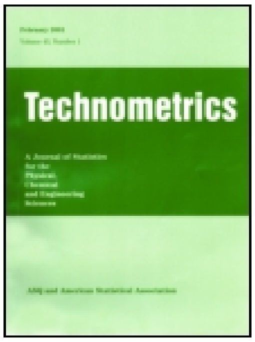

Taylor & Francis
Taylor & Francis Group

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Application of Stepwise Regression to Non-Linear Estimation

R. I. Jennrich & P. F. Sampson

To cite this article: R. I. Jennrich & P. F. Sampson (1968) Application of Stepwise Regression to Non-Linear Estimation, Technometrics, 10:1, 63-72, DOI: 10.1080/00401706.1968.10490535

To link to this article: https://doi.org/10.1080/00401706.1968.10490535

Published online: 30 Apr 2012.

TECHNOMETRICS VOL. 10, NO. 1 FEBRUARY 1968

Application of Stepwise Regression to Non-Linear Estimation*

R. I. JENNRICH AND P. F. SAMPSON

University of California, Los Angeles

This paper reviews three basic methods of non-linear least squares estimation and the best known modifications to the popular Gauss-Newton procedure. With regard to the Gauss-Newton procedure the paper identifies the problem of poor parameterization and gives simple examples to show how it may arise. A simple example is also given to show that one popular method of handling constrained boundaries can lead to erroneous results. A modification, based on stepwise linear regression techniques, to handle both of these problems is presented and discussed and an example is given to illustrate its effectiveness. The required fundamentals of stepwise linear regression are reviewed.

1. INTRODUCTION

Let y_t denote a set of n responses which have an assumed structure

$$y_t = f_t(\boldsymbol{\theta}) + e_t \qquad t = 1, \cdots, n. \tag{1}$$

The response function $f_t(\boldsymbol{\theta})$ is a known function of t and a parameter vector $\boldsymbol{\theta}$ which ranges over a parameter space $\boldsymbol{\Theta}$ of the form

$$\boldsymbol{\Theta} = \{(\theta_1, \cdots, \theta_p): a_i \leq \theta_i \leq b_i ; i = 1, \cdots, p\}. \tag{2}$$

A parameter vector $\hat{\boldsymbol{\theta}}$ is called a least squares estimate of some true and usually unknown parameter vector $\boldsymbol{\theta}_0$ if $\hat{\boldsymbol{\theta}}$ minimizes the residual sum of squares

$$Q(\boldsymbol{\theta}) = \sum_{t=1}^{n} (y_t - f_t(\boldsymbol{\theta}))^2. \tag{3}$$

We will summarize briefly three methods of minimizing $Q(\boldsymbol{\theta})$. These are steepest descent, quadratic approximation, and the Gauss-Newton method.

The steepest descent method consists in moving the parameter vector $\boldsymbol{\theta}$ a short distance in the direction opposite to that of the gradient vector $\mathbf{Q}'(\boldsymbol{\theta}) = ((\partial/\partial\theta_i)Q(\boldsymbol{\theta}))$. The process is repeated until $\boldsymbol{\theta}$ hopefully converges to a least squares estimate.

Quadratic approximation consists of approximating $Q(\boldsymbol{\theta})$ locally by a quadratic function and minimizing this function. The appropriate modification is made to the parameter vector and the whole process is repeated until it hopefully converges. Quadratic approximation is mathematically equivalent to using the Newton-Raphson method to solve the gradient equation $\mathbf{Q}'(\boldsymbol{\theta}) = \mathbf{0}$.

Received October 1966; Revised July 1967.

* This research was supported by the National Institutes of Health Grant FR-3.

The Gauss–Newton method consists in linearly approximating the response function $f_t(\boldsymbol{\theta})$ viewed as a function of $\boldsymbol{\theta}$ and solving the resulting linear least squares problem by the standard techniques of linear regression. To be more specific let $\mathbf{y}$, $\mathbf{f}(\boldsymbol{\theta})$, $\mathbf{e}$ denote the vectors (y_t), $(f_t(\boldsymbol{\theta}))$, (e_t) and let the vector $\mathbf{f}_i'(\boldsymbol{\theta}) = ((\partial/\partial\theta_i)f_t(\boldsymbol{\theta}))$ denote the partial derivative of the vector valued function $\mathbf{f}(\boldsymbol{\theta})$ with respect to the ith component of $\boldsymbol{\theta}$. Equation (1) becomes $\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}) + \mathbf{e}$. Replacing $\mathbf{f}(\boldsymbol{\theta})$ in this equation by its local linear approximation

$$\mathbf{f}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) \approx \mathbf{f}(\boldsymbol{\theta}) + \Delta\theta_1\mathbf{f}_1'(\boldsymbol{\theta}) + \cdots + \Delta\theta_p\mathbf{f}_p'(\boldsymbol{\theta}) \tag{4}$$

gives

$$\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}) = \Delta\theta_1\mathbf{f}_1'(\boldsymbol{\theta}) + \cdots + \Delta\theta_p\mathbf{f}_p'(\boldsymbol{\theta}) + \mathbf{e} \tag{5}$$

which may be viewed as a standard linear regression equation with the $\mathbf{f}_i(\boldsymbol{\theta})$ playing the role of independent variables, the $\Delta\theta_i$ the role of regression coefficients and $\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})$ the role of the dependent variable. Using this identification the $\Delta\theta_i$ are obtained by standard regression techniques. Replacing $\boldsymbol{\theta}$ by $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ the whole process is repeated until it hopefully converges.

Let $\mathrm{Q}''(\boldsymbol{\theta})$ denote the matrix $((\partial/\partial\theta_i)\ (\partial/\partial\theta_j)Q(\theta))$. The modification $\Delta\boldsymbol{\theta}$ of the vector $\boldsymbol{\theta}$ prescribed by each of the three methods on a given step is for steepest descent:

$$\Delta\boldsymbol{\theta} = -\alpha\mathrm{Q}'(\boldsymbol{\theta}), \qquad \alpha > 0 \tag{6}$$

quadratic approximation:

$$\Delta\boldsymbol{\theta} = -(\mathrm{Q}''(\boldsymbol{\theta}))^{-1}\mathrm{Q}'(\boldsymbol{\theta}) \tag{7}$$

Gauss–Newton:

$$\Delta\boldsymbol{\theta} = -\tfrac{1}{2}(\mathbf{f}'(\boldsymbol{\theta})^T\mathbf{f}'(\boldsymbol{\theta}))^{-1}\mathrm{Q}'(\boldsymbol{\theta}). \tag{8}$$

The step $\Delta\boldsymbol{\theta}$ will in general be distinct in both length and direction for each of the three methods. The steepest descent equation prescribes only the direction and not the length of $\Delta\boldsymbol{\theta}$. The latter is usually obtained by some form of a one dimensional search. Of the three the Gauss–Newton method seems to be by far the most popular probably because 1) it specifies the step size as well as its direction, 2) it does not require the computation of the $p(p + 1)/2$ second partial derivatives of each of the functions $f_t(\boldsymbol{\theta})$ which is required by the quadratic approximation method and 3) in the special case when the response function is linear in the parameters it converges in a single step.

2. Modifications of the Basic Gauss-Newton Procedure

The most commonly used, and perhaps most important, modification of the basic Gauss–Newton procedure is the use of partial steps $\pi\Delta\boldsymbol{\theta}$ in place of the full step $\Delta\boldsymbol{\theta}$ specified by (8). When the linear approximations on which the Gauss–Newton method is based fail the full step $\Delta\boldsymbol{\theta}$ frequently results in an increase in the residual sum of squares $Q(\boldsymbol{\theta})$. It can be shown however that if $\boldsymbol{\theta}$ is not already a minimizing value a sufficiently small step in the direction of $\Delta\boldsymbol{\theta}$ will reduce the residual sum of squares and it is this observation which motivates the use of partial steps. The proportion π of the step to be used may be obtained by trying an arbitrary sequence of values say 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{16}$, $\cdots$ until one which reduces the residual sum of squares is found or by some other one dimensional search procedure [4].

It can, and often does, happen that the vectors $\mathbf{f}_1'(\boldsymbol{\theta}), \cdots, \mathbf{f}_p'(\boldsymbol{\theta})$ are, or are nearly, linearly dependent. This is equivalent to saying that the matrix $\mathbf{A}(\boldsymbol{\theta}) = \mathbf{f}'^T(\boldsymbol{\theta})\mathbf{f}'(\boldsymbol{\theta})$ is, or is nearly, singular. Under these circumstances it may be difficult or impossible to compute the basic step $\Delta\boldsymbol{\theta}$ accurately. One could describe the problem by saying the response function is poorly parameterized. One solution is to reparameterize ([1], p. 803) but this is often difficult and usually inconvenient. Even in the case of linear regression where it is comparatively simple to reparameterize most workers prefer to use the parameters which are natural to the problem at hand and, if necessary, employ a modified form of regression called "stepwise regression" to avoid problems of singularity. In this modification variables are selected one at a time in the order which "does the most good" in reducing the residual sum of squares. If singularity problems arise (i.e., the tolerance discussed below becomes too low) one or more of the variables may be omitted. In this paper we propose the local application of these same stepwise techniques to deal with the problem of poor parameterization which frequently arises in non-linear regression. It turns out that stepwise techniques also provide a convient and sensible way to handle boundary restrictions on the components of the parameter vector. This is done by simply not modifying a parameter which has a boundary value unless the resulting modification is toward the interior of the parameter region. This amounts to "solving the problem on the boundary" whenever it appears that the absolute best fit parameter vector lies outside the parameter region.

Another solution to the poor parameterization problem can be found in the modification of Marquardt [5] which consists in replacing the matrix $\mathbf{A}(\boldsymbol{\theta})$ in the Gauss-Newton method by $\mathbf{A}(\boldsymbol{\theta}) + \lambda\mathbf{I}$ where $\mathbf{I}$ denotes an identity matrix. This method may be viewed as a compromise between the Gauss-Newton and steepest descent methods. When $\lambda = 0$ it gives a Gauss-Newton step. As $\lambda \to \infty$ the step shortens and its direction approches that of the steepest descent method. The point to note here is that even if the matrix $A(\theta)$ is singular, the matrix $\mathbf{A}(\boldsymbol{\theta}) + \lambda\mathbf{I}$ is non-singular whenever $\lambda > 0$ and can be made arbitrarily well conditioned by choosing λ sufficiently large. Values of λ which are too large, however, lead to slow convergence. The usual practice is to adjust the value of λ dynamically from one iteration to the next by a rule which is based on the past behavior of the algorithm.

3. Near-Singularity

There are many reasons why the sum of cross products matrix $\mathbf{A}(\boldsymbol{\theta})$ is frequently singular or nearly singular. In this section we will point out how this difficulty may arise when attempting to fit a sum of exponentials. Let

$$f_t(\boldsymbol{\theta}) = e^{\theta_1 t} + e^{\theta_2 t}. \tag{9}$$

In the notation of Section 1

$$\mathbf{f}_1'(\boldsymbol{\theta}) = (te^{\theta_1 t}) \quad \text{and} \quad \mathbf{f}_2'(\boldsymbol{\theta}) = (te^{\theta_2 t}). \tag{10}$$

When $\theta_1 = \theta_2$ these two vectors are equal and thus linearly dependent. If the data y_t follow a straight line or any concave-down path as t ranges from 1 to n it can be shown that the least squares values $\hat{\theta}_1$ and $\hat{\theta}_2$ must be equal and hence that $\mathbf{A}(\boldsymbol{\theta})$ is singular at a least squares solution and nearly singular in a neighborhood of it. Less precise but perhaps more general insight can be obtained by considering the slightly more general sum of exponentials

$$f_t(\boldsymbol{\theta}) = \theta_1 e^{\theta_3 t} + \theta_2 e^{\theta_4 t}. \tag*{11)}$$

If the fitted response function $f_t(\hat{\boldsymbol{\theta}})$ follows the path of a single exponential, say $\alpha e^{\beta t}$, it can be shown that $\mathbf{A}(\hat{\boldsymbol{\theta}})$ is singular and if the fitted response function nearly follows such a path $A(\hat{\boldsymbol{\theta}})$ is nearly singular. The same statement can be made when the fitted response follows or nearly follows any striaght line path.

A still less precise and more general statement is that singularity problems are likely to arise whenever there is a curve in the parameter space along which the response function $f_t(\theta)$ differs little from the least squares response $f_t(\hat{\boldsymbol{\theta}})$; that is, whenever it is possible to vary $\boldsymbol{\theta}$ substantially without substantially affecting the best fit. A sometimes useful, though obviously over-simplified, statement is that trouble arises when there are too many parameters.

4. Stepwise Regression

In this section we will review the fundamentals of stepwise regression. The method is based on the Gauss–Jordan pivot. Let (a_{ij}) be an n by n matrix whose kth diagonal element a_{kk} is non-zero. The result of a Gauss–Jordan pivot on this element is a new matrix whose ijth element

$$\tilde{a}_{ij} \begin{cases} a_{ij} - a_{ik}a_{kj}/a_{kk} & i \neq k, \quad j \neq k \\ -a_{ik}/a_{kk} & i \neq k, \quad j = k \\ a_{kj}/a_{kk} & i = k, \quad j \neq k \\ 1/a_{kk} & i = k, \quad j = k. \end{cases} \tag{12}$$

Gauss–Jordan pivots commute. That is, a pivot on the ith diagonal element followed by a pivot on the jth diagonal element produces the same result as pivots on the ith and jth diagonal elements in reverse order. Moreover a Gauss–Jordan pivot is its own inverse. That is, two applications of the same pivot leaves the matrix (a_{ij}) unchanged. If (a_{ij}) is partitioned as the matrix on the left below

$$\begin{bmatrix} \mathbf{A}_{11}, & \mathbf{A}_{12} \\ \mathbf{A}_{21}, & \mathbf{A}_{22} \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{A}_{11}^{-1}, & \mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ -\mathbf{A}_{21}\mathbf{A}_{11}^{-1}, & \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{bmatrix} \tag*{13)}$$

and if it is possible to pivot on all of the diagonal elements of $\mathbf{A}_{11}$ (i.e., no zeros are encountered) then $\mathbf{A}_{11}$ is non-singular and the result of pivoting on its diagonal elements is given on the right in (13). In particular if a matrix is pivoted on all of its diagonal elements the result is its inverse.

If a matrix is non-negative (i.e., positive semi-definite) then its diagonal elements are non-negative and they remain so after any number of pivots. A non-negative matrix is singular if and only if for any order of pivoting at least one diagonal element becomes zero before all diagonal elements have been used as pivots. Let $\bar{a}_{ii}$ denote the value of the ith diagonal element of a non-negative matrix (a_{ij}) after pivoting on a number of other diagonal elements. The ratio $\bar{a}_{ii}/a_{ii}$ is called the tolerance of the ith diagonal element at this state. Since $\bar{a}_{ii}$ is obtained from a_{ii} by a sequence of subtractions of non-negative quantities the tolerance of the ith diagonal element in a sense measures its precision. The use of an element with low tolerance as a pivot may well destroy the precision of the entire matrix. For this reason, in practice, elements with low tolerance (say 10^{-5} on an 8 decimal place computer) are not used as pivots.

Let $\mathbf{X} = (x_{ti})$ denote an n by p matrix and let $\mathbf{Y} = (y_{ti})$ denote an n by q matrix. The columns $\mathbf{x}_i = (x_{ti})$ and $\mathbf{y}_i = (y_{ti})$ of these matrices will be called variables. Assume the matrix $\mathbf{X}^T\mathbf{X}$ is non-singular and let the matrix on the right in (14) denote the result of pivoting the matrix on the left on the diagonal elements of $\mathbf{X}^T\mathbf{X}$

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X}, & \mathbf{X}^T\mathbf{Y} \\ \mathbf{Y}^T\mathbf{X}, & \mathbf{Y}^T\mathbf{Y} \end{bmatrix} \to \begin{bmatrix} (\mathbf{X}^T\mathbf{X})^{-1}, & \boldsymbol{\beta} \\ -\boldsymbol{\beta}, & \mathbf{S} \end{bmatrix}. \tag{14}$$

It follows from (13) and the standard formulas of regression theory that the ijth component β_{ij} of the matrix $\boldsymbol{\beta}$ is the regression coefficient for variable $\mathbf{x}_i$ which results from regressing variable $\mathbf{y}_j$ on the entire set of x-variables. That is

$$\hat{\mathbf{y}}_j = \beta_{1j}\mathbf{x}_1 + \cdots + \beta_{pj}\mathbf{x}_p \tag{15}$$

is the linear regression of $\mathbf{y}_j$ on $\mathbf{x}_1, \cdots, \mathbf{x}_p$. Moreover the matrix $\mathbf{S} = (s_{ij})$ is the sum of cross products matrix for the residuals $\mathbf{y}_i - \hat{\mathbf{y}}_i$. In particular its ith diagonal element is the residual sum of squares resulting from regressing $\mathbf{y}_i$ on all of the x-variables. Thus all the essential building blocks of a standard regression analysis are readily available in the matrix on the right in expression (14).

Stepwise regression [3] is based on the observation that a variable can be moved from the set of y-variables to the set of x-variables or from the set of x-variables to the set of y-variables by simply pivoting on the corresponding diagonal element of the matrix on the right in expression (14). In standard applications one of the y-variables, say $\mathbf{y}_d$, is the real dependent variable. The object at each step is to move that y-variable (other than $\mathbf{y}_d$) into the set of x-variables (the set of variables regressed upon) which gives the greatest possible reduction in the residual sum of squares of the dependent variable $\mathbf{y}_d$. (Equivalently this is the y-variable whose partial correlation with $\mathbf{y}_d$, when partialed on the x-variables, is greatest; or the y-variable whose regression coefficient after entering the set of x-variables has the greatest F-value.) Since the reduction in the residual sum of squares resulting from entering the variable $\mathbf{y}_i$ is β_{id}^2/s_{ii} it is a simple matter to locate the required variable. The variable is entered providing the tolerance of the corresponding diagonal element of the matrix on the right in expression (14) is above a specified threshold value. Thus the variable $\mathbf{y}_i$ actually entered, if any, is the one among all $\mathbf{y}_i \neq \mathbf{y}_d$, whose tolerance exceeds the prescribed threshold and whose β_{id}^2/s_{ii} value is maximum. The process begins with all variables in the y-set. Entering one variable at a time into the x-set the process ends when no available variables satisfying the tolerance criterion remains or when the reduction in the residual sum of squares fails to exceed a specified threshold. (This latter threshold may alternatively be stated in terms of partial correlation or F-value thresholds).

Roughly speaking variables are entered in the order that (one step at a time at least) does the most good providing their entry does not do too much damage to the numerical presicion of the results. In ordinary applications to linear regression ([2], p. 233), regression coefficients and other intermediate results are produced at each step and variables whose removal does not greatly increase the residual sum of squares may be removed. In the applications below no such intermediate output is produced and variable removal is used only to prevent boundary violations.

5. Details of the Stepwise Modification

Augment the matrix $\mathbf{A}(\boldsymbol{\theta})$ by a $d = p + 1$st row and column so that its components become

$$\begin{aligned} a_{ij}(\boldsymbol{\theta}) &= \sum_{t=1}^{n} \frac{\partial}{\partial \theta_i} f_t(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} f_t(\boldsymbol{\theta}) \qquad i, j = 1, \cdots, p \\ a_{id}(\boldsymbol{\theta}) &= a_{di}(\boldsymbol{\theta}) = \sum_{t=1}^{n} \frac{\partial}{\partial \theta_i} f_t(\boldsymbol{\theta})(y_t - f_t(\boldsymbol{\theta})) \qquad i = 1, \cdots, p \\ a_{dd}(\boldsymbol{\theta}) &= \sum_{t=1}^{n} (y_t - f_t(\boldsymbol{\theta}))^2. \end{aligned} \tag{16}$$

A given iteration step proceeds as follows. Identify the dth diagonal element of the augmented matrix $\mathbf{A}(\boldsymbol{\theta})$ with the dependent variable and perform Gauss–Jordon pivots in the order specified by stepwise regression. While the threshold tolerance should be a parameter of the computational procedure, a value of about 10^{-5} on an 8 decimal place computer will assure about 3 decimal places of precision in the pivots used and seems to be a satisfactory threshold. The components of $\Delta\boldsymbol{\theta}$ are given by

$$\Delta\theta_i = \begin{cases} \bar{a}_{id} & \text{if } \mathbf{A}(\boldsymbol{\theta}) \text{ has been pivoted on its } i\text{th diagonal element} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Here $\bar{a}_{id}$ denotes the value of the idth element of $\mathbf{A}(\boldsymbol{\theta})$ after pivoting. Next the maximum value $\alpha \leq 1$ for which $\boldsymbol{\theta} + \alpha\,\Delta\boldsymbol{\theta}$ satisfies the boundary conditions (2) is obtained. If $\alpha = 0$ then $\boldsymbol{\theta}$ lies on at least one coordinate boundary and the corresponding coordinate of $\Delta\boldsymbol{\theta}$ is positive if the boundary is an upper boundary and negative if it is a lower boundary. As soon as such a coordinate of $\Delta\boldsymbol{\theta}$ is encountered the corresponding diagonal element of $A(\boldsymbol{\theta})$ is unpivoted by means of an inverse Gauss–Jordan pivot. Using the resulting $\Delta\boldsymbol{\theta}$ a new α is obtained as before and the process is continued until an $\alpha > 0$ is obtained or all diagonal elements are unpivoted. If an $\alpha > 0$ is obtained $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\theta} + \alpha\,\Delta\boldsymbol{\theta}$ completing the iteration step. If all diagonal elements are unpivoted (in the experience of the authors this has never happened) the entire iteration process is terminated. In effect what has been done is that the coordinates of $\boldsymbol{\theta}$, which already have boundary values and whose modification would lead to boundary violations, are held fixed at their boundary values while a standard Gauss–Newton iteration step is performed on remaining coordinates which previously satisfied the tolerance criterion.

While this completes the description of the stepwise modification of a Gauss–Newton iteration step it is probably a good idea to also use some form of the partial step modification described in Section 2. Beginning with the step $\alpha\,\Delta\boldsymbol{\theta}$ the value of α is further reduced, if necessary, until a reduction in the residual sum of squares is obtained.

The entire procedure then consists of selecting a starting value for $\boldsymbol{\theta}$ and applying iteration steps until some scale free convergence criterion (such as a relative change in the residual sum of squares of less than 10^{-7}) is satisfied.

6. Some Practical Considerations

In most implementations of the Gauss–Newton algorithm the user supplies one or more subroutines to evaluate the functions $f_t(\boldsymbol{\theta})$ (which almost always have the form $f_t(\boldsymbol{\theta}) = f(\mathbf{x}_t\ ,\ \boldsymbol{\theta})$) and their partial derivatives. There is an advantage to doing this in a single subroutine since for many functions it is computationally only a little more expensive to evaluate the function and its partial derivatives than to evaluate the function alone. For example, if $f_t(\theta) = \theta_1 e^{\theta_2 t}$ then both of its partial derivatives $e^{\theta_2 t}$ and $t\theta_1 e^{\theta_2 t}$ can be obtained by one additional multiplication. This observation also suggests that extra evaluations of the residual sum of squares may be relatively expensive. In many applications they are nearly as expensive as an entire Gauss–Newton iteration step and for this reason the number of extra residual sum of squares evaluations in the partial step modification of Section 2 should be kept to a minimum.

A number of popular non-linear estimation routines handle boundary restrictions by computing an unconstrained step and projecting the results onto the appropriate boundary whenever a boundary violation would otherwise occur. When properly handled [6] this technique works for the steepest descent method. As can be seen from Figure 1 however it does not work for the Gauss–Newton method and most of its modified forms. The line in Figure 1 represents an upper boundary of a two-dimentional parameter region and the ellipse represents a constant value surface for $Q(\boldsymbol{\theta})$ corresponding to a response function which is linear in its parameters. Under these conditions the unconstrained minimum of $Q(\boldsymbol{\theta})$ occurs at the center $\boldsymbol{\theta}_c$ of the ellipse which is outside of the parameter region. Moreover the constrained best fit occurs at the point of tangency $\boldsymbol{\theta}$. Assume that an iteration step begins at any boundary point $\boldsymbol{\theta}$. An unconstrained Gauss–Newton step will take it to $\boldsymbol{\theta}_c$ and projection onto

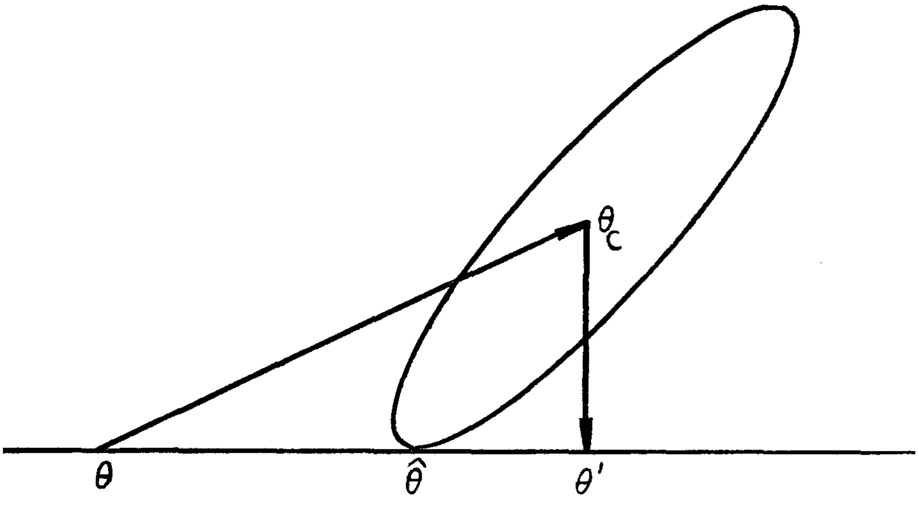

FIGURE 1—Figure showing the projection method of handling boundary restrictions is inappropriate when used with the Gauss-Newton method. The symbols are defined in Section 6.

the boundary to $\boldsymbol{\theta}'$. Thus $\boldsymbol{\theta}'$ is a limit point of such a procedure but, unfortunately, not a least squares solution.

Whenever possible it is a good idea to plot the original data and the fitted model or at least to list the residuals $y_t - f_t(\hat{\boldsymbol{\theta}})$. This gives some indication of the appropriateness of the assumed structure and may indicate a failure of the algorithm to converge to a minimizing solution (e.g., when it converges to a local, rather than a global, minimum). Statistical information, such as an estimate of the variance of the errors e_t and estimates of the variances of the parameter estimates $\hat{\theta}_1, \cdots, \hat{\theta}_p$, is also helpful in interpreting and evaluating results (formulas for these and other statistical parameters can be found in the writeup of the computer program discussed in the next section).

7. Example

To demonstrate that the stepwise modification discussed in the previous section works when the matrix $\mathbf{A}(\boldsymbol{\theta})$ is singular or nearly singular we will use it to fit the response function (9) to the data

$$y_t = 2 + 2t, \qquad t = 1, \cdots, 10. \tag{18}$$

For this problem, as pointed out in Section 3, the least squares values $\hat{\theta}_1$ and $\hat{\theta}_2$ are equal, $\mathbf{A}(\boldsymbol{\theta})$ is singular at the least squares solution and nearly singular in a neighborhood of it.

Table 1 shows the convergence of the stepwise modification of the Gauss–Newton algorithm using a tolerance of 10^{-5}. The accuracy of the final values $\hat{\theta}_1 = .25785$ and $\hat{\theta}_2 = .25780$ is demonstrated by the fact that they agree to four decimal places with the value $\hat{\theta} = .25783$ obtained by replacing θ_1 and θ_2 by a single parameter θ and solving the resulting well behaved one parameter least squares problem.

TABLE 1

Convergence in the Example of Section 7 Using the Stepwise Modification

Iteration	Step Halving Freq.	θ_1	θ_2	Residual Mean Sq.
0	0	.30000	.40000	521.41
1	2	.35468	.35515	429.84
2	0	.35468	.22546	88.154
3	4	.34180	.25697	83.743
10	0	.22799	.30565	21.330
30	0	.25790	.25774	15.545
60	10	.25785	.25780	15.545
100	10	.25785	.25780	15.545

The results of applying the Gauss–Newton algorithm without the stepwise modification (this was done using the stepwise alogrithm and setting the tolerance equal to zero) to the same problem are given in Table 2. The final residual mean square value 429.99 of Table 2 when compared with the corresponding value 15.545 of Table 1 indicates clearly that after 100 steps the iteration is far from a least squares solution. The iteration, unfortunately, did not obviously "blow-up" but instead looks like it has converged rapidly. After the second step, at the beginning of which the parameters have nearly equal values, only very small changes occur. While the step halving frequency is high (it had an imposed maximum of 11 corresponding to one 2048th of a full step) this is also a characteristic of convergence.

A computer program (7094 FORTRAN IV) which implements the stepwise method discussed in Section 5 has been written by one of the authors. This program was used to perform the calculations required in the previous example and has also been used quite successfully for the past two years on a large variety of problems arising in medical research. The program and its writeup may be obtained by writing to the Program Librarian at the Health Sciences Computing Facility, University of California, Los Angeles, California 90024.

TABLE 2

Convergence in the Example of Section 7 Without the Stepwise Modification

Iteration	Step Halving Freq.	θ_1	θ_2	Residual Mean Sq.
0	0	.30000	.40000	521.41
1	2	.35468	.35515	429.84
2	10	.35503	.35481	429.84
3	11	.35468	.35515	429.85
10	10	.35503	.35481	429.85
30	10	.35503	.35481	429.88
60	10	.35504	.35481	429.93
100	10	.35505	.35481	429.99

References

1. Box, G. E. P., 1960. Fitting empirical data. *Annals of the New York Academy of Sciences, 86,* 792–816.
2. Dixon, W. J., 1965. Editor. *BMD Biomedical Computer Programs,* UCLA Student Store, 308 Westwood Blvd., Los Angeles, California 90024.
3. Efroymsen, M. A., 1960. Multiple regression analysis. *Mathematical Methods for Digital Computers.* Edited by A. Ralston and H. S. Wilf, John Wiley and Sons, New York, 191–203.
4. Hartley, H. O., 1961. The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. *Technometrics, 3,* 269–280.
5. Marquardt, D. W., 1963. An algorithm for the estimation of non-linear parameters. *Society for Industrial and Applied Mathematics Journal, 11,* 431–441.
6. Rosen, J. B., 1960. The gradient projection method for non-linear programming. Part I. Linear Constraints. *Society for Industrial and Applied Mathematics Journal, 8,* 181–217.