Journal of Econometrics 23 (1983) 337–342. North-Holland

A REMARK ON SERIAL CORRELATION IN MAXIMUM LIKELIHOOD

David LEVINE*

University of California, Los Angeles, CA 90024, USA

Received September 1981, final version received April 1983

Maximum likelihood estimation can be consistent and asymptotically normal despite serial correlation in the residuals. The usual estimator of the asymptotic covariance of the parameter estimator is inconsistent, but an alternative consistent estimator is derived.

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 *not* imply that $\exp TL^T(\theta) \equiv \prod_{t=1}^T f(y^t|z^t, \theta_0)$ is the joint density of the y^t (conditional, or otherwise), nor that $L^T(\theta)$ is the log-likelihood function for *any* model. 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 *only* on $f(y|z, \theta_0)$ being the actual density of y conditional on z and *not* on $\prod_{t=1}^T f(y^t|z^t, \theta_0)$ 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 $L^T(\theta) \equiv (1/T)\sum_{t=1}^T \lambda^t(\theta)$. Subscript θ's denote differentiation. Thus

*I would like to thank Jerry Hausman, Paul Ruud, Michael Veall, Halbert White, Sweder van Wijnbergen, and two anonymous referees.

[1]See Hansen (1979) or White and Domowitz (1981).

the score contributions are $\lambda_\theta^t(\theta)$. Associated with these are the autocorrelation functions

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

$$R_k(\theta) \equiv \mathrm{E} r_k^t(\theta),$$

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

For pedagogical purposes we make the following assumptions:

(1) x^t is stationary and strong α-mixing with exponentially declining weights α.

(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[2] to $f(y^t|z^t,\theta)$ unless $\theta=\theta_0$ (global identification).

(4) $\theta_0 \in$ interior (Θ).

(5) λ^t is a twice continuously differentiable function of θ.

(6) For some $\delta>0$,

$$\mathrm{E}\sup_\theta |\lambda^t(\theta)|^{2+\delta} \leqq B, \quad \mathrm{E}\sup_\theta |\lambda_\theta^t(\theta)|^{4+\delta} \leqq B', \quad \mathrm{E}\sup_\theta |\lambda_{\theta\theta}^t(\theta)|^{2+\delta} \leqq B''.$$

(7) $L_{\theta\theta}(\theta_0)$ is non-singular[3] (local identification).

The mixing condition (1) requires a word of explanation. A stochastic process x^t is called strong α-mixing where α is an infinite sequence of scalars $\alpha=(\alpha_0,\alpha_1,\ldots)$ with $\lim \alpha_k=0$ 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

$$|\mathrm{pr}(F^{t+k},F_t) - \mathrm{pr}(F^{t+k})\,\mathrm{pr}(F_t)| \leqq \alpha_k.$$

Since in the independent case $\alpha_k \equiv 0$, this asserts that the distant future is largely independent of the past. Strong mixing is a weak condition in that most common processes such as the normal ARMA processes satisfy (1).

[2]Stochastic equivalence means that the two densities are almost everywhere equal. In practice, since we do not know θ, we must check that, for all $\theta \neq \theta'$, $f(x^t,\theta) \neq f(x^t,\theta')$ a.e.

[3]In practice, we must check that $L_{\theta\theta}(\theta)$ is non-singular for all θ.

Indeed, most processes which are observed can reasonably be argued to satisfy this condition. An extensive discussion of conditions (1) and (6), which can be weakened substantially, can be found in White and Domowitz (1981).

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.

Theorem 1. If (1)–(4) and (6) hold, θ^T is strong consistent.[4]

Proof. By the uniform strong law of large numbers of White and Domowitz (1981)[5] together with assumptions (1) and (6), $L^T(\theta)$ almost surely (a.s.) converges uniformly to $L(\theta)$. By a trivial modification of a classic argument (3) implies $L(\theta)$ has a unique maximum at θ_0.[6] These facts imply, via an argument due to Frydman (1980), assumption (2) and the definition of θ^T, that $\theta^T \overset{\text{a.s.}}{\to} \theta_0$. Q.E.D.

Turning to asymptotic normality by assumptions (2), (4), (5) and the usual Taylor series expansion there exists a $\bar{L}_{\theta\theta}^T$ such that

$$(8) \qquad \sqrt{T}(\theta^T - \theta_0) = [\bar{L}_{\theta\theta}^T]^{-1}(1/\sqrt{T}) \sum_{t=1}^{T} \lambda_\theta^t(\theta_0).$$

Here the rows of $\bar{L}_{\theta\theta}^T$ are the rows of $L_{\theta\theta}^T$ evaluated at a point (which may depend upon the row) between θ_0 and θ^T.[7] By assumption (6) the uniform strong law of large numbers implies each row of $L_{\theta\theta}^T(\theta)$ converges uniformly to the corresponding row of $E\lambda_{\theta\theta}^T(\theta)$. Also by (6), $\lambda_\theta^t(\theta)$ and $\lambda_{\theta\theta}^t(\theta)$ are absolutely integrable[8] so that

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

Thus plim $\bar{L}_{\theta\theta}^T = L_{\theta\theta}(\theta_0)$ and by assumption (7) plim $[\bar{L}_{\theta\theta}^T]^{-1} = L_{\theta\theta}^{-1}(\theta_0)$.

As in (9), since assumption (6) permits the exchange of differentiation and integration, $L(\theta)$ is twice continuously differentiable by assumption (5) and

$$(10) \qquad L_\theta(\theta) = \mathrm{E}\lambda_\theta^t(\theta) = \partial[\mathrm{E}\lambda^t(\theta)]/\partial\theta.$$

[4]An alternative proof for a special class of time series models is in Kohn (1978).

[5]McLeish (1975) shows how to prove the strong law for mixing processes; Hoadley (1971) shows how to extend the strong law in the independent case to a uniform strong law; White and Domowitz (1981) show how to combine the two proofs to get a uniform strong law for mixing processes.

[6]See Wald (1949).

[7]The measurability of these random points is demonstrated in Lemma 3 of Jennrich (1969).

[8]Interchange of differentiation and expectation is discussed in Cramer (1946).

Since by assumption (4) θ_0 (the unique maximizing parameter in Θ) is in the interior of Θ, $L_\theta(\theta_0)=0$. Thus by (10) $E\lambda_\theta^t(\theta_0)=0$. This and assumptions (1) and (6) show that $(1/\sqrt{T})\sum_{t=1}^{T}\lambda_\theta^t(\theta_0)$ satisfies the hypotheses of the Rosenblatt Central Limit Theorem[9] for strong mixing processes so that

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

$$(12)\qquad V=R_0(\theta_0)+2\sum_{k=1}^{\infty}R_k(\theta_0).$$

From this follows:

Theorem 2. *If (1)–(7) hold,* $\sqrt{T}(\theta^T-\theta_0)\xrightarrow{D}N(0,L_{\theta\theta}^{-1}(\theta_0)VL_{\theta\theta}^{-1}(\theta_0))$.

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 $[\bar{L}_{\theta\theta}^T]^{-1}$ is consistent. Thus we must find a consistent estimator for V. Define

$$(13)\qquad V_k^T(\theta)\equiv R_0^T(\theta)+2\sum_{j=1}^{k}R_j^T(\theta),\qquad V_k(\theta)\equiv R_0(\theta)+2\sum_{j=1}^{k}R_j(\theta).$$

From the mixing assumption (1) it follows that

$$(14)\qquad V(\theta)\equiv\lim_{k\to\infty}V_k(\theta).$$

exists and that the convergence is uniform in θ.

Furthermore for fixed k it follows from assumptions (1) and (6) and the uniform strong law of large numbers that $V_k^T(\theta)\xrightarrow{a.s.}V_k(\theta)$ uniformly.[10] The difficult question is how to choose k, both as a function of T and possibly of the sample as well. Note that if we can find $k(T)$ such that $V_{k(T)}^T(\theta)\xrightarrow{a.s.}V(\theta)$ uniformly then $V_{k(T)}^T(\theta^T)$ will almost surely converge to V. White and Domowitz (1981) show for a special class of distributions that, if $k(T)\alpha T^\gamma$ for $0<\gamma<\frac{1}{2}$, then $V_{k(T)}^T(\theta)\xrightarrow{a.s.}V(\theta)$ uniformly and conjecture this result holds more generally.[11] If we are intested only in the *weak* consistency of $V_{k(T)}^T(\theta^T)$,

[9]Blum and Rosenblatt (1956) show that in the case where $\lambda_\theta^t(\theta_0)$ is a scalar $C_n\sum_{t=1}^{\infty}\lambda_\theta^t(\theta_0)$ converges to normality for some weights C_n. The Cramer–Wald device and computation of the variance of $\sum_{t=1}^{n}\lambda_\theta^t(\theta_0)$ yields the stated result.

[10]See White (1980) 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.

[11]In a conversation with the author.

it is straightforward to show $0<\gamma<\frac{1}{3}$ works in general. By Amemiya's Lemma and Chebychev's Inequality is suffices to show that

$$\mathrm{E}\left|V_{k(T)}^{T}(\theta)-V_{k(T)}(\theta)\right|^{2}\to 0 \quad \text{uniformly in } \theta.$$

However, $V_k^T(\theta)$ is the sum of the $R_j^T(\theta)$; by McLeish's (1975) covariance bound and assumption (6),

$$\mathrm{E}\left|R_j^T(\theta)-R_j(\theta)\right|^2 \leqq CT^{\gamma}/(T-T^{\gamma}) \quad \text{for a fixed constant } C.$$

Thus regardless of how highly correlated the R_j^T are,

$$\mathrm{E}\left|V_{k(T)}^{T}(\theta)-V_{k(T)}(\theta)\right|^{2} \leqq CT^{3\gamma}/(T-T^{\gamma})\to 0. \tag{15}$$

As a practical matter knowledge of the rate at which k should be increased with sample size isn't very helpful in dealing with a sample of fixed size. The issues in choosing k are these: if k is too large then the estimates R_j^T for j near k will not be very reliable as they are based on only $T-j$ observations. On the other hand, if k is too small V_k is not going to approximate V very well since $2\sum_{j=k+1}^{\infty} R_j$ is omitted. Indeed the error in estimating V is at *least* $2\sum_{j=T}^{\infty} R_j$, which among other things is unobservable. A crude rule of thumb is this: under the mixing assumption (1) $|R_j|$ must decline exponentially.[12] Choose k so that there are 'reasonably' many observations with which to estimate R_k. Then use the estimated $|R_0^T|,|R_1^T|,\ldots,|R_k^T|$ to find an exponentially declining upper bound: this can be extrapolated to give some idea of how large the error $2\sum_{j=k+1}^{\infty}|R_j|$ might be. I should note that the problem of choosing k correctly is a fundamental problem with asymptotic theory: how large is large? Ultimately only finite sample theory can answer this question.

[12]See McLeish (1975).

References

Blum, J. and M. Rosenblatt, 1956, A class of stationary processes and a central limit theorem, Proceedings of the National Academy of Science 42, 412–413.

Cramer, H., 1946, Mathematical methods of statistics (Princeton University Press, Princeton, NJ).

Frydman, R., 1980, A proof of the consistency of maximum likelihood estimators of non-linear regression models with autocorrelated errors, Econometrica 48, 853–860.

Hansen, L.P., 1979, Asymptotic distribution of least squares with endogenous regressors and dependent residuals, March (Carnegie-Mellon University, Pittsburgh, PA).

Hoadley, B., 1971, Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case, Annals of Mathematical Statistics 42, 1977–1991.

Jennrich, R., 1969, Asymptotic properties of non-linear least squares estimators, Annals of Mathematical Statistics 40, 633–643.

Kohn, R., 1978, Local and global identification and strong consistency in time series models, Journal of Econometrics 8, 269–294.
McLeish, D., 1975, A maximal inequality and strong dependent laws, Annals of Probability 3, 826–836.
Wald, A., 1949, Note on the consistency of MLE, Annals of Mathematical Statistics 20, 595–601.
White, H., 1980, Heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica 48, 817–838.
White, H. and I. Domowitz, 1981, Nonlinear regression with dependent observations (University of California, San Diego, CA).