The following is from my book and is the introduction to the chapter entitled: The Method of Least Squares

2.1 Introduction

The first published treatment of the method of least squares was included in an appendix to Adrien Marie Legendre's book *Nouvelles methods pour la determination des orbites des cometes.* The 9 page appendix was entitled *Sur la methode des moindres quarres.* The book and appendix was published in 1805 and included only 80 pages but gained a 55 page supplement in 1806 and a second 80 page supplement in 1820 [ST86]. It has been said that the method of least squares was to statistics what calculus had been to mathematics. The method became a standard tool in astronomy and geodesy throughout Europe within a decade of its publication. The method was also the cause of a dispute between two giants of the scientific world of the 19th century: Legendre and Gauss. Gauss in 1809 in his famous *Theoria Motus* claimed that he had been using the method since 1795. That book was first translated into English in 1857 under the authority of the United States Navy by the Nautical Almanac and Smithsonian Institute [GA57]. Another interesting aspect of the method is that it was rediscovered in a slightly different form by Sir Francis Galton. In 1885 Galton introduced the concept of regression in his work on heredity. But as Stigler says: "Is there more than one way a sum of squared deviations can be made small?" Even though the method of least squares was discovered about 200 years ago, it is still "the most widely used nontrivial technique of modern statistics" [ST86].

The least squares method is discussed in many books but the treatment is usually limited to linear least squares problems. In particular, the emphasis is often on fitting straight lines or polynomials to data. The multiple linear regression problem (described below) is also discussed extensively (e.g., [FR92, WA93]). Treatment of the general nonlinear least squares problem is included in a much smaller number of books. One of the earliest books on this subject was written by W. E. Deming and published in the pre-computer era in 1943 [DE43]. An early paper by R. Moore and R. Zeigler discussing one of the first general purpose computer programs for solving nonlinear least squares problems was published in 1960 [MO60]. The program described in the paper was developed at the Los Alamos Laboratories in New Mexico. Since then general least squares has been covered in varying degrees and with varying emphases by a number of authors (e.g., DR66, WO67, BA74, GA94, VE02).

For most quantitative experiments, the method of least squares is the "best" analytical technique for extracting information from a set of data. The method is best in the sense that the parameters determined by the least squares analysis are normally distributed about the true parameters with the least possible standard deviations. This statement is based upon the assumption that the uncertainties (i.e., errors) in the data are uncorrelated and normally distributed. A complete and detailed proof of this statement is included in a very old text by Mansfield Merriman (*The Elements of the Method of Least Squares*, J.Wiley and Sons, 1877). For most quantitative experiments this is usually true or is a reasonable approximation. When the curve being fitted to the data is a straight line, the term **linear regression** is often used. For the more general case in which a plane based upon several independent variables is used instead of a simple straight line, the term **multiple linear regression** is often used [FR92, WA93]. Prior to the advent of the digital computer, curve fitting was usually limited to lines and planes. For the simplest problem (i.e., a straight line), the assumed relationship between the dependent variable $\boldsymbol{y}$ and the independent variable $\boldsymbol{x}$ is:

$$\boldsymbol{y} = \boldsymbol{a}_1 + \boldsymbol{a}_2 \boldsymbol{x} \qquad (2.1.1)$$

For the case of more than one independent variable (multiple linear regression), the assumed relationship is:

$$y = a_1 x_1 + a_2 x_2 + ... + a_m x_m + a_{m+1} \qquad (2.1.2)$$

For this more general case each data point includes $m+1$ values: $y_i, x_{1i}, x_{2i}, \ldots, x_{mi}$.

The least squares solutions for problems in which Equations 2.1.1 and 2.1.2 are valid fall within the much broader class of **linear least squares** problems. In general, all linear least squares problem are based upon an equation of the following form:

$$y = f(\mathbf{X}) = \sum_{k=1}^{k=p} a_k g_k(\mathbf{X}) = \sum_{k=1}^{k=p} a_k g_k(x_1, x_2 .... x_m) \qquad (2.1.3)$$

In other words, y is a function of **X** (a vector with m terms). Any equation in which the p unknown parameters (i.e., the a_k's) are coefficients of functions of only the independent variables (i.e., the m terms of the vector **X**) can be treated as a linear problem. For example in the following equation, the values of a_1, a_2, and a_3 can be determined using linear least squares:

$$y = a_1 \, sin(x^{3/2}) / cosh(x-1) + a_2 (cos(x^{5/2} - x^3))^{3/2}$$
$$+ a_3 / ln(1/x + 1/x^2)$$

This equation is nonlinear with respect to x but the equation is linear with respect to the a_k's. In this example, the **X** vector contains only one term so we use the notation x rather than x_1. The following example is a linear equation in which **X** is a vector containing 2 terms:

$$y = a_1 \, sin(x_1^{3/2}) / cosh(x_2 - 1) + a_2 (cos(x_1^{5/2} - x_2^3))^{3/2}$$
$$+ a_3 / ln(1/x_1 + 1/x_1^2)$$

The following example is a nonlinear function:

$$y = a_1 \, sin(x_1^{3/2}) / cosh(x_2 - 1) + a_2 (cos(x_1^{5/2} - x_2^3))^{3/2}$$
$$+ a_3 / ln(1/x_1 + a_4 / x_1^2)$$

The fact that a_4 is embedded within the last term makes this function incompatible with Equation 2.1.3 and therefore it is nonlinear with respect to the a_k's.

For both linear and nonlinear least squares, a set of p equations and p unknowns is developed. If Equation 2.1.3 is applicable then this set of equations is linear and can be solved directly. However, for nonlinear equations, the p equations require estimates of the a_k's and therefore iterations are required to achieve a solution. For each iteration, the a_k's are updated, the terms in the p equations are recomputed and the process is continued until some convergence criterion is achieved. Unfortunately, achieving convergence is not a simple matter for some nonlinear problems.

For some problems our only interest is to compute $y = f(\mathbf{X})$ and perhaps some measure of the uncertainty associated with these values (e.g., σ_f) for various values of **X**. This is what is often called the **prediction** problem. We use measured or computed values of x and y to determine the parameters of the equation (i.e., the a_k's) and then apply the equation to calculate values of y for any value of x. For cases where there are several (let us say m) independent variables, the resulting equation allows us to predict y for any combination of x_1, x_2, .. x_m. The least squares formulation developed in this chapter also includes the mthodology for prediction problems.