ASYMPTOTIC DISTRIBUTION THEORY FOR A CLASS OF NONLINEAR ESTIMATION METHODS

by

Christopher A. Sims

Discussion Paper No. 76 - 69, July 1976

Center for Economic Research
Department of Economics
University of Minnesota
Minneapolis, Minnesota 55455

Asymptotic Distribution Theory for a Class of Nonlinear Estimation Methods

This paper, prepared as notes for an econometrics course, applies to a class of estimation methods including least squares, maximum likelihood, and "quasi-maximum likelihood" (maximum likelihood with an incorrect likelihood function) the recent developments in martingale central limit theorems discussed, e.g., in Scott [1973]. Such theorems allow us to avoid strong assumptions on the form of the distributions of random variables being observed or on the nature of serial interdependence. In particular they allow a unified treatment of regression models with or without lagged dependent variables.

The formulation of the theorems also aims at avoiding convenient assumptions which are likely to be violated in practice, yet are not essential to the results. While the extent to which the theorems succeed in this is a matter of judgment, they do have a useful property in this regard: the theorems cover the case of the standard linear regression model without restricting the parameter space or the range of variation of the independent variables, so that at least all models which are in some sense "close" to the linear model are covered, for the stationary case.

No great originality is claimed for these results. They are largely embellishments or combinations of ideas I have seen elsewhere. Many of the most interesting time series applications of the results are covered by Hannan [1973a] [1973b]. However the full extent of the results unoriginality remains to be explored.

Preliminaries

In consistency proofs, we shall assume that the parameter space S is a compact, separable metric space under the metric $d(\ ,\ )$. In many applications, it will be natural to take S to be the whole of R^n under a metric which allows us to compactify R^n by adding "infinity" points to it. Thus in applying Theorem 4 to a linear model where b is unrestricted and one-dimensional one might take $d(b_1,b_2)$ to be $|\mathrm{Tan}^{-1}b_1 - \mathrm{Tan}^{-1}b_2|$. The real line, expanded to include $\pm\infty$, is a compact separable metric space under this metric, and the conditions of the theorem are verifiable for the linear model under this interpretation.

Definition: The function $f(y(t),b)$ is continuous in the mean in b at b_0 iff $E[\max|f(y(t),b) - f(y(t),b_o|\text{s.t. } d)b, b_o) < \delta] \to 0$ as $\delta \to 0$.

Definition: The function $f(y(t),b)$ is continuous in b, uniformly in the mean, iff for every $\varepsilon > 0$ there is a $\delta > 0$ such that

$$E[\max\{|f(y(t),b_1) - f(y(t),b_2)| \text{ s.t. } d(b_1,b_2) < \delta\}] < \varepsilon .$$ [1]

Definition: For any matrix A, $||A|| = (\mathrm{tr}(A'A))^{1/2}$.

We assume we have a function $g(y(t),b)$ such that $E[g(y(t),b)]$ has a unique minimum at $b = b_0$. We then look at $g_T(b) = T^{-1}\sum_{s=1}^{T} g(y(t),b)$. In hopes that g_T might look like the expectation of g, we consider estimators of b_0 obtained by setting $\hat{b}_T$ equal to that value of b which minimizes g_T. These notes discuss the properties of these and certain related estimators.

---

[1]Note that under these definitions continuity in the mean implies continuity for fixed y with probability one and continuity uniformly in the mean implies uniform continuity for fixed y with probability one.

Theorem 1: Suppose:

i) $y(t)$ is stationary and ergodic;

ii) b lies in a compact, separable metric space S;

iii) $g_\infty(b) = E[g(y(t),b)]$ exists on a dense subset of S;

iv) $\bar{g}(y(t),b) = (\, g(y(t),b) - g_\infty(b_0)\,) \,/\, (g_\infty(b)-g_\infty(b_0) + 1)$ is continuous in b, uniformly in the mean on S;

v) $g_\infty(b)$ has a unique minimum at $b=b_0$.

Then $\hat{b}_T \xrightarrow{a.s.} b_0$.

Proof: First we show that $\bar{g}_\infty(b) = E[\bar{g}(y(t),b)]$ is uniformly continuous. Assumption (iv) gives us uniform continuity of $\bar{g}$, and hence of $\bar{g}_T(b) = T^{-1}\sum_{s=1}^{T} \bar{g}(y(s),b)$ with probability one for each T. the function

$$\varepsilon(y(t),\delta) = \max\,\{|\bar{g}(y(t),b_1)-\bar{g}(y(t),b_2)| \quad \text{s.t.} \quad d(b_1,b_2) < \delta\}$$

can be averaged over the sample to give us a bound $\varepsilon_T(\delta)$ on $|\bar{g}_T(b_1)-\bar{g}_T(b_2)|$ for $d(b_1,b_2) < \delta$. Since ε has an expectation which converges to zero with δ, and since y is ergodic, we can, given any $\mu > 0$, with probability one pick a T_0 large enough and a δ small enough such that for all $T>T_0$, $\varepsilon_T(\delta) < \mu$. By (iii) and (i), $\bar{g}_T$ converges a.s. for each b in a dense subset of S. Since S is separable, this dense subset has a countable dense subset. Now for any b in S and any $\mu > 0$, we can with probability one pick T_0 large enough and δ small enough so that for all $T_i > T_0$, and for any b_1 in the dense convergence set, $d(b_1,b) < \delta$ implies

$$|\bar{g}_{T_1}(b)-\bar{g}_{T_2}(b)| < |\bar{g}_{T_1}(b_1)-\bar{g}_{T_2}(b_1)| + |\bar{g}_{T_1}(b)-\bar{g}_{T_1}(b_1)| + |\bar{g}_{T_2}(b)-\bar{g}_{T_2}(b_1)| < \mu.$$

This makes $\bar{g}_T$ Cauchy for each b, and hence convergent. Further, a standard argument, very similar to that used to show that a uniformly continuous sequence of functions which converges pointwise converges to a continuous limit, shows that $\bar{g}_T$ converges a.s. to a continuous limit.

Since on the dense subset where $E[g] = g_\infty$ exists, g_T converges to it a.s., $\bar{g}_\infty$ has a unique extension to a continuous function on all of S, and we shall use $\bar{g}_\infty$ to refer to this extension henceforth. Clearly $\bar{g}_T$ converges a.s. to $\bar{g}_\infty$.

Now we show that given any $\varepsilon > 0$, we can choose a T_0 such that for all $T > T_0$, $|\hat{b}_T - b_0| < \varepsilon$, which will complete the proof. It follows from the continuity of $\bar{g}_\infty$, the compactness of S, and the uniqueness of the minimum of g_∞ at $b=b_0$ that $\min_{b \text{ in } S_\varepsilon} \{\bar{g}_\infty\} = \delta$ for some positive δ where $S_\varepsilon = \{b \text{ in } S \mid d(b,b_0) \geq \varepsilon\}$. Now choose T_0 large enough so that for all $T>T_0$, $|\bar{g}_T - \bar{g}_\infty| < \delta/2$ uniformly over S. Then $\bar{g}_T(b_0) < \bar{g}_\infty(b_0) + \delta/2$; and $\bar{g}_T(b) > \delta - \delta/2 = \delta/2$, for b in S. This latter inequality is easily seen to imply that $g_T(b) > g_\infty(b_0) + \delta/2$ for b in S_ε. Thus for all b in S_ε, $g_T(b) > g_T(b_0)$, and the minimum of g_T must occur in $S - S_\varepsilon$. Q.E.D.

<u>Theorem 2</u>

Suppose

i) $\hat{b}_T \xrightarrow[T\to\infty]{} b_o$;

ii) $||D_{bb'}g(y(t),b) - D_{bb'}g(y(t),b_0)||$ is continuous in the mean at $b = b_0$;

iii) $E[D_{bb'}g(y(t),b)]$ is non-singular;

iv) $E[D_b g(y(t),b_0) | y(t-s), \text{ all } s > 0] = 0$;

v) $y(t)$ is stationary and ergodic.

vi) $E[D_b g(y(t),b_o)\ D_b g(y(t),b_o)']$ exists and is non-singular.

Then $\sqrt{T}(\hat{b}_T - b_0) \xrightarrow{D} N(0, E[D_b g(y(t),b_0) D_b g(y(t),b_0)']\ \{E[D_{bb'}g(y(t),b_0)]\}^{-1}$
$E[D_b g(y(t),b_0)\ D_b g(y(t),b_0)'])$.

Proof: Implicit in (ii) and (iii) is the assumption that $D_{bb'}g_T(b)$ exists a.s. for b close enough to b_0, hence that $D_b g_T(\hat{b}_T) = 0$ a.s., at least for large T. But $D_b g_T = D_b g_T(b_0) + \Lambda_T(\hat{b}_T - b_0)$, where Λ_T lies between $D_{bb'}(b_0)$ and $D_{bb'}(b_T)$

(*) $\qquad \hat{b}_T - b_0 = -[\Lambda_T]^{-1} D_b g_T(b_0)$,

assuming that the inverse on the right-hand-side of this expression exists. But now $\Lambda_T \xrightarrow{P} E[D_{bb'}g(y(t),b_0)]$. This follows because, first, $D_{bb'}g_T(b) \xrightarrow{a.s.} E[D_{bb'}g(y(t),b)]$ for any b such that the expectation on the right exists, by ergodicity. Define $\varepsilon(y(t),\delta) = \max\{||D_{bb'}g(y(t),b) - D_{bb'}g(y(t),b_0)||$ s.t. $d(b,b_0) < \delta$. Now $||D_{bb'}g_T(\hat{b}_T) - D_{bb'}g_T(b_0)||$ $< \frac{1}{T}\sum_{t=1}^{T} \varepsilon(y(t),\delta)$ for $d(\hat{b}_T,b_0) < \delta$. By ergodicity, $\frac{1}{T}\sum_{t=1}^{T} \varepsilon(y(t),\delta) \xrightarrow{a.s.}$ $E[\varepsilon(y(t),\delta)]$ as $T \to \infty$. By (i) and (ii), then, we can choose, for any $\mu > 0$, $\alpha > 0$, δ small enough, then T_1 large enough, such that for all $T > T_1$ we have both $P[d(\hat{b}_T,b_0) < \delta] > 1 - \alpha$ and $||D_{bb'}g_T(\hat{b}_T) - D_{bb'}g_T(b_0)|| < \mu$ for $d(\hat{b}_T,b_0) < \delta$. Thus $\Lambda_T \xrightarrow{P} E[D_{bb'}g(y(t),b_0)]$.

With (iii) this implies that for T large enough the inverse on the right of (*) will exist, with arbitrarily high probability and that it will converge in probability to $\{E[D_{bb'}g(y(t),b_0)]\}^{-1}$.

Now consider

(**) $\qquad \sqrt{T}\, D_b g_T(b_0) = \frac{1}{\sqrt{T}} \sum_{s=1}^{T} D_b g(y(t), b_0)$.

We know by (iv) that the sum on the right of (**) forms a martingale, and by (v) that it has ergodic increments. Application of a Martigale Central Limit Theorem, as appears e.g. in Scott [1973], yields the result that (**) is asymptotically $N(0, E[D_b g(y(t),b_0)\, D_b g(y(t),b_0)'])$ and the conclusion of the theorem then follows immediately.

<u>Theorem 3</u>: Suppose the assumptions of Theorem 2 hold and that in addition $\bar{\bar{b}}_T$ is a consistent estimator with the property that

$$\lim_{\lambda\to\infty} \overline{\lim_{T\to\infty}} P[\sqrt{T}\,||\bar{\bar{b}}_T - b_0|| > \lambda] = 0.$$ [1/]

Then $\sqrt{T}(\hat{\hat{b}}_T - \hat{b}_T) \xrightarrow{P} 0$, where $\hat{\hat{b}}_T = \bar{\bar{b}}_T - [D_{bb'}g_T(\bar{\bar{b}}_T)]^{-1} D_b g_T(\bar{\bar{b}}_T)$.

Proof: $\hat{\hat{b}}_T - \hat{b}_T = \bar{\bar{b}}_T - [(D_{bb'}g_T(\bar{\bar{b}}_T)]^{-1} D_b g_T(\bar{\bar{b}}_T) - \bar{\bar{b}}_T + [\bar{\Lambda}_T]^{-1} D_b g_T(\bar{\bar{b}}_T)$, where $\bar{\Lambda}_T$ lies between $D_{bb'}g_T(\bar{\bar{b}}_T)$ and $D_{bb'}g_T(\hat{b}_T)$, and a first-order Taylor expansion of $D_b g_T(b)$ about $\bar{\bar{b}}_T$ has been applied in obtaining the expression for $\hat{b}_T$.

Thus, clearly

$$(+)\quad \sqrt{T}\,(\hat{\hat{b}}_T - \hat{b}_T) = \sqrt{T}\,\{-[D_{bb'}g_T(\bar{\bar{b}}_T)]^{-1} + [\bar{\Lambda}_T]^{-1}\}\, D_b g_T(\bar{\bar{b}}_T) = \{[\bar{\Lambda}_T]^{-1} - [D_{bb'}g_T(\bar{\bar{b}}_T)]^{-1}\}\,[\sqrt{T}\, D_b g_T(b_0) + \bar{\bar{\Lambda}}_T\,(\bar{\bar{b}}_T - b_0)\,\sqrt{T}] ,$$

where $\bar{\bar{\Lambda}}_T$ lies between $D_{bb'}g_T(\bar{\bar{b}}_T)$ and $D_{bb'}g_T(b_0)$.

Since $\bar{\bar{b}}_T$ is consistent for b_0, the first bracket on the right of (+) converges in probability to zero, by the same sort of argument used in Theorem 2 to show $D_{bb'}g_T(\hat{b}_T) \xrightarrow{P} D_{bb'}g_T(b_0)$. The first term within the second bracket, we showed in Theorem 2, has a limiting distribution. But then finally the assumption of this theorem on $\bar{\bar{b}}_T$'s rate of convergence to b_0 guarantees that for any A_T s.t. $A_T \xrightarrow{P} 0$, $A_T(\bar{\bar{b}}_T - b_0)\sqrt{T} \to 0$. Thus the conclusion of this theorem is proved. Q.E.D.

[1/] A sufficient condition for this is that $\sqrt{T}\,(\bar{\bar{b}}_T - b_0)$ have a limiting distribution.

Theorem 4: If

i) $y(t) = f(x(t),b_0) + u(t)$, all t;

ii) $E[u(t)f(x(t),b)] = 0$, all b in S, and $E[u(t)^2] = \sigma^2$.

iii) $y(t),x(t)$ are jointly stationary and ergodic;

iv) $$\frac{[u(t) + f(x(t),b_0) - f(x(t),b)]^2 - \sigma^2}{E[(f(x(t),b_0) - f(x(t),b))^2] + 1}$$ is continuous in b uniformly in the mean for all b in S under the metric d;

v) b_0 lies in S and S is a compact separable metric space under d;

vi) $E[(f(x(t),b) - f(x(t),b_0))^2]$ exists on a dense subset of S and is positive whenever it exists and $b \neq b_0$.

Then non-linear least squares estimates of b are strongly consistent estimates of b_0.

Proof: The theorem's conditions are almost direct translations of those of Theorem 1, once the identification $g(y_1(t),b) = (y(t) - f(x(t),b))^2$ is made, with $y_1(t) = (y(t), x(t))$.

Theorem 5: Suppose:

i) $y(t) = f(x(t),b) + u(t)$;

ii) $E[u(t)|x(t),u(t-1),x(t-1),u(t-2)\ldots] = 0$;

iii) y, x jointly stationary and ergodic;

iv) $E[D_b f(y(t),b_0)\ D_b f(y(t),b_0)']$ is non-singular;

v) $D_b f$, $D_{bb'} f$, and f itself are all continuous in the mean at $b=b_0$;

vi) $E[u(t)^2|x(t)] = \sigma^2 < \infty$;

vii) The least-squares estimate $\hat{b}_T$ is consistent.

Then $\sqrt{T}\,(\hat{b}_T - b_0) \xrightarrow{D} N(0,\ \sigma^2\ [D_b g(y(t),b_0) D_b g(y(t),b_0)']$.

Proof: The proof is only a matter of verifying the assumptions of Theorem 2 from this Theorem's assumptions with $g = (y-f)^2$. Assumption 2(iii) follow from 5.v, and the Schwarz inequality once we write out an explicit expression for $D_{bb'}g$. Assumption 2.i is 5.vii. 2.ii follows from 5.ii and 5.iv once we write out an explicit expression for $D_{bb'}g$. 2.iv follows from 5.ii, 2.v is 5.iii, and 2.vi follows from 5.vi and 5.iv. Q.E.D.

Theorem 6: Suppose the hypotheses of Theorem 5 hold and in addition

$$\lim_{\lambda\to 0}\ \lim_{T\to\infty}\ P[\sqrt{T}\ ||\bar{b}_T - b_0|| > \lambda] = 0.$$

Then if $\hat{\hat{b}}_T = \bar{b}_T + (\sum_{t=1}^{T} D_b f(x(t),\bar{b}_T)\ D_b f(x(t),\bar{b}_T)')^{-1} \sum_{t=1}^{T} D_b f(x(t),\bar{b}_T)$

$(y(t)-f(x(t),\bar{b}_T)$, $\sqrt{T}\,(\hat{\hat{b}}_T - \hat{b}_T) \xrightarrow{P} 0$.

Proof: The proof follows almost precisely the lines of the proof of Theorem 3. However it is important to note that the $\hat{\hat{b}}_T$ of this Theorem is not the $\hat{\hat{b}}_T$ of Theorem 3. In this theorem we consider the estimator obtained by linearizing the equation in 5.i, the regression equation, about $\bar{b}_T$. The estimator of Theorem 3 would be obtained by linearizing the first derivative of the sum of squared residuals with respect to b, setting the resulting expression to zero. The latter estimator, though asymptotically equivalent to that obtained by linearizing the regression equation under the regularity conditions imposed in this theorem, requires use of the second derivatives of f.

Extensions

Though the class of estimators considered in this paper is broad, it does not include an important class of estimators which might be characterized as generalizations of the classical method of moments. Instead of using the existence of a $g(y,b)$ function whose expected value is minimized at the true b as a basis for estimation, this other class begins from a vector of functions $h(y,b)$ whose expected value is zero when b is set at its true value. One then takes sample averages of h and chooses b to set h_T to zero (if h and b have the same length) or to minimize a suitably defined "length" of h_T. Econometricians call such methods "instrumental variables" or "two-stage least squares" when h is a vector of cross-moments between "instruments" and a "residual"[2] The methods of this paper will work in application to such generalized method of moments (GMM) estimators, and it is intended that a future draft of this paper will deal with GMM estimators. Proofs of results like Theorem 2 (asymptotic distribution theory) carry over with little modification to GMM estimators. Consistency proofs do not extend so directly to GMM estimates, but the GMM results resemble those Theorems 1 and 4 in relaxing the strong serial independence assumptions made in most of the econometric literature on this subject.

---

[2] See, e.g., Jorgenson and Laffont [1974].

REFERENCES

Hannan, E. J. (1973a), "The Asymptotic Theory of Linear Time-Series Models", Journal of Applied Probability, 10, 130-145.

________ (1973b), "Central Limit Theorems for Time Series Regression", Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete, 26. 157-170.

Jorgenson, Dale W., and Jean-Jacques Laffont, (1974), "Efficient Estimation of Nonlinear Simultaneous Equations with Additive Disturbances", Annals of Economic and Social Measurement, 3, 615-640.

Scott, D. J., (1973), "Central Limit Theorems for Martingales and for Processes with Stationary Increments Using a Skorokhod Representation Approach", Advances in Applied Probability, 5, 119-137.