

The Method of Least Squares

Steven J. Miller*

Mathematics Department
Brown University
Providence, RI 02912

Abstract

The Method of Least Squares is a procedure to determine the best fit line to data; the proof uses simple calculus and linear algebra. The basic problem is to find the best fit straight line $y = ax + b$ given that, for $n \in \{1, \dots, N\}$, the pairs (x_n, y_n) are observed. The method easily generalizes to finding the best fit of the form

$$y = a_1 f_1(x) + \dots + c_K f_K(x); \quad (0.1)$$

it is not necessary for the functions f_k to be linearly in x – all that is needed is that y is to be a linear combination of these functions.

Contents

1	Description of the Problem	1
2	Probability and Statistics Review	2
3	The Method of Least Squares	4

1 Description of the Problem

Often in the real world one expects to find linear relationships between variables. For example, the force of a spring linearly depends on the displacement of the spring: $y = kx$ (here y is the force, x is the displacement of the spring from rest, and k is the spring constant). To test the proposed relationship, researchers go to the lab and measure what the force is for various displacements. Thus they assemble data of the form (x_n, y_n) for $n \in \{1, \dots, N\}$; here y_n is the observed force in Newtons when the spring is displaced x_n meters.

*E-mail: sjmiller@math.brown.edu

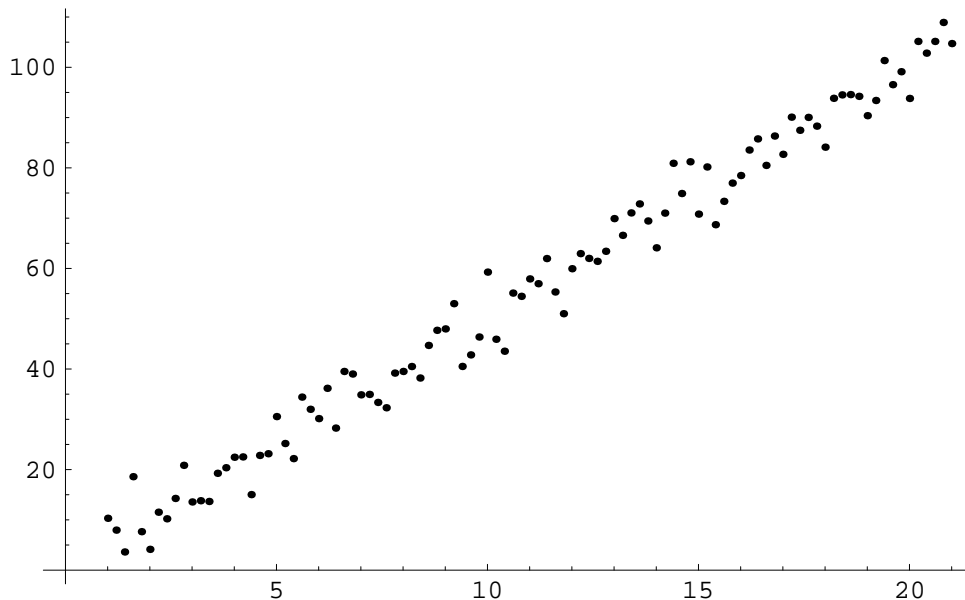


Figure 1: 100 “simulated” observations of displacement and force ($k = 5$).

Unfortunately, it is extremely unlikely that we will observe a perfect linear relationship. There are two reasons for this. The first is experimental error; the second is that the underlying relationship may not be exactly linear, but rather only approximately linear. See Figure 1 for a simulated data set of displacements and forces for a spring with spring constant equal to 5.

The Method of Least Squares is a procedure, requiring just some calculus and linear algebra, to determine what the “best fit” line is to the data. Of course, we need to quantify what we mean by “best fit”, which will require a brief review of some probability and statistics.

A careful analysis of the proof will show that the method is capable of great generalizations. Instead of finding the best fit line, we could find the best fit given by *any* finite linear combinations of specified functions. Thus the general problem is given functions f_1, \dots, f_K , find values of coefficients a_1, \dots, a_K such that the *linear* combination

$$y = a_1 f_1(x) + \dots + a_K f_K(x) \tag{1.1}$$

is the best approximation to the data.

2 Probability and Statistics Review

We give a quick introduction to the basic elements of probability and statistics which we need for the Method of Least Squares; for more details see [BD, CaBe, Du, Fe, Kel, LF, MoMc].

Given a sequence of data x_1, \dots, x_N , we define the **mean** (or the **expected value**) to be

$(x_1 + \dots + x_N)/N$. We denote this by writing a line above x : thus

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n. \quad (2.2)$$

The mean is the average value of the data.

Consider the following two sequences of data: $\{10, 20, 30, 40, 50\}$ and $\{30, 30, 30, 30, 30\}$. Both sets have the same mean; however, the first data set has greater variation about the mean. This leads to the concept of variance, which is a useful tool to quantify how much a set of data fluctuates about its mean. The variance of $\{x_1, \dots, x_N\}$, denoted by σ_x^2 , is

$$\sigma_x^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2; \quad (2.3)$$

the standard deviation σ_x is the square root of the variance:

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2}. \quad (2.4)$$

Note that if the x 's have units of meters then the variance σ_x^2 has units of meters², and the standard deviation σ_x and the mean \bar{x} have units of meters. Thus it is the standard deviation that gives a good measure of the deviations of the x 's around their mean.

There are, of course, alternate measures one can use. For example, one could consider

$$\frac{1}{N} \sum_{n=1}^N (x_n - \bar{x}). \quad (2.5)$$

Unfortunately this is a signed quantity, and large positive deviations can cancel with large negatives. In fact, the definition of the mean immediately implies the above is zero! This, then, would be a terrible measure of the variability in data, as it is zero regardless of what the values of the data are.

We can rectify this problem by using absolute values. This leads us to consider

$$\frac{1}{N} \sum_{n=1}^N |x_n - \bar{x}|. \quad (2.6)$$

While this has the advantage of avoiding cancellation of errors (as well as having the same units as the x 's), the absolute value function is not a good function analytically. It is not differentiable. This is primarily why we consider the standard deviation (the square root of the variance) – this will allow us to use the tools from calculus.

We can now quantify what we mean by “best fit”. If we believe $y = ax + b$, then $y - (ax + b)$ should be zero. Thus given observations

$$\{(x_1, y_1), \dots, (x_N, y_N)\}, \quad (2.7)$$

we look at

$$\{y_1 - (ax_1 + b), \dots, y_N - (ax_N + b)\}. \quad (2.8)$$

The mean should be small (if it is a good fit), and the variance will measure how good of a fit we have.

Note that the variance for this data set is

$$\sigma_{y-(ax+b)}^2 = \frac{1}{N} \sum_{n=1}^N (y_n - (ax_n + b))^2. \quad (2.9)$$

Large errors are given a higher weight than smaller errors (due to the squaring). Thus our procedure favors many medium sized errors over a few large errors. If we used absolute values to measure the error (see equation (2.6)), then all errors are weighted equally; however, the absolute value function is not differentiable, and thus the tools of calculus become inaccessible.

3 The Method of Least Squares

Given data $\{(x_1, y_1), \dots, (x_N, y_N)\}$, we may define the error associated to saying $y = ax + b$ by

$$E(a, b) = \sum_{n=1}^N (y_n - (ax_n + b))^2. \quad (3.10)$$

This is just N times the variance of the data set $\{y_1 - (ax_1 + b), \dots, y_N - (ax_N + b)\}$. It makes no difference whether or not we study the variance or N times the variance as our error, and note that the error is a function of two variables.

The goal is to find values of a and b that minimize the error. In multivariable calculus we learn that this requires us to find the values of (a, b) such that

$$\frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b} = 0. \quad (3.11)$$

Note we do not have to worry about boundary points: as $|a|$ and $|b|$ become large, the fit will clearly get worse and worse. Thus we do not need to check on the boundary.

Differentiating $E(a, b)$ yields

$$\begin{aligned} \frac{\partial E}{\partial a} &= \sum_{n=1}^N 2(y_n - (ax_n + b)) \cdot (-x_n) \\ \frac{\partial E}{\partial b} &= \sum_{n=1}^N 2(y_n - (ax_n + b)) \cdot 1. \end{aligned} \quad (3.12)$$

Setting $\partial E/\partial a = \partial E/\partial b = 0$ (and dividing by 2) yields

$$\begin{aligned}\sum_{n=1}^N (y_n - (ax_n + b)) \cdot x_n &= 0 \\ \sum_{n=1}^N (y_n - (ax_n + b)) &= 0.\end{aligned}\tag{3.13}$$

We may rewrite these equations as

$$\begin{aligned}\left(\sum_{n=1}^N x_n^2\right) a + \left(\sum_{n=1}^N x_n\right) b &= \sum_{n=1}^N x_n y_n \\ \left(\sum_{n=1}^N x_n\right) a + \left(\sum_{n=1}^N 1\right) b &= \sum_{n=1}^N y_n.\end{aligned}\tag{3.14}$$

We have obtained that the values of a and b which minimize the error (defined in (3.10)) satisfy the following matrix equation:

$$\begin{pmatrix} \sum_{n=1}^N x_n^2 & \sum_{n=1}^N x_n \\ \sum_{n=1}^N x_n & \sum_{n=1}^N 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^N x_n y_n \\ \sum_{n=1}^N y_n \end{pmatrix}.\tag{3.15}$$

We will show the matrix is invertible, which implies

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^N x_n^2 & \sum_{n=1}^N x_n \\ \sum_{n=1}^N x_n & \sum_{n=1}^N 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{n=1}^N x_n y_n \\ \sum_{n=1}^N y_n \end{pmatrix}.\tag{3.16}$$

Denote the matrix by M . The determinant of M is

$$\det M = \sum_{n=1}^N x_n^2 \cdot \sum_{n=1}^N 1 - \sum_{n=1}^N x_n \cdot \sum_{n=1}^N x_n.\tag{3.17}$$

As

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n,\tag{3.18}$$

we find that

$$\begin{aligned}\det M &= N \sum_{n=1}^N x_n^2 - (N\bar{x})^2 \\ &= N^2 \left(\frac{1}{N} \sum_{n=1}^N x_n^2 - \bar{x}^2 \right) \\ &= N^2 \cdot \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2,\end{aligned}\tag{3.19}$$

where the last equality follows from simple algebra. Thus, as long as all the x_n are not equal, $\det M$ will be non-zero and M will be invertible.

Thus we find that, so long as the x 's are not all equal, the best fit values of a and b are obtained by solving a linear system of equations; the solution is given in (3.16).

Remark 3.1. The data plotted in Figure 1 was obtained by letting $x_n = 5 + .2n$ and then letting $y_n = 5x_n$ plus an error randomly drawn from a normal distribution with mean zero and standard deviation 4 ($n \in \{1, \dots, 100\}$). Using these values, we find a best fit line of

$$y = 4.99x + .48; \tag{3.20}$$

thus $a = 4.99$ and $b = .48$. As the expected relation is $y = 5x$, we expected a best fit value of a of 5 and b of 0.

While our value for a is very close to the true value, our value of b is significantly off. We deliberately chose data of this nature to indicate the dangers in using the Method of Least Squares. Just because we know 4.99 is the best value for the slope and .48 is the best value for the y -intercept does *not* mean that these are good estimates of the true values. The theory needs to be supplemented with techniques which provide error estimates. Thus we want to know something like, given this data, there is a 99% chance that the true value of a is in (4.96, 5.02) and the true value of b is in $(-.22, 1.18)$; this is far more useful than just knowing the best fit values.

If instead we used

$$E_{\text{abs}}(a, b) = \sum_{n=1}^N |y_n - (ax_n + b)|, \tag{3.21}$$

then numerical techniques yield that the best fit value of a is 5.03 and the best fit value of b is less than 10^{-10} in absolute value. The difference between these values and those from the Method of Least Squares is in the best fit value of b (the least important of the two parameters), and is due to the different ways of weighting the errors.

Exercise 3.2. *Generalize the method of least squares to find the best fit quadratic to $y = ax^2 + bx + c$ (or more generally the best fit degree m polynomial to $y = a_mx^m + a_{m-1}x^{m-1} + \dots + a_0$).*

While for any real world problem, direct computation determines whether or not the resulting matrix is invertible, it is nice to be able to prove the determinant is always non-zero for the best fit line (if all the x 's are not equal).

Exercise 3.3. *If the x 's are not all equal, must the determinant be non-zero for the best fit quadratic or the best fit cubic?*

Looking at our proof of the Method of Least Squares, we note that it was not essential that we have $y = ax + b$; we could have had $y = af(x) + bg(x)$, and the arguments would have

proceeded similarly. The difference would be that we would now obtain

$$\begin{pmatrix} \sum_{n=1}^N f(x_n)^2 & \sum_{n=1}^N f(x_n)g(x_n) \\ \sum_{n=1}^N f(x_n)g(x_n) & \sum_{n=1}^N g(x_n)^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^N f(x_n)y_n \\ \sum_{n=1}^N g(x_n)y_n \end{pmatrix}. \quad (3.22)$$

Exercise 3.4. Consider the generalization of the Method of Least Squares given in (3.22). Under what conditions is the matrix invertible?

Exercise 3.5. The method of proof generalizes further to the case when one expects y is a linear combination of K fixed functions. The functions need not be linear; all that is required is that we have a linear combination, say $a_1 f_1(x) + \cdots + a_K f_K(x)$. One then determines the a_1, \dots, a_K that minimize the variance (the sum of squares of the errors) by calculus and linear algebra. Find the matrix equation that the best fit coefficients (a_1, \dots, a_K) must satisfy.

Exercise 3.6. Consider the best fit line from the Method of Least Squares, so the best fit values are given by (3.16). Is the point (\bar{x}, \bar{y}) , where $\bar{x} = \frac{1}{n} \sum_{n=1}^N x_n$ and $\bar{y} = \frac{1}{n} \sum_{n=1}^N y_n$, on the best fit line? In other words, does the best fit line go through the “average” point?

References

- [BD] P. Bickel and K. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco, 1977.
- [CaBe] G. Casella and R. Berger, *Statistical Inference*, 2nd edition, Duxbury Advanced Series, Pacific Grove, CA, 2002.
- [Du] R. Durrett, *Probability: Theory and Examples*, 2nd edition, Duxbury Press, 1996.
- [Fe] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd edition, Vol. II, John Wiley & Sons, New York, 1971.
- [Kel] D. Kelley, *Introduction to Probability*, Macmillan Publishing Company, London, 1994.
- [LF] R. Larson and B. Farber, *Elementary Statistics: Picturing the World*, Prentice-Hall, Englewood Cliffs, NJ, 2003.
- [MoMc] D. Moore and G. McCabe, *Introduction to the Practice of Statistics*, W. H. Freeman and Co., London, 2003.