NONLINEAR LEAST SQUARES ESTIMATION[1]

By H. O. Hartley[2] and Aaron Booker[3]

Iowa State University

1. Introduction and summary. We are given a set of N responses Y_t which have arisen from a nonlinear regression model

$$Y_t = f(x_t, \theta) + e_t; \qquad t = 1, 2, \cdots, N. \tag{1.1}$$

Here x_t denotes the tth *fixed* input vector of k elements giving rise to Y_t, whilst θ is an m-element unknown parameter vector with elements θ_i and the e_t are a set of N independent error residuals from $N(0, \sigma^2)$ with σ^2 unknown. The expectations of the Y_t, are therefore the functions $f(x_t, \theta)$ which will be assumed to satisfy certain regularity conditions. The problem is to estimate θ notably by *least squares*.

In this paper we shall develop an iterative method of solution of the least squares equations which has the following properties:

(a) the computational procedure is convergent for finite N;

(b) the resulting estimators are asymptotically *100% efficient* as $N \to \infty$.

In Sections 2–4 we give a survey of our results leaving the mathematical proofs to Sections 5–7 whilst in Section 8 we illustrate our method with an example.

Although our theoretical development is oriented towards our specific goals certain results are proved in a somewhat more general form. Some of our theory will be seen to correspond to well known theorems on stochastic limits which have to be reproved because of certain modifications which we require.

2. The formulation of the large sample theory of the least square point estimator. The estimation of θ by nonlinear least squares (here identical with maximum likelihood estimation) gives rise to the minimization of

$$Q(\theta) \equiv \sum_t (Y_t - f(x_t, \theta))^2 \tag{2.1}$$

with associated *least squares equations*

$$(\partial Q/\partial \theta_i)(\theta) = Q_i(\theta) = 0, \qquad i = 1, 2, \cdots, m. \tag{2.2}$$

Whilst there are iterative methods of solving the nonlinear least squares equations (2.2) (see e.g. Hartley, 1961) it will, in general, not be known whether the solution θ of (2.2) so obtained is a local minimum of (2.1) or the absolute minimum, and it is only for this absolute minimum of $Q(\theta)$ that asymptotic optimality properties have been established. The exhaustive scanning of the parameter space is usually computationally impractical, particularly when the number of param-

Received 16 August 1963; revised 21 October 1964.

[1] Prepared with the partial support of the National Science Foundation Grant GP-2240.

[2] Now at Texas A and M University, College Station, Texas.

[3] Now at United Electrodynamics, Inc., Alexandria, Virginia.

eters is >3 and the conditions on Q ensuring uniqueness of the solutions of Equations (2.2) are usually not satisfied. A method of estimation is therefore developed which avoids the search for the absolute minimum of (2.1) and yet yields two estimators, $\tilde{\theta}$ and $\dot{\theta}$, which are asymptotically 100% efficient under fairly general assumptions. The method consists of splitting the N observations into m groups of (say) n observations each so that $N = mn$ and the responses $Y_{h\tau}$ arise from k-dimensional inputs $x_{h\tau}$ ($h = 1, 2, \cdots, m; \tau = 1, 2, \cdots, n$). The convex closures $C_h(n)$ of the $x_{h\tau}$ must be disjoint. (See more specific formulation in Section 6.) The method then consists of two steps:

STEP (i) Construct a consistent estimator θ^* of θ by solving the system of m nonlinear equations

$$\bar{Y}_h = \bar{f}(h, \theta^*) \tag{2.3}$$

where $\bar{Y}_h = n^{-1} \sum_\tau Y_{h\tau}$; $\bar{f}(h, \theta) = n^{-1} \sum_\tau f(x_{h\tau}, \theta)$.

STEP (ii) Using θ^* as a *starting* value carry out *one* iteration step of the standard Gauss Newton iteration applied to (2.1) to obtain the 100% efficient estimator $\tilde{\theta}$. As an alternative starting with θ^* the modified Gauss Newton iteration (Hartley, 1961) may be carried to convergence yielding a local (or absolute) minimum of $Q(\theta)$ at $\theta = \dot{\theta}$ which is likewise asymptotically 100% efficient. In fact under certain additional assumptions $\dot{\theta}$ (coincident with $\hat{\theta}$) will be asymptotically the unique consistent solution of the likelihood Equations (2.2), see Huzurbazar, (1948), and hence yield, asymptotically the absolute minimum of $Q(\theta)$.

The main result lies in establishing the consistency of θ^* under very general conditions on f. Computationally the solution of (2.3) is achieved by driving $\bar{Q}(\theta) = \sum_h (\bar{Y}_h - \bar{f}(h, \theta))^2$ to its minimum value of zero with the help of the *modified Gauss Newton* iteration. Certain computational shortcuts are introduced. It will be noted that when $f(x, \theta)$ is linear in θ the estimators $\tilde{\theta}$ and $\dot{\theta}$ agree with the standard BLUE least squares estimator irrespective of what starting value is used. Briefly we make the following assumptions. (For a more specific statement of our assumptions see Section 6.)

The first derivatives, $f_i \equiv (\partial f/\partial\theta_i)(x, \theta)$, are continuous functions of x and θ where θ is confined to a certain closed, bounded, convex region, S, of the m-space and the x_t are confined to certain convex closures $C_h(n)$ of the k-space. We also assume that the $N \times m$ matrix of first derivatives f_i has full rank, viz

$$F = (f_i(x_t, \theta)) \tag{2.4}$$

has rank m for all θ in S and any set of vectors x_t, $t = 1, 2, \cdots, N$ of which at least m are distinct. Certain minor additional assumptions concerning the function $f(x, \theta)$ will be described in Section 3 below.

We shall be concerned with the asymptotic behavior of the above estimators of θ as the sample size $N \to \infty$. More specifically we shall assume for convenience that N is a multiple of m, i.e. that $N = nm$ and $n \to \infty$. Moreover, we shall assume that it will be possible to split the set of x_t vectors into m groups of n vectors

$x_{h\tau}$ $(h = 1, 2, \cdots, m; \tau = 1, 2, \cdots, n)$ in such a way that the convex closures $C_h(n)$ containing the $x_{h\tau}$ are disjoint, uniformly bounded in n and that the minimum distance of any two points lying in different $C_h(n)$ is bounded away from zero as $n \to \infty$. These restrictions can usually be satisfied in a great variety of ways.

A method of finding a solution θ^* of (2.3) will be given in Section 3 and the consistency of θ^* will be proved in Section 6.

3. The consistent estimator θ^*. For the computation of the consistent estimator θ^* we require the following lemma, the proof of which is given in Section 6.

LEMMA. *If we denote the first derivatives of the group averages (see* (2.3)*) by*

$$\bar{f}_i(h, \theta) = (\partial \bar{f}/\partial \theta_i)(h, \theta)$$

then the $m \times m$ matrix of the $\bar{f}_i$ has rank m, i.e.

$$\text{Rank} \quad (\bar{f}_i(h, \theta)) = m \tag{3.1}$$

for $i = 1, 2, \cdots, m; h = 1, 2, \cdots, m$, and for all θ in S and all $x_{h\tau}$ sets with properties specified in Section 2.

The estimator θ^* has been defined as a solution of the m nonlinear Equations (2.3) and will be obtained as the absolute minimum of the least squares form

$$\bar{Q}(\theta) = \sum_{h=1}^{m} (\bar{Y}_h - \bar{f}(h, \theta))^2$$

It is clear that any stationary point of $\bar{Q}(\theta)$ is a solution of (2.3). For a stationary point θ must satisfy the equations

$$0 = \bar{Q}_i(\theta) = (\partial \bar{Q}/\partial \theta_i)(\theta) = -2 \sum_{h=1}^{m} (\bar{Y}_h - \bar{f}(h, \theta)) \bar{f}_i(h, \theta) \tag{3.2}$$

for $i = 1, 2, \cdots, m$.

Now since the matrix $(\bar{f}_i(h, \theta))$ has rank m (see (3.1)) any root of the system (3.2) must satisfy $\bar{Y}_h - \bar{f}(h, \theta) = 0$ that is, Equations (2.3). Various iterative methods are now available for computing a stationary point of the least squares form $\bar{Q}(\theta)$. For example the *modified Gauss Newton iteration* (Hartley, 1961) will converge to a stationary point and hence in the present case to the absolute minimum $\bar{Q}(\theta) = 0$ if, in addition to (3.1), the following assumptions are made about $\bar{Q}(\theta)$.

(3.3) Assume that S is convex, closed, and bounded such that it is possible to find a *starting value* ${}_0\theta$ in S such that $\bar{Q}({}_0\theta) < \liminf \bar{Q}(\theta)$ for θ in $\bar{S}$ where $\bar{S}$ is the complement of S.

(3.4) No two stationary points of $\bar{Q}(\theta)$ yield identical values of $\bar{Q}(\theta)$, which means that (2.3) has a unique solution.

The above two assumptions ((3.3) and (3.4)), in conjunction with (3.1), are sufficient to ensure the convergence of the modified Gauss Newton iteration to a solution of (3.2) and hence of (2.3). If Assumption (3.4) is not satisfied, the sequence of θ vectors generated by the modified Gauss Newton method may have

more than one point of accumulation in S but we can still select one of the points of accumulation to which a subsequence will converge and which will be the absolute minimum of $\bar{Q}(\theta)$. For a description of these see Hartley (1961).

4. The asymptotically efficient estimators $\tilde{\theta}$ and $\dot{\theta}$. The estimator $\tilde{\theta}$ is obtained from θ^* by adding a correction vector $D = \tilde{\theta} - \theta^*$ derived from the m linear equations.

$$(4.1) \quad \sum_{j=1}^{m} [\sum_{h\tau} f_i(x_{h\tau}, \theta^*) f_j(x_{h\tau}, \theta^*)] D_j = \sum_{h\tau} (Y_{h\tau} - f(x_{h\tau}, \theta^*)) f_i(x_{h\tau}, \theta^*).$$

The rank of Equations (4.1) is m by the Assumption (2.4). The estimator $\dot{\theta}$ is the limit of the modified Gauss Newton iteration (see Hartley, 1961) with θ^* as starting point. Both $\tilde{\theta}$ and $\dot{\theta}$ can be shown to be asymptotically 100% efficient, for $\tilde{\theta}$ this is done in Section 7. No such properties can be assured for a stationary point of $Q(\theta)$ (i.e. solution of (3)) or indeed for a local minimum of $Q(\theta)$. The asymptotic and approximate variances and covariances of both $\tilde{\theta}$ and $\dot{\theta}$ are given by

$$\operatorname{Cov} \dot{\theta}_i \dot{\theta}_j \doteq \operatorname{Cov} \tilde{\theta}_i \tilde{\theta}_j \doteq \sigma^2 (\sum_t f_i(x_t, \theta) f_j(x_t, \theta))^{-1}$$

and may be estimated by substituting $\tilde{\theta}$ or $\dot{\theta}$ for θ and $Q(\dot{\theta})/(N - m)$ or $Q(\tilde{\theta})/(N - m)$ for σ^2.

5. Some theorems on stochastic limits. The following theorem is a slight modification of Theorem 1 given in Mann and Wald (1943). We understand that the present modification is fully proved in lecture notes by H. Chernoff.

THEOREM 1. *A sequence of scalar (vector) functions $f_n(x_n)$ of a random vector x_n is such that*

$$f_n(x_n) = o_p(r(n))(O_p)$$

if and only if for every $\epsilon > 0$ there is a sequence of regions $R_n(\epsilon)$ such that;

(i) $f_n(a_n) = o(r(n))(O)$ *when* $a_n \varepsilon R_n$ *and*

(ii) $P(x_n \varepsilon R_n(\epsilon)) \geqq 1 - \epsilon$ *for* $n > N(\epsilon)$. (*The O_p inside ( ) and subsequent symbols inside ( ) represent alternative forms of the theorems.*)

Note that the dimension of $R_n(\epsilon)$ is the number of components in the vector x_n which may depend on n. The following corollary is an obvious generalization of Corollary 1 in Mann and Wald.

COROLLARY 1.1. *Let $x_n^{(j)} = O_p(f_j(n))$ or $x_n^{(j)} = o_p(f_j(n))$ for $j = 1, 2, \cdots, r$ and $R_n(\epsilon)$ be a sequence of subsets of the $k(n)$-dimensional space where $y_n = (y_n^{(1)}, y_n^{(2)}, \cdots, y_n^{(k(n))})$ is such that $P(y_n \varepsilon R_n(\epsilon)) \geqq 1 - \epsilon$ for sufficiently large n. Let $g_n(x^{(1)}, \cdots, x^{(r)}, y_n)$ be a sequence of functions such that for every $\epsilon > 0$, $g_n(a_n, b_n) = O(f(n))(o)$ if $a_n^{(j)} = O(f_j(n))$ or $a_n^{(j)} = o(f_j(n))$ and $b_n \varepsilon R_n(\epsilon)$. Then $g_n(x_n, y_n) = O_p(f(n))(o_p(f(n)))$.*

PROOF. Let $f_n(x_n)$ of Theorem 1 be $f_n(x_n) = x_n$. Then by (ii) there exist regions for x_n which can be combined with the given regions for y_n to satisfy (ii) for (x_n, y_n). Condition (i) of Theorem 1 is given by the hypothesis of Corollary 1.1 and the result follows directly from Theorem 1. Next we prove

COROLLARY 1.2. *Let y_n, x_n, z_n be sequences of stochastic vectors with dimension $k(n)$, r, r respectively. Let $R_n(\epsilon)$ be a sequence of regions such that $P(y_n \varepsilon R_n(\epsilon)) \geqq 1 - \epsilon$ for sufficiently large n and $x_n = O_p(1)$, $z_n - x_n = O_p(f(n))$ where $\lim_{n\to\infty} f(n) = 0$. Let $G_n(y_n, x_n)$ be a sequence of functions and define $H_n(y_n, x_n, z_n)$ by*

$$H_n(y_n, x_n, z_n) = G_n(y_n, x_n) - G_n(y_n, z_n) - T_{sn}(y_n, x_n, z_n)$$

where T_{sn} is the sth *order multiple Taylor expansion of $G_n(y_n, x_n)$ with respect to x_n evaluated at $x_n = z_n$. If $G_n(y_n, x_n)$ has continuous and uniformly bounded* $(s + 1)$th *order partial derivatives with respect to x provided y_n is in $R_n(\epsilon)$, then $H_n(y_n, x_n, z_n) = o_p(f^s(n))$.*

PROOF. Make the following identification of the quantities in Corollary 1.1 and 1.2.

Corollary 1.1	Corollary 1.2
x_n	x_n, $x_n - z_n$
y_n	y_n
$f_j(n)$ for x_n	1
$f_j(n)$ for $x_n - z_n$	$f(n)$
$f(n)$	$f(n)$
$g_n(x_n, y_n)$	$H_n(y_n, x_n, z_n)$

Thus it is only necessary to show that for every $\epsilon > 0$ and for any sequence a_n, b_n, c_n such that $a_n \varepsilon R_n(\epsilon)$, $c_n - b_n = O(f(n))$, it follows that $H_n(a_n, b_n, c_n) = o(f^s(n))$. That is, we must verify the property of our function H_n which is stipulated by the O property of the g_n function in Corollary 1.1 to which it corresponds.

Since H_n is the remainder term in Taylor's formula for functions of several variables and the mixed $(s + 1)$th order partial derivatives are bounded by, say B, the sequence H_n can be written

$$|H_n(a_n, b_n, c_n)| \leqq (\textstyle\sum_{i=1}^{r} |c_n^{(i)} - b_n^{(i)}|)^{s+1}/(s + 1)!.$$

Using the inequality $(\sum_{i=1}^{r} u_i)^N \leqq (\sum_{i=1}^{r} u_i^2)^{N/2} r^N$ it follows that

$$|H_n(a_n, b_n, c_n)| \leqq O(|c_n - b_n|^{s+1}) = o(f^s(n))$$

where $|c_n - b_n|$ denotes the modulus of the vector $c_n - b_n$.

6. The consistency of the estimator θ^*. We now return to the model of Section 1, that is we consider the nonlinear regression law

$$Y_{hr} = f(x_{hr}, \theta) + e_{hr}$$

under the following assumptions:

(i) The convex closures $C_h(n)$ of the x_{hr} in the k-dimensional space are contained for all n in convex compact bounded spaces C_h which (for different h) are disjoint.

(ii) The functions $f_i(x, \theta)$, $f_{ij}(x, \theta)$, and $f_{ijk}(x, \theta)$ are continuous, bounded functions of x and θ for all $x \varepsilon C_h(n)$ and $\theta \varepsilon S$, where S is a bounded convex space which contains the true θ as an interior point.

(iii) The N by m matrix with elements $f_i(x_\tau, \theta)$ has rank m for all $\theta \varepsilon S$ provided at least m of the vectors x_τ, $\tau = 1, 2, \cdots, N$, are distinct.

Note that the lemma of Section 3 implies

(iii′) The m by m matrix $F_n = (\bar{f}_i(h, \theta))$ has rank m for $\theta \varepsilon S$ and $x_{h\tau}$ satisfying (i).

PROOF OF LEMMA: Suppose that the $\bar{f}_i(h, \theta)$ had a rank $< m$ for some point θ in S and for some set of $x_{h\tau}$. Then we would have a set of u_i with $\sum_{i=1}^m u_i^2 > 0$ and

$$\sum_{i=1}^m u_i \bar{f}_i(h, \theta) = 0 \tag{6.1}$$

for all $h = 1, 2, \cdots, m$. Consider the function $G(x) = \sum_{i=1}^m u_i f_i(x, \theta)$. Now from (6.1) we have for every $h = 1, 2, \cdots, m$ that

$$n^{-1} \sum_\tau G(x_{h\tau}) = n^{-1} \sum_i u_i \sum_\tau f_i(x_{h\tau}, \theta) = \sum_i u_i \bar{f}_i(h, \theta) = 0. \tag{6.2}$$

But (6.2) implies that the m group means of the n values of $G(x_{h\tau})$ are zero for every h. It follows that $\min_\tau G(x_{h\tau}) \leqq 0 \leqq \max_\tau G(x_{h\tau})$. Since $G(x)$ is continuous it must take on the value zero at some point $\bar{x}_h$ in the closure $C_h(n)$. That is, we must have

$$G(\bar{x}_h) = \sum_{i=1}^m u_i f_i(\bar{x}_h, \theta) = 0 \tag{6.3}$$

for $h = 1, 2, \cdots, m$.

Now since the closures $C_h(n)$ are disjoint, Equations (6.3) would contradict Assumption (iii). This proves the Lemma. We now prove

THEOREM 2. *For any $\theta \varepsilon S$, the determinant $|f_i(x_h, \theta)|$, which we denote below by $|F_1|$, has the same sign for any two sets of vectors ${}_1x_h$ and ${}_2x_h$ $(h = 1, 2, \cdots, m)$ with ${}_1x_h \varepsilon C_h$ and ${}_2x_h \varepsilon C_h$.*

PROOF. Suppose ${}_1x_h, {}_2x_h \varepsilon C_h$ for $h = 1, 2, \cdots, m$ and $|F_1({}_1x_h, \theta)| > 0$, $|F_1({}_2x_h, \theta)| < 0$. Then consider the function of q

$$G(q) = |F_1(({}_1x_h(1 - q) + {}_2x_h q), \theta)|.$$

We have $G(0) > 0$ and $G(1) < 0$ and hence, because of the convexity of each C_h, there is a q^* such that $G(q^*) = 0$. Thus

$$|F_1({}_1x_h(1 - q^*) + {}_2x_h q^*, \theta)| = 0$$

which contradicts (iii′) for $n = 1$ since ${}_1x_h(1 - q) + {}_2x_h q$ is in C_h. Next we prove

THEOREM 3. *There is no subsequence F_k of the sequence F_n such that $\lim_{k\to\infty} |F_k| = 0$.*

PROOF. Suppose

$$F_k = n_k^{-1} \sum_{\tau=1}^{n_k} |f_i(x_{h\tau}, \theta)|$$

is such that $\lim_{k\to\infty} |F_k| = 0$. The determinant $|F_k|$ may be expressed as the sum of $(n_k)^m$ determinants, say F_p, where $p = 1, \cdots, (n_k)^m$ corresponding to the

$(n_k)^m$ ways of choosing t from each column. Thus $|F_k|$ is the mean of $(n_k)^m$ determinants each in the form of an F_1. By Theorem 2, all these F_1 values which we may denote by ${}_1F_1$ to ${}_{n_k}F_1$, have the same sign and hence

$$\text{mod } |F_k| \geqq \min (\text{mod } |{}_1F_1|, \cdots, \text{mod } |{}_{n_k}F_1|) = \text{mod } |{}_{l(k)}F_1|.$$

Thus $\lim_{k\to\infty} |{}_{l(k)}F_1| = 0$ which contradicts Theorem 2 and the compactness of the C_h. It follows that $|F_n|$ is bounded away from zero. Next we prove

THEOREM 4. *Let θ^* be any consistent estimate of θ and define*

$$\phi_n(\theta) = -n^{-1}\sigma^{-2} \textstyle\sum_{h\tau} f_i(x_{h\tau}, \theta) f_j(x_{h\tau}, \theta).$$

Then we have

$$n^{-1}(L''(\theta) - \phi_n(\theta)) = o_p(1),$$
$$n^{-1}(L''(\theta^*) - L''(\theta)) = o_p(1),$$

and

$$n^{-1}(L''(\theta^*)) = O_p(1),$$

where $L''(\theta)$ is the m by m matrix of second partial derivatives of the likelihood function

$$L(\theta) = \log \textstyle\prod_{h\tau} (2\pi)^{-\frac{1}{2}}\sigma^{-2} \exp [-\frac{1}{2}\sigma^{-2}(Y_{h\tau} - f(x_{h\tau}, \theta))^2].$$

PROOF. Define

$$z_\tau = \textstyle\sum_{h=1}^{m} [(Y_{h\tau} - f(x_{h\tau}, \theta)) f_{ij}(x_{h\tau}, \theta) - f_i f_j]/\sigma^2$$

so that $n^{-1}L''(\theta) = n^{-1} \sum_\tau z_\tau$. Since

$$Ez_\tau = -\textstyle\sum_h f_i f_j/\sigma^2$$
$$\text{Var } z_\tau = \textstyle\sum_h f_{ij}^2/\sigma^2$$

it follows from Assumption (ii) that Var (z_τ) is bounded and since the z_τ are independent (Loeve, M., 1955, p. 234)

$$\bar{z} - E(\bar{z}) = n^{-1}L''(\theta) - \phi_n(\theta) = o_p(1).$$

Also from Assumption (ii) it follows that $\phi_n(\theta)$ is bounded so that

$$n^{-1}L''(\theta) = O_p(1). \tag{6.4}$$

Denote the element ij of $n^{-1}L''(\theta)$ by $u_n^{(ij)}(y, x, \theta)$. Identify the functions G_n of Corollary 1.2 with the sequence $u_n^{(ij)}(y, x, \theta)$ and also

(6.5)

Corollary 1.2	Theorem 4
y_n	$(y_{11}, x_{11}, y_{12}, x_{12}, \cdots, y_{mn}, x_{mn})$
x_n	θ
z_n	θ^*

The conditions of Corollary 1.2 are satisfied for $s = 0$ since by Assumption (ii) the elements of $n^{-1}L''(\theta)$ are continuous, bounded functions and $(z_n - x_n) = o_p(1)$. Thus

$$u_n^{(ij)}(y, x, \theta^*) - u_n^{(ij)}(y, x, \theta) = o_p(1)$$

or equivalently

$$n^{-1}(L''(\theta^*) - L''(\theta)) = o_p(1) \tag{6.6}$$

and the combination of (6.4) and (6.6) by the rules of algebra concerning O_p and o_p (see Chernoff, 1956, p. 1) obtains

$$n^{-1}L''(\theta^*) = O_p(1).$$

Next we prove

THEOREM 5. *With the elements of the matrix $\phi_n(\theta)$ defined in Theorem 4, the elements of the inverse matrix $\phi_n^{-1}(\theta)$ are bounded.*

PROOF. Suppose the determinant $|\phi_\nu(\theta)| \to 0$ as $\nu \to \infty$ where ν is a subsequence of n, then the lowest characteristic value of $\phi_\nu(\theta)$ would tend to zero, i.e. we would have $\inf_{|u|=1} u'\phi_\nu(\theta)u \to 0$. But $u'\phi_\nu(\theta)u$ is an average of ν values of the form

$$u'[\textstyle\sum_h f_i(x_{h\tau}, \theta)f_j(x_{h\tau}, \theta)]u = \sum_h [\sum_i u_i f_i(x_{h\tau}, \theta)]^2.$$

Hence we have

$$\inf_{|u|=1, x_h \varepsilon C_h} [\textstyle\sum_i u_i f_i(x_h, \theta)] = 0 \qquad \text{for} \quad h = 1, 2, \cdots, m. \tag{6.7}$$

But since the C_h are compact and disjoint and since $|u| = 1$ is compact (6.7) would imply the existence of m distinct vectors x_h and a unitary vector u such that $\sum_i u_i f_i(x_h, \theta) = 0$ for $h = 1, 2, \cdots, m$, which would contradict (iii).

Next we prove

THEOREM 6. *Under Conditions* (i), (ii) and (iii) *of Section* 6 *and Condition* (3.3) *the equations*

$$\bar{Y}_h = \bar{f}(h, \theta^*)$$

have at least one solution θ^*. The present theorem establishes that for any such statistic θ^*

$$\theta^* - \theta = O_p(n^{-\frac{1}{2}}).$$

PROOF. Write $\delta = \theta^* - \theta$ and note that

$$\bar{e}_h = \bar{Y}_h - \bar{f}(h, \theta) = O_p(n^{-\frac{1}{2}}). \tag{6.8}$$

We first prove that for any α with $0 < \alpha < \frac{1}{2}$, $\delta = O_p(n^{-\alpha})$.

From arguments developed in the lemma proved earlier in this section it follows that

$$\inf_{|u| \geqq 1} \textstyle\sum_h [\sum_i u_i \bar{f}_i(h, \theta)]^2 = \lambda^2 > 0.$$

Consider now any parameter value ${}_1\theta$ with $|{}_1\theta - \theta| \geqq \epsilon$ and ${}_1\theta \varepsilon S$ and expand

$\bar{f}(h, {}_1\theta) - \bar{f}(h, \theta)$ in a second order Taylor series. This yields

$$(6.9)\quad \inf_{|{}_1\theta-\theta| \geqq \epsilon} \sum [\bar{f}(h, {}_1\theta) - \bar{f}(h, \theta)]^2 = \inf_{|u| \geqq 1} \sum_h \{\sum_i \epsilon u_i \bar{f}_i(h, \theta) + O(\epsilon^2)\}^2 \geqq \epsilon^2[\lambda^2 + O(\epsilon)].$$

Let now $\epsilon = n^{-\alpha}$ and substitute $\bar{f}(h, {}_1\theta) - \bar{f}(h, \theta) = \bar{f}(h, {}_1\theta) - \bar{Y}_h + \bar{e}_h$ in (6.9). This yields

$$(6.10)\quad \inf_{|{}_1\theta-\theta| \geqq \epsilon} |\bar{Y}_h - \bar{f}(h, {}_1\theta) - \bar{e}_h| \geqq \lambda n^{-\alpha}[1 + O(1)].$$

But since $\bar{e}_h = O_p(n^{-\frac{1}{2}})$ any statistic θ^* which is a solution of $\bar{Y}_h - \bar{f}(h, \theta^*) = 0$ must satisfy with probability approaching one

$$(6.11)\quad |\theta^* - \theta| < \epsilon = n^{-\alpha}$$

as otherwise Equation (6.10) would contradict $\bar{e}_h = O_p(n^{-\frac{1}{2}})$. Equation (6.11) establishes $\theta^* - \theta = O_p(n^{-\alpha})$.

From a second order Taylor expansion of (6.8) we obtain

$$(6.12)\quad \begin{aligned} \bar{e}_h &= \bar{Y}_h - \bar{f}(h, \theta) = \bar{f}(h, \theta^*) - \bar{f}(h, \theta) \\ &= \sum_i \bar{f}_i(h, \theta)\delta_i + \tfrac{1}{2} \sum_{ij} \bar{f}_{ij}(h, \bar{\theta})\delta_i\delta_j \end{aligned}$$

where $\bar{\theta}$ is on the line segment joining θ and θ^* but depends on h. Inverting (6.12) we obtain

$$(6.13)\quad \delta = F_n^{-1}(\bar{e} + \delta' A(h)\delta) = F_n^{-1}(O_p(n^{-\frac{1}{2}}))$$

where $A(h) = (-\frac{1}{2}\bar{f}_{ij}(h, \bar{\theta}))$ and, as before, $F_n = (\bar{f}_i(h, \theta))$.

Now since the smallest characteristic root of F_n exceeds a lower bound λ^2 for all n we have that $F_n^{-1} = O(1)$ and hence from (6.13) that

$$\delta = O_p(n^{-\frac{1}{2}}).$$

7. Asymptotic 100% efficiency of the estimator $\tilde{\theta}$. Consider the statistic $\tilde{\theta} = \theta^* + D$ where the quantity D is defined by the equation

$$(7.1)\quad -L''(\theta^*)D = L'(\theta^*).$$

Let $\hat{\theta}$ be the 'asymptotically efficient statistic' which satisfies $L'(\hat{\theta}) = 0$. For the asymptotic distribution of the maximum likelihood estimator we refer to the literature. We are here merely concerned with proving that the asymptotic variance of $\tilde{\theta}$ agrees with that of $\hat{\theta}$ to order $O(n^{-1})$. It is sufficient for asymptotic efficiency of $\tilde{\theta}$ to show that $n^{+\frac{1}{2}}(\theta^* + D - \hat{\theta}) = o_p(1)$ since it then follows that (Doob, 1935)

$$\lim L(n^{\frac{1}{2}}(\hat{\theta} - \theta)) = \lim L(n^{\frac{1}{2}}(\hat{\theta} - \theta) + n^{\frac{1}{2}}(\theta^* + D - \hat{\theta})).$$

We now identify the vectors y, x, z of Corollary 1.2 with $(y_{11}, x_{11}, \cdots, y_{mn}, x_{mn})$, $\hat{\theta}$, θ^* and let $G_n = n^{-1}L'(\theta)$. The sequence of functions G_n has bounded and continuous second order partial derivatives under Assumption (ii). Also, $(\theta^* - \hat{\theta}) = O_p(n^{-\frac{1}{2}})$ so that $f(n) = n^{-\frac{1}{2}}$ and $\lim f(n) = 0$. For $s = 1$, Corollary

1.2 obtains

$$(7.2)\quad H_n(y, \hat{\theta}, \theta^*, x) = n^{-1}L'(\hat{\theta}) - n^{-1}L'(\theta^*) - n^{-1}L''(\theta^*)(\hat{\theta} - \theta^*) = o_p(n^{-\frac{1}{2}}).$$

Since $L'(\hat{\theta}) = 0$, (7.2) can be written using (7.1)

$$(7.3)\qquad n^{\frac{1}{2}}(\theta^* + D - \hat{\theta}) = n(L''(\theta^*))^{-1}o_p(1).$$

From Theorem 4, $n^{-1}L''(\theta^*) - \phi_n(\theta) = o_p(1)$. Since the elements of $n(L''(\theta^*))^{-1}$ are rational functions of the elements of $n^{-1}L''(\theta^*)$ and $\phi_n^{-1}(\theta)$ was shown to be bounded in Theorem 5, it follows by Slutsky's theorem (Cramér, H., 1945) (a modification of the result is necessary to cover the case where the constants vary with n but satisfy the regularity Assumptions (i) and (ii) and no alteration of the proof is necessary) that

$$n(L''(\theta^*))^{-1} - \phi_n^{-1}(\theta) = o_p(1)$$

and consequently $n(L''(\theta^*))^{-1} = O_p(1)$. Now (7.3) can be written $n^{\frac{1}{2}}(\theta^* + D - \hat{\theta}) = o_p(1)$ which establishes the asymptotic efficiency of $\tilde{\theta}$.

The modified Gauss Newton method employs the corrective vector D^* defined by

$$(7.4)\quad [\textstyle\sum_{hr} f_i(x_{hr}, \theta^*)f_j(x_{hr}, \theta^*)]D^* = \sum_{hr}(Y_{hr} - f(x_{hr}, \theta^*))f_i(x_{hr}, \theta^*).$$

By adding (7.1) and (7.4), we obtain

$$(7.5)\qquad n^{\frac{1}{2}}(D^* - D) = (\sigma^{-2}\phi_n^{-1}(\theta^*))(n^{-1}\textstyle\sum_{hr}(Y - f)f_{ij})(n^{\frac{1}{2}}D).$$

The three terms in parentheses of (7.5) will now be examined one at a time. Let G_n of Corollary 1.2 be $\phi(\theta^*)$ where the correspondence between variables is given by (6.5). Since $(\theta^* - \theta) = O_p(n^{-\frac{1}{2}})$, $f(n) = n^{-\frac{1}{2}}$ and G_n has continuous bounded first order partial derivatives with respect to θ, it follows that for $s = 0$,

$$\phi_n(\theta^*) - \phi_n(\theta) = o_p(1)$$

Again using Slutsky's theorem,

$$\phi_n^{-1}(\theta^*) - \phi_n^{-1}(\theta) = o_p(1)$$

but $\phi_n^{-1}(\theta) = O_p(1)$ from Theorem 5 so that

$$(7.6)\qquad \phi_n^{-1}(\theta^*) = O_p(1).$$

Now we apply Corollary 1.2 to the expression

$$G_n(y, x, \theta^*) = n^{-1}\textstyle\sum_{hr}(y_{hr} - f(x_{hr}, \theta))f_{ij}(x_{hr}, \theta^*)$$

again using the correspondence (6.5) and it follows that

$$G_n(y, x, \theta^*) - G_n(y, x, \theta) = o_p(1).$$

Since $G_n(y, x, \theta)$ is the average of independent normal random variables $\sum_h (Y_{hr} - f(x_{hr}, \theta))f_{ij}(x_{hr}, \theta)$ each having mean zero and bounded variance,

it follows that

$$n^{-1}\sum_{hr}(Y_{hr} - f(x_{hr}, \theta))f_{ij}(x_{hr}, \theta) = o_p(1)$$

and consequently

$$G_n(y, x, \theta^*) = o_p(1). \tag{7.7}$$

Since $n^{\frac{1}{2}}(\theta^* + D) = O_p(1)$ and $n^{\frac{1}{2}}(\theta^*) = O_p(1)$, it follows that $n^{\frac{1}{2}}D = O_p(1)$ The right hand side of (7.5) now can be written as follows by using (7.6), (7.7)

$$n^{\frac{1}{2}}(D^* - D) = O_p(1)o_p(1)O_p(1) = o_p(1).$$

Thus

$$\lim L(n^{\frac{1}{2}}(\theta^* + D - \hat{\theta})) = \lim L(n^{\frac{1}{2}}(\theta^* + D^* - \hat{\theta}))$$

so that the correction vector D^* could be used instead of D and still retain the asymptotic properties of $\hat{\theta}$.

Since it has been shown that $(\tilde{\theta} - \theta) = O_p(n^{-\frac{1}{2}})$, a correction of $\tilde{\theta}$ by D defined by $-L''(\tilde{\theta})D = L'(\tilde{\theta})$ will produce an asymptotic 100% efficient estimate of θ. Thus each step in the Gauss Newton iterative procedure is consistent and asymptotically 100% efficient, provided we *fix* an upper bound for the maximum number of steps as $n \to \infty$. For all applications of the Gauss Newton iteration it is completely satisfactory to assume that the number of steps is held below finite, although possibly large, upper bound.

TABLE 1

Data

t	h	r	x_t	e_t	exp $(-x_t)$	Y_t
1	1	1	.04	−0.84	.96 079	.12 079
2	1	2	.09	1.65	.91 393	2.56 393
3	1	3	.14	−0.38	.86 936	.48 936
4	1	4	.19	−0.38	.82 696	.44 696
5	1	5	.24	−0.74	.78 663	.04 663
6	1	6	.29	0.20	.74 826	.94 826
7	1	7	.34	−1.13	.71 177	−.41 823
8	1	8	.39	0.31	.67 706	.98 706
9	1	9	.44	−0.33	.64 404	.31 404
10	1	10	.49	0.18	.61 263	.79 263
11	2	1	.54	−0.99	.58 275	−.40 725
12	2	2	.59	−0.64	.55 433	−.08 567
13	2	3	.64	−0.26	.52 729	.26 729
14	2	4	.69	0.00	.50 158	.50 158
15	2	5	.74	1.75	.47 711	2.32 711
16	2	6	.79	−1.89	.45 384	−1.43 616
17	2	7	.84	−0.88	.43 171	−.44 829
18	2	8	.89	−0.64	.41 066	−.22 934
19	2	9	.94	−0.74	.39 063	−.34 937
20	2	10	.99	1.08	.37 158	1.45 158

TABLE 2

Computation of θ^*

Iteration Number	${}_i\theta_1$	${}_i\theta_2$	${}_i\delta_1$	${}_i\delta_2$	$\bar{Q}({}_i\theta)$
1	1.05 34	−1.98 88	.05 377	−.98 88	.85 110
2	1.18 86	−2.68 68	.13 524	−.69 79	.02 164
3	1.23 76	−2.87 04	.04 902	−.18 35	4.13E-5
4	1.24 04	−2.87 88	.00 281	−.00 847	1.8E-10
5	1.24 04	−2.87 88	5.7E-6	−1.4E-5	1.7E-20
6	1.24 04	−2.87 88	5.2E-11	−1.2E-10	0

TABLE 3

Computation of $\tilde{\theta}$ and $\dot{\theta}$

Iteration Number	${}_i\theta_1$	${}_i\theta_2$	${}_i\delta_1$	${}_i\delta_2$	$Q({}_i\theta)$
1	1.09 69	−2.59 22	−.14 353	.28 660	15.512
2	1.09 51	−2.56 29	−.00 185	.02 933	15.512
3	1.09 45	−2.56 06	−.00 055	.00 235	15.512
4c	1.09 45	−2.56 04	−.00 005	.00 021	15.512

8. A numerical example. In order to illustrate the algorithms described in Sections 2 to 3 we use the exponential law with zero asymptotes.

$$Y_{h\tau} = \theta_1 \exp(\theta_2 x_{h\tau}) + e_{h\tau}$$

for $\theta_1 = 1$ and $\theta_2 = -1$. Using a table of random normal deviates $N(0, 1)$ for the $e_{h\tau}$ and the equidistant series of x-values $x_{h\tau} = (.04) + (\tau - 1)(.05) + .5(h - 1)$ for $h = 1, 2$ and $\tau = 1, \cdots, 10$, we obtained the data shown in Table 1.

The linear equations in δ of the modified Gauss Newton method for θ^* are

$$\sum_h (\bar{Y}_h - \bar{f}(h, {}_0\theta)\bar{f}_1(h, {}_0\theta) = \delta_1 \sum_h \bar{f}_1^2 + \delta_2 \sum_h \bar{f}_1 \bar{f}_2$$
$$\sum_h (\bar{Y}_h - \bar{f}(h, {}_0\theta)\bar{f}_2(h, {}_0\theta) = \delta_1 \sum_h \bar{f}_1 \bar{f}_2 + \delta_2 \sum_h \bar{f}_2^2$$

where

$$\bar{Y}_h = (\tfrac{1}{10}) \sum_{\tau=1}^{10} Y_{h\tau}$$
$$\bar{f}(h, {}_0\theta) = (\tfrac{1}{10}) \sum_{\tau=1}^{10} {}_0\theta_1 \exp({}_0\theta_2 x_{h\tau})$$
$$\bar{f}_1(h, {}_0\theta) = (\tfrac{1}{10}) \sum_{\tau=1}^{10} \exp({}_0\theta_2 x_{h\tau})$$
$$\bar{f}_2(h, {}_0\theta) = (\tfrac{1}{10}) \sum_{\tau=1}^{10} {}_0\theta_1 x_{h\tau} \exp({}_0\theta_2 x_{h\tau}).$$

The form $\bar{Q}(\theta)$ is evaluated at ${}_0\theta + \delta$, ${}_0\theta + .9\delta$, $\cdots$, ${}_0\theta + .1\delta$, ${}_0\theta + .09\delta$, $\cdots$ accepting ${}_1\theta$ as the first value such that $\bar{Q}(\theta)$ is reduced. The values obtained in the iterative computation of θ^* are shown in Table 2.

The computed θ^* is then taken as the starting point in the solution of

$$\sum_t (Y_t - f(x_t, \theta^*) f_1(x_t, \theta^*) = \delta_1 \sum_t f_1^2 + \delta_2 \sum_t f_1 f_2$$
$$\sum_t (Y_t - f(x_t, \theta^*) f_2(x_t, \theta^*) = \delta_1 \sum_t f_1 f_2 + \delta_2 \sum_t f_2^2.$$

The form $Q(\theta)$ is evaluated at $\theta = \theta^* + 2^{-k}\delta$ for $k = 0, 1, \cdots$, until a value is obtained for which $Q(\theta)$ is reduced. This value then becomes the starting point for the next iteration in the solution of $\dot{\theta}$. The calculation of $\dot{\theta}$ is shown in Table 3.

Whilst $\tilde{\theta}_1 = 1.0969$ and $\dot{\theta}_1 = 1.0945$ are fairly close to the true parametric value $\theta_1 = 1$. The discrepancy for $\tilde{\theta}_2 = -2.5922$ and $\dot{\theta}_2 = -2.5604$ from $\theta_2 = -1$ appears to be considerably larger. Approximate large sample formulas for the variances and covariances of $\dot{\theta}_1$, $\dot{\theta}_2$ (or of $\tilde{\theta}_1$, $\tilde{\theta}_2$) reveal the following results:

$$\text{Var } \dot{\theta}_1 \doteq 0.312, \quad \text{Var } \dot{\theta}_2 \doteq 1.68, \quad \text{and} \quad \text{Cov } \dot{\theta}_1, \dot{\theta}_2 = -.53$$

indicating (among others) that the difference $|\dot{\theta}_2 - \theta_2| = 1.56$ is about 1.2 times its estimated standard deviation.

9. Acknowledgment. The authors wish to thank the referee for his criticisms and suggestions for improving the original version of the paper.

REFERENCES

[1] CHERNOFF, H. (1956). Large sample theory: parametric case. *Ann. Math. Statist.* **27** 1–22.

[2] CRAMÉR, H. (1945). *Mathematical Methods of Statistics.* Princeton Univ. Press, Princeton.

[3] DOOB, J. L. (1935). Limiting distributions of certain statistics. *Ann. Math. Statist.* **6** 160–169.

[4] HARTLEY, H. O. (1961). Modified Gauss Newton method for the fitting of nonlinear regression functions. *Technometrics* **3** 269–280.

[5] HUZURBAZAR, V. S. (1948). The likelihood equation consistency and the maxima of the likelihood function. *Ann. Eugenics* **14** 185–198.

[6] LOÈVE, M. (1955). *Probability Theory.* Van Nostrand, Princeton.

[7] MANN, H. B. and WALD, A. (1943). On stochastic limit and order relationships. *Ann. Math. Statist.* **13** 217–226.