PROBLEMS IN LEAST SQUARES

by

EDWARD LEON ARMBRUST

B. S., Kansas State University, 1963

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1965

Approved by:

Leslie F. Marcus

Major Professor

TABLE OF CONTENTS

INTRODUCTION

As a result of the availability of electronic computers many more relationships between dependent and independent variables are being examined. A specific problem brought to the attention of the author was the estimation of the coefficients θ_1, θ_2, θ_3, and θ_4 of the equation

$$y = \theta_1 e^{-\theta_2 x} + \theta_3 e^{-\theta_4 x} \quad -\infty < \theta_i < \infty \quad .$$
$$i = 1, \ldots, 4$$

This equation has been used to describe the amount, y, of a drug in the blood of a human being after time x has elapsed (Wagner (1964)).

With this problem in mind the development of least squares is discussed. Several methods for obtaining least squares estimates, with emphasis on the non-linear case are summarized. Some discussion concerning the statistical properties of least squares estimates is also presented in this paper.

HISTORICAL DEVELOPMENT

The theory of least squares has developed over the past 243 years and the method of least squares as a method of estimating coefficients has been in use for over 160 years. Merriman (1877) lists 408 titles of papers between 1722 and 1876 which relate to least squares. Todhunter (1865) is another work which considers the early development of least squares.

The method of least squares as a method of estimating coefficients was first stated by Legendre (1806). This formulation is described below.

Observations which are assumed to be functions of other variables and coefficients frequently can be assumed to be related by a set of linear equations:

$$\theta_1 x_{11} + \ldots + \theta_k x_{1k} = y_1$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$\theta_1 x_{s1} + \ldots + \theta_k x_{sk} = y_s.$$

The θ_i, $i = 1, \ldots, k$, are the unknown coefficients to be estimated. The x_{ij}, $i = 1, \ldots, s$, $j = 1, \ldots, k$, are known quantities, and the y_i, $i = 1, \ldots, s$, are observations which are subject to error and estimate the true value of the function Y_i, say. Legendre called this set of linear equations, "the equations of condition." Usually s k and the equations are independent.

It is desired to find values of the coefficients $\theta_1, \ldots, \theta_k$ such that these equations are satisfied as nearly as possible. To "satisfy as nearly as possible" means to minimize the errors, or possibly a function of the errors, $e_1, \ldots, e_s$, where

$$e_i = \theta_1 x_{i1} + \ldots + \theta_k x_{ik} - Y_i,\ i = 1, \ldots, s.$$

That is, the differences between the observed values and the functional values or some function of these differences is to be minimized.

Legendre proposed that $e_1^2 + \ldots + e_s^2$, the sum of the squared errors, be minimized. A solution $\theta_1, \ldots, \theta_k$ is obtained by setting

$$\frac{\Sigma e_i^2}{\theta_1} = \ldots = \frac{\Sigma e_i^2}{\theta_k} = 0.$$

These resulting equations Legendre denoted as the normal equations.

JUSTIFICATION OF LEAST SQUARES

Gauss first attempted to justify the use of the method of least squares by demonstrating that the errors of observation for a quantity to be estimated are distributed normally and that, if this is true, the most probable estimate is the estimate obtained by the method of least squares. This justification by Gauss is somewhat intuitive. It is presented, however, because it is one of the first attempts to justify the use of the method of least squares.

This early justification assumes the postulate of the arithmetic mean (Whittaker and Robinson (1924)): "When any number of equally good direct observations M, M', M'', . . . of an unknown magnitude x are given, the most probable value for x is their arithmetic mean." Gauss deduced this postulate from four elementary axioms.

> Axiom I - The differences between the most probable value and the individual measures do not depend on the position of the null point from which they are reckoned.
>
> Axiom II - The ratio of the most probable value to any individual measure does not depend on the unit in terms of which measures are reckoned.
>
> Axiom III - The most probable value is independent of the order in which the measurements are made, and so is a symmetric function of the measures.
>
> Axiom IV - The most probable value, regarded as a function of the individual measures, has one-valued and continuous first derivatives with respect to them.

The first two axioms propose the most probable value to be invariant under a linear transformation. The third axiom means the observations are a random sample. The fourth axiom explains itself.

Suppose on some small interval between Δ and $\Delta + d\Delta$ the probability of error is $\phi(\Delta)d\Delta$. Then, $\phi(\Delta)$ is the relative frequency of error. Let denote the least possible error, for any one measurement. The case is being considered where $d\Delta = \epsilon$ and the probability of error associated with any Δ is $\emptyset(\Delta)\epsilon$.

If M, M', M", . . . ,$M^{(s)}$ are s observations which are used to calculate a quantity x whose true value is p, the errors of observation are

$$\Delta = M - p$$

$$\Delta' = M' - p$$

$$\Delta'' = M'' - p.$$

The probability of a certain sequence M, M', M", . . . ,$M^{(s)}$ occurring, $\epsilon^{s}\emptyset(M - p)\emptyset(M' - p) \ldots ,\emptyset(M^{(s)} - p)$. To illustrate, consider the problem of estimating the length of a desk with true length p. The accuracy of the yardstick being used is ¼". Therefore ϵ = ¼" if no human error. The s measurements M, M', M", . . . are measurements of a quantity x, which estimates p, such that E(x) = p. The probability of a certain sequence M, M', M", . . . is $\frac{1}{4}^{s}\emptyset(M - p) \cdot \emptyset(M' - p) \ldots$ where $\emptyset(M - p)\frac{1}{4}$ is the probability of an error of size M - p.

An expression for the probability of the true value of x lying between p and p + dp can be obtained by considering Bayes' Theorem,

$$P(A_i | B) = \frac{P(B|A_i)P(A_i)}{\Sigma P(B|A_j)P(A_j)} .$$

For this problem A_i is the event x for the given sequence being between p and p + dp. The event B can be considered as the particular sequence of observations. In words, the probability that the true x is between p and p + dp for a given sequence of observations is equal to the probability of a certain sequence given that x is between p and p + dp multiplied by the probability x is between p and p + dp, all divided by the sum of the above product for all possible values of x.

The $P(B|A_i)$ is $\epsilon^s \phi(M - p)\phi(M' - p) \ldots$ and $P(A_i)$ is dp. The right hand side of the expression above could be written in the notation previously introduced as

$$\frac{\epsilon^s \phi(M - p)\phi(M' - p) \ldots dp}{\epsilon^s \int \phi(M - p)\phi(M' - p) \ldots dp}.$$

To maximize the probability that x is between p and p + dp the value of x which maximized $\phi(M - x)\phi(M' - x) \ldots$ should be adopted. The function $\phi(M - x)\phi(M' - x) \ldots$ being positive, the logarithm of the function will occur at the same point as for the first. Therefore that value of x for which

$$\Sigma d \log \phi(M - x)/dx = 0$$

is desired. But, assuming the postulate of the arithmetic mean, the most probable value for x is the arithmetic mean. Then $x = (1/s)(M + M' + \ldots)$, and $\Sigma(M - x) = 0$. Therefore

$$\Sigma d/dx \log \phi(M - x) = \Sigma(M - x) = 0$$

and $\quad d/dx \log \phi(M - x) = c(M - x)$ for c a constant. Performing the integration, $\log \phi(M - x) = -c(M - x)^2/2 + b$ and, therefore, $\phi(M - x) = Ae^{-\frac{1}{2}c(M - x)}$, where A a constant.

Since the sum of the probabilities of all the errors is unity, A can be found as $\sqrt{c/2\pi}$. Writing $h = \sqrt{c/2}$ one has

$$\phi(\Delta) = \frac{he^{-h^2\Delta^2}}{\sqrt{\pi}}$$

which means the distribution of measurements about the true value is a normal frequency distribution. The probability of a certain set of errors, therefore, is

$$\frac{\epsilon_1 h_1 e^{-h_1^2(M - p)^2}}{\sqrt{\pi}} \quad . \; . \; . \quad \frac{\epsilon_s h_s e^{-h_s^2(M'\cdots' - p)^2}}{\sqrt{\pi}} \quad .$$

The most probable value of x is that which maximizes the above expression if p is replaced by x, or it is that value of x which minimizes $h_1^2(M - x)^2 + \ldots + h_s^2(M'\cdots' - x)^2$, a sum of squared deviations. Then

$$x = \frac{(h_1^2 M + \ldots + h_s^2 M'\cdots')}{(h_1^2 + \ldots + h_s^2)} \; .$$

Suppose $w_i = h_i^2$, $i = 1, \ldots, s$; then,

$$x = \frac{(w_1 M + \ldots + w_s M'\cdots')}{(w_1 + \ldots + w_s}$$

is the least squares solution.

The quantities $w_1, \ldots, w_s$ are called the weights of the observations, M, M', M'', . . . and frequently the sum of the w_i is set equal to W.

Gauss again justified the use of the method of least squares with a less heuristic argument in 1821. This justification has been summarized by Plackett (1949), who emphasizes that this justification is very different from the justification of Laplace. Plackett used matrix notation to summarize these results as defined below.

Let $\theta(s \times 1)$ be a vector of unknown coefficients, y $(n \times 1)$ a vector of observations, e $(n \times 1)$ a vector of errors and X $(n \times s)$ a matrix of known quantities. A set of equations relating these quantities can be expressed as $X\theta = y + e$. Further, assume that W $(n \times n)$ is a diagonal matrix whose elements are the reciprocals of the variances of the e vector.

The method of least squares leads to estimates of the coefficients which satisfy the relationship $X'WX\theta^* = X'Wy$. That is $\theta^* = (X'WX)^{-1}X'Wy$, where θ^* is the vector which estimates θ.

Gauss's 1821 justification was written in Latin and a French translation was published by Bertrand in 1855. The following represents this justification as summarized by Plackett (1949).

From the above statement $\theta^* = (X'WX)^{-1}X'Wy$ let $W = I$, the identity matrix. If $\theta^* = By$ is unbiased and $E(e) = 0$ for all θ then $E(\theta^*) = E(By) = BE(y) = BE(X\theta + e) = BXE(\theta) = BX\theta = \theta$. Therefore $BX = I$. Let $S = X'X$. It follows that $S^{-1} = BXS^{-1}$ and $BB' = (S^{-1}X')(S^{-1}X')' + (B - S^{-1}X')(B - S^{-1}X')'$, and the diagonal elements of BB' are least when $B = S^{-1}X'$, which is the least squares solution.

This important theorem was discussed by Markoff in 1912 and the theorem was extended by Aitken in 1934 to consider not only the diagonal elements as in W, but the entire variance-covariance matrix.

Laplace in 1811 established the method of least squares in a different manner. Plackett (1949) summarized this and the work of Laplace from 1812-1820 concerning least squares in the following way.

Among all s x n matrices F leading to estimates of the form $FX\theta^* = Fy$, the expected values of the elements $|\theta - \theta^*|$ are minimized as $n \to \infty$ when $F = @X'W$, @ being an arbitrary multiplier.

In other words the solution of the equation $@X'WX\theta^* = @X'Wy$ for the estimate θ^* will be an unbiased estimate of θ. As noted earlier the above estimate for θ is the least squares estimate.

PROPERTIES OF ESTIMATES IN THE LINEAR MODEL

If one considers the general linear hypothesis a model of full rank, that is, the model $X\theta - y = e$, where the matrix X is of full rank, then the method of least squares gives rise, with two additional assumptions, to estimates whose statistical properties are given in the Gauss-Markoff theorem. Because of the importance of this theorem, both the theorem and a proof are given below modified after Graybill (1961).

If the general-linear-hypothesis model of full rank $y = X\theta + e$ is such that the following two conditions on the random vector e are met:

1) $E(e) = 0$

2) $E(ee') = \sigma^2 I$,

the best (minimum-variance) linear (linear function of the elements of y) unbiased estimate of θ is given by least squares.

The proof is as follows. Let B be any s x n constant matrix and let $\theta^* = By$. Suppose $B = S^{-1}X' + A$. $S^{-1}X'$ is known, but A must be found in order to specify B. For unbiasedness, $E(\theta^*) = \theta$. Therefore $E(\theta^*) = E[(S^{-1}X' + A)y] = E(S^{-1}X'y) + E(Ay) = S^{-1}X'X\theta + E(Ay) = \theta + AX\theta$. Thus to be unbiased, $AX = 0$.

For the property of minimum-variance the matrix A must be found that minimizes the variance of θ^* subject to the restriction $AX = 0$.

Consider the covariance of θ^*,

$$\mathrm{cov}(\theta^*) = E[(\theta^* - \theta)(\theta^* - \theta)']$$

$$= E\,[(S^{-1}X' + A)y - \theta][(S^{-1}X' + A)y - \theta]'$$

Substituting $X\theta + e$ for y and recalling that $AX = 0$,

$$\text{cov}(\theta^*) = E(S^{-1}X'ee'XS^{-1}) + Aee'A' + S^{-1}X'ee'A' + Bee'XS^{-1}$$

$$= \sigma^2(S^{-1} + AA')$$

Let $AA' = G = (g_{ij})$. To minimize the var (θ^*), the diagonal elements of $\sigma^2(S^{-1} + AA')$ must be minimized. Since σ^2 and S^{-1} are constants, the diagonal elements of G must be minimized. But G is a positive semidefinite matrix and hence $g_{ii} \geqslant 0$. The diagonal elements obtain a minimum when $g_{ii} = 0$. This implies that all the elements of A = 0 and A = 0. Therefore $B = S^{-1}X$ and $\theta^* = S^{-1}Xy$ are the least squares estimate for θ.

Aitken (1948) generalized this theorem to consider the situation where $E(ee') = \sigma^2 V$, V being in a completely general matrix. A discussion of this topic can be found in Kendall and Stuart (1961).

Plackett (1950) considers the case when X is of less than full rank. The minimum-variance, unbiased solution for the estimate of θ may be obtained by modifying the method of least squares. Plackett's 1950 paper also includes a procedure, which requires a minimum of calculation, for estimating the coefficients and sums of squares when additional observations occur.

If the additional assumption that the vector of errors is normally distributed can be made, the least squares estimates and the estimates obtained by the method of maximum likelihood are identical for the general linear hypothesis of full rank. These estimates have the statistical properties of being consistent, efficient, unbiased, sufficient, complete, and have minimum variance (Graybill (1961)).

FIRST ORDER TAYLOR APPROXIMATION OF NON-LINEAR FUNCTIONS

Suppose it is required to find estimates of the coefficients θ_1 and θ_2 from a set of equations $f_i(\theta_1, \theta_2) = y_i \quad i = 1, \ldots, s$ where $f_i(\theta_1, \theta_2)$ are known functions not necessarily linear in θ_1 and θ_2 and $y_1, \ldots, y_s$ are observations subject to error.

Whittaker and Robinson (1924) first suggested an approximation by a first order Taylor series,

$$f_i(\theta_1^*, \theta_2^*) = f_i(\theta_{10}^* + \Delta_1, \theta_{20}^* + \Delta_2)$$

$$= f_i(\theta_{10}^*, \theta_{20}^*) + \frac{\partial f_i(\theta_{10}^*, \theta_{20}^*)\Delta_1}{\partial \theta_{10}^*} + \frac{\partial f_i(\theta_{10}^*, \theta_{20}^*)\Delta_2}{\partial \theta_{20}^*}$$

where θ_{10}^* and θ_{20}^* are initial approximations for θ_1 and θ_2 and $\Delta_1 = \theta_1^* - \theta_{10}^*$, $\Delta_2 = \theta_2^* - \theta_{20}^*$. The resulting series of equations can be solved for values of Δ_1 and Δ_2 by the method of least squares. The initial approximations can be corrected by these values and used as new approximations for the coefficients. The equations then can be solved for new values of Δ_1 and Δ_2. The iteration continues until Δ_1 and Δ_2 equal zero. Convergence to solutions is not guaranteed. It has been assumed that the first order approximation is an appropriate approximation for the original equation.

ESTIMATION UTILIZING THE JACOBIAN

Wynn (1962) considers a special non-linear function and gives an organized treatment of this function applying the first order Taylor approximation. If the coefficients to be estimated occur in the relationship

$$Y_i = \theta' + \Sigma_j \theta_{1j} e^{-\theta_{2j} x_i}$$

and one lets

$$v_i = y_i - \theta' - \Sigma_j \theta_{1j} e^{-\theta_{2j} x_i}$$

and $S = \Sigma_i v_i^2$, then S is to be a minimum.

By taking partial derivatives of S with respect to the coefficients one can derive the 2h + 1 equations

$$\emptyset_0 = \frac{\partial S}{\partial \theta'} = -2 \Sigma_i v_i$$

$$\emptyset_{2j-1} = \frac{\partial S}{\partial \theta_{1j}} = -2 \Sigma_i e^{-\theta_{2j} x_i} v_i \quad j = 1, \ldots, h$$

$$\emptyset_{2j} = \frac{\partial S}{\partial \theta_{2j}} = 2\theta_{1j} \Sigma_i e^{-\theta_{2j} x_i} v_i x_i \quad j = 1, \ldots, h$$

Let the sequence of estimates be related by the equations

$$\theta'^{(r+1)} = \theta'^{(r)} + \Delta\theta'^{(r)}, \quad \theta_{1j}^{(r+1)} = \theta_{1j}^{(r)} + \Delta\theta_{1j}^{(r)}$$

$\theta_{2j}^{(r+1)} = \theta_{2j}^{(r)} + \Delta\theta_{2j}^{(r)}$ and let the notation

$$\emptyset_u^{(r)} = \emptyset_u \begin{bmatrix} \theta'^{(r)} \\ \cdot \\ \cdot \\ \cdot \\ \theta_{2h}^{(r)} \end{bmatrix} \qquad u = 0, 1, \ldots, 2h$$

be adopted.

$\emptyset_u^{(r)}$ is one of the 2h + 1 partial derivatives above evaluated at the r^{th} estimate of the coefficients. The notation for $\emptyset_u^{(r)}$ is adopted to convey this meaning.

Neglecting second and higher order terms one obtains

$$(\emptyset_0^{(r+1)}, \emptyset_1^{(r+1)}, \ldots, \emptyset_{2h}^{(r+1)}) =$$

$$(\emptyset_0^{(r)}, \emptyset_1^{(r)}, \ldots, \emptyset_{2h}^{(r)}) + J^{(r)}(\Delta\theta'^{(r)}, \ldots, \Delta\theta_{2h}^{(r)}).$$

The dimensions of this operation are

$$[(2h+1)\cdot(1)] = [(2h+1)\cdot(1)] + [(2h+1)\cdot(2h+1)][(2h+1)\cdot(1)].$$

The vectors in the equation are column vectors and J is the Jacobian which can be expressed as

$$\begin{vmatrix} \frac{\partial\emptyset_0}{\partial\theta'} & \frac{\partial\emptyset_0}{\partial\theta_{11}} & \cdots & \frac{\partial\emptyset_0}{\partial\theta_{2h}} \\ \frac{\partial\emptyset_1}{\partial\theta'} & \frac{\partial\emptyset_1}{\partial\theta_{11}} & \cdots & \\ \vdots & & & \vdots \\ \frac{\partial\emptyset_{2h}}{\partial\theta'} & & \cdots & \frac{\partial\emptyset_{2h}}{\partial\theta_{2h}} \end{vmatrix}$$

and $J^{(r)}$ implies the r^{th} set of vectors in the Jacobian.

If the set $\emptyset_u^{(r)}$ is computed by the derived equations and the values $\Delta\theta'$, $\Delta\theta_{1j}^{(r)}$, $\Delta\theta_{2j}^{(r)}$ are required which make $\emptyset_u^{(r+1)}$ equal to zero, then from the above equation

$$(\Delta\theta'^{(r)}, \Delta\theta_{1i}^{(r)}, \ldots, \Delta\theta_{2h}^{(r)}) \doteq (-J^{(r)})^{-1}(\emptyset_0^{(r)}, \ldots, \emptyset_{2h}^{(r)})$$

Improved values of θ', θ_{1j}, θ_{2j} are obtained by iteration.

An alternate procedure which demands less computation is to compute $J^{(0)}$ and invert this matrix. The iteration then proceeds in conjunction with

$$(\Delta\theta_1^{(r)}, \ldots, \Delta\theta_{2h}^{(r)}) = (J^{(0)})^{-1}(\phi_0^{(r)}, \ldots, \phi_{2h}^{(r)}).$$

Although less computation is necessary as far as inverting the Jacobian with each iteration, this latter technique will converge much more slowly than if the Jacobian is calculated and inverted each time.

WEIGHTING AND HIGHER ORDER TAYLOR APPROXIMATIONS

It has been observed that in some problems the first order Taylor approximation fails to improve the initial solution (Wilson and Puffer (1933)). Some alternatives to the first order approximation are considered.

Consider the non-linear equation $Y_i = \theta_1 e^{\theta_2 x}$, where θ_1 and θ_2 are the coefficients to be estimated. If the method of least squares is used, the $\Sigma(Y_i - y_i)^2$ is to be minimized where the y_i are observations which estimate Y_i. The problem is that θ_1 enters linearly but θ_2 enters non-linearly. If one utilizes logarithms and minimizes $\Sigma(\theta_2 x_i + \log \theta_1 - \log y_i)^2$, θ_2 enters linearly and θ_1 enters non-linearly. Although the adoption of logarithms for this type of problem is not uncommon, it does not necessarily provide for a better solution.

If one can assume that the error of observation is relatively small the difference $y_i - Y_i$ could be considered as a differential of Y. Since $d\log Y = dY/Y$, $\Sigma(y_i - Y_i)^2$ is approximated by either $\Sigma(\log y_i - \log Y_i)^2 y_i^2$ or $\Sigma(\log y_i - \log Y_i)^2 Y_i^2$ if y_i and Y_i are nearly equal. Of the two forms

Wilson and Puffer (1933) point out that the second form is quadratic in the unknown coefficients and also leads to linear normal equations.

A third method of solving non-linear equations is to expand the function by a Taylor series of order greater than one.

$$S = \Sigma(y_i - Y_i - \frac{\partial Y_i}{\partial \theta_1}\Delta\theta_1 + \frac{\partial Y_i}{\partial \theta_2}\Delta\theta_2 - \frac{1}{2}\frac{\partial^2 Y_i}{\partial \theta_2}\Delta\theta_1^2 \ldots)^2$$

"DAMPING" THE NORMAL EQUATIONS

Levenberg (1944) modified the method of estimating coefficients utilizing the first order Taylor approximation.

Let $g_i(\theta) = f_i(\theta^*) - f_i(\theta)$. The least squares criterion requires that $S(\theta) = \Sigma g_i^2(\theta)$ be minimized. The function

$$g_i(\theta) \doteq G_i(\theta) = g_i({}_0\theta) + \frac{\partial g_i}{\partial \theta_1}\Delta_1 + \frac{\partial g_i}{\partial \theta_2}\Delta_2 + \ldots .$$

Using this approximating function $G_i(\theta)$, it is now required to minimize $S(\theta) = \Sigma G_i^2(\theta)$, where ${}_0\theta$ is an initial approximation vector.

By taking partial derivatives of $\Sigma G_i^2(\theta)$ with respect to the various coefficients and setting the resulting expressions equal to zero the following equations may be obtained.

$$\Sigma \frac{(\partial G_i)^2}{\partial \theta_1}\Delta_1 + \Sigma \frac{(\partial G_i \cdot \partial G_i)}{\partial \theta_1 \; \partial \theta_2}\Delta_2 + \ldots + \Sigma \frac{\partial G_i}{\partial \theta_1} G_i = 0$$

$$\Sigma \frac{(\partial G_i \cdot \partial G_i)}{\partial \theta_1 \; \partial \theta_2}\Delta_1 + \Sigma \frac{(\partial G_i)^2}{\partial \theta_2}\Delta_2 + \ldots + \Sigma \frac{\partial G_i}{\partial \theta_2} G_i = 0$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

Levenberg points out that the Δ_i may be so large that successive approximations for Θ^* will yield a $\Sigma G_i^2(\Theta^*)$ which is larger than the initial solution.

Let $\bar{S}(\Theta^*) = WS(\Theta^*) + w_1\Delta_1^2 + w_2\Delta_2^2 + \ldots$, where $w_1, w_2, \ldots$ are the weights of the Δ_i and W expresses the relative importance of the residuals and the Δ_i. Suppose the function $\bar{S}(\Theta^*)$ takes its minimum at $\Theta+ = (\Theta_{1+}, \Theta_{2+}, \ldots)$ and set $Q({}_0\Theta) = w_1\Delta_1^2 + w_2\Delta_2^2 + \ldots$. Under the assumption $S(\Theta)$ is not stationary at $\Theta = \Theta_0$, Levenberg obtained the following inequalities. $S(\Theta+) < S({}_0\Theta)$ and $Q(\Theta+) < Q(\Theta_\infty)$, where $Q(\Theta_\infty)$ denotes the standard least squares solution.

The normal equations which result from minimizing the expression for $\bar{S}(\Theta^*)$ above become

$$\Sigma \frac{(\partial G_i)^2}{\partial \Theta_1} + w_1W^{-1})\,\Delta_1 + \Sigma \frac{(\partial G_i\,\partial G_i)}{\partial \Theta_1\,\partial \Theta_2}\,\Delta_2 + \ldots + \Sigma \frac{\partial G_i}{\partial \Theta_1} G_i = 0$$

$$\Sigma \frac{(\partial G_i\,\partial G_i)}{\partial \Theta_1\,\partial \Theta_2}\,\Delta_1 + (\Sigma \frac{(\partial G_i)^2}{\partial \Theta_2} + w_2W^{-1})\,\Delta_2 + \ldots + \Sigma \frac{\partial G_i}{\partial \Theta_2} G_i = 0$$

etc.

The best value of W may theoretically be determined by solving $dS(\Theta^*)/dW = 0$. $S(\Theta^*)$ may be approximated by a Taylor expansion gives $S(\Theta^*) \doteq S({}_0\Theta) + W(dS/dW)\Big|_{W = 0}$. And setting this expression equal to zero on the assumption that ${}_0\Theta$ was chosen so that the decreased value $S(\Theta^*)$ is small, the result

$$W = \frac{S({}_0\Theta)}{(dS/dW)_{W = 0}} = \frac{\frac{1}{2}S({}_0\Theta)}{((\Sigma \frac{\partial G_i}{\partial \Theta_i} f_i)^2 w_1^{-1} + (\Sigma \frac{\partial G_i}{\partial \Theta_i} f_i)^2 w_2^{-1} + \ldots .)}$$

was obtained by Levenberg. If necessary, the value for W may be improved by calculating $S({}_0\theta)$ for several values of ${}_0\theta$ so that an approximate minimum may be located.

Concerning the weights $w_1, w_2, \ldots$, a system which has been successfully used is $w_1 = w_2 = \ldots$. Another system which has also proven useful is $w_1 = \Sigma \frac{(\partial G_i)^2}{\partial \theta_1}$, $w_2 = \Sigma \frac{(\partial G_i)^2}{\partial \theta_2}, \ldots$.

THE MODIFIED GAUSS-NEWTON METHOD

Hartley (1961) named the method previously described as the Gauss-Newton method and proposed another modification to the method. If the assumptions, (a) the non-linear function is continuous and first and second derivatives exist, (b) the observed X matrix is of full rank, and (c) if $Q = \underset{\bar{S}}{\lim} \inf Q(x;\theta)$ where $\bar{S}$ is the compliment of S which is a bounded convex set of the coefficient space $\theta_1, \ldots, \theta_m$, it is possible to find a vector ${}_0\theta$ in the interior of S such that $Q(x;{}_0\theta) < Q$, are satisfied. Hartley (1961) proved it is possible to describe an iterative process which will always converge to the least squares solution.

Under the above assumptions Hartley (1961) proposed to start with the usual normal equations obtained by utilizing a first order Taylor series approximation. However, instead of letting the second approximation for θ by ${}_0\theta + \Delta$ consider $Q(v) = Q(x;{}_0\theta + v\Delta) \quad 0 \leq v \leq 1$ and denote by v' the value of v for which Q(v) is a minimum on the interval $0 \leq v \leq 1$. Then v' may be found approximately if one evaluates Q(v) for $v = 0$, $v = \frac{1}{2}$, and $v = 1$ and determines the level of v for which the parabola through which $Q(0)$, $Q(\frac{1}{2})$, $Q(1)$ attains its minimum from

$v = \frac{1}{2} + \frac{1}{4}(Q(0) - Q(1))/(Q(1) - 2Q(\frac{1}{2}) + Q(0))$.

The above procedure is illustrated by Hartley with an example. The authors experience indicates that the direct application of Hartley's modification does not in all cases lead to a solution in a reasonable amount of time on the International Business Machines 1620 computer.

Further modifications to the method of Hartley are discussed in a master's report by Pence (1963). Pence mentions that Mr. Carlton Hassell and Dr. Dale Cooper, both of the Mathematical Research Section of Continental Oil Company have worked for some time in this area and plan to publish their results soon. Their computer program, NONLN, was written for the International Business Machines 7090 computer. Some of the additional provisions of the modified Gauss-Newton method discussed by Pence include elimination of the need for having a starting value for any of the linear coefficients. Also, graphing the results of the modified Gauss-Newton method and then by inspection estimating new values for the coefficients and calculating the new sum of squares, and adjusting the non-linear coefficient without having to estimate any linear coefficients are considered.

GENERAL PROPERTIES OF LEAST SQUARES ESTIMATES

Least squares estimates for the coefficients of any function have been shown to be sufficient estimates by Barnard (1963). In general, however, the estimates obtained by the method of least squares in the non-linear problem have no other general optimum properties (Kendall and Stuart (1961)).

ACKNOWLEDGMENTS

The author wishes to express his sincere appreciation to his major professor, Dr. Leslie F. Marcus, for his advice and assistance during the preparation of this report.

The help of Dr. Kali S. Banerjee and Dr. S. Thomas Parker in reviewing the manuscript is gratefully acknowledged.

BIBLIOGRAPHY

Aitken, A. C. (1948). "On the Estimation of Many Statistical Parameters." Proc. Roy. Soc. Edin. (A), 62, 369.

Barnard, G. A. (1963). "The Logic of Least Squares." J. R. Statist. Soc. B, 25, 124-127.

Graybill, F. A. (1961). An Introduction to Linear Statistical Models, Vol. 1, McGraw-Hill Book Co., Inc.

Hartley, H. O. (1961). "The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Functions by Least Squares." Ann. Math. Statist., 3, 269-280.

Kendall, M. G. and Stuart, A. (1961). The Advanced Theory of Statistics, Vol. 2, Charles Griffin & Company Limited.

Legendre. (1806). Nouvelles Methodes pour la determination des orbites des cometes, Paris.

Levenberg, K. (1944). "A Method for the Solution of Certain Non-Linear Problems in Least Squares." Quarterly of Applied Math., 2, 164-168.

Merriman, M. (1877). "A List of Writings Relating to the Method of Least Squares with Historical and Critical Notes." Transactions of the Connecticut Academy, Vol. IV, 151-232.

Pence, J. (1963). "Non-Linear Estimators." Unpublished Master's Report, Kansas State University Library, Manhattan, Kansas.

Plackett, R. L. (1949). "A Historical Note on the Method of Least Squares." Biometrika, 36, 458.

______________. (1950). "Some Theorems in Least Squares." Biometrika, 37, 149.

Wagner, J. (1964). Untitled and unpublished report, The Upjohn Co., Kalamazoo, Michigan.

Whittaker, E. T. and Robinson, G. (1924). The Calculus of Observations, Blackie and Sons, Limited.

Wilson, E. B. and Puffer, R. R. (1933). "Least Squares and Laws of Population Growth." Proc. Amer. Acad. Arts and Sci., 68, 285-382.

Wynn, P. (1962). "A Note on the Fitting of Certain Types of Experimental Data." Statist. Neerlandica, 16, 143-150.

PROBLEMS IN LEAST SQUARES

by

EDWARD LEON ARMBRUST

B. S., Kansas State University, 1963

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1965

The purposes of this report are to examine the early important developments of the method of least squares, to show the statistical properties of the estimates obtained by the method of least squares, and to present the techniques used in applying the method of least squares. Special consideration is given to the non-linear problem.

Legendre was the first to state the method of least squares. Gauss made one of the first attempts to put the method of least squares on a logical foundation. Laplace later added another justification.

There exists some disagreement among authors concerning the results implied by these early papers. R. L. Plackett has used matrix notation to clarify and summarize many of the early results in a much referred to paper published in 1949.

Most of the developers of least squares focused on the problem of minimization of the sum of squares of a linear function of known quantities and unknown coefficients. The estimates for this linear problem have desirable statistical properties.

Today, however, with the availability of high-speed electronic computers and the demand for more sophisticated mathematical models in the investigation of new problems, non-linear models are being considered and thoroughly investigated.

One approach to the non-linear problem is the approximation of a non-linear function by applying Taylor's series. Another method is the use of logarithms. Still another technique is to weight the observations or functions of the observations in some manner. A weighting was developed by K. Levenberg of Frankford Arsenal in 1944. This procedure has been called the Gauss-Newton method by H. O. Hartley of Texas A. & M. who modified it and called it the Modified Gauss-Newton method. The authors experience and that of others shows that this Modified Gauss-Newton method will not always yield a solution within a reasonable length of time. Therefore further modifications are needed. This problem has been more recently studied by Dr. Dale Cooper of Continental Oil Company.