The Annals of Statistics
1986, Vol. 14, No. 1, 355–357

PSEUDO MAXIMUM LIKELIHOOD ESTIMATION: THE ASYMPTOTIC DISTRIBUTION

By William R. Parke

University of California, Santa Barbara

Gong and Samaniego (1981) define pseudo maximum likelihood estimation and derive the asymptotic distribution of the resulting estimates. This note gives a simpler and more elegant expression for the asymptotic variance of a pseudo maximum likelihood estimate.

1. Introduction. In the presence of computational and algebraic obstacles to computing maximum likelihood estimates for a given model, the following pseudo maximum likelihood estimation procedure may be of some value. Let the likelihood function $L_n(\theta, \pi)$ for a sample of size n be defined over two parameter vectors, θ and π. Suppose that maximizing $L_n(\theta, \pi)$ over both θ and π is difficult, but that some alternative procedure (other than maximizing $L_n(\theta, \pi)$) yields an estimate $\tilde{\pi}_n$. A pseudo MLE $\hat{\theta}_n(\tilde{\pi}_n)$ can then be computed by maximizing $L_n(\theta, \tilde{\pi}_n)$ over θ. Gong and Samaniego (1981) discuss this general procedure and derive the asymptotic distribution of the pseudo MLE $\hat{\theta}_n(\tilde{\pi}_n)$.

This note improves upon the expression that Gong and Samaniego give for the asymptotic variance–covariance matrix of the pseudo MLE $\hat{\theta}_n(\tilde{\pi}_n)$. Their expression contains a term that is shown here to equal zero for all pseudo maximum likelihood estimation problems satisfying the standard regularity conditions that they specified. While this point could be established for each particular application (as Gong and Samaniego do for their signal plus noise application), the present proof covers the general case.

The possible practical advantages of pseudo maximum likelihood estimation are realized for a diverse range of estimation problems. Gong and Samaniego use the technique to simplify the calculations for a signal plus noise problem. Other applications include estimating a linear reduction of an otherwise nonlinear least squares estimation problem [Durbin (1960), Wallis (1967)], pooling time series and cross section data [Maddala (1971)], and avoiding likelihood function singularities in estimating large economic forecasting models [Fair and Parke (1980) and Parke (1985)].

2. Asymptotic theory. The information matrix ϑ for the vector $(\theta, \pi)'$ can be partitioned as

$$\vartheta = \begin{bmatrix} \vartheta_{11} & \vartheta_{12} \\ \vartheta_{21} & \vartheta_{22} \end{bmatrix}.$$

Let $l_n(\theta_0, \pi_0)$ denote $\log(L_n(\theta_0, \pi_0))$, let $\bar{l}_\theta(\theta_0, \pi_0)$ denote $n^{-1}\,\partial l_n(\theta, \pi)/\partial\theta$

Received November 1984; revised September 1985.
AMS 1980 *subject classifications*. Primary 62F12; secondary 62A10.
Key words and phrases. Likelihood, estimation, asymptotics, relative efficiency.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*. www.jstor.org

evaluated at (θ_0, π_0), and let the matrix Σ be defined by

$$\sqrt{n}\begin{bmatrix} \bar{l}_\theta(\theta_0, \pi_0) \\ \tilde{\pi}_n - \pi_0 \end{bmatrix} \to_d N\left(0, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right).$$

Gong and Samaniego show that, for the case of scalars θ and π,

$$\sqrt{n}\,(\hat{\theta}_n(\tilde{\pi}_n) - \theta_0) \to_d N(0, \sigma^2),$$

where

$$\sigma^2 = 1/\vartheta_{11} + (\vartheta_{12}/\vartheta_{11})^2\Sigma_{22} - 2(\vartheta_{12}/\vartheta_{11}^2)\Sigma_{12}.$$

In fact, $\Sigma_{12} = 0$, and

$$\text{(2.1)}\quad \sqrt{n}\begin{bmatrix} \hat{\theta}_n(\tilde{\pi}_n) - \theta_0 \\ \tilde{\pi}_n - \pi_0 \end{bmatrix} \to_d N\left(0, \begin{bmatrix} \vartheta_{11}^{-1} + \vartheta_{11}^{-1}\vartheta_{12}\Sigma_{22}\vartheta_{21}\vartheta_{11}^{-1} & -\vartheta_{11}^{-1}\vartheta_{12}\Sigma_{22} \\ -\Sigma_{22}\vartheta_{21}\vartheta_{11}^{-1} & \Sigma_{22} \end{bmatrix}\right).$$

The proof of this result is given in Section 3.

The simplicity of this result arises from the surprising fact that Σ_{12} equals zero for any consistent estimate $\tilde{\pi}_n$. In general, statistics computed using the same data, as are $\bar{l}_\theta(\theta_0, \pi_0)$ and $\tilde{\pi}_n$, will be asymptotically correlated. A general result that this correlation is zero eliminates the potentially difficult task of deriving an expression for Σ_{12} for a particular estimation problem.

The asymptotic distributions of the pseudo MLE $\hat{\theta}_n(\tilde{\pi}_n)$ and the MLE $\hat{\theta}_n(\hat{\pi}_n)$ can be compared by inverting the information matrix ϑ to obtain

$$\text{(2.2)}\quad \sqrt{n}\begin{bmatrix} \hat{\theta}_n(\hat{\pi}_n) - \theta_0 \\ \hat{\pi}_n - \pi_0 \end{bmatrix} \to_d N\left(0, \begin{bmatrix} \vartheta_{11}^{-1} + \vartheta_{11}^{-1}\vartheta_{12}\Sigma_{22}^*\vartheta_{21}\vartheta_{11}^{-1} & -\vartheta_{11}^{-1}\vartheta_{12}\Sigma_{22}^* \\ -\Sigma_{22}^*\vartheta_{21}\vartheta_{11}^{-1} & \Sigma_{22}^* \end{bmatrix}\right),$$

where $\Sigma_{22}^* = (\vartheta_{22} - \vartheta_{21}\vartheta_{11}^{-1}\vartheta_{12})^{-1}$ is the asymptotic variance of the MLE $\hat{\pi}_n$. The expression (2.1) can in fact be obtained algebraically from (2.2) by simply substituting the asymptotic variance Σ_{22} of the alternative estimate $\tilde{\pi}_n$ for the asymptotic variance of the MLE $\hat{\pi}_n$. In both (2.1) and (2.2), the term ϑ_{11}^{-1} is the asymptotic variance of the estimate $\hat{\theta}_n(\pi_0)$ that could be computed if π_0 were known. The term $\vartheta_{11}^{-1}\vartheta_{12}\Sigma_{22}^*\vartheta_{21}\vartheta_{11}^{-1}$ in (2.2) is the minimum possible additional asymptotic variance in an estimate of θ_0 if π_0 is not known. This minimum is attained for the MLE $\hat{\theta}_n(\hat{\pi}_n)$ computed jointly with the MLE $\hat{\pi}_n$. For the pseudo MLE $\hat{\theta}_n(\tilde{\pi}_n)$, the corresponding term $\vartheta_{11}^{-1}\vartheta_{12}\Sigma_{22}\vartheta_{21}\vartheta_{11}^{-1}$ in (2.1) is the asymptotic variance that can be attributed to using the estimate $\tilde{\pi}_n$ of π_0. The asymptotic inefficiency of the pseudo MLE $\hat{\theta}_n(\tilde{\pi}_n)$ relative to the MLE $\hat{\theta}_n(\hat{\pi}_n)$ is thus given by $\vartheta_{11}^{-1}\vartheta_{12}(\Sigma_{22} - \Sigma_{22}^*)\vartheta_{21}\vartheta_{11}^{-1}$. The pseudo MLE $\hat{\theta}_n(\tilde{\pi}_n)$ is asymptotically efficient if ϑ_{12} equals zero (because the MLEs $\hat{\theta}_n(\hat{\pi}_n)$ and $\hat{\pi}_n$ are asymptotically uncorrelated) or if $\tilde{\pi}_n$ is asymptotically as efficient as the MLE $\hat{\pi}_n$.

3. Proof of (2.1). Consider the hypothetical estimate $\hat{\theta}_n(\pi_0)$ that could be computed if π_0 were known. Under the regularity conditions given in Gong and Samaniego (1981), the equation

$$\bar{l}_\theta(\theta, \pi_0) = 0$$

has a consistent root $\hat{\theta}_n(\pi_0)$. Expanding $\bar{l}_\theta(\hat{\theta}_n(\pi_0), \pi_0)$ about θ_0 yields

$$\sqrt{n}\,(\hat{\theta}_n(\pi_0) - \theta_0) = -\sqrt{n}\,\vartheta_{11}^{-1}\bar{l}_\theta(\theta_0, \pi_0) + o_p(1).$$

Differentiating with respect to π yields

$$\sqrt{n}\ \partial\hat{\theta}_n(\pi)/\partial\pi = \sqrt{n}\,\vartheta_{11}^{-1}\vartheta_{12} + o_p(1), \tag{3.1}$$

where $\partial\hat{\theta}_n(\pi)/\partial\pi$ is evaluated at π_0.

The asymptotic distribution of $\hat{\theta}_n(\tilde{\pi}_n)$ then follows from the Taylor series approximation:

$$\sqrt{n}\,(\hat{\theta}_n(\tilde{\pi}_n) - \theta_0) = \sqrt{n}\,(\hat{\theta}_n(\pi_0) - \theta_0) + \sqrt{n}\ \partial\hat{\theta}_n(\pi)/\partial\pi\,(\tilde{\pi}_n - \pi_0) + o_p(1).$$

Using (3.1),

$$\sqrt{n}\,(\hat{\theta}_n(\tilde{\pi}_n) - \theta_0) = \sqrt{n}\,(\hat{\theta}_n(\pi_0) - \theta_0) + \sqrt{n}\,\vartheta_{11}^{-1}\vartheta_{12}(\tilde{\pi}_n - \pi_0) + o_p(1). \tag{3.2}$$

Pierce (1982) shows that $\sqrt{n}\,(\hat{\theta}_n(\pi_0) - \theta_0)$ and $\sqrt{n}\,(\tilde{\pi}_n - \pi_0)$ must be asymptotically independent because $\hat{\theta}_n(\pi_0)$ is asymptotically efficient and $\sqrt{n}\,(\tilde{\pi}_n - \pi_0)$ has asymptotic mean zero. (Hausman (1978) gives a similar application of asymptotic efficiency to establish asymptotic independence.) The conclusion (2.1) follows from this last result and (3.2). □

Acknowledgment. The author would like to thank an associate editor for very helpful comments on an earlier version of this paper.

REFERENCES

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

FAIR, R. C. and W. R. PARKE (1980). Full information estimates of a nonlinear macroeconometric model. *J. Econometrics* **13** 269–291.

GONG, G. and F. J. SAMANIEGO (1981). Pseudo maximum likelihood estimation: Theory and applications. *Ann. Statist.* **9** 861–869.

HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica* **46** 1251–1271.

MADDALA, G. S. (1971). The likelihood approach to pooling cross-section and time-series data. *Econometrica* **39** 939–953.

PARKE, W. R. (1985). Two modified FIML estimators for use in small samples. To appear in *Internat. Econom. Rev.*

PIERCE, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Ann. Statist.* **10** 475–478.

WALLIS, K. F. (1967). Lagged dependent variables and serially correlated errors: A reappraisal of three-pass least squares. *Rev. Econom. Statist.* **49** 555–567.

DEPARTMENT OF ECONOMICS
UNIVERSITY OF CALIFORNIA
SANTA BARBARA, CALIFORNIA 93106