NBER WORKING PAPER SERIES

The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model

Takeshi Amemiya*

Working Paper No. 90

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

June 1975

*Stanford University and NBER Computer Research Center
Research supported by NSF Grant DCR 70-03456 A04 to the National Bureau of Economic Research, Inc.

Abstract

The consistency and the asymptotic normality of the maximum likelihood estimator in the general nonlinear simultaneous equation model are proved. It is shown that the proof depends on the assumption of normality unlike in the linear simultaneous equation model. It is proved that the maximum likelihood estimator is asymptotically more efficient than the nonlinear three-stage least squares estimator if the specification is correct. However, the latter has the advantage of being consistent even when the normality assumption is removed. Hausman's instrumental-variable-interpretation of the maximum likelihood estimator is extended to the general nonlinear simultaneous equation model.

Acknowledgement

The author had stimulating discussions with David Belsley, Michio Hatanaka, and Jerry Hausman.

Contents

1. Introduction

In this paper we obtain the asymptotic properties of the maximum likelihood estimator in the general nonlinear simultaneous equation model and compares them with those of the nonlinear three-stage least squares estimator. The main results of the paper are the following:

1) The proof of the consistency and the asymptotic normality of the maximum likelihood estimator in the general nonlinear simultaneous equation model crucially depends on the assumption of normality of the error term unlike in the linear case.

2) All the third-order derivatives can be asymptotically ignored either in the iterative method for obtaining the maximum likelihood estimator or in the computation of the asymptotic variance-covariance matrix.

3) The maximum likelihood estimator is asymptotically more efficient than the nonlinear three-stage least squares estimator.

4) Hausman's iteration method for the computation of the maximum likelihood estimator in the linear case (see Hausman [1975]) is generalized to the nonlinear case. Unlike in the linear case, it does not produce an asymptotically efficient second-round estimator even if the initial estimator is consistent, but, like in the linear case, it illustrates the similarity and the difference between the maximum likelihood and the nonlinear three-stage least squares estimator.

2. Model

We will consider the nonlinear simultaneous equation model defined by the following system of n equations:

$$f_i(y_t, x_t, \alpha_i) = u_{it}, \quad i=1,2,\ldots,n \tag{2.1}$$

where y_t is a n-dimensional vector of endogenous variables, x_t is a vector of exogenous variables, and α_i is a vector of unknown parameters. Not all of the elements of vectors y_t and x_t may actually appear in the arguments of each f_{it}. Define a n-dimensional vector u_t as $(u_{1t}, u_{2t}, \ldots, u_{nt})'$. Then we assume $\{u_t\}$ is independently and identically distributed as multivariate $N(0,\Omega)$. We assume that there are no constraints among α_i's, but the results we subsequently obtain are not affected by the removal of this assumption as we will show at the end of Section 5. We assume either that f_i defines a one-to-one mapping between y_t and u_t or that the researcher can apriori specify a particular root of y_t for a given value of u_t so that the density of y_t can be obtained by the usual way as the product of the Jacobian and the density of u_t. Finally we assume that all the partial derivatives of f_i with respect to α_i and y_t that appear in equation (3.5) in Section 3 exist and are continuous and that $\frac{\partial f_t}{\partial y_t'}$ and $\sum_{t=1}^{T} f_t f_t'$, where $f_t=(f_{1t},\ldots,f_{nt})'$, are nonsingular. These assumptions enable us to define the maximum likelihood estimator. The other conditions needed for the consistency and asymptotic normality of the maximum likelihood estimator are given in Section 3.

3. Maximum Likelihood Estimator

Because of the basic assumptions of Section 2, we can write the logarithmic likelihood function as

$$L^* = -\frac{T}{2}\log|\Omega| + \sum_{t=1}^{T}\log\left|\left|\frac{\partial f_t}{\partial y_t'}\right|\right| - \frac{1}{2}\sum_{t=1}^{T} f_t'\,\Omega^{-1}\,f_t \tag{3.1}$$

where we defined $f_t=(f_{1t},f_{2t},\ldots,f_{nt})'$. Equating the partial derivatives of L^* with respect to Ω to zero, we obtain

$$\Omega = \frac{1}{T}\ \Sigma\, f_t\, f_t' \tag{3.2}$$

where we will abbreviate $\sum_{t=1}^{T}$ as Σ from now on. Putting (3.2) into (3.1) we obtain the concentrated likelihood function

$$L = \Sigma\log\left|\left|\frac{\partial f_t}{\partial y_t'}\right|\right| - \frac{T}{2}\log\left|T^{-1}\,\Sigma\, f_t\, f_t'\right| \quad . \tag{3.3}$$

We define a vector $g_{it} = \dfrac{\partial f_{it}}{\partial \alpha_i}$ and a matrix $g_{ijt} = \dfrac{\partial^2 f_{it}}{\partial\alpha_i\partial\alpha_j'}$.

We will write the partial derivatives of L using these symbols below. To avoid the excessive subscripts, we will omit the subscript t from f, y, u, and g whenever they appear inside the summation. We have

$$\frac{\partial L}{\partial \alpha_i} = \Sigma\ \frac{\partial g_i}{\partial u_i} - T\,\Sigma\, g_i\, f'(\Sigma ff')^{-1}_i \tag{3.4}$$

where we used $\dfrac{\partial g_i}{\partial u_j} = \dfrac{\partial g_i}{\partial y'}\left[\dfrac{\partial f}{\partial y'}\right]^{-1}_j$ and wrote $(\ )^{-1}_i$ for the i^{th} column of the inverse of the matrix within the bracket. We have

$$\frac{\partial^2 L}{\partial \alpha_i \partial \alpha_j'} = \Sigma \frac{\partial g_{ij}}{\partial u_i} - T \Sigma g_{ij} f'(\Sigma ff')^{-1}_{i}$$

$$- \Sigma \frac{\partial g_i}{\partial u_j} \frac{\partial g_j'}{\partial u_i} - T (\Sigma ff')^{-1}_{ij} \Sigma g_i g_j' \qquad (3.5)$$

$$+ T \Sigma g_i f'(\Sigma ff')^{-1}_{j} (\Sigma ff')^{-1'}_{i} \Sigma f g_j'$$

$$+ T(\Sigma ff')^{-1}_{ij} \Sigma g_i f'(\Sigma ff')^{-1} \Sigma f g_j'$$

where we used $\frac{\partial g_{ij}}{\partial u_i} = \frac{\partial g_{ij}}{\partial y'} \left[\frac{\partial f}{\partial y'}\right]^{-1}_{i}$ and wrote $(\ )^{-1}_{ij}$ for the i, j^{th} element of the inverse of the matrix within the bracket.

We define the maximum likelihood estimator of α as a root of equation $\frac{\partial L}{\partial \alpha_i} = 0$. Given assumptions A through E in the appendix, one of the roots is consistent and if we denote the consistent root by $\hat{\alpha}$, we have

$$\sqrt{T}\,(\hat{\alpha} - \alpha_o) \to N\left(0, - \text{plim} \left[\frac{1}{T} \frac{\partial^2 L}{\partial \alpha \partial \alpha'}\bigg|_{\alpha_o}\right]^{-1}\right) . \qquad (3.6)$$

The proof is given in the appendix. The above result is of course not a surprising one. Our main reason for writing down the assumptions explicitly is that checking some of these conditions, especially B and E, is instructive in our model: It will show that the consistency proof depends crucially on the normaility assumption and that the terms involving g_{ij} in (4.5) can asymptotically be ignored. Also, it will aid

us later when we compare the maximum likelihood estimator with the nonlinear three-stage least squares estimator.

We will consider each of the assumptions in the appendix and indicate what conditions on the function f are implied by each. We will not make a great effort to find the minimum set of assumptions needed on f since that is not likely to be a useful exercise. As it was stated earlier, assumptions B and E are most interesting to verify and we will devote most of our time on their verification. But since assumption C requires the greatest number of conditions on f, we will state a sweeping set of conditions on f to make assumption C satisfied. After this is done, only a small number of additional conditions is needed to satisfy the remaining four assumptions. Thus we assume

Condition 1. The probability limit of T^{-1} times every summation that occurs in the right-hand side of (3.5) is finite and is equal to the limit of T^{-1} times its expectation. Moreover, the convergence is uniform in a neighborhood of α_o. In addition, plim T^{-1} $\Sigma ff'$ is nonsingular.

Note that the uniform boundedness of the third-order derivatives may be substituted for assumption C.

Before proceeding further, we will prove the following important lemma which will be frequently used.

Lemma. Suppose $u_1, u_2, \ldots u_n$ are jointly normal with mean 0 and $h(u_1, u_2, \ldots u_n)$ is such that E h and $E \frac{\partial h}{\partial u_i}$ are finite. Then, $Ehu_1 = \sum_{i=1}^{n} E \frac{\partial h}{\partial u_i} \sigma_{1i}$, where σ_{1i} is the covariance between u_1 and u_i.

Proof. Replace u_i in h with $\frac{\sigma_{il}}{\sigma_1^2} u_1 + w_i$ for i=2,...,n and treat h as a function of $u_1, w_2, w_3, \ldots, w_n$. Then, $Eh u_1 = E_w E_{u_1} h u_1$ where $w = (w_2, \ldots, w_n)$. But using integration by parts we have

$$\begin{aligned} E_{u_1} h u_1 &= \int_{-\infty}^{\infty} h u_1 \phi \, du_1 \\ &= -\sigma_1^2 [h\phi]_{-\infty}^{\infty} + \sigma_1^2 \int_{-\infty}^{\infty} \frac{dh}{du_1} \phi \, du_1 \end{aligned} \tag{3.7}$$

where ϕ is the density of $N(0,\sigma_1^2)$. But the first term of the right-hand side of (3.7) is zero because Eh is finite. Note $\frac{dh}{du_1} = \sum_{i=1}^{n} \frac{\partial h}{\partial u_i} \frac{\sigma_{1i}}{\sigma_1^2}$.

Therefore, taking the expectation of both sides of (3.7) with respect to w, we get the desired result.

Now we will consider assumption B. Using (3.4) we have

$$\begin{aligned} \frac{1}{\sqrt{T}} \frac{\partial L}{\partial \alpha_i}\bigg|_{\alpha_o} &= \frac{1}{\sqrt{T}} \Sigma \left[\frac{\partial g_i}{\partial u_i} - g_i \, u' \, \sigma^i\right] \\ &\quad - \frac{1}{T} \Sigma \, g_i \, u' \cdot \sqrt{T} \left[\left(\frac{\Sigma uu'}{T}\right)_i^{-1} - \sigma^i\right] \end{aligned} \tag{3.8}$$

where σ^i is the i^{th} column of Ω^{-1}. We immediately see that the mean of the first term of the right-hand side of (3.8) is zero since g_i satisfies the condition of the lemma because of condition 1. Also using the lemma we have

$$\text{plim } \frac{1}{T} \Sigma\, g_i\, u' = \lim \frac{1}{T} \Sigma\, E\, \frac{\partial g_i}{\partial u'}\, \Omega \tag{3.9}$$

since $\{g_{it}\, u_{jt}\}$ satisfies the conditions for a law of large numbers because of conditon 1. Therefore, denoting the equivalence of the limit distribution by the symbol $\underline{\underline{LD}}$, we have

$$\frac{1}{\sqrt{T}}\, \frac{\partial L}{\partial \alpha_i}\bigg|_{\alpha_o} \underline{\underline{LD}}\; p_{i1} + p_{i2} \tag{3.10}$$

where

$$p_{i1} = \frac{1}{\sqrt{T}}\, \Sigma \left[\frac{\partial g_i}{\partial u_i} - g_i\, u'\, \sigma^i\right] \tag{3.11}$$

and

$$p_{i2} = \lim \frac{1}{T} \Sigma\, E\, \frac{\partial g_i}{\partial u'} \cdot \frac{1}{\sqrt{T}} \Sigma\, (uu' - \Omega)\sigma^i\ . \tag{3.12}$$

Written thus, it is clear that a certain essential boundedness of g_i and $\frac{\partial g_i}{\partial u_i}$ is sufficient to let (3.10) follow a central limit theorem.

For example, the following condition is certainly sufficient:

Condition 2. $E|g_{it}|^3$ and $E\left|\frac{\partial g_{it}}{\partial u_{it}}\right|^3$ are uniformly bounded for all t, where g_{it} and $\frac{\partial g_{it}}{\partial u_{it}}$ are evaluated at α_o.

Next we will verify assumption E. Taking the probability limit of T^{-1} times (3.5) evaluated at α_o, we have

$$\begin{aligned}\operatorname{plim} T^{-1} \frac{\partial^2 L}{\partial \alpha_i \partial \alpha_j'}\bigg|_{\alpha_o} &= \operatorname{plim} T^{-1} \Sigma \left[\frac{\partial g_{ij}}{\partial u_i} - g_{ij}\, u'\, \sigma^i\right] \\ &- \operatorname{plim} T^{-1} \Sigma \frac{\partial g_i}{\partial u_j} \frac{\partial g_j'}{\partial u_i} - \sigma^{ij} \operatorname{plim} T^{-1} \Sigma\, g_i g_j' \\ &+ \operatorname{plim} T^{-1} \Sigma\, g_i\, u' \cdot \sigma^j\, \sigma^{i\prime} \cdot \operatorname{plim} T^{-1} \Sigma\, u g_j' \\ &+ \sigma^{ij} \operatorname{plim} T^{-1} \Sigma\, g_i\, u' \cdot \Omega^{-1} \cdot \operatorname{plim} T^{-1} \Sigma\, u g_j' \quad .\end{aligned} \tag{3.13}$$

Because of condition 1, we can replace plim in the right-hand side of (3.13) with lim E. But, then, the first term drops out provided g_{ij} satisfies the condition of the lemma. So we impose

Condition 3. E g_{ijt} is finite, where g_{ijt} is evaluated at α_o.

Thus, either in performing Newton's iteration to obtain the maximum likelihood estimator or in obtaining its asymptotic variance-covariance matrix one need not compute g_{ij}. Also we can apply the result of the lemma to each term involving the product of u and g_i. Thus we have

$$\text{plim } T^{-1} \left. \frac{\partial^2 L}{\partial \alpha_i \partial \alpha_j'} \right|_{\alpha_o} = - \lim T^{-1} \Sigma E \frac{\partial g_i}{\partial u_j} \frac{\partial g_j'}{\partial u_i}$$

$$- \sigma^{ij} \lim T^{-1} \Sigma E g_i g_j' \qquad (3.14)$$

$$+ \lim T^{-1} \Sigma E \frac{\partial g_i}{\partial u_j} \cdot \lim T^{-1} \Sigma E \frac{\partial g_j'}{\partial u_i}$$

$$+ \sigma^{ij} \lim T^{-1} \Sigma E \frac{\partial g_i}{\partial u'} \cdot \Omega \cdot \lim T^{-1} \Sigma E \frac{\partial g_j'}{\partial u} \quad .$$

We must compare the above with $\lim T^{-1} E\left[\left. \frac{\partial L}{\partial \alpha_i} \right|_{\alpha_o} \cdot \left. \frac{\partial L}{\partial \alpha_j'} \right|_{\alpha_o} \right]$.

We impose

Condition 4. $E\left[\left. \frac{\partial g_{it}}{\partial u_{it} \partial u_{jt}} g_j' \right|_{\alpha_o} \right]$ is finite.

Then, by the repeated application of the lemma, we have

$$E\left[\frac{\partial g_i}{\partial u_i} - g_i u' \sigma^i \right] \left[\frac{\partial g_j'}{\partial u_j} - g_j' u' \sigma^j \right]$$

$$= - E\left[\frac{\partial g_i}{\partial u_i \partial u_j} - \frac{\partial g_i}{\partial u_j} u' \sigma^i - \sigma^{ij} g_i \right] g_j' \qquad (3.15)$$

$$= E \frac{\partial g_i}{\partial u_j} \frac{\partial g_j'}{\partial u_i} + \sigma^{ij} E g_i g_j' \quad .$$

Therefore, from (3.11) and (3.15) we have

$$\lim E\, P_{il}\, P_{jl} = \lim T^{-1} \Sigma\, E \frac{\partial g_i}{\partial u_j} \frac{\partial g_j'}{\partial u_i} + \sigma^{ij} \lim T^{-1} \Sigma\, E\, g_i\, g_j' \quad . \tag{3.16}$$

We have

$$E\,(uu' - \Omega)\,\sigma^i\,\sigma^{j'}\,(uu' - \Omega) = \sigma^{ij}\,\Omega + \ell_j\,\ell_i' \tag{3.17}$$

where ℓ_j is a n-dimensional vector with 1 in the j^{th} place and 0 elsewhere. Therefore, from (3.12) and (3.17) we have

[Continued on page 11]

$$\lim E\, p_{i2}\, p_{j2} = \sigma^{ij} \lim T^{-1} \Sigma E \frac{\partial g_i}{\partial u'} \cdot \Omega \cdot \lim T^{-1} \Sigma E \frac{\partial g_j'}{\partial u} + \lim T^{-1} \Sigma E \frac{\partial g_i}{\partial u_j} \cdot \lim T^{-1} \Sigma E \frac{\partial g_j'}{\partial u_i} \quad . \tag{3.18}$$

By the application of the lemma we have

$$E \left[\frac{\partial g_i}{\partial u_i} - g_i\, u'\, \sigma^i \right] \sigma^{j'} (uu' - \Omega) = - \sigma^{ij} E \frac{\partial g_i}{\partial u'}\, \Omega - E \frac{\partial g_i}{\partial u_j} \ell_i' \quad . \tag{3.19}$$

Therefore from (3.11), (3.12), and (3.19) we have

$$\lim E\, p_{i1}\, p_{j2} = - \sigma^{ij} \lim T^{-1} \Sigma E \frac{\partial g_i}{\partial u'} \cdot \Omega \cdot \lim T^{-1} \Sigma E \frac{\partial g_j'}{\partial u} - \lim T^{-1} \Sigma E \frac{\partial g_i}{\partial u_j} \cdot \lim T^{-1} \Sigma E \frac{\partial g_j'}{\partial u_i} \quad . \tag{3.20}$$

Similarly we have

$$E(uu' - \Omega)\, \sigma^i \left[\frac{\partial g_j}{\partial u_j} - g_j\, u'\, \sigma^j \right] = - \sigma^{ij}\, \Omega\, E \frac{\partial g_j'}{\partial u} - \ell_j\, E \frac{\partial g_j'}{\partial u_i} \quad . \tag{3.21}$$

Therefore we have

$$\lim E\, p_{i2}\, p_{j1} = \lim E\, p_{i1}\, p_{j2} \quad . \tag{3.22}$$

Finally, assumption E follows from (3.10), (3.14), (3.16), (3.18), (3.20), and (3.22).

This leaves assumptions A and D. Assumption A requires only one additional condition:

Condition 5. $\operatorname{plim} T^{-1} \Sigma \log \left|\left|\frac{\partial f}{\partial y'}\right|\right|$ exists in a neighborhood of α_o.

As we will show in Section 5, assumption D is implied by

Condition 6. $\lim T^{-1} \Sigma E \left.\frac{\partial f_i}{\partial \alpha_i}\right|_{\alpha_o} E \left.\frac{\partial f_i}{\partial \alpha_i'}\right|_{\alpha_o}$ is finite and nonsingular

for every i.

To sum up, conditions 1 through 6 imply assumptions A through E in the appendix.

Note that the proof of both consistency and asymptotic normality crucially depends on the normality assumption unlike in the linear case where the maximum likelihood estimator can be easily shown to be consistent for general specifications on the error term. This fact increases the usefulness of such an estimator as the nonlinear two-stage or three-stage least squares estimator which has been shown to be consistent for general specifications on the error term.

4. Iterative Methods

Consider the class of gradient methods of iteration defined by

$$\hat{\alpha}_2 = \hat{\alpha}_1 - A \left.\frac{\partial L}{\partial \alpha}\right|_{\hat{\alpha}_1} \tag{4.1}$$

where $\hat{\alpha}_1$ is an initial estimator and A is some matrix which may be stochastic. Using a Taylor expansion of $\left.\frac{\partial L}{\partial \alpha}\right|_{\hat{\alpha}_1}$ around α_o, the true value, we have from (4.1)

$$\sqrt{T}\,(\hat{\alpha}_2-\alpha_o) = -\sqrt{T}\; A \left.\frac{\partial L}{\partial \alpha}\right|_{\alpha_o} + \left[I - A \left.\frac{\partial^2 L}{\partial \alpha \partial \alpha'}\right|_{\alpha^*}\right] \cdot \sqrt{T}\,(\hat{\alpha}_1-\alpha_o) \tag{4.2}$$

where α^* lies between $\hat{\alpha}_1$ and α_o. Suppose that $\hat{\alpha}_1$ is a consistent estimator of α_o such that $\sqrt{T}\,(\hat{\alpha}_1-\alpha_o)$ has a proper limit distribution. It is apparent from (4.2) that the asymptotic distribution of the second-round estimator does not depend upon the asymptotic distribution of the first-round estimator if and only if

$$\operatorname{plim} T^{-1} A = \operatorname{plim} T^{-1} \left.\frac{\partial^2 L}{\partial \alpha \partial \alpha'}\right|_{\alpha_o}. \tag{4.3}$$

Moreover, it is apparent from (4.2) that in this case the limit distribution of $\sqrt{T}\,(\hat{\alpha}_2-\alpha_o)$ is the same as that of the maximum likelihood estimator. We will call the gradient method satisfying (4.3) the efficient Newton iteration.

Next consider the iteration that can be derived from the equation obtained by putting (3.4) equal to zero. We can rewrite the equation as

$$\left[T^{-1} \Sigma \frac{\partial g_i}{\partial u'} \cdot F' - G_i' \right] F(T^{-1} F'F)_i^{-1} = 0 \tag{4.4}$$

where F' is the nxT matrix whose i, t^{th} element is $f_i(y_t, x_t, \alpha_i)$ and G_i' is the matrix whose t^{th} column is $\frac{\partial f_i(y_t, x_t, \alpha_i)}{\partial \alpha_i}$. Define

$$\hat{G}_i' = G_i' - T^{-1} \Sigma \frac{\partial g_i}{\partial u'} \cdot F' \tag{4.5}$$

and

$$\underset{\sim}{\hat{G}}' = \begin{bmatrix} \hat{G}_1' & 0 & . & . & 0 \\ 0 & \hat{G}_2' & & & \\ . & & . & & \\ . & & & . & \\ 0 & & & & \hat{G}_n' \end{bmatrix} . \tag{4.6}$$

Also define $\underset{\sim}{f}$ as the (nxT)-dimensional vector obtained by stacking the columns of F. Then, all the n equations in (4.4) for i=1,2,...,n can be combined as

$$\underset{\sim}{\hat{G}}' (\hat{\Omega}^{-1} \otimes I) \underset{\sim}{f} = 0 \tag{4.7}$$

where we have written $\hat{\Omega}$ for $T^{-1}F'F$. Expanding $\underset{\sim}{f}(\hat{\alpha}_1)$ in a Taylor series around α_o we finally obtain the iteration

$$\hat{\alpha}_2 = \hat{\alpha}_1 - [\hat{\underset{\sim}{G}}'(\hat{\Omega}^{-1} \otimes I)\ \underset{\sim}{G}]^{-1}\ \hat{\underset{\sim}{G}}'(\hat{\Omega}^{-1} \otimes I)\ \underset{\sim}{f} \tag{4.8}$$

where

$$\underset{\sim}{G} = \begin{bmatrix} G_1 & 0 & . & . & 0 \\ 0 & G_2 & & & \\ . & & . & & \\ . & & & . & \\ 0 & & & & G_n \end{bmatrix} \tag{4.9}$$

and every variable in the right-hand side of (4.8) is evaluated at $\hat{\alpha}_1$. Equation (4.8) is the generalization of the formula expounded by Hausman [1975] for the linear case. Note that (4.8) belongs to the class of iteration defined by (4.1) with $A = [\hat{\underset{\sim}{G}}'(\hat{\Omega}^{-1} \otimes I)\underset{\sim}{G}]^{-1}$.
By the application of the lemma we can easily show

$$\begin{aligned} \operatorname{plim} T^{-1}\ \hat{G}_i'(\hat{\Omega}^{-1} \otimes I)\ G_j &= -\ \sigma^{ij}\ \lim T^{-1}\ \Sigma\ E\ g_i g_j' \\ &+ \sigma^{ij}\ \lim T^{-1}\ \Sigma\ E\ \frac{\partial g_i}{\partial u'} \cdot \Omega \cdot \lim T^{-1}\ \Sigma\ E\ \frac{\partial g_j'}{\partial u} \quad . \end{aligned} \tag{4.10}$$

By comparing (3.14) with (4.10) we see that condition (4.3) is violated. Thus we conclude that the asymptotic distribution of the second-round estimator in this iteration depends on the asymptottic distribution of the initial estimator and is not asymptotically efficient. Note that the result is not changed if $[\hat{G}_i'(\hat{\Omega}^{-1} \otimes I)\hat{G}_j]$ is used instead because its probability limit can be shown to be equal to (4.10). Note also that in the linear case the sum of the first term and the third term of the right-hand side of (3.14) is zero so that conditon (4.3) gets satisfied.

Although (4.8) may not be a good method of iteration, it does serve a useful pedagogical purpose as Hausman's linear case does, for it demonstrates a certain similarity between the maximum likelihood estimator and the nonlinear three-stage least squares estimator.

5. Nonlinear Three-Stage Least Squares Estimator

Jorgenson and Laffont [1974] defined the nonlinear three-stage least squares estimator (henceforth to be abbreviated as NL3S) and proved its consistency and asymptotitc normality, extending the result of Amemiya [1974] obtained for the nonlinear two-stage least squares estimator. They defined the NL3S as the value of α that minimizes

$$\underset{\sim}{f}(\alpha)'[\hat{\Omega}^{-1} \otimes X(X'X)^{-1} X']\underset{\sim}{f}(\alpha) \tag{5.1}$$

where $\hat{\Omega}$ is some consistent estimate of Ω and X is a matrix of exogenous variables which may not coincide with the exogenous variables that appear originally in the arguments of f. Its asymptotic variance-covariance matrix is given by

$$\left[\operatorname{plim} T^{-1} \frac{\partial \underset{\sim}{f}'}{\partial \underset{\sim}{\alpha}} \bigg|_{\alpha_o} [\Omega^{-1} \otimes X(X'X)^{-1} X'] \frac{\partial \underset{\sim}{f}}{\partial \underset{\sim}{\alpha}'} \bigg|_{\alpha_o} \right]^{-1} . \tag{5.2}$$

In this paper we will define the NL3S more generally as the value of α that minimizes $\underset{\sim}{f}'\, A\, \underset{\sim}{f}$ where A could take any one of the following three forms:

$$A_1 = \hat{\Lambda}^{-\frac{1}{2}} S_1 (S_1'S_1)^{-1} S_1' \hat{\Lambda}^{-\frac{1}{2}} , \tag{5.3}$$

$$A_2 = S_2 (S_2' \hat{\Lambda} S_2)^{-1} S_2' \tag{5.4}$$

and

$$A_3 = \hat{\Lambda}^{-1} S_3 (S_3' \hat{\Lambda}^{-1} S_3)^{-1} S_3' \hat{\Lambda}^{-1} \tag{5.5}$$

where S_1, S_2, and S_3 are matrices of at least asymptotically nonstochastic variables and $\hat{\Lambda} = \hat{\Omega} \otimes I$. The asymptotic variance-covariance matrix is given by

$$\left[\text{plim } T^{-1} \frac{\partial f'}{\partial \alpha}\bigg|_{\alpha_o} A \frac{\partial f}{\partial \alpha'}\bigg|_{\alpha_o} \right]^{-1} . \tag{5.6}$$

All the three formulations are equivalent in the sense that A_1, A_2, and A_3 can be made equal by appropriately choosing S_1, S_2, and S_3. If we take

$$S_1 = S_2 = S_3 = \begin{bmatrix} X & 0 & . & . & 0 \\ 0 & X & & & \\ . & & . & & \\ . & & & . & \\ 0 & & & & X \end{bmatrix} ,$$

all the three are reduced to the Jorgenson-Laffont NL3S. It is apparent from (5.6) that for all A_i, i=1, 2, 3, its lower bound is equal to

$$\left[\lim T^{-1} E \frac{\partial f'}{\partial \alpha} \Big|_{\alpha_o} (\Omega^{-1} \otimes I) E \frac{\partial f}{\partial \alpha'} \Big|_{\alpha_o} \right]^{-1} . \tag{5.7}$$

The lower-bound is attained when $S_1 = \hat{\Lambda}^{-\frac{1}{2}} E \frac{\partial f}{\partial \alpha'}$, $S_2 = \hat{\Lambda}^{-1} E \frac{\partial f}{\partial \alpha'}$, and $S_3 = E \frac{\partial f}{\partial \alpha'}$, where we are implicitly assuming that the α that appears in $E \frac{\partial f}{\partial \alpha'}$ must be estimated consistently. We will call the resulting NL3S estimator where any of these optimal S's is used as the best nonlinear three-stage least squares estimator (abbreviated as BNL3S). This is often not a practical estimator because $E \frac{\partial f}{\partial \alpha'}$ is usually difficult to obtain in explicit form, but the consideration of BNL3S is theoretically useful as it provides something to aim at.

One can also attain the lower bound (5.7) using the Jorgenson-Laffont NL3S, but that is possible if and only if the space spanned by the column vectors of X contains the union of the spaces spanned by the column vectors of $E \frac{\partial f_i}{\partial \alpha_i'} \Big|_{\alpha_o}$ for i = 1, 2,...,n. This necessitates including many columns in X, which is likely to increase the finite sample variance of the estimator although it has no affect asymptotically. This is the disadvantage of the Jorgenson-Laffont definition compared to the definition of this paper.

We will next show that the BNL3S is asymptotically less efficient than the maximum likelihood estimator. Using the lemma, we have

$$E\left\{\left[\frac{\partial g_i}{\partial u_i} - g_i\, u'\, \sigma^i\right] \sigma^{j'}\, u\, E\, g_j'\right\} = -\sigma^{ij}\, E\, g_i\, E\, g_j' \quad . \tag{5.8}$$

Similarly we have

$$E\left\{E\, g_i \cdot u'\, \sigma^i \left[\frac{\partial g_j}{\partial u_j} - g_j\, u'\, \sigma^j\right]\right\} = -\sigma^{ij}\, E\, g_i\, E\, g_j' \quad . \tag{5.9}$$

We have

$$E(E\, g_i \cdot u'\, \sigma^i\, \sigma^{j'}\, u\, E\, g_i) = \sigma^{ij}\, E\, g_i\, E\, g_j' \quad . \tag{5.10}$$

We obviously have

$$E[(uu' - \Omega)\sigma^i\, \sigma^{j'}\, u\, E\, g_i] = 0 \tag{5.11}$$

and

$$E[E\, g_i \cdot u'\, \sigma^i\, \sigma^{j'}(uu' - \Omega)] = 0 \quad . \tag{5.12}$$

Therefore, from (3.10), (3.11), (3.12), and (5.8) through (5.12) we have

$$\lim E(p_{i1} + p_{i2} - \frac{1}{\sqrt{T}}\, E\, g_i \cdot u'\, \sigma^i)(p_{j1}' + p_{j2}' - \frac{1}{\sqrt{T}}\sigma^{j'}\, u\, E\, g_j) \tag{5.13}$$

$$= \lim T^{-1}\, E\left[\left.\frac{\partial L}{\partial \alpha_i}\right|_{\alpha_o} \cdot \left.\frac{\partial L}{\partial \alpha_j'}\right|_{\alpha_o}\right] - \sigma^{ij} \lim T^{-1} \Sigma\, E\, g_i\, E\, g_j' \quad .$$

The first term of the right-hand side of (5.13) is the i-jth block of the inverse of the asymptotic variance-covariance matrix of the maximum likelihood estimator and the second term is that of the BNL3S as it is evident from (5.7). But the matrix whose i-jth block is given in the left-hand side of (5.13) is clearly nonnegative definite. Moreover, since the matrix is nonzero with probability one in general, we conclude that the BNL3S is asymptotically less efficient than the maximum likelihood estimator.

Although the NL3S is asymptotically less efficient than the maximum likelihood estimator, it is more robust against non-normality because it is consistent provided the error term has mean zero and certain higher-order finite moments whereas the concistency of the maximum likelihood estimator in the nonlinear model depends crucially on the normality assumption as we have seen in Section 3 above.

A necessary and sufficient condition for the matrix to be inverted in (5.7) to be nonsingular is easily seen to be condition 6 of Section 3. In the linear case this condition implies the usual rank condition of identifiability for each equation. However in the nonlinear case the above condition is likely to be met even if all the exogenous variables appear in each f_i provided f_i is sufficiently nonlinear. Because of (5.13), condition 6 implies assumption D of the appendix.

The Gauss-Newton iteration to obtain the BNL3S can be written as

$$\hat{\alpha}_2 = \hat{\alpha}_1 - [\bar{G}' (\hat{\Omega}^{-1} \otimes I) G]^{-1} \bar{G}' (\hat{\Omega}^{-1} \otimes I) f \qquad (5.15)$$

where

$$\bar{\underset{\sim}{G}}' = \begin{bmatrix} \bar{G}_1' & 0 & . & . & 0 \\ 0 & \bar{G}_2' & & & \\ . & & . & & \\ . & & & . & \\ 0 & & & & \bar{G}_n' \end{bmatrix} \tag{5.16}$$

and

$$\bar{G}_i' = E\, G_i' \tag{5.17}$$

Equation (5.15) differs from (4.8) only in the respective "instrumental variables" used defined by (5.17) and (4.5) respectively. Intuitively speaking, $\hat{G}_i$ catches more of the essentially nonstochastic part of G_i than $\bar{G}_i$ does. Note that by a Taylor expansion we have

$$g_{it}(u_t) = g_{it}(0) + \frac{\partial g_{it}}{\partial u_t'} \cdot u_t \quad . \tag{5.18}$$

But (4.5) can be written as

$$g_{it}(u_t) = \hat{g}_{it} + T^{-1} \Sigma \frac{\partial g_{it}}{\partial u_t'} \cdot u_t \quad . \tag{5.19}$$

The similarity between (5.18) and (5.19) provides some justification of $g_{it}(0)$ as the alternative instrumental variable. The α_i that appears in $g_{it}(0)$ must be consistently estimated. The resulting NL3S is

asymptotically less efficient than the BNL3S but is much more practical. An even more practical choice of the instrument is to use $g_{it}(\hat{y}_t, x_t, \hat{\alpha}_i)$ where $\hat{y}_t$ is calculated simply as the predictor of y_t obtained by the linear least squares regression of y_t on all the exogenous variables. A definite comparison between this choice and the use of $g_{it}(0)$ can not be easily made.

So far in this paper we have assumed that there are no constraints among α_i's. The removal of this assumption, however, causes no difficult problem. If there are constraints among α_i's, we can express each α_i parametically as $\alpha_i(\beta)$ where the number of elements in β is fewer than those in $\alpha=(\alpha_1', \alpha_2', \ldots, \alpha_n')'$. Thus, one can simply premultiply the inverse of the asymptotic variance-covariance matrix of the maximum likelihood estimator or the NL3S by $\frac{\partial \alpha'}{\partial \beta}$ and postmultiply by $\frac{\partial \alpha}{\partial \beta'}$. Hence, all the results of the paper hold.

6. Conclusions

We have proved that the maximum likelihood estimator is asymptotically more efficient than the nonlinear three-stage least squares estimator. However we have also shown that the consistency of the maximum likelihood estimator depends on the assumption of normality whereas that of the nonlinear three-stage least squares is not. This fact increase the attractiveness of the latter. The following are some important topics for further research:

1) Evaluate the degree of the relative inefficiency of the best nonlinear three-stage least squares estimator as compared to the maximum likelihood in specific models.
2) Evaluate the degree of the realtive inefficiency of several versions of the computationally practical nonlinear three-stage least squares estimator as compared to the best nonlinear three-stage least squares estimator in specific models.
3) Is there an estimator, possibly even better than the best nonlinear three-stage least squares estimator, which is computationally simpler than the maximum likelihhod estimator? Can that estimator remain consistent when the normality assumption is removed?

References

1. Amemiya, T. [1974], "The Nonlinear Two-Stage Least-Squares Estimator," Journal of Econometrics, Vol. 2, 105-110.

2. Hausman, J. [1975], "An Instrumental Variable Approach to Full Information Estimators for Linear and Non-Linear Econometric Models," Econometrica, forthcoming.

3. Jorgenson, D.W., and J. Laffont [1974], "Efficient Estimation of Nonlinear Simultaneous Equations with Additive Disturbances," Annals of Economic and Social Measurement, Vol. 3, 615-640.

APPENDIX

Assumpitons

We make the following assumptions in additon to the basic assumptions of the model stated in Section 2.

A. $\operatorname{plim} \frac{1}{T} L_T(\alpha)$ exists in a neighborhood of α_o.

B. $$\frac{1}{\sqrt{T}} \left.\frac{\partial L_T}{\partial \alpha}\right|_{\alpha_o} \to N\left(0, \lim T^{-1} E\left[\left.\frac{\partial L_T}{\partial \alpha}\right|_{\alpha_o} \cdot \left.\frac{\partial L_T}{\partial \alpha'}\right|_{\alpha_o}\right]\right).$$

C. $\operatorname{plim} T^{-1} \frac{\partial^2 L_T}{\partial \alpha \partial \alpha'}$ exists in a neighborhood of α_o and the convergence is uniform in the neighborhood.

D. $\operatorname{plim} T^{-1} \left.\frac{\partial^2 L_T}{\partial \alpha \partial \alpha'}\right|_{\alpha_o}$ is negative definite.

E. $$\lim T^{-1} E\left[\left.\frac{\partial L_T}{\partial \alpha}\right|_{\alpha_o} \cdot \left.\frac{\partial L_T}{\partial \alpha'}\right|_{\alpha_o}\right] = -\operatorname{plim} T^{-1} \left.\frac{\partial^2 L_T}{\partial \alpha \partial \alpha'}\right|_{\alpha_o}$$

Theorem. Under the basic assumptions of the model stated in Section 2 and assumptions A through E above, a root of the equation $\frac{\partial L}{\partial \alpha} = 0$ is consistent and the consistent root $\hat{\alpha}$ satisfies $\sqrt{T}(\hat{\alpha}-\alpha_o) \to N\left(0, -\operatorname{plim} T\left[\left.\frac{\partial^2 L}{\partial \alpha \partial \alpha'}\right|_{\alpha_o}\right]^{-1}\right)$.

Proof. Expanding $T^{-1} L_T(\alpha)$ in a Taylor series around the true value α_o, we have

$$T^{-1} L_T(\alpha) = T^{-1} L_T(\alpha_o) + T^{-1} \left.\frac{\partial L_T}{\partial \alpha'}\right|_{\alpha_o} (\alpha - \alpha_o) + \frac{1}{2} (\alpha - \alpha_o)' T^{-1} \left.\frac{\partial^2 L_T}{\partial \alpha \partial \alpha'}\right|_{\alpha_T^*} (\alpha - \alpha_o) \tag{A.1}$$

where α_T^* lies between α and α_o. Taking the probability limit of both sides of (A.1) and using assumptions A, B, and C, we have

$$\operatorname{plim} T^{-1} L_T(\alpha) = \operatorname{plim} T^{-1} L_T(\alpha_o) + \frac{1}{2} (\alpha - \alpha_o)' \operatorname{plim} T^{-1} \left.\frac{\partial^2 L_T}{\partial \alpha \partial \alpha'}\right|_{\alpha_T^*} (\alpha - \alpha_o) \quad . \tag{A.2}$$

Since $\frac{\partial^2 L_T}{\partial \alpha \partial \alpha'}$ is continuous by a basic assumption stated in Section 2, assumption C implies that plim $T^{-1} \frac{\partial^2 L_T}{\partial \alpha \partial \alpha'}$ is continuous in a neighborhood of α_o. Therefore, by (D), the second term of the right-hand side of (A.1) is negative for all α in a neighborhood of α_o. Therefore, plim $T^{-1} L_T(\alpha)$ attains a local maximum at α_o. This implies that a root of equation $\frac{\partial L}{\partial \alpha_i} = 0$ is consistent. The asymptotic normality follows easily from assumptions B through E using the Taylor expansion

$$\left.\frac{\partial L_T}{\partial \alpha}\right|_{\hat{\alpha}} = \left.\frac{\partial L_T}{\partial \alpha}\right|_{\alpha_o} + \left.\frac{\partial^2 L_T}{\partial \alpha \partial \alpha'}\right|_{\alpha_T^{**}} (\hat{\alpha} - \alpha_o) \qquad (A.3)$$

where $\hat{\alpha}$ is the consistent root and α_T^{**} lies between $\hat{\alpha}$ and α_o and noting the left-hand side of (A.3) is zero by the definition of $\hat{\alpha}$.