

Least Squares Approximation

Lecture notes for Access 2008

Let's say we have a collection of points, and they are "roughly" in a line. How do we find the equation of the line which best fits the data points?

First, what do we mean when we say that a line "best fits" the data points? We need a particular way of quantifying the distance between a collection of points and a line so that there is a usable formula to tell us the closest line. It turns out that Hausdorff distance, which worked so well in talking about fractals, is NOT the correct distance for this problem!

Example 1. Let's say we have the data points $(0,1)$, $(2,2)$, and $(4,1)$. We want the equation of a line $y = mx + b$ so that the line is close to our points.

We want to find the vertical distances from our points to the line, so we need to find the y -value for points on the line $y = mx + b$ which correspond to the x -values 0, 2, and 4. The points on the line are

$$(0, b), (2m + b), (4m + b).$$

The vertical distances from our points to the line are given by the equation $|y_2 - y_1|$, so we get the following vertical distances:

$$d_1 = |b - 1|, d_2 = |2m + b - 2|, d_3 = |4m + b - 1|.$$

To analyze this data, we want positive distances, which is why we use the absolute values. But absolute values can make some things difficult. Instead, let's just square each of the quantities. Our distance analyzing function is then

$$R(m, b) = (b - 1)^2 + (2m + b - 2)^2 + (4m + b - 1)^2.$$

Notice that m and b are now our variables, or our unknown values. This function $R(m, b)$, often called the "sum of the squared vertical deviations," is how we will measure the distance from the given points to the line $y = mx + b$. We want to minimize $R(m, b)$.

When we minimize/maximize a function, we want to find where the slope of the function is zero, i.e., where the derivative of the function is zero.

If you have not had calculus, and don't know what a derivative is, don't worry about it. You will learn about it eventually. In the next section just pay attention to the equations we get, even if you don't know how we got them.

For those of you who know how to take derivatives, the only difference with taking a derivative of a function with two variables is we pretend that one variable is really just a constant. This way we have a function in one variable and can take derivatives normally. The only difference is that when we do this, we use a fancy symbol which says we are only concentrating on one of the variables. This is called a partial derivative.

In our example we have:

$$\frac{\partial R}{\partial m} = 2(2m + b - 2)(2) + 2(4m + b - 1)(4) = 40m + 12b - 16$$

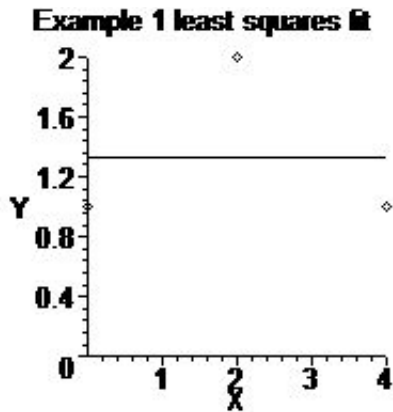
$$\frac{\partial R}{\partial b} = 2(b - 1) + 2(2m + b - 2) + 2(4m + b - 1) = 12m + 6b - 8.$$

We want to find the values of m and b which make the derivatives zero, so we solve the following equations:

$$0 = 40m + 12b - 16$$

$$0 = 12m + 6b - 8$$

We can solve this with substitution or elimination or matrices, and we get that $m = 0$ and $b = \frac{4}{3}$. Therefore the equation of the line which best fits the data is $y = \frac{4}{3}$.



Symbolically

1. Obtain data points, and possibly graph them to see if they roughly look like a line. Our data points will be $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n)$.
2. Find the vertical distances from the points to a line $y = mx + b$ with the following equation: For each i , find $d_i = mX_i + b - Y_i$.
3. Make the analyzing function $R(m, b)$ which is the sum of the squares of our vertical distances, i.e., $R(m, b) = \sum_{i=1}^n (mX_i + b - Y_i)^2$.
4. Find the partial derivatives of $R(m, b)$.

$$\frac{\partial R}{\partial m} = 2 \sum_{i=1}^n (mX_i + b - Y_i)X_i$$

$$\frac{\partial R}{\partial b} = 2 \sum_{i=1}^n (mX_i + b - Y_i)$$

5. Solve the equations $\frac{\partial R}{\partial m} = 0$ and $\frac{\partial R}{\partial b} = 0$ to find the values of m and b which minimizes the distance from the points to the line.
6. Using those values for m and b , write the equation $y = mx + b$ and graph it on the same graph with the data points.

Example 2. Find the line which best fits the data points $(0,3)$, $(1,5)$, and $(2,5)$.

Now that we know how to do this process, and why, it is good to know that most mathematical programs have this Least Squares function built into it. For example, Maple has this feature. We will now do some examples using Maple so we don't have to get bogged down in the calculations.

Open up Maple, and open the URL
<http://www.math.utah.edu/~erin/Access/LeastSquares.mws>