THREE STAGE LEAST SQUARES ESTIMATION FOR A SYSTEM OF SIMULTANEOUS, NONLINEAR, IMPLICIT EQUATIONS

by

A. RONALD GALLANT

Institute of Statistics
Mimeograph Series No. 1032
Raleigh, N. C. 1975

H.G.B. ALEXANDER RESEARCH FOUNDATION

GRADUATE SCHOOL OF BUSINESS

UNIVERSITY OF CHICAGO

THREE STAGE LEAST SQUARES ESTIMATION FOR A SYSTEM OF SIMULTANEOUS, NONLINEAR, IMPLICIT EQUATIONS

by

A. Ronald Gallant

Preliminary September, 1975

THREE STAGE LEAST SQUARES ESTIMATION FOR A SYSTEM OF SIMULTANEOUS, NONLINEAR, IMPLICIT EQUATIONS

A. Ronald GALLANT*

North Carolina State University
Raleigh, N. C. 27607, U. S. A.

The article describes a nonlinear three stage least squares estimator for a system of simultaneous, nonlinear, implicit equations. The estimator is shown to be strongly consistent, asymptotically normally distributed, and more efficient than the nonlinear two stage least squares estimator.

1. Introduction

Recently, Amemiya (1974) set forth a nonlinear two stage least squares estimator and derived its asymptotic properties. This article is an extension of his work. A nonlinear three stage least squares estimator is proposed, and its asymptotic properties are derived.

A salient characteristic of most nonlinear systems of equations is that it is impossible to obtain the reduced form of the system and costly, if not actually impossible, to obtain it implicitly using numerical methods.[1] For this reason, the estimation procedure proposed here does not require the reduced form--the system may remain in implicit form throughout. In this sense, the nonlinear two stage estimator used as the first step of the procedure is a generalization of Amemiya's estimator since he assumes that at least one endogenous variable in the structural equation to be estimated can be written explicitly in terms of the rest--that assumption is not made here.

---

*Presently on leave with the Graduate School of Business, University of Chicago, 5836 Greenwood Avenue, Chicago, Illinois 60637. The author wishes to thank Professor Arnold Zellner for helpful suggestions.

The estimation procedure is a straightforward generalization of the linear three stage least squares estimator.[2] The estimator thus obtained is shown to be strongly consistent, asymptotically normally distributed, and more efficient than the nonlinear two stage least squares estimator. The regularity conditions used to obtain these results--while standard[3]--are somewhat abstract from the point of view of one whose interest is in the applications. This point of view is kept in mind throughout the development; the practical implications of the regularity conditions--and some pitfalls--are discussed and illustrated by example.[4] Also, the means by which the estimators can be computed using readily available nonlinear regression programs is mentioned.

2. The statistical model

The structural model is the simultaneous system consisting of M (possibly) nonlinear equations in implicit form

$$q_\alpha(y, x, \theta_\alpha^*) = 0 \qquad (\alpha = 1, 2, \ldots, M)$$

where y is an M by 1 vector of endogenous variables, x is a k by 1 vector of exogenous variables, and θ_α^* is a p_α by 1 vector of unknown parameters contained in the compact parameter space Θ_α. An example is:

$$a_0^* + a_1^* \ln y_1 + a_2^* \ln y_2 + a_3^* x = 0 ,$$

$$b_0^* + b_1^* y_1 + b_2^* y_2 + b_3^* x = 0 .$$

The correspondence with the above notation is $\theta_1 = (a_0, a_1, a_2, a_3)$ and $\theta_2 = (b_0, b_1, b_2, b_3)$.

It is assumed throughout that all a priori within-equation parametric restrictions and the normalization rule have been eliminated by reparameterization.

For the example, the a priori information $a_0^* + a_1^* = 0$ and the normalization rule $a_2^* = 1$ would be incorporated into the first equation by rewriting it as

$$a_1^* (\ell n\, y_1 - 1) + \ell n\, y_2 + a_3^* x = 0 ;$$

whence, $\theta_1 = (a_1, a_3)$. In applications this convention will rarely be inconvenient because most a priori within-equation information consists of exclusion restrictions; reparameterization amounts to no more than simple omissions when writing the equation. A priori across-equation restrictions are explicitly incorporated into the estimation procedure and theoretical development.

The formal assumptions, set forth in Section 4, indirectly imply that $q_\alpha(y, x, \theta_\alpha)$ cannot depend trivially on any component of θ_α; i.e., $q_\alpha(y, x, \theta_\alpha)$ must actually vary with each component of θ_α. However, y and x include all variables in the model and trivial dependencies are permitted.

Due to slight errors in specification and/or errors of observation, the data (y_t, x_t) available for estimation of the structural parameters are assumed to follow the statistical model

$$q_\alpha(y_t, x_t, \theta_\alpha^*) = e_{\alpha t} \quad (\alpha = 1, 2, \ldots, M; \quad t = 1, 2, \ldots, n)$$

where the M-variate errors

$$e_t = (e_{1t}, e_{2t}, \ldots, e_{Mt})'$$

are independent and identically distributed each having mean vector zero and positive definite variance-covariance matrix Σ.[5]

Instrumental variables--a sequence of K by 1 vectors $\{z_t\}$--are assumed available for estimation. Mathematically, the instrumental variables need only satisfy the regularity conditions. However, it seems reasonable to at least insist, as does Fisher (1966, Ch. 5), that instrumental variables be restricted to functions of the exogenous variables; viz., $z_t = Z(x_t)$. The usual convention in linear systems of simultaneous equations is to require that the instrumental variables be the exogenous variables themselves; $z_t = x_t$. Unfortunately, such a restriction would destroy identification in many nonlinear systems. On the other hand, permitting wide latitude in the choice of instruments for the purpose of identification and estimation introduces a disturbing element of variability in results--different results will obtain for various choices of instruments even though the data, model specification, and normalization rules are the same.

Fisher (1966, p. 131) states: "Since the model is supposed to be a theoretical statement of the way in which the endogenous variables are determined, given the predetermined variables and disturbances, we shall assume that it does so." In the present context, this requirement translates into the existence of a vector valued function $Y(x, e)$ such that $y_t = Y(x_t, e_t)$, i.e., the existence of a reduced form. This requirement, as the requirement that $z_t = Z(x_t)$, is unnecessary for the mathematical development in the later sections. However, as will be seen in the next section, the existence of $Y(x, e)$ and the insistence that the instruments be functions of x certainly makes attainment of the formal regularity conditions more plausible.[6] It should be emphasized, however, that the user is not required to find $Y(x, e)$ in closed form or even to be able to compute $Y(x, e)$ for given (x, e) using numerical methods in order to apply the statistical methods set forth here.

One detail should be mentioned before describing the estimation procedure. The sequence of exogenous variables $\{x_t\}$ is assumed to be either a sequence of constants or, if some coordinates are random variables, the assumptions and results are conditional probability statements given the sequence $\{x_t\}$--lagged endogenous variables are not permitted. Nevertheless, the proofs, in fact, only require the stated regularity conditions with the assumptions on the error process $\{e_t\}$ modified to read: The conclusions of Lemma A.3 are satisfied. Be that as it may, there are no results available in the probability literature, to the author's knowledge, which yield the conclusions of Lemma A.3 for lagged endogenous instrumental variables generated by a system of simultaneous, nonlinear, implicit equations. (The requisite theory for the linear case is spelled out in Section 10.1 of Theil (1971).) The reader who feels that these modified assumptions are reasonable, in the absence of such results, may include lagged endogenous variables as components of x_t.

3. Estimation procedure

The observables, corresponding to a trial value of θ_α, may be written in convenient vector form as:

$$q_\alpha(\theta_\alpha) = (q_\alpha(y_1, x_1, \theta_\alpha), q_\alpha(y_2, x_2, \theta_\alpha), \ldots, q_\alpha(y_n, x_n, \theta_\alpha))' \quad (n \times 1) ,$$

$$Z = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_n' \end{bmatrix} \qquad (n \times K) .$$

The first step of the procedure is to obtain the nonlinear two stage least squares estimators $\hat{\theta}_\alpha$ by minimizing

$$S_\alpha(\theta_\alpha) = (1/n)\, q_\alpha'(\theta_\alpha)\, Z(Z'Z)^{-1}Z'\, q_\alpha(\theta_\alpha)$$

over $(H)_\alpha$, equation by equation.

The second step is to estimate the elements $\sigma_{\alpha\beta}$ of the variance-covariance matrix Σ by

$$\hat{\sigma}_{\alpha\beta} = (1/n)\, q_\alpha'(\hat{\theta}_\alpha)\, q_\beta(\hat{\theta}_\beta) \quad (\alpha, \beta = 1, 2, \ldots, M) .$$

To carry out the next step, "stack" the parameters and observables as:

$$\theta = (\theta_1', \theta_2', \ldots, \theta_M')' \quad (p = \sum_{\alpha=1}^{M} p_\alpha \times 1) ,$$

$$q(\theta) = (q_1'(\theta_1), q_2'(\theta_2), \ldots, q_M'(\theta_M))' \quad (nM \times 1) .$$

The third step is to obtain the nonlinear three stage least squares estimator by minimizing

$$S(\theta) = (1/n)\, q'(\theta)(I \otimes Z)(\hat{\Sigma} \otimes Z'Z)^{-1}(I \otimes Z')\, q(\theta)$$

over $(H) = X_{\alpha=1}^{M} (H)_\alpha$ where I is the M by M identity matrix.

Define:

$\nabla_\alpha\, q_\alpha(y,x,\theta_\alpha)$ = the p_α by 1 vector whose $i^{\underline{th}}$ element is $(\partial/\partial\theta_{i\alpha})\, q_\alpha(y, x, \theta_\alpha)$,

$Q_\alpha(\theta_\alpha)$ = the n by p_α matrix whose $t^{\underline{th}}$ row is $\nabla_\alpha'\, q_\alpha(y_t, x_t, \theta_\alpha)$,

$$Q(\theta) = \text{diag}\,(Q_1(\theta_1), Q_2(\theta_2), \ldots, Q_M(\theta_M)) \quad (nM \times p) .$$

The fourth and final step is to obtain the inverse of the matrix

$$\tilde{\Omega} = (1/n)\,[Q'(\tilde{\theta})\,(I \otimes Z)\,(\hat{\Sigma} \otimes Z'Z)^{-1}\,(I \otimes Z')\,Q(\tilde{\theta})]\ .$$

In Section 5 it is shown that $\sqrt{n}\,(\tilde{\theta} - \theta^*)$ is distributed asymptotically as a p-variate normal with a variance-covariance matrix for which $(\tilde{\Omega})^{-1}$ is a strongly consistent estimator.

One may wish to impose restrictions across equations in the third stage. In the present context, the most convenient way to represent these restrictions is by reparameterization. Let ρ be an r by 1 vector of new parameters and let g_α be a p_α by 1 vector valued function relating the original parameters to the new parameters according to $\theta_\alpha = g_\alpha(\rho)$. It is assumed that $r \leq p$ and that ρ is contained in the compact parameter space P.

Define:[7]

$$g(\rho) = (g_1'(\rho),\ g_2'(\rho),\ \ldots,\ g_M'(\rho))' \quad (p \times 1)\ ,$$

$$\nabla_\rho\, g_{i\alpha}(\rho) = \text{the } r \text{ by } 1 \text{ vector whose } j^{\underline{\text{th}}} \text{ element is } (\partial/\partial\rho_j)\, g_{i\alpha}(\rho),$$

$$G_\alpha(\rho) = \text{the } p_\alpha \text{ by } r \text{ matrix whose } i^{\underline{\text{th}}} \text{ row is } \nabla_\rho'\, g_{i\alpha}(\rho),$$

$$G(\rho) = (G_1'(\rho),\ G_2'(\rho),\ \ldots,\ G_M'(\rho))' \quad (p \times r)\ .$$

The third step of the procedure is modified to read: Obtain the estimator $\tilde{\rho}$ by minimizing

$$S(g(\rho)) = (1/n)\, q'(g(\rho))\,(I \otimes Z)\,(\hat{\Sigma} \otimes Z'Z)^{-1}\,(I \otimes Z')\, q(g(\rho))\ .$$

The fourth step is modified to read: Obtain the inverse of the matrix

$$\tilde{G}'\tilde{\Omega}\tilde{G} = (1/n)\,[G'(\tilde{\rho})\,Q'(g(\tilde{\rho}))\,(I \otimes Z)\,(\hat{\Sigma} \otimes Z'Z)^{-1}\,(I \otimes Z')\,Q(g(\tilde{\rho}))\,G(\tilde{\rho})]\ .$$

In Section 5 it is shown that $\sqrt{n}\,(\tilde{\rho} - \rho^*)$ is asymptotically distributed as an r-variate normal with mean vector zero and a variance-covariance matrix for which $(\tilde{G}'\tilde{\Omega}\tilde{G})^{-1}$ is a strongly consistent estimator.

If desired, results obtained subject to the restrictions $\theta = g(\rho)$ may be reported in terms of the original parameters: Put $\bar{\theta} = g(\tilde{\rho})$. The estimator $\bar{\theta}$ is strongly consistent for θ^* and $\sqrt{n}\,(\bar{\theta} - \theta^*)$ is asymptotually distributed as a p-variate normal with mean vector zero and a variance covariance matrix for which $G(\tilde{\rho})\,(\tilde{G}'\tilde{\Omega}\tilde{G})^{-1}\,G'(\tilde{\rho})$ is a strongly consistent estimator. This matrix will be singular when $r < p$.

The computations may be performed using either Hartley's modified Gauss-Newton method or Marquardt's algorithm. A program using one of these algorithms is the preferred choice for the computations because they: are widely available, perform well without requiring the user to supply second partial derivatives, and will print the asymptotic variance-covariance matrix needed to obtain standard errors. Our discussion of how they are to be used depends on the notation in Section 1 of Gallant (1975b) and the description of these algorithms in Section 3 of the same reference; the reader will probably need this reference in hand to read the next few paragraphs.

To use the algorithms as they are usually implemented, factor the MK by MK matrix $(\hat{\Sigma} \otimes Z'Z)^{-1}$ to obtain R such that $R'R = (\hat{\Sigma} \otimes Z'Z)^{-1}$; then use the algorithms with these substitutions: $y = 0$, $f(\theta) = R(I \otimes Z')\,q(\theta)$, and $F(\theta) = R(I \otimes Z')\,Q(\theta)$. Most implementations of these algorithms expect the user to supply a PL/1 or FORTRAN subroutine to compute a row of $f(\theta)$ or $F(\theta)$ given θ and the row index. (To be given an input vector x_t of $f(x_t, \theta)$ is equivalent to having been given the row index.) This is needlessly costly, for the present purpose, because it requires recomputation of $q(\theta)$ and $Q(\theta)$ for every row. The difficulty can be circumvented since, for each trial value of θ, the subroutine is always called sequentially with

the row index running from $I = 1, 2, \ldots, nM$. The user-supplied subroutine can be written so as to compute and store $f(\theta)$ and $F(\theta)$ when $I = 1$ and then supply entries from these arrays for the subsequent calls: $I = 2, \ldots, nM$. The matrix $\hat{C}$ printed by the program will satisfy $n\hat{C} = (\tilde{\Omega})^{-1}$; that is, the diagonal entries of $\hat{C}$ are the estimated variances of $\tilde{\theta}$, whereas the diagonal entries of $(\tilde{\Omega})^{-1}$ are the estimated variances of $\sqrt{n}\,\tilde{\theta}$. The standard errors for $\tilde{\theta}$ printed by the program will have been computed from the diagonal entries of $s^2\,\hat{C}$ and may be used after division by $\sqrt{s^2}$; s^2 is, in this instance, a strongly consistent estimator of one. A good choice for the starting value is $\theta_0 = \hat{\theta} = (\hat{\theta}_1', \hat{\theta}_2', \ldots, \hat{\theta}_M')'$.

Similarly, one may use these algorithms to compute $\tilde{\rho}$ by putting $y = 0$, $f(\rho) = R(I \otimes Z')\, q(g(\rho))$, and $F(\rho) = R(I \otimes Z')\, Q(g(\rho))\, G(\rho)$; the matrix $\hat{C}$ will satisfy $(\tilde{G}'\tilde{\Omega}\tilde{G})^{-1} = n\hat{C}$. To compute the nonlinear two stage least squares estimators $\hat{\theta}_\alpha$, factor the K by K matrix $(Z'Z)^{-1}$ to obtain R such that $R'R = (Z'Z)^{-1}$ then put $y = 0$, $f(\theta) = R\ Z'\ q_\alpha(\theta_\alpha)$, and $F(\theta) = R\ Z'\ Q_\alpha\ (\theta_\alpha)$.

4. Exogenous variables, assumptions, and identification

The assumptions, given later in this section, used to obtain the asymptotic theory of the nonlinear three stage least squares estimator, require that sample moments such as $(1/n) \sum_{t=1}^{n} z_t\, q_\alpha(y_t, x_t, \theta_\alpha)$ converge almost surely uniformly in θ_α. It is obligatory, therefore, to set forth reasonable conditions on the exogenous variables so that such limiting behavior is achieved in applications. To illustrate one of the problems involved, assume that the errors

$$e_{\alpha t} = q_\alpha(y_t, x_t, \theta_\alpha^*)$$

are normally distributed. This implies that the function $q_\alpha(y, x, \theta_\alpha)$ cannot be bounded, which effectively rules out the use of weak convergence of measures as a means of imposing conditions on the sequence of exogenous variables $\{x_t\}$ (Malinvaud, 1970). We are led, therefore, to define a similar but stronger notion.

Definition. A sequence $\{v_t\}$ of points from a Borel set $\mathcal{V}$ is said to generate convergent Cesàro sums with respect to a probability measure ν defined on the Borel subsets of $\mathcal{V}$ if, for every real valued continuous function f with $\int |f(v)| d\nu(v) < \infty$, the limit

$$\lim_{n \to \infty} (1/n) \sum_{t=1}^{n} f(v_t) = \int f(v)\, d\nu(v) .$$

There are two simple ways to generate such sequences: fixed in repeated samples and random sampling from a probability measure. In the former case, where a sequence $\{x_t\}$ is generated from the points $a_0, a_1, \ldots, a_{T-1}$ according to

$$x_t = a_{(t \bmod T)}$$

the relevant measure is defined by

$$\mu(A) = (1/T) \sum_{t=0}^{T-1} I_A(a_t)$$
$$= \text{the proportion of the } a_t \text{ in } A$$

where I_A is the indicator function of the set A. In the latter case, where a sequence $\{e_t\}$ is generated by random sampling from a probability measure P, the relevant measure is P itself; almost every realization of $\{e_t\}$ will generate convergent Cesàro sums with respect to P by the Strong Law of Large Numbers. The joint sequence of exogenous variables and errors

$$v_t = (x_t, e_t) \quad (t = 1, 2, \ldots)$$

will generate convergent Cesàro sums with respect to the product measure

$$\nu(A) = \iint I_A(x, e)\, d\mu(x)\, dP(e)$$
$$= (1/T) \sum_{t=0}^{T-1} \int I_A(a_t, e)\, dP(e) .$$

There are, of course, many other ways to generate a sequence of exogenous variables $\{x_t\}$ such that the joint sequence $\{v_t = (x_t, e_t)\}$ generates convergent Cesàro sums with respect to a product measure $\nu = \mu \times P$. Another example, "near replicates," is obtained by letting the $\{a_{ij}\}$ be sequences converging to a_j $(j = 1, 2, \ldots, T)$ as i tends to infinity and putting $x_t = a_{ij}$ where $t = T(i - 1) + j$.

The attainability of the limit assumptions obtains from:

Theorem 1. Let $f(v, \theta)$ be a real valued continuous function on $\mathcal{V} \times \textcircled{H}$ where $\mathcal{V}$ is a Borel set and $\textcircled{H}$ is compact. Let $\{v_t\}$ generate convergent Cesàro sums with respect to a probability measure ν defined on the Borel subsets of $\mathcal{V}$. Let $h(v)$ be a real valued continuous function on $\mathcal{V}$ such that $|f(v, \theta)| \leq h(v)$ and $\int h(v)\, d\nu(v) < \infty$. Then:

i) The sum $(1/n) \sum_{t=1}^{n} f(v_t, \theta)$ converges to the integral $\int f(v, \theta)\, d\nu(v)$ uniformly for all θ in $\textcircled{H}$.

ii) The sum $(1/n) \sum_{t=1}^{n} \sup_{\textcircled{H}} |f(v_t, \theta)|$ is bounded.

Proof. The proof of the first conclusion is, word for word, the same as for Jennrich's (1969) Theorem 1, excepting that $(1/n) \sum_{t=1}^{n} f(v_t, \theta)$ replaces $\int g(x, \theta)\, dF_n(x)$, $\int f(v, \theta)\, d\nu(v)$ replaces $\int g(x, \theta)\, dF(x)$, and the $\overline{\lim}$ is obtained from the definition of convergent Cesàro sums with respect to ν rather than the Helly-Bray theorem. The second conclusion

follows from the bound $\sup_{(H)} |f(v, \theta)| \leq h(v)$ and the assumption that $(1/n) \sum_{t=1}^{n} h(v_t)$ converges. □

Recall that earlier the existence of a reduced form $Y(x, e)$ such that $y_t = Y(x_t, e_t)$ was assumed and that the instruments were assumed to be related to the exogenous variables according to $z_t = Z(x_t)$. If it is further assumed that $Y(x, e)$, $Z(x)$, and $q_\alpha(y, x, \theta_\alpha)$ are continuous, then, by Theorem 1, the uniform limit of

$$(1/n) \sum_{t=1}^{n} z_t \, q_\alpha(y_t, x_t, \theta_\alpha)$$

exists if $\{v_t = (x_t, e_t)\}$ generates convergent Cesàro sums with respect to ν and there is a ν-integrable continuous function $h(v)$ which dominates

$$f(v, \theta_\alpha) = Z_{\ell}(x) \, q_\alpha(Y(x, e), x, \theta_\alpha) .$$

We illustrate with the example, imposing $a_2^* = 0$ and the normalization rules $a_1^* = 1$, $b_2^* = 1$:

$$a_0^* + \ln y_{1t} + a_3^* x_t = e_{1t} ,$$

$$b_0^* + b_1^* y_{1t} + y_{2t} + b_3^* x_t = e_{2t} .$$

The reduced form is:

$$Y_1(x, e) = \exp(e_1 - a_0^* - a_3^* x) ,$$

$$Y_2(x, e) = e_2 - b_0^* - b_1^* \exp(e_1 - a_0^* - a_3^* x) - b_3^* x .$$

Assume normally distributed errors and that $\{x_t\}$ is fixed in repeated samples due to replication of the points $(0, 1, 2, 3)$. To evaluate, say, $(1/n) \sum_{t=1}^{n} x_t \, q_1(y_t, x_t, \theta_1)$ consider

$$f(v, \theta_1) = x\, q_1 (Y(x, e), x, \theta_1)$$
$$= x\, a_0 + x\, e_1 - x\, a_0^* - a_3^* x^2 + a_3 x^2$$

which is dominated by the continuous function

$$h(v) = 3|e_1| + \sup_{(a_0, a_3)} [3|a_0 - a_0^*| + 9|a_3 - a_3^*|] ;$$

recall that $\textcircled{H}_1$ is compact. Thus, the sum $(1/n) \sum_{t=1}^{n} x_t\, q_1(y_t, x_t, \theta_1)$ has uniform limit

$$\int\int (x\, a_0 + x\, e_1 - x\, a_0^* - a_3^* x^2 + a_3 x^2)\ d\mu(x)\ dP(e)$$
$$= (1/4) \sum_{x=0}^{3} \int (x\, a_0 + x\, e_1 - x\, a_0^* - a_3^* x^2 + a_3 x^2)\, n(e; 0, \Sigma)\ de$$
$$= (1.5)\ (a_0 - a_0^*) + (3.5)\ (a_3 - a_3^*) .$$

This limit is an almost sure uniform limit because the Strong Law of Large Numbers was used to deduce that $\{e_t\}$ generated convergent Cesàro sums; hence, there is an exceptional set E occurring with probability zero corresponding to realizations for which the conclusion of the Strong Law of Large Numbers fails to hold.

The regularity conditions imposed on the system are:

Assumptions. The moment matrix of the instrumental variables $(1/n)\ Z'Z$ converges to a positive definite matrix P. The errors $\{e_t\}$ are independently and identically distributed each with mean vector zero and positive definite variance-covariance matrix Σ. Each parameter space $\textcircled{H}_\alpha$ is compact; the true parameter value θ_α^* is contained in an open sphere O_α which is, in turn, contained in $\textcircled{H}_\alpha$ Each function $q_\alpha(y, x, \theta_\alpha)$ and its first and second partial derivatives with respect to θ_α are continuous in θ_α

for fixed (y, x). The Cesàro sums $(1/n) \sum_{t=1}^{n} q_\alpha(y_t, x_t, \theta_\alpha)\, q_\beta(y_t, x_t, \theta_\beta)$, $(1/n) \sum_{t=1}^{n} z_t\, q_\alpha(y_t, x_t, \theta_\alpha)$, and $(1/n) \sum_{t=1}^{n} z_t\, (\partial/\partial \theta_{i\alpha})\, q_\alpha(y_t, x_t, \theta_\alpha)$ converge almost surely uniformly in $(\theta_\alpha, \theta_\beta)$. The sums $(1/n) \sum_{t=1}^{n} \sup_{\textcircled{H}_\alpha} |z_{\ell t}\, (\partial/\partial \theta_{i\alpha})\, q_\alpha(y_t, x_t, \theta_\alpha)|$ and $(1/n) \sum_{t=1}^{n} \sup_{\textcircled{H}_\alpha} |z_{\ell t}\, (\partial^2/\partial \theta_{i\alpha}\, \partial \theta_{j\alpha})\, q_\alpha(y_t, x_t, \theta_\alpha)|$ are bounded almost surely ($\ell = 1, \ldots, K$; $i, j = 1, \ldots, p_\alpha$; $\alpha = 1, \ldots, M$). The p by p matrix

$$\Omega = \begin{bmatrix} \sigma^{11} A_{11} & \sigma^{12} A_{12} & \cdots & \sigma^{1M} A_{1M} \\ \sigma^{21} A_{21} & \sigma^{22} A_{22} & \cdots & \sigma^{2M} A_{2M} \\ \vdots & \vdots & & \vdots \\ \sigma^{M1} A_{M1} & \sigma^{M2} A_{M2} & \cdots & \sigma^{MM} A_{MM} \end{bmatrix}$$

is nonsingular where the $\sigma^{\alpha\beta}$ are the elements of Σ^{-1} and

$$A_{\alpha\beta} = \lim_{n \to \infty} [(1/n)\, Q'_\alpha(\theta^*_\alpha) Z]\, [(1/n) Z'Z]^{-1}\, [(1/n)\, Z' Q_\beta(\theta^*_\beta)] .$$

Identification. The structural equation

$$q_\alpha(y_t, x_t, \theta^*_\alpha) = e_{\alpha t}$$

from a system satisfying the assumptions is said to be identified by the instruments z_t if the only solution of the almost sure limit

$$\lim_{n \to \infty} (1/n) \sum_{t=1}^{n} z'_t\, q_\alpha(y_t, x_t, \theta_\alpha) = 0$$

is $\theta_\alpha = \theta^*_\alpha$.

Note that since $\textcircled{H}_\alpha$ has been previously constrained to be compact this is a local, not global, concept of identifiability. That is, $\textcircled{H}_\alpha$ is essentially of the form $\{\theta_\alpha: \ ||\theta_\alpha - \theta^*_\alpha|| < B\}$; if there is a point θ_α in $\mathbb{R}^{p_\alpha}$ satisfying the test criterion it is ruled out if $||\theta_\alpha - \theta^*_\alpha|| > B$.

The bound B may be reduced, if necessary, so as to rule out all extraneous, isolated solutions.

This definition is compatible with the instrumental variables approach for linear systems (Theil, 1971, Sec. 9.4) and Fisher's (1966, Ch. 5) approach to systems linear in the parameters but nonlinear in the variables because a system which is globally identified must be locally identified; recall that a priori within-equation restrictions and a normalization rule have already been incorporated into $q_\alpha(y_t, x_t, \theta_\alpha)$ by reparameterization.

To illustrate, we examine the identification status of the second equation of the example $(a_2^* = 0, a_1^* = 1, b_2^* = 1)$ with respect to the instrumental variables $z_t = (1, x_t, x_t^2)'$. The almost sure limit of $(1/n) \sum_{t=1}^{n} z_t' q_2(y_t, x_t, \theta_2)$ is

$$\begin{bmatrix} 1 & c \sum_{x=0}^{3} e^{-a_3^* x} & 1.5 \\ 1.5 & c \sum_{x=0}^{3} x\, e^{-a_3^* x} & 3.5 \\ 3.5 & c \sum_{x=0}^{3} x^2 e^{-a_3^* x} & 9 \end{bmatrix} \begin{bmatrix} b_0 - b_0^* \\ b_1 - b_1^* \\ b_3 - b_3^* \end{bmatrix}$$

where $c = (1/4) \exp(\sigma_{11}/2 - a_0^*)$. The equation will be identified by the instruments when the 3 by 3 matrix W_2 in the limit equations has full rank--when $a_3^* \neq 0$. Note that if $a_3^* = 0$ the equation will remain underidentified even if more instruments are added; note also that the restriction $z_t = x_t$, customary in linear systems, would destroy identification.

This illustrates a point which should not be overlooked when considering identification for systems linear in the parameters but nonlinear in the variables:

$$q_\alpha(y_t, x_t, \theta_\alpha) = B'(y_t, x_t) \begin{bmatrix} 1 \\ \theta_\alpha \end{bmatrix} .$$

(Recall that the normalization rule has been incorporated into $q_\alpha(y_t, x_t, \theta_\alpha)$.) The K by $p_\alpha + 1$ almost sure limit matrix

$$W = [w_1 \vdots W_2] = \lim_{n \to \infty} (1/n) \sum_{t=1}^{n} z_t B'(y_t, x_t)$$

appearing in the equations for checking identification

$$w_1 + W_2 \theta_\alpha = W_2(\theta_\alpha - \theta_\alpha^*)$$

will, in general, depend on all parameters (θ^*, Σ) of the system. Thus, one is at risk in asserting that W_2 has full rank without giving the matter some thought; see, e.g., Fisher (1966, Ch. 5).

Except for the very simplest nonlinear models, it is all but impossible to obtain the almost sure limits for checking identification. The best one can do is to try not to overlook situations affecting identification which can be spotted by inspection.

To illustrate, consider how one could deduce that $a_3^* = 0$ destroys identification in the example without computing the limit. If $a_3^* = 0$ the structural equations are

$$a_0^* + \ell n\, y_{1t} = e_{1t} ,$$

$$b_0^* + b_1^* y_{1t} + y_{2t} + b_3^* x_t = e_{2t} .$$

Observe from the first equation that $\ell n\, y_{1t}$ is independently and identically distributed; hence, y_{1t} itself must be independently and identically distributed. If $z_{\ell t}$ is an instrumental variable which satisfies the assumptions, it must follow that the almost sure limit of $(1/n) \sum_{t=1}^{n} z_{\ell t}\, y_{1t}$

is $[E(y_1)] \lim_{n\to\infty} (1/n) \sum_{t=1}^{n} z_t$; see Theorem 3 of Jennrich (1969). Consequently, the first and second columns of

$$W_2 = \lim_{n\to\infty} (1/n) \sum_{t=1}^{n} z_t(1, y_{1t}, x_t)$$

must be proportional and W_2 cannot have full rank.[8]

A trap has been set! A natural inclination would be to test the hypothesis H: $a_3^* = 0$ for the system

$$a_0^* + \gamma_1 y_{1t} + a_3^* x_t = e_{1t} ,$$

$$b_0^* + b_1^* y_{1t} + y_{2t} + b_3^* x_t = e_{2t}$$

using the nonlinear three stage least squares estimate $\tilde{a}_3$ divided by its estimated standard error. But, if the hypothesis is true, then identification in the second equation is destroyed. Consequently, the assumptions used to obtain the asymptotic null distribution of the test statistic are not satisfied. The null hypothesis may, however, be tested using the nonlinear two stage estimate for the first equation; see Amemiya (1974).

In the case when the restrictions $\theta = g(\rho)$ are imposed on the parameters of the system, the following additional assumptions are required.

Assumptions. (Continued) The function $g(\rho)$ is a twice continously differentiable mapping of a compact set P into the parameter space $H = \chi_{\alpha=1}^{n} H_\alpha$. There is only one point ρ^* in P which satisfies $g(\rho) = \theta^*$, and ρ^* is contained in an open sphere O which is, in turn, contained in P. The p by r matrix $G(\rho^*)$ has rank r.

5. Strong consistency and asymptotic normality

Two theorems establishing the strong consistency and asymptotic normality of the estimator $\tilde{\rho}$ are proved in this section; the strong

consistency and asymptotic normality of $\tilde{\theta}$ follow as corollaries to these two theorems by taking $g(\rho)$ to be the identity transformation. In order to simplify the notation in this section, the function $q_\alpha(\theta_\alpha)$ will be written as q_α when it is evaluated at the true value of the parameter $\theta_\alpha = \theta_\alpha^*$. Similarly, we will write Q_α for $Q_\alpha(\theta_\alpha^*)$, q for $q(\theta^*)$, Q for $Q(\theta^*)$, and G for $G(\rho^*)$.

Theorem 2. If the assumptions listed in Section 4 are satisfied and each equation in the system is identified, then $\tilde{\rho}$ converges almost surely to ρ^*.

Proof. The matrix $[\hat{\Sigma} \otimes (1/n)\, Z'Z]^{-1}$ converges almost surely to $(\Sigma \otimes P)^{-1}$ by the assumptions and Lemma A.4. The term $(1/n)\, q'(g(\rho))\, (I \otimes Z)$ converges almost surely uniformly in ρ in consequence of the assumptions and Lemma A.1. Thus, $S[g(\rho)]$ converges almost surely uniformly in ρ to, say, $\bar{S}[g(\rho)]$.

Consider a sequence of points $\{\tilde{\rho}_n\}$ minimizing $S(g(\rho))$ corresponding to a realization of the error process $\{e_t\}$. Since $\underline{P}$ is compact there is at least one limit point ρ^0 and one subsequence $\{\rho_{n_m}\}$ such that $\lim_{m\to\infty} \tilde{\rho}_{n_m} = \rho^0$. Excepting realizations corresponding to an event E, which occurs with probability zero, $S(g(\rho))$ converges uniformly whence

$$0 \leq \bar{S}(g(\rho^0)) = \lim_{m\to\infty} S(g(\tilde{\rho}_{n_m})) \leq \lim_{m\to\infty} S(\theta^*) = \bar{S}(\theta^*) .$$

But $\bar{S}(\theta^*) = \lim_{m\to\infty} [(1/n_m)\, q'(\theta^*)\, (I \otimes Z)]\, (\Sigma \otimes P)^{-1}\, [(1/n_m)\, (I \otimes Z')\, q(\theta^*)]$ which equals zero by Lemma A.3--excepting realizations in E. Consequently, excepting realizations in E, it follows that $\bar{S}(g(\rho^0)) = 0$ which implies $\lim_{m\to\infty} (1/n_m)\, (I \otimes Z')\, q(g(\rho^0)) = 0$ because $(\Sigma \otimes P)$ is positive definite. The assumption that every equation in the system is identified implies that

$g(\rho^0) = \theta^*$ which, by assumption, implies $\rho^0 = \rho^*$. Thus, excepting realizations in E, $\{\tilde{\rho}_n\}$ has only one limit point which is ρ^*. □

Corollary. If the assumptions listed in Section 4 are satisfied and each equation in the system is identified, then $\tilde{\theta}$ converges almost surely to θ^*.

Theorem 3. If the assumptions listed in Section 4 are satisfied and each equation in the system is identified, then $\sqrt{n}\ (\tilde{\rho} - \rho^*)$ converges in distribution to an r-variate normal with mean vector zero and variance-covariance matrix $(G'\Omega G)^{-1}$. The matrix

$$(\tilde{G}'\tilde{\Omega}\tilde{G}) = (1/n)\ G'(\tilde{\rho})\ Q'(g(\tilde{\rho}))\ (I \otimes Z)\ (\hat{\Sigma} \otimes Z'Z)^{-1}\ (I \otimes Z')\ Q(g(\tilde{\rho}))\ G(\tilde{\rho})$$

converges almost surely to $G'\Omega G$.

Proof. Define $\dot{\rho} = \tilde{\rho}$ if $\tilde{\rho}$ is in O and $\dot{\rho} = \rho^*$ if $\tilde{\rho}$ is not in O. Set $\dot{\theta} = g(\dot{\rho})$. Since $\sqrt{n}\ (\dot{\rho} - \tilde{\rho})$ converges almost surely to zero by Theorem 2, it will suffice to prove the theorem for $\dot{\rho}$.

The first order Taylor's expansion of $q(\dot{\theta})$ may be written as $q(\dot{\theta}) = q + Q\ G(\dot{\rho} - \rho^*) + H(\dot{\rho} - \rho^*)$ where H is the nM by r matrix $H = (H_1', H_2', \ldots, H_M')'$; the t^{th} row of the n by r submatrix H_α is $(1/2)\ (\dot{\rho} - \rho^*)'\ \nabla_\rho^2\ q_\alpha(y_t, x_t, g_\alpha(\bar{\rho}))$ where $\bar{\rho}$ varies with t and $\dot{\rho}$ and lies on the line segment joining $\dot{\rho}$ to ρ^*. A typical element of $(1/n)\ Z'H_\alpha$ is a finite sum of products composed of the terms $(\dot{\rho}_i - \rho_i^*)$, the first and second partial derivatives of $g_{i\alpha}(\rho)$ evaluated at $\rho = \bar{\rho}$, $(1/n) \sum_{t=1}^{n} z_{(t} (\partial/\partial\theta_{j\alpha})\ q_\alpha(y_t, x_t, g_\alpha(\bar{\rho}))$, and $(1/n) \sum_{t=1}^{n} z_{(t}(\partial^2/\partial\theta_{i\alpha}\partial\theta_{j\alpha})\ q_\alpha(y_t, x_t, g_\alpha(\bar{\rho}))$. As a direct consequence of the almost sure bounds imposed on these latter terms by assumption and the

almost sure convergence of $\dot{\rho}$ to ρ^*, it follows that $(1/n)\ (I \otimes Z')H$ converges almost surely to the zero matrix.

The vector $\tilde{\rho}$ minimizes $S(g(\rho))$ whence

$$(\sqrt{n}/2)\ \nabla_\rho\ S(\dot{\theta}) = (1/\sqrt{n})\ G'(\dot{\rho})\ Q'(\dot{\theta})\ (I \otimes Z)\ (\hat{\Sigma} \otimes Z'Z)^{-1}\ (I \otimes Z')\ q(\dot{\theta})$$

converges almost surely to the zero vector. By substituting the Taylor's expansion of $q(\dot{\theta})$ in this expression, the right-hand side becomes

$$(1/\sqrt{n})\ G'(\dot{\rho})\ B(\dot{\theta})\ U + G'(\dot{\rho})\ B(\dot{\theta})\ [(1/n)\ (I \otimes Z')\ Q\ G + (1/n)\ (I \otimes Z')H]\sqrt{n}\ (\dot{\rho} - \rho^*)$$

where B and U are defined in Lemmas A.2 and A.3, respectively. In consequence of the almost sure convergence of $\dot{\rho}$ to ρ^*, the continuity of $g(\rho)$ and $G(\rho)$, Lemma A.2, and Lemma A.3 the first term, $(1/\sqrt{n})\ G'(\dot{\rho})\ B(\dot{\theta})\ U$, converges in distribution to an r-variate normal with mean vector zero and variance-covariance matrix $G'\overline{B}\ (\Sigma \otimes P)\ \overline{B}'G = G'\Omega G$. Similarly, the matrix premultiplying $\sqrt{n}\ (\dot{\rho} - \rho^*)$ in the second term converges almost surely to the nonsingular matrix $G'\Omega G$ in consequence of the (trivial) almost sure convergence of ρ^* to ρ^*, Lemma A.2, and our previous remarks concerning $(1/n)\ (I \otimes Z')H$. It follows, by Slutzky's theorem, that $\sqrt{n}\ (\dot{\rho} - \rho^*)$ converges in distribution to the r-variate normal with mean vector zero and variance-covariance matrix $(G'\Omega G)^{-1}\ (G'\Omega G)\ (G'\Omega G)^{-1} = (G'\Omega G)^{-1}$.

The second conclusion of the theorem is a direct consequence of Theorem 2 and Lemma A.2. □

Corollary. If the assumptions listed in Section 4 are satisfied and each equation in the system is identified, then $\sqrt{n}\ (\tilde{\theta} - \theta^*)$ converges in distribution to a p-variate normal distribution with mean vector zero and variance-covariance matrix Ω^{-1}. The matrix

$$\tilde{\Omega} = (1/n)\ Q'(\tilde{\theta})\ (I \otimes Z)\ (\Sigma \otimes Z'Z)^{-1}\ (I \otimes Z')\ Q(\tilde{\theta})$$

converges almost surely to Ω.

6. Asymptotic relative efficiency

The asymptotic variance-covariance matrix of the nonlinear two stage least squares estimator is $A^{-1} T A^{-1}$ given by Lemma A.4. The three stage nonlinear least squares estimator is more efficient than the nonlinear two stage least squares estimator in the sense that the difference between their asymptotic variance-covariance matrices $D = A^{-1} T A^{-1} - \Omega^{-1}$ is a positive semi-definite matrix.

To see that this is so, let $X = (I \otimes Z(Z'Z)^{-1}Z')Q$ and $V = \Sigma \otimes I$. It can be verified by straightforward matrix algebra that

$$D = \lim_{n \to \infty} n \left[(X'X)^{-1} X' V X(X'X)^{-1} - (X' V^{-1} X)^{-1}\right] .$$

Viewing this expression in the context of the regression equations $y = X\beta + u$ where $C(u, u') = V$, the implication of Aitken's theorem is that the matrix in brackets is positive semi-definite. As a consequence, D is positive semi-definite.

Footnotes

[1] See Eisenpress and Greenstadt (1966, Sec. 6).

[2] Berndt, Hall, Hall, and Hausman (1974) consider the computations for this estimator and recommend a method which is, in essence, a modified Gauss-Newton algorithm. Here, we discuss the practical aspects of using existing first derivative nonlinear regression programs for the computations, which, of course, includes this method. Jorgenson and Laffont (1974) consider some asymptotic properties of this estimator assuming the existence of an explicitly defined reduced form.

[3] See, e.g., Jennrich (1969), Malinvaud (1970), Amemiya (1974), and Gallant (1975a).

[4] The example used throughout is linear in the parameters but nonlinear in the variables. Such models have received considerable attention, see Goldfeld and Quandt (1972, Ch. 8) and their references.

[5] Since the model remains in implicit form, transformations $\phi[q_\alpha(y_t, x_t, \theta_\alpha)] = \phi(e_{\alpha t})$ may be employed, in applications, to make the residuals more nearly normally distributed. One would expect that such transformations, by improving the rate at which $(1/\sqrt{n})\ U$ of Lemma A.3 approaches the normal, would improve the rate at which the nonlinear two and three stage least squares estimators approach the normal.

[6] There may be several functions Y such that $y_t = Y(x_t, e_t)$; e.g., put $a_1^* = a_2^* = b_2^* = 1$ in the example. The situation is analogous to the regression $y_t^2 = x_t \theta^* + e_t$ adequate for estimating θ but inadequate for predicting y_t without additional information.

[7] In the case of linear restrictions $T_2 \theta = t_2$, choose T_1 so that $T = \begin{bmatrix} T_1 \\ \cdots \\ T_2 \end{bmatrix}$ is nonsingular and let $W = [W_1 \vdots W_2]$ be the inverse of T. The transformation is $g(\rho) = W_1 \rho + W_2 t_2$ and $G(\rho) = W_1$.

[8] The conclusion that the second equation is not identified may, alternatively, be deduced by applying Theorem 5.4.3 of Fisher (1966, Ch. 5).

Appendix

The following lemmas are variations on known results which are enumerated here in the notation of this article for ease of reference in the proofs. The conclusion of Lemma A.1 follows as a direct consequence of uniform convergence; Lemma A.2 follows immediately from Lemma A.1; and Lemma A.3 is proved by applying Theorem 3 and Corollary 1 of Jennrich (1969). The proof of Lemma A.4 is entirely analogous to the proof of Theorems 2 and 3 with the identity transformation replacing $g(\rho)$ and I replacing $\hat{\Sigma}$; the strong consistency of $\hat{\sigma}_{\alpha\beta}$ follows directly from Lemma A.1. See Amemiya (1974) for similar results.

Lemma A.1. For each fixed v let $f(v, \lambda)$ be a continuous function defined on the compact set Λ. Let $(1/n) \sum_{t=1}^{n} f(v_t, \lambda)$ converge to $\bar{f}(\lambda)$ uniformly in λ. If $\hat{\lambda}_n$ converges almost surely to λ^* in Λ, then $(1/n) \sum_{t=1}^{n} f(v_t, \hat{\lambda}_n)$ converges almost surely to $\bar{f}(\lambda^*)$.

Lemma A.2. Let the assumptions of Section 4 hold, let θ_n^0 and θ_n^{00} converge almost surely to θ^*, and let the elements of $\hat{\Sigma}$ converge almost surely to the elements of Σ. Then

$$B(\theta_n^0) = [(1/n)\, Q'(\theta_n^0)\, (I \otimes Z)]\, [\hat{\Sigma} \otimes (1/n)\, Z'Z]^{-1}$$

converges almost surely to, say, $\bar{B}$ and $(1/n)\, B(\theta_n^0)\, (I \otimes Z')\, Q(\theta_n^{00})$ converges almost surely to Ω. Note, in addition, that $\Omega = \bar{B}\, (\Sigma \otimes P)\bar{B}'$.

Lemma A.3. Let the assumptions of Section 4 hold and let

$$U = (I \otimes Z')\, q(\theta^*) = \sum_{t=1}^{n} e_t \otimes z_t .$$

Then $(1/n)\, U$ converges almost surely to the zero vector and $(1/\sqrt{n})\, U$ converges in distribution to a normal with mean vector zero and variance-covariance matrix $\Sigma \otimes P$.

Lemma A.4. Let the assumptions of Section 4 hold. Combine the two stage least squares estimators into the p by 1 vector $\hat{\theta} = (\hat{\theta}_1', \hat{\theta}_2', \ldots, \hat{\theta}_M')'$. If every equation in the system is identified then: $\hat{\theta}$ converges almost surely to θ^*, $\hat{\sigma}_{\alpha\beta}$ converges almost surely to $\sigma_{\alpha\beta}$ $(\alpha, \beta = 1, 2, \ldots, M)$, and $\sqrt{n}\,(\hat{\theta} - \theta^*)$ converges in distribution to a normal with mean vector zero and variance-covariance matrix $A^{-1} T A^{-1}$ where $A = \text{diag}\,(A_{11}, A_{22}, \ldots, A_{MM})$ and

$$T = \begin{bmatrix} \sigma_{11}A_{11} & \sigma_{12}A_{12} & \cdots & \sigma_{1M}A_{1M} \\ \sigma_{21}A_{21} & \sigma_{22}A_{22} & \cdots & \sigma_{2M}A_{2M} \\ \vdots & \vdots & & \vdots \\ \sigma_{M1}A_{M1} & \sigma_{M2}A_{M2} & & \sigma_{MM}A_{MM} \end{bmatrix}$$

References

Amemiya, T., 1974, "The nonlinear two-stage least-squares estimator," Journal of Econometrics 2, 105-110.

Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hausman, 1974, "Estimation and inference in nonlinear structural models," Annals of Economic and Social Measurement 3, 653-666.

Eisenpress, H., and J. Greenstadt, 1966, "The estimation of nonlinear econometric systems," Econometrica 34, 851-861.

Fisher, F. M., 1966, The identification problem in econometrics (New York: McGraw-Hill).

Gallant, A. R., 1975a, "Seemingly unrelated nonlinear regressions," Journal of Econometrics, 3, 35-50.

Gallant, A. R., 1975b, "Nonlinear regression," The American Statistician, 29, 73-81.

Goldfeld, S. M., and R. E. Quandt, 1972, Nonlinear methods in econometrics (Amsterdam: North Holland).

Jennrich, R. I., 1969, "Asymptotic properties of non-linear least squares estimators," The Annals of Mathematical Statistics 40, 633-643.

Jorgenson, D. W., and J. Laffont, 1974, "Efficient estimation of nonlinear simultaneous equations with additive disturbances," Annals of Economic and Social Measurement 3, 615-640.

Malinvaud, E., 1970, "The consistency of nonlinear regressions," The Annals of Mathematical Statistics 41, 956-969.

Theil, H., 1971, Principles of econometrics (New York: John Wiley and Sons).