The Annals of Mathematical Statistics
1969, Vol. 40, No. 2, 633–643

ASYMPTOTIC PROPERTIES OF NON-LINEAR LEAST SQUARES ESTIMATORS[1]

By Robert I. Jennrich

University of California, Los Angeles

1. Introduction. The purpose of this paper is to set forth conditions for the consistency and asymptotic normality of least squares estimators of non-linear parameters and to show that the Gauss-Newton iteration method of estimation is asymptotically numerically stable. Assume that

(a) *a sequence of real valued responses y_t has the structure*

$$y_t = f_t(\theta_0) + e_t, \qquad t = 1, 2, 3, \cdots,$$

where the f_t are known continuous functions on a compact subset Θ of a Euclidean space and the e_t are independent identically distributed errors with zero mean and finite variance $\sigma^2 > 0$. (The values of θ_0 and σ^2 are unknown.)

Any vector $\hat{\theta}$ in Θ which minimizes

$$(1) \qquad Q_n(\theta) = n^{-1} \sum_{t=1}^{n} (y_t - f_t(\theta))^2$$

will be called a least squares estimate of θ_0 based on the first n values of y_t. It is natural to ask if, under assumption (a), there always exists a least squares estimator, i.e., a measurable function of $y_1, \cdots, y_n$ whose values are least squares estimates. As will be shown in Lemma 2, the answer is "yes." We seek conditions which will guarantee the consistency and asymptotic normality of a sequence of least squares estimators.

The problem arises already in the case of elementary linear regression. Let $x_1, x_2, \cdots$ be a sequence of real numbers and let

$$y_t = \beta x_t + e_t, \qquad t = 1, \cdots, n.$$

Under the assumption of normally distributed errors the least squares estimator $\hat{\beta}$ of β is also normally distributed. Most of us believe that for large n the same result holds, at least approximately, even if the errors are not normally distributed. But this is not automatic. If for example $x_t = 1/t$ and the e_t have a uniform distribution, the estimator $\hat{\beta}$ fails to be consistent and fails to be asymptotically normally distributed. On the other hand if $x_t \equiv 1$, and the e_t have any distribution satisfying assumption (a), then $\hat{\beta}$ is consistent and asymptotically normally distributed.

Results for linear least squares estimation are given by Eicker [3] and by Grenander and Rosenblatt ([4], p. 244). These authors make weaker assumptions

Received 14 September 1966; revised 9 October 1968.

[1] This research was supported in part by the Air Force Office of Scientific Research Grant AF-AFOSR-4.59-66 and by the National Institutes of Health Grant FR-3.

about the errors e_t. Relaxing the assumption that the errors are identically distributed Eicker gives conditions on the functions f_t for the asymptotic normality of the estimators. Assuming the errors are weakly stationary Grenander and Rosenblatt give conditions for asymptotic efficiency by which they mean conditions under which the least squares estimates have the same limiting variance as the Gauss-Markov estimates.

For the non-linear case Theorems 6 and 7 set forth sufficient conditions for the consistency and asymptotic normality of a sequence of least squares estimators. Under the same conditions Theorem 8 shows that the Gauss-Newton iteration procedure is asymptotically numerically stable and that all sequences of least squares estimators are tail equivalent ([7], p. 233).

Under the assumption of normally distributed errors least squares estimators become maximum likelihood estimators. In this context Hartley and Booker [5] have studied the asymptotic efficiency of an estimator $\hat{\theta}$ obtained by applying a finite number of steps of the Gauss-Newton non-linear estimation procedure to a consistent starting estimate θ^*. They show that if an asymptotically efficient maximum likelihood estimator exists then $\hat{\theta}$ is also asymptotically efficient. As our first example shows, however, asymptotically efficient maximum likelihood estimators do not always exist. Our first example also gives a counter example for Hartley and Booker's Theorem 6 which claims the consistency of their estimator θ^*. Results related to those of Hartley and Booker are referred to at the ends of Sections 4 and 5.

2. Tail products. Unless otherwise specified all limits will be taken as $n \to \infty$. Let $x = (x_t)$ and $y = (y_t)$ be two sequences of real numbers and let $(x, y)_n = n^{-1} \sum_{t=1}^{n} x_t y_t$. If $(x, y)_n$ converges to a real number its limit (x, y) will be called the tail product of x and y. Let g and h be two sequence valued functions on Θ. If $(g(\alpha), h(\beta))_n \to (g(\alpha), h(\beta))$ uniformly for all α and β in Θ, let $[g, h]$ denote the function on $\Theta \times \Theta$ which takes $\langle \alpha, \beta \rangle$ into $(g(\alpha), h(\beta))$. This function will be called the tail cross product of g and h. Note that if in addition the components of g and h are continuous then $[g, h]$, as a uniform limit of continuous functions, is also continuous.

For example let $\mathfrak{X}$ be a Euclidean space, let $\tilde{g}$ and $\tilde{h}$ be bounded and continuous functions on $\mathfrak{X} \times \Theta$, and let $g_t(\theta) = \tilde{g}(x_t, \theta)$ and $h_t(\theta) = \tilde{h}(x_t, \theta)$ where $x_1, x_2, \cdots$ is a sequence of vectors in $\mathfrak{X}$ whose sample distribution function F_n approaches a distribution function F completely. ($F_n(x) = k/n$ where k is the number of points $x_1, \cdots, x_n$ less than or equal to x.) Then

$$(g(\alpha), h(\beta))_n = \int \tilde{g}(x, \alpha)\tilde{h}(x, \beta)\, dF_n(x).$$

According to Theorem 1 below, which is a simple extension of the Helly-Bray theorem,

$$(g(\alpha), h(\beta))_n \to \int \tilde{g}(x, \alpha)\tilde{h}(x, \beta)\, dF(x)$$

uniformly for all α and β in Θ. Hence the tail cross product $[g, h]$ exists.

As a second example let $x_1, x_2, \cdots$ be a random sample from a probability distribution function F on $\mathfrak{X}$. Then the sample distribution function $F_n \to F$ with probability one and hence $[g, h]$ exists with probability one. A stronger result however can be obtained. With regard to $\tilde{g}$ and $\tilde{h}$ assume only that $\tilde{g}(x, \theta)$ and $\tilde{h}(x, \theta)$ are measurable in x, continuous in θ, and bounded uniformly in θ by a function of x which is square integrable with respect to F. Then by Theorem 2 below, which is a simple extension of the strong law of large numbers,

$$(g(\alpha), h(\beta))_n \to \int \tilde{g}(x, \alpha)\tilde{h}(x, \beta)\, dF(x)$$

uniformly for all α and β in Θ for almost every sequence $x_1, x_2, \cdots$. Thus $[g, h]$ exists almost surely.

To prove Theorem 1 we need the following lemma.

LEMMA 1. *If g is a real valued function which is continuous on the Cartesian product $\mathfrak{X} \times \mathfrak{Y}$ of two Euclidean spaces and if Y is a bounded subset of $\mathfrak{Y}$ then $\sup_{y \in Y} g(x, y)$ is a continuous function of x.*

PROOF. Let $\bar{Y}$ be the closure of Y. Since g is uniformly continuous on compact subsets of $\mathfrak{X} \times \mathfrak{Y}$, for any x_0 in $\mathfrak{X}$ and any $\epsilon > 0$ there is a $\delta > 0$ such that

$$g(x_0, y) - \epsilon < g(x, y) < g(x_0, y) + \epsilon$$

for all y in $\bar{Y}$ and $|x - x_0| < \delta$. Thus, for $|x - x_0| < \delta$,

$$\sup_{y \in Y} g(x_0, y) - \epsilon \leqq \sup_{y \in Y} g(x, y) \leqq \sup_{y \in Y} g(x_0, y) + \epsilon.$$

Since ϵ is arbitrary $\sup_{y \in Y} g(x, y)$ is continuous at $x = x_0$. Since x_0 is arbitrary it is continuous for all x in $\mathfrak{X}$.

THEOREM 1. *If $\mathfrak{X}$ is a Euclidean space, if Θ is a compact subset of a Euclidean space, if g is a bounded and continuous function on $\mathfrak{X} \times \Theta$, and if $F_1, F_2, \cdots$ are distribution functions on $\mathfrak{X}$ which converge completely to a distribution function F then*

$$\int g(x, \theta)\, dF_n(x) \to \int g(x, \theta)\, dF(x)$$

uniformly for all θ in Θ.

PROOF. Let

$$h_n(\theta) = \int g(x, \theta)\, dF_n(x) - \int g(x, \theta)\, dF(x)$$

and let N be a neighborhood of a point θ_0 in Θ. (Unless otherwise specified a neighborhood of a point in Θ is a subset of Θ which contains the point and is open in the relative topology on Θ.) Obviously

$$\sup_{\theta \in N} h_n(\theta) \leqq \int \sup_{\theta \in N} g(x, \theta)\, dF_n(x) - \int \inf_{\theta \in N} g(x, \theta)\, dF(x).$$

Since g is bounded $\sup_{\theta \in N} g(x, \theta)$ is a bounded and, by Lemma 1, continuous function of x. It follows from the Helly-Bray theorem that

$$\overline{\lim}_{n \to \infty} \sup_{\theta \in N} h_n(\theta) \leqq \int (\sup_{\theta \in N} g(x, \theta) - \inf_{\theta \in N} g(x, \theta))\, dF(x).$$

For each x the integrand above approaches zero as the diameter of N approaches

zero and hence, by the dominated convergence theorem, for any $\epsilon > 0$ there is a neighborhood N of θ_0 such that

$$\overline{\lim}_{n\to\infty} \sup_{\theta\in N} h_n(\theta) < \epsilon.$$

It follows that Θ may be covered by such neighborhoods and, since it is compact, by a finite collection of such neighborhoods. Thus $h_n(\theta) < 2\epsilon$ for all θ when n is sufficiently large or, in terms of g,

$$\int g(x, \theta)\, dF_n(x) < \int g(x, \theta)\, dF(x) + 2\epsilon$$

for all θ when n is sufficiently large. The same argument applied to $-g$ gives

$$\int g(x, \theta)\, dF_n(x) > \int g(x, \theta)\, dF(x) - 2\epsilon$$

for all θ when n is sufficiently large. The theorem follows from the fact that ϵ is arbitrary.

Theorem 2, given by Mickey ([8], p. 40), may be proved in a manner similar to that of Theorem 1.

THEOREM 2. *Let g be a function on $\mathfrak{X} \times \Theta$ where $\mathfrak{X}$ is a Euclidean space and Θ is a compact subset of a Euclidean space. Let $g(x, \theta)$ be a continuous function of θ for each x and a measurable function of x for each θ. Assume also that $|g(x, \theta)| \leqq h(x)$ for all x and θ, where h is integrable with respect to a probability distribution function F on $\mathfrak{X}$. If $x_1, x_2, \cdots$ is a random sample from F then for almost every sequence (x_t)*

$$n^{-1} \sum_{t=1}^{n} g(x_t, \theta) \to \int g(x, \theta)\, dF(x)$$

uniformly for all θ in Θ.

For any sequence x of real numbers let $|x|_n = (x, x)_n^{\frac{1}{2}}$ and when it exists call $|x| = (x, x)^{\frac{1}{2}}$ the tail norm of x. Note that if $[g, h]$, $[g, g]$, and $[h, h]$ exist then $|g(\alpha) - h(\beta)|_n \to |g(\alpha) - h(\beta)|$ uniformly for all α and β in Θ.

3. Tail products and random samples. The following result, which is a strong law of large numbers follows directly from a theorem by Chow [1]. Let $e = (e_t)$.

THEOREM 3. *If e satisfies assumption* (a) *and if the tail norm of a sequence x of real numbers exists then $(x, e)_n \to 0$ for almost every e.*

This theorem extends to the following.

THEOREM 4. *If e satisfies assumption* (a), *if g is a sequence of continuous functions on Θ, and if $[g, g]$ exists, then for almost every e, $(g(\theta), e)_n \to 0$ uniformly for all θ in Θ.*

PROOF. Since g is continuous and $|g(\alpha) - g(\beta)|_n$ converges uniformly in α there is, for every $\epsilon > 0$ and every β in Θ, a neighborhood N of β such that $|g(\alpha) - g(\beta)|_n < \epsilon$ for all α in N and all n sufficiently large. It follows from Theorem 3 and the inequality

$$|(g(\alpha), e)_n| \leqq |g(\alpha) - g(\beta)|_n |e|_n + |(g(\beta), e)_n|,$$

that for almost every e, Θ is covered by neighborhoods N such that $|(g(\theta), e)_n| < \epsilon$ for all θ in N, when n is sufficiently large. Since Θ is compact there is a finite collection of such neighborhoods which cover Θ and hence for almost every

e, $|(g(\theta), e)_n| < \epsilon$ for all θ in Θ, when n is sufficiently large. The theorem follows from the fact that ϵ is arbitrary.

The following form of the central limit theorem follows easily from Theorem 2 of Eicker ([2], p. 444).

THEOREM 5. *If e satisfies assumption* (a) *and the tail norm of a sequence x of real numbers exists, then*

$$\mathfrak{L}(n^{\frac{1}{2}}(x, e)_n) \to \mathfrak{N}(0, |x|^2\sigma^2).$$

Using **2c.4** (xi) of Rao ([9], p. 103), this theorem may be extended, as follows, to the multivariate case.

COROLLARY 1. *If e satisfies assumption* (a) *and if all possible tail products of the sequences $x_1, \cdots, x_p$ of real numbers exist, then*

$$(n^{\frac{1}{2}}((x_1, e)_n, \cdots, (x_p, e)_n)) \to \mathfrak{N}(0, \hat{\sigma}^2((x_i, x_j))).$$

4. The asymptotic normality of least squares estimators. The following lemma shows that assumption (a) is sufficient to guaranty the existence of (measurable) least squares estimators in the non-linear case.

LEMMA 2. *Let Q be a real valued function on $\Theta \times Y$ where Θ is a compact subset of a Euclidean space and Y is a measurable space. For each θ in Θ let $Q(\theta, y)$ be a measurable function of y and for each y in Y a continuous function of θ. Then there exists a measurable function $\hat{\theta}$ from Y into Θ such that for all y in Y,*

$$Q(\hat{\theta}(y), y) = \inf_\theta Q(\theta, y).$$

PROOF. Let (Θ_n) be an increasing sequence of finite subsets of Θ whose limit is dense in Θ. For each n there is a measurable function $\tilde{\theta}_n$ from Y into Θ_n such that

$$Q(\tilde{\theta}_n(y), y) = \inf_{\theta \in \Theta_n} Q(\theta, y)$$

for all y in Y. Let $\tilde{\theta}_{n1}$ denote the first component of $\tilde{\theta}_n$. Let $\hat{\theta}_1 = \underline{\lim}_n \tilde{\theta}_{n1}$ and note that $\hat{\theta}_1$ is measurable. For each y in Y there is a subsequence $(\tilde{\theta}_{n_i}(y))$, of $(\tilde{\theta}_n(y))$, which converges to a point $\bar{\theta}$ in Θ of the form $(\hat{\theta}_1(y), \bar{\theta}_2, \cdots, \bar{\theta}_p)$.

$$\begin{aligned}\inf_{(\theta_1, \cdots, \theta_p) \in \Theta} Q((\hat{\theta}_1(y), \theta_2, \cdots, \theta_p), y) &\leqq Q(\bar{\theta}, y) = \lim_i Q(\tilde{\theta}_{n_i}(y), y) \\ &= \lim_i \inf_{\theta \in \Theta_{n_i}} Q(\theta, y) = \inf_\theta Q(\theta, y).\end{aligned}$$

The last equality follows from the fact that $\lim_n \Theta_n$ is dense in Θ. Thus

$$\inf_{(\theta_1, \cdots, \theta_p) \in \Theta} Q((\hat{\theta}_1(y), \theta_2, \cdots, \theta_p), y) = \inf_\theta Q(\theta, y)$$

for all y in Y. Let

$$Q'((\theta_1, \cdots, \theta_p), y) = Q((\hat{\theta}_1(y), \theta_2, \cdots, \theta_p), y),$$

then $Q'(\theta, y)$ is a continuous function of θ for all y in Y and a measurable function of y for all θ in Θ. Applying the same argument to Q' that was applied to Q gives a measurable real valued function $\hat{\theta}_2$ such that

$$\inf_{(\theta_1, \cdots, \theta_p) \in \Theta} Q((\hat{\theta}_1(y), \hat{\theta}_2(y), \theta_3, \cdots, \theta_p), y) = \inf_\theta Q(\theta, y)$$

for all y in Y. Continuing in this manner produces measurable real valued functions $\hat{\theta}_1, \cdots, \hat{\theta}_p$ such that

$$Q((\hat{\theta}_1(y), \cdots, \hat{\theta}_p(y)), y) = \inf_\theta Q(\theta, y)$$

for all y in Y. Thus $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_p)$ is a measurable function from Y into Θ with the required property.

Lemma 2 together with the familiar mean value theorem from advanced calculus gives rise to the following mean value theorem for random functions.

LEMMA 3. *Let f be a real valued function on $Y \times \Theta$ where Y is a measurable space and Θ is a convex compact subset of a Euclidean space. For each θ in Θ let $f(y, \theta)$ be a measurable function of y and for each y in Y a continuously differentiable function of θ. Let θ_1 and θ_2 be measurable functions from Y into Θ. Then there exists a measurable function $\bar{\theta}$ from Y into Θ such that*

(i) $f(y, \theta_1(y)) - f(y, \theta_2(y)) = (\partial/\partial\theta)f(y, \bar{\theta}(y))(\theta_1(y) - \theta_2(y))$,

(ii) $\bar{\theta}(y)$ *lies on the segment joining $\theta_1(y)$ and $\theta_2(y)$,*

for all y in Y.

PROOF. Let $D(y, \theta)$ denote the distance from θ to the segment joining $\theta_1(y)$ and $\theta_2(y)$ and let

$$Q(\theta, y) = |f(y, \theta_1(y)) - f(y, \theta_2(y)) - (\partial/\partial\theta)f(y, \theta)(\theta_1(y) - \theta_2(y))| + D(y, \theta).$$

Viewing θ_1 and θ_2 as pointwise limits of sequences of simple functions and $(\partial/\partial\theta)f(y, \theta)$ as the limit of a sequence of difference quotients shows that $Q(\theta, y)$ is a measurable function of y for each θ in Θ. Since $Q(\theta, y)$ is clearly a continuous function of θ for each y in Y, Lemma 2 applies and hence there exists a Θ-valued measurable function $\bar{\theta}$ on Y with the property that for each y, $\bar{\theta}(y)$ minimizes $Q(\theta, y)$ viewed as a function of θ. Moreover it follows from the mean value theorem of calculus that when $\theta = \bar{\theta}(y)$ the first and second terms in the definition of $Q(\theta, y)$ are zero. The assertions of the theorem follow at once.

In addition to the assumptions on the functions f_t given in (a) assume that

(b) *the tail cross product of $f = (f_t)$ with itself exists and that*

$$Q(\theta) = |f(\theta) - f(\theta_0)|^2 \text{ has a unique minimum at } \theta = \theta_0.$$

Note that under assumptions (a) and (b), Q is continuous.

THEOREM 6. *Let $(\hat{\theta}_n)$ be a sequence of least squares estimators. Under assumptions* (a) *and* (b), *$\hat{\theta}_n$ and $\hat{\sigma}_n^2 = Q_n(\hat{\theta}_n)$ are strongly consistent estimators of θ_0 and σ^2.*

PROOF. It follows from the strong law of large numbers and Theorem 4 that for almost every e, $|e| = \sigma$ and $(f(\theta) - f(\theta_0), e)_n \to 0$ uniformly for all θ in Θ. Choose such an e. Then

$$Q_n(\theta) = |f(\theta_0) - f(\theta) + e|_n^2 \to |f(\theta_0) - f(\theta)|^2 + \sigma^2 = Q(\theta) + \sigma^2$$

uniformly for θ in Θ. Let $\theta_n = \hat{\theta}_n(f(\theta_0) + e)$, let θ' be a limit point of the sequence (θ_n) and let (θ_{n_t}) be any subsequence which converges to θ'. By the continuity of Q and the uniform convergence of Q_n to $Q + \sigma^2$, $Q_{n_t}(\theta_{n_t}) \to Q(\theta') + \sigma^2$ as

$t \to \infty$. Since θ_{n_t} is a least squares estimate, $Q_{n_t}(\theta_{n_t}) \leqq Q_{n_t}(\theta_0) = |e|^2_{n_t}$. It follows by letting $t \to \infty$ that $Q(\theta') + \sigma^2 \leqq \sigma^2$. Hence $Q(\theta') = 0$. Since Q has a unique minimum at θ_0, $\theta' = \theta_0$. Thus $\theta_n \to \theta_0$. Since this result holds for almost every e, $\hat{\theta}_n \to \theta_0$ and $\hat{\sigma}_n{}^2 = Q_n(\hat{\theta}_n) \to Q(\theta_0) + \sigma^2 = \sigma^2$ almost surely.

To establish the asymptotic normality of a sequence of least squares estimators we need the derivatives

$$f'_{ti}(\theta) = (\partial/\partial\theta_i)f_t(\theta) \quad \text{and} \quad f''_{tij}(\theta) = (\partial/\partial\theta_i)(\partial/\partial\theta_j)f_t(\theta).$$

Let $f_i' = (f'_{ti})$, $f''_{ij} = (f''_{tij})$ and assume that for $i, j = 1, \cdots, p$,

(c) *the derivatives f'_{ti} and f''_{tij} exist and are continuous on Θ and that all tail cross products of the form $[g, h]$, where $g, h = f, f_i', f''_{ij}$, exist.*

For each θ in Θ let

$$a_{nij}(\theta) = (f_i'(\theta), f_j'(\theta))_n, \qquad \mathbf{a}_n(\theta) = (a_{nij}(\theta)), \tag{2}$$

$$a_{ij}(\theta) = (f_i'(\theta), f_j'(\theta)), \qquad \mathbf{a}(\theta) = (a_{ij}(\theta)),$$

and assume that

(d) *the true parameter vector θ_0 is an interior point of Θ and the matrix $\mathbf{a}(\theta_0)$ is non-singular.*

The following theorem provides conditions for the asymptotic normality of a sequence of least squares estimators.

THEOREM 7. *Let $(\hat{\theta}_n)$ be a sequence of least squares estimators of θ_0. Under assumptions* (a) *through* (d)

$$\mathfrak{L}(n^{\frac{1}{2}}(\hat{\theta}_n - \theta_0)) \to \mathfrak{N}(0, \sigma^2\mathbf{a}^{-1}(\theta_0)).$$

Moreover $\mathbf{a}_n(\hat{\theta}_n)$ is a strongly consistent estimator of $\mathbf{a}(\theta_0)$.

PROOF. Since $\hat{\theta}_n \to \theta_0$ almost everywhere there is a sequence $(\tilde{\theta}_n)$ which is tail-equivalent to $(\hat{\theta}_n)$ and such that each $\tilde{\theta}_n$ takes its values in a convex compact neighborhood of θ_0 which is interior to Θ. Letting $\tilde{\theta}_n = (\tilde{\theta}_{n1}, \cdots, \tilde{\theta}_{np})$ and $\theta_0 = (\theta_{01}, \cdots, \theta_{0p})$ it follows from Lemma 3 that there exists a measurable Θ-valued function $\bar{\theta}_n$ such that

$$\begin{aligned}(f_i'(\tilde{\theta}_n), y - f(\tilde{\theta}_n))_n &- (f_i'(\theta_0), y - f(\theta_0))_n \\ &= \textstyle\sum_{j=1}^p ((f''_{ij}(\bar{\theta}_n), y - f(\tilde{\theta}_n))_n - (f_i'(\bar{\theta}_n), f_j'(\bar{\theta}_n))_n)(\tilde{\theta}_{nj} - \theta_{0j})\end{aligned}$$

where $|\bar{\theta}_n - \theta_0| \leqq |\tilde{\theta}_n - \theta_0|$. Multiplying by $n^{\frac{1}{2}}$ and using assumption (a) gives,

$$\begin{aligned}n^{\frac{1}{2}}(f_i'(\theta_0), e)_n = &\textstyle\sum_{j=1}^p ((f_i'(\bar{\theta}_n), f_j'(\bar{\theta}_n))_n \\ &- (f''_{ij}(\bar{\theta}_n), y - f(\tilde{\theta}_n))_n)n^{\frac{1}{2}}(\tilde{\theta}_{nj} - \theta_{0j}) \\ &+ n^{\frac{1}{2}}(f_i'(\tilde{\theta}_n), y - f(\tilde{\theta}_n))_n.\end{aligned} \tag{3}$$

Note that the last term is a multiple of $\partial/\partial\theta_i Q_n(\tilde{\theta}_n)$. Since, almost surely, $\tilde{\theta}_n = \hat{\theta}_n$ when n is sufficiently large and since $\partial/\partial\theta_i Q_n(\hat{\theta}_n) = 0$ whenever the value of $\hat{\theta}_n$ is an interior point of Θ it follows from the strong convergence of $\hat{\theta}_n$ to θ_0 and from Theorem 4 that the inner products on the right converge to $a_{ij}(\theta_0)$,

0, and 0 almost surely. By Corollary 1 of Theorem 5 the joint distribution of the inner products on the left, i.e., $n^{\frac{1}{2}}(f_1'(\theta_0), e)_n, \cdots, n^{\frac{1}{2}}(f_p'(\theta_0), e)_n$, approaches $\mathfrak{N}(0, \sigma^2\mathbf{a}(\theta_0))$. Since $\mathbf{a}(\theta_0)$ is non-singular

$$\mathfrak{L}(n^{\frac{1}{2}}(\tilde{\theta}_n - \theta_0)) \to \mathfrak{N}(0, \sigma^2\mathbf{a}^{-1}(\theta_0)).$$

The first assertion of the theorem follows from the fact that $(\hat{\theta}_n)$ is tail-equivalent to $(\tilde{\theta}_n)$. The strong consistency of $\mathbf{a}_n(\hat{\theta}_n)$ as an estimator of $\mathbf{a}(\theta_0)$ follows from the uniform convergence of $\mathbf{a}_n$ to $\mathbf{a}$ and the strong convergence of $\hat{\theta}_n$ to θ_0.

Under assumptions (a) through (d) a sequence of least squares estimators will be asymptotically normally distributed but it will not, in general, be asymptotically efficient. If the errors e_t are normally distributed, least squares estimators become maximum likelihood estimators and although they may not be based on a random sample (the y_t not being in general identically distributed) they can nevertheless be shown to be asymptotically efficient. This may be done, using the definition of Rao ([9], p. 285), by showing that Rao's condition **(5c.2.6)** follows from equation (3).

5. Asymptotic numerical stability of the Gauss-Newton iteration. The Gauss-Newton iteration procedure [5] consists of linearizing the functions f_t about some point θ and finding a least square fit to the data by minimizing

$$|y - f(\theta) - f_1'(\theta)\, d\theta_1 - \cdots - f_p'(\theta)\, d\theta_p|_n$$

with respect to $d\theta = (d\theta_1, \cdots, d\theta_p)$. Replacing θ by $\theta + d\theta$ the entire process is repeated until it, hopefully, converges. More specifically the iteration is of the form $\theta^{(k+1)} = g_n(\theta^{(k)})$ where the ith component of g_n at $\theta = (\theta_1, \cdots, \theta_p)$ is given by

$$g_{ni}(\theta) = \theta_i + \textstyle\sum_j a_n^{ij}(\theta)(f_j'(\theta), y - f(\theta))_n, \quad i = 1, \cdots, p.$$

Here $a_n^{ij}(\theta)$ is the ijth component of the inverse of the matrix $\mathbf{a}_n(\theta)$. If the sequence $\theta^{(k)}$ converges then its limiting value will be called the Gauss-Newton estimate of θ_0 based on the first n sample values and the starting value $\theta^{(1)}$. In general an iteration procedure of the form $x_{k+1} = h(x_k)$ is said to be numerically stable in a neighborhood N if the iteration converges to the same fixed point whenever the starting value is in N. A sufficient condition for numerical stability in a spherical neighborhood N of a fixed point is that

$$|\mathbf{h}'(x)| \leqq c < 1, \tag{4}$$

for all x in N. Here $\mathbf{h}'(x)$ denotes the differential of h at x and $|\mathbf{h}'(x)|$ denotes the norm, subordinate to the Euclidean norm, of $\mathbf{h}'(x)$ (i.e., $|\mathbf{h}'(x)| = \sup\{|\mathbf{h}'(x)\, dx| : |dx| \leqq 1\}$). This condition will be used to prove that the Gauss-Newton iteration is asymptotically numerically stable.

THEOREM 8. *Let $(\hat{\theta}_n)$ be a sequence of least squares estimators of θ_0 and let assumptions* (a) *through* (d) *hold. Then there exists a neighborhood N of θ_0 such that for almost every y there is an n_y such that the Gauss-Newton iteration will converge to $\hat{\theta}_n(y)$ from any starting value in N whenever $n \geqq n_y$.*

PROOF. Let (a^{ij}) be the inverse of a matrix (a_{ij}) of functions of a real variable α. When it exists

$$(d/d\alpha)a^{ij} = -\sum_{kl} a^{ik}((d/d\alpha)a_{kl})a^{lj}.$$

Making use of this formula we find that when it exists the ikth component of the differential $\mathbf{g}_n'(\theta)$ of g_n at θ is given by

$$(5)\quad g'_{nik}(\theta) = (\partial/\partial\theta_k)g_{ni}(\theta) = \sum_j a_n{}^{ij}(\theta)(f''_{jk}(\theta), y - f(\theta))_n - \sum_{jlm} a_n{}^{il}(\theta)a'_{nlmk}(\theta)a_n{}^{mj}(\theta)(f_j'(\theta), y - f(\theta))_n$$

where

$$a'_{nlmk}(\theta) = (f''_{lk}(\theta), f_m'(\theta))_n + (f_l'(\theta), f''_{mk}(\theta))_n.$$

By Theorems 4 and 6 there exists a set A of measure one such that for all y in A

(i) $(f_j'(\theta), y - f(\theta))_n \to (f_j'(\theta), f(\theta_0) - f(\theta))$,

(ii) $(f''_{jk}(\theta), y - f(\theta))_n \to (f''_{jk}(\theta), f(\theta_0) - f(\theta))$,

(iii) $\hat{\theta}_n(y) \to \theta_0$,

uniformly in θ as $n \to \infty$. Since $\mathbf{a}$ is continuous and non-singular at θ_0 it is non-singular on a compact neighborhood M of θ_0. Since $\mathbf{a}_n \to \mathbf{a}$ uniformly, $\mathbf{a}_n$ is non-singular on M when n is sufficiently large. Thus for any y in A and θ in M,

$$g'_{nik}(\theta) \to \sum_j a^{ij}(\theta)(f''_{jk}(\theta), f(\theta_0) - f(\theta)) - \sum_{jlm} a^{il}(\theta)a'_{lmk}(\theta)a^{mj}(\theta)(f_j'(\theta), f(\theta_0) - f(\theta))$$

where

$$a'_{lmk}(\theta) = (f''_{lk}(\theta), f_m'(\theta)) + (f_l'(\theta), f''_{mk}(\theta)).$$

Moreover, the convergence is uniform in θ. Since, by inspection, the limit of g'_{nik} is zero at θ_0 the differential $\mathbf{g}_n' = (g'_{nik})$ of g_n has norm $|\mathbf{g}_n'| \leqq \frac{3}{4}$ on a spherical neighborhood L of θ_0 when n is sufficiently large.

Let r be the radius of L, let N be a spherical neighborhood of θ_0 of radius $r/2$, and let N_n be a spherical neighborhood of $\hat{\theta}_n(y)$ of radius $3r/4$. Since $\hat{\theta}_n(y) \to \theta_0$, $N \subset N_n \subset L$ when n is sufficiently large. For large n, $\hat{\theta}_n(y)$ is a zero of $(\partial/\partial\theta_j)Q_n(\theta) = -2(f_j'(\theta), y - f(\theta))_n$ and hence a fixed point of g_n. Thus for n sufficiently large

(i) $N \subset N_n$,

(ii) $|\mathbf{g}_n'(\theta)| \leqq \frac{3}{4}$ for all θ in N_n,

(iii) $\hat{\theta}_n(y)$ is a fixed point of g_n.

It follows from (4) that the Gauss-Newton iteration converges to $\hat{\theta}_n(y)$ from every starting value in N when n is sufficiently large. The theorem follows from the fact that this last statement holds for every y in A.

It follows from Theorem 8 that for any given starting value in N the Gauss-Newton iteration will converge to a least squares estimate with probability approaching one as $n \to \infty$. The result is of some interest since there exist examples with well behaved functions f_t for which the Gauss-Newton iteration

will not converge no matter how good the starting value. Theorem 8 implies that such difficulties are not likely to arise when n is sufficiently large. Under the conditions of Theorem 8 and using a fixed starting value in N, let $\bar{\theta}_n(y)$ denote the converged value of the Gauss-Newton iteration if it converges and its starting value if it does not. Then $(\bar{\theta}_n)$ is a sequence of estimators of θ_0 which is tail-equivalent to any sequence $(\hat{\theta}_n)$ of least squares estimators. Thus $(\bar{\theta}_n)$ is, by Theorem 6, strongly consistent and, by Theorem 7, asymptotically normally distributed. Moreover, all sequences of least squares estimators are tail equivalent.

Unfortunately, Theorem 8 requires that the starting value of the Gauss-Newton iteration be sufficiently close to θ_0. H. O. Hartley and Aaron Booker [5] have suggested a solution to this problem which consists in replacing the starting value by a consistent estimator of θ_0. If follows, from Theorem 8, that the probability that such a two-step procedure will produce a least squares estimate approaches one as $n \to \infty$.

It is easy, in theory, to produce consistent estimators of θ_0. Under assumptions (a) through (d) a random search of n values of θ for the one θ_n^* which minimizes Q_n is a strongly consistent estimator. Starting values of this type are sometimes used in practice.

6. Examples. Let $f_t(\theta) = \theta_1 e^{\theta_2 x_t}$ where $\theta = (\theta_1, \theta_2)$ ranges over the unit rectangle Θ and $x_1, x_2, \cdots$ is a bounded sequence of real numbers whose sample distribution function F_n approaches a distribution function F completely. Assume that $\theta_0 = (\alpha_1, \alpha_2)$ is an interior point of Θ and that F is not degenerate. If $e_1, e_2, \cdots$ are independent identically distributed errors with zero mean and finite variance, assumption (a) holds for the model

$$y_t = f_t(\theta_0) + e_t, \qquad t = 1, 2, 3, \cdots.$$

Moreover the derivatives f'_{ti}, f''_{tij} exist and are continuous on Θ. By Theorem 1 all tail cross products of the form $[g, h]$, where $g, h = f, f'_i, f''_{ij}$, exist. Hence, assumption (c) holds. It is easy to show that

$$Q(\theta) = \int (\alpha_1 e^{\alpha_2 x} - \theta_1 e^{\theta_2 x})^2 \, dF(x).$$

This expression is zero if and only if $\alpha_1 e^{\alpha_2 x} = \theta_1 e^{\theta_2 x}$ on a set of x with dF measure one. Since $\alpha_1 \neq 0$ and since F is not degenerate this can happen only when $\alpha_1 = \theta_1$ and $\alpha_2 = \theta_2$. Hence Q has a unique minimum at θ_0. This establishes assumption (b).

Let β_1 and β_2 be any real numbers.

$$\sum_{ij} \beta_i a_{ij}(\theta_0) \beta_j = \int (\beta_1 + \beta_2 \alpha_1 x)^2 e^{2\alpha_2 x} \, dF(x).$$

This expression equals zero only if $\beta_1 + \beta_2 \alpha_1 x = 0$ on a set of x with dF measure one. Since $\alpha_1 \neq 0$ and since F is not degenerate this can happen only when $\beta_1 = \beta_2 = 0$. Thus $\mathbf{a}(\theta_0)$ is non-singular. Assumption (d) follows at once and with it the results of Sections 4 and 5.

Alternatively in the previous example let $x_1, x_2, \cdots$ be a random sample from, say, a normal distribution. In this case the $f_t(\theta)$, in all probability, are not bounded but using Theorem 2 it can be shown that assumptions (a) through (d) hold for almost every sequence $x_1, x_2, \cdots$. Thus the results of Sections 4 and 5 apply not only to fixed sequences f of functions but also to sequences of functions selected randomly from a suitable population of functions.

REFERENCES

[1] CHOW, Y. S. (1967). On the strong law of large numbers for martingales. *Ann. Math. Statist.* **38** 610.

[2] EICKER, F. (1963). Central limit theorems for families of sequences of random variables. *Ann. Math. Statist.* **34** 439–446.

[3] EICKER, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann. Math. Statist.* **34** 447–456.

[4] GRENANDER, U. and ROSENBLATT, M. (1957). *Statistical Analysis of Stationary Time Series*. Wiley, New York.

[5] HARTLEY, H. O. and BOOKER, A. (1965). Non-linear least squares estimation. *Ann. Math. Statist.* **36** 638–650.

[6] JENNRICH, R. I. and SAMPSON, P. F. (1968). Application of stepwise regression to non-linear least squares. *Technometrics* **10** 63–72.

[7] LOÈVE, M. (1955). *Probability Theory*. Van Nostrand, New York.

[8] MICKEY, M. R., MUNDLE, P. M., WALKER, D. N. and GLINSKI, A. M. (1963). *Test criteria for Pearson type III distributions*. Aeronautical Research Laboratories Office of Aerospace Research, Wright Patterson Air Force Base ARL63-100, Dayton, Ohio.

[9] RAO, C. R. (1965). *Linear Statistical Inference and Its Applications*. Wiley, New York.