Least Mean Squares

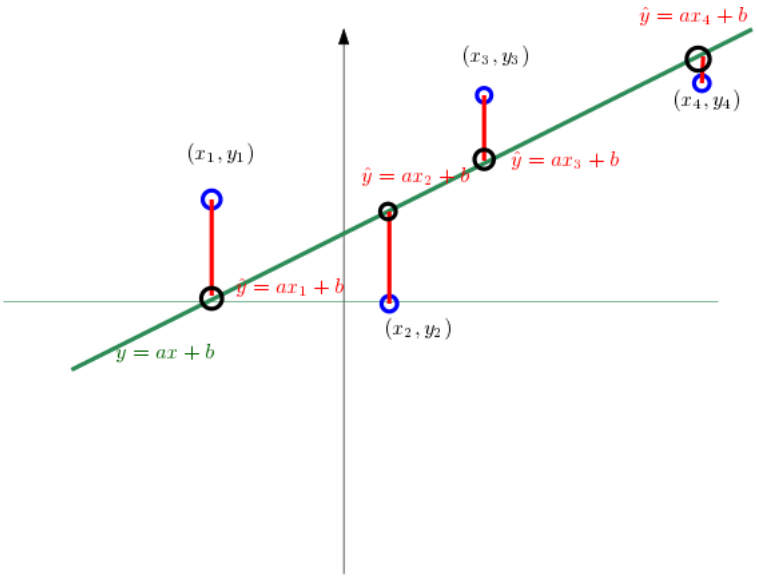
1 What is the "line of best fit"?

The book suggests that, given a scatterplot of data in a plane that we would like to summarize with a straight line that "approximates as best as we can" the data, we use a calculator with a dedicated key (probably labeled *LMS*) to find an equation for this purpose. But what is the calculator doing?

The logic behind the *Least Mean Squares* or *Regression* line is usually discussed in introductory Statistics classes, since the most stringent argument for its use relies on a specific (probabilistic) modeling for the data. We will not go into the logic of the argument, but the result is the following.

Suppose we have a number of data points of the form $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. In many cases, we think of the first coordinates as fixed (for example, they may be time points, as in years), and of the second coordinates as subject to deviations from an exact linear dependence on the first. For example, we might think that there should be an approximate linear relation between education level and average salary, but the observations will certainly not lead to an exact line. We now try to find a line such that, the y's corresponding to the x_i's by the line will differ "as little as possible for the observed y_i's". There is more than one way to define "as little as possible", but the most common one is to choose the line $y = ax + b$ that makes the sum of the squares of the differences $y_i - ax_i - b$ as small as possible (hence *Least Mean Squares*).

Why the squares? When the statistical model we mentioned applies, there is a rigorous mathematical reason for this choice, but this method is applied much more broadly, in which case, the best justification is that the math is much more convenient than with other choices (as in the sum of the absolute values of the differences, which can also be used, but is more cumbersome to work with).



In this picture, the blue circles are the data points, the green line is a candidate for the LMS line, and the black circles are the points on this line aliasing the actual data points. The red segments are the discrepancy between the "theoretical" points $(x_i, \hat{y}_i = ax_i + b)$ and the true data points (x_i, y_i). The line we will choose will be the one that makes the sum of the squares of the lengths of the red segments smallest.

This means, of four data points as in the picture, to find a and b such that

$$\begin{aligned}(y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + (y_3 - ax_3 - b)^2 + (y_4 - ax_4 - b)^2 = \\ = a^2\left(x_1^2 + x_2^2 + x_3^2 + x_4^2\right) + 4b^2 - 2ab\left(x_1 + x_2 + x_3 + x_4\right) - \\ -2a\left(x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4\right) - 2b\left(y_1 + y_2 + y_3 + y_4\right) + \\ + y_1^2 + y_2^2 + y_3^2 + y_4^2\end{aligned}$$

is as small as possible. Let's write this function of a and b in abbreviated notation as

$$F(a, b) = a^2X_2 + 4b^2 - 2abX_1 - 2aXY - 2bY_1 + Y_2 \tag{1}$$

This is a quadratic function, but it depends on two variables, so we don't really have a tool to find its lowest values right now (in your second Precalculus class you will be introduced to quadratic functions in two variables). We can find the lowest value in a special case: suppose that, for general reasons, we are looking for a line that has to go though the origin $(0, 0)$ (e.g., we are matching observations to a theoretical model that expects the relation to be of the form $y = ax$ – an example would be Ohm's law connecting current intensity and voltage as $I = \frac{1}{R}V$, where R is the resistance of the circuit, and using measurements of current corresponding to various voltages to determine $\frac{1}{R}$). Since we are forcing

$b = 0$, we now have a quadratic function in a, $a^2 X_2 - 2aXY + Y_2$, and it slowest value (the quadratic function has a positive coefficient for a^2, so its vertex is a minimum) occurs at

$$a = \frac{2XY}{2X_2} = \frac{x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4}{x_1^2 + x_2^2 + x_3^2 + x_4^2}$$

To find a formula for a and b in the general case, while we wait for additional tools, we might accept, on trust, that a and b should be related by the equation

$$\frac{y_1 + y_2 + y_3 + y_4}{4} = a\frac{x_1 + x_2 + x_3 + x_4}{4} + b \tag{2}$$

so that

$$b = \frac{y_1 + y_2 + y_3 + y_4}{4} - a\frac{x_1 + x_2 + x_3 + x_4}{4} = \frac{Y_1}{4} - a\frac{X_1}{4} \tag{3}$$

For a plausibility argument why they should be connected by the averages of the data look at the end of this file.

Substituting in the expression (1) results in a quadratic function of a only, whose minimum we can again find using the vertex formula. The end result is what your calculator computes when asked for the regression equation:

$$a = \frac{4XY - X_1 Y_1}{4X_2 - (X_1)^2}$$

and b given by (3). Of course, all the occurrences of "4" in these formulas are simply the size of the data set. If we had 20 points, we would get the same formulas, but with 20 in place of 4.

Why "Regression"?

The name has nothing to do with the mathematical tool we are using. It refers to one of the first, and famous, applications of the method, where the heights of children were compared with the heights of their parents. The data suggested that children of tall parents tend to be taller than the mean, and children of short parents to be shorter than the mean, but in both cases the deviation from the mean tended to be less pronounced, hence to exhibit a *regression to the mean.*

Why equation (3)?

It is not unreasonable that averages are the quantities that minimize squared discrepancies. A simple motivation can be found in this fact:

Fact: *Given data points $x_1, x_2, \ldots, x_n$, the number m minimizing the sum of the squares of the differences $x_i - m$ is the average $\bar{x} = \frac{x_1 + x_2 \ldots + x_n}{n}$.*

The proof is again in the vertex formula:

$$(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_n - m)^2 = nm^2 - 2m\,(x_1 + x_2 + \ldots + x_n) + x_1^2 + \ldots x_n^2$$

a quadratic function (in m), whose vertex is at

$$m = \bar{x} = \frac{x_1 + \ldots + x_n}{n} \tag{4}$$

Another rough argument for (3)

The idea (from the statistics approach) of regression is that the y_i's and x_i's should be related by

$$y_i = ax_i + b + \varepsilon_i \tag{5}$$

, where the ε_i's are "errors" that average out to zero. So, taking the mean of the two sides in (5) we would have (add all terms and divide by n), using "sigma-notation"[1]

$$\sum y_i = a \sum x_i + nb + \sum \varepsilon_i$$

$$\frac{1}{n} \sum y_i = a \frac{1}{n} \sum x_i + b + \frac{1}{n} \sum \varepsilon_i$$

Since we assumed that $\overline{\varepsilon} \equiv \frac{1}{n} \sum \varepsilon_i = 0$, we have the generalization of (2) to n data points. The argument is not really a proof, since the assumption of zero error average is not really solid for a finite number of data points (again, see a statistics discussion for this).

Can we address the general case?

You will notice from (1) that, in the particular case when $X_1 = x_1 + x_2 + x_3 + x_4 = 0$, the function can be seen as the sum of two separate quadratic functions, one of of a and one of function of b. It is intuitive (and true) that the lowest value can be found by minimizing the two functions separately. And, indeed, in this case, the resulting formulas are the *LMS* solution. Still, what if that is not the case? We can still find the appropriate formula if we *shift* the x_k data by their mean, that is working the regression line for the pairs $(z_k = x_k - \overline{x}, y_k)$ ($\overline{x}$ is the mean in (4)). You can check yourself that there will be no ab term in this new coordinates, and that we can recover the "true" line by shifting back the line computed in terms of the z_k's.

One last remark

The common way to justify the linear Least Mean Squares formula is through *differential* calculus (you need calculus for functions of two variable). However, this problem is about a polynomial (even if in two variables), and it turns out that what you can do about polynomials with calculus, you can do with a little algebra and a few simple observations, without the need to bring in the heavy artillery. You may find a discussion of this in another complementary file.

[1] the shorthand $\sum x_i$ stands for $x_1 + x_2 + \ldots + x_n$

2 Fitting Nonlinear Data

2.1 Polynomial Fits

Finding a line of best fit, as discussed above, is a common method for identifying trends, even when the scatterplot does not really suggest a linear model. However, there are many cases where a linear approximation will just not be a good summary of the data.

The book suggests to extend the Least Mean Squares Method and approximate data with peaks and valleys with a polynomial. This is technically easy (although it requires to find the minimum of a function of more than two variables, and is thus definitely out of our toolbox), but is often not a very good approach.

For one thing, the rigorous underpinning of linear regression is no longer available. That is, least mean squares rests on a much more ad-hoc basis.

Mainly, though, the problem is that polynomials, while very important as tools, are rarely a theoretically justified model, as opposed to linear functions and, coming up soon, exponential and logarithmic functions. Even quadratic functions have a relatively limited scope: notice, from examples and exercises, how they work for the motion of a body thrown in a gravity field (and then, only ignoring air friction, and only for trajectories close to earth), and for planning rectangular fences, but little else.

It is true that data exhibiting a turnaround (a maximum or a minimum) can often be well approximated with a quadratic function, but this approximation breaks down almost inevitably as soon as we look farther from the turnaround, hence, such a model has almost no value for prediction purposes (which is, after all, the main advantage of having a model).

In fact, trying to find a polynomial that will mimic a number of wiggles in the data may easily lead to *overfitting*, that is coming up with a function that chases after every little perturbation in the data, masking whatever core behavior may lie behind, and failing dramatically as soon as the model is applied beyond the range of the available data. After all, n data points can always be *exactly* fitted with a polynomial of degree $n-1$, but such a fit is completely useless as it will miss any subsequent additional data, and will lead to completely unrealistic forecasts.

2.2 Exponential and power fits

Many situations lead to exponential models (e.g., population dynamics, radioactive decay, compound interest, and more), so fitting data that should be modeled this way is useful. Once again, the least mean squares approach is convenient, but has little theoretical underpinning. One way is to fit data that we hope should fit like $\left(x, Ae^{kx}\right)$, by first taking the logarithm of the second component, as in $(x, \ln A + kx)$ that corresponds to a linear model. Though fairly common, the logic of using an LMS approach is even weaker, since it is not "natural" to assume that it is the logarithm of the data that is affected by an error with

common spread.

While polynomial fitting is usually completely ad-hoc, it may happen that some probabilistic models will lead to *power* functions being a good candidate. These are especially popular with modeling of rare events, like the time of occurrence of earthquakes. This is an interesting area, but you would have to go beyond even elementary statistics for an overview.