 times that of a point one unit away from the LSQR. This fact points out the necessity of visually inspecting the data by graphical means.