LEAST SQUARES LINE FITTING

Charles R. O'Neill

School of Mechanical and Aerospace Engineering

Oklahoma State University

Stillwater, OK 74078

Computer Project in MAE 3403

Computer Methods in Analysis and Design

April 1999

ABSTRACT

A formula for the least-squares best fit line was derived. A computer program was made to calculate the best fit line for a given input file. The computer program was used in two test cases to test the validity of the method to retrieve an input function from data points. The results show that the least-squares method does correctly find the input function but can have problems with certain types of data. Implications of these results are discussed.

Jan 2003: Thank you to Jeremy Connell for finding and reporting a mistake in regression coefficient b.

INTRODUCTION

Experimental data are subject to physical laws. The purpose of data collection is to establish a trend or relationship between input and output variables. Because of inaccurate equipment and imperfect operators, all measured data contains errors. Curve fitting attempts to find a best-fit function of the data. Two methods of curve fitting involve either fitting a curve through all data points or fitting a curve through an approximate function. Due to the inherent errors present, a best-fit function for establishing a trend should not attempt to pass through all of the data points. Instead, the function should attempt to find a good approximate curve which can capture the behavior of the experiment without explaining all of the points.

The objective is to use the least-squares method of curve fitting to generate a best fit line function. At least two data points must be provided. A computer program will use the derived formula to calculate a best-fit line of the input data. Two test cases will be used to establish the validity of the method and computer program used and to identify possible problems with the method.

THEORY

The least squares method of curve fitting generates a line which can approximate the true function of the data points. From the general equation of a line,

$$y = ax + b$$

any point (Y,x) can be represented as a distance from the line at the point's particular x value. This error distance of the ith point is

$$e_i = Y_i - (ax_i + b)$$

The least squares method attempts to minimize the square of the error distances of all the data points. So, data with N points has a total error S.

$$S = e_1^2 + e_2^2 + \ldots e_n^2 = \sum_{i=1}^{N} e_i^2$$

Substituting the equation of e_i yields,

$$S = \sum_{i=1}^{N} (Y - (ax_i + b))^2$$

In order to find the best fit, S is minimized. The partial derivatives of S with respect to the two variables a and b minimize S when set to zero.

$$\frac{\partial S}{\partial a} = 0 = \sum_{i}^{N} 2(Y_i - ax_i - b)(-x_i)$$

$$\frac{\partial S}{\partial b} = 0 = \sum_{i}^{N} 2(Y_i - ax_i - b)(-1)$$

solving for a and b and entering into matrix form yields,

$$a\,[\sum x_i^2\,] + b\,[\sum x_i\,] = \sum x_i Y_i$$

$$a\,[\sum x_i\,] + b\,[\,N\,] = \sum Y_i$$

Cramer's rule allows a set of 2 equations to be easily solved for 2 unknowns. If $ax + by = e$ and $cx + dy = f$ then,

$$x = \frac{ed - fb}{ad - bc}$$

$$y = \frac{af - ec}{ad - bc}$$

Using Cramer's rule for the solution of the 2x2 matrix yields,

$$a = \frac{N(\sum x_i Y_i) - (\sum x_i)(\sum Y_i)}{N(\sum x_i^2) - (\sum x_i)^2}$$

$$b = \frac{(\sum Y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i Y_i)}{N(\sum x_i^2) - (\sum x_i)^2}$$

Thus, the least square best fit line with a and b as calculated above is,

$$y = ax + b$$

The least square method of fitting a line can be used with other than linear data provided a transform of one or both of the axes is made. Properties of the log and ln function are useful. For example, data of the form $y = ax^b$ can be transformed as $log(y) = log(ax^b) = b \cdot log(ax)$. Chapra (1998) gives further examples of these transforms.

METHOD OF CALCULATION

A computer program written in FORTRAN 77 and given in Appendix A was created to output the parameters of a least squares best fit line of n data points given in an external file. The program first initializes the variables to be used: number of points, sum of x, sum of y, sum of $x \cdot y$ and sum of x^2. All summation variables are implicitly declared as floating point single precision and the index variable is declared implicitly as an integer. All variables are initially assigned a value of 0. Next, a loop is used to read a single line of data consisting of an x and y value. Each iteration of the loop calculates, prints and stores the current values of the variables. When the end of the data file is reached, the program exits the loop and closes the data file. Next using the formulas for a and b derived in the theory section, the program calculates the slope and y intersect of the best fit line. The best fit line is printed as a slope and y intercept. Finally, the program terminates.

RESULTS AND DISCUSSION

Calculations have been performed as described above on two test cases. Case 1 is the fitting of line to a linear equation $y(x) = x + 1$. Case 2 involves fitting a line to the exponential equation $y(x) = e^x$.

For Case 1, the x values ranged from 0 to 5 in steps of 1. Using the definition of $y(x) = x + 1$, the values for x and y were entered into the fl.dat data file (APPENDIX B). The program was run and output the equation $y(x) = 1.00000x + 1.0000$. This is the function used to create the data file.

For Case 2, the x values ranged from 0 to 5 in steps of 1. From Chapra(1998), plotting log(y) versus x will result in a line for the function $y(x) = ae^{bx}$. The values of x were entered into the data file (APPENDIX B). Using the definition of $y(x) = e^x$, the values for y were used to compute $log(y)$ and these values were also entered into the data file. The program was run and output the equation $y(x) = 0.43430x + -0.00004$. Since the y axis was scaled according to the function $y_{new} = log(y(x))$, the slope equals $b * log(e)$ and the y intersect equals $log(a)$. Transforming back, $y(x) = 10^{-.00004} \cdot e^{(\frac{0.43430}{log(e)}) \cdot x} = 0.9999e^{1.000013 * x}$. This is close to the correct function $y(x) = e^x$.

These two cases show that the least sum of the squares method can retrieve the original functions. This success indicates that other data can be fitted by this method with reasonable assurance of a good linear fit. In the first case, the original function was exactly found by the line fitting routine, which indicates that no round off errors were encountered. The second test case did have round off error and truncation due to entering irrational values into the data file. However, the original function was easily deduced from the returned values.

The least squares method does assume that some transform of the data will result in a near linear data set. With data that does not conform to this assumption, this method will fail to find the true relationship of the data points. For example, any nonlinear curve such as a parabola or sine curve can not be represented with this linear best fit method.

CONCLUSION

The least-squares method does find a trend function. The accuracy of this function depends upon the input data. Certain types of nonlinear trends can be transformed to yield linear approximations making it adaptable to types of data caused by physical processes. Thus, the least-squares best fit is a curve fitting method useful for the identification and determination of approximation functions for experimental data sets.

REFERENCES

Chapra, Steven C., Raymond P. Canale (1998)

Numerical methods for engineers: with programming and software applications,

WCB/McGraw-Hill.

APPENDIX A

Computer Program

APPENDIX B

Sample Input and Output