Least Squares Analysis and Curve Fitting

Don C. Warrington
Departments of Civil and Mechanical Engineering
University of Tennessee at Chattanooga

This is a brief overview of least squares analysis. It begins by explaining the difference between interplation and least squares analysis using basic linear algebra. From there vector norms and their relationship with the residual is discussed. Non-linear regression, such as is used with exponential and power regression, is explained. Finally a worked example is used to show the various regression schemes applied to a data set.

Keywords: least squares, regression, residuals, linear algebra, logarithmic plotting

Introduction

Curve fitting–particularly linear "curve" fitting–is a well known technique among engineers and scientists. In the past the technique was generally applied graphically, i.e., the engineer or scientist would plot the points and then draw a "best fit" line among the points, taking into consideration outliers, etc.

In reality, curve fitting is a mathematical technique which involves the solution of multiple equations, invoking the use of linear algebra and statistical considerations. This is the way a spreadsheet would look at the problem, and students and practitioners alike utilize this tool without really understanding how they do their job. That understanding, however, can be critical; numerical methods, while capable of excellent results, can also veer into poor ones without much warning. Avoiding problems such as this requires that the engineer look at the results before he or she uses them.

This is a brief introduction to the subject. Here we attempt to present the concepts in a way which utilizes basic concepts to understand some relatively advanced ones. Many presentations get lost in the theory, and the students are likewise lost; we attempt to avoid this here.

Linear Interpolation

Let us begin by considering the equation of a line, thus

$$y = mx + b \tag{1}$$

It's worth stopping here and noting that there are only two mathematical operations going one here: addition and scalar multiplication. In linear algebra, a vector space is defined as a set where all the elements are either sums of two elements, scalar multiples of two elements, or a combination of both (Gelfand (1961).) Thus this simple equation is an excellent illustration of the connection between linear algebra–to which we will have recourse–and basic graphical concepts.

We know we can define a line using two points. Working in two dimensions, we can write these two equations as follows:

$$\begin{aligned} y_1 &= mx_1 + b \\ y_2 &= mx_2 + b \end{aligned} \tag{2}$$

In matrix form, this is

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{3}$$

We can also write this as

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = 0 \tag{4}$$

which will become handy shortly.

Equation 3 is easy to solve; we can invert the matrix thus:

$$A^{-1} = \begin{bmatrix} (x_1 - x_2)^{-1} & -(x_1 - x_2)^{-1} \\ -\frac{x_2}{x_1 - x_2} & \frac{x_1}{x_1 - x_2} \end{bmatrix} \tag{5}$$

We then premultiply the right hand side of Equation 3 by this to obtain

$$\begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \frac{y_1 - y_2}{x_1 - x_2} \\ \frac{x_1 y_2 - x_2 y_1}{x_1 - x_2} \end{bmatrix} \tag{6}$$

From this we can compute the slope m and y-intercept b. We should also note that the line passes through both points perfectly; this is the physical meaning of Equation 4. This is illustrated in Figure 1. In this case $m = 2$, $b = 1$.

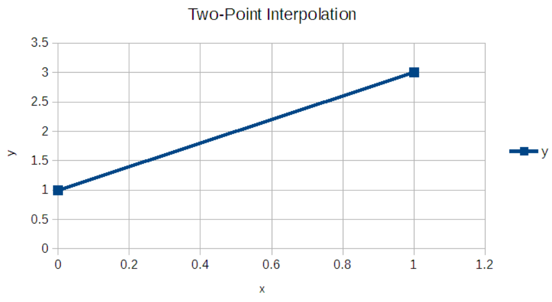

Figure 1. Two-Point Interpolation

Adding Points

Now let us consider the situation where we have three (3) points. Equation 3 then becomes

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \tag{7}$$

The matrix-vector multiplication is fine, but we can't invert the matrix on the left hand side. What we have here is a situation where, for a linear solution, we have overdefined the problem: we have more equations than we have unknowns. This situation is illustrated in Figure 2.

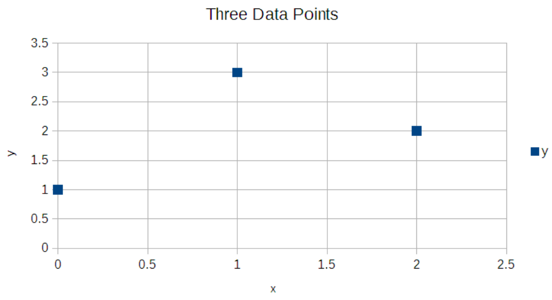

Figure 2. Three Data Points

So what is to be done? We have two choices. The first is to "square up" the matrix on the left hand side of Equation 7, which would allow us to solve the problem in the same way as we had with Equation 3. We could define a new "line" (it's actually a parabola) as follows:

$$y = nx^2 + mx + b \tag{8}$$

In this case, Equation 7 becomes

$$\begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ x_3^2 & x_3 & 1 \end{bmatrix} \begin{bmatrix} n \\ m \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \tag{9}$$

This would result in a second-order equation that would pass through the three points. We could keep expanding this with an increasing number of points, although there are practical limits. This is true interpolation: the curve that results passes through all of the points. Doing this is illustrated in 3, with the equation that results from this type of interpolation.

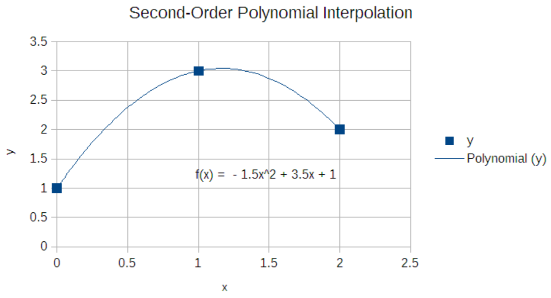

Figure 3. Second-Order Polynomial Interpolation

There are many applications for true interpolation; the best known (albeit invisible to the user) are the Bézier curves and cubic splines that are used in programs such as Adobe Photoshop®.

The second way is to do something that students are all too familiar with: cheat. Experimental data is subject to many variations in instrumentation and data collection to the point that, depending upon the situation, trying to correlate the data to some kind of simple expression neither warrants nor deserves the use of anything more than a linear expression. In this case we can use Equation 7, but with the caveat that that we're not looking for the unique solution of the equation but the best solution, i.e., one where the line we generate comes closest to the points without necessarily passing through any of them.

We start this process as we did before, by rewriting Equation 7 as

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} \tag{10}$$

First note: for more than three data points, we simply expand the number of rows in both the matrix and the y vector, both being the same. Beyond that, unlike Equation 4, the right hand side is nonzero. This is a residual; our goal is to come up with values of m and b that minimize that residual.

Vector Norms

But what does it mean to minimize a vector, which has more than one scalar? At this point we introduce the concept of a vector norm, which is best explained using illustrations. The idea of a vector norm is to express the "size" of a vector in a scalar form.

Before we start, it's worth noting that the values of r can be positive, negative, or zero. We're normally (sorry!) not interested in the sign of these vector entries; we will only consider them in a positive way. There is more than one way to define these norms; the methods we show here can be found in Gourdin and Boumahrat (1989).

One way would be to simply add up the absolute values of the entries of the norm. This is referred to as the 1-norm, given by the equation

$$||r||_1 = \sum_{i=1}^{n} |r_i| \tag{11}$$

Another way would be to minimize the entry with the largest absolute value. This is referred to as the infinity norm, or

$$||r||_\infty = max|r_i|,\ 1 \leq i \leq n \tag{12}$$

The last one is referred to as the 2-norm or Euclidean norm, given by the equation

$$||r||_2 = \sqrt{\sum_{i=1}^{n} r_i^2} \tag{13}$$

For two data points, the Euclidean norm is the length of the hypotenuse of a triangle with side lengths r_1 and r_2 (the Pythagorean theorem.) What this physically means for more than three dimensions is physically iffy, but the Euclidean norm is the most commonly used for a wide variety of reasons.

Minimizing the Residual

Now that we've presented vector norms, we're ready to do something with them. It should be obvious that our goal is to minimize the norm of the right hand side of Equation 10 using the definition of Equation 13. But how? One way is to employ a technique which students use frequently: guess. Actually this isn't as stupid as it sounds, because guessing is pretty much what drives most non-linear solution and optimization techniques. The idea is that our guessing scheme is reasonably methodical and grounded in the mathematical reality, which speeds up getting to the answer.

In this case, however, we can skip the guessing, because the solution can be found in "closed form," as shown by Wylie (1951). Consider Equation 10; the equation for any right-hand side residual is

$$r_i = mx_i + b - y_i \tag{14}$$

Squaring this per Equation 13 yields

$$r_i^2 = m^2 {x_i}^2 + 2\,mb\,x_i - 2m\,x_i y_i + b^2 - 2\,by_i + {y_i}^2 \tag{15}$$

What we want to do is to find the values of b and m so that Equation 13 is minimized. We can skip the square root operation; the result will be the same and it only complicates the differentiation. To accomplish this we take two partial differentials:

$$\begin{aligned} \frac{\partial \sum r^2}{\partial b} &= 0 \\ \frac{\partial \sum r^2}{\partial m} &= 0 \end{aligned} \tag{16}$$

Doing the summations and differentiations yields

$$\begin{aligned} nb \quad &+m\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i &= 0 \\ b \quad &+m\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i &= 0 \end{aligned} \tag{17}$$

Solving these two equations simultaneously will yield values for b and m. Fortunately this is built into spreadsheets so explicit calculation is not necessary.

One quantity you will see frequently with least squares fitting (linear regression) with spreadsheets is R^2, generally referred to as the coefficient of determination. This is generally computed by

$$R^2 = 1 - \frac{||r||_2^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{18}$$

where $\bar{y}$ is the mean of all the values of y. It is a handy way of measuring the quality of the fit of the data to the line you have computed. If our goal is to minimize the Euclidean norm, then R^2 will approach 1 as the fit improves.

Doing this for the example in Figures 2 and 3 is shown in Figure 4. Although for this data it is not the best way to do this (as evidenced by the low value of R^2,) as the number of data points increase (in statistics, the sample size) the value of this type of least-squares analysis–also known as linear regression–increases.

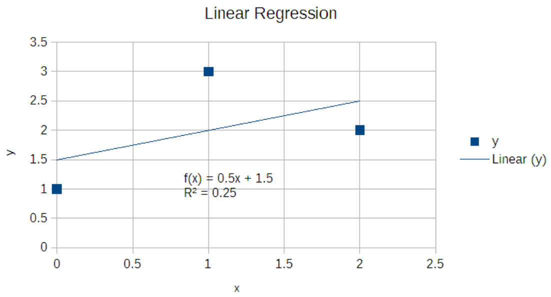

Figure 4. Linear Regression Analysis

Non-Linear Equations

Now that we've gone through linear regression, if we stop and think we realize that we cannot define every set of data we might have by a line, or at least not very well. So what can we do? From where we are there are two paths we can take.

The first is polynomial least squares fitting. If we look at Equation 9, and if we add a row to each matrix or vector in the equation, we will end up with the same "problem" as we had with Equation 7: too many equations for the unknowns we have. But that can be solved in a manner similar to the one we used to solve Equation 10. We can also add columns to the matrix (and rows to the left hand side vector) and, if we construct the matrix properly, increase the degree of the polynomial we wish to use to fit the data. Most spreadsheets have polynomial regression; the one thing you need to be careful about is attempting too high of a polynomial degree. One handy rule you can use is to look at the number of times the curve indicated by the data crosses the x-axis; that is the number of roots of the equation, which in turn indicates the degree of the polynomial to be used. (That rule assumes all of the roots are real, which isn't always the case.)

To use the second, we resort to another technique: cheating. Consider the exponential equation

$$\hat{y} = \hat{b}e^{mx} \tag{19}$$

Taking the natural logarithm of both sides yields

$$\ln\hat{y} = mx + \ln\hat{b} \tag{20}$$

This look suspiciously like Equation 1, and in fact can be considered that with the following change of variables:

$$\begin{aligned} y &= \ln\hat{y} \\ b &= \ln\hat{b} \end{aligned} \tag{21}$$

In the past this was plotted on semi-logarithmic paper, i.e., one scale ruled logarithmically and the other linearly. This can also be done with common logarithms; Equation 19 must be written thus

$$\hat{y} = \hat{b}10^{mx} \tag{22}$$

in which case

$$\log\hat{y} = mx + \log\hat{b} \tag{23}$$

and

$$\begin{aligned} y &= \log\hat{y} \\ b &= \log\hat{b} \end{aligned} \tag{24}$$

Another common extension of linear regression involves the power equation

$$\bar{y} = \bar{b}\bar{x}^{m} \tag{25}$$

which, taking the logarithm (natural or common) of both sides, yields

$$\log\bar{y} = \log\bar{b} + m\log\bar{x} \tag{26}$$

with change of variables

$$\begin{aligned} y &= \log\bar{y} \\ b &= \log\bar{b} \\ x &= \log\bar{x} \end{aligned} \tag{27}$$

to Equation 1.

Again, in the past this was plotted on logarithmic paper, with both scales ruled logarithmically. It is possible to get your spreadsheet to scale in this manner, but in reality this is not always the best way to present the data.

Example

In the past, most linear regression was done using a "best-fit" plotting, which in reality an intuitive way of minimizing the residual. Today a more customary way of doing this is to use a spreadsheet. This has advantages but some serious pitfalls which most students do not take into consideration.

As an example, consider the data in Table 1, which is taken from Thomas (1953):

Table 1
Sample Data for Test Case

x	y
0	1
1	3
2	2
3	4
4	5

Plotted results for this, with a number of different types of regression, is shown in Figure 5.

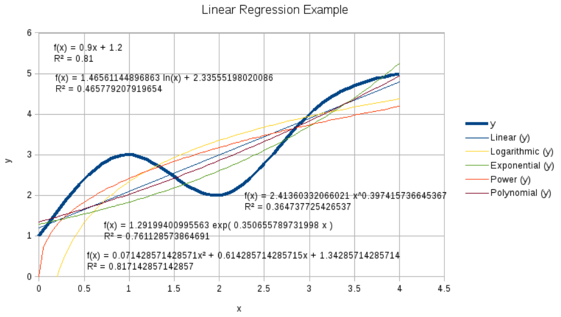

Figure 5. Results of Linear Regression Analysis

Some things to note:

1. Although the results are fairly "busy," each regression has its trend line, equation and coefficient of determination included in the graph. All of this information is essential in a proper linear regression plot.

2. The initial data curve is "smoothed." Generally the spreadsheet does this using interpolation. As a caution, when you first plot the data do so with straight lines connecting the dots and then apply smoothing. Smoothing generally works well but, as is the case with least squares solutions, interpolation can yield unrealistic results.

3. None of the curves chosen really do the data justice, although there is some apparent scatter in the data. The second-order polynomial regression is only marginally better than the linear one, and the others are in varying degrees of uninspiring. It's possible to improve on the results with techniques such as forcing the curve through one or more data points or weighting some data points more than others, but these can backfire as well.

4. The best advice for the student is simple: *look at the data and the relationship between that and the regressions/trend lines you are generating*. Don't just assume a regression equation is going to be the best; do multiple regressions and see which is best. The worst thing you can do is to simply run one trend line, throw it up and expect it to be fine.

Conclusion

Least squares curve fitting/regression analysis is a powerful tool for the engineer to reduce data and observe its trends. In the hands of the unobservant, it can be misused, as is the case with any other statistical method. Care should be used when reviewing the results and coming to conclusions.

References

Gelfand, I. M. (1961). *Lectures on linear algebra* (No. 9). New York, NY: Interscience Publishers, Inc.

Gourdin, A., & Boumahrat, M. (1989). *Méthodes Numériques Appliquées*. Paris, France: Téchnique et Documentation-Lavoisier.

Thomas, G. (1953). *Calculus and analytic geometry*. Reading, MA: Addison-Wesley Publishing Company, Inc.

Wylie, C. (1951). *Advanced Engineering Mathematics*. New York, NY: McGraw-Hill Book Company, Inc.