NBER TECHNICAL PAPER SERIES

USING INFORMATION ON THE MOMENTS OF DISTURBANCES TO INCREASE THE EFFICIENCY OF ESTIMATION

Thomas E. MaCurdy

Technical Paper No. 22

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge MA 02138

May 1982

This work was support by NSF grant No. SES-8023043. Special thanks are given to Takeshi Amemiya for many valuable discussions on the contents of this paper. 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 #22
May 1982

Using the Information on the Moments of Disturbances To Increase the Efficiency of Estimation

Abstract

This study considers the estimation of both regression and simultaneous equations that may involve nonlinearities in parameters and variables. For a wide range of assumptions concerning the distributional properties of disturbances, this analysis develops new estimators whose efficiencies dominate those associated with the estimators obtained by familiar least squares and instrumental variable procedures. Essentially the only time one cannot improve the efficiency of estimation using the methods proposed in this paper corresponds to those special situations in which the familiar procedures yield estimators that are asymptotically efficient.

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

Introduction

This study develops new estimators for regression and simultaneous equations that are in general strictly efficient relative to the conventional least squares and two-stage least squares estimators. The analysis encompasses both linear and nonlinear estimation in situations where random variables are distributed independently across observations. The discussion deals with many alternative assumptions concerning the specific distributional properties of the disturbances associated with the equation under consideration, ranging from a situation in which the variances of disturbances vary across observations in an unknown and arbitrary fashion, to one in which errors are identically distributed over the sample. For each of these distributional assumptions, estimation procedures are formulated that yield a gain in efficiency over the familiar ordinary, weighted, and two-stage least squares procedures except when these latter procedures correspond directly to the application of maximum likelihood methods. Thus, unless the exact distribution of random variables is known, one can improve the efficiency of estimation using the procedures described here.

The properties of the second and higher order moments of the disturbances constitutes the source of information used to improve the efficiency of estimation in this analysis. Estimators are formulated by jointly estimating the regression or simultaneous equation under consideration with various transformations of this equation which are obtained by weighting schemes and/or by raising both sides of an equation to a power (e.g., squaring or cubing both sides). The particular transformations included in the joint estimation procedure are chosen

in a way to exploit the distributional properties of the disturbances for the sample under investigation. As a by-product of this work, general methods are developed for testing the hypotheses that any particular moment or combination of moments of the disturbances are constant across observations for both the cases of regression and simultaneous equations. This analysis also provides a natural framework for discussing recent important contributions by Chamberlain (1982), related work by Amemiya (1982), and results reported in White (1982).

Section I presents notation and results used throughout the paper. Section II discusses estimating the parameters of a linear multiple regression model, and Section III considers the general problem of estimating the parameters of a nonlinear simultaneous equation.

I. Combining Information to Compute Estimators

This section reviews a familiar procedure for computing parameter estimates that optimally combines information and constraints from various sources. The purpose of this discussion is to establish the notation and derive formulas used in the subsequent analysis.

Suppose one is interested in obtaining a consistent estimate of the "true" value of a p x 1 parameter vector θ that is an unknown determinant of the distribution generating a random vector Y. Denote this true value as θ_o which is assumed to be an interior point in a compact set Θ, and let Y_i and Z_i, i = 1,...,N, denote N observations on Y and on a vector of measured characteristics Z. The Y_i's are assumed to be independently distributed across observations after conditioning on the Z_i's, or when these characteristics are treated as known constants.

The most widely used method for computing a consistent estimate for θ_o is to solve a system of equations that implicitly defines a value for θ and is known to be satisfied when $\theta = \theta_o$ as the sample size goes to infinity. In particular, let $\ell_i(\theta) \equiv \ell(\theta, Y_i, Z_i)$, i = 1,...,N, represent a r x 1 vector of known functions with $r \geq p$, and consider the system of equations

$$(1) \qquad L_N(\theta) \equiv \frac{1}{N} \sum_{i=1}^{N} \ell_i(\theta) = 0.$$

Assuming each ℓ_i possesses a sufficiently well-behaved distribution and is chosen so that $E(\ell_i(\theta_o)) = 0$, one can show that setting $\theta = \theta_o$ solves (1) in

the sense that $L_N(\theta_o)$ converges in probability to zero as the sample size goes to infinity. For most estimation procedures, the ℓ_i's are gradient vectors associated with the optimization of a particular function of the data defined over the parameter space Θ. By introducing additional assumptions which guaranteee the satisfaction of a set of regularity conditions, one may further demonstrate that computing a solution to (1) yields a strongly consistent estimate for θ_o that is asymptotically normally distributed.

Some notation and assumptions are needed for the following discussion. The matrices of first partials $\frac{\partial \ell_i}{\partial \theta'}$, $i = 1,\ldots,N$, are assumed to exist with each element uniformly continuous in θ. Denote the average of these partials by $S_N \equiv \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_i}{\partial \theta'}$, assumed to possess full column rank. Define the matrix $V_N(\theta) \equiv \frac{1}{N} \sum_{i=1}^{N} \ell_i(\theta)\ell_i'(\theta)$ as an average of outer products; and further define $L(\theta) \equiv \lim E(L_N(\theta))$, $S(\theta) \equiv \lim E(S_N(\theta))$, and $V(\theta) \equiv \lim E(V_N(\theta))$ with limits computed as $N \to \infty$. To derive the asymptotic results cited below, the distributions associated with the ℓ_i's and the matrices of first partials cannot have too much weight in the tails.[1] Given such distributional assumptions and the independence of the ℓ_i's, one may show the following strong convergence results: $L_N(\theta) \xrightarrow{s} L(\theta)$; $S_N \xrightarrow{s} S(\theta)$; and $V_N(\theta) \xrightarrow{s} V(\theta)$, where $\xrightarrow{s}$ designates almost sure convergence with these results holding for each $\theta \in \Theta$. Thus, if $E(\ell_N(\theta_o)) = 0$, it follows that $L_N(\theta_o) \xrightarrow{s} 0$ implying θ_o solves (1) as the sample size goes to infinity. Furthermore, one may show that $\sqrt{N}\, L_N(\theta_o) \xrightarrow{d} N(0, V(\theta_o))$, where $\xrightarrow{d}$ denotes convergence in distribution, and $N(\cdot,\cdot)$ signifies a normal probability

law. In many applications, one cannot rule out the possibility that values of θ other than θ_o may also satisfy (1) in the limit. One can, however, easily resolve this issue for the estimation problems considered below, and for simplicity this analysis assumes the solution to (1) is unique. To prove consistency and asymptotic normality of this solution, the convergence of L_N, S_N, and V_N to their respective limits must be uniform in θ, and this assumption is maintained throughout the discussion.

When the number of equations in (1) used to compute estimates exceeds the number of parameters, one requires a weighting scheme for comparing the errors obtained in solving the various equations; one cannot typically find a value for θ that solves all equations exactly in finite samples. This is, of course, a familiar problem in statistics. The theory of ordinary least squares derives an estimate for θ by minimizing the sum of squared errors associated with the r equations appearing in (1), which implies minimization of the quantity $L_N'(\theta)L_N(\theta)$. It is well known, however, that applying generalized least squares yields a more efficient parameter estimate. According to this procedure, one solves equation (1) by choosing θ to minimize the quantity $L_N'(\theta)[E(V_N(\theta_o))]^{-1} L_N(\theta)$, where this expression uses the relation $E(L_N(\theta_o)L_N'(\theta_o)) = \frac{1}{N} E(V_N(\theta_o))$ following from the independence of observations. The matrix $E(V_N(\theta_o))$ is unknown, but, as with many generalized least squares analyses, a consistent estimate for this matrix is easily constructed and the asymptotic properties of estimators are unaffected if one substitutes this consistent estimate for the true value of the matrix. Accordingly, when computing an estimate for θ_o, one sacrifices no estimation efficiency by instead minimizing the quadratic form

$$L_N'(\theta)\ \hat{V}_N^{-1}\ L_N(\theta), \tag{2}$$

where $\hat{V}_N \equiv V_N(\hat{\theta})$ with $\hat{\theta}$ representing any strongly consistent estimate for θ_o implying $\hat{V}_N \xrightarrow{s} V(\theta_o)$. Essentially any solution to (1) or to any subset of this system of equations may serve as $\hat{\theta}$. Let $\tilde{\theta}$ denote that value of θ minimizing (2).

Identifying the asymptotic properties of the estimator $\tilde{\theta}$ is a straightforward task once one recognizes that (2) is in the form of the quantities minimized to compute conventional nonlinear two- and three-stage least squares estimators. Following the work of Amemiya (1974, 1977), who initiated the study of this class of estimators, one can readily verify that $\tilde{\theta} \xrightarrow{s} \theta_o$ and

$$(3) \qquad \sqrt{N}\,(\hat{\theta} - \theta_o) \xrightarrow{d} N(0,\ [S'(\theta_o)V^{-1}(\theta_o)S(\theta_o)]^{-1}).$$

Thus, the approximate distribution for $\tilde{\theta}$ in large samples is

$$\tilde{\theta} \overset{\cdot}{\sim} N(\theta_o, \tfrac{1}{N}[\tilde{S}_N'\ \hat{V}_N^{-1}\ \tilde{S}_N]^{-1}).$$

where $\tilde{S}_N \equiv S_N(\tilde{\theta})$. Recent work by Hansen (1981) also examines this method for computing estimators and derives (3) using an alternative set of assumptions.

The interesting problem in estimation concerns the choice of the equations included in system (1). When the distributions of random variables are known, setting L_N equal to the gradient associated with the log of the likelihood function is well known to produce an estimator that is asymptotically efficient. In this instance, adding more equations to (1) obviously provides for no gain in estimation efficiency, and changing any one of the equations leads, in general, to a loss in efficiency. Without knowledge of distributions, however, the choice of equations used to compute estimators is an open question.

The subsequent analysis examines the strategy of adding equations to those typically used in the conventional application of least squares or instrumental variable estimation methods as a way of deriving more efficient estimators.

To determine the gain in efficiency as one introduces more information or equations to compute estimators, suppose one has two sets of equations available for estimating a parameter vector α: one system of equations associated with the i^{th} observation consists of the vector of functions $g_i(\alpha, \mu)$ which also depends on unknown parameters μ; and a second system consists of the vector of functions $h_i(\alpha, \psi)$ depending on a third set of parameters ψ. The expectations of both g_i and h_i are assumed to vanish at the true parameter values for each i, making it possible to use the averages of either vector to compute a consistent estimate for α_o. Combining these equations to obtain a single estimate for α_0, one sets $\theta' = (\alpha', \mu', \psi')$ and $\ell_i' = (g_i', h_i')$, and minimizes (2) with respect to θ. According to (3), the covariance matrix associated with the resulting estimate is $[S' V^{-1} S]^{-1}$, where the argument θ_o indicating the point of evaluation is dropped for notational convenience. Consider the following partitions of the matrices S and V:

$$S \equiv \lim E\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ell_i}{\partial \theta'}\Big|_{\theta_o}\right] = \lim E\begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}\frac{\partial g_i}{\partial \alpha'}\Big|_{\theta_o} & \frac{1}{N}\sum_{i=1}^{N}\frac{\partial g_i}{\partial \mu'}\Big|_{\theta_o} & \frac{1}{N}\sum_{i=1}^{N}\frac{\partial g_i}{\partial \psi'}\Big|_{\theta_o} \\ \frac{1}{N}\sum_{i=1}^{N}\frac{\partial h_i}{\partial \alpha'}\Big|_{\theta_o} & \frac{1}{N}\sum_{i=1}^{N}\frac{\partial h_i}{\partial \mu'}\Big|_{\theta_o} & \frac{1}{N}\sum_{i=1}^{N}\frac{\partial h_i}{\partial \psi'}\Big|_{\theta_o} \end{bmatrix}$$

$$\equiv \begin{bmatrix} S_{11} & S_{12} & 0 \\ S_{21} & 0 & S_{23} \end{bmatrix};$$

and

$$V \equiv \lim E\frac{1}{N}\left[\sum_{i=1}^{N} \ell_i(\theta_o)\ell_i'(\theta_o)\right] = \lim E\begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} g_i(\theta_o)g_i'(\theta_o) & \frac{1}{N}\sum_{i=1}^{N} g_i(\theta_o)h_i'(\theta_o) \\ \frac{1}{N}\sum_{i=1}^{N} h_i(\theta_o)g_i'(\theta_o) & \frac{1}{N}\sum_{i=1}^{N} h_i(\theta_o)h_i'(\theta_o) \end{bmatrix}$$

$$\equiv \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}.$$

The matrix V is assumed to be nonsingular, which in essence requires each of the functions added to the analysis contained in h_i to constitute a unique piece of information. Defining the matrices $V_{22\cdot 1} \equiv V_{22} - V_{21}V_{11}^{-1}V_{12}$ and $F \equiv (S_{21}\ 0) - V_{21}V_{11}^{-1}(S_{11}\ S_{12}) \equiv (G\ \ F_2)$ with $G \equiv S_{21} - V_{21}V_{11}^{-1}S_{11}$ and $F_2 \equiv -V_{21}V_{11}^{-1}S_{12}$, use of the partion inverse formula for V^{-1} yields

$$[S'\ V^{-1}\ S]^{-1} = \left[\begin{pmatrix} S_{11}' & S_{21}' \\ S_{12}' & 0 \\ 0 & S_{23}' \end{pmatrix}\begin{pmatrix} I & -V_{11}^{-1}V_{12} \\ 0 & I \end{pmatrix}\begin{pmatrix} V_{11}^{-1} & 0 \\ 0 & V_{22\cdot 1}^{-1} \end{pmatrix}\begin{pmatrix} I & 0 \\ -V_{21}V_{11}^{-1} & I \end{pmatrix}\begin{pmatrix} S_{11} & S_{12} & 0 \\ S_{21} & 0 & S_{23} \end{pmatrix}\right]^{-1}$$

$$= \begin{bmatrix} \begin{pmatrix} S_{11}' \\ S_{12}' \end{pmatrix} V_{11}^{-1}(S_{11}S_{12}) + F'V_{22\cdot 1}^{-1}F & F'V_{22\cdot 1}^{-1}S_{23} \\ S_{23}'V_{22\cdot 1}^{-1}F & S_{23}'V_{22\cdot 1}^{-1}S_{23} \end{bmatrix}^{-1}.$$

Given this partition, the implied precision matrix (i.e., the inverse of the covariance matrix) associated with estimates for α and μ is

$$(4) \qquad \begin{bmatrix} S_{11}' \\ S_{12}' \end{bmatrix} V_{11}^{-1}(S_{11} \; S_{12}) + \begin{bmatrix} G' \\ F_2' \end{bmatrix} A(G \;\; F_2)$$

with $A \equiv V_{22\cdot1}^{-1} - V_{22\cdot1}^{-1} S_{23}(S_{23}' \; V_{22\cdot1}^{-1} \; S_{23})^{-1} \; S_{23}' \; V_{22\cdot1}^{-1}$. Since the first matrix in this expression is the precision matrix obtained if one uses only the vector of functions g_i to estimate the parameters α and μ, the second matrix shows the gain in precision achieved in estimating α and μ by also using the equations h_i in the computation of estimates. In particular, the matrix determining the efficiency gain in estimating the parameters of interest α is $G' \; A \; G$. Assuming the new parameters ψ appearing in the functions included in h_i are identified, the matrix S_{23} possesses full column rank; and assuming the inclusion of h_i in computing estimates adds more equations than new parameters into the analysis (i.e., the number of elements in the vector h_i exceeds the number in ψ), the matrix A is positive definite.[2] Thus, using the additional information conveyed in the h_i's to estimate α_0 leads to increased efficiency unless the matrix G is of rank zero, which essentially occurs only when

$$(5) \qquad G \equiv S_{21} - V_{21} \; V_{11}^{-1} \; S_{11} = 0.$$

It is clear from this analysis, then, that introducing more information to compute estimates never results in a loss of efficiency and achieves a gain in efficiency unless condition (5) is satisfied.

In a situation where the h_i's do not depend on a new set of parameters ψ, one can readily verify that the above formulas indicating efficiency gains are applicable if one simply sets $A = V_{22\cdot 1}^{-1}$. Thus, condition (5) once again determines whether there is an increase in the precision of estimation.

II. Least Squares Estimation

Suppose observations on the random variable Y are generated by the multiple repression model

$$(6) \qquad Y_i = X_i'\beta_0 + \varepsilon_i, \qquad i = 1,\ldots,N,$$

where X_i is a column vector of explanatory variables selected from a set of measured characteristics Z_i associated with the i^{th} observation, β_0 is an unknown parameter vector, and the ε_i's are independently distributed errors with $E(\varepsilon_i) = 0$ where expectations here implicitly condition on the Z_i's. The following analysis primarily considers two distinct assumptions relating to the distributional properties of the disturbances: first, the ε_i's are not identically distributed with variances $\sigma_i^2 \equiv E(\varepsilon_i^2)$ differing across observations; and, second, variance and higher order moments of the errors are constant across observations.

Using Exclusion Restrictions to Gain Efficiency

In the presence of heteroscedasticity, Chamberlain (1982) develops the first estimator of which I am aware that is efficient relative to the least squares estimator and whose computation does not rely on specific knowledge or modeling of the form taken by the heteroscedasticity. A discussion of this estimator offers a convenient opportunity for emphasizing several features of the estimation framework outlined in the previous section before proceeding to the alternative estimators proposed below. Amemiya (1982) offers

a more straightforward interpretation and derivation of Chamberlain's estimator, and it is Amemiya's simpler formulation of this estimator that is presented here.

Instead of directly estimating equation (7), Chamberlain suggests adding variables to this equation and forming a constrained estimator using a particular variant of the procedure outlined in Section I. In particular, letting Q_i denote a vector of measured variables linearly independent of X_i, consider the expanded model

$$(7) \qquad Y_i = X_i'\beta_0 + Q_i'\gamma_0 + \varepsilon_i, \qquad i = 1,\ldots,N,$$

where $\gamma_0 = 0$ according to (7). In terms of the notation of the previous section, with $Z_i' = (X_i', Q_i')$, one can view constrained least squares estimation of (7) with the restriction $\gamma = 0$ as choosing an estimate for β_0 by minimizing a quadratic form like (2) with $\ell_i(\beta) = Z_i(Y_i - X_i'\beta)$ and with the matrix $\hat{V}_N$ in (2) replaced by $M_N \equiv \frac{1}{N}\sum_{i=1}^{N} Z_i Z_i'$.
Chamberlain in effect uses the same expressions for the $\ell_i(\beta)$'s, but derives his estimator by directly minimizing (2) with no substitution for $\hat{V}_N$. For this problem, $\hat{V}_N = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\hat{\beta})\ell_i'(\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N} Z_i Z_i'\hat{\varepsilon}_i^2$ where $\hat{\varepsilon}_i$ represents a residual consistent for $Y_i - X_i'\beta_0$. For the class of estimators defined by minimizing a quadratic form in the vector L_N, the use of (2) and the matrix $\hat{V}_N$ produces the most efficient estimate for β_0 when constraints make it impossible to solve the equations $L_N = 0$ exactly. A standard application of asymptotic theory implies $\hat{V}_N \xrightarrow{s} \lim\left[\frac{1}{N}\sum_{i=1}^{N} Z_i Z_i'\sigma_i^2\right]$. When the variances of disturbances vary across observations, it is evident that the matrices $\hat{V}_N$ and M_N do not have asymptotic limits that are proportional to one another; and, as a consequence, the use of constrained least squares implies an efficiency loss.

Since unconstrained least squares estimation of (6) and constrained least squares estimation of (7) produces the same estimate for β_0, it follows that Chamberlain's estimator in general achieves greater efficiency than the conventional least squares estimator when heteroscedasticity exists.

Identifying the exact gain in efficiency is easily accomplished using the results of the previous section. Unconstrained least squares applied to equation (6) in effect computes an estimate for β_0 by minimizing the quadratic form given by (2) with the vector of functions $g_i(\beta) = X_i(Y_i - X_i'\beta)$ used in place of the ℓ_i's. One may, then, view Chamberlain's procedure as combining these g_i's with the information contained in the relations $h_i(\beta) = Q_i(Y_i - X_i'\beta)$ and minimizing (2) with $\ell_i' = (g_i', h_i')$. According to formula (4), the matrix G' A G shows the gain in precision achieved by including the h_i's in the computation of an estimate for β_0. For this problem, G is the asymptotic limit of the matrix

$$\left[-\frac{1}{N}\sum_{i=1}^{N} Q_i X_i'\right] - \left[\frac{1}{N}\sum_{i=1}^{N} Q_i X_i' \sigma_i^2\right]\left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \sigma_i^2\right]^{-1}\left[-\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right];$$

and A is the asymptotic limit of

$$\left[\left[\frac{1}{N}\sum_{i=1}^{N} Q_i Q_i' \sigma_i^2\right] - \left[\frac{1}{N}\sum_{i=1}^{N} Q_i X_i' \sigma_i^2\right]\left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \sigma_i^2\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} X_i Q_i' \sigma_i^2\right]\right]^{-1}.$$

Obviously, when the σ_i's are constant across observations, G = 0, and there is no gain in precision as expected.

A major deficiency with this approach concerns the absence of a strategy for choosing the variables to be included in Q; it is unclear

how to go about choosing an optimal Q. For this reason, the following estimation procedure offers an attractive alternative to the above approach.

Weighted Regressions

With heteroscedasticity present, the most familiar method for improving the efficiency of estimating β_0 is to consider a weighted version of equation (6) and apply least squares to this new equation. Let the variables ω_i, $i = 1,\ldots,N$, denote measured quantities depending on the characteristics Z and on a set of known or estimated coefficients. Using the ω_i's to weight (6) yields the equations

$$(8) \qquad \omega_i Y_i = \omega_i X_i'\beta_0 + \omega_i \varepsilon_i, \qquad i = 1,\ldots,N.$$

Obviously, if one knows how the σ_i's vary across observations and chooses the ω_i's to adjust for this variation, then least squares estimation of equation (8) rather than (6) generates the more efficient estimate for β_0. Without this knowledge, however, or given an arbitrary choice for the ω_i's, there is no a priori reason for preferring either the weighted or the ordinary least squares estimate on efficiency grounds.

Combining the information used to compute these distinct estimates offers an alternative procedure for constructing an estimate for β_0. In terms of the framework outlined in Section I, unconstrained least squares estimation of equation (6) computes an estimate for β_0 using the vector of functions $g_i(\beta) \equiv X_i(Y_i - X_i'\beta)$; and weighted least squares uses the vectors $h_i(\beta) \equiv \omega_i^2 X_i(Y_i - X_i'\beta)$.[3] To compute a single estimate for β_0 that combines these different relations, one sets $\ell_i' = (g_i', h_i')$ and minimizes (2). Satisfaction of

condition (5) determines whether adding the h_i's leads to a more efficient estimator. For this problem, the matrix G appearing in this condition is the asymptotic limit of

$$\left[-\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \omega_i^2\right] - \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \omega_i^2 \sigma_i^2\right]\left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \sigma_i^2\right]^{-1}\left[-\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right],$$

which does not in general equal the zero matrix unless $\sigma_i^2 = \sigma_0^2$ for all i. In the presence of heteroscedasticity, then, the new estimator produced by jointly estimating equations (6) and (9) is strictly more efficient than the ordinary least squares estimator. Furthermore, it is evident that this new estimator also possesses greater efficiency than the familiar weighted least squares estimator unless the weights are optimally chosen (i.e., $\omega_i = \frac{1}{\sigma_i}$).

In contrast to the previous approach based on the use of exclusion restrictions, this procedure for computing an estimator offers a natural strategy for choosing the weights and the additional relations to be combined with those used in conventional least squares estimation. While an optimal choice for the weights is assumed not to be an option in this analysis, one intuitively wants to choose each ω_i so that it approximates $\frac{1}{\sigma_i}$ as close as possible. One procedure for obtaining such weights involves a simple regression analysis. In particular, letting $b_i \equiv (Z_i,\lambda$ denote a function designed to capture the suspected variability of the σ_i's across observations, a regression of the squared least squares residuals $\hat{\varepsilon}_i^2$ on $b(\cdot)$ estimates the unknown parameters λ and provides for the construction of a fitted value b_i and a weight $\omega_i = 1/\sqrt{b_i}$.

To the extent that this choice for the weights fails to adjust for the variability of the σ_i's, one may construct an estimator for β_0 with improved efficiency by introducing yet another weighted variant of the regression equation. Joint least squares estimation of this new equation along with (6) and (9)

using the above procedure to combine the relations defining the individual least squares estimates for each equation provides the framework for computing this new estimator. A natural choice for this new equation is to consider a weighted version of (8) with the new weights chosen to reduce the variation in the variances $\omega_i^2\sigma_i^2$ associated with the disturbances of this equation.

Using Higher Order Moments

When disturbances are homoscedastic, there are other equations one may jointly estimate along with the regression equation to obtain an estimator more efficient than the one produced by ordinary least squares. This analysis works directly with equation (6) presuming that weights are not needed to induce homoscedasticity. If this presumption is false, then one must interpret equation (6) in this discussion as the appropriately weighted variant of the original regression equation. Squaring both sides of equation (6) yields

$$(9) \qquad Y_i^2 = \sigma_0^2 + (X_i'\beta_0)^2 + \nu_i, \qquad i = 1,\ldots,N,$$

where σ_0^2 is an intercept term, and $\nu_i \equiv 2(X_i'\beta_0)\varepsilon_i + \varepsilon_i^2 - \sigma_0^2$ is a disturbance with mean zero. Assuming the orthogonality conditions $\lim E(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'(\varepsilon_i^2 - \sigma_0^2)) = 0$, which obviously follows if $\sigma_i^2 = \sigma_0^2$ for all i, nonlinear constrained least squares applied to (9) offers an alternative method for computing a consistent estimate of β_0. Except for a sign convention, the components of β are in general identified using the information of equation (9) alone assuming X_i includes more than an intercept; fixing the sign of one nonzero component of β solves the sign convention problem. While least squares estimation of (9) yields a less efficient estimate for β_0 than the one obtained by least squares applied to (6), combining the relations used to compute these distinct estimates provides for the formulation

of a new estimator that is in general efficient relative to either of the estimators computed using a single equation alone.

Jointly estimating equations (6) and (9) without restricting the β coefficients to be equal for the two equations provides sufficient information to test: (i) whether the second moments of disturbances satisfy the orthgonality conditons needed to justify an equality restriction for the estimates of β across equations; and (ii) whether the variances of disturbances are constant over the sample. Least squares applied to equation (6) uses the vector of functions $g_i(\beta_1) = X_i(Y_i - X_i'\beta_1)$ to compute an estimate for β_0; and, with $v_i^* \equiv Y_i - \sigma^2 - (X_i'\beta_2)^2$, nonlinear least squares applied to (9) calculates estimates for β_0 and σ_0^2 using the relations

$$h_i(\beta_2, \sigma^2) = -\frac{1}{2}\frac{\partial v_i^{*2}}{\partial \begin{pmatrix}\beta \\ \sigma^2\end{pmatrix}} = \begin{pmatrix} 2X_i(X_i'\beta_2) \\ \\ 1 \end{pmatrix}(Y_i^2 - \sigma^2 - (X_i'\beta_2)^2).$$

Joint least squares estimation of these equations means in this analysis that one sets $\ell_i'(\beta_1, \beta_2, \sigma^2) = (g_i', h_i')$ and minimizes (2) with respect to β_1, β_2, and σ^2. Using the implied joint asymptotic distribution for the estimates $\tilde{\beta}_1$ and $\tilde{\beta}_2$ given by (3), one may use a standard Wald statistic to test the null hypothesis that the probability limit of $\tilde{\beta}_1 - \tilde{\beta}_2$ is zero. Acceptance of this hypothesis indicates that the moments $E(\varepsilon_i^2)$ satisfy the orthogonality conditions needed for $\tilde{\beta}_2 \xrightarrow{s} \beta_0$. However, more is required to test the assumption of homoscedasticity because the restriction $\beta_1 = \beta_2$ does not rule out the possibility that the variances σ_i^2 may depend on a set of measured variables that are orthogonal to X_iX_i'. Including these other variables in equation (9) permits one to estimate their relationship to the σ_i^2's. A test for homoscedasticity, then, involves a null hypothesis with two parts: the quantity $\tilde{\beta}_1 - \tilde{\beta}_2$ converges in probability to zero; and other variables assumed to determine the σ_i^2's have zero coefficients when entered into equation (9).[4]

Rejection of the restriction $\beta_1 = \beta_2$ means that one cannot use information from equation (9) to attain an efficiency gain for the estimate of β_0. At the same time, however, rejection of this hypothesis ensures that the joint estimation scheme described above combining the regression equation and its weighted variants yields an improvement in efficiency over conventional least squares estimation.

Acceptance of the restriction $\beta_1 = \beta_2$ indicates that one can carry out joint estimation of equations (6) and (9) imposing this equality constraint. In particular, one sets $\ell_i'(\beta,\sigma^2) = (g_i'(\beta), h_i'(\beta, \sigma^2))$ using the specifications of g_i and h_i cited directly above and minimizes quadratic form (2) with respect to β and σ^2. Since ordinary least squares only uses the g_i relations to compute an estimate for β_0, forming the matrix G identifies the conditions under which this joint estimation procedure leads to an efficiency gain over the usual least squares procedure. The implied specification for G is the asymptotic limit of

$$\left[-\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}4(X_i'\beta_0)^2X_iX_i'\\ \\ 2(X_i'\beta_0)X_i'\end{pmatrix}\right] - \left[\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}2(X_i'\beta_0)X_iX_i'\\ \\ X_i'\end{pmatrix}E(\varepsilon_i\nu_i)\right]\left[\frac{1}{N}\sum_{i=1}^{N}X_iX_i'E(\varepsilon_i^2)\right]^{-1}\left[-\frac{1}{N}\sum_{i=1}^{N}X_iX_i'\right].$$

Given the definition of ν_i, $E(\varepsilon_i\nu_i) = 2(X_i'\beta_0)E(\varepsilon_i^2) + E(\varepsilon_i^3)$. In the presence of homoscedasticity with $E(\varepsilon_i^2) = \sigma_0$ for all i, $G \neq 0$ if $E(\varepsilon_i^3) \neq 0$; so, one achieves an efficiency gain if the third moments of disturbances are nonzero. In those rare situations where the disturbances ε_i are heteroscedastic and still satisfy the orthogonality conditions needed to estimate β_0 consistently using equation (9), $G \neq 0$ even if $E(\varepsilon_i^3) = 0$, implying that joint estimation in these instances virtually always leads to a gain in efficiency over conventional least squares.

Accounting for the heteroscedasticity associated with the disturbances of the equations used in this joint estimation procedure provides for the formulation of an even more efficient estimator β_0. The covariance matrix associated with the errors of equations (6) and (9) for the i^{th} observation is

$$\Phi_i \equiv E\left[\begin{pmatrix}\varepsilon_i\\ \nu_i\end{pmatrix}(\varepsilon_i \nu_i)\right] = \begin{bmatrix} E(\varepsilon_i^2) & 2(X_i'\beta_0)E(\varepsilon_i^2) + E(\varepsilon_i^3) \\ 2(X_i'\beta_0)E(\varepsilon_i^2) + E(\varepsilon_i^3) & 4(X_i'\beta_0)^2E(\varepsilon_i^2) + 2(X_i'\beta_0)E(\varepsilon_i^3) + E(\varepsilon_i^4)\end{bmatrix}.$$

The disturbances of the squared regression equation have nonconstant variances by construction. In addition, there exists several other potential sources of heteroscedasticity: the moments $E(\varepsilon_i^3)$ and $E(\varepsilon_i^4)$ may vary across observations; and, as noted above, the ability to use the information from equation (9) in computing an estimate for β_0 does not rule out the possibility that the second moments $E(\varepsilon_i^2)$ are nonconstant. One method of improving the efficiency of estimation by introducing adjustments for heteroscedasticity involves the implementation of an approach like the one followed above; that is, in addition to equations (6) and (9), one may also use information from weighted variants of these equations in the computation of an estimate for β_0.

If one is willing to assume that the disturbances ε_i have constant moments up to fourth order, then there exists a more efficient method of accounting for the remaining heteroscedasticity in estimation. Treating equations (6) and (9) as a seemingly unrelated regression model, one can estimate the parameters of these equations by joint generalized least squares. Using least squares residuals and fitted values, one can easily form consistent estimates for the parameters $E(\varepsilon^2)$, $E(\varepsilon^3)$, and $E(\varepsilon^4)$, and for the quantities $X_i'\beta_0$. With these estimates, it is possible to construct a consistent estimate for each covariance matrix Φ_i,

denoted by $\hat{\Phi}_i$. Defining $\varepsilon_i^* \equiv (Y_i - X_i'\beta)$ and $\nu_i^* \equiv (Y_i^2 - \sigma^2 - (X_i'\beta)^2)$, joint generalized least squares in effect sets

$$\ell_i(\beta, \sigma^2) = -\frac{\partial(\varepsilon_i^* \ \nu_i^*)}{\partial\begin{pmatrix}\beta \\ \sigma^2\end{pmatrix}} \hat{\Phi}_i^{-1}\begin{pmatrix}\varepsilon_i^* \\ \nu_i^*\end{pmatrix} = \begin{pmatrix}X_i & 2X_i(X_i'\beta) \\ 0 & 1\end{pmatrix}\hat{\Phi}_i^{-1}\begin{pmatrix}(Y_i - X_i\beta) \\ (Y_i^2 - \sigma^2 - (X_i'\beta)^2)\end{pmatrix}$$

and computes an estimate for β_0 and σ_0^2 by minimizing (2). While this procedure yields a more efficient estimator for β_0 than the one based on the joint least squares estimation of (6) and (9), it is still necessary for third moments to be nonzero before an efficiency gain is attained over the conventional least squares estimator.

As soon as one assumes that moments of the ε_i's above second order are constant across observations, there are even more sources of information available for improving estimation efficiency. The constancy of third moments suggests consideration of a cubic version of the regression equation. Cubing both sides of (7) yields

$$(10) \qquad Y_i^3 = \tau_0 + 3(X_i'\beta_0)\sigma_0^2 + (X_i'\beta_0)^3 + \eta_i, \qquad i = 1,\ldots,N,$$

where τ_0 and σ_0^2 are unknown parameters, and $\eta_i \equiv 3(X_i'\beta_0)^2\varepsilon_i + 3(X_i\beta_0)(\varepsilon_i^2 - \sigma_0^2) + (\varepsilon_i^3 - \tau_0)$ is a disturbance with zero mean. Nonlinear least squares applied to (10) yields consistent estimates for β_0, σ_0^2 and τ_0 assuming satisfaction of the orthogonality conditions $\lim E(\frac{1}{N}\sum_{i=1}^{N} X_iX_i'(X_i'\beta_0)(\varepsilon_i^2 - \sigma_0^2)) = 0$ and $\lim E(\frac{1}{N}\sum_{i=1}^{N} X_iX_i'(X_i'\beta_0)(\varepsilon_i^3 - \tau_0)) = 0$. These orthogonality conditions obviously follow if $E(\varepsilon_i^2)$ and $E(\varepsilon_i^3)$ do not vary with i. It is also interesting to note that these conditions necessarily follow in those situations where the combined estimation of

(6) and (9) produces a consistent estimate for β_0, but yields no efficiency gain.

Joint least squares estimation of equations (6), (9) and (10), then, imposing all the constraints implied across equations, offers yet another procedure for constructing an estimator for β_0. Before constraining the estimates of β_0 obtained from these three equations to be equal, it is a straightforward matter to test the orthogonality conditions required for imposing this constraint using a direct analogue of the testing procedure described above. In particular, with the ℓ_i relations made up of the gradient vectors associated with least squares estimation of equations (6), (9), and (10), and with the coefficients representing β_0 and σ_0^2 made distinct across equations, one can test the orthogonality conditions by checking for equality of the probability limits for the different estimates of β and σ^2 using a Wald statistic and the asymptotic distribution implied by (3). The above discussion also indicates what is required to test the hypothesis that third moments are constant across observations. In addition to testing the restriction that the estimates of β a σ^2 obtained from equations (6), (9), and (10) are the same, a test for the constancy of third moments also implies that any other measured variables entered into equation (10) have zero coefficients.

As in the previous case dealing with the joint estimation of only equations (6) and (9), several options are available for simultaneously estimating the regression equation and its squared and cubic variants. Assuming the ε_i's have constant moments across the sample up to sixth order, joint generalized least squares offers the most efficient estimation procedure. Treating equations (6), (9), and (10) as a seemingly unrelated regression model, let $e_i' = (\varepsilon_i, \nu_i, \eta_i)$ denote the vector of disturbances associated with this

model for the i^{th} observation; define $\Psi_i \equiv E(e_i e_i')$ as the implied covariance matrix; let $\hat{\Psi}_i$ represent a consistent estimate for Ψ_i; and define $e_i^{*\prime} = (\varepsilon_i^*, \nu_i^*, \eta_i^*)$ where $\eta_i^* \equiv Y_i^3 - \tau - 3(X_i'\beta)\sigma^2 - (X_i'\beta)^3$. Joint generalized least squares applied to this system of equations in effect computes estimates for β_0, σ_0^2 and τ_0 by minimizing quadratic form (2) with

$$\ell_i(\beta, \sigma^2, \tau) = C_i \hat{\Psi}_i^{-1} e_i^*, \qquad i = 1,\ldots,N$$

where

$$C_i \equiv -\frac{\partial e_i^{*\prime}}{\partial \begin{pmatrix} \beta \\ \sigma^2 \\ \tau \end{pmatrix}} \equiv \begin{pmatrix} X_i & 2X_i(X_i'\beta) & 3X_i(\sigma^2 + (X_i'\beta)^2) \\ 0 & 1 & 3(X_i'\beta) \\ 0 & 0 & 1 \end{pmatrix}.$$

To discover the conditions under which this procedure yields an improved efficiency for the estimate of β_0 over ordinary least squares, it is once again convenient to check a condition like (5). Least squares applied to (6) computes estimates using the relations $g_i(\beta) = X_i(Y_i - X_i'\beta)$; and joint generalized least squares applied to (6), (9), and (10) uses the relations $h_i(\beta, \sigma^2, \tau) = C_i \hat{\Psi}_i^{-1} e_i^*$ to calculate estimates. Partition the matrix C_i as $C_i' = [C_{1i}' \; C_{2i}']$ with $C_{1i} = [X_i \quad 2X_i(X_i'\beta) \quad 3X_i\sigma^2 + (X_i'\beta)^2]$, and let a "o" superscript on these matrices indicate evaluation at the true parameter values. It is possible to show that the matrix G associated with this problem is the asymptotic limit of

$$\left[-\frac{1}{N}\sum_{i-1}^{N} C_i^o \Psi_i^{-1} C_{1i}^{o\prime}\right] + \left[\frac{1}{N}\sum_{i=1}^{N} C_i^o \Psi_i^{-1} K_i\right],$$

where

$$K_i \equiv \frac{1}{\sigma_0^2} E(e_i \varepsilon_i) X_i' = \frac{1}{\sigma_0^2} \begin{pmatrix} \sigma_0^2 X_i' \\ 2\sigma_0^2 (X_i'\beta_0) X_i' + E(\varepsilon_i^3) X_i' \\ 3\sigma_0^2 (X_i'\beta_0)^2 X_i' + 3E(\varepsilon_i^3)(X_i'\beta_0) X_i' + E(\varepsilon_i^4) X_i' \end{pmatrix}.$$

This latter matrix involves both the third and the fourth moments of the ε_i's. Accordingly, the matrix G does not in general vanish even when the distribution of ε is symmetric, and condition (5) is typically not satisfied implying an efficiency gain for the estimate of β_0. In one special case, it is possible to show that G = 0 and the gain in efficiency is lost. Not surprisingly, this case corresponds to a situation in which disturbances are normally distributed, implying $E(\varepsilon_i^3) = E(\varepsilon_i^5) = 0$, $E(\varepsilon_i^4) = 3\sigma_0^4$, and $K_i = C_{1i}^o$, as a consequence. According to this analysis, then, joint estimation of the regression equation and its squared and cubed variants yields an estimator for β_0 that is strictly efficient relative to the conventional least squares estimator in essentially every instance except when least squares corresponds to the application of maximum likelihood.

One can, of course, continue and consider a quartic variant and higher order powers of the regression equation. Joint least squares estimation of these new equations along with those introduced above without constraints across equations offers a framework for testing: (i) whether one can justify imposing an equality restriction for the estimates of β obtained from the different equations; and (ii) whether one can accept the stronger hypotheses that fourth and possibly higher order moments of the disturbances are constant across observations. If the estimates of β computed using the different equations are

all consistent for β_0, then joint estimation of these equations constraining all the estimates of β to be equal will almost certainly lead to a more efficient estimator for β_0 than the one computed using fewer equations and, consequently, less information on the moments of disturbances. Knowing the ε_i's have constant moments across the sample offers yet another source of information to attain further gains in efficiency. Such information makes it possible to estimate at least a subset of the equations under consideration by generalized rather than least squares procedures.

Nonlinear Regression

All the procedures outlined in this section for improving on the least squares estimate may be applied in the nonlinear regression case as well. In particular, instead of (6), suppose observations on Y are generated by the equation

$$(11) \qquad Y_i = \phi(X_i, \beta_0) + \varepsilon_i, \qquad i = 1,\ldots,N,$$

where $\phi(\cdot)$ is a known function. One may readily verify that the results and the relations derived above remain valid after making the following modifications: replace $X_i'\beta$ and $X_i'\beta_0$ by $\phi(X_i, \beta)$ and $\phi(X_i, \beta_0)$, respectively; and substitute $\frac{\partial \phi_i}{\partial \beta}\Big|_{\hat{\beta}}$ for X_i where the point of evaluation $\hat{\beta}$ is consistent for β_0.[5] Thus, when the ε_i's are heteroscedastic, joint nonlinear least squares estimation of (11) and weighted versions of this equation yields a more efficient estimator for β_0 than conventional nonlinear least squares. If, on the other hand, the ε_i's have moments above first order that satisfy the orthogonality conditions associated with the nonlinear model or that are constant across observations, then joint least or generalized least squares

estimation of (11) and its squared and cubic variants creates an estimator for β_0 whose efficiency in general dominates the one obtained from a standard application of nonlinear least squares. Because the demonstration of these results for nonlinear regression involves no new concepts and follows directly from the previous analysis, the discussion instead considers the applications of the above estimation procedures to a more general representation of a nonlinear equation of the sort encountered in simultaneous equation analysis which nests (11) as a special case.

III. Instrumental Variable Estimation

Suppose observations on a random vector Y obey the equation

$$(12) \qquad f(Y_i, X_i, \gamma_0) = \varepsilon_i, \qquad i = 1,\ldots,N,$$

where $f(\cdot)$ is a known function, and γ_0 is an unknown parameter vector. As in the previous analysis, the ε_i's are assumed to be distributed independently across observations with $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) \equiv \sigma_i^2$ where expectations are calculated given a measured set of characteristics Z.

Given a vector of instruments q_i for each observation, nonlinear two-stage least squares estimation of (12) uses the orthogonality conditions $E(q_i f(Y_i, X_i, \gamma_0)) = 0$ to compute an estimate for γ_0. Formally, the elements of q_i may depend on estimated coefficients as well as on the measured characteristics Z as long as they are asymptotically nonstochastic. A standard application of nonlinear two-stage least squares sets $\ell_i(\gamma) = q_i f_i$ with $f_i \equiv f(Y_i, X_i, \gamma)$ and calculates an estimate by minimizing a quadratic form like (2) with the matrix $B_N = \frac{1}{N} \sum_{i=1}^{N} q_i q_i'$ replacing the matrix $\hat{V}_N$. Observing that the implied specification for $\hat{V}_N$ is $\frac{1}{N} \sum_{i=1}^{N} q_i q_i' \hat{\varepsilon}_i^2$ where $\hat{\varepsilon}_i$ is a residual consistent for $f(Y_i, X_i, \gamma_0)$, it is not surprising to find that the direct minimization of (2) in general yields a more efficient estimator for γ_0. Referring to the discussion of Chamberlain's approach for improving on least squares estimation, it is evident that conventional instrumental variable techniques construct estimators in a way analogous to constrained least squares procedures and, thus, suffer from the same deficiencies.

In particular, one can demonstrate that conventional two-stage least squares yields a less efficient estimator for γ_0 than the one obtained by directly minimizing (2) when disturbances are heteroscedastic and there are exclusion or over-identifying restrictions. Chamberlain (1982), White (1982) and Amemiya (1982) each propose estimators for linear simultaneous equations that exploit exactly this observation. Clearly, one can exploit this same source of efficiency gain when constructing estimators for nonlinear simultaneous equations as well.

The optimal choice for the vectors of instruments varies according to whether disturbances are homoscedastic or heteroscedastic. For the class of estimators defined by minimizing (2) with $\ell_i = q_i f_i$, Amemiya (1975) shows that setting q_i equal to $E\left(\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right)$ or to a consistent estimate of this quantity yields the most efficient estimator for γ_0 when errors are homoscedastic. An estimator with approximately the same efficiency is obtained if one instead sets q_i equal to any vector of explanatory variables W_i where a regression of the elements of $\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}$ on W_i would in principal produce fitted values that closely approximate $E\left(\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right)$.[6] While these alternative choices for the vector of instruments are essentially equivalent in the homoscedastic case, choosing $q_i = W_i$ will in general yield the more efficient parameter estimate in the heteroscedastic case. Using $q_i = E\left(\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right)$ creates a just-identified model (i.e., the dimensions of q and γ are equal), and minimizing (2) for such a model yields the same estimator for γ_0 regardless of whether or not errors are heteroscedastic. The use of W_i as instruments, on the other hand, typically involves some degree of over-identification, and as noted above the minimization of (2) exploits over-identifying restrictions in the presence of heteroscedasticity

to develop an estimator with greater efficiency than one that ignores such restrictions. While one can argue that W_i dominates $E\left[\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right]$ as instruments in the heteroscedastic case, there is, unfortunately, no clear strategy for choosing the optimal W_i in this case if one purely views the problem as introducing over-identifying restrictions.

The following analysis applies the strategies proposed in the previous section to formulate estimators for γ_0 that possess greater efficiency than those obtained from the application of conventional two-stage least squares. In this analysis, the optimal choice for instruments is the same as in the homoscedastic case. Thus, it is assumed throughout this discussion that one sets the vector of instruments equal either to the expected value of the gradient vector associated with the nonlinear equation under consideration or to a vector W_i designed to predict this gradient vector accurately.

Non-constant Moments Case

To construct an estimator for γ_0 with improved efficiency when the disturbances of equation (12) are heteroscedastic, one may jointly estimate (12) and a weighted variant of this equation. Letting the weights ω_i represent asymptotically nonstochastic variables designed to eliminate the variation in the transformed variances $\omega_i^2\varepsilon_i^2$ across observations, weighting (12) yields

$$\omega_i f(Y_i, X_i, \gamma_0) = \omega_i \varepsilon_i, \qquad i = 1,\ldots,N. \tag{13}$$

Nonlinear two-stage least squares applied to (12) uses the vector of functions $g_i(\gamma) = q_i f_i$ to compute an estimate for γ_0; and applying this procedure to estimate (13) uses the vector of functions $h_i(\gamma) = \omega_i^2 q_i f_i$. Joint two-stage

least squares estimation of equations (12) and (13) in this analysis means that one sets $\ell_i' = (g_i', h_i')$ and minimizes (2) with respect to γ.

The resulting estimator is strictly efficient relative to the conventional two-stage least squares estimator in the presence of heteroscedasticity; and it also dominates the weighted two-stage least squares estimator unless the weights are optimally chosen (i.e., $\omega_i = \frac{1}{\sigma_i}$) in which case the efficiencies of these two estimators are the same. Similar to the analysis above, one only needs to consider the form of the matrix G to verify these propositions concerning efficiency. The matrix G for this problem is the asymptotic limit of

$$\left[\frac{1}{N}\sum_{i=1}^{N}\omega_i^2 q_i E\left(\frac{\partial f_i}{\partial \gamma'}\Big|_{\gamma_0}\right)\right] - \left[\frac{1}{N}\sum_{i=1}^{N}\omega_i^2 q_i q_i' \sigma_i^2\right]\left[\frac{1}{N}\sum_{i=1}^{N} q_i q_i' \sigma_i^2\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} q_i E\left(\frac{\partial f_i}{\partial \gamma'}\Big|_{\gamma_0}\right)\right].$$

As expected, given an optimal choice for the vector of instruments $q_i = E\left(\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right)$, $G = 0$ when variances are constant across observations implying no efficiency gain. In general, however, $G \neq 0$ and the joint estimation of equations (12) and (13) leads to an improvement in efficiency over the estimation of equation (12) alone.

As noted previously, regressing the squared residuals $\hat{\varepsilon}_i^2$ on a function of explanatory variables and treating the resulting fitted values as a measure of $1/\omega_i^2$ offers a simple and natural procedure for forming the weights used in (13). If this choice for the weights fails to adjust for the variability of the σ_i's, one may introduce a weighted version of equation (13) and jointly estimate this new equation along with the others to obtain a further improvement in efficiency.

Constant Moments Case

Given the knowledge that the disturbances of the equation under consideration are homoscedastic, the analysis of the previous section suggests that jointly estimating this equation along with its squared variant will typically produce an increased efficiency in estimating the parameters of interest. Squaring both sides of equation (12) and introducing an intercept yields

$$(14) \qquad f^2(Y_i, X_i, \gamma_0) - \sigma_0^2 = u_i, \qquad i = 1,\ldots,N,$$

where $u_i \equiv \varepsilon_i^2 - \sigma_0^2$. If the variances of the ε_i's are constant across observations and equal to σ_0^2, then $E(u_i) = 0$ and (14) is in the form of a nonlinear structural equation. One must exercise caution, however, in applying a nonlinear two-stage least squares routine to estimate the parameters of this equation. Such a routine uses the vector of functions $\ell_i(\gamma, \sigma^2) = q_i^s(f_i^2 - \sigma^2)$ to compute estimates for γ_0 and σ_0^2 with q_i^s representing a vector of instruments. A problem encountered with implementing this routine relates to the properties of the matrix of the averaged partial derivatives associated with vectors of functions $\ell_i(\gamma, \sigma^2)$. As noted in the discussion of this general class of estimation schemes in Section I, application of these schemes requires the matrix $S_N \equiv \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_i}{\partial \theta'}$ to possess full column rank over the relevant portion of the parameter space. For the estimation problem considered here, the form of this matrix evaluated at the true parameter values is $S_N^o \equiv S_N(\gamma_0, \sigma_0^2) = \frac{1}{N} \sum_{i=1}^{N} q_i^s r_i'$, where $r_i' = (r_{1i}' \; r_{2i})$ denotes a row vector with $r_{1i} = 2\varepsilon_i \left.\frac{\partial f_i}{\partial \gamma}\right|_{\gamma_0}$ and $r_{2i} = -1$. It is easy to find cases for which S_N^o possesses

less than full column rank in the limit. Such situations arise, for example, whenever equation (14) refers to a regression relation (e.g., $f(Y_i, X_i, \gamma_0) = Y_i - X_i'\gamma_0$ or $= Y_i - \phi(X_i, \gamma_0)$), in which case $E(r_{1i}) = 0$ implying $S_N^o \overset{s}{\to} [0\ \bar{q}^s]$ where $\bar{q}^s \equiv \lim \frac{1}{N} \sum_{i=1}^{N} q_i^s$.

While difficulties may arise with regard to the direct estimation of equation (14), forming a linear combination of equations (12) and (14) creates an estimable specification that depends on the squared representation of the structural equation and incorporates second moment information. Instead of (14), then, consider the linear combination

$$(15) \qquad f^2(Y_i, X_i, \gamma_0) - \sigma^2 + a_i f(Y_i, X_i, \gamma_0) = u_i + a_i \varepsilon_i, \qquad i = 1,\ldots,N,$$

where the a_i's denote asymptotically nonstochastic variables. Nonlinear two-stage least squares applied to (15) uses the vectors of functions $\ell_i(\gamma, \sigma^2) = q_i^a(f_i^2 - \sigma^2 + a_i f_i)$ to calculate estimates for γ_0 and σ_0^2 with q_i^a representing a vector of instruments. Assuming the original structural equation given by (12) is estimable and the a_i's are nonzero, one may readily verify that the matrix of averaged partials S_N associated with this estimation problem possesses full column rank. Equation (15), then, may be directly estimated using nonlinear two-stage least squares, and this routine produces consistent estimates for γ_0 and σ_0^2 if the ε_i's are homoscedastic. If one sets the a_i's equal to the constant $\bar{a}$, the optimal choice for instruments is

$$q_i^a = E\left[\left.\frac{\partial(f_i^2 - \sigma^2 + \bar{a}f_i)}{\partial\begin{pmatrix}\gamma\\ \sigma^2\end{pmatrix}}\right|_{\gamma_0,\sigma_0}\right] = \begin{pmatrix} 2E\left[\varepsilon_i \left.\frac{\partial f_i}{\partial\gamma}\right|_{\gamma_0}\right] + \bar{a}\, E\left[\left.\frac{\partial f_i}{\partial\gamma}\right|_{\gamma_0}\right] \\ -1 \end{pmatrix}.$$

Notice that the squared variant of the linear regression equation considered in the previous section given by (9) is in the form of (15) with $f_i = Y_i - X_i'\beta_0$ and $a_i = 2X_i'\beta_0$.

Jointly estimating equations (12) and (13) without constraints across equations provides a framework for testing both: (i) the orthogonality conditions needed for the estimate of γ computed using equation (13) alone to be consistent for γ_0; and (ii) whether the disturbances of the original structural equation have constant variances across observations. Joint nonlinear two-stage least squares estimation of these equations means that one sets $\ell_i'(\gamma_1, \gamma_2, \sigma^2) = (g_i', h_i')$ with $g_i(\gamma_1) = q_i f_i$ and $h_i(\gamma_2, \sigma^2) = q_i^a(f_i^2 - \sigma^2 + a_i f_i)$, where γ_1 and γ_2 represent distinct coefficient vectors, and minimizes (2) with respect to γ_1, γ_2, and σ^2. The orthogonality conditions imply that the estimates $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ have the same probability limits, and one may test this hypothesis by computing a Wald statistic using the asymptotic distribution of estimates implied by (3). In addition to this equality restriction, a homoscedasticity assumption also implies that no other function of the explanatory variables enters equation (13). If one considers only relationships between the σ_i^2's and the explanatory variables that may be described by functions linear in parameters, then checking for homoscedasticity amounts to performing tests of hypotheses consisting of the constraint $\gamma_1 = \gamma_2$ and zero restrictions for all parameters other than γ_1, γ_2, and σ^2. Once again, then, standard Wald statistics offer a simple method for testing the homoscedasticity assumption in the case of simultaneous equations.

Given acceptance of the equality restriction for the γ coefficients across equations, joint nonlinear two-stage least squares estimation of equations (12) and (15) imposing the constraint $\gamma_1 = \gamma_2$ generates an estimator

for γ_0 that is typically more efficient than the conventional two-stage least squares estimator. Inspecting the matrix G for the choice of g and h used by this estimation procedure reveals the exact conditions needed to achieve this efficiency gain. Assuming $a_i = 1$ for all i, G is the asymptotic limit of

$$\left[\frac{1}{N}\sum_{i=1}^{N} q_i^a E\left((2\varepsilon_i+1)\frac{\partial f_i}{\partial \gamma'}\Big|_{\gamma_0}\right)\right] - \left[\frac{1}{N}\sum_{i=1}^{N} q_i^a q_i' E(\varepsilon_i^3+\varepsilon_i^2)\right]\left[\frac{1}{N}\sum_{i=1}^{N} q_i q_i' E(\varepsilon_i^2)\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} q_i E\left(\frac{\partial f_i}{\partial \gamma'}\Big|_{\gamma_0}\right)\right].$$

Evaluating this expression assuming $E(\varepsilon_i^2) = \sigma_0^2$ and for the optimal choice of instruments (i.e.,

$q_i = E\left(\frac{\partial f}{\partial \gamma}\Big|_{\gamma_0}\right)$ and $q_i^{a'} = (q_{1i}^{a'}\ q_{2i}^{a})$ with $q_{1i}^{a} = 2E\left(\varepsilon_i \frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right) + q_i$ and $q_{2i}^{a} = -1$),

it is evident that $G_1 \neq 0$ if at least one of the following conditions are satisfied: (i) $E(\varepsilon_i^3) \neq 0$; or (ii) $E\left(\varepsilon_i \frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right) \neq 0$. In contrast, then, to the above findings on least squares estimation, jointly estimating a structural equation and its squared variants in the presence of homoscedasticity can yield a more efficient estimator for the parameters of interest even if the third moments of the ε_i's are zero for all observations; a nonzero correlation between ε_i and the gradient vector $\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}$ also implies that joint estimation yields an efficiency gain.

Extending the analysis to consider cubic and higher order powers of the structural relation involves no concepts not already discussed above. Joint nonlinear two-stage least squares estimation of these new equations along with (12) and (15) without equality constraints across equations offers a framework for testing the appropriateness of imposing such constraints. Given one can justify these equality restrictions, joint estimation imposing these constraints

will yield further efficiency gains for the resulting estimator of γ_0. One may also test hypotheses concerning the constancy of third and higher order moments of the disturbances using this joint estimation framework. As outlined above, hypotheses of this nature translate into equality constraints across equations and into zero restrictions for the effects of any other explanatory variables one might choose to include in these equations. Information of this nature determines whether it is possible to attain additional improvements in efficiency by applying three-stage rather than two-stage least squares procedures. Knowing, for example, that moments of the ε_i's up to fourth order are constant permits one to estimate equations (12) and (15) using nonlinear three-stage least squares, which makes the choice for the a_i's irrelevant and provides for an optimal combination of these equations in the calculation of estimates.

To illustrate the use of nonlinear three-stage least squares procedures in this analysis, consider the joint estimation of the original specification of the structural equation given by (12), its squared representation given by (14), and its cubic variant given by

$$(16) \qquad f^3(Y_i, X_i, \gamma_0) - \tau_0 = v_i, \qquad i = 1,\ldots,N,$$

where $v_i \equiv \varepsilon_i^3 - \tau_0$. Stacking equations (12), (14), and (16) creates a model with $(\varepsilon_i\ u_i\ v_i)$ representing the disturbance vector for the i^{th} observation. This disturbance vector is homoscedastic assuming the moments $E(\varepsilon_i^k)$ for $k \leq 6$ are constant across the sample. Letting Ω_0 denote the covariance matrix associated with this disturbance vector, and $\hat{\Omega}$ denote a consistent estimate for Ω_0, nonlinear three-stage least squares applied to this three equation model computes estimates for the parameters γ_0, σ_0^2, and τ_0 by minimizing quadratic form (2) with

$$\ell_i(\gamma, \sigma^2, \tau) = R_i \hat{\Omega}^{-1} \begin{bmatrix} f_i \\ f_i^2 - \sigma^2 \\ f_i^3 - \tau \end{bmatrix}$$

and

$$R_i = \begin{bmatrix} q_i & q_{1i}^s & q_{1i}^c \\ 0 & q_{2i}^s & 0 \\ 0 & 0 & q_{2i}^c \end{bmatrix},$$

where q_i, q_{1i}^s, and q_{1i}^c are vectors of instruments with the same dimension as γ, and q_{2i}^s and q_{2i}^c are scalar instruments. According to the results of Amemiya (1977), the optimal choice for these instruments is $q_i = E\left[\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right]$, $q_{1i}^s = 2E\left[\varepsilon_i \frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right]$, $q_{1i}^c = 3E\left[\varepsilon_i^2 \frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right]$, and $q_{2i}^s = q_{2i}^c = -1$. To identify the conditions under which this three-stage procedure yields an efficiency gain over conventional two-stage least squares estimation of equation (12), one may set $g_i = q_i f_i$ (used in the estimation of (12)), h_i to the specification for ℓ_i listed directly above, and determine the form for the matrix G. When evaluated at the optimal choice for instruments, this matrix is the asymptotic limit of

$$\left[\frac{1}{N} \sum_{i=1}^{N} R_i \Omega_0^{-1} (q_i \; q_{1i}^s \; q_{1i}^c)'\right] - \left[\frac{1}{\sigma_0^2 N} \sum_{i=1}^{N} R_i (q_i \; 0 \; 0)'\right].$$

Except in those cases where two-stage least squares estimation of (12) corresponds directly to the application of maximum likelihood, this matrix does not vanish. Thus, one again finds that using information on the higher

order moments of the disturbances provides for an increase in the efficiency of estimation.

Conclusion

The overall estimation strategy suggested by this analysis may be summarized as follows. Starting with a form of the regression (or simultaneous) equation whose disturbances are believed to be homoscedastic, one introduces a new equation obtained by squaring both sides of this equation. Joint unconstrained estimation of the regression equation and its squared variant offers a framework for testing: (i) whether the second moments of disturbances satisfy the orthogonality conditions needed to justify the imposition of equality constraints across equations; and (ii) whether one can accept the stronger hypothesis that disturbances have constant variances across observations.

Given rejection of the equality restrictions relating the parameters of the regression equation and its squared variant implied by hypothesis (i), one eliminates the equation relating the squares of variables from the analysis and implements the procedure that jointly estimates the regression equation and weighted variants of this equation. More than one weighted regression equation may be included in this joint estimation procedure. While one in principal wants to choose weights in a way to induce homoscedasticity, it is unnecessary to know the form of heteroscedasticity to implement this procedure, and almost any choice for weights will yield an improvement in the efficiency of estimation.

Given acceptance of the equality restrictions implied by hypothesis (i), one acquires further information by considering cubic and possibly higher order powers of the regression relation. Jointly estimating these new equations along

with the regression equation and its squared variant initially without constraints across equations allows one to test: (iii) whether third and higher order moments of disturbances satisfy the orthogonality conditions required to restrict the estimates associated with the different equations; and (iv) whether disturbances have constant third and higher order moments over the sample. Assuming tests of (iii) support the imposition of equality restrictions across equations, carrying out joint estimation imposing these constraints will almost always leads to a more efficient estimator for the parameters of interest than one computed using fewer equations and, consequently, less information on the moments of disturbances. If one finds that moments vary over the sample but still satisfy conditions (i) and (iii), then one can achieve further increases in the efficiency of estimation by also including weighted variants of equations with heteroscedastic errors in the joint estimation procedure as well. If, on the other hand, tests also support property (iv) and the disturbances have constant moments over the sample, one can exploit this information to improve estimation efficiency by using generalized or three-stage least squares methods when jointly estimating equations.

Following this estimation strategy, one can literally proceed indefinitely in finding another equation which when jointly estimated along with the equations currently considered leads generally to a further gain in the efficiency of estimation. Essentially, whenever it can be argued that including an equation of one type does not yield an efficiency gain, this fact itself suggests a new source of information that may be exploited by introducing an equation of another type. This study offers no rule indicating where one should stop adding equations to the joint estimation scheme. While specific distributional assumptions provide the basis for formulating such rules, knowledge of these assumptions often means

that it is possible to use maximum likelihood techniques to estimate parameters; in which case, on efficiency grounds, there is no reason to consider the estimation procedures proposed in this paper.

Footnotes

1. Letting ℓ_{ji} and s_{jki} denote the j and (j, k) elements of ℓ_i and $\frac{\partial \ell_i}{\partial \theta'}$, respectively, sufficient conditions restricting the tails of distributions are: $E\left|\ell_{ji}\right|^{2+\delta_2} \leq C_1 < \infty$ and $E\left|s_{jki}\right|^{1+\delta_1} \leq C_2 < \infty$ for some $\delta_1, \delta_2 > 0$ and all $\theta \in \Theta$.

2. For the matrix algebra theorem needed here, see Rao (1973. p. 77, problem 33).

3. If the ω_i's depend on estimated coefficients, it is, of course, not in general true that $E(h_i(\beta_0)) = 0$. All the consistency and asymptotic normality results of Section I, however, remain valid in this case as long as: each $\omega_i \xrightarrow{s} \bar{\omega}_i$ with $\bar{\omega}_i$ nonstochastic; and the vectors $\bar{h}_i = \bar{\omega}_i^2 X_i(Y_i - X_i'\beta)$ satisfy the properties of the ℓ_i's outlined in the previous section.

4. For the current problem, there is a simpler procedure for testing the hypotheses concerning the orthogonality conditions and/or the homoscedasticity assumption that avoids the need for jointly estimating equations (6) and (9). Following the work of White (1980), one may carry out these tests using a multiple regression framework with squares of residuals treated as dependent variables and with the unique components of X_iX_i' and possibly other explanatory variables serving as regressors. Performing a conventional F-test of the hypothesis that coefficients other than the intercept are zero yields an asymptotically valid test for either of the hypotheses cited above, assuming the constancy of the fourth moments of disturbances. One cannot, however, use this simpler testing framework for any of the similar situations considered below, including those concerned with testing for the constancy of third or higher order moments in the regression case or for the constancy of any moments in the simultaneous equation case.

5. Alternatively, one may substitute the vector $\frac{\partial \phi_i}{\partial \beta}$ viewed as a function of β for X_i when forming the ℓ_i's and replace X_i by $\frac{\partial \phi_i}{\partial \beta}\Big|_{\beta_0}$ in those expressions for the matrices G and A.

6. According to formula (3), the use of $q_i = E\left\{\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right\}$ to compute an estimate for γ_0 implies a covariance matrix equal to the limit of

$$\sigma_0^2\left[\frac{1}{N}\sum_{i=1}^{N} E\left(\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right) E\left(\frac{\partial f_i}{\partial \gamma'}\Big|_{\gamma_0}\right)\right]^{-1}$$; and setting $q_i = W_i$ yields a covariance matrix that may be written as the asymptotic limit of the matrix $\sigma_0^2\left[\frac{1}{N}\sum_{i=1}^{N} P_i P_i'\right]^{-1}$ where $P_i = \left[\sum_{j=1}^{N} \frac{\partial f_j}{\partial \gamma}\Big|_{\gamma_0} W_j'\right]\left[\sum_{k=1}^{N} W_k W_k'\right]^{-1} W_i$ represents the vector of fitted values obtained from regressing each of the elements of $\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}$ on W_i. Thus, as stated in the text, if $P_i \approx E\left(\frac{\partial f_i}{\partial \gamma}\Big|_{\gamma_0}\right)$, these two covariance matrices are approximately equal.

References

Amemiya, T. "The Nonlinear Two-stage Least-squares Estimator." Journal of Econometrics (1974): 105-10.

_____. "The Nonlinear Limited-information Maximum-likelihood Estimator and the Modified Nonlinear Two-stage Least-squares Estimator." Journal of Econometrics (1975): 375-86.

_____. "The Maximum Likelihood and the Nonlinear Three-stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model." Econometrica (May 1977): 955-68.

_____. "Partially Generalized Least Squares and Two-stage Least Squares Estimators." Technical Report No. 368, February 1982, IMSSS, Stanford University.

Chamberlain, Gary. "Multivariate Regression Models for Panel Data." Forthcoming Journal of Econometrics 1982.

Hansen, Lars Peter. "Large Sample Properties of Generalized Methods of Moments Estimators." Forthcoming Econometrica. Mimeo 1981 Carnegie-Mellon University.

Rao, C. R. Linear Statistical Inference and Its Applications, 2nd ed. New York: John Wiley & Sons, Inc., 1973.

White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity." Econometrica 48 (1980): 721-46.

_____. "Instrumental Variable Regression with Independent Observations." Econometrica (1982): 483-500.

NBER TECHNICAL WORKING PAPER SERIES

Number	Author	Title	Date
12	Willem H. Buiter	A Note on the Solution of a Two-Point Boundary Value Problem Frequently Encountered in Rational Expectations Models	6/81
13	Willem H. Buiter	Macroeconometric Modelling for Policy Evaluation and Design	6/81
14	Thomas E. MaCurdy	Asymptotic Properties of Quasi-Maximum Likelihood Estimators and Test Statistics	6/81
15	Gary Chamberlain and Michael Rothschild	Arbitrage and Mean-Variance Analysis on Large Asset Markets	7/81
16	Mervyn A. King	Welfare Analysis of Tax Reforms Using Household Data	7/81
17	Jerry A. Hausman, Bronwyn Hall and Zvi Griliches	Econometric Models for Count Data with an Application to the Patents-R&D Relationship	8/81
18	James N. Brown and Harvey S. Rosen	On the Estimation of Structural Hedonic Price Models	11/81
19	David S. Jones and V. Vance Roley	Bliss Points in Mean-Variance Portfolio Models	12/81
20	Willem H. Buiter	Saddlepoint Problems in Continuous Time Rational Expectations Models: A General Method and Some Macroeconomic Examples	2/82
21	Willem H. Buiter	Predetermined and Non-Predetermined Variables in Rational Expectations Models	3/82
22	Thomas E. MaCurdy	Using Information on the Moments of Disturbances to Increase the Efficiency of Estimation	5/82

Note: Copies of the above technical working papers can be obtained by sending $1.50 per copy to Working Papers, NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138. Advance payment is required on orders totaling less than $10.00. Please make check payable to National Bureau of Economic Research, Inc.