NBER TECHNICAL PAPER SERIES

ASYMPTOTIC PROPERTIES OF QUASI-MAXIMUM LIKELIHOOD ESTIMATORS AND TEST STATISTICS

Thomas E. MaCurdy

Technical Paper No. 14

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge MA 02138

June 1981

Portions of this paper were presented at the Econometric Society Meetings held in Montreal, Canada, June 1979. I am grateful to Tom Mroz for comments. The research reported here is part of the NBER's research program in Labor Studies. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

NBER Technical Working Paper #14
June, 1981

Asymptotic Properties of Quasi-Maximum Likelihood Estimators and Test Statistics

Abstract

This paper examines the consequences of using maximum likelihood estimation techniques based on the assumption of joint normality when the error distribution does not necessarily belong to the family of normal distributions. A nonlinear seemingly unrelated regression model with covariance restrictions provides the basic statistical framework considered in this analysis. In addition to discussing methods where one simultaneously estimates all parameters, this study examines computationally efficient procedures designed to estimate only regression coefficients or only parameters of the covariance matrix. All estimation methods are shown to generate estimators that are consistent and normally distributed in large samples even in the absence of normality. The following analysis also derives the general asymptotic properties of statistics typically used to test composite hypotheses in a large sample setting, including the Wald, the likelihood ratio, and the Lagrange multiplier test statistics. Without the assumption of normality the likelihood ratio and Lagrange multiplier statistics still converge to the usual chi-squared distribution when used to test restrictions on regression coefficients, but diverge from this distribution when used to test any sort of covariance restrictions.

Professor Thomas E. MaCurdy
Department of Economics
Stanford University
Stanford, Ca. 94305
(415) 497-3983

Introduction

When analyzing a multivariate regression model with nonlinearities in variables and parameters, a researcher typically assumes that disturbances follow a normal distribution and applies the theory of maximum likelihood to estimate parameters and to test hypotheses. The purpose of this paper is to examine the asymptotic properties of the estimators and the test statistics derived from this estimation method when disturbances are not in fact normally distributed.

The following discussion shows that these "quasi-maximum likelihood" estimators, including those for parameters of the covariance matrix, are consistent and are also asymptotically normally distributed. Falsely assuming that error vectors follow a multivariate normal distribution leads to the incorrect computation of standard errors for some parameter estimates, but it does not invalidate many of the large sample properties usually associated with maximum likelihood estimators.

In addition to examining the asymptotic properties of estimators derived from full-information methods where one simultaneously estimates regression coefficients and covariance parameters, this study also explores the properties of estimators derived from "conditional" or limited-information quasi-maximum likelihood procedures designed to estimate only subsets of parameters. One such procedure substitutes consistent estimates for the regression coefficients and estimates only parameters of the covariance matrix. Estimates based on this procedure are shown to have the same asymptotic distribution as the estimator derived from full-information methods. Thus, when

estimating parameters of the covariance matrix, it is possible to treat residuals as if they are the true values of the disturbances. Similar procedures are examined for the estimation of regression coefficients.

An important by-product of these findings concerns the use of quasi-maximum likelihood techniques as a general method for estimating the parameters of a system of nonlinear seemingly unrelated regression equations with covariance restrictions. While there are well known procedures for estimating the regression coefficients of such a model which do not rely on specific distributional assumptions for large sample properties, the estimation of covariance parameters is another matter. The theoretical results developed below offer robust procedures for estimating both regression coefficients and covariance parameters accounting for all nonlinear restrictions, including those relating regression coefficients and elements of the covariance matrix. The limited-information procedures offer a computationally efficient method of estimation.

One of these procedures is especially attractive since it provides for the estimation of regression coefficients in the presence of arbitrary and unknown forms of heteroscedasticity. Application of this method of estimation does not require any assumptions concerning the form or the absence of heteroscedasticity. The large sample properties of the estimators and the standard errors computed by this procedure remain valid no matter what the actual form of the heteroscedasticity.

When applying the theory of maximum likelihood to test hypotheses, a researcher is led to one of three test statistics: the Wald test, the likelihood ratio test, and the Rao or the Lagrange multiplier test. This

study examines the asymptotic properties of these tests when disturbances do not follow a multivariate normal distribution. In many instances, the large sample distribution associated with these test statistics is unaffected when the normality assumption is violated.

Section I outlines the basic statistical model considered in this paper. Section II develops the asymptotic properties of quasi-maximum likelihood estimators including those derived from both full- and limited-information procedures. Section III explores the large sample properties of various test statistics.

I. A Nonlinear Multivariate Regression Model with Heteroscedasticity

The basic model considered in this study consists of the following system of T equations

$$(1)\qquad \begin{pmatrix} Y_i(1) \\ \cdot \\ \cdot \\ \cdot \\ Y_i(T) \end{pmatrix} = \begin{pmatrix} f(1, X_i(1), \gamma_1) \\ \cdot \\ \cdot \\ \cdot \\ f(T, X_i(T), \gamma_T) \end{pmatrix} + \begin{pmatrix} \varepsilon_i(1) \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_i(T) \end{pmatrix} \qquad i = 1,\ldots,N,$$

where $Y_i(j)$, $j = 1,\ldots,T$, is the dependent variable associated with equation j and observation i, the $X_i(j)$'s are vectors of exogenous variables, the γ_j's are parameter vectors, $f(j, X_i(j), \gamma_j)$ is the expectation of $Y_i(j)$ conditional on $X_i(j)$, and the $\varepsilon_i(j)$'s are disturbances distributed independently across the observations, $i = 1,\ldots,N$, but correlated across equations for a given observation. In vector notation we may write (1) as

$$Y_i = f(X_i, \gamma) + \varepsilon_i,$$

or simply as

$$(2)\qquad Y_i = f_i + \varepsilon_i \qquad i = 1,\ldots,N,$$

where X_i and γ are vectors including all the unique elements of the $X_i(j)$'s and the γ_j's, respectively, and Y_i, $f(X_i, \gamma)$, f_i, and ε_i are T x 1 vectors defined by $Y_i' = (Y_i(1),\ldots,Y_i(T))$, $f_i' \equiv f'(X_i, \gamma) = (f(1, X_i(1), \gamma_1),\ldots, f(T, X_i(t), \gamma_T))$, and $\varepsilon_i' = (\varepsilon_i(1),\ldots,\varepsilon_i(T))$. It is assumed that

$$(3) \qquad E(\varepsilon_i \varepsilon_j') = \begin{cases} \Omega_i \equiv \Omega(X_i, \omega) & i = j, \\ 0 & \text{otherwise,} \end{cases}$$

where ω is a vector of parameters, and $\Omega(\cdot,\cdot)$ is a matrix of functions of X_i and ω forming a T x T positive definite symmetric matrix for each observation i. In the general case, the parameter vectors γ and ω may contain common elements, or there may exist nonlinear restrictions relating the components of γ and ω.

Model (2) combined with (3) constitutes a system of nonlinear seemingly unrelated regression equations with heteroscedasticity and covariance restrictions. This system of equations provides a rich statistical framework, and it accommodates many models familiar in econometrics including models with random coefficients and simultaneous equation models which are linear in endogenous variables. The dependent variables $Y_i(j)$ may be either continuous or discrete in nature. Hence, it is possible to analyze models of discrete choice or quantal response within the framework of model (2), including some models that relate discrete and continuous endogenous variables.

II. Estimation Methods

The technique usually applied to estimate models like (2) where one is also interested in estimating parameters of a covariance matrix is the method of maximum likelihood. Typically, a researcher assumes that disturbances are normally distributed and computes estimates by maximizing the kernel of a multivariate normal density function. Such a procedure provides for the simultaneous estimation of regression coefficients and covariance parameters; it permits the imposition of almost any sort of restrictions including constraints relating elements of the covariance matrix and regression coefficients; and it provides an estimation method when heteroscedasticity is present.

Under the assumption of joint normality, the maximum likelihood estimates of the regression coefficients and the covariance parameters of model (2) are defined as those values of γ and ω that maximize the function

$$(4) \qquad Q_N(\gamma, \omega) = \frac{1}{N} \sum_{i=1}^{N} q_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} [- \ell n|\Omega_i| - (Y_i - f_i)'\Omega_i^{-1} (Y_i - f_i)]$$

where q_i is the function of γ, ω, Y_i, and X_i defined by the second expression for Q_N.[1] The function Q_N has as its arguments the p elements of the parameter vector $\theta' = (\gamma', \omega')$ which is assumed to lie in the interior of a convex parameter space Θ. In the following analysis, it

[1] The reader will immediately recognize that Q_N is proportional to the kernel of a multivariate density function.

is important to distinguish between the vector of parameters θ which can take any value in the set Θ, and the true value of θ, denoted below as θ_0, which actually generates the data. Formally, in the specification of the system of equations assumed to generate Y_i given by (1) and (2) and in the specification of the covariance matrix associated with the disturbance vector ε_i given by (3), the vector θ_0, (i.e., γ_0 and ω_0) should appear rather than θ.[1] To simplify the following exposition, it is initially assumed that γ and ω contain no common elements, and there are no constraints relating the components of these vectors. Furthermore, it is convenient to introduce the following notation: denote the gradient vector associated with Q_N evaluated at $\gamma = \gamma^*$ and $\omega = \omega^*$ and partitioned subvectors of this gradiant vector as

$$\ell(\theta^*) \equiv \frac{\partial Q_N}{\partial \theta}\Big|_{\theta^*} \equiv \begin{pmatrix} \ell_\gamma(\gamma^*, \omega^*) \\ \\ \ell_\omega(\gamma^*, \omega^*) \end{pmatrix} \equiv \begin{pmatrix} \dfrac{\partial Q_N}{\partial \gamma}\Big|_{\theta^*} \\ \\ \dfrac{\partial Q_N}{\partial \omega}\Big|_{\theta^*} \end{pmatrix};$$

denote minus the matrix of second partials and partitioned submatrices as

$$H(\theta^*) \equiv -\frac{\partial^2 Q_N}{\partial\theta\partial\theta'}\Big|_{\theta^*} \equiv \begin{bmatrix} H_{\gamma\gamma}(\gamma^*, \omega^*) & H_{\gamma\omega}(\gamma^*, \omega^*) \\ \\ H_{\omega\gamma}(\gamma^*, \omega^*) & H_{\omega\omega}(\gamma^*, \omega^*) \end{bmatrix} \equiv \begin{bmatrix} -\dfrac{\partial^2 Q_N}{\partial\gamma\partial\gamma'}\Big|_{\theta^*} & -\dfrac{\partial^2 Q_N}{\partial\gamma\partial\omega'}\Big|_{\theta^*} \\ \\ -\dfrac{\partial^2 Q_N}{\partial\omega\partial\gamma'}\Big|_{\theta^*} & -\dfrac{\partial^2 Q_N}{\partial\omega\partial\omega'}\Big|_{\theta^*} \end{bmatrix};$$

---

[1] Only at $\gamma = \gamma_0$ and $\omega = \omega_0$ will the vector ε_i have zero mean. Unfortunately, the notation θ is sometimes used to represent an unknown parameter vector and, at other times, it is used to denote the true values of these parameters. To avoid confusion, θ and θ_0 will be sharply distinguished in the remaining analysis.

and denote the matrix of outer partials as

$$G(\theta^*) \equiv \frac{1}{N}\sum_{i=1}^{N} \frac{\partial q_i}{\partial \theta}\Big|_{\theta^*} \frac{\partial q_i}{\partial \theta'}\Big|_{\theta^*}$$

$$\equiv \begin{bmatrix} G_{\gamma\gamma}(\gamma^*, \omega^*) & G_{\gamma\omega}(\gamma^*, \omega^*) \\ G_{\omega\gamma}(\gamma^*, \omega^*) & G_{\omega\omega}(\gamma^*, \omega^*) \end{bmatrix} \equiv \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} \frac{\partial q_i}{\partial \gamma}\Big|_{\theta^*} \frac{\partial q_i}{\partial \gamma'}\Big|_{\theta^*} & \frac{1}{N}\sum_{i=1}^{N} \frac{\partial q_i}{\partial \gamma}\Big|_{\theta^*} \frac{\partial q_i}{\partial \omega'}\Big|_{\theta^*} \\ \frac{1}{N}\sum_{i=1}^{N} \frac{\partial q_i}{\partial \omega}\Big|_{\theta^*} \frac{\partial q_i}{\partial \gamma'}\Big|_{\theta^*} & \frac{1}{N}\sum_{i=1}^{N} \frac{\partial q_i}{\partial \omega}\Big|_{\theta^*} \frac{\partial q_i}{\partial \omega'}\Big|_{\theta^*} \end{bmatrix}.$$

The function Q_N must satisfy a set of regularity conditions in order to prove that the estimate for θ obtained by maximizing Q_N has the asymptotic properties one usually associated with maximum likelihood estimators. The analysis below assumes the following two conditions:

(I) The gradient vector for each observation $\frac{\partial q_i}{\partial \theta}$ exists and is uniformly continuous in θ,[1] and any linear combination of these vectors denoted as $z_i = \lambda' \frac{\partial q_i}{\partial \theta}$, where λ is any conformable vector of real constants, satisfies

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} E\left|z_i - E(z_i)\right|^{2+\delta} < \infty$$

for some $\delta > 0$ and all $\theta \in \Theta$; and

(II) The matrix of second partials $\frac{\partial^2 q_i}{\partial\theta\partial\theta'}$ exists and is uniformly continuous in θ with each element of this matrix, denoted as h_i, satisfying

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} E|h_i| < \infty \quad \text{for all } \theta \in \Theta,$$

[1] A function $f(a)$ is said to be uniformly continuous on the set A if, for every $\phi > 0$, there exists a vector $\bar{\alpha} > 0$ such that $|f(a+\alpha) - f(a)| \leq \phi$ for any $|\alpha| < \bar{\alpha}$, and for any $a \in A$.

and the average of these matrices $\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 q_i}{\partial\theta\partial\theta} \equiv -H(\theta)$ satisfies

$$\lim_{N\to\infty} E(H) = \bar{H}(\theta) \qquad \text{for all } \theta \in \Theta,$$

where $\bar{H}(\theta)$ is a nonsingular matrix whose elements are strictly less than infinity in absolute value.

It is important to emphasize that the above conditions restrict the form of the parameter space Θ and, in general, limit Θ to be a closed ball in R^p where p is the dimension of θ.[1] These two conditions guarantee that one can apply the asymptotic theorems cited below. While these conditions are sufficient to prove most of the following results, they are not necessary.[2] They are, however, satisfied for most problems. Their main attraction over weaker conditions concerns the fact that they are typically easier to verify in most applications.

---

[1]The assumption that moments exist for first and second partials for all $\theta \in \Theta$ restricts Θ to be strictly bounded from above, at least in some dimensions, and requires other dimensions to be bounded away from zero. A more subtle, but possibly more important limitation on the form on the set Θ is imposed by the assumption that the matrix of second partials $-\bar{H}(\theta)$ is nonsingular for all $\theta \in \Theta$. This assumption rules out the possibility that the function $Q_N(\theta)$ (formally its limit) possesses more than one maximum on the convex set Θ where $\frac{\partial Q_N}{\partial\theta} = 0$. If there were two points, θ_1 and θ_2, satisfying this property, then, according to Rolle's theorem, we know there exists a $\theta_3 \in \Theta$ such $Q_N(\theta)$ achieves a minimum. If this were true, then $\bar{H}(\theta)$ must switch from being positive to negative definite which ensures that it is singular from some $\theta \in \Theta$. Thus, the regularity conditions require Θ to be defined so that $Q_N(\theta)$ achieves at most one maximum in the interior of Θ.

[2]To prove consistency of estimators, for example, one does not require the existence of any moments or partial derivatives. Using theorems due to Amemiya (1973) and Jenrich (1969), consistency of an estimator $\hat{\theta}$ that maximizes $Q_N(\theta)$ follows if $Q_N(\theta)$ converges in probability to a function $\bar{Q}(\theta)$ uniformly in θ, and $\bar{Q}(\theta)$ attains a unique maximum in the interior of Θ. This paper follows a more traditional method of analysis of the sort found in Rao (1973, Ch. 6) to prove both consistency and asymptotic normality.

Combining the above conditions with the assumption that the disturbance vectors ε_i are jointly normally distributed, two well known results can be shown to follow. Letting $\hat{\theta}$ denote the maximum likelihood estimator for θ dervied by maximizing Q_N, the first result states that $\hat{\theta}$ is consistent for θ_0; or

$$\text{plim}\{\hat{\theta}\} = \theta_0.$$

The second result asserts that the random vector $\sqrt{N}(\hat{\theta} - \theta_0)$ possesses an asymptotic normal distribution; in particular,

$$\text{(5)} \qquad \text{dlim}\{\sqrt{N}(\hat{\theta} - \theta_0)\} = N(0, \bar{H}^{-1}(\theta_0)),$$

where the notation dlim denotes convergence in distribution, $\bar{H}(\theta_0) \equiv \text{plim}(H(\theta_0))$ is the probability limit of the matrix of second partials known as the information matrix, and $N(\mu, V)$ signifies distribution according to a multivariate normal probability law with mean vector μ and covariance matrix V. On the basis of result (5), one concludes that in large samples

$$\text{(6)} \qquad \hat{\theta} \overset{\cdot}{\sim} \left(\theta_0, \frac{1}{N} H^{-1}(\hat{\theta})\right);$$

or, $\hat{\theta}$ is approximately normally distributed with a mean vector equal to the true values of the parameters and a covariance matrix equal to minus the inverse of the matrix of second partials evaluated at the maximum likelihood estimates divided by the sample size. All maximum likelihood computer packages report standard errors and test statistics based on (6), and it is these quantities that most researchers use in their empirical analyses.

The following discussion examines the asymptotic properties of $\hat{\theta}$ without the assumption that the disturbance vectors are distributed according to a multivariate normal distribution. Instead, it only assumes that the ε_i's are independently distributed across the observations i = 1,...,N, and the two regularity conditions specified above. These conditions, of course, do restrict the distribution generating the ε_i's. The second condition, for example, requires some absolute moment greater than fourth order to exist for any linear combination of the ε_i's. This condition ensures that one can apply the multivariate central limit theorems used in the following analysis; it essentially limits the amount of weight in the tails of the distribution functions generating the random vectors Y_i and X_i.

Full-Information Estimation

Using the relation for Y_i given by model (2), which formally should be written as $Y_i = f_i^0 + \varepsilon_i$ where $f_i^0 \equiv f(X_i, \gamma_0)$ is a vector of conditional means evaluated at the true parameter values, one can eliminate Y_i in the expression of Q_N to obtain

$$Q_N = \frac{1}{N} \sum_{i=1}^{N} (- \ln|\Omega_i| - ((f_i^0 - f_i) + \varepsilon_i)'\Omega_i^{-1}((f_i^0 - f_i) + \varepsilon_i)).$$

Differentiating Q_N with respect to an arbitrary element of γ and ω yields

$$\frac{\partial Q_N}{\partial \gamma_k} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial q_i}{\partial \gamma_k} = \frac{1}{N} \sum_{i=1}^{N} \left(2 \frac{\partial f_i'}{\partial \gamma_k} \Omega_i^{-1} (f_i^0 - f_i + \varepsilon_i) \right)$$

and

$$\frac{\partial Q_N}{\partial \omega_k} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial q_i}{\partial \omega_k} = \frac{1}{N} \sum_{i=1}^{N} \left[- \operatorname{tr}\left(\Omega_i^{-1} \frac{\partial \Omega_i}{\partial \omega_k} \right) + \operatorname{tr}\left(\Omega_i^{-1} \frac{\partial \Omega_i}{\partial \omega_k} \Omega_i^{-1} (f_i^0 - f_i + \varepsilon_i)(f_i^0 - f_i + \varepsilon_i)' \right) \right]$$

where we have used the facts $\frac{\partial \ell n |\Omega_i|}{\partial \omega_k} = \operatorname{tr}\left(\Omega_i^{-1} \frac{\partial \Omega_i}{\partial \omega_k} \right)$, $\frac{\partial \Omega_i^{-1}}{\partial \omega_k} = - \Omega_i^{-1} \frac{\partial \Omega_i}{\partial \omega_k} \Omega_i^{-1}$, and $\operatorname{tr}(AB) = \operatorname{tr}(BA)$. Define $\Omega_i^0 = \Omega(X_i, \omega_0)$ where ω_0 is the true value of ω. Noting that $E(\varepsilon_i \varepsilon_i') = \Omega_i^0$ and ε_i and X_i are independent which implies $E(f_i' \Omega_i \varepsilon_i) = 0$ for all $\theta \in \Theta$, one can verify that the expectations of these derivatives are

$$(7) \qquad E\left(\frac{\partial Q_N}{\partial \gamma_k} \right) = \frac{1}{N} \sum_{i=1}^{N} E\left(\frac{\partial q_i}{\partial \gamma_k} \right)$$

$$= \frac{2}{N} \sum_{i=1}^{N} E\left[\frac{\partial f_i'}{\partial \gamma_k} \Omega_i^{-1} (f_i^0 - f_i + \varepsilon_i) \right]$$

$$= \frac{2}{N} \sum_{i=1}^{N} E\left[\frac{\partial f_i'}{\partial \gamma_k} \Omega_i^{-1} (f_i^0 - f_i) \right]$$

and

$$(8) \qquad E\left(\frac{\partial Q_N}{\partial \omega_k} \right) = \frac{1}{N} \sum_{i=1}^{N} E\left(\frac{\partial q_i}{\partial \omega_k} \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} E\left[\operatorname{tr}\left(\Omega_i^{-1} \frac{\partial \Omega_i}{\partial \omega_k} \Omega_i^{-1} \left(-\Omega_i^{-1} + (f_i^0 - f_i + \varepsilon_i)(f_i^0 - f_i + \varepsilon_i)' \right) \right) \right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} E\left[\operatorname{tr}\left(\Omega_i^{-1} \frac{\partial \Omega_i}{\partial \omega_k} \Omega_i^{-1} \left(-\Omega_i^{-1} + \Omega_i^0 + (f_i^0 - f_i)(f_i^0 - f_i)' \right) \right) \right]$$

where expectations in the last lines of (7) and (8) are computed only with respect to the random elements of X_i (i.e., expressions for expectations with respect to ε_i have been explicitly substituted into these last lines).

The estimator $\hat{\theta}$ is defined by the system of equations

$$\ell(\hat{\theta}) \equiv \frac{\partial Q_N}{\partial \theta}\Big|_{\hat{\theta}} = 0.$$

Formally, it can be shown that a solution to this system of equations lies in the interior of the parameter space Θ with probability one which ensures $\hat{\theta} \in \Theta$ in sufficiently large samples.[1] An exact first order Taylor's expansion of $\ell(\hat{\theta})$ around the true parameter value θ_0 yields

$$\ell(\theta_0) - H(\theta_f)(\hat{\theta} - \theta_0) = 0$$

where θ_f lies between $\hat{\theta}$ and θ_0 and in the set Θ. Solving this system of equations for $\hat{\theta} - \theta_0$, one obtains

$$(9) \qquad (\hat{\theta} - \theta_0) = H^{-1}(\theta_f)\ell(\theta_0).^2$$

---

[1] Inspection of equations (7) and (8) reveals that $E\left(\frac{\partial Q_N}{\partial \theta}\right) = 0$ if one sets $\theta = \theta_0$ which lies in the interior of the set Θ. The regularity conditions listed above guarantee that $\frac{\partial Q_N}{\partial \theta}$ converges strongly to $E\left(\frac{\partial Q_N}{\partial \theta}\right)$. Thus, we know that there exists at least one $\theta \in \Theta$ (namely $\theta = \theta_0$) such that $\frac{\partial Q_N}{\partial \theta}$ converges to zero with probability one. Formally, in the following analysis we only require that there exists a $\theta \in \Theta$ such that $\operatorname{plim}\left(\frac{\partial Q_N}{\partial \theta}\right) = 0$ which is assured.

[2] The following analysis formally only requires $H(\theta_f)$ to be nonsingular in the limit.

The regularity conditions specified above are sufficient to prove that $\hat{\theta}$ is strongly consistent for θ_0. Where d_i, h_i, and g_i denote elements of the gradient vector $\frac{\partial q_i}{\partial \theta}$, the hessian matrix $\frac{\partial^2 q_i}{\partial \theta \partial \theta'}$, and the matrix of outer partials $\frac{\partial q_i}{\partial \theta}\frac{\partial q_i}{\partial \theta'}$, respectively, these conditions require $\lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N} E|d_i|\right) < \infty$, $\lim_{N\to\infty}\frac{1}{N}\left(\sum_{i=1}^{N} E|h_i|\right) < \infty$, and $\lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N} E|g_i|\right) < \infty$ for each $\theta \in \Theta$. Combined with the assumption of independence, these restrictions directly imply

$$\operatorname{slim}(\ell(\theta)) = \lim_{N\to\infty} E(\ell(\theta)),$$

$$\operatorname{slim}(H(\theta)) = \lim_{N\to\infty} E(H(\theta)) \equiv \bar{H}(\theta),$$

and

$$\operatorname{slim}(G(\theta)) = \lim_{N\to\infty} E(G(\theta)) \equiv \bar{G}(\theta),$$

where slim denotes strong convergence (i.e., convergence almost everywhere),[1] and this convergence is uniform in θ on the set Θ.[2] Inspection of the

---

[1] For any sequence of random variables a_i, $i = 1,\ldots,$ the condition $\lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N} E|a_i|\right) < \infty$ implies that the quantity $\frac{1}{N}\sum_{i=1}^{N} a_i$ converges strongly (or converges almost everywhere) to $\lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N} E(a_i)\right)$ as N goes to infinity (see Chung, p. 125 [1974]).

[2] Considering for simplicity the single parameter case where $H(\theta)$ and $\bar{H}(\theta)$ are scalars, the quantity $H(\theta)$ converges strongly to $\bar{H}(\theta)$ uniformly in θ on the set Θ if, for each $\delta > 0$, there exists a $\bar{N}(\delta)$ (which does not depend on θ) such that $|H(\theta) - \bar{H}(\theta)| < \delta$ for $N > \bar{N}$ with probability one. As noted above, the regularity conditions restrict the set Θ. Combined with the independence assumption, they directly imply that, for each $\delta > 0$ and $\theta \in \Theta$, there exists $N^*(\theta, \delta) < \infty$ such that $|H(\theta) - \bar{H}(\theta)| < \delta$ for $N > N^*$ with probability one. Uniform convergence, then, follows by taking $\bar{N}(\delta) = \sup_{\theta\in\Theta} N^*(\theta, \delta)$.

expectations of derivatives given by (7) and (8) reveals that $E(\ell(\theta_0)) = 0$, so

$$\text{(10)} \qquad \text{slim}(\ell(\theta_0)) = 0.$$

Combining these results, it is possible to show that

$$\text{slim}(\hat{\theta} - \theta_0) = [\text{slim } H(\theta_f)]^{-1} \text{ slim}(\ell(\theta_0))$$

$$= 0.^{1}$$

Since θ_f is between $\hat{\theta}$ and θ_0, it is also evident that θ_f is strongly consistent for θ_0 (i.e., $\text{slim}(\theta_f) = \theta_0$). Furthermore, one can show that $\text{slim}(H(\theta_f)) = \text{slim}(H(\theta)) = \bar{H}(\theta_0)$ and $\text{slim}(G(\theta)) = \bar{G}(\theta_0)$.[2]

---

[1] Considering for simplicity the single parameter case, verification of this result requires that there exists for each $\delta > 0$, a $\bar{N}(\delta)$ so $|H^{-1}(\theta_f)\ell(\theta_0)| < \delta$, or equivalently $|\ell(\theta_0)| < \delta\,|H(\theta_f)|$, with probability one (wpo) whenever $N > \bar{N}$. Define $b = \inf_{\theta \in \Theta} |\bar{H}(\theta)|$ and observe $b > 0$ since $\bar{H}(\theta)$ is invertable by assumption for all $\theta \in \Theta$. Given the restriction that the root $\hat{\theta}$, and thus θ_f, lie in the set Θ wpo for sufficiently large N, $\text{slim}(\bar{H}(\theta) - H(\theta)) = 0$ uniformly in θ on Θ implies the existence of a $N^*(\phi)$ for any ϕ with $0 < \phi < b$ such that $|\bar{H}(\theta_f) - H(\theta_f)| < \phi$ wpo for $N > N^*$. Since $b \leq |\bar{H}(\theta_f)| = |H(\theta_f) + \bar{H}(\theta_f) - H(\theta_f)| \leq |H(\theta_f)| + |\bar{H}(\theta_f) - H(\theta_f)|$, we have $|H(\theta_f)| > b - \phi > 0$ wpo for $N > N^*$. Since $\text{slim}(\ell(\theta_0)) = 0$, we know there is a $N^{**}(\varepsilon)$ for any $\varepsilon > 0$ so that $|\ell(\theta_0)| < \varepsilon$ wpo for $N > N^{**}$. Taking $\varepsilon = \delta(b - \phi)$ and observing $\varepsilon < \delta|H(\theta_f)|$, we see that $\bar{N} = \max(N^*, N^{**}) < \infty$ guarantees $|\ell(\theta_0))| < \delta|H(\theta_f)|$ wpo for $N > \bar{N}$.

[2] Since $\text{slim}(H(\theta) - \bar{H}(\theta)) = 0$ uniformly in θ on the set Θ and $\hat{\theta} \in \Theta$ for sufficiently large N with probability one, it follows that $\text{slim}(H(\hat{\theta}) - \bar{H}(\hat{\theta})) = 0$. Given uniform continuity of $\bar{H}(\theta)$, one can show $\text{slim}(\bar{H}(\hat{\theta}) - \bar{H}(\hat{\theta}_0)) = 0$. Combining results we conclude $\text{slim}(H(\hat{\theta}) - \bar{H}(\theta_0)) = 0$. The same argument applies to show $\text{slim}(H(\theta_f) - \bar{H}(\theta_0)) = 0$ and $\text{slim}(G(\hat{\theta}) - \bar{G}(\theta_0)) = 0$.

Since $\ell(\theta_0)$ is an average of independently distributed random vectors which have zero mean (see formulas (7) and (8)) and which satisfy the above regularity conditions, one can apply Liapounov's central limit theorem to obtain

$$\text{(11)} \qquad \text{dlim}(\sqrt{N}\,\ell(\theta_0)) = N(0,\ \bar{G}(\theta_0)),$$

where $\bar{G}(\theta_0) \equiv \text{plim}(G(\theta_0))$ is the probability limit of the matrix of outer partials.[1] Combining these results, a standard application of asymptotic theory yields

$$\text{(12)} \qquad \text{dlim}(\sqrt{N}(\hat{\theta} - \theta_0)) = [\text{plim}\ (H(\theta_f))]^{-1}\ \text{dlim}(\sqrt{N}\,\ell(\theta_0))$$

$$= N(0,\ \bar{H}^{-1}(\theta_0)\bar{G}(\theta_0)\bar{H}^{-1}(\theta_0));$$

so $\hat{\theta}$ possesses an asymptotic normal distribution. In large samples, then, $\hat{\theta}$ is approximately distributed according to a normal probability law of the form

$$\text{(13)} \qquad \hat{\theta} \overset{\cdot}{\sim} N(\theta_0,\ \frac{1}{N} H^{-1}(\hat{\theta})G(\hat{\theta})H^{-1}(\hat{\theta})).$$

---

[1] Given a sequence of independent random variables a_i, $i = 1,\ldots,$ with zero means, application of this central limit theorem requires $\lim \frac{1}{N^\delta}\left(\frac{M_N}{S_N}\right) = 0$ for $\delta > 0$ where $M_N = \frac{1}{N}\sum_{i=1}^{N} E|a_i|^{2+\delta}$ and $S_N = \frac{1}{N}\sum_{i=1}^{N} E(a_i)^2$ (see Loeve [1977, p. 287]). The regularity conditions specified above ensure satisfaction of this requirement for all linear combinations of the random vectors $\frac{\partial q_i}{\partial \theta}$, so one can apply a multivariate version of this theorem (see Rao, p. 128 [1973]).

If disturbances are distributed according to a multivariate normal distribution, then we have the well known result $E(G(\theta_0)) = E(H(\theta_0))$ which directly implies $\bar{G}(\theta_0) = \bar{H}(\theta_0)$. In this case the asymptotic covariance matrix for $\sqrt{N}(\hat{\theta} - \theta_0)$ reduces to $\bar{H}^{-1}(\theta_0)$ and the asymptotic distribution assumed in maximum likelihood estimation given by (5) applies. Most researchers assume normality and draw inferences using output from a standard maximum likelihood computer package which reports $\frac{1}{N} H^{-1}(\hat{\theta})$ as the approximate covariance matrix for parameter estimates. Using instead the covariance matrix $\frac{1}{N} H^{-1}(\hat{\theta})G(\hat{\theta})H^{-1}(\hat{\theta})$ -- which is readily computable -- avoids the need for any specific distributional assumptions.

While use of the matrix $H^{-1}(\hat{\theta})$ leads to the incorrect computation of standard errors for the estimates $\hat{\omega}$ of the covariance parameters in the absence of normality, it produces the correct standard errors for the estimates $\hat{\gamma}$ of the regression coefficients. Even without normality of disturbances, one can easily verify that the matrices $G_{\gamma\gamma}(\gamma_0, \omega_0)$, $H_{\gamma\gamma}(\gamma_0, \omega_0)$, and $H_{\gamma\omega}(\gamma_0, \omega_0)$, defined in the above partitions of $G(\theta_0)$ and $H(\theta_0)$, satisfy $E(H_{\gamma\omega}(\gamma_0, \omega_0)) = 0$ and $E(G_{\gamma\gamma}(\gamma_0, \omega_0)) = E(H_{\gamma\gamma}(\gamma_0, \omega_0))$. Given the above regularity conditions, it follows that

$$\text{(14)} \qquad \bar{H}_{\gamma\omega}(\theta_0) \equiv \operatorname{slim}(H_{\gamma\omega}(\theta_0)) = 0$$

and

$$\text{(15)} \qquad \bar{G}_{\gamma\gamma}(\theta_0) \equiv \operatorname{slim}(G_{\gamma\gamma}(\gamma_0, \omega_0)) = \operatorname{slim}(H_{\gamma\gamma}(\gamma_0, \omega_0)) \equiv \bar{H}_{\gamma\gamma}(\theta_0).$$

The first relation implies

$$(16)\qquad \bar{H}^{-1}(\theta_0) = \begin{bmatrix} \bar{H}_{\gamma\gamma}^{-1}(\theta_0) & 0 \\ 0 & \bar{H}_{\omega\omega}^{-1}(\theta_0) \end{bmatrix},$$

and combining this result with the second relation yields

$$(17)\qquad \bar{H}^{-1}(\theta_0)\bar{G}(\theta_0)\bar{H}^{-1}(\theta_0) = \begin{bmatrix} \bar{H}_{\gamma\gamma}^{-1}(\theta_0) & \bar{H}_{\gamma\gamma}^{-1}(\theta_0)\bar{G}_{\gamma\omega}(\theta_0)\bar{H}_{\omega\omega}^{-1}(\theta_0) \\ \bar{H}_{\omega\omega}^{-1}(\theta_0)\bar{G}_{\omega\gamma}(\theta_0)\bar{H}_{\gamma\gamma}^{-1}(\theta_0) & \bar{H}_{\omega\omega}^{-1}(\theta_0)\bar{G}_{\omega\omega}(\theta_0)\bar{H}_{\omega\omega}^{-1}(\theta_0) \end{bmatrix}.$$

On the basis of these formulas, we see that using the matrix $\frac{1}{N} H^{-1}(\hat{\theta})$ rather than $\frac{1}{N} H^{-1}(\hat{\theta})G(\hat{\theta})H^{-1}(\hat{\theta})$ gives a valid approximation for the asymptotic covariance matrix associated with $\hat{\gamma}$, but the wrong approximation for $\hat{\omega}$. One also obtains the wrong approximation for the covariance between $\hat{\gamma}$ and $\hat{\omega}$ unless third moments of all linear combinations of the disturbances are zero in which case $\bar{G}_{\gamma\omega}(\theta_0) = 0$, and the off diagonal blocks of the matrix in (16) and (17) are equal.

This finding provides the basis for a simple proof of the well known proposition that maximum likelihood estimation of a linear simultaneous equations model yields consistent estimates of structural coefficients and standard errors for these estimates that are asymptotically valid even if disturbances are not in fact normally distributed.[1] Consider the system of simultaneous equations

$$(18)\qquad \Gamma Y_i = \Pi X_i + \nu_i, \qquad i = 1,\ldots,N,$$

where Γ and Π are matrices containing the structural coefficients, and ν_i

[1] This result was first shown by Anderson-Rubin (1949).

is a disturbance vector with $E(\nu_i \nu_j') = \Phi$ if $i = j$ and $= 0$ otherwise. The implied reduced form associated with (18) is clearly in the form of model (2) with $f_i = \Gamma^{-1} \Pi X_i$, $\varepsilon_i = \Gamma^{-1} \nu_i$, and $\Omega_i = \Gamma^{-1} \Phi \Gamma^{-1}'$. Maximum likelihood estimation of (18) amounts to applying the estimation procedure outlined above. A limited-information maximum likelihood method estimates the parameters of the reduced form model imposing the restrictions implied by a single structural equation, and full-information methods impose restrictions implied by several structural equations when estimating these parameters. Typically, all identifying restrictions are introduced through the matrices Γ and Π, and the covariance matrix Φ is left unrestricted. Since $f_i = \Gamma^{-1} \Pi X_i$, we see that the parameter vector γ includes all the coefficients of Γ and Π as its elements, and the usual type of identifying restrictions limits the number of elements in γ and constrains the form of the expected value f_i. There are two choices for the parameterization of the vector ω: one may include the coefficients of Γ and Φ as elements of ω and treat Ω in estimation as a function of the form $\Omega = \Gamma^{-1} \Phi \Gamma^{-1}'$; or, one may simply include the elements of Ω in ω and estimate Ω directly forcing the matrix $\Gamma \Omega \Gamma'$ to satisfy the same restrictions as the matrix Φ. As long as one accounts for all constraints in estimation, maximization of the function Q_N yields exactly the same $\hat{\gamma}$ and covariance matrix for these estimates no matter how one chooses to parameterize ω.[1] This observation reflects the well known invariance property of maximum likelihood estimates.

[1] To verify this result, let the parameter vector ϕ include all the unknown elements of Φ and define $\sigma' = (\gamma', \phi')$. The parameter vectors θ and σ are related by the functions $\theta = r(\sigma) \equiv \begin{bmatrix} \gamma \\ b(\gamma, \phi) \end{bmatrix}$ where $\omega = b(\gamma, \phi)$ is determined by the system of equations $\Omega = \Gamma^{-1} \Phi \Gamma^{-1}'$. Define the matrix

Considering the latter parameterization of ω, the absence of restrictions on the covariance matrix Φ implies that there are no constraints relating the elements of γ and ω. In this case, since $\hat{\gamma}$ is the maximum likelihood estimate of Γ and Π, it follows immediately from the results derived above that maximum likelihood estimation of (18) produces consistent estimates and appropriate large sample standard errors for structural coefficients even in the absence of normally distributed error terms. It is further evident that this proposition remains valid for any simultaneous equations model that is linear in endogenous variables; in particular, one can generalize (18) to permit heteroscedasticity and nonlinearities in parameters and in exogenous variables. If, however, there are any sort of restrictions on the covariance matrix Φ which translate into constraints relating the elements of γ to those of ω, then the proposition no longer holds, and one must use the asymptotic distribution for estimates given by (13) which does not rely on the normality assumption.

Permitting the existence of constraints relating the elements of γ and ω introduces no significant complications in the above analysis. No modifications are required in the derivation of the asymptotic properties of the full-information estimator $\hat{\theta}$. The approximate distribution given by (12) and (13) still applies with partials computed accounting for the

---

of first partials $R(\sigma) = \frac{\partial r'}{\partial \sigma} = \begin{bmatrix} I & B_\gamma \\ 0 & B_\phi \end{bmatrix}$ with $B_\gamma = \frac{\partial b'}{\partial \gamma}$ and $B_\phi = \frac{\partial b'}{\partial \phi}$, and denote the reparameterized "likelihood" function as $L_N(\sigma) \equiv Q_N(r(\sigma))$. Since $\frac{\partial L_N}{\partial \sigma} = R\, \ell(\theta)$ and R is nonsingular, $\frac{\partial L_N}{\partial \sigma} = 0$ only when $\ell(\theta) = 0$, which implies $\hat{\theta} = r(\hat{\sigma})$. Thus, maximizing Q_N with respect to θ yields exactly the same value for $\hat{\gamma}$ as maximizing L_N with respect to σ. Furthermore, one can readily verify that $K(\hat{\sigma}) = -\frac{\partial^2 L}{\partial \sigma \partial \sigma'}\Big|_{\hat{\sigma}} = R(\hat{\sigma})H(\hat{\theta})R(\hat{\sigma})'$ using the fact $\ell(\hat{\theta}) = 0$. Letting V_1 and V_2 denote the (1, 1) block of the matrices $K^{-1}(\hat{\sigma})$ and $H^{-1}(\hat{\sigma})$, respectively, which correspond to the covariance matrices of $\hat{\gamma}$ for the alternative parameterizations, it is possible to show that $V_1 = V_2$.

restrictions relating γ and ω. The imposition of such restrictions invalidates relations (14) and (15), and it, of course, in general yields more efficient parameter estimates.

We next discuss quasi-maximum likelihood techniques that can be applied to estimate subsets of parameters rather than all parameters simultaneously. These limited-information procedures are particularly useful if the aim of an analysis is to estimate only covariance parameters or only regression coefficients. These methods decrease the amount of computational burden required to estimate either the elements of ω or γ, and, in the absence of constraints relating γ and ω, these limited-information procedures permit one to subdivide the estimation problem and estimate ω conditioning on a value of γ and vise verse.

Estimating Covariance Parameters

Consider first a simple limited-information procedure for obtaining an estimate of ω. Given any consistent estimate of γ_0, denoted as $\tilde{\gamma}$, define the distance function

$$(19) \qquad \tilde{Q}_N(\omega) \equiv Q_N(\tilde{\gamma}, \omega)$$

$$= \frac{1}{N} \sum_{i=1}^{N} (- \ell n|\Omega_i| + \tilde{\varepsilon}_i' \Omega_i^{-1} \tilde{\varepsilon}_i)$$

where $\tilde{\varepsilon}_i \equiv Y_i - f(X_i, \tilde{\gamma})$ is a vector of fitted residuals. A standard nonlinear least squares estimator is a logical candidate for $\tilde{\gamma}$. The function $\tilde{Q}_N$ looks like Q_N except that known residuals $\tilde{\varepsilon}_i$ replace unknown disturbances $Y_i - f_i$. Q_N may be interpreted as a "likelihood function" that treats residuals as if they were the true values of the disturbances. Below we investigate the asymptotic properties of the estimator $\tilde{\omega}$ derived

by maximizing the function $\tilde{Q}_N$. $\tilde{\omega}$ is easier to compute than $\hat{\omega}$ defined above because one does not simultaneously calculate estimates for γ; optimization is carried out over fewer parameters.

Given the same conditions assumed in proving (12), one can show that $\tilde{\omega}$ is a consistent estimate for the true parameter values ω_0, and it possesses an asymptotic normal distribution. The estimator $\tilde{\omega}$ is determined by the system of equations

$$\frac{\partial \tilde{Q}_N}{\partial \omega}\Big|_{\tilde{\omega}} = \frac{\partial Q_N}{\partial \omega}\Big|_{\tilde{\gamma},\tilde{\omega}} \equiv \ell_\omega(\tilde{\gamma}, \tilde{\omega}) = 0.$$

An exact first order Taylor's expansion admits a solution of the form

$$(20) \qquad (\tilde{\omega} - \omega_0) = H^{-1}_{\omega\omega}(\tilde{\gamma}, \omega_g)(\ell_\omega(\gamma_0, \omega_0) - H_{\omega\gamma}(\gamma_g, \omega_g)(\tilde{\gamma} - \gamma_0))$$

where ω_g and γ_g lie between $\tilde{\omega}$ and ω_0 and between $\tilde{\gamma}$ and γ_0, respectively, and we have used the obvious facts $-\frac{\partial^2 \tilde{Q}_N}{\partial\omega\partial\omega'} = H_{\omega\omega}$ and $\frac{\partial^2 \tilde{Q}_N}{\partial\omega\partial\gamma'} = H_{\omega\gamma}$.[1] The above regularity conditions, relation (10), and the consistency of the estimator $\tilde{\gamma}$ directly lead to the conclusion

$$\text{plim}(\tilde{\omega}) = \omega_0.^{2}$$

This convergence is strong (i.e., $\text{slim}(\tilde{\omega}) = \omega_0$) if $\text{slim}(\tilde{\gamma}) = \gamma_0$ rather than

---

[1] Expanding $\ell_\omega(\tilde{\gamma}, \tilde{\omega}) = 0$ around ω_0 yields $(\tilde{\omega} - \omega_0) = H^{-1}_{\omega\omega}(\tilde{\gamma}, \omega_g)\ell(\tilde{\gamma}, \omega_0)$, and expanding $\ell(\tilde{\gamma}, \omega_0)$ around γ_0 yields (20).

[2] Because $\text{slim}(H_{\omega\gamma}) = 0$, the second regularity condition specifying $\text{slim}(H)$ is a nonsingular matrix necessarily implies that $\text{slim}(H_{\omega\omega}) = \bar{H}_{\omega\omega}$ is nonsingular which is required to prove the consistency and the asymptotic normality of $\tilde{\omega}$.

$\text{plim}(\tilde{\gamma}) = \gamma_0$. Combining this result with relations (11) and (14) further implies

$$(21) \qquad \text{dlim}(\sqrt{N}(\tilde{\omega} - \omega_0)) = N(0, \bar{H}_{\omega\omega}^{-1}(\theta_0)\bar{G}_{\omega\omega}(\theta_0)\bar{H}_{\omega\omega}^{-1}(\theta_0))$$

where $\bar{H}_{\omega\omega}(\theta_0) = \text{slim}(H_{\omega\omega}(\gamma_0, \omega_0))$ and $\bar{G}_{\omega\omega}(\theta_0) \equiv \text{slim}(G_{\omega\omega}(\gamma_0, \omega_0))$. The estimator $\tilde{\omega}$, then, in large samples is approximately normally distributed. In particular,

$$(22) \qquad \tilde{\omega} \overset{\cdot}{\sim} N(\omega_0, \frac{1}{N} H_{\omega\omega}^{-1}(\tilde{\gamma}, \tilde{\omega})G_{\omega\omega}(\tilde{\gamma}, \tilde{\omega})H_{\omega\omega}^{-1}(\tilde{\gamma}, \tilde{\omega})).$$

There are two characteristics of this distribution that are important to recognize. First, notice that one can compute an estimate of the covariance matrix of $\tilde{\omega}$ neglecting the fact that $\tilde{\gamma}$ is an estimated value of γ_0. The covariance matrix appearing in (21) and (22) depends only on the matrix of second partials and the average of the matrix of outer partials associated with the function $\tilde{Q}_N$. It is, then, possible to treat $\tilde{\gamma}$ as if it were the true value of γ, and residuals may be treated as if they were the true disturbances. In fact, inspection of (21) reveals that there would be no gain in estimation efficiency if γ_0 were actually known.

Second, the limited-information estimator $\tilde{\omega}$ has the same asymptotic distribution as the full-information estimator $\hat{\omega}$ whose distribution is given by (13). Both estimators of ω are equally efficient, so there is no gain in simultaneously estimating the elements of γ and ω.

Estimating Regression Coefficients

There is an analogous procedure for obtaining an estimate of γ. This procedure is, of course, nothing more than nonlinear generalized least squares.

The nonlinear generalized least square estimate of γ, denoted as γ^*, is defined as that value of γ that maximizes the function

$$(23) \qquad Q_N^*(\gamma) = Q_N(\gamma, \omega^*)$$

$$= \frac{1}{N} \sum_{i=1}^{N} [- \ell n|\Omega(X_i, \omega^*)| - (Y_i - f_i)'\Omega^{-1}(X_i, \omega^*)(Y_i - f_i)]$$

where ω^* is any consistent estimate of ω_0. Instead, then, of evaluating Q_N as a consistent estimate of γ_0 as we did above to obtain the estimate $\tilde{\omega}$, we evaluate it at a consistent estimate of ω_0. Switching the roles of γ and ω in the above derivation of $\tilde{\omega}$ and using relations (14) and (15), one can easily verify the well known large sample result

$$(24) \qquad \gamma^* \overset{\cdot}{\sim} N(\gamma_0, \frac{1}{N} H_{\gamma\gamma}^{-1}(\gamma^*, \omega^*))$$

where $H_{\gamma\gamma}$ is minus the Hessian matrix of the function $Q_N^*(\gamma)$.

Similar to $\tilde{\omega}$, the nonlinear generalized least squares estimator γ^* has the same asymptotic distribution as its full-information counterpart $\hat{\gamma}$. Also, one neglects the fact that ω^* is an estimated quantity when computing the covariance matrix of γ^*. None of these results assumes that the estimates γ to be asymptotically uncorrelated or independent of the estimates of ω. The estimates of γ and ω obtained by any of the procedures outlined above will not be independently distributed in large samples unless the third moments of all linear combinations of disturbances are zero. The crucial fact responsible for the equivalent asymptotic properties of limited- and full-information estimators is that the matrix of cross partials evaluated at the true parameter values, $H_{\gamma\omega}(\gamma_0, \omega_0)$ converges to zero in probability.

Estimating Regression Coefficients with Unknown Forms of Heteroscedasticity

Often the functional form of the covariance matrix $\Omega(X_i, \omega)$ is unknown, and a researcher wants to obtain consistent estimates of the regression coefficients and the appropriate standard errors of these estimates without making invalid or arbitrary assumptions concerning the covariance structure of disturbances. The above analysis provides a methodology for accomplishing this task.

A nonlinear least squares procedure offers a natural method for estimating γ without the need for any specification of the covariance matrix. Instead of the distance function Q_N given by (14), set $\Omega_i = I$ for all i in Q_N, and define the new function

$$(25) \qquad M_N(\gamma) = \frac{1}{N} \sum_{i=1}^{N} m_i$$

$$= -\frac{1}{N} \sum_{i=1}^{N} (Y_i - f_i)'(Y_i - f_i)$$

where m_i is the function of γ, Y_i, and X_i defined by the second expression for M_N. Maximizing the function M_N with respect to γ yields the nonlinear least squares estimator $\hat{\gamma}^*$. Following the steps presented above in the derivation of the asymptotic properties of the full-information estimator with M_N, m_i and $\hat{\gamma}^*$ replacing Q_N, q_i, and $\hat{\theta}$, respectively, it is straightforward to show that

$$(26) \qquad \hat{\gamma}^* \overset{\cdot}{\sim} N\left(\gamma_0, \frac{1}{N}\left[-\frac{\partial^2 M_N}{\partial\gamma\partial\gamma'}\Big|_{\hat{\gamma}^*}\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial m_i}{\partial\gamma}\Big|_{\hat{\gamma}^*}\frac{\partial m_i}{\partial\gamma'}\Big|_{\hat{\gamma}^*}\right]\left[-\frac{\partial^2 M_N}{\partial\gamma\partial\gamma'}\Big|_{\hat{\gamma}^*}\right]^{-1}\right).$$

According to this formula, the large-sample covariance matrix of $\hat{\gamma}^*$ depends simply on the matrix of second partials and the average of the

matrix of outer partials associated with the function M_N evaluated at the nonlinear least squares estimate $\hat{\gamma}^*$. In contrast to the function Q_N for the function M_N we in general have

$$\text{slim}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial m_i}{\partial \gamma}\Big|_{\hat{\gamma}^*}\frac{\partial m_i}{\partial \gamma'}\Big|_{\hat{\gamma}^*}\right) \neq \text{slim}\left(-\frac{\partial^2 M_N}{\partial \gamma \partial \gamma'}\Big|_{\hat{\gamma}^*}\right);$$

so the covariance matrix associated with estimates of regression coefficients does not simplify as it did in the previous analysis. The result given by (26) represents a multi-equation extension of a result due to White (1980).

This is a very useful finding. The distribution for $\hat{\gamma}^*$ given by (26) is valid in large samples no matter what the actual form of the covariance structure associated with disturbances. Using this distribution to compute standard errors and to test hypotheses does not rely on any assumptions concerning the form or the absence of heteroscedasticity, and, as with all the results developed in this section, it does not rely on specific distributional assumptions for disturbances. Besides the luxury of avoiding the difficulty of correctly specifying covariance matrices, this method of estimation permits one to analyze many complicated models that are not easily estimated using alternative procedures. An example of such a model is one in which the dependent variables $Y_i(t)$ appearing in (1) are nonindependent discrete random variables. Using (26), it is possible to estimate the parameters determining the probabilities generating these variables and to test hypotheses concerning these parameters without modeling how these variables are correlated with one another which requires the specification of joint probabilities.

III. Hypothesis Tests

This section examines the general asymptotic properties of statistics derived from the theory of maximum likelihood to test hypotheses in a large sample setting. Specifically, assuming the parameter vector θ contains p elements, we consider tests of the hypothesis

$$b(\theta) = 0, \tag{27}$$

where b is a s x 1 vector with $s < p$, which may equivalently be written as

$$\theta = r(\alpha), \tag{28}$$

where α is a n x 1 vector of parameters with $n = p - s$. The components of the vectors b and r are assumed to be functions admitting uniformly continuous partial derivatives of the first order. Below we denote the matrices of these partials as $B(\theta^*) \equiv \frac{\partial b}{\partial \theta'}\Big|_{\theta^*}$ and $R(\alpha^*) \equiv \frac{\partial r'}{\partial \alpha}\Big|_{\alpha^*}$.

Tests Based on Full-Information Estimation

There are three well-known large sample tests of (27) and (28) justified using maximum likelihood theory: Wald's test (W), the likelihood ratio test (LRT), and Rao's test also known as the Lagrange multiplier test (LMT).[1] Letting $\hat{\theta}$ denote the unconstrained estimator that maximizes

[1] The test statistic used in this paper (see (31) below) is from Rao (1947). It is, however, equivalent to the Lagrange multiplier test developed by Silvey (1959). Maximizing $Q_N(\theta)$ subject to (27) is achieved by unconstrained maximization of $Q_N(\theta) + \lambda(B(\theta))$ where λ is a vector of Lagrange multipliers. Optimality implies $\frac{\partial Q_N}{\partial \theta} = B(\theta)\lambda$. Inserting this relation into (31) yields the test proposed by Silvey (1959).

the function $Q_N(\theta)$ defined by (4), and letting $\hat{\theta}_c = r(\hat{\alpha})$ denote the constrained estimator where $\hat{\alpha}$ maximizes $Q_N(r(\alpha))$, the statistics associated with each of these tests are:

$$W = N\, b(\hat{\theta})'[B(\hat{\theta})V(\hat{\theta})B(\theta)']^{-1}\, b(\hat{\theta}), \tag{29}$$

$$LRT = 2N[Q_N(\hat{\theta}) - Q_N(\hat{\theta}_c))], \tag{30}$$

and

$$LMT = N\, \ell(\hat{\theta}_c)'[H^{-1}(\hat{\theta}_c)]\, \ell(\hat{\theta}_c), \tag{31}$$

where $V(\hat{\theta})$ in (29) is the approximate asymptotic covariance matrix of $\sqrt{N}(\hat{\theta} - \theta_0)$ with $V(\hat{\theta}) = H^{-1}(\hat{\theta})$ when applying the theory of maximum likelihood. The computation of W requires estimation of only the unconstrained model, while the computation of LMT requires only estimation of the constrained model. To calculate LRT one obviously needs to estimate both models.

To determine the asymptotic distributions of W, LRT, and LMT, it is convenient to write these statistics in terms of the gradient vector $\ell(\theta_0)$ evaluated at the true value of θ, which under the null hypothesis satisfies the system of equations $\theta_0 = r(\alpha_0)$. Denote the vector of first partials and the matrix of second partials of $Q_N(r(\alpha))$ with respect to α as $e(\alpha) \equiv \frac{\partial Q_N}{\partial \alpha}\Big|_{\alpha}$ and $F(\alpha) \equiv -\frac{\partial^2 Q_N}{\partial \alpha \partial \alpha'}\Big|_{\alpha}$, respectively. Applying the chain rule, we have $e(\alpha) = R(\alpha)\, \ell(\theta)$ where $\theta = r(\alpha)$. Further application of this rule and using (10), it is possible to show that the matrix $F(\alpha_1)$ is asymptotically equivalent to the matrix $R(\alpha_2)H(\theta_1)R(\alpha_2)'$ when

$\text{plim}(\alpha_1) = \text{plim}(\alpha_2) = \alpha_0$ and $\text{plim}(\theta_1) = \theta_0$ (i.e., $\text{plim}[F(\alpha_1) - R(\alpha_2)H(\theta_1)R(\alpha_2)'] = 0$).[1] Using the argument employed to derive (9), one can show $(\hat{\alpha} - \alpha_0) = F^{-1}(\alpha_f)e(\alpha_0)$; so $(\hat{\alpha} - \alpha_0) = F^{-1}(\alpha_f)R(\alpha_0)\ell(\alpha_0)$.

First, consider the Wald test statistic. Expanding $b(\hat{\theta})$ around the parameter value θ_0, we have under the null hypothesis

$$b(\hat{\theta}) = B(\theta_h)(\hat{\theta} - \theta_0)$$

where θ_h lies between $\hat{\theta}$ and θ_0. Using (9) to eliminate $(\hat{\theta} - \theta_0)$ and substituting the resulting expression into (29) yields

$$W = N\, \ell(\theta_0)'\, K\, \ell(\theta_0)$$

with

$$K = H^{-1}(\theta_f)B(\theta_h)'\, [B(\hat{\theta})V(\hat{\theta})B(\hat{\theta})']^{-1}\, B(\theta_h)H^{-1}(\theta_f).$$

Observing that θ_h, θ_f, and $\hat{\theta}$ are all consistent for θ_0, it is possible to show that the matrix K converges strongly to $\bar{K}$ where

$$\bar{K} = \bar{H}^{-1}(\theta_0)B(\theta_0)'\, [B(\theta_0)\bar{V}(\theta_0)B(\theta_0)']^{-1}\, B(\theta_0)\bar{H}^{-1}(\theta_0) \tag{32}$$

with $\bar{V}(\theta_0) = \bar{H}^{-1}(\theta_0)$ in maximum likelihood estimation. A standard application of large sample theory implies that under the null hypothesis W has the same asymptotic distribution as the statistic

$$W_0 = (\sqrt{N}\, \ell(\theta_0))'\, \bar{K}(\sqrt{N}\, \ell(\theta_0)). \tag{33}$$

Consider next the test statistic LRT. An exact second order Taylor's expansion of the function $Q_N(\hat{\theta})$ around θ_0 yields

---

[1] To demonstrate this result formally, one must assume that the estimators θ_1, $r(\alpha_1)$, and $r(\alpha_2)$ lie in the set Θ with probability one for sufficiently large N, or these estimators are defined to lie in this set. This assumption combined with the strong convergence of the matrix $H(\theta)$ to $\bar{H}(\theta)$ uniformly in θ are sufficient to prove $\text{plim}[F(\alpha_1) - R(\alpha_2)H(\theta_1)R(\alpha_2)'] = 0$.

$$Q_N(\hat{\theta}) = Q_N(\theta_0) + \ell(\theta_0)'(\hat{\theta} - \theta_0) - \frac{1}{2}(\hat{\theta} - \theta_0)' H(\theta_a)(\hat{\theta} - \theta_0)$$

where θ_a lies between $\hat{\theta}$ and θ_0. Using (9) to eliminate $(\hat{\theta} - \theta_0)$, one obtains

$$Q_N(\hat{\theta}) = Q_N(\theta_0) + \ell(\theta_0)'[H^{-1}(\theta_f) - \frac{1}{2} H^{-1}(\theta_f)H(\theta_a)H^{-1}(\theta_f)]\ell(\theta_0).$$

An analogous expansion of $Q_N(r(\hat{\alpha}))$ yields

$$Q_N(r(\hat{\alpha})) = Q_N(\theta_0) + e(\alpha_0)'[F^{-1}(\alpha_f) - \frac{1}{2} F^{-1}(\alpha_f)F(\alpha_a)F^{-1}(\alpha_f)]e(\alpha_0).$$

Since $e(\alpha_0) = R(\alpha_0)\ell(\theta_0)$, we have

$$LRT = N\, \ell(\theta_0)'\, A_1 \ell(\theta_0),$$

where

$$A_1 = 2\, H^{-1}(\theta_f) - H^{-1}(\theta_f)H(\theta_a)H^{-1}(\theta_f) - R(\alpha_0)'[2F^{-1}(\alpha_f) - F^{-1}(\alpha_f)F(\alpha_a)F^{-1}(\alpha_f)]R(\alpha_0).$$

Using the above results, it can be shown that A_1 converges strongly to a matrix $\bar{A}$ where

$$\bar{A} = \bar{H}^{-1}(\theta_0) - R(\alpha_0)'[R(\alpha_0)\bar{H}(\theta_0)R(\alpha_0)']^{-1}R(\alpha_0). \tag{34}$$

Thus, under the null hypothesis, LRT is asymptotically equivalent to

$$LRT_0 = (\sqrt{N}\, \ell(\theta_0))'\, \bar{A}(\sqrt{N}\, \ell(\theta_0)). \tag{35}$$

Finally, consider the LMT statistic. Expanding $\ell(r(\hat{\alpha}))$ around α_0 yields

$$\ell(r(\hat{\alpha})) = \ell(\theta_0) - H(\theta_d)R(\alpha_d)'(\hat{\alpha} - \alpha_0)$$

$$= (I - H(\theta_d)R(\alpha_d)'F^{-1}(\alpha_f)R(\alpha_0))\ell(\theta_0)$$

where $\theta_d = r(\alpha_d)$ lies between $\hat{\theta}_c = r(\hat{\alpha})$ and θ_0. Substituting this relation into (31), one obtains

$$LMT = N\, \ell(\theta_0)'\, A_2\, \ell(\theta_0)$$

where

$$A_2 = [I - R(\alpha_0)'F^{-1}(\alpha_f)R(\alpha_d)H(\theta_d)]H^{-1}(\hat{\theta}_c)[I - H(\alpha_d)R(\alpha_d)'F^{-1}(\alpha_f)R(\alpha_0)].$$

The matrix A_2 can be shown to converge strongly to $\bar{A}$ defined by (34). Thus, LMT is asymptotically equivalent to

$$(36) \qquad LMT_0 = (\sqrt{N}\, \ell(\theta_0))'\, \bar{A}(\sqrt{N}\, \ell(\theta_0))$$

under the null hypothesis.

Regardless of specific distributional assumptions, we know from the previous section that the vector $\sqrt{N}\, \ell(\theta_0)$ converges to a normal distribution as N goes to infinity with covariance matrix $\bar{G}(\theta_0)$ (see (11)). Inspection of the formulas given by (33), (35), and (36) indicates the statistics W_0, LRT_0, and LMT_0 are quadratic forms in the vector $\sqrt{N}\, \ell(\theta_0)$. Thus, the problem of determining the asymptotic distribution of the test statistics W, LRT, and LMT reduces to one of determining the distribution of a quantity of the form $\eta' \wedge \eta$ where the vector $\eta \sim N(0, \Sigma)$ and $\wedge$ is a symmetric matrix of constants.

Assuming disturbances in model (2) follow a multivariate normal distribution, it is well known that each of these test statistics possesses a large sample chi-squared distribution with degrees of freedom equal to the number of restrictions s. A necessary and sufficient condition for the quadratic form $\eta' \wedge \eta$ to be distributed as a χ^2_k variate is for the matrix $\wedge\Sigma \equiv \wedge E(\eta\eta')$ to be idemponent (i.e., $\wedge\Sigma = \wedge\Sigma\wedge\Sigma$) with $tr(\wedge\Sigma) = k$.[1] When

[1] A reference to this theorem can be found in Rao (1973, p. 188). The statement of this theorem in the text uses the fact that the rank of an idemponent matrix equals its trace.

error terms are normally distributed, we have $\bar{G}(\theta_0) = \bar{H}(\theta_0)$ so the asymptotic covariance matrix associated with $\sqrt{N}\ \ell(\theta_0)$ is $\bar{H}(\theta_0)$. One can readily verify that the matrices $\bar{K}\ \bar{H}(\theta_0)$ and $\bar{A}\ \bar{H}(\theta_0)$ are each idemponent with traces equal to s. This simple observation, then, directly leads to the conclusion that W, LRT, and LMT converge to a χ^2_s distribution under the null hypothesis.

Relaxing the normality assumption means that $\bar{G}(\theta_0) \neq \bar{H}(\theta_0)$, and this alters the asymptotic properties of the test statistics LRT and LMT in some instances. No real problems arise with regard to the Wald statistic. It is well known that W is distributed as a large sample χ^2 variate if $V(\hat{\theta})$ in (29) is a consistent estimate of the asymptotic covariance matrix of $\sqrt{N}(\hat{\theta} - \theta_0)$. In the absence of normality, then, one simply sets $V(\hat{\theta}) = H^{-1}(\hat{\theta})G(\hat{\theta})H^{-1}(\hat{\theta})$. Unfortunately, no simple modifications of this sort exist for the statistics LRT and LMT. Before examining the cases in which these two statistics diverge asymptotically from the χ^2 distribution, the following analysis first identifies those instances where the familiar large sample properties apply.

When hypotheses involve only constraints on regression coefficients γ and do not involve covariance parameters ω, relaxing the normality assumption has no effect on the asymptotic properties of the test statistics LRT and LMT. Hypotheses of this nature imply that one can specialize the restrictions given by (27) and (28) to read

$$b_\gamma(\gamma) = 0 \tag{37}$$

and

$$\begin{bmatrix}\gamma \\ \omega\end{bmatrix} = r(\alpha) = \begin{bmatrix} r_\gamma(\alpha_\gamma) \\ \alpha_\omega \end{bmatrix} \tag{38}$$

where $r_\gamma(\alpha_\gamma)$ is a vector of functions with a dimension equal to that of γ, and α_γ and α_ω are parameter vectors. While this analysis allows for the existence of constraints on the covariance matrix Ω which are not part of the hypothesis, it assumes that there are no constraints relating the elements of Ω to those of γ. The matrix of first partials associated with $r(\alpha)$ in (38) is

$$R(\alpha) = \begin{pmatrix} R_\gamma(\alpha_\gamma) & 0 \\ 0 & I \end{pmatrix}$$

where $R_\gamma(\alpha_\gamma) \equiv \frac{\partial r_\gamma(\alpha_\gamma)'}{\partial \alpha_\gamma}$, and $I = \frac{\partial \omega'}{\partial \alpha_\omega}$ is the identity matrix. Substituting this matrix of partials into the expression for $\bar{A}$ given by (34) and using relations (14) and (15) to simplify the resulting expression, one obtains

$$\bar{A} = \begin{pmatrix} \bar{A}_{\gamma\gamma} & 0 \\ 0 & 0 \end{pmatrix}$$

with

$$(39) \qquad \bar{A}_{\gamma\gamma} = \bar{G}^{-1}_{\gamma\gamma}(\theta_0) - R_\gamma(\alpha_{\gamma 0})[R_\gamma(\alpha_{\gamma 0})G_{\gamma\gamma}(\theta_0)R_\gamma(\alpha_{\gamma 0})']^{-1}R_\gamma(\alpha_{\gamma 0})',$$

where $\alpha_{\gamma 0}$ is the true value of α_γ. In this case the matrix $\bar{G}(\theta_0)\bar{A}$ is idemponent with a trace equal to s (i.e., the number of restrictions imposed in (37)) which implies that the statistics LRT_0 and LMT_0 converge to a chi-squared distribution with s degrees of freedom. Thus, when testing hypotheses concerning regression coefficients the statistics LRT and LMT both possess a large sample χ^2_s distribution under the null hypothesis even when disturbances are not normally distributed.

An interesting implication of this finding concerns the applicability of the likelihood ratio statistic for testing hypotheses involving structural coefficients of a simultaneous equations model. To be concrete, consider the system of simultaneous equations given by (18) where the matrices Γ and Π contain the structural coefficients of the model, and Φ is the covariance matrix associated with disturbances. As outlined in the previous section, both limited- and full-information maximum likelihood estimation of a simultaneous equation model is equivalent to applying the full-information estimation procedure discussed in this paper to estimate a reduced form model in the form of (2) with γ containing all the coefficients of Γ and Π, and with ω parameterized to include either the elements of Γ and Φ or the elements of the reduced form covariance matrix Ω. The particular parameterization chosen for ω is obviously irrelevant when computing the value of the likelihood function Q_N under various hypotheses. Considering this latter parameterization for ω, it is evident that constraints on the coefficients of Γ and Π imply constraints on the elements of γ, but do not imply constraints on the elements of ω assuming the covariance matrix Φ is unrestricted. In the absence of covariance restrictions, then, all hypotheses concerning the form of the matrices Γ and Π translate into restrictions on γ which may be written as (37) and (38). Thus, it follows immediately from the above results that the LRT statistic for testing restrictions on these structural coefficients is distributed in large samples according to the χ^2 distribution usually associated with this statistic as long as there are no constraints on the covariance matrix Φ. One can in effect use this LRT statistic to test hypotheses in a simultaneous equation framework as if disturbances are normally distributed.

The asymptotic properties of the test statistics LRT and LMT are fundamentally changed when used to test hypotheses concerning covariance parameters in the absence of normality. The large sample distrubiton of these statistics for this case can easily be determined using theoretical results from the statistical literature on the distribution of quadratic forms of normal variants. The distribution of the quantity $\eta' \wedge \eta$ is known when $\wedge$ is a symmetric positive semi-definite matrix, which is the case of concern for determining the distribution of LRT_0 and LMT_0. Defining $\lambda_1,\ldots,\lambda_k$ as the positive characteristics roots of the matrix $\wedge\Sigma$, the quadratic form $\eta' \wedge \eta$ with $\eta \sim N(0, \Sigma)$ can be shown to have a distribution with a characteristic function given by

$$\prod_{j=1}^{k} \left(1 - 2it\lambda_j\right)^{-1/2}.$$

For notational convenience we denote a distribution with this characteristic function as $QFD(\lambda_1,\ldots,\lambda_k)$. A comprehensive discussion of the properties of this distribution can be found in Johnson-Kotz (1970, Ch. 29).[1] While the cumulative distribution function associated with $QFD(\lambda_1,\ldots,\lambda_k)$ has a closed-form solution, it involves an infinite series which makes it difficult to use for computing critical values needed in hypothesis testing without resorting to a computer. Difficulties arise in tabulating this function because of the number of parameters involved. Each of the k roots $\lambda_1,\ldots,\lambda_k$ constitutes a parameter which may take any positive value. Johnson-Kotz (1970, Ch. 29) list several sources with tables for selected values of the cumulative function for $k \leq 5$, and they present formulas for approximating this function for an aribitrary number of roots. In an article, Johnson-Kotz

[1] Kendall-Stuart (1969, Ch. 15) also presents a discussion of this distribution.

(1969) provide a computer program for calculating the cumulative function in the general case.

Inspection of the formulas for LRT_0 and LMT_0 given by (35) and (36) reveals that both of these statistics converge in distribution to a $QFD(\lambda_1,\ldots,\lambda_s)$ distribution where $\lambda_1,\ldots,\lambda_s$ are the s positive characteristic roots of the matrix $\bar{A}\,\bar{G}(\theta_0)$. In those cases discussed above where $\bar{A}\,\bar{G}(\theta_0)$ is idemponent, all the λ_j's, $j = 1,\ldots,s$, are equal to one and, accordingly, LRT_0 and LMT_0 converge to a $QFD(1,\ldots,1)$ distribution which is, of course, identical to a χ^2_s distribution. When testing any sort of covariance restrictions, however, these roots may take values other than one, and a method is required for computing these roots. Referring to the definition of the matrix $\bar{A}$ given by (33), it is evident that calculating the s positive characteristics roots of the matrix

$$[H^{-1}(\hat{\theta}) - R(\hat{\alpha})'[R(\hat{\alpha})H(\hat{\theta})R(\hat{\alpha})']^{-1}R(\hat{\alpha})]G(\hat{\theta})$$

yields consistent estimates of the parameters $\lambda_1,\ldots,\lambda_s$. Using these estimates to evaluate the cumulative distribution function associated with a $QFD(\lambda_1,\ldots,\lambda_s)$ variate, one can compute critical values that may be compared with the realized values of the statistics LRT and LMT to test hypotheses concerning covariance restrictions, including constraints relating covariance parameters and regression coefficients.

Tests Based on Limited-Information Estimation

Three methods for estimating subsets of parameters are discussed in the previous section: one is the traditional generalized least squares

procedure; another conditions on consistent estimates of regression coefficients and provides a technique for estimating covariance parameters; and the third procedure amounts to a least squares method of estimation. Below we consider the asymptotic properites of test statistics analogous to those presented above for each of these estimation methods.

The generalized least squares procedure uses the distance function $Q_N^*(\gamma)$ defined by (23) to obtain estimates of the regression coefficients. To test hypotheses concerning the structure of γ which may be written in the form of equations (37) and (38), the statistics corresponding to the Wald test, the likelihood ratio test, and the Lagrange multiplier test are:

$$W^* = N\, b_\gamma(\gamma^*)' [B_\gamma(\gamma^*)V(\gamma^*)B_\gamma(\gamma^*)']^{-1} b_\gamma(\gamma^*), \tag{40}$$

$$LRT^* = 2N(Q_N^*(\gamma^*) - Q_N^*(\gamma_c^*)), \tag{41}$$

and

$$LMT^* = N \left.\frac{\partial Q_N^*}{\partial \gamma'}\right|_{\gamma_c^*} \left[- \left.\frac{\partial^2 Q_N^*}{\partial \gamma \partial \gamma'}\right|_{\gamma_c^*} \right]^{-1} \left.\frac{\partial Q_N^*}{\partial \gamma}\right|_{\gamma_c^*}, \tag{42}$$

where γ^* and γ_c^* are the unconstrained and the constrained estimates of γ_0 derived by maximizing the function $Q_N^*(\gamma)$, $B_\gamma(\gamma^*) \equiv \left.\frac{\partial b_\gamma}{\partial \gamma'}\right|_{\gamma^*}$, and $V(\gamma^*) = H^{-1}(\gamma^*, \omega^*)$ is the approximate asymptotic covariance matrix associated with the quantity $\sqrt{N}(\gamma^* - \gamma_0)$.

Following the above derivations of the quadratic forms W_0, LRT_0, and LMT_0, one can derive similar relationships for W*, LRT*, and LMT*.

Obtaining an expression analogous to (20) for γ^* and using relation (14), it can be shown that W* is asymptotically equivalent to

$$W_0^* = (\sqrt{N}\, \ell_\gamma(\theta_0))' \bar{K}_{\gamma\gamma}(\sqrt{N}\, \ell_\gamma(\theta_0))$$

where $\bar{K}_{\gamma\gamma}$ has the same expression as $\bar{K}$ given by (32) except $H_{\gamma\gamma}(\theta_0)$ and $B_\gamma(\theta_0)$ replace the matrices $H(\theta_0)$ and $B(\theta_0)$, respectively; and LRT* and LMT* are asymptotically equivalent to

$$LRT_0^* = LMT_0^* = (\sqrt{N}\, \ell_\gamma(\theta_0))' \bar{A}_{\gamma\gamma}(\sqrt{N}\, \ell_\gamma(\theta_0))$$

where $\bar{A}_{\gamma\gamma}$ is given by (39). When computing the likelihood ratio statistic, one is not required to evaluate Q_N at the same value of ω to derive the above equivalence relationship. In particular, if instead of the expression for LRT* given by (41), one defines

$$LRT^* = 2N[Q_N(\gamma^*, \omega_1^*) - Q_N(\gamma_c^*, \omega_2^*)]$$

where ω_1^* and ω_2^* are both consistent estimates of ω_0 under the null hypothesis, then LRT* is still asymptotically equivalent to LRT_0^*.

We know that $\text{dlim}(\sqrt{N}\, \ell_\gamma(\theta_0)) = N(0, \bar{G}_{\gamma\gamma}(\theta_0))$, and using (15), one can readily verify that the matrices $\bar{K}_{\gamma\gamma}\bar{G}_{\gamma\gamma}(\theta_0)$ and $\bar{A}_{\gamma\gamma}\bar{G}_{\gamma\gamma}(\theta_0)$ are both idemponent with traces equal to the number of restrictions. Hence, the statistics W*, LRT*, and LMT* are clearly each distributed according to a χ^2_s variate in large samples under the null hypothesis. In the context of a generalized least squares procedure, then, we find once again that asymptotic tests may be carried out as if disturbances are normally distributed, and furthermore, one can neglect the fact that the covariance matrix is an estimated quantity rather than equal to the true matrix.

The limited-information procedure for estimating covariance parameters offers a framework for testing hypotheses concerning the structure of ω which may be written as $b_\omega(\omega) = 0$ or as $\omega = r_\omega(\alpha_\omega)$. This procedure uses the distance function $\tilde{Q}_N(\omega)$ to obtain estimates for ω and suggests the following three statistics:

$$\tilde{W} = N\, b_\omega(\tilde{\omega})'[B_\omega(\tilde{\omega})V(\tilde{\omega})B_\omega(\tilde{\omega})']^{-1} b_\omega(\tilde{\omega}), \tag{43}$$

$$\tilde{LRT} = 2N[\tilde{Q}_N(\tilde{\omega}) - \tilde{Q}_N(\tilde{\omega}_c)], \tag{44}$$

and

$$\tilde{LMT} = N \frac{\partial \tilde{Q}_N}{\partial \omega'}\Big|_{\tilde{\omega}_c} \left[- \frac{\partial^2 \tilde{Q}_N}{\partial \omega \partial \omega'}\Big|_{\tilde{\omega}_c} \right]^{-1} \frac{\partial \tilde{Q}_N}{\partial \omega}\Big|_{\tilde{\omega}_c} \tag{45}$$

where $\tilde{\omega}$ and $\tilde{\omega}_c$ are the unconstrained and the constrained estimates of ω_0 derived by maximizing the function $\tilde{Q}_N(\omega)$, $B_\omega(\tilde{\omega}) \equiv \frac{\partial b_\omega}{\partial \omega'}\Big|_{\tilde{\omega}}$, and $V(\tilde{\omega}) = H^{-1}_{\omega\omega}(\tilde{\gamma}, \tilde{\omega})G^{-1}_{\omega\omega}(\tilde{\gamma}, \tilde{\omega})H^{-1}_{\omega\omega}(\tilde{\gamma}, \tilde{\omega})$ is the approximate asymptotic covariance matrix associated with $\sqrt{N}(\tilde{\omega} - \omega_0)$.

Under the null hypothesis, $\tilde{W}$ obviously possesses a large sample χ^2_s where s is the number of restrictions being tested. Furthermore, it follows directly from the above analysis that $\tilde{LRT}$ and $\tilde{LMT}$ are approximately distributed as a $QFD(\lambda_1,\ldots,\lambda_s)$ variate, where the parameters $\lambda_1,\ldots,\lambda_s$ may be consistently estimated by computing the s positive characteristic roots of the matrix

$$[H^{-1}_{\omega\omega}(\tilde{\gamma}, \tilde{\omega}) - R_\omega(\tilde{\alpha}_\omega)'[R_\omega(\tilde{\alpha}_\omega)H_{\omega\omega}(\tilde{\gamma}, \tilde{\omega})R_\omega(\tilde{\alpha}_\omega)']^{-1}R_\omega(\tilde{\alpha}_\omega)]G_{\omega\omega}(\tilde{\gamma}, \tilde{\omega}),$$

with $R_\omega(\tilde{\alpha}_\omega) = \frac{\partial r'_\omega}{\partial \alpha_\omega}\Big|_{\tilde{\alpha}_\omega}$, and $\tilde{\alpha}_\omega$ determines the constrained estimate $\tilde{\omega}_c = r_\omega(\tilde{\alpha}_\omega)$.

Finally, to test hypotheses concerning the regression coefficients in the presence of heteroscedasticity of an unknown form, one can use the least squares procedure discussed above which employs the function $M_N(\gamma)$ given by (25). Statistics corresponding to the Wald test, the likelihood ratio test, and the Lagrange multiplier test are:

$$\hat{W}^* = N\, b_\gamma(\hat{\gamma}^*)'[B_\gamma(\hat{\gamma}^*)V(\hat{\gamma}^*)B_\gamma(\hat{\gamma}^*)']^{-1}b_\gamma(\hat{\gamma}^*), \tag{46}$$

$$\hat{LRT}^* = 2N[M_N(\hat{\gamma}^*) - M_N(\hat{\gamma}_c^*)], \tag{47}$$

and

$$\hat{LMT}^* = N\,\frac{\partial M_N}{\partial \gamma'}\Big|_{\hat{\gamma}_c^*}\left[-\frac{\partial^2 M_N}{\partial\gamma\partial\gamma'}\Big|_{\hat{\gamma}_c^*}\right]^{-1}\frac{\partial M_N}{\partial\gamma}\Big|_{\hat{\gamma}_c^*}, \tag{48}$$

where $\hat{\gamma}^*$ and $\hat{\gamma}_c^*$ are the unconstrained and the constrained values of γ_0 that maximize $M_N(\gamma)$, and

$$V(\hat{\gamma}^*) = \left[-\frac{\partial^2 M_N}{\partial\gamma\partial\gamma'}\Big|_{\hat{\gamma}^*}\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial m_i}{\partial\gamma}\Big|_{\hat{\gamma}^*}\frac{\partial m_i}{\partial\gamma'}\Big|_{\hat{\gamma}^*}\right]\left[-\frac{\partial^2 M_N}{\partial\gamma\partial\gamma'}\Big|_{\hat{\gamma}^*}\right]^{-1}$$

is a consistent estimate for the asymptotic covariance matrix of $\sqrt{N}(\hat{\gamma}^* - \gamma_0)$.

$\hat{W}^*$ is distributed according to a χ^2_s distribution in large samples under the null hypothesis. The above results further imply that the statistics $\hat{LRT}^*$ and $\hat{LMT}^*$ both converge to a $QFD(\lambda_1,\ldots,\lambda_s)$ distribution, where the s positive characteristics roots of the matrix

$$\left[\left[-\frac{\partial M_N}{\partial\gamma\partial\gamma'}\Big|_{\hat{\gamma}^*}\right]^{-1} - R_\gamma(\hat{\alpha}_\gamma^*)'\left[R_\gamma(\hat{\alpha}^*)\left(-\frac{\partial M_N}{\partial\gamma\partial\gamma'}\Big|_{\hat{\gamma}^*}\right)R_\gamma(\hat{\alpha}_\gamma^*)'\right]R_\gamma(\hat{\alpha}_\gamma^*)\right]\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial m_i}{\partial\gamma}\Big|_{\hat{\gamma}^*}\frac{\partial m_i}{\partial\gamma'}\Big|_{\hat{\gamma}^*}\right]$$

provide consistent estimates of the parameters $\lambda_1,\ldots,\lambda_s$.

Conclusion

We have seen that applying maximum likelihood techniques which assume normality to estimate the parameters of a nonlinear multivariate regression model with covariance restrictions, including general forms of heteroscedasticity, yields parameter estimates that are both consistent and asymptotically normally distributed under fairly general conditions. In addition to considering full-information methods where one simultaneously estimates all parameters, three limited-information procedures are examined which provide for the separate estimation of regression coefficients and covariance parameters. One procedure treats consistent estimates of regression coefficients as true parameter values and offers a simple method for estimating covariance parameters; and a second procedure does just the opposite. The estimators obtained from both procedures are shown to have the same asymptotic distribution as the full-information estimator. The third limited-information procedure offers a robust technique for estimating regression coefficients in the presence of arbitrary and unknown forms of heteroscedasticity. The general asymptotic properties of several statistics used to test composite hypotheses in a large sample setting are derived for both full- and limited-information methods of estimation.

An important finding of this paper concerns the robustness of results for regression coefficients. The application of standard maximum likelihood techniques not only yields estimates for these coefficients that are consistent and normally distributed in large samples, these techniques also report standard errors for these estimates that are asymptotically valid regardless of whether or not disturbances actually follow a joint normal distribution. Furthermore, the normality assumption is not needed to prove

that the likelihood ratio and the Lagrange multiplier statistics for testing restrictions on regression coefficients are distributed according to the large sample chi-squared distribution usually associated with these statistics. An interesting by-product of this result concerns the implication that one can use likelihood ratio statistics to test restrictions on structural coefficients of simultaneous equation models as if the normality assumption were satisfied as long as there are no covariance restrictions. This large sample robustness of standard errors and the test statistics for regression coefficients holds when applying either full-information estimation methods or standard nonlinear generalized least squares procedures.

In contrast to the above finding, the standard errors for estimates of covariance parameters reported by a maximum likelihood routine which assumes joint normality are asymptotically invalid if in fact the normality assumption is violated. Computing the appropriate standard errors in the general case requires the use of the matrix of outer partials as well as the inverse of the matrix of second partials. The above analysis also explicitly derives the general asymptotic distributions of the likelihood ratio and the Lagrange multiplier statistics associated with testing the covariance restrictions, including restrictions relating covariance parameters to regression coefficients. These asymptotic distributions are shown to depend crucially on the normality assumption and will in general diverge from the familiar chi-squared distribution.

References

Amemiya, T. "Regression Analysis when the Dependent Variable is Truncated Normal." Econometrica 41 (1973): 997-1016.

Anderson, T. W. and H. Rubin. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." Annals of Mathematical Statistics 20, No. 1 (March 1949).

Chung, K. L. A Course in Probability Theory, 2nd Ed. New York: Academic Press, 1974.

Jenrich, R. I. "Asymptotic Properties of Non-linear Least Squares Estimators." Annals of Mathematical Statistics 40 (1969): 633-643.

Johnson, N. L. and S. Kotz. "Tables of Distributions of Quadratic Forms in Central Normal Variables, I and II." Institute of Statistics, Mimeo Series No. 543 and No. 557, University of North Carolina at Chapel Hill, (also Sankhya, Series B, 30: 303-14).

_____. Distribution in Statistics: Continuous Univariate Distributions 2. New York: John Wiley & Sons, 1970.

Kendall, M. and A. Stuart. The Advanced Theory of Statistics, Vol. 1 Distribution Theory. New York: Macmillan Publishing Co., 1969.

Loeve, M. Probability Theory, I, 4th Ed. Princeton, N.J.: Van Nostrand Company, Inc., 1977.

Rao, C. R. "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation." Proceedings of Cambridge Philosophical Society 44 (1947): 50-57.

_____. Linear Statistical Inference and its Applications, 2nd Ed. New York: John Wiley & Sons, Inc., 1973.

Silvey, S. D. "The Lagrangian Multiplier Test." Annals of Mathematical Statistics 30 (1959): 389-407.

White, H. "Nonlinear Regression on Cross-Section Data." Econometrica 48 (1980): 721-46.