A Monte Carlo study on two methods of calculating the MLEs covariance matrix in a seemingly unrelated nonlinear regression

Jensen, Mark J

1995

ECONOMETRIC REVIEWS, 14(3), 315-330 (1995)

A MONTE CARLO STUDY ON TWO METHODS OF CALCULATING THE MLE'S COVARIANCE MATRIX IN A SEEMINGLY UNRELATED NONLINEAR REGRESSION*

Mark J. Jensen

Department of Economics, Southern Illinois University
Carbondale, IL 62901

Keywords and Phrases: nonlinear SURE models, covariance estimates, hypothesis testing of neoclassical theory

JEL Classification: C30

ABSTRACT

Econometric techniques to estimate output supply systems, factor demand systems and consumer demand systems have often required estimating a nonlinear system of equations that have an additive error structure when written in reduced form. To calculate the ML estimate's covariance matrix of this nonlinear system one can either invert the Hessian of the concentrated log likelihood function, or invert the matrix calculated by pre-multiplying and post-multiplying the inverted MLE of the disturbance covariance matrix by the Jacobian of the reduced form model. Malinvaud has shown that the latter of these methods is the actual limiting distribution's covariance matrix, while Barnett has shown that the former is only an approximation.

In this paper, we use a Monte Carlo simulation study to determine how these two covariance matrices differ with respect to the nonlinearity of the model, the number of observations in the data set, and the residual process. We find that the covariance matrix calculated from the Hessian of the concentrated likelihood function produces Wald statistics that are distributed above those calculated with the other covariance matrix. This difference becomes insignificant as the sample size increases to one-hundred or more observations, suggesting that the asymptotics of the two covariance matrices are quickly reached.

*The author has greatly benefited from discussions with Prof. William Barnett and Prof. Robert Parks and especially thank them for the encouragement and guidance.

The two Wald statistic's empirical size and range of values were found to be unacceptably large when the data set consisted of sixty or less observations. These small sample properties reinforce those previously found by others that the standard errors for this class of models are overly optimistic and underestimate the parameter's true variance. The small sample results suggest that an alternative method of calculating the parameter estimates' covariance is needed in order for stronger statistical inference to be made with this class of models.

1 Introduction

Econometric techniques employing neoclassical theory to estimate output supply, factor demand, or consumer demand systems have often required the empirical economist to estimate a nonlinear system of equations that is additive in its error structure when written in reduced form. With the advancements in computers, economist can easily estimate such nonlinear systems with many statistical computer packages. Most of these packages provide the user with the facilities to compute a list of various parameter estimates of these nonlinear systems. Often this list includes the maximum likelihood estimator (MLE) of the model's parameters and its residual covariance matrix.[1] However, even though the calculation of the MLE is provided by the statistical package the method by which the MLE and the estimates' covariance matrix is computed often differs from software to software.

Many statistical packages use a Gauss-Newton procedure to calculate the MLE of a nonlinear system. These packages often estimate the MLE's covariance matrix with the inverse of the inverted residual covariance matrix, pre- and post-multiplied by a matrix containing the gradients of the nonlinear system of equations, i.e. by the system's Jacobian.[2] Since this method of calculating the MLE's asymptotic covariance matrix involves the gradient of each equation in the nonlinear system, we refer to it as the Gradient approach.

Other statistical packages use a form of one of the following Newton-like algorithms to calculate the MLE for a nonlinear system of equations:

- Newton-Raphson
- DFP (Davidon (1959)-Fletcher-Powell (1963))
- BHHH (Brendt, Hall, Hall, and Hausman (1974)).

Each of these algorithms use an iterative procedure where the estimate of the system's parameters are updated and improved on through a function of the log likelihood function's

[1] SAS, TSP, and SHAZAM have procedures in there PC versions that calculate the MLE of a nonlinear system.

[2] PROC MODEL in SAS with the FIT option ITSUR uses this method to calculate the MLE's covariance matrix.

Hessian matrix. The only difference between these Newton-like methods is the manner in which the Hessian is calculated.

With a Newton-like procedure, a consistent estimate of the MLE's asymptotic covariance matrix is calculated from the inverted Hessian of the likelihood function. We call this method of calculating the MLE's asymptotic covariance matrix the Hessian approach. However, because each of the above algorithms calculates the Hessian matrix differently each method may obtain a different estimate of the covariance matrix. Fortunately, Belsley (1980) has shown that there is little difference in the Hessian approach due to the different Newton-like methods.

To date, the Gradient and Hessian methods have not been compared. Such a comparison is this paper's objective with the intent of shedding some light on the properties of the two covariance matrices. By comparing these two methods under different models, sample sizes, and additive error structures we are able to draw some conclusions about the strengths and weaknesses of the statistical inference made with each covariance matrix. For instance, we find that the asymptotic properties of the two covariance matrices are reached with relatively modest data sets containing 120 or more observations, and that these covariance matrices preform poorly in describing the variability of the parameter estimates with sixty or less observations.

In this paper we construct a Monte Carlo simulation study of two systems of equations. For each model, sample size, and error structure both the Gradient and the Hessian MLE's covariance matrices are used to compute the estimated model's Wald statistic. We then plot the two Wald statistic's box-plots and compare their empirical size. In Section II we present the class of models that we are interested in and review the asymptotic properties of their MLE. The results of Section II are based on the work of Malinvaud (1970) and Barnett (1976). Section III presents the Monte Carlo simulations and their results for the MLE's covariance matrix from a absolute price version Rotterdam model and a linear expenditure system. Section IV summarizes our findings.

2 Asymptotic Properties and Results

We are concerned with the following class of models

$$\mathbf{y}_t = \mathbf{g}(\mathbf{x}_t, \gamma_0) + \epsilon_t \qquad t = 1, 2, 3, \ldots, T \tag{1}$$

where the unobservable $\epsilon_t \sim i.i.d.\ \mathcal{N}(\mathbf{0}, \Omega_0)$, and a matrix or vector subscripted by 0 denotes its true value. The $m \times 1$ vector $\mathbf{y}_t$ is comprised of the observable endogenous variables at time t, whereas $\mathbf{x}_t$ is a $p \times 1$ vector of the model's observable exogenous variables. $\mathbf{g}$ is an $m \times 1$ vector of known functions where each function is continuous in its arguments $\mathbf{x}_t$ and γ_0.

Gallant (1975) refers to this class of models as a seemingly unrelated nonlinear equation after Zellner's (1962) seminal article. In addition to the above additive error structure, this class of models also includes those nonlinear structural system of equations with a reduced form additive error structure. The above class of models also represent those systems of equations which are linear in their parameters, but have explicit analytic constraints. For example, a linear consumer demand system with the theoretical parameter restrictions of symmetry and homogeneity belongs to this class of models.

Let S_T be a positive definite $m \times m$ matrix for any positive integer T. The GLS estimator of γ in Eq. (1) is

$$\gamma(S_T) = \arg\min_{\gamma} \left\{ \sum_{t=1}^{T} [y_t - g(x_t, \gamma)]' S_T [y_t - g(x_t, \gamma)] \right\}$$

for any T and any S_T. The GLS estimate of the residual's covariance matrix is the matrix $\tilde{\Omega}_T = \frac{1}{T} \sum_{t=1}^{T} \tilde{\epsilon}_t \tilde{\epsilon}_t'$, where $\tilde{\epsilon}_t = y_t - g_t(\gamma(S_T))$. With $\tilde{\Omega}_T^{-1}$ the economist can obtain the finite-step GLS estimator, $\tilde{\gamma}$, by calculating $\gamma(\tilde{\Omega}_T^{-1})$.

Repeating the above steps until convergence, and assuming they converge uniformly in T, results in the iterated GLS estimator, IGLS, $\hat{\gamma}$.[3] Barnett (1976) has shown that as $T \to \infty$, $\hat{\gamma}$ tends in probability to its true value. Barnett also provided the important result that the IGLS estimator of the residual's covariance matrix, $\hat{\Omega}$, also tends to its true value as $T \to \infty$.

Under certain regularity conditions (see pp. 331-332 of Malinvaud), Malinvaud shows that $\tilde{\gamma}$ is consistent, is asymptotically efficient in the class of all consistent regular estimators, and has an asymptotic distribution of $\sqrt{T}(\tilde{\gamma} - \gamma_0) \to \mathcal{N}(0, [M_T(\tilde{\tilde{\Omega}}_T^{-1})]^{-1})$ as $T \to \infty$, where $\tilde{\tilde{\Omega}}_T = \frac{1}{T} \sum_{t=1} T \tilde{\tilde{\epsilon}}_t \tilde{\tilde{\epsilon}}_t'$ and $\tilde{\tilde{\epsilon}}_t = y_t - g_t(\gamma(\tilde{\Omega}_T^{-1}))$.

The inverted covariance matrix of the asymptotic distribution is defined by

$$M_T(S) = \frac{1}{T} \sum_{t=1}^{T} Z_t' S Z_t \tag{2}$$

where $Z_t = \partial g(x_t, \gamma)/\partial \gamma'$, and S is any positive definite symmetric matrix. Malinvaud shows that the finite-step GLS's $\tilde{\tilde{\Omega}}_T^{-1}$ tends in probability to Ω_0^{-1}, and that $T \cdot M_T(\Omega_0^{-1})$ is equal to Fischer's information matrix. Hence, the matrix $[T \cdot M_T(\tilde{\tilde{\Omega}}_T^{-1})]^{-1}$ is a consistent estimate of the lower bound for the covariance matrix of a consistent regular estimate of γ_0. This is the Gradient method of calculating the MLE's asymptotic covariance matrix.

Barnett (1976) extends Malinvaud's finite-step GLS asymptotic results to include the IGLS estimator and hence, the MLE for this class of models. Let $\hat{\theta}$ be a $p + m(m+1)/2 \times 1$ vector that contains the IGLS estimators $\hat{\gamma}$ and the nonredundant elements of the covariance

[3] The limit of this procedure is a critical point of the likelihood function.

matrix $\hat{\Omega}$. Barnett's Theorem 1 establishes that the consistency of $\hat{\theta}$ is sufficient for the negative Hessian of the log likelihood function to be a consistent estimator of the limiting information matrix

$$-\frac{1}{T}\cdot\left[\frac{\partial^2(\ln L(\theta|\mathbf{y},\mathbf{x})}{\partial\theta\partial\theta'}\right]\Bigg|_{\theta=\hat{\theta}}$$

where $\mathbf{y}=(\mathbf{y}_1',\ldots,\mathbf{y}_T')'$, $\mathbf{x}=(\mathbf{x}_1',\ldots,\mathbf{x}_T')'$ and L is the likelihood function. Furthermore, Barnett establishes that the inverted limiting information matrix evaluated at θ_0 is the lower bound covariance matrix of any consistent asymptotic estimator of θ. Hence, the asymptotic covariance matrix may be estimated by

$$-\left[\frac{\partial^2(\ln L(\theta|\mathbf{y},\mathbf{x})}{\partial\theta\partial\theta'}\right]^{-1}\Bigg|_{\theta=\hat{\theta}}.$$

Because the asymptotic covariance matrix of $\hat{\Omega}_T$ is rarely desired and the computations required to invert the entire Hessian is large and burdensome, Barnett provides a simplification to calculating the asymptotic covariance matrix of $\hat{\gamma}$. One of the many properties established by Malinvaud is the block diagonal matrix pattern of Eq. (1)'s information matrix. Barnett shows that the upper left diagonal block of the asymptotic covariance matrix associated with the elements of γ can be calculated by inverting the Hessian of the concentrated log likelihood function. Hence, the asymptotic covariance matrix of $\hat{\gamma}$ is equal to

$$-\left[\frac{\partial^2\ln L(\gamma,k(\gamma|\mathbf{y},\mathbf{x})|\mathbf{y},\mathbf{x})}{\partial\gamma'\partial\gamma}\right]^{-1}\Bigg|_{\gamma=\hat{\gamma}} \tag{3}$$

where $L(\gamma,k(\gamma|\mathbf{y},\mathbf{x})|\mathbf{y},\mathbf{x})$is the concentrated likelihood function and $k(\gamma|\mathbf{y},\mathbf{x})$ is an $m(m+1)/2\times 1$ vector of the nonredundant individual elements of $\hat{\Omega}_T$ that maximizes the likelihood function conditional on γ.

When a statistical package uses a Newton-like method to solve for the MLE it is $\partial^2\ln L(\gamma,k(\gamma|\mathbf{y},\mathbf{x})|\mathbf{y},\mathbf{x})/\partial\gamma\partial\gamma'$, or an approximation of it, that is used to determine which direction the estimated parameters should move for convergence. Because this matrix is asymptotically equivalent to the limiting information matrix and is computed with each iteration in a Newton-like algorithm, in the final iteration the negative inverse of this matrix is readily available as an estimate of the MLE covariance matrix. This method of calculating the MLE covariance matrix is the Hessian approach.

3 Monte Carlo Simulation

The details of our Monte Carlo study are as follows. We replicate the following systems of equations

$$y_{it} = p_{it}\alpha_i + \beta_i \left[m_t - \sum_{j=1}^{4} p_{jt}\alpha_j \right] + \epsilon_{it} \qquad i = 1,2,3,4 \tag{4}$$

$$y_{it} = \mu_i \left[Dm_t - \sum_{j=1}^{4} w_{jt}^{*} Dp_{jt} \right] + \sum_{j \in S} \pi_{ij} \left(Dp_{jt} - Dp_{kt} \right) + \epsilon_{it} \qquad i = 1,2,3,4 \tag{5}$$

where $t = 1,2,\ldots,T$ and $\epsilon_t = [\epsilon_{1t}, \epsilon_{2t}, \epsilon_{3t}, \epsilon_{4t}]' \sim i.i.d.\mathcal{N}(\mathbf{0}, \mathbf{\Omega}_0)$ is a multivariate error process, for different values of T and $\mathbf{\Omega}_0$.

Eq. (4) is the linear expenditure system derived from the Klein-Rubin utility function $U(q_{1t}, q_{2t}, q_{3t}, q_{4t}) = \sum_{i=1}^{4} \beta_i \ln(q_{it} - \alpha_i)$, where $y_{it} = p_{it}q_{it}$. In order for the utility function to remain consistent with theory, the parameters of Eq. (4) are restricted to $0 \leq \beta_i \leq 1$, and $0 \leq \alpha_i \leq q_{it}$ for $i = 1,2,3,4$ and $\forall t$. However, as a result of these restrictions Eq. (4) will be over-identified and $\mathbf{\Omega}_0$ singular.[4] This problem is overcome by eliminating any one of the four equations of Eq. (4) and using the constraint $\sum_{i=1}^{4} \beta_i = 1$ to estimate the β_i of the deleted equation.

In our Monte Carlo experiment we have chosen to eliminate the fourth equation of Eq. (4). Hence, the true parameter vector of Eq. (4) is

$$\begin{aligned} \gamma_0 &= [\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3]' \\ &= [6.5, 3.0, 1.5, 6.0, 0.366, 0.069, 0.121]'. \end{aligned}$$

The other system of equations, Eq. (5), represents the absolute price version of the Rotterdam model, where $w_{it}^{*} = (w_{it-1} + w_{it})/2$ is the average expenditure share of good i in time t, and D is the finite natural log change operator, i.e. $Dp_{it} = \ln p_{it} - \ln p_{it-1}$. In the second term of Eq. (5)'s righthand side the Dp_{jt}'s are normalized by Dp_{kt} for an arbitrarily chosen good k. These terms are summed over the set $S = \{1,2,3,4\} - \{k\}$. In our simulation study we have chosen $k = 4$, hence, $S = \{1,2,3\}$.

In order for the absolute price version of the Rotterdam model to agree with economic demand theory its parameters are restricted so that the model is homogeneous and its Slutsky matrix symmetrical and negative semidefinite. This is accomplished by letting $\sum_{i=1}^{4} \mu_i = 1$, $\pi_{ij} = \pi_{ji}$ for $i,j = 1,2,3,4$ and $\sum_{i=1}^{4} \pi_{ij} = 0$ for $j = 1,2,3,4$. As with the restricted linear expenditure system, these parameter constraints cause the Rotterdam model to be over- identified, and requires one of the system's equations to be removed. To be consistent with the linear expenditure system we remove the fourth equation from Eq. (5).

[4] See Barten (1969).

With these parameter restrictions and eliminations Eq. (5)'s true parameter vector is

$$\begin{aligned}\gamma_0 &= [\mu_1, \mu_2, \mu_3, \pi_{11}, \pi_{12}, \pi_{13}, \pi_{22}, \pi_{23}, \pi_{33}]' \\ &= [0.292, 0.056, 0.098, -0.039, 0.01, 0.008, -0.028, -0.003, -0.023]'\end{aligned}$$

for our simulation study.[5]

To calculate the y_{it}'s of Eq. (4) and (5) we use Citibase consumer price index (1982-84 base) and consumer expenditure (1982 dollars) data. The data set contains 240 monthly observations from January, 1967 to January, 1987 of service goods, perishable goods, semi-durable goods, and durable goods. We respectively set these prices and expenditure levels equal to p_{it} and q_{it} for $i = 1, 2, 3, 4$. The exogenous variable m_t are set equal to the corresponding Citibase monthly personal income data (1982 dollars).

Eq. (4) and (5) are simulated 500 times for $T = 30, 60, 120, 240$ and the two error processes with covariance matrices

$$\Omega_1 = \begin{bmatrix} 1.00 & 0.75 & 0.66 \\ 0.75 & 1.00 & 0.25 \\ 0.66 & 0.25 & 1.00 \end{bmatrix}$$

$$\Omega_2 = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{bmatrix}.$$

We then estimated Eq. (4) and (5)'s MLEs for the 500 samples and used these estimates to calculate the Wald statistics $[\hat{\gamma} - \gamma_0]' B^{-1} [\hat{\gamma} - \gamma_0]$ and $[\hat{\gamma} - \gamma_0]' M^{-1} [\hat{\gamma} - \gamma_0]$ where

$$B = - \left[\partial^2 \ln L(\gamma, k(\gamma|\mathbf{y}, \mathbf{x})|\mathbf{y}, \mathbf{x}) / \partial\gamma' \partial\gamma \right]^{-1} \Big|_{\gamma=\hat{\gamma}}$$

is the covariance matrix calculated with the Hessian approach, and $M = [T \cdot M_T(\hat{\Omega}^{-1})]^{-1}$ is the covariance matrix from the Gradient approach. Asymptotically, the Wald statistics should be distributed $\chi^2_{(7)}$ for Eq. (4) and $\chi^2_{(9)}$ for Eq. (5).

In Figure 1 through Figure 4 the Wald statistics from the Monte Carlo experiments are arranged as box-plots. The box-plots show the Gradient approach being distributed lower than the Hessian, regardless of the model, sample size, or residual's covariance matrix. The difference, however, is negligible when $T \geq 120$.

In Section II the negative Hessian of the concentrated likelihood function was shown to be a consistent estimator of the limiting information matrix, while the matrix $[T \cdot M_T(\Omega_0^{-1})]$ equaled the limiting information matrix. This, coupled with the similarity of the Gradients

[5]Simulations with different parameter values for both Eq. (4) and (5) did not obtain different results.

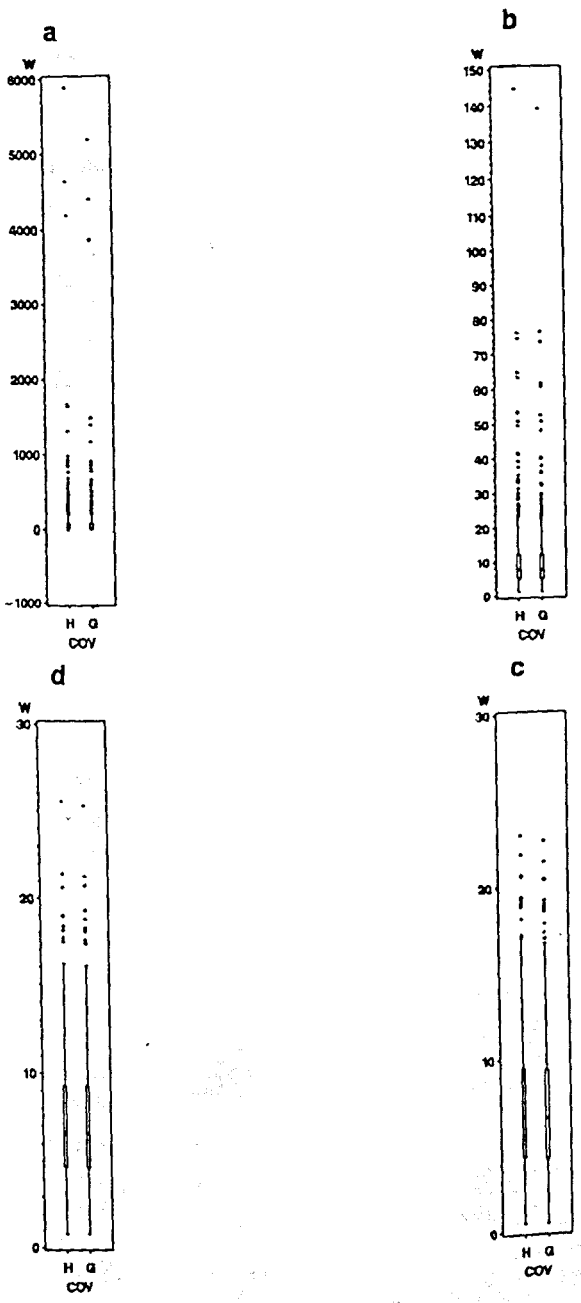

Figure 1: Box-Plots of Wald Stat. for Eq. (4), Ω_1, a) $T = 30$, b) $T = 60$, c) $T = 120$, and d) $T = 240$.

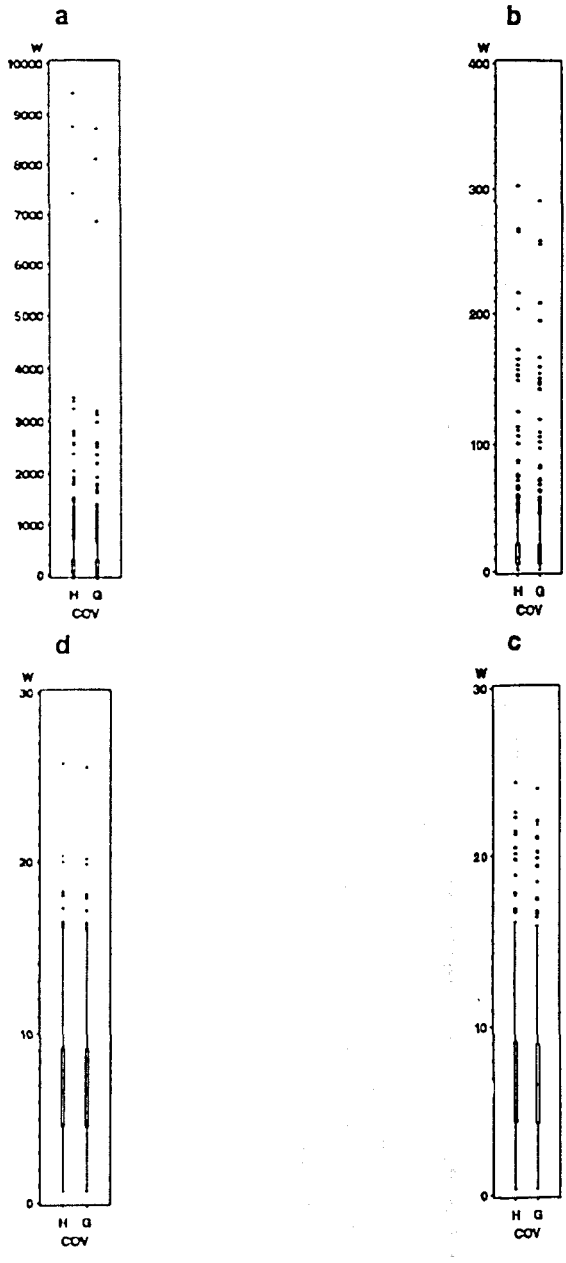

Figure 2: Box-Plots of Wald Stat. for Eq. (4), Ω_2, a) $T = 30$, b) $T = 60$, c) $T = 120$, and d) $T = 240$.

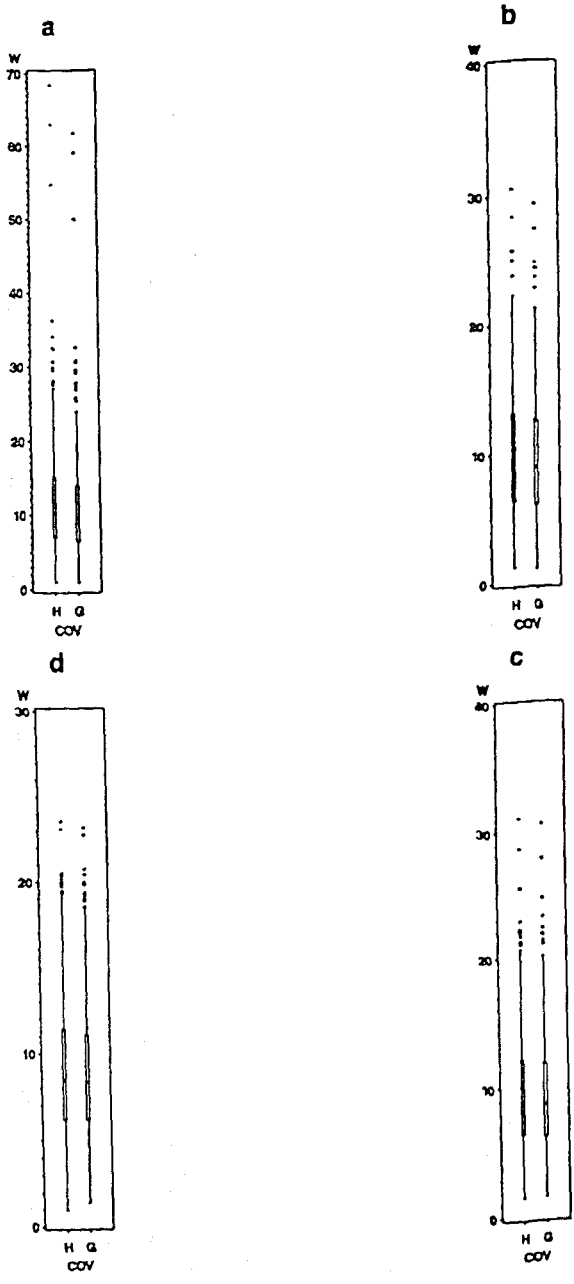

Figure 3: Box-Plots of Wald Stat. for Eq. (5), Ω_1, a) $T = 30$, b) $T = 60$, c) $T = 120$, and d) $T = 240$.

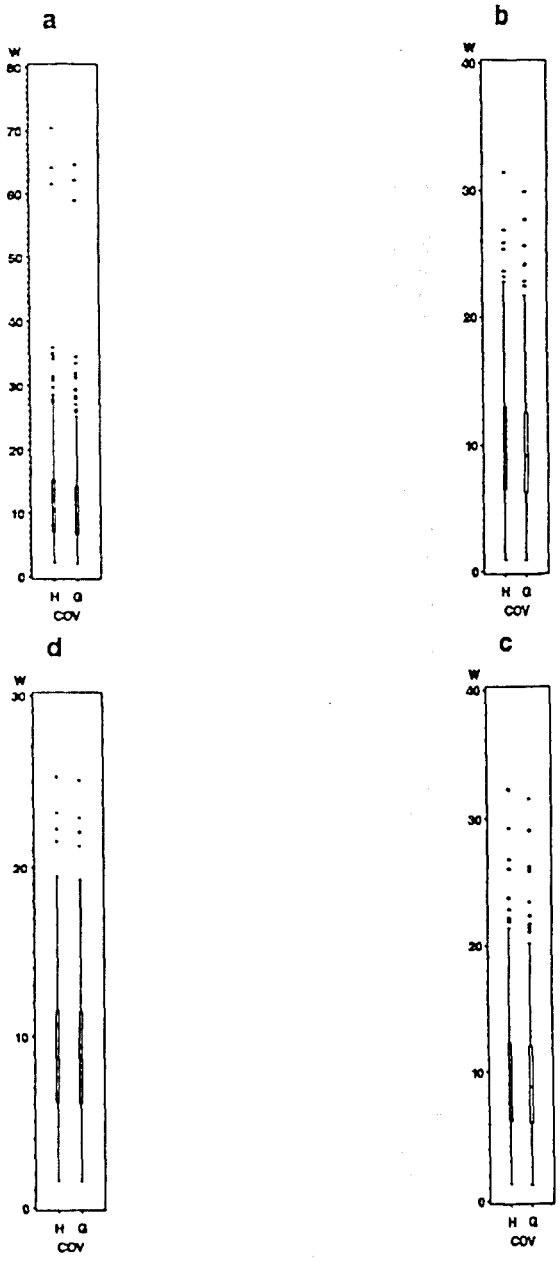

Figure 4: Box-Plots of Wald Stat. for Eq. (5), Ω_2, a) $T = 30$, b) $T = 60$, c) $T = 120$, and d) $T = 240$.

Table 1: Size 5%-Level of the Wald Statistics

	Ω_1		Ω_2	
T	α_M	α_B	α_M	α_B
Klein 30	0.634	0.622	0.764	0.778
60	0.174	0.178	0.398	0.404
120	0.054	0.058	0.064	0.068
240	0.048	0.052	0.064	0.066
Rotter 30	0.130	0.180	0.132	0.188
60	0.074	0.088	0.076	0.088
120	0.064	0.064	0.070	0.072
240	0.056	0.056	0.040	0.046

The standard error of the empirical size is 9.75E-3.

and Hessian's box-plots, suggests that the asymptotic properties of the two covariance matrices are available at relatively small sample sizes, i.e. $T \geq 120$.

The empirical size of the Wald statistics in Table 1 also lends credence to this conclusion. In both of the cases, $T = 120, 240$, the Gradient's size is never significantly different from the Hessians. When $T = 120$ the size of the two Wald statistics are insignificantly larger than the nominal 5% level, and when $T = 240$ three of the eight cases have empirical sizes that are insignificantly smaller than 5%. Hence, the empirical size and box-plots for $T \geq 120$ suggest that an econometrician need not worry whether the Hessian or Gradient method is being used to calculate the MLE's covariance matrix when her data set contains approximately one-hundred and twenty observations.

A number of the empirical tests for homogeneity of demand and Slutsky symmetry have been conducted on data sets with $T \leq 60$, and in most of these studies the null hypothesis of demand theory have been rejected.[6] The results from our Monte Carlo experiment provide one reason for the poor statistical inference with systems of demand equations: inadequate sample size. Our finding suggest that a data set with $T \geq 120$ would enable stronger statistical inference to be made about the neoclassical hypothesis of demand theory.

Because a smaller standard error, on average, causes the Wald statistic to increase, the a ($T = 30$) and b ($T = 60$) box-plots in Figures 1 through 4 support Rosalsky et. al. (1984) and Theil and Rosalsky's (1985) argument that the small sample properties of the MLE's asymptotic standard errors are overly optimistic and under estimate the true variability of the parameter estimate. In the extreme case of Figure 1a the Hessian's Wald statistic has a minimum of -7.71 and a maximum of 5893.83, as opposed to 0.081 and 25.53 in Figure 1d. The small sample box-plots also allows Gallant's (1975) findings that the finite-step GLS

[6]See Theil and Clements (1987), Chapter 3 for a survey of these empirical demand studies.

asymptotic standard errors underestimate the actual asymptotic standard errors to include the Hessians and Gradient's ML standard errors.[7]

An important observation concerning the relationship between the Hessian and Gradient approaches can be shown in the mathematical expression of the log likelihood's Hessian matrix. Malinvaud (1970) derived the property

$$\begin{aligned} -d^2 \ln \mathcal{L}(\theta|\mathrm{y},\mathrm{x}) &= \frac{T}{2} tr[(\Omega^{-1}d\Omega)^2(2\Omega^{-1}\mathrm{N}(\gamma) - \mathrm{I})] \\ &\quad -Ttr[\Omega^{-1}d\Omega\Omega^{-1}d\mathrm{N}] + \frac{T}{2} tr(\Omega^{-1}d^2\mathrm{N}) \end{aligned} \tag{6}$$

where $d^i\mathrm{N}$ is the ith differential of the function $\mathrm{N}(\gamma) = (1/T)\sum_{t=1}^{T}[\mathrm{y}_t - \mathrm{g}_t(\gamma)][\mathrm{y}_t - \mathrm{g}_t(\gamma)]'$. Our concern is with the γ's MLE covariance matrix, hence we set $d\Omega = 0$, eliminating the first and second terms of Eq. (6). Malinvaud has shown that the remaining term is equal to [8]

$$\begin{aligned} \frac{T}{2} tr(\Omega^{-1}d^2\mathrm{N}) &= \frac{T}{2} tr \left\{ \Omega^{-1} \left[\frac{1}{T} \sum_{t=1}^{T} \sum_{j,k=1}^{p} \left(2 \frac{\partial \mathrm{g}_t(\gamma)}{\partial \gamma_j} \frac{\partial \mathrm{g}_t(\gamma)'}{\partial \gamma_k} \right. \right. \right. \\ &\quad \left. \left. \left. - \epsilon_t \frac{\partial^2 \mathrm{g}_t'}{\partial \gamma_k \partial \gamma_j} - \frac{\partial^2 \mathrm{g}_t}{\partial \gamma_k \partial \gamma_j} \epsilon_t' \right) d\gamma_k d\gamma_j \right] \right\}. \end{aligned} \tag{7}$$

Algebraically manipulating Eq. (7) we rewrite it as

$$\begin{aligned} \frac{T}{2} tr(\Omega^{-1}d^2\mathrm{N}) &= \frac{T}{2} \left\{ \sum_{k,j=1}^{p} 2 \left[\frac{1}{T} \sum_{t=1}^{T} \left(\frac{\partial \mathrm{g}_t(\gamma)'}{\partial \gamma_k} \Omega^{-1} \frac{\partial \mathrm{g}_t(\gamma)}{\partial \gamma_j} \right) d\gamma_k d\gamma_j \right] \right. \\ &\quad - \sum_{k,j=1}^{p} \left[\frac{1}{T} \sum_{t=1}^{T} \left(\frac{\partial^2 \mathrm{g}_t'}{\partial \gamma_j \partial \gamma_k} \Omega^{-1} \epsilon_t \right) d\gamma_k d\gamma_j \right] \\ &\quad \left. - \sum_{k,j=1}^{p} \left[\frac{1}{T} \sum_{t=1}^{T} \left(\epsilon_t' \Omega^{-1} \frac{\partial^2 \mathrm{g}_t}{\partial \gamma_k \partial \gamma_j} \right) d\gamma_k d\gamma_j \right] \right\}. \end{aligned} \tag{8}$$

From the definition of $M_T(\mathrm{S})$ in Eq. (2), the first righthand side term of Eq. (8) is equal to the k,jth element of the matrix $M_T(\Omega^{-1})$. Hence,

$$\begin{aligned} \frac{T}{2} tr(\Omega^{-1}d^2\mathrm{N}) &= T \left\{ \sum_{k,j=1}^{p} \left[M_T(\Omega^{-1}) \right]_{k,j} d\gamma_k d\gamma_j \right. \\ &\quad \left. - \sum_{k,j=1}^{p} \frac{1}{T} \sum_{t=1}^{T} \left[\sum_{h=1}^{m} \sum_{i=1}^{m} \epsilon_{ht} \omega_{hi} \frac{\partial^2 g_{it}}{\partial \gamma_k \partial \gamma_j} \right] d\gamma_k d\gamma_j \right\} \end{aligned} \tag{9}$$

[7] Gallant (1975) noted that this reduction in the standard error only occurs when the functions of g differ or when cross-equation restrictions are imposed. In our Monte Carlo study, the functions of Eq. (4) and (5) do not differ, but cross-equations restrictions have been imposed through the symmetry and homogeneity conditions.

[8] Eq. (7) corrects for two of Malinvaud's original sign errors.

where ω_{hi} is the h,ith element of $\mathbf{\Omega}^{-1}$. Thus, the Hessian of the log likelihood function is equal to the inverse of the Gradient approaches' covariance matrix minus an additional term.

To derive the consistency of the log likelihood function as an estimate of the limiting information matrix, Barnett (1976) showed that for a consistent estimate of γ and $\mathbf{\Omega}$ the additional term in Eq. (9) tended in probability to zero as $T \rightarrow \infty$. This result equates the Gradient and Hessian's asymptotic covariance matrices for sufficiently large T, which our simulations have shown to be close to $T = 120$. If T is small the Hessian approach will differ from the Gradient approach by the magnitude of the additional term. Fortunately, even for small T the biasness caused by the additional term will be zero if the model is linear in its parameters, i.e. $\partial^2 \mathbf{g}_t / \partial \gamma_k \partial \gamma_j = 0$ for $k, j = 1, 2, \ldots, p$.

Since Eq. (5) is linear in its parameters the difference between the Hessian and Gradient approach should be insignificant. Most of the box-plots in Figure 3 and 4 show this to be the case, but in Figure 3a the Hessian's Wald statistics are significantly higher than the Gradients. This pattern is also found for the empirical sizes listed in Table 1 when $T = 30$. However, from previous studies we believe that this behavior is attributed to the small sample shortcomings of the $\mathbf{\Omega}$'s MLE [Leitinen (1978) and Meisner (1979)].

Because Eq. (4) is a nonlinear model the difference between the Hessian and Gradient covariance matrices is dependant on T, the estimator of $\mathbf{\Omega}$, and the severity of model's nonlinearity. The box-plots in Figure 1a and Figure 2a shows the Hessian's Wald statistics to be distributed significantly greater than the Gradients. Furthermore, Eq. (4)'s empirical size in Table 1 points out that in most of the small sample cases the Hessian methods Wald statistics are greater than the Gradients.

Although the Gradient approach seems to produce smaller Wald statistics, regardless of the model or the residual's covariance matrix, the magnitude of its size when $T \leq 60$ suggests that an alternative method is needed to calculate the MLE's covariance matrix when T is small. Perhaps bootstrapping or jackknifing would allow stronger statistical inference to be made with nonlinear systems. How much these alternative methods would improve on those approaches discussed in this paper is not known, but would be interesting research.

4 Conclusion

In this paper we have constructed a Monte Carlo experiment of two systems of equations with additive error terms and subjected the true models to two types of disturbances for $T = 30, 60, 120, 240$. After replicating the models 500 times we calculated the MLE's asymptotic covariance matrix with the Hessian and Gradient method and their respective Wald statistics.

From these simulations we found that for either model, disturbance, or sample size the Gradient approach produced Wald statistics that were distributed below that of the Hessian. This difference became insignificant for the two models and error structures as the sample size increased to $T = 120$, and suggest that for relatively small data sets the econometrician does not have to worry about which method is used to calculate the MLE covariance matrix.

The simulation results when $T \leq 60$ showed that both types of covariance matrices produced standard errors that were overly optimistic and under estimated the variance of the parameter estimates. The two covariances poor small sample properties also provided a strong reason for why a large number of empirical demand studies reject neoclassical theory. The data sets have been to small. It was also the small sample experiments with the linear system that displayed a significant difference between the Hessian and Gradient covariance even though in theory they should be the same.

Overall the small sample properties of the two approaches suggested that strong statistical inference could not be made with either method when $T \leq 60$. Thus, an alternative method to the Hessian and Gradient approach is needed when calculating the MLE's covariance matrix with a small data set. Two possibilities are the bootstrapping and jackknifing methods, who's statistical properties have yet to be tested.

REFERENCES

[1] Barnett, William A. (1976), "Maximum Likelihood and Iterated Aitken Estimation of Nonlinear System of Equations", *Journal of the American Statistical Association*, 71, 354-360.

[2] Barten, A. P. (1969), "Maximum Likelihood Estimation of a Complete Demand System," *European Economic Review*, 1, 7-73.

[3] Belsley, David A. (1980), "On the Efficient Computation of the Nonlinear Full-Information Maximum-Likelihood Estimator," *Journal of Econometrics*, 14, 203-225.

[4] Berndt E. K., B. H. Hall, R. E. Hall and J. A. Hausman (1974), "Estimation and Inference in Nonlinear Structural Models," *Annals of Economics and Social Measurement*, Oct., 653-665.

[5] Davidon, W. C. (1959), "Variable Metric Method for Minimizations," *A. E. C. Research and Development Report*, ANL5990, revised.

[6] Fletcher, R. and M. J. D. Powell (1963), "A Rapidly Convergent Descent Method for Minimization," *Computer Journal*, 6, 163-168.

[7] Gallant, A. Ronald (1975), "Seemingly Unrelated Nonlinear Regressions," *Journal of Econometrics*, 3, 35-50.

[8] Laitinen, K. (1978), "Why is Demand Homogeneity So Often Rejected?", *Economics Letters*, 1, 187-191.

[9] Malinvaud, E. (1970), *Statistical Methods of Econometrics*, (North-Holland, Amsterdam).

[10] Meisner, J. F. (1979), "The Sad Fate of the Asymptotic Slutsky Symmetry Test for Large Systems," *Economics Letters*, 2, 231-233.

[11] Rosalsky, Mercede C., Renate Finke, and Henri Theil, (1984) "The Downward Bias of Asymptotic Standard Errors of Maximum Likelihood Estimates of Non- Linear Systems," *Economics Letters*, 14, 207-211.

[12] Theil, Henri and Kenneth W. Clements (1987), *Applied Demand Analysis: Results from System-Wide Approaches*, (Ballinger: Cambridge, Mass.).

[13] Theil, Henri and Mercede C. Rosalsky (1985), "Least Squares and Maximum Likelihood Estimation of Non-Linear Systems," *Economics Letters*, 17, 119-122.

[14] Zellner, Arnold (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Test for Aggregation Bias," *Journal of the American Statistical Association*, 57, 348-368.