Annals of Economic and Social Measurement, 3/4, 1974

ESTIMATION AND INFERENCE IN NONLINEAR STRUCTURAL MODELS*

BY E. K. BERNDT, B. H. HALL, R. E. HALL, AND J. A. HAUSMAN

Maximum likelihood and minimum distance estimators are specified for nonlinear structural econometric models. A theorem is proven which insures convergence to a local maximum of the respective likelihood function and distance function. Techniques of inference are developed for both estimators. The maximum likelihood theorem and algorithm are based on the fundamental statistical relation that the covariance matrix of the estimator is equal to the covariance matrix of the gradient of the likelihood function. The algorithm requires much less computation than previous algorithms and, unlike previous algorithms, is guaranteed to converge.

Econometric methods of structural estimation generally assume linearity of the model in both variables and parameters. On the other hand, many contemporary models of economic behavior are both nonlinear and simultaneous. Modern demand analysis, for example, starts from a rich specification of individual tastes for a variety of goods and deals with the structural relation it implies among prices, quantities, and income. This relation is nonlinear in both variables and parameters in all but the simplest cases. Similarly, models of production with many factors are invariably nonlinear in their variables, and are frequently nonlinear in their parameters as well, especially when trends in productivity are present. In this paper we deal with the practical issues that arise in estimating nonlinear structural models. We review the statistical theory of estimation in these models and draw attention to some important recent advances. We also summarize some useful results from the theory of maximization. Our main contribution is a set of algorithms for estimation and inference in nonlinear structural models. These carry out statistical procedures with known desirable properties, embody modern numerical techniques, and are organized to conserve both computation and storage.

1. MODEL AND STATISTICAL THEORY

Throughout we are concerned with estimation and inference in the multivariate structural model,

$$F_t(y_t, \beta) = \varepsilon_t. \tag{1.1}$$

Here y_t is a $1 \times M$ row vector of jointly dependent variables, F_t is a twice-differentiable function whose value is a $1 \times M$ vector, and ε_t is a $1 \times M$ vector of random disturbances, assumed to be distributed according to the multivariate normal probability law, with expectation zero and variance-covariance matrix Σ. The model may involve exogenous variables as well, but these are subsumed under

* The authors are grateful to J. E. Dennis and D. W. Jorgenson for helpful suggestions at several points in this research. Proofs of theorems are omitted due to space constraints but are available from the authors upon request. The authors' affiliations, in the order in which their names are listed above, are as follows: University of British Columbia, Harvard University, Massachusetts Institute of Technology, and Massachusetts Institute of Technology.

the dependence of F_t on the time index, t. The model contains a $K \times 1$ vector of unknown parameters, β. We make no assumptions about the assignment of parameters to the M elements of F_t, so, for example, the same parameter may appear in more than one equation. We assume that β is identifiable; see Fisher (1966), Chapter 5, for a discussion of identification problems in nonlinear structural models.

We discuss two estimators of β: maximum likelihood and minimum distance. Problems of estimation and inference are well understood for maximum likelihood, but the estimator has received little use in econometric work because of the apparent complexity of the calculations it requires. Previous discussions of maximum likelihood (Eisenpress and Greenstadt (1966) and Chow (1973)) do not employ a numerical method of maximization that guarantees convergence. Further, their use of Newton's method requires the formation and calculation of an enormous number of third derivatives of the model, effectively restricting their method to small models. We show in this paper that the third derivatives are both unnecessary and dangerous. By eliminating them we bring about a great simplification of the computations and at the same time achieve a method whose convergence is guaranteed.

Minimum distance methods have formed the basis of most practical work to date on simultaneous estimation of linear structural models. Three-stage least squares is a minimum distance estimator. Recently Amemiya (1974) has extended the theory of minimum distance to nonlinear models. We discuss the application of his method in simultaneous estimation. The minimum distance estimator is substantially easier to compute than is maximum likelihood, but is not generally statistically efficient. However, the estimates are consistent and asymptotically normal, so the method can form the basis for a complete approach to estimation and inference in nonlinear structural models.

2. Gradient Methods for Maximization

In this section we review results on numerical methods of maximization that are familiar to applied mathematicians but have been overlooked in most statistical work. Our essential point is that methods are available whose convergence to at least a critical point is guaranteed in theory. A serious defect of many applications of the method of scoring and other statistical maximization procedures is their failure to use methods with assured convergence.

In general we deal with the maximization of the scalar function $V(x)$ of the vector x of length K. We assume that V is twice continuously differentiable and has compact upper contour sets. The starting point for our analysis is the

Gradient Theorem

Consider the gradient of V at x, $g = \partial V(x)/\partial x$. Then any vector, d, in the same halfspace as g (that is, with $d'g > 0$) is a direction of increase of $V(x)$, in the sense that $V(x + \lambda d)$ is an increasing function of the scalar λ, at least for small enough λ.

This classical result in maximization theory is proved, for example, in Jacoby *et al.*, (1972), p. 97.

A successful method chooses a direction at each iteration that lies in the halfspace defined by the gradient. In general, each iteration consists in computing the gradient, g, deriving from it a direction, d, and then finding a value of λ that maximizes $V(x + \lambda d)$. Any method following this procedure is assured of convergence.

The set of directions, d, that are in the gradient halfspace consists precisely of those that can be derived from the gradient by multiplying it by a positive definite matrix, say Q. Alternative gradient methods are specified succinctly by providing rules for forming Q at each iteration. In general, convergence is speeded by a choice of Q that is close to the inverse of the Hessian matrix of second derivatives of $V(x)$, especially in the neighborhood of the optimum where the use of the Hessian makes final convergence quadratic. If $V(x)$ is concave, then the inverse of the Hessian matrix itself can serve as Q, and we have Newton's method. Even in that case, convergence is guaranteed only if a suitable method is employed for searching for λ at each iteration. In most statistical applications, however, the objective function $V(x)$ cannot be relied upon to be concave, and Newton's method is unsuitable. Dependence on Newton's method is a shortcoming of the work of Eisenpress and Greenstadt (1966) and of Chow (1973) on nonlinear structural estimation.

In statistical work, it is usually convenient to choose Q in a way that makes it approximate the variance-covariance matrix of the estimates. Since the latter is necessarily positive definite, it is eligible as a choice of Q. Later in the paper we will derive easily computed Q's that serve as well as variance-covariance matrices for the maximum likelihood and minimum distance cases. It is necessary, however, to rule out the possibility that Q approaches a singular matrix as the process iterates. For this purpose we state the

Restriction on Q

Let α be a prescribed positive constant less than one. At each iteration we require

$$r = \frac{d'g}{d'd} > \alpha. \tag{2.1}$$

If r drops below α on a particular iteration, we should replace Q by a matrix with larger diagonal elements. Note that the restriction can always be satisfied by $Q = I$, which is an admissible choice.

All gradient methods require a "λ-search" after determining the direction, d. The choice of method for selecting λ involves some subtle issues—not every method yields guaranteed convergence. For example, trying out decreasing values of λ until one is found that gives a higher value of $V(x + \lambda d)$ is inadequate; it can generate an infinite sequence of iterations that do not converge to a point where g is zero. However, a choice of λ that maximizes $V(x + \lambda d)$, while guaranteeing convergence, often imposes an unacceptable computational burden (Powell, 1971). Convergence is assured in the class of problems we consider under the

Criterion for Choice of λ

Let δ be a prescribed constant in the interval $(0, \frac{1}{2})$. Define

$$\gamma(x, \lambda) = \frac{V(x + \lambda d) - V(x)}{\lambda d'g}. \tag{2.2}$$

If $\gamma(x, 1) \geq \delta$, take $\lambda = 1$. Otherwise, choose λ to satisfy

$$\delta \leq \gamma(x, \lambda) \leq 1 - \delta. \tag{2.3}$$

Under our assumptions about $V(x)$, a λ satisfying this criterion will always exist.

Now we can state the

Convergence Theorem

Assume $V(x)$ is twice continuously differentiable and is defined over a compact upper contour set.

Consider the sequence $x^{(1)}, x^{(2)}, \ldots$, where

$$x^{(i+1)} = x^{(i)} + \lambda^{(i)} d^{(i)}, \tag{2.4}$$

$$d^{(i)} = Q^{(i)} g^{(i)}, \tag{2.5}$$

and $Q^{(i)}$ obeys the restriction (2.1) and $\lambda^{(i)}$ satisfies the criterion (2.3). Then $\lim_{i \to \infty} g^{(i)} = 0$.

The proof of this theorem follows Goldstein (1967), page 31, generalized along the lines he suggests on page 36.

Not every critical point of $V(x)$ is a local maximum. If the iterative process chooses a value of x where $V(x)$ has a local minimum or a saddle point, the iterative process will stall, as $g = 0$ at such points. Since the process moves intentionally toward a critical point only if it is a local maximum, stalling elsewhere is only a very remote possibility. The safeguard against this possibility is precisely the same as against convergence to a local maximum that is not a global maximum: choose several initial values of x. If they do not all lead to convergence to the same point, investigate the actual shape of the function with care until the global maximum is located.

3. Estimation and Inference by Maximum Likelihood

Maximum likelihood estimates are known to be statistically efficient; see, for example, Rao (1965), pp. 299–302, and Hausman (1975), who discusses regularity conditions for the structural model. Further, the likelihood ratio test provides a powerful and general method of inference. In structural estimation, however, maximum likelihood has seen little practical use to date because of the apparent complexity of the computations necessary to find the maximum of the likelihood function. Until Hausman's recent work (1974, 1975), maximum likelihood seemed impractical even for linear structural models. Hausman demonstrates that iteration of an instrumental variables estimator with suitably chosen instruments converges to maximum likelihood if it converges at all. However, he does not establish that his method converges.[1] Further, it is still unclear how his method could be extended to models that are nonlinear in both parameters and variables.

In this section we develop a practical approach to maximum likelihood within the framework of gradient methods. Our approach has two substantial advantages over the application of Newton's method advocated by Eisenpress and Greenstadt (1966) and Chow (1973). First, its convergence is assured. Newton's method uses a Q matrix that may not be positive definite and thus fails to confine the direction vector to the gradient halfspace. Second, our method requires the evaluation of derivatives of the model up to second order only, while Newton's method requires certain third derivatives. The sheer number of third derivatives makes Newton's method suitable only for small structural models. We eliminate the third derivatives by taking advantage of a fundamental statistical relation: the asymptotic variance-covariance matrix of a maximum likelihood *estimator* is equal to the variance-covariance matrix of the *gradient* of the likelihood function (Kendall and Stuart (1967), Vol. II, p. 9). As we remarked earlier, it is natural and convenient to use a variance-covariance matrix as the Q matrix in a gradient method. The relevance of this statistical relation to numerical maximization of likelihood functions in econometric applications has apparently not been pointed out before.

We need to maximize the concentrated log-likelihood function of the coefficients, β:

$$L(\beta) = k + \sum_t \log|\det J_t| - \frac{1}{2}\log\det F'F. \tag{3.1}$$

Here k is an inessential constant, and J_t is the Jacobian matrix of the transformation from the underlying disturbances to the observed random variables, y_t,

$$J_t = \frac{\partial F_t(y_t, \beta)}{\partial y_t}. \tag{3.2}$$

The gradient is

$$\frac{\partial L}{\partial \beta} = \sum_t \frac{\partial}{\partial \beta}\log|\det J_t| - \frac{1}{2}\frac{\partial}{\partial \beta}\log\det \sum_t F_t'F_t = p - q. \tag{3.3}$$

The variance-covariance matrix of the gradient is

$$E\left[\left(\frac{\partial L}{\partial \beta}\right)\left(\frac{\partial L}{\partial \beta}\right)'\right] = E[(p - q)(p - q)']. \tag{3.4}$$

Our strategy is to replace the expectation by a statistic with equal expectation.

[1] However, the use of an appropriate λ-search guarantees convergence of Hausman's procedure in the case of linear structural models. His method is related to ours, but he is able to simplify the expression for the gradient by using the linearity of the model.

The natural choice is the sample variance-covariance matrix of the gradient. We define

$$p_t = \frac{\partial}{\partial\beta}\log|\det J_t| \tag{3.5}$$

$$= \sum_k \sum_l (J_t')^{-1}_{k,l} \frac{\partial J_{t,k,l}}{\partial\beta}$$

and

$$q_t = \left(\frac{\partial F_t}{\partial\beta}\right)\left(\sum_\tau F_\tau' F_\tau\right)^{-1} F_t'. \tag{3.6}$$

Then we define R_T as the sample covariance matrix of the gradient multiplied by T^2:

$$R_T = T\sum_t (p_t - q_t)(p_t - q_t)'. \tag{3.7}$$

It is not hard to show that

$$\operatorname{plim} \quad R_T = \lim_{T\to\infty} E\left(\frac{1}{T}(p - q)(p - q)'\right). \tag{3.8}$$

Thus R_T^{-1} is a suitable choice for Q in a gradient method. We summarize our proposed approach in a

Theorem on Computation of Maximum Likelihood Estimators and Their Covariance Matrices

Consider the following iteration:

$$\beta^{(i+1)} = \beta^{(i)} + \lambda^{(i)}(R_T^{(i)})^{-1}(p^{(i)} - q^{(i)}) \tag{3.9}$$

where $\lambda^{(i)}$ is computed by the method of Section 2, $R_T^{(i)}$ from equation (3.7), and $p^{(i)}$ and $q^{(i)}$ from equation (3.3). Then

(i) the method converges to a stationary point of the likelihood function as $i \to \infty$,

(ii) the method is close to Newton's method in that R_T converges to the Hessian matrix of the likelihood function as $T \to \infty$, and

(iii) $(1/T)R_T^{-1}$ is a consistent estimate of the variance-covariance matrix of the estimated parameters.

Note that the assumptions on $V(x)$ place no important restrictions on the likelihood functions encountered with nonlinear structural models.

The following application of the theorem yields the maximum likelihood estimate and a consistent estimate of the associated variance-covariance matrix:

Maximum Likelihood Algorithm

1. Compute the variance-covariance matrix of the residuals from equation (1.1) using the estimated parameter values from the previous iteration.

For the first iteration, arbitrary initial values may be used.

2. Over each observation, compute the Jacobian term, p_t, and the "sum of squares" term, q_t. Update R, p, and q.
3. Calculate the new direction vector, $d = R^{-1}(p - q)$.
4. Check for convergence, defined as

$$\max_i \frac{|d_i|}{\max(1, |\beta_i|)} < \text{prescribed tolerance}$$

5. Search for λ and update β using equation (3.1) and equation (3.9). Return to step 1.
6. Report β and its estimated variance–covariance matrix, $(1/T)R^{-1}$.

The maximum likelihood algorithm may be modified at Step 2 by recalculating R only after several iterations have been carried out. Convergence of the modified algorithm is also assured, and the computational effort in forming R will be reduced. In many cases, however, most of the effort will be consumed in forming p_t and q_t at each iteration, so it is probably better to use the best available approximation to the curvature of the likelihood function.

Likelihood ratio tests are the natural method of inference when maximum likelihood estimates are available. These tests are known to have many desirable properties (Kendall and Stuart, 1967, Vol. II, pp. 224–247). Briefly, if L^* is the maximum of the likelihood function of a structural model which is nested in a larger structural model with maximized likelihood L, then the statistic

$$-2(\log L^* - \log L) \tag{3.10}$$

is distributed asymptotically as χ^2, with degrees of freedom equal to the difference in the numbers of parameters in the two models.

We conclude our discussion of maximum likelihood with the remark that iteration of our algorithm is not required to achieve any of the known desirable properties of the resulting estimator, provided that the initial parameter values are consistent estimates. The asymptotic equivalence of maximum likelihood estimates and estimates obtained from one iteration of Newton's method is well known.[2] Since the matrix R in our procedure converges to the matrix of second derivatives, it follows that one step of our method is asymptotically equivalent to maximum likelihood as well. The step requires much less computation than one step of Newton's method. This justifies the

One-Step Efficient Estimation Algorithm

1. Use the minimum distance algorithm of Section 4 to obtain consistent parameter estimates. β^{MD}. Use these to evaluate p_t and q_t and thus to form R, p, and q.
2. Calculate the direction vector, $d = R^{-1}(p - q)$.

[2] See, for example, Zacks (1971), pp. 250–251. The equivalence was pointed out by Rothenberg and Leenders (1964) for the linear structural model.

3. Calculate the efficient estimates,

$$\beta = \beta^{MD} + d \tag{3.11}$$

Note that λ is taken as one.

4. Calculate the variance-covariance matrix $(1/T)R^{-1}$ using β, and, if needed for inference, the value of the likelihood functions.

Inference is again based on likelihood ratio tests as described earlier.

4. Estimation and Inference by the Minimum Distance Method

The maximum likelihood method discussed in Section 3 yields efficient estimates and powerful tests. These properties are achieved at the computational cost of evaluating second derivatives of the structural model arising from the presence of the Jacobian matrix in the likelihood function. In an important recent paper, Amemiya (1974) has developed a class of estimators for nonlinear structural models that requires the minimization of a quadratic distance function. The distance function contains instrumental variables but no explicit Jacobian matrix. In the linear case, Hausman (1975) shows that a particular set of instruments substitutes exactly for the Jacobian and thus he provides an interpretation of maximum likelihood in terms of instrumental variables. The relation between Amemiya's instrumental variables estimator and maximum likelihood is less clear in the nonlinear case. Amemiya demonstrates only that for arbitrarily chosen instruments, the minimum distance estimator is consistent and asymptotically normal.

It is easiest to deal with Amemiya's prodedure in a "stacked" version of the model:

$$f(y, \beta) = \varepsilon \tag{4.1}$$

where f, y, and ε are $T \cdot M \times 1$ vectors. His estimator minimizes the distance,

$$\Delta(\beta) = \tfrac{1}{2}(f(y, \beta))'Df(y, \beta) \tag{4.2}$$

where D is defined as[3]

$$D = (S^{-1} \otimes I)H(H'(S^{-1} \otimes I)H)^{-1}H'(S^{-1} \otimes I), \tag{4.3}$$

S is an arbitrary $M \times M$ symmetric positive definite matrix, and H is an $MT \times N$ matrix of instrumental variables. Amemiya proves that the value of β that minimizes the distance is a consistent and asymptotically normal estimator of the true β, provided the instruments, H, are independent of the structural disturbances, ε. It is not, in general, an efficient estimator. If f is linear in both y and β, if S is a consistent estimator of the structural variance-covariance matrix Σ, and if H consists of all of the exogenous variables in the model (all of the derivatives of f with respect to β that do not involve y), then the minimum distance estimator is three-stage least squares and is known to be asymptotically efficient. No precise information about efficiency is available when f is nonlinear.[4] Presumably S should be as close

[3] Amemiya deals with the univariate case where D has a simpler form. We start from an obvious multivariate generalization of his results.

[4] Hausman (1975) does prove efficiency of his instrumental variables estimator in the case of a model that is nonlinear in parameters but linear in variables.

as possible to Σ and the instruments should resemble the derivatives of f with respect to β.

The gradient of the distance function is

$$g = G'Df \tag{4.4}$$

where G is the matrix of derivatives of f with respect to β. Again, we seek a point where $g = 0$. Amemiya demonstrates that the asymptotic variance-covariance matrix of the minimum distance estimator is $(G'DG)^{-1}$. As before, this is a suitable choice for the Q matrix in a gradient method: it is positive definite, and its computation is necessary in any case at the conclusion to provide an indication of the sampling dispersion of the estimates. It is possible to show that $G'DG$ converges in probability to the Hessian matrix of the distance function, so its use gives a Newton-like method. We summarize our conclusions about the minimum distance estimator in a

Theorem on Computation of the Minimum Distance Estimator and Its Variance-Covariance Matrix

Consider the following iteration:

$$\beta^{(i+1)} = \beta^{(i)} - \lambda^{(i)}[G'^{(i)}DG^{(i)}]^{-1}g^{(i)} \tag{4.5}$$

where $\lambda^{(i)}$ is computed by the method of Section 2 and G. D, and g are as defined above. Then

(i) the method converges to a stationary point of the distance function as $i \to \infty$,

(ii) the method is close to Newton's method in that $G'DG$ converges to the Hessian matrix of the distance function as $T \to \infty$, and

(iii) $[G'DG]^{-1}$ is a consistent estimate of the variance-covariance matrix of the estimated parameters.

Practical application of the minimum distance estimator for models of any size requires careful organization of the computations. It is desirable to avoid recomputing the distance function after the calculations begin, but the matrix D as defined in equation (4.3) is much too large to store in memory. Our approach is based on two preliminary transformations. We premultiply the instruments by the matrix square root of $S^{-1} \otimes \mathrm{I}$ and postmultiply by the matrix square root of $H'(S^{-1} \otimes I)H$ (the second transformation has the effect of orthonormalizing the instruments). Then at each iteration we perform the first of these transformations on the derivatives of the model and on the residuals. This process is described more precisely in the

Minimum Distance Algorithm

1. Calculate a consistent estimate of Σ, S. Calculate the square root or Choleski factorization of S^{-1}, W:

$$S^{-1} = WW'. \tag{4.6}$$

2. Form $H'(S^{-1} \otimes I)H$ and calculate the Choleski factorization: $(H'(S^{-1} \otimes I)H)^{-1} = VV'$. Form transformed instruments,

$$\tilde{H} = (W' \otimes I)HV. \tag{4.7}$$

3. At each iteration, form the matrix G of values of the derivative of f.
4. Form the transformed derivatives and residuals as

$$\tilde{G} = (W' \otimes I)G \tag{4.8}$$

$$\tilde{f} = (W' \otimes I)f. \tag{4.9}$$

5. Calculate the direction,

$$d = (\tilde{G}'\tilde{H}\tilde{H}'\tilde{G})^{-1}\tilde{G}'\tilde{H}\tilde{H}'\tilde{f}. \tag{4.10}$$

6. Check for convergence.
7. Search for λ, update β, and return to Step 3.

The reader should have no trouble verifying that the expression for d in terms of transformed G, H, and f is the same as that specified in the theorem in equation (4.5).

Inference for minimum distance estimators is based on the asymptotic normality of the estimates. We consider the following rather general class of nonlinear null hypotheses:

$$\beta^{(2)} = \Phi(\beta^{(1)}). \tag{4.11}$$

Here $\beta^{(2)}$ is a vector of length n and $\beta^{(1)}$ is a vector of the remaining $K - n$ parameters. We assume that Φ is an analytic function; often it will be linear or even constant. The statistic in the sample corresponding to equation (4.11) is

$$z = b^{(2)} - \Phi(b^{(1)}) \tag{4.12}$$

which will be close to a zero vector if the null hypothesis is true. The statistic is asymptotically normal (Malinvaud, 1970, Chapter 9), with variance-covariance matrix

$$V(z) = V^{(2,2)} - V^{(2,1)}\left(\frac{\partial\Phi}{\partial b^{(1)}}\right)' - \left(\frac{\partial\Phi}{\partial b^{(1)}}\right)V^{(1,2)} + \left(\frac{\partial\Phi}{\partial b^{(1)}}\right)V^{(1,1)}\left(\frac{\partial\Phi}{\partial b^{(1)}}\right)' \tag{4.13}$$

where $V^{(i,j)}$ are the blocks of the asymptotic variance-covariance matrix of b. Then inference is based on the quadratic form,

$$F = z'[V(z)]^{-1}z \tag{4.14}$$

which is distributed asymptotically as $\chi^2(n)$ under the null hypothesis.

Computation of the test statistic appears a formidable task, but in fact a method exists for computing it as simply as the likelihood ratio statistic for maximum likelihood. F is equal to twice the increase in the distance function when the null hypothesis is imposed as a constraint on the parameters in the way described in the algorithm below. This method is familiar to econometricians in the linear regression model, especially in the form of the "Chow test", but its applicability to simultaneous estimation apparently has not been noted previously.

Algorithm for Computing the χ^2 Test Statistic from the Minimum Distance Estimator

1. Estimate the parameters of the unconstrained model corresponding to the maintained hypothesis by the minimum distance algorithm. Let Δ be the value of the distance function at the minimum.
2. Substitute the constraint $b(^2) = \Phi(b^{(1)})$ into the model.
3. Starting from the value of $h^{(1)}$ from Step 1 and, using the same variance-covariance matrix S as in Step 1, take one additional iteration. Set $\lambda = 1$.
4. Let Δ^* be the value of the linearized distance function:

   $$\Delta^* = \tfrac{1}{2}\hat{f}'D\hat{f} \tag{4.15}$$

   where $\hat{f}$ are the residuals around the linearized model:

   $$\hat{f} = f(y, b) + \left(\frac{\partial f}{\partial b^{(1)}} + \frac{\partial f}{\partial b^{(2)}}\frac{\partial \Phi}{\partial b^{(1)}}\right)(b^{*(1)} - b^{(1)}) \tag{4.16}$$

   and b and b^* are the estimates from Steps 1 and 3, respectively.
5. Calculate F as $2(\Delta^* - \Delta)$.

Inference by this method requires no additional computations beyond those of estimation except for the calculation of the linearized distance.

It is difficult to compare the power of this test relative to the corresponding likelihood ratio test. Since the minimum distance estimator is not generally efficient, the test based on it is probably usually less powerful than the likelihood ratio test. However, the minimum distance estimates are consistent, so the χ^2 test is consistent as well—the probability of rejecting null hypothesis approaches one as the sample becomes large.

5. Application to Nonlinear Multivariate Regression

Multivariate regression is an important special case of the general structural model. In the case of regression, the derivatives of the model with respect to the parameters do not depend on the endogenous variables, y. This has two implications for our methods. First, and most important, the Jacobian determinant J_t in the likelihood function of equation (3.1) equals unity and the troublesome term $\Sigma \log |\det J_t|$ disappears from the equation. The gradient of the log-likelihood function is just

$$\frac{\partial L}{\partial \beta} = -q = -\frac{1}{2}\frac{\partial}{\partial \beta}\log\det \sum_t F_t'F_t. \tag{5.1}$$

It is not hard to show that

$$q = \left(\frac{\partial f}{\partial \beta}\right)'(\Sigma^{-1} \otimes I)f. \tag{5.2}$$

The true (not sample) variance-covariance matrix of the gradient is then

$$(5.3) \qquad R = E(qq') = E\left[\left(\frac{\partial f}{\partial \beta}\right)'(\Sigma^{-1} \otimes I)ff'(\Sigma^{-1} \otimes I)\frac{\partial f}{\partial \beta}\right].$$

The second implication of the nonstochastic nature of $\partial f/\partial \beta$ is that we can pass the expectation operator through to the middle of this expression:

$$(5.4) \qquad \begin{aligned} R &= \left(\frac{\partial f}{\partial \beta}\right)'(\Sigma^{-1} \otimes I)E(ff')(\Sigma^{-1} \otimes I)\frac{\partial f}{\partial \beta} \\ &= \left(\frac{\partial f}{\partial \beta}\right)'(\Sigma^{-1} \otimes I)\frac{\partial f}{\partial \beta}. \end{aligned}$$

Since this can be computed exactly, while the alternative R_T of equation (3.7) is only an estimate, it appears a better choice of the Q matrix in a gradient method and a better estimate of the variance-covariance matrix of the maximum likelihood estimator. This choice of Q is well known in univariate nonlinear regression as the Gauss-Newton method. For multivariate regression, the theorem on maximum likelihood estimators and the maximum likelihood algorithm in Section 3 continue to apply if R is substituted for R_T.

When the minimum distance estimator is applied in the case of multivariate regression, the matrix of instrumental variables, H, is superfluous and the distance matrix should be taken as

$$(5.5) \qquad D = S^{-1} \otimes I.$$

Malinvaud (1970, Chapter 9) has studied the minimum distance estimator in considerable detail. He has shown that for an arbitrary positive definite S the estimator is consistent and asymptotically normal, and further, that if S is any consistent estimator of Σ, the minimum distance estimator is asymptotically efficient.

With the redefinition of D given above, the theorem on the computation of minimum distance estimators and the minimum distance algorithm of Section 4 apply without change.

Although the maximum likelihood and minimum distance approaches yield asymptotically equivalent estimators, they are not generally numerically identical in finite samples. Maximum likelihood updates the estimate of Σ at each iteration, while minimum distance holds S constant. At the conclusion of maximum likelihood, Σ is exactly the sample variance-covariance matrix of the residuals, but minimum distance lacks this consistency between S and the residuals. If the minimum distance algorithm is modified to update S at each iteration, it becomes precisely the same as the maximum likelihood algorithm.

The one-step efficient method, using R from equation (5.4), proceeds as before. An initial consistent estimate can be obtained by applying univariate regression separately to each equation, or by minimum distance.[5] Then a single iteration

[5] If there are no parameter constraints across equations, minimum distance with $S = I$ is exactly the same as univariate regression applied separately.

with $\lambda = 1$ provides estimates that are asymptotically equivalent to full maximum likelihood.

Finally, inference in multivariate regression follows the rules set out at the ends of Sections 3 and 4. For maximum likelihood, the likelihood ratio is

$$-2(\log L - \log L^*) = \det \frac{1}{T} U^{*\prime} U^* - \det \frac{1}{T} U' U. \tag{5.6}$$

For minimum distance, the difference between the linearized constrained distance and the unconstrained distance is again χ^2 with n degrees of freedom under the null hypothesis. Asymptotically the two methods of inference are equivalent, but will differ in finite samples because S will not be the sample variance-covariance matrix of the residuals in the minimum distance case.

References

[1] Amemiya, T., "The Nonlinear Two-stage Least-squares Estimator," *Journal of Econometrics*, July 1974, pp. 105–110.

[2] Chow, G., "On the Computation of Full-Information Maximum Likelihood Estimates for Nonlinear Equation Systems," *Review of Economics and Statistics*, February 1973, pp. 104–109.

[3] Daniel, J. W., "Convergent Step-Sizes for Gradient-Like Feasible Direction Algorithms for Constrained Optimization," in J. B. Rosen, O. L. Mangasarian, and K. Ritter (eds.), *Nonlinear Programming*, Academic Press, New York, 1970, pp. 245–274.

[4] Eisenpress, H. and J. Greenstadt, "The Estimation of Nonlinear Econometric Systems," *Econometrica*, October 1966, pp. 851–861.

[5] Fisher, F., *The Identification Problem in Econometrics*, McGraw-Hill, New York, 1966.

[6] Goldstein, A., *Constructive Real Analysis*, Harper & Row, New York, 1967.

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

[8] Hausman, J. A., "Full Information Instrumental Variable Estimation of Simultaneous Equation Models," this issue, 1974.

[9] Jacoby, S. L. S., J. S. Kowalik, and J. T. Pizzo, *Iterative Methods for Nonlinear Optimization Problems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1972.

[10] Kendall, M. G. and A. Stuart, *Advanced Theory of Statistics*, Griffin, London, 1967.

[11] Malinvaud, E., *Statistical Methods of Econometrics*, second ed., North-Holland, Amsterdam, 1970.

[12] Powell, M. J. D., "Recent Advances in Unconstrained Optimization," *Mathematical Programming*, Vol. 1, pp. 26–57, October 1971.

[13] Rao, C. R., *Linear Statistical Inference and its Applications*, Wiley, New York, 1965.

[14] Rothenberg, T., and C. Leenders, "Efficient Estimation of Simultaneous Equation Systems," *Econometrica*, January 1964, pp. 57–76.

[15] Zacks, S., *The Theory of Statistical Inference*, Wiley, New York, 1971.