# Least Squares Regression

Harini Chandramouli
MATH2374

chand409@umn.edu

How do we find the best-fit line? Let's derive the least squares regression method!

First, what do we mean by best? Here we are choosing "best" to mean that we will minimize the sum of the squares of the vertical distances between the points and the line we want to relate. So let $(x_i, y_i)$ represent our points in question that we want to find a line for. We know our best fit line will be of the form $y = mx + b$, we just need to find $m$ and $b$. Let $\hat{y}$ represent the $y$ value on our line for any given $x_i$. Then, the quantity we want to minimize, if we have $n$ points is

$$E \coloneqq \sum_{i=1}^{n} (\hat{y} - y_i)^2.$$

Let's write the above expression with $m$ and $b$ in it, since these are the quantities we want to find. Well, $\hat{y}$ is the point that lies on the line when I plus in $x_i$, so it should be equal to $mx_i + b$, thus,

$$\begin{aligned}
\sum_{i=1}^{n} (\hat{y} - y_i)^2 &= \sum_{i=1}^{n} (mx_i - b - y_i)^2 \\
&= \sum_{i=1}^{n} (m^2 x_i^2 + 2bmx_i + b^2 - 2mx_i y_i - 2by_i + y_i^2) \\
&= m^2 \sum_{i=1}^{n} x_i^2 + 2bm \sum_{i=1}^{n} x_i + b^2 \sum_{i=1}^{n} 1 - 2m \sum_{i=1}^{n} x_i y_i - 2b \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i^2 \\
&= m^2 \sum_{i=1}^{n} x_i^2 + 2bm \sum_{i=1}^{n} x_i + b^2 \sum_{i=1}^{n} 1 - 2m \sum_{i=1}^{n} x_i y_i - 2b \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i^2.
\end{aligned}$$

So we want to minimize $m$ and $b$ so that they are the optimal numbers. The word "minimize" should ring some calculus bells in our head! We want to set the derivative equal to 0 and solve! If we want to minimize $m$, we take the derivative with respect to $m$. If we want to minimize $b$, we take the derivative with respect to $b$. Thus, we get

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

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

We notice that from above, there is only one $m$ and one $b$ that will come out of these equations since they are linear in $m$ and $b$. How do we know that these are not maximums or inflection points? We use the handy second derivative test!

$$\frac{\partial^2 E}{\partial m^2} = 2\sum_{i=1}^{n} x_i^2 > 0 \qquad\qquad \frac{\partial^2 E}{\partial b^2} = 2n > 0$$

Thus, we have minima since we are given that these functions are concave up. We rewrite (1) and (2) so that all the terms not involving $b$ and $m$ are on the right hand side. We simplify a bit as well and we obtain

$$m\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \tag{3}$$

$$m\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} 1 = \sum_{i=1}^{n} y_i. \tag{4}$$

We can rewrite this system of equations as

$$\begin{pmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} b \\ m \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

Which is exactly the formula found in your textbook (pg. 163, eq. (9))!

If you think about it, we have two equations and two unknowns, so we could solve for $m$ and $b$ explicitly. In fact, why don't we just to practice our skills? Solving (4) for b, we get that

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

Plugging this into (3) we get

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

$$\implies mn \sum_{i=1}^{n} x_i^2 + \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right) - m \left( \sum_{i=1}^{n} x_i \right)^2 = n \sum_{i=1}^{n} x_i y_i$$

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

Plugging this back into our formula for $b$,

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

Well that looks messy, but it's the right answer!