A REMARK ON SERIAL CORRELATION IN MAXIMUM LIKELIHOOD*

by

David Levine

Working Paper #215
University of California, Los Angeles
Department of Economics
August 1981

*I would like to thank Jerry Hausman, Paul Ruud, Michael Veall and Sweder van Wijnbergen.

It is well known that OLS can be consistent and asymptotically normal despite serial correlation in the residuals. Although the usual estimator of the asymptotic covariance of the parameter estimator is inconsistent there is an alternative covariance estimator which is consistent.[1] The purpose of this note is to sketch how and why these results extend to MLE.

Let y^t be endogenous and z^t predetermined at time t with $x^t \equiv (y^t, z^t)$. For notational simplicity suppose $\{x^t\}$ is stationary. Let $f(y^t,z^t,\theta)$ be a family of conditional density functions for y^t and suppose $f(y^t,z^t,\theta_0)$ is the actual density of y^t conditional on the predetermined variables z^t. Notice that this does <u>not</u> imply that exp $T\,L^T(\theta) \equiv \Pi_{t=1}^T f(y^t,z^t,\theta_0)$ is the joint density of the y^t (conditional, or otherwise). If the z^t are exogenous this is true only if the y^t are independent. Define the partial MLE θ^T to be the estimator that maximizes $L^T(\theta)$. Note that MLE under the assumption of independence is partial MLE if there is serial correlation. We shall extend the usual consistency argument to show that the consistency of partial MLE depends <u>only</u> on $f(y,z,\theta_0)$ being the actual density of y conditional on z and <u>not</u> on $\Pi_{t=1}^T f(y^t,z^t,\theta_0)$ being a joint density for the y^t.

We make use of the following notation. The log-likelihood contribution is $\lambda^t(\theta) = \log f(x^t,\theta)$. Associated with λ^t are $L(\theta) \equiv E\lambda^t(\theta)$ and its empirical counterpart $L^T(\theta) \equiv (1/T)\Sigma_{t=1}^T \lambda^t(\theta)$. Subscript θ's denote differentiation. Thus the score contributions are $\lambda_\theta^t(\theta)$. Associated with these are autocorrelation functions

$$r_k^t(\theta) \equiv (1/2)\left\{[\lambda_\theta^t(\theta)]\,[\lambda_\theta^{t-k}(\theta)]' + [\lambda_\theta^{t-k}(\theta)]\,[\lambda_\theta^t(\theta)]'\right\}$$

$$R_k(\theta) \equiv Er_k^t(\theta) \text{ and}$$

$$R_k^T(\theta) \equiv (1/(T-k))\Sigma_{t=k}^T r_k^t(\theta).$$

For pedagogical purposes we make the following assumptions:

(1) x^t is stationary and φ-mixing with

$$\Sigma_{t=0}^{\infty} \varphi_t^{1/2} \leq B$$

(2) Θ is a compact convex set

(3) $f(y^t, z^t, \theta_0)$ is the true conditional density of y^t given z^t and is not stochastically equivalent to $f(y^t, z^t, \theta)$ unless $\theta = \theta_0$ [global identification]

(4) $\theta_0 \in$ interior (Θ)

(5) λ^t and L are C^2 and R_k C^o functions of θ

(6) $E|\lambda^t(\theta)|^2 \leq B$

$$E|\lambda_\theta^t(\theta)|^4 \leq B$$

$$E|\lambda_{\theta\theta}^t(\theta)|^2 \leq B$$

(7) $L_{\theta\theta}(\theta_0)$ is non-singular [local identification]

The mixing condition (1) requires a word of explanation.[2] A stochastic process x^t is called φ-mixing where φ is an infinite sequence $\varphi = (\varphi_0, \varphi_1, \ldots)$ iff any event F_t defined by $x^1, \ldots, x^t$ and event F^{t+k} defined by $x^{t+k}, x^{t+k+1}, \ldots$ satisfy $|pr(F^{t+k}|F_t) - pr(F^{t+k})| \leq \varphi_k$.

Roughly this asserts that the distant future is largely independent of the past. Mixing is restrictive in that some common processes such as a normal AR(1) aren't mixing. However, a stationary AR(1) with bounded innovations is mixing. Indeed, all Markov processes which satisfy Doeblin's condition and have a single ergodic class without cyclically moving subsets satisfy

assumption (1).[3] Thus, most processes which are observed can reasonably be argued to satisfy this condition.

An important (and obvious) fact about mixing processes is that functions of mixing processes depending on a fixed finite number of lagged observations are also mixing. Thus, the fact that x^t satisfies assumption (1) implies that $\lambda^t(\theta)$, $r_k^t(\theta)$, etc. all satisfy (1) as well. This is very convenient in a non-linear context.

First we show that θ^T is consistent.[4] By the uniform weak law of large numbers in the appendix together with assumptions (1) and (6) $L^T(\theta)$ converges uniformly in probability to $L(\theta)$. By a standard argument (3) implies $L(\theta)$ has a unique maximum at θ_0.[5] These facts imply via another standard argument, assumption (2) and the definition of θ^T that plim $(\theta^T) = \theta_0$.[6]

Turning to asymptotic normality by (4),(5) and the usual Taylor series expansion

$$(8)\ \sqrt{T}\,(\theta^T-\theta_0) = [L^T_{\theta\theta}(\bar{\theta}^T)]^{-1}\ (1/\sqrt{T})\Sigma_{t=1}^T \lambda^t_\theta(\theta_0)$$

where plim $(\bar{\theta}^T) = \theta_0$. By assumption (6) the uniform weak law of large numbers implies $L^T_{\theta\theta}(\theta)$ converges uniformly to $E\lambda^t_{\theta\theta}(\theta)$. Also by (6) $\lambda^t_\theta(\theta)$ and $\lambda^t_{\theta\theta}(\theta)$ are absolutely integrable so that[7]

$$(9)\ L_{\theta\theta}(\theta) = E\lambda^t_{\theta\theta}(\theta) = \partial^2[E\lambda^t(\theta)]/\partial\theta^2.$$

Thus, by assumption (7) and a standard lemma, plim $[L^T_{\theta\theta}(\bar{\theta}^T)]^{-1} = L^{-1}_{\theta\theta}(\theta_0)$.[8]

As in (9) assumption (6) implies

$$(10)\ L_\theta(\theta) = E\lambda^t_\theta(\theta) = \partial[E\lambda^t(\theta)]/\partial\theta.$$

Since $L(\theta)$ is C^2 by assumption (5) and θ_0 (the unique maximizing parameter in Θ) is in the interior of Θ by assumption (4) $L_\theta(\theta_0) = 0$. Thus by (10) $E\lambda_\theta^t(\theta_0) = 0$. This and assumptions (1) and (6) show that $(1/\sqrt{T})\ \Sigma_{t=1}^T \lambda_\theta^t(\theta_0)$ satisfies the hypotheses of the central limit theorem for mixing processes so that[9]

$$(11)\quad (1/\sqrt{T})\ \Sigma_{t=1}^T \lambda_\theta^t(\theta_0) \xrightarrow{D} N(0,V)$$

$$(12)\quad V = R_0(\theta_0) + 2\,\Sigma_{k=1}^{\infty} R_k(\theta_0).$$

From this it follows that

$$(13)\quad \sqrt{T}(\theta^T-\theta_0) \xrightarrow{D} N(0,\ L_{\theta\theta}^{-1}(\theta_0)VL_{\theta\theta}^{-1}(\theta_0)).$$

It remains to provide a consistent estimator of $L_{\theta\theta}^{-1}(\theta_0)VL_{\theta\theta}^{-1}(\theta_0)$. The matrix $L_{\theta\theta}^{-1}(\theta_0)$ can be consistently replaced with $[L_{\theta\theta}^T(\theta^T)]^{-1}$ for the same reason discussed above that $[L_{\theta\theta}^T(\bar{\theta}^T)]^{-1}$ is consistent. To estimate V observe that by (6) and line (A-1) in appendix (A) the uniform weak law of large numbers implies plim $R_k^T(\theta^T) = R_k(\theta_0)$. Suppose then that we wish to approximate V to within ε with probability $1-\alpha$. We choose K large enough that $[|V-R_0(\theta_0) - 2\Sigma_{j=1}^{k} R_j(\theta_0)|<\varepsilon/2.]$ Then we choose the sample size T large enough that $0 \le j \le k$ with probability $1-\alpha/(k+1)$ or more $|R_k^T(\theta^T) - R_k(\theta_0)|<\varepsilon/2(k+1)$. We conclude that

$$(14)\quad V^T = R_o^T(\theta^T)+ 2\Sigma_{j=1}^{k} R_j^T(\theta^T)$$

is the required estimator. As a practical rule the number of autocorrelations (k) which are used must be a fairly small fraction of the sample size actually available.

Assumptions (1) - (7) have been chosen for pedagogical reasons: the results described above hold under much weaker circumstances. The nature of the proof is such that these assumptions may easily be replaced. For example, the statement that mixing implies the central limit theorem can be replaced by any other assumptions (e.g. ergodicity) which imply the central limit theorem.

A useful feature of the variance correction formula (14) is that it protects against heteroskedasticity as well as serial correlation.[10] I hope that these results will be useful in models such as switching regressions where it is impractical to explicitly model serial correlation in the likelihood function.

APPENDIX

Uniform Weak Law of Large Numbers [for Stationary Mixing Processes]:

Let $h^t(\theta)$ be stationary and $\varphi(\theta)$-mixing with $\Sigma_{t=0}^{\infty}\varphi_t^{1/2}(\theta) \leq B$ $E|h^t(\theta)|^2 \leq B$ for $\theta\epsilon\Theta$. Define $H^T(\theta) \equiv (1/T)\Sigma_{t=1}^{T} h^t(\theta)$ and $H(\theta) \equiv Eh^t(\theta)$. Then $H^T(\theta)$ converges in probability to $H(\theta)$ uniformly over Θ.

proof:

The essential fact from lemma (1) p. 170 in Billingsley [2] is that

$$(A\text{-}1) \quad |\mathrm{cov}(h^t(\theta), h^{t+k}(\theta))| \leq 2\varphi_k^{1/2}B.$$

We can now apply Chebychev's inequality

$$\begin{aligned}(A\text{-}2) \quad & \mathrm{pr}(|H^T(\theta)-H(\theta)|>\epsilon \\ & \leq \mathrm{var}(H^T(\theta))/\epsilon^2 \\ & = (1/T^2)\Sigma_{t=1}^{T}\Sigma_{s=1}^{t}\mathrm{cov}(h^t(\theta),h^s(\theta))/\epsilon^2 \\ & \leq (4/T^2)\Sigma_{s=1}^{T}\Sigma_{s=1}^{t}\varphi_{t-s}^{1/2}B/\epsilon^2 \\ & \leq (4/T)B^2/\epsilon^2.\end{aligned}$$

Since the final expression converges to zero uniformly in θ as $T\to\infty$ QED.

NOTES

(1) See Hansen [7].

(2) A discussion of mixing is found in Billingsley [2] chapter 4.

(3) See Doob [5] chapter 6.

(4) An alternative proof for a special class of time series models is in Kohn [8].

(5) See Bowden [3].

(6) A proof of this using strong convergence is in Frydman [6]. The weak proof can be found in the appendix to Levine [9].

(7) Interchange of differentiation and expectation is discussed in Cramer [4] section 7.3.

(8) See Amemiya [1] or the appendix to Levine [9].

(9) Billingsley [2] Theorem 20.1.

(10) See White [10] for the application of this estimator to provide robustness against heteroskedasticity in OLS. MLE in discrete choice models is ordinarily inconsistent if heteroskedasticity is present.

REFERENCES

[1] Amemiya, T., "Non-linear 2SLS," Journal of Econometrics, 2 (1974).

[2] Billingsley, Patrick, Convergence of Probability Measures, John Wiley & Sons (New York: 1968).

[3] Bowden, R., "Theory of Parametric Identification," Econometrica, 1973.

[4] Cramer, H., Mathematical Methods of Statistics, Princeton University Press (Princeton: 1946).

[5] Doob, J.L., Stochastic Processes, John Wiley & Sons (New York: 1953).

[6] Frydman, R., "Consistency of MLE in Non-Linear Regression with Autocorrelation," Econometrica, May 1980.

[7] Hansen, Lars Peter, "Asymptotic Distribution of Least Squares with Endogenous Regressors and Dependent Residuals," Carnegie-Mellon University, March 1979.

[8] Kohn, R., "Local and Global Identification and Strong Consistency in Time Series Models," J. Econometrics, 8, 1978, 269-94.

[9] Levine, David, "Impact of Small Errors-in-Variables on MLE," MIT, July 1980.

[10] White, Halbert, "Heteroskedasticity Consistent Covariance Matrix Estimator," Econometrica, May 1980.