MONASH

WP 6-98

ISSN 1440-771X
ISBN 0 7326 1043 5

MONASH UNIVERSITY

ANCORA · IMPARO

AUSTRALIA

Comparisons of estimators and tests based on modified likelihood and message length functions

Mizan R. Laskar and Maxwell L. King

Working Paper 6/98
June 1998

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

COMPARISONS OF ESTIMATORS AND TESTS BASED ON MODIFIED LIKELIHOOD AND MESSAGE LENGTH FUNCTIONS

Mizan R. Laskar and Maxwell L. King
Department of Econometrics and Business Statistics
Monash University
Clayton, Victoria 3168
Australia

Abstract

The presence of nuisance parameters causes unexpected complications in econometric inference procedures. A number of modified likelihood and message length functions have been developed for better handling of nuisance parameters but all of them are not equally efficient. In this paper, we empirically compare different modified likelihood and message length functions in the context of estimation and testing of parameters from linear regression disturbances that follow either a first-order moving average or first-order autoregressive error processes. The results show that estimators based on the conditional profile likelihood and tests based on the marginal likelihood are best. If there is a minor identification problem, the sizes of the likelihood ratio and Wald tests based on simple message length functions are best. The true sizes of the Lagrange multiplier tests based on message length functions are rather poor because the score functions of message length functions are biased.

1. Introduction

Satisfactory statistical analysis of non-experimental data, is an important problem in econometrics. Econometric models usually involve a large number of influences, most of which are not of immediate interest. This means that such models contain two kinds of parameters, those of interest and those not of immediate interest that are known as nuisance parameters. Their presence causes unexpected complications in econometric inference. A fairly standard procedure in likelihood based statistical inference is to concentrate the likelihood function by replacing nuisance parameters by their respective maximum likelihood (ML) estimators conditional on the parameters of interest. In such situations, estimators and tests can perform poorly in small samples (Bewley 1986, Cox and Reid 1987, King 1987, King and McAleer 1987, Moulton and Randolph 1989, Chesher and Austin 1991). Earlier, Neyman and Scott (1948) warned that nuisance parameters can seriously compromise likelihood based inference. In relation to this, King (1996) observed that when nuisance parameters are present, statistical theory is generally less helpful in suggesting reliable diagnostic tests. Also, Cordus (1986) noted that the presence of nuisance parameters causes a shift in the estimated mean of the null distribution of the likelihood ratio test.

The question which then arises is which methods should be used to tackle the problem of nuisance parameters in order to improve estimators and tests. The marginal likelihood is one such method for handling nuisance parameters. Estimators and tests based on this likelihood have better small sample properties compared to those based on the classical likelihood function (Ara 1995, Cordus 1986, Rahman and King 1998). In the context of estimating variance components in the linear regression model, a related approach known as residual (or restricted) maximum likelihood (REML) (Patterson and Thompson 1971) has gained considerable importance. The marginal likelihoods cannot be constructed in all situations and REML applies only to the disturbances parameters in the linear model. As an alternative, Barndorff-Nielsen (1983) proposed the modified profile likelihood (MPL) and Cox and Reid (1987) initiated the idea of the conditional profile likelihood (CPL) which requires that the parameter(s) of interest and nuisance parameters are orthogonal. Also, using the combination of REML and CPL, Laskar and King (1998) derived the conditional profile restricted log-likelihood function (CPRL) for better handling of nuisance parameters. They investigated the small sample properties of estimators and tests based on this likelihood function and three other modified likelihood functions and compared with those based on the profile likelihood function.

An alternative approach, known as minimum message length (MML), is a information theoretic criteria for parameter estimation and model selection. The MML principle needs a prior distribution of the parameters, the square root of the determinant of the information matrix for the parameters and a likelihood function. In this context, Wallace and Dowe (1993) mentioned that the inclusion of the first two factors helps reduce the measure of uncertainty, their ratio is dimension free and invariant to reparameterization. Extending their research, Laskar and King (1996) derived six different message length functions using different prior distributions of the parameters and combinations of CPL and message length functions. They investigated the small sample properties of estimators based on these message length functions. Moreover, Laskar and King (1997) investigated the small sample properties of different tests based on these message length functions. There are many different modified likelihood and message length functions for handling nuisance parameters but for econometric problems where estimation and diagnostic testing are of main interest, all of them are not equally efficient. Thus, it is important to investigate and find out the best approaches for handling nuisance parameters.

The aim of this paper is to empirically compare all the likelihood and message length functions in the context of estimation and testing of parameters involved in the variance-covariance matrix of linear regression disturbances. We extend and compare the Monte Carlo results of Laskar and King (1996, 1997a, 1997b, 1998). This will enable us to recommend the best functions in estimation and testing problems. In section 2, different likelihood and message length functions are presented. A Monte Carlo experiment, conducted to compare the estimators and tests based on all the likelihood and message length functions are reported in section 3. Some concluding remarks are made in section 4.

2. Theory

Consider the linear regression model with non-spherical disturbances

$$y = X\beta + u;\ u \sim N(0, \sigma^2\Omega(\theta)) \qquad (1)$$

where y is $n\times 1$, X is $n\times k$, nonstochastic and of rank $k < n$, β is a $k\times 1$ vector, $\Omega(\theta)$ is a symmetric matrix and θ is a $p\times 1$ vector. This model generalizes a wide range of disturbance processes of the linear regression model of particular interest to statisticians and econometricians. These include all parametric forms of autocorrelated disturbances, all parametric forms of heteroscedasticity (in which case $\Omega(\theta)$ is a diagonal matrix), and error components models including those that result from random regression coefficients. The likelihood and log-likelihood for this model (excluding constants) are respectively

$$L(y;\theta,\sigma^2,\beta) \propto \sigma^{-n}|\Omega(\theta)|^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2}(y-X\beta)'\Omega(\theta)^{-1}(y-X\beta)\}, \tag{2}$$

$$l(y;\theta,\sigma^2,\beta) \propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|\Omega(\theta)| - \frac{1}{2\sigma^2}(y-X\beta)'\Omega(\theta)^{-1}(y-X\beta) \tag{3}$$

and the log profile (or concentrated) likelihood is

$$l_p(y;\theta) \propto -\frac{n}{2}\log\hat{\sigma}_\theta^2 - \frac{1}{2}\log|\Omega(\theta)| \tag{4}$$

where $\hat{\sigma}_\theta^2 = (y - X\hat{\beta}_\theta)'\Omega(\theta)^{-1}(y - X\hat{\beta}_\theta)/n$ and $\hat{\beta}_\theta = (X'\Omega(\theta)^{-1}X)^{-1}X'\Omega(\theta)^{-1}y$.

2.1. Modified Likelihood Functions

Tunnicliffe and Wilson (1989) derived the marginal likelihood for θ in (1) as

$$\ell_m(y;\theta) = -\frac{1}{2}\log|\Omega(\theta)| - \frac{1}{2}\log|X'\,\Omega(\theta)^{-1}X| - \frac{m}{2}\log(\hat{u}'\,\Omega(\theta)^{-1}\hat{u}) \tag{5}$$

where $m = n - k$. Using the combination of REML and CPL, Laskar and King (1998) derived the CPRL function of θ for model (1) as

$$\bar{\ell}^*_{cpe}(y;\theta) = -\frac{(m-2)}{2m}\left[\log|\Omega(\theta)| - \log|X'\,\Omega(\theta)^{-1}X| - m\log(\hat{u}'\,\Omega(\theta)^{-1}\hat{u})\right] \tag{6}$$

Using the idea of Cox and Reid (1987), Laskar (1998) derived the CPL for θ in (1) as

$$l_{cp}(y;\theta) = -\frac{1}{2}\log|X'\,\Omega(\theta)^{-1}X| - \frac{(n-2)}{2n}\log|\Omega(\theta)| - \frac{(m-2)}{2}\log(\hat{u}'\Omega(\theta)^{-1}\hat{u}). \tag{7}$$

Based on the idea of Cox and Reid (1993), Laskar (1998) also derived an approximate conditional profile likelihood (ACPL) for θ in (1) as

$$l_{acp}(y;\theta) = -\frac{m-2}{2}\log(\hat{u}'\Omega(\theta)^{-1}\hat{u}) - \frac{1}{2}\log|\Omega(\theta)| - \frac{1}{2}\log|X'\Omega(\theta)^{-1}X|$$

$$+\frac{1}{n}tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta}). \qquad (8)$$

From (5) and (6)

$$\bar{l}^{*}_{cpr}(y;\theta)=\frac{(m-2)}{m}l_m(y;\theta)$$

so that for the purposes of estimating θ, the marginal likelihood function and the CPRL are equivalent. This is not necessarily true for likelihood based tests of θ, because scores, Hessians and maximized likelihood will be different, although any differences will obviously disappear as n increases.

2.2. Message Length Functions

Minimum message length is a Bayesian method which chooses estimators to minimize the length of an encoded form of the data made up of a model and the deviations from that model (residuals). Wallace and Dowe (1993) state that the MML principle is that the best possible conclusion to draw from the data is the theory which maximizes the product of the probability of the data occurring in the light of the theory with the prior probability of that theory.

For model (1), an approximate message length function given by Wallace and Freeman (1987) and accurate to $\delta=1/\sqrt{K_s^s F(\theta,\sigma^2,\beta)}$ is

$$-\log\left[\frac{\pi(\theta,\sigma^2,\beta)L(\theta,\sigma^2,\beta)}{\sqrt{F(\theta,\sigma^2\beta)}}\right]+\frac{s}{2}(1+\log K_s) \qquad (9)$$

where $\pi(\theta,\sigma^2,\beta)$ is a prior density for $\gamma=(\theta',\sigma^2,\beta',)'$, $F(\theta,\sigma^2,\beta)$ is the determinant of the information matrix, $s=k+p+1$, K_s is the s dimensional lattice constant which is independent of parameters, as given by Conway and Sloan (1988, p. 59-61). For example $K_1=\frac{1}{12}$, $K_2=\frac{5}{36\sqrt{3}}$ and $K_3=\frac{19}{36.\sqrt[3]{2}}$. Wallace and Dowe (1994) mentioned, maximizing (9) is equivalent to maximizing the average of the log-likelihood function over region of size proportional to $1/\sqrt{F(\theta,\sigma^2,\beta)}$ while the ML estimator maximizes the likelihood function at a single point. The value of θ which minimizes (9) is the MML estimate of θ with accuracy $\delta=1/\sqrt{K_s^s F(\theta,\sigma^2,\beta)})$.

Inclusion of $\pi(\theta,\sigma^2,\beta)$ and $\sqrt{F(\theta,\sigma^2,\beta)}$ help reduce the measure of uncertainty, their ratio is dimension free and invariant to reparameterization (Wallace and Dowe 1993). Since MML is a Bayesian method and depends on the choice of prior density of the parameters, there is scope in selecting the prior. Using different prior densities and combinations of CPL and message length functions, Laskar and King (1996) derived six different message length functions which are

$$\begin{aligned} ML_1 = {} & \frac{m-1}{2}\log\sigma^2 + \frac{1}{2}\log|\Omega(\theta)| + \frac{1}{2\sigma^2}u'\Omega(\theta)^{-1}u + \frac{1}{2}\log\left|X'\Omega(\theta)^{-1}X\right| \\ & + \frac{1}{2}\log\left(n \times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2\right) \\ & + \frac{s}{2}(1+\log K_s) - \log 2, \end{aligned} \tag{10}$$

$$\begin{aligned} ML_2 = {} & \frac{m}{2}\log\sigma^2 + \frac{1}{2}\log|\Omega(\theta)| + \frac{1}{2\sigma^2}u'\Omega(\theta)^{-1}u + \frac{1}{2}\log\left|X'\Omega(\theta)^{-1}X\right| \\ & + \frac{1}{2}\log\left(n \times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2\right) \\ & + \frac{s}{2}(1+\log K_s) - \log 2, \end{aligned} \tag{11}$$

$$\begin{aligned} CPML_1 = {} & \frac{m-k-3}{2}\log\hat{\delta}_1 + \frac{k+1}{n+k+1}\log|\Omega(\theta)| + \log\left|X_\theta^{\dagger\prime}X_\theta^\dagger\right| \\ & + \frac{1}{2}\log\left(n \times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2\right) \end{aligned} \tag{12}$$

where $u_\theta^\dagger = y_\theta^\dagger - X_\theta^\dagger\beta$, $X_\theta^\dagger = D(\theta)^{-\frac{1}{2}}X$, $D(\theta) = \Omega(\theta)/|\Omega(\theta)|^{\frac{1}{(n+k+1)}}$, $y_\theta^\dagger = D(\theta)^{-\frac{1}{2}}y$, $\hat{\delta}_1 = \hat{u}_\theta^{\dagger\prime}\hat{u}_\theta^\dagger/(n-k-1)$, $\hat{u}_\theta^\dagger = y_\theta^\dagger - X_\theta^\dagger\hat{\beta}_\theta^\dagger$ and $\hat{\beta}_\theta^\dagger = (X_\theta^{\dagger\prime}X_\theta^\dagger)^{-1}X_\theta^{\dagger\prime}y_\theta^\dagger$.

$$\begin{aligned} CPML_2 = {} & \frac{m-k-2}{2}\log\hat{\delta}_2 + \frac{k}{n+k}\log|\Omega(\theta)| + \log\left|X_\theta^{*\prime}X_\theta^*\right| \\ & + \frac{1}{2}\log\left(tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2\right) \end{aligned} \tag{13}$$

where $\hat{\delta}_2 = \hat{u}_\theta^{*\prime}\hat{u}_\theta^*/m$, $\hat{u}_\theta^* = y_\theta^* - X_\theta^*\hat{\beta}_\theta^*$, $\hat{\beta}_\theta^* = (X_\theta^{*\prime}X_\theta^*)^{-1}X_\theta^{*\prime}y_\theta^*$, $X_\theta^* = G_1(\theta)^{-\frac{1}{2}}X$, $y_\theta^* = G_1(\theta)^{-\frac{1}{2}}y$ and $G_1(\theta) = \Omega(\theta)/|\Omega(\theta)|^{\frac{1}{(n+k)}}$.

$$AML_1 = \frac{m-1}{2}\log\delta + \frac{1}{2\delta}u_\theta' u_\theta + \frac{1}{2}\log\left|X_\theta' X_\theta\right| + \frac{1}{2}\log\left|C(\theta)\right|, \tag{14}$$

$$AML_2 = \frac{m}{2}\log\delta + \frac{1}{2\delta}u_\theta' u_\theta + \frac{1}{2}\log\left|X_\theta' X_\theta\right| + \frac{1}{2}\log\left|C(\theta)\right| \tag{15}$$

where $\sigma^2 = \delta / |\Omega(\theta)|^{\frac{1}{n}}$, $G(\theta)$ is an $n\times n$ matrix comprised of $\Omega(\theta)$ with each element divided by $|\Omega(\theta)|^{\frac{1}{n}}$ and the $(i,j)^{\text{th}}$ element of the $p\times p$ matrix $C(\theta)$ is $\frac{1}{2}tr\left[\frac{\partial^2 G(\theta)^{-1}}{\partial\theta_i\partial\theta_j}G(\theta)\right]$.

Details of the LR, LM, Wald, AW and NW tests based on all the likelihood and message length functions in the context of testing H$_0$: $\theta = \theta_0$ against H$_a$: $\theta \neq \theta_0$ in (1) are given in Laskar and King (1998), Laskar and King (1997a) and Laskar and King (1997b). Laskar and King (1998) estimated the MA(1) disturbances parameter constrained between -1 to 1, because of the identification problem for MA(1) disturbances. It is well known that there is a non-zero probability of getting ML estimators of -1 or 1 for MA(1) disturbances parameter (Shephard 1993). The score with respect to the MA(1) parameter is discontinuous and the information matrix is not well defined at those two points. As a result, Laskar and King (1998) faced the problem of nonmonotonicity of the power curve of the Wald test. They initially tackled this problem by rejecting the null hypothesis whenever the estimate of the MA(1) disturbance parameter is ±1 and called this the AW test. Unfortunately the AW test cannot totally solve this problem because it takes into account boundary values of the parameter estimates only. The power curve may be nonmonotonic at some other points of the parameter space. Laskar and King (1997a) fully overcame this problem by replacing the unknown parameter values in the variance component of the Wald test with their null hypothesis values rather than their estimated values and denoted it as the NW test.

3. Monte Carlo Experiment

Laskar and King (1998) investigated the small sample properties of estimators and LR, LM, Wald and AW tests based on different modified likelihood functions in the context of MA(1) and AR(1) regression disturbances. Also, Laskar and King (1997a) investigated the small sample properties of NW tests based on different modified likelihood functions in the context of MA(1) regression disturbances. When message length functions based estimation and testing are concern, Laskar and King (1996) investigated the small sample properties of estimators in the context of MA(1) regression disturbances and Laskar and King (1997b) investigated the small sample properties of tests in the context of MA(1) regression disturbances.

In order to compare the small sample properties of estimators and small sample size and power properties of the LR, LM, Wald, AW and NW tests for testing H$_0$: $\gamma = 0$ for MA(1) regression disturbances or H$_0$: $\rho = 0$ for AR(1) regression disturbances i.e. H$_0$: $\theta = 0$ based on different modified likelihoods, classical (profile) likelihood and message length functions, we considered results from above papers and further a Monte Carlo experiment was conducted for computing the estimators and small sample sizes and powers based on message length functions with the disturbances of (1) generated by the AR(1) process

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad (16)$$

in which $\varepsilon_t \sim IN(0,\sigma^2)$, $t = 0,1,\ldots,n$. Under (16), $u \sim N(0,\sigma^2\Omega(\rho))$, where $u_0 \sim N(0,\sigma^2/(1-\rho^2))$, $\Omega(\rho)$ is the $n \times n$ symmetric matrix whose $(i,j)^{\text{th}}$ element is $\rho^{|i-j|}/(1-\rho^2)$. For the model (16), all the message length functions are not defined at $\rho = \pm 1$. So, the best way of tackling this problem is to restrict ρ to the interval

$$-0.9999 \le \rho \le 0.9999. \qquad (17)$$

For our purposes, the need to impose the restrictions (17), has a positive implication. Often when estimators are being investigated, there is uncertainty about which moments of the estimator's distribution exist. If, for example, the second-order moment does not exist, then any estimate of it obtained from a Monte Carlo experiment will be finite but meaningless. In our case, while we do not know the distributions of our estimators, the restrictions (17) implies that all moments will exist.

3.1. Experimental Design

The first part of the study covered a comparison of the different MML estimators for the AR(1) parameter. The estimates based on (i) ML$_1$, (ii) ML$_2$, (iii) CPML$_1$, (iv)

CPML$_2$, (v) AML$_1$ and (v) AML$_2$ when ρ = -0.8, -0.4, 0, 0.4, 0.8 were used for the first comparison. The second part involved a comparison of sizes of different tests using asymptotic critical values at the five percent level. The third part of the experiment was divided into two parts. In first part, the Monte Carlo method was used to estimate appropriate critical values of each of the tests in order to compare the powers of all tests at approximately the same level. These critical values were calculated using 2000 replications. In second part, powers of all the tests were calculated using these (simulated) critical values. The tests involved LR, LM, Wald and NW tests.

All the calculations were repeated 2000 times using the GAUSS (1996) constrained optimization routine but with particular care taken in choosing starting valus (see Laskar, 1998). The following X matrices were used with n = 30 and n = 60:

$X1$: (k = 5). A constant, quarterly Australian private capital movements, Government capital movements commencing 1968(1) and these two variables lagged one quarter as two additional regressors.

$X2$: (k = 3). A constant, quarterly seasonally adjusted Australian household disposable income and private final consumption expenditure commencing 1959(4).

$X3$: (k = 3). The regressors are the eigenvectors corresponding to the three smallest eigenvalues of the $n \times n$ tridiagonal matrix whose main diagonal elements are 2, except for the top left and bottom right elements which are both 1 and whose elements in the leading off-diagonals are all -1.

$X4$: (k = 2). A constant and a linear trend.

These matrices reflect a variety of behaviour. The capital movements regressors in $X1$ are rapidly changing with a high degree of seasonality. This is in contrast to the relatively smooth regressors $X2$ (seasonally adjusted quarterly data). The regressors in $X3$ are smoothly evolving and include an intercept. They cause the Durbin-Watson statistic, which is an approximately locally best one-sided test against both MA(1) and AR(1) disturbances (King and Evans 1988), to attain its upper bound. Also Laskar and King (1998), Laskar and King (1997a), Laskar and King (1997b) and Laskar and King (1996) considered the same set of X matrices.