The Annals of Statistics
1978, Vol. 6, No. 3, 629–642

ON CONDITIONAL LEAST SQUARES ESTIMATION FOR STOCHASTIC PROCESSES

BY LAWRENCE A. KLIMKO AND PAUL I. NELSON

University of Wisconsin and Bucknell University

An estimation procedure for stochastic processes based on the minimization of a sum of squared deviations about conditional expectations is developed. Strong consistency, asymptotic joint normality and an iterated logarithm rate of convergence are shown to hold for the estimators under a variety of conditions. Special attention is given to the widely studied cases of stationary ergodic processes and Markov processes with are asymptotically stationary and ergodic. The estimators and their limiting covariance matrix are worked out in detail for a subcritical branching process with immigration. A brief Monte Carlo study of the performance of the estimators is presented.

1. Introduction and notation. We develop an estimation procedure for dependent observations based on the minimization of a sum of squared deviations about conditional expectations. This approach, which we call "conditional least squares" (CLS), provides a unified treatment of estimation problems for widely used classes of stochastic models. The method is implicit in the observation of Mann and Wald (1943), Durbin (1960) and others that the assumption of normally distributed error terms in autoregressive models renders maximum likelihood estimation equivalent to the minimization of a sum of squares.

The method of CLS is motivated by the interpretation of conditional expectation as an orthogonal projection on L^2. Under a variety of conditions the CLS estimators are shown to be strongly consistent and asymptotically jointly normally distributed. The rate of convergence of the estimators is found to be $(((\log \log n)/n)^{\frac{1}{2}})$. The proofs of these results are presented in a very general setting in Section 2. The assumptions made concern the application of strong laws, central limit theorems and laws of an iterated logarithm to sums of dependent random variables. A wide variety of conditions under which these hold may be found in Stout (1974), McLeish (1974), and Heyde and Scott (1973). These conditions are generally a trade-off among moment assumptions, stationarity, the martingale property, mixing, ergodicity and the Markov property; with no one set of assumptions being "most" universal. The presentation of Section 2 provides an uncluttered exposition of the basic ideas involved.

In Sections 3 and 4 the general results are shown to apply, under some moment conditions, to processes which are stationary and ergodic and to Markov

Received June 1976; revised June 1977.

AMS 1970 *subject classifications*. Primary 62M05, 62M10; Secondary 62F10.

Key words and phrases. Estimation, ergodic Markov processes, stationary processes, consistency, asymptotic normality, iterated logarithm, branching process with immigration, time series.

processes which are asymptotically stationary and ergodic. The proofs of the theorems in this section owe much to the work of Billingsley (1961a).

In Section 5 we apply the method of CLS to branching processes with immigration. Estimation for these processes has been treated in Heyde and Seneta (1972, 1974) and Quine (1976). A correction to the asymptotic variance of the estimator of the mean of the immigration process in Quine (1976) is given in Quine (1977). A brief summary of a Monte Carlo study of the behavior of the CLS estimators for branching processes is also given in Section 5 where it is seen that the variances of the estimators are converging to zero at the rates given by our formulas.

Let y_t, $t = 1, 2, \cdots$ be a stochastic process defined on a probability space $(\Omega, \mathscr{F}, P_\alpha)$, whose distribution depends on a (column) vector $\alpha = (\alpha_1, \cdots, \alpha_p)'$ of unknown parameters with α lying in some open set A of Euclidean p-space. Let $E_\alpha(\cdot)$ and $E_\alpha(\cdot \mid \cdot)$ denote expectation and conditional expectation under P_α. Denote the "true" value of α by $\alpha^\circ = (\alpha_1{}^\circ, \cdots, \alpha_p{}^\circ)'$. This latter statement is taken to mean that all probabilities, a.e.'s and nonsubscripted expectations and conditional expectations are taken relative to the measure determined by α°. Let $\{F_t\}_{t=0}^\infty$ denote a sequence of sub-sigma fields with F_t generated by an arbitrary subset of $\{y_1, y_2, \cdots, y_t\}$, $t \geqq 1$, and let F_0 denote the trivial sigma field. Assume that $y_t \in L^1$, $t = 1, 2, \cdots$. Define the functions $g(\cdot, \cdot)$ by

$$g(\alpha, F_t) = E_\alpha(y_{t+1} \mid F_t), \qquad t \geqq 0. \tag{1.1}$$

Given a set of observations y_t, $t = 1, 2, \cdots, n$, we estimate α by trying to minimize the conditional sum of squares

$$Q_n(\alpha) = \sum_{t=0}^{n-1} [y_{t+1} - g(\alpha, F_t)]^2 \tag{1.2}$$

with respect to α. The estimates are actually obtained by solving the "least squares" equations.

$$\partial Q_n(\alpha)/\partial \alpha_i = 0, \qquad i = 1, 2, \cdots, p. \tag{1.3}$$

In specific applications it may be convenient to start the summation defining $Q_n(\alpha)$ at a positive integer greater than one. This is the case, for example, if $\{y_t\}$ is an mth order Markov process (see Section 3). Define the difference $u_t(\alpha)$ by

$$u_t(\alpha) = y_t - g_t(\alpha, F_{t-1}), \qquad t = 1, 2, \cdots. \tag{1.4}$$

Note that if $F_t = \sigma(y_1, y_2, \cdots, y_t)$, $\{u_t(\alpha^\circ)\}$ is a sequence of martingale differences.

2. General results. The basic technique of proof is to control the behavior of the first and second order terms in a Taylor expansion of $Q_n(\alpha)$ about α°. It is assumed throughout that $g(\alpha, F_t)$ is twice continuously differentiable with respect to α a.e., in some neighborhood S of α°. Without further notice, all neighborhoods defined below will be taken to be contained in S. Then, for $\delta > 0$, $\|\alpha - \alpha^\circ\| < \delta$, for some α^*, $0 < \|\alpha^\circ - \alpha^*\| < \delta$ (henceforth, α^* denotes

an appropriate intermediate point not necessarily the same from line to line),

$$\begin{aligned} Q_n(\alpha) &= Q_n(\alpha^\circ) + (\alpha - \alpha^\circ)' \, \partial Q_n(\alpha^\circ)/\partial\alpha + \tfrac{1}{2}(\alpha - \alpha^\circ)' \, \partial^2 Q_n(\alpha^*)/\partial\alpha^2 \\ &= Q_n(\alpha^\circ) + (\alpha - \alpha^\circ)' \, \partial Q_n(\alpha^\circ)/\partial\alpha \\ &\quad + \tfrac{1}{2}(\alpha - \alpha^\circ)' V_n(\alpha - \alpha^\circ) + \tfrac{1}{2}(\alpha - \alpha^\circ)' T_n(\alpha^*)(\alpha - \alpha^\circ) , \end{aligned} \tag{2.1}$$

where V_n is the $p \times p$ matrix of second partials of $Q_n(\alpha^\circ)$

$$V_n = (\partial^2 Q_n(\alpha^\circ)/\partial\alpha_i \, \partial\alpha_j) , \qquad \text{and}$$

$$T_n^{p\times p}(\alpha^*) = (\partial^2 Q_n(\alpha^*)/\partial\alpha^2 - V_n) .$$

Note that

$$\begin{aligned} \tfrac{1}{2}V_n = &(\textstyle\sum_{t=0}^{n-1} (\partial g(\alpha^\circ, F_t)/\partial\alpha_i \cdot \partial g(\alpha^\circ, F_t)/\partial\alpha_j))_{i\leqq p; j\leqq p} \\ &- (\textstyle\sum_{t=0}^{n-1} (\partial^2 g(\alpha^\circ, F_t)/\partial\alpha_i \, \partial\alpha_j) u_{t+1}(\alpha^\circ))_{i\leqq p; j\leqq p} . \end{aligned}$$

The following is a requirement of all our results and is henceforth assumed to hold.

(2.2) ASSUMPTION. $(2n)^{-1}V_n \to V^{p\times p}$ a.e., a positive definite (symmetric) matrix of constants. This condition can be verified by showing that

$$(n^{-1} \textstyle\sum_{t=0}^{n-1} (\partial^2 g(\alpha^\circ, F_t)/\partial\alpha_i \, \partial\alpha_j) u_{t+1}(\alpha^\circ))^{p\times p} \to 0^{p\times p} , \quad \text{a.e.} \tag{2.3}$$

(strong laws for martingales can be used here when $F_t = \sigma(y_1, y_2, \cdots, y_t)$, $t \geqq 1$ and integrability is assumed; see Stout (1974)) and

$$(n^{-1} \textstyle\sum_{t=0}^{n-1} \partial g(\alpha^\circ, F_t)/\partial\alpha_i \cdot \partial g(\alpha^\circ, F_t)/\partial\alpha_j)^{p\times p} \to V^{p\times p} , \quad \text{a.e.}$$

where the limit is assumed to be positive definite.

Strong consistency of the CLS estimators can now be shown.

THEOREM 2.1. *Assume that*:

(i) $\lim_{n\to\infty} \sup_{\delta\to 0} (|T_n(\alpha^*)_{ij}|/n\delta) < \infty$ a.e., $i \leqq p$, $j \leqq p$,
(ii) (2.2) *holds*,
(iii) $n^{-1} \, \partial Q_n(\alpha^\circ)/\partial\alpha_i \to 0$, a.e., $i \leqq p$.

(*The comment under* (2.3) *is also relevant here.*)

Let $\varepsilon > 0$, $\delta > 0$, *be given and let* N_δ *denote the open sphere of radius* δ *centered at* α°. *Then, for some* δ^*, $0 < \delta^* < \delta$, *there exists an event* E *with* $P(E) > 1 - \varepsilon$ *and an* n_0 *such that on* E, *for any* $n > n_0$, *the least squares equations* (1.3) *have a solution* $\{\hat{\alpha}_n\}$ *in* N_{δ^*} *at which point* $Q_n(\alpha)$ *attains a relative minimum.*

PROOF. Using (i)—(iii) we can find by Egoroff's theorems an event E with $P(E) > 1 - \varepsilon$, a positive δ^* less than δ, $M > 0$ and an n_0 such that on E, for any $n > n_0$, $\alpha \in N_{\delta^*}$ the following three conditions hold: (a) $|(\alpha - \alpha^\circ)' \, \partial Q_n(\alpha^\circ)/\partial\alpha| < n\delta^3$, (b) the minimum eigenvalue of $(2n)^{-1}V_n$ is greater than some $\Delta > 0$ (recall that $V^{p\times p} = \lim (2n)^{-1}V_n$ is positive definite), (c) $\frac{1}{2}(\alpha - \alpha^\circ)' T_n(\alpha^*)(\alpha - \alpha^\circ) < nM\delta^3$. Hence, using the Taylor expansion (2.1), for α on the boundary of N_{δ^*},

$$\begin{aligned} Q_n(\alpha) &\geqq Q_n(\alpha^\circ) + n(-\delta^3 + \delta^2\Delta - M\delta^3) \\ &= Q_n(\alpha^\circ) + n\delta^2(\Delta - \delta - M\delta) . \end{aligned}$$

Since $\Delta - \delta - M\delta$ can be made positive by initially choosing δ sufficiently small, $Q_n(\alpha)$ must attain a minimum at some $\hat{\alpha}_n = (\hat{\alpha}_{n1}, \hat{\alpha}_{n2}, \cdots, \hat{\alpha}_{np})'$ in N_{δ^*}, at which point the least squares equations (1.3) must be satisfied.

COROLLARY 2.1. *Under the assumptions of Theorem* 2.1, *there exists a sequence of estimators* $\hat{\alpha}_n$ *such that* $\hat{\alpha}_n \to \alpha^\circ$ a.e., *and for* $\varepsilon > 0$, *there is an event* E *with* $P(E) > 1 - \varepsilon$ *and an* n_0 *such that on* E, *for* $n > n_0$, $\hat{\alpha}_n$ *satisfies the least squares equations* (1.3) *and* Q_n *attains a relative minimum at* $\hat{\alpha}_n$.

PROOF. Apply Theorem 2.1 with $\varepsilon_k = 2^{-k}$ and $\delta_k = k^{-1}$, $k = 1, 2, \cdots$ to determine a sequence of events $\{E_k\}$ and an increasing sequence $\{n_k\}$ having the properties specified in the theorem. For $n_k < n \leqq n_{k+1}$ define $\hat{\alpha}_n$ on E_k to be a root of (1.3) within δ_k of α° at which Q_n attains a relative minimum and define $\hat{\alpha}_n$ to be zero otherwise. Then $\hat{\alpha}_n \to \alpha^\circ$ on $\liminf E_k$ and this set has probability one. The latter assertion in the corollary clearly holds.

The joint asymptotic normality of the estimators $\{\hat{\alpha}_n\}$ obtained in Corollary 2.1 follows from the assumption that the linear term in the Taylor expansion (2.1) has asymptotically a joint normal distribution. This assumption may be verified by using the Cramér–Wold technique (see [6] or [5], page 48) and an appropriate central limit theorem on

$$n^{-\frac{1}{2}}c'(\alpha - \alpha^\circ)'\,\partial Q_n(\alpha^\circ)/\partial\alpha = -2n^{-\frac{1}{2}}\sum_{t=0}^{n-1}\left(\sum_{i=1}^{p} c_i\,\partial g(\alpha, F_t)/\partial\alpha_i\right)u_{t+1}(\alpha^\circ)\,,$$

where $c' = (c_1, c_2, \cdots, c_p)$ is an arbitrary nonzero vector of constants.

THEOREM 2.2. *In addition to the conditions of Theorem* 2.2, *assume that*

$$(\tfrac{1}{2})n^{-\frac{1}{2}}\,\partial Q_n(\alpha^\circ)/\partial\alpha \to_{\mathscr{L}} MVN(0^{p\times 1}, W)\,,$$

where $W^{p\times p}$ *is a positive definite matrix. Then,*

$$n^{\frac{1}{2}}(\hat{\alpha}_n - \alpha^\circ) \to_{\mathscr{L}} MVN(0^{p\times 1}, V^{-1}WV^{-1})\,.$$

PROOF. Since we are dealing with an asymptotic result, we may assume that $\{\hat{\alpha}_n\}$ satisfies the least squares equations (1.3). Expand the vector $n^{-\frac{1}{2}}\,\partial Q_n(\alpha^\circ)/\partial\alpha$ in a Taylor series about α° to obtain (using the notation of (2.1))

$$\begin{aligned}(2.4)\qquad 0^{p\times 1} &= n^{-\frac{1}{2}}\,\partial Q_n(\hat{\alpha}_n)/\partial\alpha \\ &= n^{-\frac{1}{2}}\,\partial Q_n(\alpha_n{}^\circ)/\partial\alpha + n^{-1}(V_n + T_n(\alpha^*))\cdot n^{\frac{1}{2}}(\hat{\alpha}_n - \alpha^\circ)\,.\end{aligned}$$

Since $n^{-1}(V_n + T_n(\alpha^*)) \to 2V$ a.e., the limiting distribution of $n^{\frac{1}{2}}(\hat{\alpha}_n - \alpha^\circ)$ is the same as that of $(2V)^{-1}n^{-\frac{1}{2}}\,\partial Q_n(\alpha^\circ)/\partial\alpha$. This yields the desired result.

The application of laws of the iterated logarithm to the right side of the Taylor expansion (2.4) yields rates of convergence for $n^{\frac{1}{2}}(\hat{\alpha}_n - \alpha^\circ)$. See Stout (1974) and Heyde and Scott (1973) for a variety of conditions under which such laws hold. Those conditions pertaining to martingales seem most relevant here. The proof given below indicates that the same rates of convergence can be obtained for maximum likelihood estimates. This observation has been independently made by Basawa et al. (1976).

COROLLARY 2.2. *In addition to the conditions of Theorem* 2.2, *assume that for any nonzero vector of constants* $c' = (c_1, \cdots, c_p)$,

$$\limsup_{n\to\infty} \frac{\sum_{t=0}^{n-1} h(\alpha^\circ, F_t, c)u_{t+1}(\alpha^\circ)}{(2n\sigma^2 \log\log n)^{\frac{1}{2}}} = 1 \quad \text{a.e.},$$

where

$$h(\alpha^\circ, F_t, c) = \sum_{i=1}^{p} c_i \, \partial g(\alpha^\circ, F_t)/\partial\alpha_i$$

and

$$\sigma^2 = c'V^{-1}WV^{-1}c .$$

Then

$$\limsup_{n\to\infty} \frac{n^{\frac{1}{2}}c'(\hat{\alpha}_n - \alpha^\circ)}{(2\sigma^2 \log\log n)^{\frac{1}{2}}} = 1 \quad \text{a.e.}$$

PROOF. The proof follows from (2.4).

The following corollary is the CLS analogue of the result for maximum likelihood estimation that twice the logarithm of the likelihood ratio has a limiting chi-square distribution with p degrees of freedom (Rao (1973), (6e. 1.6)).

COROLLARY 2.3. *Assume that the conditions of Theorem* 2.2 *hold. Let* $\{\chi_i\}$ *be independent chi-square variates each with one degree of freedom. Then,*

$$Q_n(\alpha^\circ) - Q_n(\hat{\alpha}_n) \to_{\mathscr{L}} \sum_{i=1}^{p} \lambda_i \chi_i , \tag{2.5}$$

where λ_i, $i = 1, 2, \cdots, p$ *are the (nonnegative) eigenvalues of* $V^{-1}W$.

PROOF. From (2.1) we have

$$\begin{aligned} Q_n(\alpha^\circ) - Q_n(\hat{\alpha}_n) = &-(\hat{\alpha}_n - \alpha^\circ)' \cdot \partial Q_n(\alpha^\circ)/\partial\alpha - \tfrac{1}{2}(\hat{\alpha}_n - \alpha^\circ)' \cdot V_n \cdot (\hat{\alpha}_n - \alpha^\circ) \\ &- \tfrac{1}{2}(\hat{\alpha}_n - \alpha^\circ)'T_n(\alpha_n{}^*)(\hat{\alpha}_n - \alpha^\circ) . \end{aligned}$$

The Taylor expansion (2.4) yields

$$-\partial Q_n(\alpha^\circ)/\partial\alpha = (V_n + T_n(\alpha_n{}^*))(\hat{\alpha}_n - \alpha^\circ) .$$

Thus

$$\begin{aligned} Q_n(\alpha^\circ) - Q_n(\hat{\alpha}_n) &= \tfrac{1}{2}(\hat{\alpha}_n - \alpha^\circ)'(V_n + T_n(\alpha^*))(\hat{\alpha}_n - \alpha^\circ) \\ &= [n^{\frac{1}{2}}(\hat{\alpha}_n - \alpha^\circ)'](V_n/2n + T_n(\alpha_n{}^*)/2n)[n^{\frac{1}{2}}(\hat{\alpha}_n - \alpha^\circ)] . \end{aligned}$$

Since $V_n/2n + T_n(\alpha_n{}^*)/2n \to V$ a.e. and $n^{\frac{1}{2}}(\hat{\alpha}_n - \alpha^\circ) \to Z$ where

$$Z \sim MVN(0^{p\times 1}, V^{-1}WV^{-1}) ,$$

the limiting distribution of $Q_n(\alpha^\circ) - Q_n(\hat{\alpha}_n)$ is given by the distribution of $Z'VZ$. The moment generating function of $Z'VZ$ is

$$\begin{aligned} E(e^{tz'Vz}) &= |1 - 2tV^{-1}W|^{-\frac{1}{2}} \\ &= (\textstyle\prod_{i=1}^{p}(1 - 2t\lambda_i))^{-\frac{1}{2}} , \end{aligned}$$

where λ_i, $i = 1, \cdots, p$ are the eigenvalues of $V^{-1}W$. Nonnegativity of the λ_i follows from Lancaster (1969), Theorem 2.15.1, and this identifies the distribution of $Z'VZ$ as that given in the right side of (2.5).

3. Stationary ergodic processes. Throughout this section it is assumed that $\{y_t\}_{t=0}^{\infty}$ is a stationary ergodic sequence of integrable random variables and for an arbitrary positive integer m, $F_t = \sigma(y_t, y_{t-1}, \cdots, y_{t-m+1})$, $t = m-1, m, \cdots$. For reasons of symmetry now define

$$Q_n(\alpha) = \sum_{t=m}^{n} [y_{t+1} - g(\alpha, F_t)]^2 .$$

The proofs in this section are based on a Taylor expansion of $Q_n(\alpha)$ carried out to third order terms.

$$(3.1) \qquad Q_n(\alpha) = Q_n(\alpha^\circ) + (\alpha - \alpha^\circ)' \, \partial Q_n/\partial\alpha + \tfrac{1}{2}(\alpha - \alpha^\circ)' V_n (\alpha - \alpha^\circ) + R_n ,$$

where as before $V_n^{p\times p} = \partial^2 Q_n(\alpha^\circ)/\partial\alpha^2$ and R_n is the usual remainder term. It is assumed that the function $g = g(\alpha, F_t)$ satisfies the following regularity conditions:

(i) $\partial g/\partial\alpha_i$, $\partial^2 g/\partial\alpha_i \, \partial\alpha_j$ and $\partial^3 g/\partial\alpha_i \, \partial\alpha_j \, \partial\alpha_k$ exist and are continuous for all $\alpha \in A$, $i \leqq p$, $j \leqq p$, $k \leqq p$;

(ii) for $i \leqq p$, $j \leqq p$, $E|(y_m - g)\, \partial g/\partial\alpha_i| < \infty$, $E|(y_m - g)\, \partial^2 g/\partial\alpha_i \, \partial\alpha_j| < \infty$ and $E|\partial g/\partial\alpha_i \cdot \partial g/\partial\alpha_j| < \infty$ where g and its partial derivatives are evaluated at α° and F_{m-1};

(iii) for $i, j, k = 1, \cdots, p$ there exist functions

$$H^{(0)}(y_{m-1}, \cdots, y_0) , \qquad H_i^{(1)}(y_{m-1}, \cdots, y_0) , \qquad H_{ij}^{(2)}(y_{m-1}, \cdots, y_0) ,$$
$$H_{ijk}^{(3)}(y_{m-1}, \cdots, y_0)$$

such that

$$|g| \leqq H^{(0)} , \qquad |\partial g/\partial\alpha_i| \leqq H_i^{(1)} , \qquad |\partial^2 g/\partial\alpha_i \, \partial\alpha_j| \leqq H_{ij}^{(2)} ,$$
$$|\partial^3 g/\partial\alpha_i \, \partial\alpha_j \, \partial\alpha_k| \leqq H_{ijk}^{(3)} \quad \text{for all} \quad \alpha \in A$$

and

$$E|y_m \cdot H_{ijk}^{(3)}(y_{m-1}, \cdots, y_0)| < \infty ,$$
$$E\{H^{(0)}(y_{m-1}, \cdots, y_0) \cdot H_{ijk}^{(3)}(y_{m-1}, \cdots, y_0)\} < \infty ,$$
$$E\{H_i^{(1)}(y_{m-1}, \cdots, y_0) \cdot H_{jk}^{(2)}(y_{m-1}, \cdots, y_0)\} < \infty .$$

Note that if y_m and all the H's in (iii) are square integrable, then the integrability requirements in (ii) and (iii) will be satisfied because of the Cauchy–Schwarz inequality. Finally we denote by V the $p \times p$ matrix of expected values

$$(3.2) \qquad V = (E(\partial g(\alpha^\circ; F_{m-1})/\partial\alpha_i \cdot \partial g(\alpha^\circ; F_{m-1})/\partial\alpha_j)) .$$

We will assume throughout that V is positive definite. Since V is always nonnegative definite, this is the same as assuming that V is nonsingular.

The approach here is to use the above assumptions to show that the conditions of the corresponding theorems of Section 2 are satisfied.

THEOREM 3.1 (consistency). *There is a sequence of estimators $\{\hat{\alpha}_n\}$ such that the conclusions of Corollary* 2.1 *hold.*

PROOF. The ergodic theorem yields:

$$n^{-1}(\alpha - \alpha^\circ)' \partial Q_n(\alpha^\circ)/\partial\alpha \to 0 \quad \text{a.e.},$$
$$(2n)^{-1}(\alpha - \alpha^\circ)' V_n(\alpha - \alpha^\circ) \to (\alpha - \alpha^\circ)' V(\alpha - \alpha^\circ), \quad \text{a.e.}$$

Together with the above regularity conditions (i)—(iii), it also implies that $R_n = \frac{1}{2}(\alpha - \alpha^\circ)' T_n(\alpha^*)(\alpha - \alpha^\circ)$ satisfies (i) of Theorem 2, where the expression on the right is the remainder term in (2.1). The assumptions of Theorem 2.1 are thereby satisfied. An application of Corollary 2.1 completes the proof.

THEOREM 3.2 (asymptotic normality). *In addition to the above, assume*:

$$E(y_t \mid y_{t-1}, \cdots, y_0) = E(y_t \mid y_{t-1}, \cdots, y_{t-m}) \quad \text{a.e.}, \qquad t \geqq m, \tag{3.3}$$

$$E(u_m^2(\alpha^\circ) | \partial g(\alpha^\circ, F_{m-1})/\partial\alpha_i \cdot \partial g(\alpha^\circ, F_{m-1})/\partial\alpha_j |) < \infty, \qquad i, j \leqq p \tag{3.4}$$

where (as before) $u_m(\alpha^\circ) = y_m - E(y_m \mid F_{m-1})$. *Define the* $p \times p$ *matrix* W *by*

$$W = (E(u_m^2(\alpha^\circ)\, \partial g(\alpha^\circ, F_{m-1})/\partial\alpha_i \cdot \partial g(\alpha^\circ, F_{m-1})/\partial\alpha_j)). \tag{3.5}$$

Let $\{\hat{\alpha}_n\}$ *be the consistent sequence of estimators obtained on Theorem* 3.1. *Then*

$$n^{\frac{1}{2}}(\hat{\alpha}_n - \alpha^\circ) \to MVN(0, V^{-1}WV^{-1}).$$

Conditions (3.3) and (3.4) are imposed so that a martingale central limit theorem may be used. The Cauchy–Schwarz inequality implies that (3.4) holds if y_t and the partial derivatives $\partial g(\alpha, F_{m-1})/\partial\alpha_i$ have finite fourth moments (under α°); but, sometimes (specifically in the case of a branching process with immigration) it is possible to get by with less than fourth moments. See Section 5.

PROOF. Expand the vector $\partial Q_n(\hat{\alpha}_n)/\partial\alpha$ in a Taylor series about α° and multiply through by $n^{-\frac{1}{2}}$

$$\begin{aligned} 0^{p\times 1} &= n^{-\frac{1}{2}}\, \partial Q_n(\hat{\alpha}_n)/\partial\alpha \\ &= n^{-\frac{1}{2}}\, \partial Q_n(\alpha^\circ)/\partial\alpha + n^{-1}(V_n + U_n) n^{\frac{1}{2}}(\hat{\alpha}_n - \alpha^\circ), \end{aligned} \tag{3.6}$$

where $U_n^{p\times p} = (2^{-1} \sum_{k=1}^p (\hat{\alpha}_{nk} - \alpha_k^\circ)\, \partial^3 Q_n(\alpha^*)/\partial\alpha_i\, \partial\alpha_j\, \partial\alpha_k)_{i\leqq p; j\leqq p}$ and we may assume without loss of generality that $\{\hat{\alpha}_n\}$ satisfies the least squares equation and α^* is an appropriate intermediate point. We have $n^{-1}U_n \to 0^{p\times p}$ a.e. Billingsley's (1961 b) central limit theorem may be applied to the martingale

$$\sum_{t=m}^n \sum_{i=1}^p (c_i\, \partial g(\alpha^\circ, F_{t-1})/\partial\alpha_i) u_t(\alpha^\circ)$$

for all nonzero vectors of constants $c = (c_1, c_2, \cdots, c_p)$. Thus, the conditions of Theorem 2.2 are satisfied and the result follows.

There is a relationship between the positive definiteness of V and that of W. It is easy to see that if

$$E([y_m - g(y_{m-1}, \cdots, y_0; \alpha^\circ)]^2 \mid F_{m-1}) > 0 \quad \text{a.e.},$$

then positive definiteness of V implies the same of W. For if $c = (c_1, \cdots, c_p)' \neq 0$

is any vector, then

$$c'Wc = E[(y_m - g)^2(\textstyle\sum_{i=1}^{p} c_i\, \partial g/\partial\alpha_i)^2]$$
$$= E[(\textstyle\sum_{i=1}^{p} c_i\, \partial g/\partial\alpha_i)^2 E((y_m - g)^2 \mid F_{m-1})]$$

and this latter quantity is positive since

$$E[(\textstyle\sum_{i=1}^{p} c_i\, \partial g/\partial\alpha_i)^2] = c'Vc > 0\,.$$

On the other hand the following example shows that it is possible to have V positive definite and W only semi-definite. Consider a Markov process with state space $\{-2, -1\} \cup (0, 1)$ and transition measure depending on a single parameter $\alpha \in (0, 1)$ as follows. If the process is in -2 or -1 it goes to α, while if the process is in $(0, 1)$, it goes to -2 or -1 each with probability $\frac{1}{2}$. Then, letting $F_t = \sigma(y_t)$,

$$g(\alpha; F_{t-1}) = E_\alpha(y_t \mid y_{t-1}) = \alpha \qquad \text{if} \quad y_{t-1} = -2 \quad \text{or} \quad -1$$
$$= -\tfrac{3}{2} \qquad \text{if} \quad y_{t-1} \in (0, 1)\,,$$

so

$$\partial g(\alpha; F_{t-1})/\partial\alpha = 1 \qquad \text{if} \quad y_{t-1} = -2 \quad \text{or} \quad -1$$
$$= 0 \qquad \text{otherwise.}$$

For this example $V = (\frac{1}{2})$ and $W = (0)$ for all $\alpha \in (0, 1)$. Note that $E((y_m - g)^2 \mid F_{m-1})$ is positive with positive probability in the example. If we had combined states -2 and -1, $E((y_m - g)^2 \mid F_{m-1})$ would have been zero a.e.

The conditions of Corollaries 2.2 and 2.3 can be seen to hold under the assumptions of Theorem 3.2. In Corollary 2.2 the remainder converges to zero as indicated above and Stout's (1970) law of the iterated logarithm for stationary ergodic sequences of martingale differences can be applied to $\{-2^{-1}V^{-1}\,\partial Q_n(\alpha^\circ)/\partial\alpha\}$ and to the individual terms of the vector $\partial Q_n(\alpha^\circ)/\partial\alpha$. The convergence to zero of the remainder term in Corollary 2.3 follows from the ergodic theorem, the stochastic boundedness of $n^{\frac{1}{2}}(\hat{\alpha}_n - \alpha^\circ)$ (Theorem 3.2) and the convergence of $\hat{\alpha}_n$ to α°.

4. The nonstationary Markov case. Let $\{y_t\}$ be a Markov process with stationary transition probabilities $p_\alpha(y, \cdot)$ for which there exists a unique stationary distribution $\pi_\alpha(\cdot)$, and assume that for *every* y the transition measure $p_{\alpha^\circ}(y, \cdot)$ is absolutely continuous with respect to the stationary distribution $p_{\alpha^\circ}(\cdot)$. Then, according to Theorem 1.1 of Billingsley (1961a) the process will be stationary and ergodic if $\pi_{\alpha^\circ}(\cdot)$ is the initial distribution, and the ergodic theorem will be in force. Moreover, the conclusion of the ergodic theorem will hold for *any* initial distribution. The proof of this theorem in [3] actually contains the result that if Λ is any set measurable on the σ-field generated by $y_1, y_2, \cdots$, and if Λ has probability zero under the stationary distribution, then Λ has probability zero for any initial distribution. This result shows that if *any* strong law holds under the stationary distribution, then, a fortiori, it must hold with the same limit for any initial distribution. These observations seem to be a part

of the folklore of the subject. In the same vein (see Billingsley (1961 b), last paragraph), if the martingale CLT of [4] is applicable under the stationary distribution, then it may be applied, yielding the same limit, with any initial distribution. Thus, if the assumptions of our theorems are satisfied when the process starts with the stationary distribution, all of the convergences used in the proofs will hold for any initial distribution, and therefore the conclusions hold regardless of the initial distribution.

5. An application. Let y_t, $t = 0, 1, \cdots$ be a branching process with immigration whose offspring distribution has mean $\lambda_1 < 1$ and finite variance σ_1^2 and whose immigration distribution has mean $\lambda_2 > 0$ and finite variance σ_2^2. Let the process have any initial distribution. Set

$$\mu = \lambda_2/(1 - \lambda_1)$$
$$c^2 = \mu\sigma_1^2 + \sigma_2^2$$

and

$$\sigma_\pi^2 = c^2/(1 - \lambda_1^2) .$$

When the offspring and immigration distributions have finite third moments, set

$$\gamma = [\Sigma(j - \lambda_2)^3 b_j + \mu\Sigma(j - \lambda_1)^3 f_j + 3\lambda_1\sigma_1^2 c^2(1 - \lambda_1^2)^{-1}]/(1 - \lambda_1^3)$$

where $\{f_j\}$ and $\{b_j\}$ are the offspring and immigration distributions respectively. It is shown in [15] and [16] that the state space for this process contains an irreducible positive recurrent class of states and that the stationary distribution has a finite second moment and even a finite third moment when the offspring and immigration distributions have finite third moments. In [16] the functional equation for cumulant generating functions is also given and may be used to show that the stationary distribution has μ, σ_π^2 and γ as its mean, variance and third central moment respectively.

Heyde and Seneta (1972), in Section 5, give an interesting account of how this model is used in the natural sciences and of attempts to estimate various parameters in the model. In particular, maximum likelihood estimation leads to a rather unwieldy equation. See Bartlett (1955), page 247. The method of conditional least squares which we have developed here can be used to estimate λ_1 and λ_2. We have that $E(y_t \mid F_{t-1}) = \lambda_1 y_{t-1} + \lambda_2$. Solving the least squares equations (1.3) yields

$$\hat{\lambda}_1 = \frac{n\Sigma y_{t-1} y_t - (\Sigma y_{t-1})(\Sigma y_t)}{n\Sigma y_{t-1}^2 - (\Sigma y_{t-1})^2} \tag{5.1}$$

and

$$\hat{\lambda}_2 = \frac{1}{n}(\Sigma y_t - \hat{\lambda}_1 \Sigma y_{t-1}) \tag{5.2}$$

where all sums run from 1 to n. These estimates are essentially the same as those in Quine (1976), page 319, the only difference being whether or not an initial term or final term is included in certain sums; hence both pairs of estimators have the same asymptotic behavior. It is easy to see that the regularity

conditions of Section 3 are satisfied in a neighborhood of the true parameters and (see below) that V is positive definite when at least one of σ_1^2 and σ_2^2 is positive; thus (see Theorem 3.1 and Section 4) $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are strongly consistent. By conditioning on y_{t-1} it can be shown that (3.4) holds when the stationary distribution has a finite third moment; this will be so when both the offspring and immigration distributions have finite third moments. In this case $n^{\frac{1}{2}}(\hat{\lambda}_1 - \lambda_1, \hat{\lambda}_2 - \lambda_2)$ will have the limiting multivariate normal distribution given in Theorem 3.2. Furthermore the law of the iterated logarithm will hold for $(\hat{\lambda}_1, \hat{\lambda}_2)$. The matrices V and W are obtained from (3.2) and (3.5). The expectation is of course computed under the stationary distribution. In evaluating (3.5) one first conditions on y_0. The results, expressed in terms of the first three moments r_1, r_2 and r_3, and the variance $\sigma_\pi^2 = r_2 - r_1^2$ of the stationary distribution, are

$$V = \begin{pmatrix} r_2 & r_1 \\ r_1 & 1 \end{pmatrix} \qquad V^{-1} = \sigma_\pi^{-2} \begin{pmatrix} 1 & -r_1 \\ -r_1 & r_2 \end{pmatrix} \tag{5.3}$$

and

$$W = \begin{pmatrix} \sigma_1^2 r_3 + \sigma_2^2 r_2 & \sigma_1^2 r_2 + \sigma_2^2 r_1 \\ \sigma_1^2 r_2 + \sigma_2^2 r_1 & \sigma_1^2 r_1 + \sigma_2^2 \end{pmatrix}. \tag{5.4}$$

For this process we have

$$E((Y_m - g)^2 \mid F_{m-1}) = \sigma_1^2 y_{m-1} + \sigma_2^2 \tag{5.5}$$

which is positive a.e. when at least one of σ_1^2 and σ_2^2 is positive. Thus, by a remark after Theorem 3.2, W is positive definite. We can multiply out $V^{-1}WV^{-1}$ and express the stationary moments in terms of stationary central moments to obtain expressions for the asymptotic variances and covariance. After a long and somewhat tedious calculation we obtain

$$(V^{-1}WV^{-1})_{11} = (\sigma_1^2\gamma + c^2\sigma_\pi^2)/\sigma_\pi^4, \tag{5.6}$$

$$(V^{-1}WV^{-1})_{21} = -(\mu\sigma_1^2\gamma + \mu c^2\sigma_\pi^2 - \sigma_1^2\sigma_\pi^4)/\sigma_\pi^4 \tag{5.7}$$

and

$$(V^{-1}WV^{-1})_{22} = (\mu^2\sigma_1^2\gamma + \mu^2 c^2\sigma_\pi^2 + \sigma_2^2\sigma_\pi^4 - \mu\sigma_1^2\sigma_\pi^4)/\sigma_\pi^4. \tag{5.8}$$

To get some feeling for how these estimators behave we have done some simulations for the case in which the offspring distribution is Bernoulli (one with probability p and zero with probability $q = 1 - p$) and the immigration distribution is Poisson with mean λ. The transition matrix for this chain is a bit complicated, but one can write it down and then use it to check that the stationary distribution is Poisson with mean $\mu = \lambda/q$. Alternatively one can check that the Poisson cumulant generating function satisfies the function equation in [16], page 321. The estimates $\hat{p}$ and $\hat{\lambda}$ are given by the right sides of (5.1) and (5.2) respectively and the limiting covariance matrix of $n^{\frac{1}{2}}(\hat{p} - p, \hat{\lambda} - \lambda)$ is found from (5.6), (5.7) and (5.8) to be

$$\begin{pmatrix} q[pq + \lambda(1+p)]/\lambda & -\lambda(1+p) \\ -\lambda(1+p) & \lambda[\lambda(1+p) + q]/q \end{pmatrix}. \tag{5.9}$$

To carry out the simulations uniform random numbers were generated by the multiplicative congruential method using the equation

$$x_{n+1} = ax_n \bmod m$$

where $a = 16807$, $m = 2^{31} - 1$ and the initial value x_0 (which can be completely arbitrary) was taken to be 142867893 on every run. This particular random number generator is described and evaluated in [12] where it is found to be "highly satisfactory." Computations were done on a Xerox Sigma-7 at the Freas-Rooke Computer Center, Bucknell University, and the random numbers were generated using double precision arithmetic (14 significant hexadecimal digits). A separate run was made for each pair of values of p and λ. The initial distribution was taken to be the stationary distribution and 100 steps of the process were generated. The estimates $\hat{p}$ and $\hat{\lambda}$ were computed successively after every 10 steps of the process. Also computed were estimates of p with λ known. 1000 repetitions were made on each run.

The simulation results for $p = .4$, $\lambda = 3$ are typical and a sampling of these results is presented below. In all of the tables n is the number of steps the process has run. All of the results in the tables were determined on the basis of 1000 repetitions. Table 1 gives the means, variances and covariance of $\hat{p}$ and $\hat{\lambda}$. It is clear from this table that $\hat{p}$ is biased down and $\hat{\lambda}$ is biased up. This inverse relationship can be anticipated from the negative covariance between $\hat{p}$ and $\hat{\lambda}$. The presence of bias is not entirely surprising in view of the close connection, already observed by Heyde and Seneta (1972), between the estimation technique used here and time series models. The problem of bias is a familiar one in the analysis of time series data. Table 1 also shows that when λ is known, $\hat{p}$ is nearly unbiased. Tables 2 and 3 are intended to provide an informal test of normality. They give the number of times the estimate fell within one, between one and two, etc., standard deviations above and below its mean, where the means and standard deviations are those determined in Table 1. It can be seen from Tables 2 and 3 that the distribution of $\hat{p}$ is skewed to the left while that of $\hat{\lambda}$ is skewed to the right.

In judging normality we have compared the estimates with their actual means

TABLE 1

Expected values, variances and covariances, as determined by 1000 *simulations*

n	$E(\hat{p})$	$E(\hat{\lambda})$	$n \cdot \mathrm{Var}(\hat{p})$	$n \cdot \mathrm{Var}(\hat{\lambda})$	$n \cdot \mathrm{Cov}(\hat{p}, \hat{\lambda})$	(λ known)	
						$E(\hat{p})$	$n \cdot \mathrm{Var}(\hat{p})$
20	.290	3.56	.907	26.85	−4.33	.380	.167
40	.344	3.28	.896	25.75	−4.31	.390	.155
60	.361	3.20	.879	24.81	−4.16	.394	.164
80	.372	3.14	.859	23.09	−3.99	.396	.159
100	.380	3.11	.868	22.82	−3.99	.397	.161
∞	.400	3.00	.888	24.00	−4.20	.400	.155

TABLE 2

Number of times $\hat{p}$ fell within 1, *between* 1 *and* 2, etc. *standard deviations above and below its mean in* 1000 *simulations*

n	-4	-3	-2	-1	1	2	3	4
20	1	26	135	321	367	140	9	1
40	5	24	133	328	348	148	14	0
60	3	27	133	317	360	145	15	0
80	2	24	135	332	356	132	19	0
100	3	32	124	326	363	137	15	0
∞	1	22	136	341	341	136	22	1

TABLE 3

Number of times $\hat{\lambda}$ fell within 1, *between* 1 *and* 2, etc. *standard deviations above and below its mean in* 1000 *simulations*

n	-4	-3	-2	-1	1	2	3	4
20	0	4	149	385	301	122	36	3
40	0	7	153	374	303	134	23	6
60	0	12	148	370	305	134	27	4
80	1	13	135	376	320	129	21	5
100	0	17	131	378	321	120	28	5
∞	1	22	136	341	341	136	22	1

TABLE 4

Number of times various confidence intervals covered their parameters in 1000 *simulations*

n	$\hat{p}$			$\hat{\lambda}$		
	90%	95%	98%	90%	95%	98%
20	856	889	893	840	885	894
40	878	935	973	880	946	971
60	884	943	978	900	948	975
80	887	951	980	901	951	980
100	888	949	980	905	952	981

and standard deviations (as determined from 1000 simulations). Of more practical interest is the question of how the estimation procedure behaves when confidence intervals are formed using normal critical values and using the variances from (5.9) with p, q and λ replaced by their estimates. A problem that can arise here is that $\hat{p}$ may not fall in (0, 1) or $\hat{\lambda}$ may not come out positive. This can result in negative estimates of variance. The problem of $\hat{\lambda} \leqq 0$ or $\hat{p} \geqq 1$ was rate, but $\hat{p} \leqq 0$ was common for small p and for small n. Table 4 gives the number of times out of 1000 simulations that various confidence intervals, when formed by the above procedure, covered their parameters. When either estimate fell out of bounds the case was not counted. From Table 4 it can be seen that the results are extremely good, and we would not hesitate to recommend this

procedure for the particular model considered here. For more extreme values of p, the results were not quite as good, but still impressive.

6. Concluding remark. It might appear that the estimation method developed here is limited to those parameters which appear in the conditional expectation (1.1) of the process. However by suitably choosing functions $\{f_i\}_{i=1}^r$ and applying the method to a conditional sum of squares defined by

$$Q_n(\alpha) = \sum_{i=1}^r \Sigma(f_i(y_{t+1}, y_t, \cdots, y_{t-q}) - E(f_i(y_{t+1}, \cdots, y_{t-q}) \mid F_{t-q-1}))^2,$$

estimates can be obtained for a wide variety of parameters. CLS estimates, for example, can be obtained for the transition probabilities of a Markov chain by choosing $\{f_{ij}\}$ as the indicator functions of jumps from state i to state j. The estimates in this case are the same as the MLE's. If the $\{y_t\}$ are i.i.d. and $f_i(x) = x^i$, $i \leqq r$, this formulation of CLS amounts to the method of moments.

REFERENCES

[1] BARTLETT, M. S. (1955). *An Introduction to Stochastic Processes.* Cambridge Univ. Press.

[2] BASAWA, I. V., FEIGIN, P. D. and HEYDE, C. C. (1978). Asymptotic properties of maximum likelihood estimators for stochastic processes. *Sankhyā.* To appear.

[3] BILLINGSLEY, P. (1961 a). *Statistical Inference for Markov Processes.* Univ. of Chicago Press.

[4] BILLINGSLEY, P. (1961 b). The Lindeberg-Lévy theorem for martingales. *Proc. Amer. Math. Soc.* **12** 788–792.

[5] BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[6] CRAMÉR, H. and WOLD, H. (1936). Some theorems on distribution functions. *J. London Math. Soc.* **11** 290–295.

[7] DURBIN, J. (1960). Estimation of parameters in time-series regression models. *J. Roy. Statist. Soc. Ser. B* **22** 139–153.

[8] HEYDE, C. C. and SENETA, E. (1972). Estimation theory for growth and immigration rates in a multiplicative process. *J. Appl. Probability* **9** 235–256.

[9] HEYDE, C. C. and SENETA, E. (1974). Note on "Estimation theory for growth and immigration rates in a multiplicative process." *J. Appl. Probability* **11** 572–577.

[10] HEYDE, C. C. and SCOTT, D. J. (1973). Invariance principles for the law of the iterated logarithm for martingales and processes with stationary increments. *Ann. Probability* **1** 428–436.

[11] LANCASTER, P. (1969). *Theory of Matrices.* Academic Press, New York.

[12] LEWIS, P. A. W., GOODMAN, A. S. and MILLER, J. M. (1969). Pseudo-random number generator for the System/360. *IBM Systems Journal* **8** (2) 136–146.

[13] MANN, H. B. and WALD, A. (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica* **11** 173–220.

[14] MCLEISH, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probability* **2** 620–628.

[15] QUINE, M. P. (1970). The multi-type Galton-Watson process with immigration. *J. Appl. Probability* **7** 411–422.

[16] QUINE, M. P. (1976). Asymptotic results for estimators in a subcritical branching process with immigration. *Ann. Probability* **4** 319–325.

[17] QUINE, M. P. (1977). Correction to "Asymptotic results for estimators in a subcritical process with immigration." *Ann. Probability* **5** 318.

[18] RAO, C. R. (1973). *Linear Statistical Inference and Its Applications II.* Wiley, New York.

[19] STOUT, W. F. (1970). The Hartman–Wintner law of the iterated logarithm for martingales. *Ann. Math. Statist.* **41** 2158–2160.

[20] STOUT, W. F. (1974). *Almost Sure Convergence.* Academic Press, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN 53706

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802