Title
Simultaneous Equations with Covariance Restrictions

Authors
Rothenberg, Thomas J.
Ruud, Paul A.

Publication Date
1989-04-01

UNIVERSITY OF CALIFORNIA, BERKELEY

Department of Economics

Berkeley, California 94720

Working Paper No. 89-106

Simultaneous Equations with Covariance Restrictions

Thomas J. Rothenberg and Paul A. Ruud

Department of Economics
University of California at Berkeley

April 1989

Key words: instrumental variables, linearized maximum likelihood, minimum distance, simultaneous equations model

Abstract

Full information maximum likelihood and minimum distance estimators are derived for the simultaneous equations model with covariance restrictions. Linearization yields estimators that are easy to compute and have an instrumental variables interpretation. The efficient three-stage least squares estimator in the presence of covariance restrictions is shown to be a special case. Modified estimators which take into account possible nonnormality in the errors are also discussed.

JEL Classification: 211

SIMULTANEOUS EQUATIONS WITH COVARIANCE RESTRICTIONS

1. Introduction

We reconsider the linear simultaneous equations system with covariance restrictions as described by Hausman, Newey, and Taylor (1987). Using both maximum likelihood and minimum distance approaches, we explore several estimation strategies. The augmented three stage least squares estimator proposed by these authors is derived as a special case.

First we review the efficient estimation of the linear simultaneous equations model without covariance restrictions. This provides the outline of the approach that we will use in the presence of covariance restrictions and it also is a useful contrast. Secondly, we provide an alternative derivation of the asymptotic covariance matrix of a subset of parameters estimated by maximum likelihood. This turns out to be much simpler than using the partitioned inverse formula on the information matrix. Thirdly, we derive the efficient 3SLS estimator in the presence of covariance restrictions as a linearized maximum likelihood and as a linearized minimum distance estimator. Finally, we discuss the properties of our estimators when the errors are not normally distributed.

2. No Covariance Restrictions

A complete linear simultaneous equations model involving G endogenous and K predetermined variables for a sample of size T can be written compactly as

$$YB + X\Gamma = U \ , \qquad (1)$$

where Y is the $T \times G$ matrix of observations on the endogenous variables, X is the $T \times K$ matrix of observations on the predetermined variables, and U is the $T \times G$ matrix of unobserved random errors. The T rows of U are mutually independent, identically distributed random vectors with mean zero and positive definite covariance matrix Σ. The matrix B is assumed to be nonsingular so that Y has the reduced-form representation

$$Y = X\Pi + V$$

where $\Pi = -\Gamma B^{-1}$ and the rows of $V = UB^{-1}$ are independent with mean zero and covariance matrix $B'^{-1}\Sigma B^{-1}$. We shall further assume that X has rank K.

Some of the elements of B and Γ are known a priori and do not have to be estimated from the data. Let $A = (B' \ \Gamma')'$ be the $(G + K) \times G$ matrix of of structural coefficients. Defining δ to be the q-dimensional vector of unknown elements of A and defining r to be the $GK + G^2 - q$ vector of known elements, we can write[1]

$$(2) \qquad r = R_\delta' \mathrm{vec}(A) \qquad \text{and} \qquad \delta = S_\delta' \mathrm{vec}(A) \ .$$

The selection matrices R_δ and S_δ consist of zeros and ones and satisfy

$$R_\delta R_\delta' + S_\delta S_\delta' = I_{G(G+K)} \ .$$

Using this fact, the prior restriction $R_\delta' \mathrm{vec}(A) = r$ can be solved as

$$\mathrm{vec}(A) = S_\delta \delta + R_\delta r \ .$$

Since equation (1) is equivalent to $[I_G \otimes (Y \ X)]\mathrm{vec}(A) = \mathrm{vec}(U)$, the

---

1 The operator *vec* transforms a matrix into a vector by stacking successive columns. We shall also use the operator *vech* which stacks the columns of the lower triangle (including diagonal) of a symmetric matrix thereby removing the redundant elements. For matrices of appropriate dimensions, $\mathrm{vec}(CDE) = (E' \otimes C)\mathrm{vec}(D)$, where $\otimes$ represents the Kronecker product.

simultaneous equations model with linear restrictions can be written in the familiar form of one stacked vector equation in GT dimensions:

$$y = Z\delta + u$$

where $u \equiv \mathrm{vec}(U)$ has mean zero and covariance matrix $(\Sigma \otimes I_T)$; the variables y and Z are given by

$$y = [I_G \otimes (Y\ X)]R_\delta r \ , \qquad Z = -\ [I_G \otimes (Y\ X)]S_\delta \ .$$

Assuming δ is identifiable, one estimation strategy is to search for a $GT \times q$ matrix W that is highly correlated with Z and (approximately) uncorrelated with u. Then, if the inverse exists, a reasonable estimate would be $(W'Z)^{-1}W'y$. This instrumental variable (IV) approach will not be explored directly, but instead will emerge from our analysis of maximum likelihood and minimum distance estimation. We discuss each in turn.

2.1 The Maximum Likelihood Approach

If the errors were normally distributed, the log likelihood function for the simultaneous equations system would be

$$(3) \qquad L(B,\Gamma,\Sigma;Y,X) = -\frac{1}{2}\, TG\cdot\log(2\pi) \ - \ \frac{1}{2}\, T\cdot\log\,\det(\Sigma) \ + \ T\cdot\log|\det(B)|$$

$$- \frac{1}{2}\,\mathrm{tr}[(YB + X\Gamma)'(YB + X\Gamma)\Sigma^{-1}] \quad .$$

For the remainder of sections 2 and 3 we shall assume that this is the actual distribution of the data. In Section 4, we explore the consequences of relaxing this assumption.

According to the method of maximum likelihood, estimators are obtained by maximizing (3) over the values of B, Γ, and Σ satisfying the prior restrictions. This can be performed by calculating the derivatives of L with respect to the unrestricted parameters and setting these scores to zero. Define $L_\delta = \partial L/\partial\delta$ and $L_\sigma = \partial L/\partial\sigma$, where $\sigma \equiv \text{vech}(\Sigma)$ is the $\frac{1}{2}G(G+1)$ vector containing the distinct unknown elements of Σ. Then, if H is the matrix of zeros and ones such that $\text{vec}(\Sigma) = H\cdot\text{vech}(\Sigma)$, the scores for δ and σ are

(4) $$L_\delta = S_\delta'\text{vec}[T(B^{-1}\ 0)' - (Y\ X)'U\Sigma^{-1}]$$

$$= S_\delta'\text{vec}[(B^{-1}\ 0)'(T\Sigma - U'U)\Sigma^{-1} - (X\Pi\ \ X)'U\Sigma^{-1}]$$

and

(5) $$L_\sigma = -\frac{1}{2}H'\text{vec}(T\Sigma^{-1} - \Sigma^{-1}U'U\Sigma^{-1})$$

$$= -\frac{1}{2}H'(\Sigma^{-1}\otimes\Sigma^{-1})H\cdot\text{vech}(T\Sigma - U'U) ,$$

where $U = YB + X\Gamma$ and $(B^{-1}\ 0)$ is a $G\times(G+K)$ matrix whose last K columns are all zero.

Since $H'(\Sigma^{-1}\otimes\Sigma^{-1})H$ is invertible, the equation $L_\sigma = 0$ has the unique solution $\Sigma = U'U/T$. Unfortunately, the equation $L_\delta = 0$ is highly nonlinear and cannot be easily solved. However, an interesting interpretation of this equation has been pointed out by Durbin (1988), Hausman (1975), and Hendry (1976). Inserting the first-order condition $T\Sigma - U'U = 0$ into (4), we obtain the *concentrated score*[2] for δ

(6) $$L_\delta^* = -S_\delta'\text{vec}[(X\Pi\ \ X)'U\Sigma^{-1}] = -S_\delta'[\Sigma^{-1}\otimes(X\Pi\ \ X)]'u$$

$$= \bar{Z}'(\Sigma^{-1}\otimes I_T)u ,$$

[2] For the log likelihood $L(\delta,\sigma)$, the concentrated function is defined as $L^*(\delta) \equiv L[\delta,h(\delta)]$ where $L_\sigma[\delta,h(\delta)] = 0$ so that $\partial h'/\partial\delta = -L_{\delta\sigma}L_{\sigma\sigma}^{-1}$. Differentiation of L^* gives $L_\delta^* = L_\delta[\delta,h(\delta)]$ and $L_{\delta\delta}^* = L_{\delta\delta} - L_{\delta\sigma}L_{\sigma\sigma}^{-1}L_{\sigma\delta}$. Thus, (6) is the score of the concentrated likelihood function and its derivative (when inverted) is $L^{\delta\delta}$, the block of the inverse Hessian of L corresponding to δ.

where $\bar{Z} = -[(I_G \otimes (X\Pi\ X)]S_\delta$. Assuming the appropriate matrices are invertible, the full information maximum likelihood (FIML) estimator which satisfies the equation $L_\delta^* = 0$ has the instrumental variable representation

$$\hat{\delta}_{ML} = [\hat{Z}'(\hat{\Sigma}^{-1} \otimes I_T)Z]^{-1}\hat{Z}'(\hat{\Sigma}^{-1} \otimes I_T)y, \tag{7}$$

where $\hat{Z} = -[I_G \otimes (X\hat{\Pi}\ \ X)]S_\delta$ and $\hat{\Sigma} = \hat{U}'\hat{U}/T$ are functions of $\hat{\delta}_{ML}$. According to the definition of $\hat{Z}$, the endogenous variables in Z are replaced by their FIML fitted values to form instruments which, when combined using generalized least squares, are interpreted as optimal instruments.

Rothenberg and Leenders (1964) derived the asymptotic covariance matrix of the FIML estimator by taking derivatives of the concentrated score function (6). Alternatively, it can be obtained from the asymptotic variance of the concentrated score. That is, the asymptotic covariance matrix for $\sqrt{T}(\hat{\delta}_{ML} - \delta)$ is given by the inverse of

$$\begin{aligned} \text{Avar}(T^{-1/2}L_\delta^*) &= S_\delta'[\Sigma^{-1} \otimes (\Pi\ \ I_K)'\text{plim}\frac{X'X}{T}(\Pi\ \ I_K)]S_\delta \\ &= \text{plim}\,\frac{1}{T}\,\bar{Z}'(\Sigma^{-1} \otimes I_T)\bar{Z} \equiv \Phi_1 . \end{aligned} \tag{8}$$

Equation (7) is only an implicit function for the FIML estimator but other asymptotically equivalent estimators take the same form and are easier to compute. For example, the three stage least squares (3SLS) estimator proposed by Zellner and Theil (1962) and the full information instrumental variables estimator (FIVE) proposed by Brundy and Jorgenson (1971) can both be written as

$$[\tilde{Z}'(\tilde{\Sigma}^{-1} \otimes I_T)Z]^{-1}\tilde{Z}'(\tilde{\Sigma}^{-1} \otimes I_T)y$$

where

$$\tilde{Z} = [(I_G \otimes (X\tilde{\Pi}\ X)]S_\delta\ , \qquad \tilde{\Sigma} = \tilde{U}'\tilde{U}/T$$

and $\tilde{\Pi}$ and $\tilde{U}$ are calculated from consistent but inefficient estimators of B and Γ. They can be interpreted as linearized maximum likelihood estimators (LMLE) since they are of the form

$$\hat{\delta}_{LML} = \tilde{\delta} + T^{-1}\tilde{\Phi}_1^{-1}\tilde{L}_\delta^* ,$$

where $\tilde{\delta}$ is some initial estimator, $T^{-1}\tilde{\Phi}_1^{-1}$ is some estimate of the variance of the MLE, and $\tilde{L}_\delta^*$ is the concentrated score for δ evaluated at $\tilde{\delta}$.[3] The asymptotic equivalence of these estimators with FIML can be demonstrated if appropriate regularity conditions are satisfied.

2.2 The Minimum Distance Approach

Rather than work with the likelihood function, a distance function in sufficient statistics provides an alternative route to asymptotically efficient estimators. Let $M = I - X(X'X)^{-1}X'$ and $n = GK + \frac{1}{2}G(G+1)$. Then, under normality, the distinct elements of the matrices $(X'X)^{-1}X'Y$ and $Y'MY/(T-K)$ are sufficient for the unknown elements of (A, Σ). These statistics can be written as an n-dimensional vector s with mean μ:

$$s = \begin{bmatrix} \text{vec } (X'X)^{-1}X'Y \\ \text{vech } \frac{Y'MY}{T-K} \end{bmatrix} \qquad \mu(A,\Sigma) = \begin{bmatrix} -\text{ vec } \Gamma B^{-1} \\ \text{vech } B'^{-1}\Sigma B^{-1} \end{bmatrix}.$$

The vector $\sqrt{T}(s - \mu)$ has mean zero and variance matrix V_s given by

$$V_s(A,\Sigma) = \begin{bmatrix} T(B\Sigma^{-1}B' \otimes X'X)^{-1} & 0 \\ 0 & \frac{T}{T-K}J'(B\Sigma^{-1}B' \otimes B\Sigma^{-1}B')^{-1}J \end{bmatrix} ,$$

where $J \equiv H(H'H)^{-1}$ has the property that $\text{vech}(\Sigma) = J'\text{vec}(\Sigma)$. The minimum

---

[3] Although the method was introduced by Rothenberg and Leenders (1964), the LMLE interpretation of 3SLS and FIVE was made by Hendry (1976).

distance approach finds parameter values that make the estimated mean $\mu(\hat{A},\hat{\Sigma})$ as close as possible to the observed s. If $\tilde{V}_s$ is some consistent estimate of V_s, it is natural to consider estimates that minimize

$$T[s - \mu(A,\Sigma)]'\tilde{V}_s^{-1}[s - \mu(A,\Sigma)]$$

subject to $R_\delta' \mathrm{vec}(A) = r$. Under regularity conditions, this minimum distance (MD) procedure produces consistent and asymptotically efficient estimates (Cf. Malinvaud (1970, 348-360) and Rothenberg (1973, 24-25)).

As in maximum likelihood, the actual solution can be rather difficult because μ is a complicated nonlinear function. A modification of the minimum distance approach is more tractable. Let $g(s,A,\Sigma)$ be a vector of n smooth functions. Suppose that, for all parameter values (A^*,Σ^*) in an open neighborhood of the true value, $g(s,A^*,\Sigma^*)$ is a locally one-to-one function of s. Suppose further that, as the sample size T tends to infinity, plim $g(s,A,\Sigma) = 0$ and $\sqrt{T}g(s,A,\Sigma)$ is asymptotically normal with zero mean and nonsingular covariance matrix $V_g(A,\Sigma)$. For some consistent estimate $\tilde{V}_g$, minimize

$$T[g(s,A,\Sigma)]'\tilde{V}_g^{-1}[g(s,A,\Sigma)] \ ,$$

subject to the constraints. Under appropriate regularity conditions, the solution to this problem is also consistent and asymptotically efficient. By cleverly choosing the function g, we can construct a computationally convenient estimate of the structural parameters.

The natural choice for g turns out to be

$$\begin{bmatrix} g_\delta \\ g_\sigma \end{bmatrix} = \begin{bmatrix} \frac{1}{T}\mathrm{vec}(X'YB + X'X\Gamma) \\ \mathrm{vech}(\frac{B'Y'MYB}{T-K} - \Sigma) \end{bmatrix} = \begin{bmatrix} \frac{1}{T}\mathrm{vec}\ X'U \\ \mathrm{vech}(\frac{U'MU}{T-K} - \Sigma) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{T}(B' \otimes X'X) & 0 \\ 0 & J'(B \otimes B)'H \end{bmatrix} [s - \mu(A,\Sigma)]$$

with

$$V_g = \begin{bmatrix} \frac{1}{T}(\Sigma \otimes X'X) & 0 \\ 0 & 2\frac{T}{T-K}J'(\Sigma \otimes \Sigma)J \end{bmatrix} .$$

In this case, the original problem has been simplified by "cancelling" the common terms involving B^{-1} that appear in both μ and V_s. For some consistent estimator $\tilde{\Sigma}$, the resulting modified distance function is[4]

$$(9) \qquad T^2 g_\delta'(\tilde{\Sigma} \otimes X'X)^{-1} g_\delta + \frac{1}{2}(T-K) g_\sigma' H'(\tilde{\Sigma} \otimes \tilde{\Sigma})^{-1} H g_\sigma .$$

When there are no restrictions on Σ, the minimization of (9) is very simple. For any value of A, g_σ can be set to zero by varying Σ. Hence, minimizing (9) over A and Σ is equivalent to minimizing

$$T^2 g_\delta'(\tilde{\Sigma} \otimes X'X)^{-1} g_\delta = (y - Z\delta)'[\tilde{\Sigma}^{-1} \otimes X(X'X)^{-1}X'](y - Z\delta)$$

over δ alone.[5] Denoting the idempotent matrix $X(X'X)^{-1}X'$ by N_X, the solution is the three stage least squares estimator

$$[Z'(\tilde{\Sigma}^{-1} \otimes N_X)Z]^{-1} Z'(\tilde{\Sigma}^{-1} \otimes N_X)y,$$

which is a LMLE using the preliminary estimates $\tilde{\Sigma}$ and $\tilde{\Pi} = (X'X)^{-1}X'Y$.

---

[4] We use that fact that the inverse of $J'(\Sigma \otimes \Sigma)J$ is $H'(\Sigma \otimes \Sigma)^{-1}H$. Cf. Richard (1975) and Rothenberg (1973, pp. 87-88). Hausman (1983, p. 421) questions this way of representing the precision of a variance matrix, apparently because he does not use our definition of J.

[5] In general, let $Q = x_1'V^{11}x_1 + x_2'V^{22}x_2 + 2x_1'V^{12}x_2$, where the V^{ij} are appropriate blocks of the inverse variance matrix. If x_2 is freely variable, the minimizing value of x_2 is $-(V^{22})^{-1}V^{21}x_1$ and hence the minimum of Q is obtained by minimizing $x_1'[V^{11} - V^{12}(V^{22})^{-1}V^{21}]x_1 = x_1'V_{11}^{-1}x_1$.

3. Partially Restricted Covariance Matrix

Suppose p of the elements of vech(Σ) are known to be zero. Then, as in (3), this information can be expressed in terms of selection matrices:

$$(10) \qquad R_\sigma' \mathrm{vec}(\Sigma) = 0 \qquad\qquad S_\sigma' \mathrm{vec}(\Sigma) = \sigma \ ,$$

where σ is the vector containing the $\frac{1}{2}G(G+1)-p$ unknown elements.[6] Again, an explicit solution is given by $\mathrm{vec}(\Sigma) = S_\sigma \sigma$. The total prior information on the structural parameters (which we assume to be sufficient for identifiability) can be expressed either by the pair of constraint equations

$$R_\delta' \mathrm{vec}(A) = r \qquad\qquad R_\sigma' \mathrm{vec}(\Sigma) = 0$$

or, equivalently, by the pair of equations relating the constrained parameters to the free parameters δ and σ:

$$\mathrm{vec}(A) = S_\delta \delta + R_\delta r \qquad\qquad \mathrm{vec}(\Sigma) = S_\sigma \sigma \ .$$

Our problem now is to maximize the likelihood function (3) or to minimize the modified distance function (9) subject to these restrictions.

3.1 Maximum Likelihood

Without covariance restrictions, concentrating the log likelihood function with respect to Σ yields the IV interpretation of FIML. It is also a route to simplifying the derivation of linearized approximations to the maximum likelihood estimator. In particular, the asymptotic covariance matrix for $\hat{\delta}_{ML}$, a key ingredient in developing such an approximation, can be calculated from the score of the concentrated log likelihood function, rather than from a partitioned inverse of the full information matrix

---

[6] R_σ and S_σ are assumed to lie in the column space of H so that σ_{ij} and σ_{ji} are treated symmetrically in (10).

of the likelihood function. In the presence of covariance restrictions, concentration of the log likelihood is much less convenient since an explicit closed form solution of the normal equations for σ does not appear to be possible. However, such solutions are not necessary to take advantage of the LML method or to obtain an IV interpretation of the FIML estimates.

The score for δ given in equation (4) remains valid in the presence of covariance restrictions, but equation (5) must be replaced by the score for the free parameters σ:

$$(11) \qquad L_\sigma = -\frac{1}{2} S_\sigma'(\Sigma^{-1} \otimes \Sigma^{-1})\text{vec}(\Sigma - U'U/T) \ .$$

Although equating L_σ to zero does not yield a simple solution for Σ in terms of U when S_σ has rank less than $\frac{1}{2}G(G+1)$, an implicit relationship can still be obtained. Substituting $\text{vec}(\Sigma) = S_\sigma \sigma$ into (11), we find the "solution"

$$\sigma = [S_\sigma'(\Sigma^{-1} \otimes \Sigma^{-1})S_\sigma]^{-1}S_\sigma'(\Sigma^{-1} \otimes \Sigma^{-1})\text{vec}(U'U/T)$$

and hence[7]

$$\begin{aligned}\text{vec}(T\Sigma - U'U) &= \{S_\sigma[S_\sigma'(\Sigma^{-1} \otimes \Sigma^{-1})S_\sigma]^{-1}S_\sigma'(\Sigma^{-1} \otimes \Sigma^{-1}) - I_{G^2}\}H\cdot\text{vech}(U'U) \\ &= -(\Sigma \otimes \Sigma)R_\sigma[R_\sigma'(\Sigma \otimes \Sigma)R_\sigma]^{-1}R_\sigma'\text{vec}(U'U) \ .\end{aligned}$$

Substituting this expression into equation (4) gives, as a generalization of (6), the concentrated score function

$$\begin{aligned}(12) \qquad L_\delta^* &= -S_\delta'\{[\Sigma^{-1} \otimes (X\Pi\ X)]' + [I_G \otimes (\Sigma B^{-1}\ 0)]'\Xi(I_G \otimes U)'\}u \\ &= (\bar{Z} + W)'(\Sigma^{-1} \otimes I_T)u\end{aligned}$$

---

7 The final expression follows, after a rotation of the coordinate system, from the identity $H(H'H)^{-1}H' = R_\sigma(R_\sigma'R_\sigma)^{-1}R_\sigma' + S_\sigma(S_\sigma'S_\sigma)^{-1}S_\sigma'$ which describes the fact that the column space of H is the direct sum of the column space of R_σ and the column space of S_σ.

where

$$\Xi = R_\sigma[R_\sigma'(\Sigma \otimes \Sigma)R_\sigma]^{-1}R_\sigma'$$

$$\overline{Z} = -[(I_G \otimes (X\Pi \ X)]S_\delta$$

$$W = -(\Sigma \otimes U)\Xi[I_G \otimes (\Sigma B^{-1} \ 0)]S_\delta .$$

The generalized IV form of FIML in (7) changes only in the instruments for the endogenous explanatory variables. With covariance restrictions, we have

$$\hat{\delta}_{ML} = [(\hat{Z} + \hat{W})'(\hat{\Sigma}^{-1} \otimes I_T)Z]^{-1}(\hat{Z} + \hat{W})'(\hat{\Sigma}^{-1} \otimes I_T)y,$$

where $\hat{Z}$, $\hat{W}$, and $\hat{\Sigma}$ are the FIML estimates of $\overline{Z}$, W, and Σ.

The matrix $\overline{Z} + W$ has a simple interpretation as an optimal instrument for Z. Using the transformation matrix $\Lambda = (\Sigma^{-1/2} \otimes I_T)$ where $\Sigma^{1/2}$ is the symmetric square root of Σ, our model can be written as

$$\Lambda y = \Lambda Z\delta + \Lambda u$$

where Λu has a scalar covariance matrix. The transformed matrix of predetermined variables are valid instruments since $\mathcal{E}[\Lambda(I_G \otimes X)]'\Lambda u = 0$. In addition, the covariance restrictions imply that $\mathcal{E}[\Lambda(\Sigma \otimes U)R_\sigma]'\Lambda u = 0$. The population projection of ΛZ onto the space spanned by the GK + p columns of $[\Lambda(I_G \otimes X) \quad \Lambda(\Sigma \otimes U)R_\sigma]$ yields $\Lambda(\overline{Z} + W)$ as the linear combination most highly correlated with ΛZ. In the presence of covariance restrictions, structural disturbances that are uncorrelated with the errors augment the reduced form regression function for the endogenous variables, thereby increasing the amount of variation in the instruments.

The asymptotic covariance matrix for the maximum likelihood estimator is, under suitable regularity assumptions, given by the inverse of the limiting Hessian of the concentrated likelihood function. Again, an easier route is via the asymptotic variance of the concentrated score. From (12), L_{δ}^{*} is the sum of two uncorrelated terms

$$L_{\delta 1}^{*} = - S_{\delta}'[\Sigma^{-1} \otimes (X\Pi\ X)]'u = \bar{Z}'(\Sigma^{-1} \otimes I_T)u$$

$$L_{\delta 2}^{*} = - S_{\delta}'[I_G \otimes (\Sigma B^{-1}\ 0)]'\Xi \mathrm{vec}(U'U) = W'(\Sigma^{-1} \otimes I_T)u.$$

Using the fact that var[vech(U'U)] is $2T\ J'(\Sigma \otimes \Sigma)J$, the asymptotic variance for $\sqrt{T}(\hat{\delta}_{ML} - \delta)$ is given by the inverse of

$$S_{\delta}'\{[\Sigma^{-1} \otimes (\Pi\ I_K)'\mathrm{plim}\frac{X'X}{T}(\Pi\ I_K)] + 2[I_G \otimes (\Sigma B^{-1}\ 0)]'\Xi[I_G \otimes (\Sigma B^{-1}\ 0)]\}S_{\delta}$$

(13)

$$= \mathrm{plim}\ \frac{1}{T}\ \bar{Z}'(\Sigma^{-1} \otimes I_T)\bar{Z} + 2\ \mathrm{plim}\ \frac{1}{T}\ W'(\Sigma^{-1} \otimes I_T)W \equiv \Phi_1 + 2\Phi_2 .$$

Note that the instruments $\bar{Z}$ and W contribute differently to the variance. Although the rows of ΛW are uncorrelated with the corresponding rows of Λu, W and u are not independent. Hence, the variance of $W'(\Sigma^{-1} \otimes I_T)u$ depends on the fourth moments of the error distribution. Under normality, this variance is exactly twice the term that would occur if W were exogenous.

3.2 Feasible IV Estimation

When there are no covariance constraints, linearized maximum likelihood simply involves replacing in (7) the FIML values $\hat{Z}$ and $\hat{\Sigma}$ by ones based on preliminary consistent (but inefficient) estimates. By analogy, one is tempted to form the estimator

$$[(\tilde{Z} + \tilde{W})'(\tilde{\Sigma}^{-1} \otimes I_T)Z]^{-1}(\tilde{Z} + \tilde{W})'(\tilde{\Sigma}^{-1} \otimes I_T)y,$$

where $\tilde{Z}$, $\tilde{W}$, and $\tilde{\Sigma}$ are estimates of $\bar{Z}$, W, and Σ based on preliminary consistent estimates of δ. This, however, will not lead to an efficient estimator in the present situation because the dependence between W and u is ignored.

A linearized maximum likelihood estimator for δ has the form

$$\hat{\delta}_{LML} = \tilde{\delta} + T^{-1}(\tilde{\Phi}_1 + 2\tilde{\Phi}_2)^{-1}\tilde{L}^*_\delta$$

where $\tilde{\delta}$ is some initial estimator, $T^{-1}(\tilde{\Phi}_1 + 2\tilde{\Phi}_2)^{-1}$ is some estimate of the variance of the MLE and $\tilde{L}^*_\delta$ is the (concentrated) score for δ evaluated at $\tilde{\delta}$. Although $(\bar{Z} + W)'(\Sigma^{-1} \otimes I_T)u$ is the correct score, $(\bar{Z} + W)'(\Sigma^{-1} \otimes I_T)Z/T$ converges in probability to $\Phi_1 + \Phi_2$ instead of the required $\Phi_1 + 2\Phi_2$. An appropriate LMLE is

$$\begin{aligned} \hat{\delta} &= \tilde{\delta} + [(\tilde{Z} + 2\tilde{W})'(\tilde{\Sigma}^{-1} \otimes I_T)Z]^{-1}(\tilde{Z} + \tilde{W})'(\tilde{\Sigma}^{-1} \otimes I_T)(y - Z\tilde{\delta}) \\ &= [(\tilde{Z} + 2\tilde{W})'(\tilde{\Sigma}^{-1} \otimes I_T)Z]^{-1}[(\tilde{Z} + 2\tilde{W})'(\tilde{\Sigma}^{-1} \otimes I_T)y - \tilde{W}'(\tilde{\Sigma}^{-1} \otimes I_T)\tilde{u}]. \end{aligned} \tag{14}$$

3.3 Minimum Distance

When there are covariance restrictions, it is no longer true that the minimum distance estimator is 3SLS. The second term in (9) must now be taken into account. Since R_σ and S_σ are selection matrices, the elements of the vector g can be grouped into three subvectors:

$$g_\delta = \mathrm{vec}\frac{X'U}{T}, \qquad g_{\sigma 1} = R_\sigma'\mathrm{vec}\frac{U'MU}{T-K}, \qquad g_{\sigma 2} = S_\sigma'\mathrm{vec}\frac{U'MU}{T-K} - \sigma .$$

Note that g_δ and $g_{\sigma 1}$ depend only on δ and that, for any value taken by δ, $g_{\sigma 2}$ can be set arbitrarily by varying σ. Hence, by the argument given in footnote 5, $g_{\sigma 2}$ can be concentrated out of (9). Then, with

$$V = T\cdot\mathrm{Var}\begin{bmatrix} g_\delta \\ g_{\sigma 1}\end{bmatrix} = \begin{bmatrix} V_{\delta\delta} & V_{\delta\sigma} \\ V_{\sigma\delta} & V_{\sigma\sigma}\end{bmatrix} = \begin{bmatrix} \frac{1}{T}(\Sigma \otimes X'X) & 0 \\ 0 & \frac{2T}{T-K} R_\sigma'(\Sigma \otimes \Sigma)R_\sigma \end{bmatrix}, \tag{15}$$

the appropriate distance function is

$$Q(\delta) = [\mathrm{vec}(X'U)]'[\Sigma \otimes X'X]^{-1}\mathrm{vec}(X'U)$$

(16)

$$+ \frac{1}{2(T-K)} [\mathrm{vec}(U'MU)]'R_\sigma[R_\sigma'(\tilde{\Sigma} \otimes \tilde{\Sigma})R_\sigma]^{-1}R_\sigma'\mathrm{vec}(U'MU) \ .$$

In the presence of covariance restrictions, the problem involves minimizing a quartic function in δ. Given the additive form of the objective function, the solution should not be difficult to calculate.

The asymptotic variance of the minimum distance estimator is given by twice the inverse of the average Hessian of the distance function. That is, it is the inverse of

(17) $$\mathrm{plim}\ \frac{1}{T} Z'(\Sigma^{-1} \otimes N_X)Z \ + \ 2\ \mathrm{plim}\ \frac{1}{T^2} Z'(I_G \otimes U)\Xi(I \otimes U)'Z$$

which is the same expression as obtained in (13) for the FIML estimator since $\mathrm{plim}\ \frac{1}{T}(I_G \otimes U)'Z = -[I_G \otimes (\Sigma B^{-1}\ 0)]S_\delta$.

Just as with maximum likelihood, there exists an efficient linearized minimum distance estimator that is easier to calculate. Let $\tilde{U} = (Y\ X)\tilde{A}$ be the residuals based on some consistent estimator $\tilde{A}$. Then, using a first-order approximation to the quadratic function, we have

$$R_\sigma'\mathrm{vec}(U'MU) \cong R_\sigma'\mathrm{vec}(\tilde{U}'MU + U'M\tilde{U} - \tilde{U}'M\tilde{U}) = R_\sigma'(I_G \otimes M\tilde{U})'(2u - \tilde{u}),$$

which can be viewed as a linearization of $g_{\sigma 1}$ around the value $A = \tilde{A}$. With $R_\sigma'\mathrm{vec}(U'MU)/\sqrt{T-K}$ replaced by the asymptotically equivalent $R_\sigma'\mathrm{vec}(U'U)/\sqrt{T}$, we have the simple quadratic distance function

$$Q^*(\delta) = u'(\tilde{\Sigma}^{-1} \otimes N_X)u$$

(18)

$$+ \frac{1}{2T} (2u - \tilde{u})'(I_G \otimes \tilde{U})R_\sigma[R_\sigma'(\tilde{\Sigma} \otimes \tilde{\Sigma})R_\sigma]^{-1}R_\sigma'(I_G \otimes \tilde{U})'(2u - \tilde{u}) \ .$$

The resulting linearized minimum distance estimator is identical to the linearized maximum likelihood estimator (14) if $\tilde{Z} = [I \otimes X(X'X)^{-1}X']Z$ and $\tilde{W} = \frac{1}{T}(\tilde{\Sigma} \otimes \tilde{U})\Xi(I_G \otimes \tilde{U})'Z$ are used to form the latter.

The linearized minimum distance estimator is also equal to a version of the augmented three stage least squares (A3SLS) estimator proposed by Hausman, Newey, and Taylor (1987). The demonstration, however, requires some additional notation. Let u_t be the t^{th} column of U' and define $e_t = R_\sigma' \text{vec}(u_t u_t')$. Form the $p \times T$ matrix $E' = [e_1, \ldots, e_T]$. Then, denoting the T-dimensional vector of ones by ι and setting $e = \text{vec}(E)$, we have $R_\sigma' \text{vec}(U'U) = E'\iota = (I_p \otimes \iota')e$. Hence, if $R_\sigma' \text{vec}(U'MU)/\sqrt{T-K}$ is replaced by $R_\sigma' \text{vec}(U'U)/\sqrt{T}$, equation (16) is equal to

$$u'(\tilde{\Sigma}^{-1} \otimes N_X)u \quad + \quad e'[(R_\sigma'(\tilde{\Sigma} \otimes \tilde{\Sigma})R_\sigma)^{-1} \otimes N_\iota]e$$

where $N_\iota = \iota(\iota'\iota)^{-1}\iota'$. Corresponding to the linearization of $R_\sigma' \text{vec}(U'U)$ around $R_\sigma' \text{vec}(\tilde{U}'\tilde{U})$, there is an equivalent linearization of e around $\tilde{e}$. Defining $Z_e = -\partial e/\partial \delta'$, e is approximated by $\tilde{e} - \tilde{Z}_e(\delta - \tilde{\delta}) \equiv \tilde{y}_e - \tilde{Z}_e \delta$. In this notation, (18) is identical to

$$(19) \quad (y-Z\delta)'(\tilde{\Sigma}^{-1} \otimes N_X)(y-Z\delta) + \frac{1}{2}(\tilde{y}_e - Z_e\delta)'[(R_\sigma'(\tilde{\Sigma} \otimes \tilde{\Sigma})R_\sigma)^{-1} \otimes N_\iota](\tilde{y}_e - \tilde{Z}_e\delta)$$

which is the A3SLS objective function when normality is imposed on the augmented error covariance matrix.

4. Estimation when the Errors are Not Normal

The quasi maximum likelihood estimator that maximizes (3) and the minimum distance estimator based on (16) are still consistent but are no longer asymptotically efficient if the errors are not normal. Furthermore, the expressions for the asymptotic variance in (13) and (17) may be wrong. Some adjustments to our statistics can get around these problems. It will be more convenient to work within the minimum distance framework, but analogous results are available using the likelihood approach.

Two potential problems occur in our analysis in section 3.3 when the errors are not normal. First, s is no longer a vector of sufficient statistics for the unknown parameters. Optimal estimates need to make use of additional sample data. Unfortunately, unless we make further parametric assumptions (or switch to semiparametric methods), it is not clear what additional sample data is relevant. Even if we restrict our attention to estimators that are functions only of s, there is a second problem. The minimum distance estimates behave asymptotically like linear functions of s, where the weights depend on the elements of the variance matrix V in equation (15). Since we have used the normality assumption to form V, the estimators derived here may be inefficient because the wrong weights are employed.

Note that $T^{-1}(\tilde{\Sigma} \otimes X'X)$ is a consistent estimate of $V_{\delta\delta}$ for any error distribution as long as $\tilde{\Sigma}$ is consistent for Σ. Hence, if $\tilde{V}_{\delta\sigma}$ is zero (because, for example, the errors are assumed to have a symmetric probability distribution), the distance function (16) is correct except for the use of the matrix $2R_\sigma'(\tilde{\Sigma} \otimes \tilde{\Sigma})R_\sigma$ as an estimate of $V_{\sigma\sigma}$. Assuming fourth moments of the error distribution exist, this matrix can be replaced in equations (16) – (19) by a more robust estimate, say $\tilde{E}'\tilde{E}/T$. No other changes need be made to obtain the optimal minimum distance estimator based on s and the correct asymptotic variance.

If asymmetry of the error distribution is suspected, matters are considerably more complicated. As pointed out by Hausman, Newey, and Taylor (1987), when third moments are not a priori zero, the fact that the elements of X are uncorrelated with certain products of the errors is new information that can be exploited. Using the matrix E defined above, it can be expressed by the fact that both X′E and X′U have mean zero. If ι is a column of X, the elements of g_δ and $g_{\sigma 1}$ are a subset of the elements of $\frac{1}{T}\text{vec}[X'(U\ E)]$. It is natural then to consider a minimum distance estimator based on the expanded set of moment restrictions.

The T(G + p) vector $\epsilon = \text{vec}(U\ E)$ has mean zero and variance matrix

$$\text{Var}\ \epsilon \ = \ \text{Var}\begin{bmatrix} u \\ e \end{bmatrix} \ = \ [\Omega \otimes I_T] \qquad \text{for } \Omega = \text{Var}\begin{bmatrix} u_t \\ e_t \end{bmatrix} = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}.$$

Of course, Ω_{11} is just Σ and Ω_{22} is $V_{\sigma\sigma}$; preliminary estimates for them have been discussed already. The matrix Ω_{12} can be estimated by $\tilde{U}'\tilde{E}/T$. Since $\text{vec}[X'(U\ E)] = (I \otimes X)'\epsilon$ has variance matrix $(\Omega \otimes X'X)$, the appropriate distance function is

$$(20) \qquad \epsilon'(\tilde{\Omega}^{-1} \otimes N_X)\epsilon \ = \ u'(\tilde{\Omega}^{11} \otimes N_X)u + e'(\tilde{\Omega}^{22} \otimes N_X)e + 2u'(\tilde{\Omega}^{12} \otimes N_X)e.$$

If e is linearized as in (19), the resulting minimum distance estimator is the A3SLS estimator.[8]

The asymptotic covariance matrix for the MD estimator based on X′(U E) is $[\text{plim}\ \frac{1}{T}Z_A'(\Omega^{-1} \otimes N_X)Z_A]^{-1}$, where $Z_A = \partial\epsilon/\partial\delta'$. Using the the same notation, the asymptotic covariance matrix for the best MD estimator based on the subset $(X'U\ \iota'E)$ can be written as the inverse of

$$\text{plim}\ \frac{1}{T}Z_A'(\Omega^{-1} \otimes N_X)Z_A \ - \ \text{plim}\ \frac{1}{T}Z'[\Omega^{12}(\Omega^{22})^{-1}\Omega^{21} \otimes N_X - N_\iota]Z\ .$$

Since the term being subtracted is positive semidefinite (and nonzero if $\Omega_{12} \neq 0$), the estimator using third-moments of the data in addition to the second moments is strictly more efficient whenever u_t is correlated with e_t.

---

[8] Cf. Hausman, Newey, and Taylor (1987), equation (4.5a).

In practice, third and fourth sample moments of regression residuals are often unreliable estimates of the corresponding population moments. Robust estimates of δ using these sample moments will probably not be an improvement over estimates based on normality unless the sample size is quite large. Furthermore, covariance restrictions often arise from the belief that certain error terms are independent of other error terms. Under independence, however, cross third moments are zero and the efficiency gain from exploiting $\mathcal{E}(X'E) = 0$ disappears. Hence, it is an open question whether the estimators using third and fourth moments will be useful in practice.

5. Conclusion

The full information maximum likelihood and minimum distance estimators for a simultaneous equations model with covariance restrictions are rather complicated. But linearization yields estimators that are easy to compute and have an instrumental variable interpretation. All of these estimators are functions of the second order moment matrices of the data and have a simple method of moments interpretation: the p covariance restrictions are added to the usual GK exogeneity conditions to form mean functions for approximate minimum distance estimation. Assuming a symmetric error distribution and a sufficiently large sample size, the estimators can easily be modified to make them robust to nonnormality. When symmetry is dropped, third order moments of the data may be used to construct improved estimators. It appears that $K(G + p)$ moment conditions are needed when cross third moments of the errors are nonzero.

6. References

Brundy, J. M., and D. W. Jorgenson (1971), "Efficient Estimation of Simultaneous Equations by Instrumental Variables," *Review of Economics and Statistics*, 53, 207-224.

Durbin, James (1988), "Maximum Likelihood Estimation of the Parameters of a System of Simultaneous Regression Equations," *Econometric Theory*, 4, 159-170.

Hausman, J. A. (1975), "An Instrumental Variable Approach to Full Information Estimators for Linear and Certain Non-Linear Econometric Models," *Econometrica*, 43, 727-738.

Hausman, J. A., W. K. Newey, and W. Taylor (1987), "Efficient Estimation and Identification of Simultaneous Equation Models with Covariance Restrictions," *Econometrica*, 55(4), 849-874.

Hausman, J. A. (1983), "Specification and Estimation of Simultaneous Equation Models," *Handbook of Econometrics, Vol. I*, (Z. Griliches and M. Intriligator, eds.), North-Holland: Amsterdam, chapter 7.

Hendry, D. F. (1976), "The Structure of Simultaneous Equations Estimators," *Journal of Econometrics*, 4, 51-88.

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

Richard, J.-F. (1975), "A Note on the Information Matrix of the Multivariate Normal Distribution," *Journal of Econometrics*, 3, 57-60.

Rothenberg, T. J. and C. T. Leenders (1964), "Efficient Estimation of Simultaneous Equation Systems," *Econometrica*, 32(1-2), 57-76.

Rothenberg, T. J. (1973), *Efficient Estimation with A Priori Information*, Yale University Press: London.

Zellner, A., and H. Theil (1962), "Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations," *Econometrica*, 30, 54-78.

March 1989

Working Paper Series
Department of Economics
University of California, Berkeley

Individual copies are available for $3.50 (in the USA or Canada) and may be obtained from the Institute of Business and Economic Research. Prepayment is required. Make checks or money orders payable to "The Regents of the University of California." Send requests to IBER, 156 Barrows Hall, University of California, Berkeley CA 94720.

8880 "Horizontal Mergers: An Equilibrium Analysis." Joseph Farrell and Carl Shapiro. June 1988.

8881 "Optimal Exclusion and Relocation of Workers in Oversubscribed Industries." Tracy R. Lewis, Roger Ware and Robert Feenstra. June 1988.

8882 "The Gold-Exchange Standard and the Great Depression." Barry Eichengreen. June 1988.

8883 "Cheap Talk, Neologisms, and Bargaining." Joseph Farrell and Robert Gibbons. July 1988.

8884 "A State Space Model of the Economic Fundamentals." Roger Craine and David Bowman. July 1988.

8885 "Settling Defaults in the Era of Bond Finance." Barry Eichengreen and Richard Portes. August 1988.

8886 "Foreign Lending in the Interwar Years: The Bondholders' Perspective." Barry Eichengreen and Richard Portes. August 1988.

8887 "Estimation of the Probability of Acquisition in an Equilibrium Setting." Brownyn H. Hall. August 1988.

8888 "Entry, Acquisition, and the Value of Shark Repellent." Richard J. Gilbert and David M. Newbery. August 1988.

8889 "The Role of Potential Competition in Industrial Organization." Richard J. Gilbert. September 1988.

8890 "Cheap Talk with Two Audiences: A Taxonomy." Joseph Farrell and Robert Gibbons. September 1988.

8891 "Management of a Common Currency." Alessandra Casella and Jonathan Feinstein. September 1988.

8892 "Economic Growth and Generalized Depreciation." Steven M. Goldman and Vai-Lam Mui. September 1988.

8893 "Raiders, Junk Bonds, and Risk." Roger Craine and Douglas Steigerwald. October 1988.

8894 "The Responsibilities of a Creditor Nation." Barry Eichengreen. October 1988.

8895 "Mobility Barriers and the Value of Incumbency." Richard J. Gilbert. October 1988.

8896 "Some Reflections on the Use of the Concept of Power in Economics." Pranab Bardhan. October 1988.

8897 "The U.S. Basic Industries in the 1980s: Can Fiscal Policies Explain Their Changing Competitive Position?" Barry Eichengreen and Lawrence H. Goulder. November 1988.

8898 "The Trend in the Rate of Labor Force Participation of Older Men, 1870-1930: A Review of the Evidence." Roger L. Ransom and Richard Sutch. November 1988.

8899 "Extensions of Estimation Methods Using the EM Algorithm.." Paul A. Ruud. January 1988.

89-100 "Deregulation and Scale Economies in the U. S. Trucking Industry: An Econometric Extension of the Survivor Principle." Theodore E. Keeler. January 20, 1989.

89-101 "Pricing in a Deregulated Environment: The Motor Carrier Experience." John S. Ying and Theodore E. Keeler. January 20, 1989.

89-102 "Optimal Patent Length and Breadth." Richard Gilbert and Carl Shapiro. January 1989.

89-103 "Product Line Rivalry with Brand Differentiation." Richard J. Gilbert and Carmen Matutues. January 1989.

89-104 "Dealing with Debt: The 1930s and the 1980s." Barry Eichengreen and Richard Portes. February 1989.

89-105 "A Comparison of the EM and Newton-Raphson Algorithms." Paul A. Ruud. February 1989.

89-106 "Simultaneous Equations with Covariance Restrictions." Thomas J. Rothenberg and Paul A. Ruud. April 1989.