Least Square Estimator

Pandega Abyan Zumarsyah

November 2023

Objective

Given some regressor data x and response data y, the Objective of Least Square is to obtain estimators b_0 and b_1 that minimize the sum of squared error (SSE):

$$SSE = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

Estimator b_1

To minimize SSE, partial derivative of SSE with respect to b_0 should be zero:

$$\begin{aligned}
\frac{\partial SSE}{\partial b_0} &= 0 \\
\frac{\partial(\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2)}{\partial b_0} &= 0 \\
\sum_{i=1}^{n}\frac{\partial(y_i - (b_0 + b_1 x_i))^2}{\partial b_0} &= 0 \\
\sum_{i=1}^{n}2(y_i - (b_0 + b_1 x_i))\frac{\partial(y_i - (b_0 + b_1 x_i))}{\partial b_0} &= 0 \\
\sum_{i=1}^{n}2(y_i - (b_0 + b_1 x_i))\left(\frac{\partial y_i}{\partial b_0} - \frac{\partial b_0}{\partial b_0} - \frac{\partial(b_1 x_i)}{\partial b_0}\right) &= 0 \\
\sum_{i=1}^{n}2(y_i - (b_0 + b_1 x_i))(0 - 1 - 0) &= 0 \\
\sum_{i=1}^{n}-2(y_i - (b_0 + b_1 x_i)) &= 0 \\
-\sum_{i=1}^{n}y_i + \sum_{i=1}^{n}b_0 + \sum_{i=1}^{n}b_1 x_i &= 0 \\
nb_0 + b_1\sum_{i=1}^{n}x_i &= \sum_{i=1}^{n}y_i
\end{aligned}$$

Not only that, partial derivative of SSE with respect to b_1 should also be zero:

$$
\begin{aligned}
\frac{\partial SSE}{\partial b_1} &= 0 \\
\frac{\partial(\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2)}{\partial b_1} &= 0 \\
\sum_{i=1}^{n} \frac{\partial(y_i - (b_0 + b_1 x_i))^2}{\partial b_1} &= 0 \\
\sum_{i=1}^{n} 2(y_i - (b_0 + b_1 x_i)) \frac{\partial(y_i - (b_0 + b_1 x_i))}{\partial b_1} &= 0 \\
\sum_{i=1}^{n} 2(y_i - (b_0 + b_1 x_i)) \left(\frac{\partial y_i}{\partial b_1} - \frac{\partial b_0}{\partial b_1} - \frac{\partial (b_1 x_i)}{\partial b_1} \right) &= 0 \\
\sum_{i=1}^{n} 2(y_i - (b_0 + b_1 x_i))(0 - 0 - x_i) &= 0 \\
\sum_{i=1}^{n} -2(x_i y_i - b_0 x_i - b_1 x_i^2) &= 0 \\
-\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} b_0 x_i + \sum_{i=1}^{n} b_1 x_i^2 &= 0 \\
b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i
\end{aligned}
$$

Previous two equations can be modified so that they are ready for elimination

The first one become:

$$
n b_0 \sum_{i=1}^{n} x_i + b_1 \left(\sum_{i=1}^{n} x_i \right)^2 = \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i
$$

While the second one become:

$$
n b_0 \sum_{i=1}^{n} x_i + n b_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i y_i
$$

Elimination can be done to obtain b_1:

$$
\begin{aligned}
b_1 \left(\sum_{i=1}^{n} x_i \right)^2 - n b_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - n \sum_{i=1}^{n} x_i y_i \\
b_1 &= \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}
\end{aligned}
$$

Estimator b_0

Since b_1 is defined, we can use it and the previous equation to define b_0:

$$\begin{aligned} nb_0 + b_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ nb_0 &= \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i \\ b_0 &= \frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} \\ b_0 &= \bar{y} - b_1 \bar{x} \end{aligned}$$