Econometrica, Vol. 50, No. 4 (July, 1982)

LARGE SAMPLE PROPERTIES OF GENERALIZED METHOD OF MOMENTS ESTIMATORS[1]

BY LARS PETER HANSEN

This paper studies estimators that make sample analogues of population orthogonality conditions close to zero. Strong consistency and asymptotic normality of such estimators is established under the assumption that the observable variables are stationary and ergodic. Since many linear and nonlinear econometric estimators reside within the class of estimators studied in this paper, a convenient summary of the large sample properties of these estimators, including some whose large sample properties have not heretofore been discussed, is provided.

1. INTRODUCTION

IN THIS PAPER we study the large sample properties of a class of generalized method of moments (GMM) estimators which subsumes many standard econometric estimators. To motivate this class, consider an econometric model whose parameter vector we wish to estimate. The model implies a family of orthogonality conditions that embed any economic theoretical restrictions that we wish to impose or test. For example, assumptions that certain equations define projections or that particular variables are predetermined give rise to orthogonality conditions in which expected cross products of unobservable disturbances and functions of observable variables are equated to zero. Heuristically, identification requires at least as many orthogonality conditions as there are coordinates in the parameter vector to be estimated. The unobservable disturbances in the orthogonality conditions can be replaced by an equivalent expression involving the true parameter vector and the observed variables. Using the method of moments, sample estimates of the expected cross products can be computed for any element in an admissible parameter space. A GMM estimator of the true parameter vector is obtained by finding the element of the parameter space that sets linear combinations of the sample cross products as close to zero as possible.

In studying strong consistency of GMM estimators, we show how to construct a class of criterion functions with minimizers that converge almost surely to the true parameter vector. The resulting estimators have the interpretation of making the sample versions of the population orthogonality conditions as close as possible to zero according to some metric or measure of distance. We use the metric to index the alternative estimators. This class of estimators includes the nonlinear instrumental variables estimators considered by, among others, Amemiya [**1**, **2**], Jorgenson and Laffont [**24**], and Gallant [**11**].[2] There the

[1]The author acknowledges helpful comments by Robert Avery, Robert Hodrick, V. Joseph Hotz, Dan Peled, Thomas Sargent, Katherine Schipper, Kenneth Singleton, Kenneth Wallis, Halbert White, and an anonymous referee. Special thanks are given to Christopher Sims who played a prominent role in the formulation of this paper.

[2]We include versions of two- and three-stage least squares under the heading of instrumental variables procedures.

population orthogonality conditions equate expected cross products of instruments and serially independent disturbances to zero. In our treatment we work directly with expressions for the population orthogonality conditions and implicitly permit the disturbance terms used in construction of the orthogonality conditions to be both serially correlated and conditionally heteroskedastic.[3] We allow ourselves flexibility in choosing the distance measure because it permits choosing measures that are computationally convenient and because the choice of distance measure influences the asymptotic distribution of the resulting estimator.

In studying asymptotic normality, we view estimation in a different but closely related fashion. We follow Sargan [**29**, **30**] and consider estimators that have the interpretation of setting linear combinations of the sample orthogonality conditions to zero, at least asymptotically, where the number of linear combinations that are set to zero is equal to the number of coordinates in the parameter vector to be estimated. We index alternative estimators by an associated weighting matrix that selects the particular linear combinations of orthogonality conditions that are used in estimation. Since alternative weighting matrices give rise to estimators with alternative asymptotic covariance matrices, we describe how to obtain an asymptotically optimal weighting matrix. The estimators considered in our treatment of consistency are shown to reside in the class of estimators considered in our treatment of asymptotic normality by examining the first-order conditions of minimization problems used to construct the class of consistent estimators. It turns out, however, that our discussion of asymptotic normality is sufficiently general to include other consistent estimators that are obtained from minimizing or maximizing other criterion functions which have first-order conditions that satisfy the specification of our generic GMM estimator, e.g., least squares or quasi-maximum likelihood estimators. Again our discussion of large sample properties permits the disturbances implicitly used in the orthogonality conditions to be both conditionally heteroskedastic and serially correlated.[4]

There are a variety of applications in which it is important to possess an asymptotic theory which accommodates these features. In testing market efficiency and the rationality of observed forecasts using least squares procedures, one oftentimes encounters situations in which the implied forecast interval

[3] Sargan [**30**] treats the case in which disturbances can follow a low-order autoregression and can be filtered to remove serial correlation prior to the construction of the orthogonality conditions. White [**34**] discusses linear instrumental variables estimation in which observation vectors are independent but not necessarily identically distributed. White allows heteroskedasticity to exist both conditionally and unconditionally, but places restrictions on higher moments of observable and unobservable variables that are not needed in this paper. Here we think of heteroskedasticity emerging because of some implicit conditioning, do not impose independence, but maintain a stationarity assumption.

[4] Engle [**9**] allows for conditional heteroskedasticity in regression models with serially uncorrelated disturbances. He proposes a maximum likelihood procedure for estimating such models when the form of the heteroskedasticity is specified *a priori*. White [**32**, **33**, **34**] has studied the asymptotic distribution of a variety of estimators for cross-sectional models which allow for both conditional and unconditional forms of heteroskedasticity. See Footnote 3.

exceeds the sampling interval giving rise to a serially correlated forecast error [**4**, **14**, **17**]. Least squares procedures can be used since the hypothetical forecast error should be orthogonal to the observed forecast and to any other variables in the information set of economic agents when the forecast is made. On the other hand, generalized least squares procedures can result in inconsistent parameter estimators (see Sims [**31**] and Hansen and Hodrick [**17**]). Brown and Maital [**4**], Hansen and Hodrick [**17**], and Hakkio [**14**] rely on the asymptotic distribution theory in this paper to carry out least squares estimation and inference for such models.

Hansen and Sargent [**18**, **19**] have considered linear rational expectations models in which economic agents are assumed to forecast infinite geometrically-declining sums of forcing variables and the econometrician employs only a subset of the variables in the information set of economic agents. The disturbance terms in these models are serially correlated but orthogonal to current and past values of a subset of variables which are not strictly exogenous. Hansen and Sargent [**18**, **19**] discuss how to apply the techniques developed in this paper to those rational expectations models. McCallum [**28**] has shown how other types of linear rational expectations models with disturbance terms that have low-order autoregressive representations lead to equations that can be estimated consistently using standard instrumental variables procedures. He notes, however, that the associated asymptotic distribution of the estimations has to be modified in the manner suggested in this paper to allow the disturbances to be serially correlated. In considering models like those studied by McCallum [**28**], Cumby, Huizinga, and Obstfeld [**5**] propose a two-step, two-stage least squares estimator that resides within the class of estimators examined in this paper.[5]

Hansen and Singleton [**20**] have studied how to test restrictions and estimate parameters in a class of nonlinear rational expectations models. They construct generalized instrumental variables estimators from nonlinear stochastic Euler equations and note that the implied disturbance terms in these models are conditionally heteroskedastic and in many circumstances serially correlated. Their estimators are special cases of the generic GMM estimator of this paper. Finally, Avery, Hansen, and Hotz [**3**] describe how to use methods in this paper to obtain computationally convenient procedures for estimating multiperiod probit models. The vector disturbance term implicit in their orthogonality conditions also is conditionally heteroskedastic.

In the examples described above, application of the techniques in this paper will not result in asymptotically efficient estimators. However, in these and other examples, a researcher may be willing to sacrifice asymptotic efficiency in exchange for not having to specify completely the nature of the serial correlation and/or heteroskedasticity or in exchange for computationally simpler estimation strategies. As noted above, we do provide a more limited optimality discussion

[5]Cumby, Huizinga, and Obstfeld [**5**] proposed their estimator independently of this paper. However, their discussion of its asymptotic distribution exploited results in a precursor to this paper written by the author.

that is patterned after an approach taken by Sargan [**29**, **30**] and can be easily exploited in practice.

The organization of the paper is as follows. The second section provides some consistency results for the GMM estimator under various assumptions about the form of the econometric model. The third section discusses the asymptotic distribution of the GMM estimator and considers the construction of an asymptotically optimal estimator among the class of estimators that exploit the same orthogonality conditions. The fourth section examines procedures for testing overidentifying restrictions using GMM estimation. Finally, the fifth section contains some concluding remarks.

2. CONSISTENCY OF THE GMM ESTIMATOR

In this section we specify our first form of the GMM estimator and provide some sufficient conditions that insure its almost sure convergence to the parameter vector that is being estimated. Let Ω denote the set of sample points in the underlying probability space used in our estimation problem, and let E denote the associated expectations operator. We will be working with a p component stochastic process $\{x_n : n \geq 1\}$ defined on this probability space. A finite segment of one realization of this process, i.e., $\{x_n(\omega_0) : 1 \leq n \leq N\}$ for sample size N and for some $\omega_0 \in \Omega$, can be thought of as the observable data series that the econometrician employs.

ASSUMPTION 2.1: $\{x_n : 1 \leq n\}$ is stationary and ergodic.

We introduce a parameter space S that is a subset of R^q (or its compactification) and let β_0 be the element of S that we wish to estimate.

ASSUMPTION 2.2: (S, σ) is a separable metric space.

One possibility is to use the standard absolute value norm on R^q to define σ. It is well known that since S is a subset of R^q the resulting metric space is separable. We do not restrict ourselves to this metric in order to allow for S to be a subset of a compactification of R^q.

We consider a function $f: R^p \times S \to R^r$ where R is the real line and r is greater than or equal to q.

ASSUMPTION 2.3: $f(\cdot, \beta)$ is Borel measurable for each β in S and $f(x, \cdot)$ is continuous on S for each x in R^p.

The function f provides an expression for the r orthogonality conditions that emerge from the econometric model in the sense indicated by Assumption 2.4.

ASSUMPTION 2.4: $Ef(x_1, \beta)$ exists and is finite for all $\beta \in S$ and $Ef(x_1, \beta_0) = 0$.

A common way to obtain orthogonality conditions is to exploit the assumption that disturbances in an econometric model are orthogonal to functions of a set of variables that the econometrician observes. For example, suppose that the econometric model is given by

$$\begin{aligned} u_n &= F(x_n, \beta_0), \\ z_n &= G(x_n, \beta_0), \end{aligned} \tag{1}$$

where

$$E[u_n \otimes z_n] = 0. \tag{2}$$

The vector functions F and G are specified *a priori*, u_n is an unobservable vector of disturbance terms, z_n is a vector of instrumental variables, and "$\otimes$" denotes the Kronecker product. The dependence of G on its second argument is oftentimes trivial. When (2) is satisfied, we can let the function f be given by

$$f(x_n, \beta_0) = F(x_n, \beta_0) \otimes G(x_n, \beta_0), \tag{3}$$

and it follows that

$$E[f(x_n, \beta_0)] = 0.$$

We proceed to describe how to use orthogonality conditions to construct an estimator of the unknown parameter vector β_0.

For our discussion of consistency, we introduce a sequence of random weighting matrices $\{a_N : N \geq 1\}$ that are dimensioned s by r where $q \leq s \leq r$. The matrices are random in order to allow for their possible dependence on sample information.

ASSUMPTION 2.5: The sequence of random matrices $\{a_N : N \geq 1\}$ converges almost surely to a constant matrix a_0.[6]

These weighting matrices are used in conjunction with a method of moments estimator of $E[f(x_n, \beta)]$ to obtain a sample objective function whose minimizer is our estimator of β_0. Let

$$f_n(\omega, \beta) = f[x_n(\omega), \beta],$$

$$g_N(\omega, \beta) = \frac{1}{N} \sum_{n=1}^{N} f_n(\omega, \beta),$$

$$h_N(\omega, \beta) = a_N(\omega)\, g_N(\omega, \beta),$$

$$B_N(\omega) = \left\{ \beta \in S : |h_N(\omega, \beta)|^2 = \inf_{\beta \in S} |h_N(\omega, \beta)|^2 \right\}.$$

[6]This matrix convergence is defined as element by element convergence using the absolute value norm on R.

The random function $g_N(\beta)$ is just the method of moments estimator of $E[f(x_n, \beta)]$, $|h_N|^2$ is the sample criterion function to be used in estimation, and B_N is the (random) set of elements in the parameter space S that minimize $|h_N|^2$. The weighting matrices $\{a_N : N \geq 1\}$ can be thought of as defining the metric by which the sample orthogonality conditions $g_N(b_N)$ are made as close as possible to zero.

To estimate β_0 we choose an element out of B_N. More precisely, we employ the following definition.

DEFINITION 2.1: The GMM estimator $\{b_N : N \geq 1\}$ is a sequence of random vectors such that $b_N(\omega) \in B_N(\omega)$ for $N \geq N^*(\omega)$ where $N^*(\omega)$ is less than infinity for almost all ω in Ω.[7]

The nonlinear instrumental variables estimators discussed by Amemiya [**1**], Jorgenson and Laffont [**24**], and Gallant [**11**] are defined in this manner for appropriate choices of a_N. Their instrumental variables estimators assume that the function f satisfies (1)–(3) and in addition that the disturbances are serially independent. They use consistent estimators of $E[u_n u_n']$ and $E[z_n z_n']$ to construct an estimator of a_0 where

$$a_0' a_0 = \{E[u_n u_n'] \otimes E[z_n z_n']\}^{-1}.$$[8]

In preparation for our first consistency theory, we introduce the notation

$$h_0(\beta) = a_0 E[f_1(\omega, \beta)],$$

$$\epsilon_1^k(\omega, \beta, \delta) = \sup\{|f_1(\omega, \beta) - f_1(\omega, \alpha)|^k : \alpha \in S, \sigma(\beta, \alpha) < \delta\}.$$

The following definition is needed for our consistency results.

DEFINITION 2.2: The random function f_1 is kth moment continuous at β if $\lim_{\delta \downarrow 0} E[\epsilon_1^k(\omega, \beta, \delta)] = 0$.[9]

Since $\{x_n : n \geq 1\}$ is stationary, it follows that if f_1 is kth moment continuous, then f_n is kth moment continuous for all n. Notice that kth moment continuity is joint property of the function f and the stochastic process $\{x_n : n \geq 1\}$. An

[7] In this definition we have imposed the requirement that the sequence of functions $\{b_N : N \geq 1\}$ be measurable. Alternatively, we could follow a suggestion of Huber [**23**] and not necessarily require that the functions be measurable and establish almost sure convergence in terms of outer probability.

[8] Amemiya [**1**], Jorgenson and Laffont [**24**], and Gallant [**11**] do not require that the instrumental variables be stationary and ergodic but instead require that the appropriate moment matrices converge. Stationarity and ergodicity coupled with finite expectations are sufficient conditions for these moment matrices to converge almost surely. Amemiya [**2**] adopts a more general representation of the orthogonality conditions than (3) to allow different disturbances to be paired with different sets of instruments.

[9] The function $\epsilon_1^k(\cdot, \beta, \delta)$ is Borel measurable under Assumptions 2.2 and 2.3. In the case in which $k = 1$, DeGroot [**6**] refers to first moment continuity as supercontinuity.

alternative characterization of kth moment continuity is established in Lemma 2.1.

LEMMA 2.1: *Under Assumption* 2.3, *if there exists a* $\delta > 0$ *such that* $E[\epsilon_1^k(\omega, \beta, \delta)] < +\infty$, *then* f_1 *is* kth *moment continuous at* β.

Using this lemma, it is apparent that kth moment continuity is implied if the random function f_1 is dominated locally by a random variable with a finite kth moment. DeGroot [**6**, p. 206] proved Lemma 2.1 for k and q equal to one, and the extension to larger specifications of k and q is immediate.

One other lemma is of use in verifying first moment continuity in the case where the function f satisfies relation (3).

LEMMA 2.2: *Suppose* (i) F_1 *and* G_1 *are second moment continuous at* β; (ii) $F_1(\cdot, \beta)$ *and* $G_1(\cdot, \beta)$ *have finite second moments. Then* $f_1 = F_1 \otimes G_1$ *is first moment continuous at* β.[10]

Lemma 2.2 may be useful in establishing that f_1 is first moment continuous at β when the orthogonality conditions are of the form (3).

We now consider our first consistency theorem for the GMM estimator.

THEOREM 2.1: *Suppose Assumptions* 2.1–2.5 *are satisfied. If* (i) f_1 *is first moment continuous for all* $\beta \in S$; (ii) S *is compact*; (iii) $h_0(\beta)$ *has a unique zero at* β_0; *then a* GMM *estimator* $\{b_N : N \geq 1\}$ *exists and converges almost surely to* β_0.

Condition (iii) of this theorem is the parameter identification requirement that the population orthogonality conditions used in estimation be satisfied only at the true parameter vector. When a_0 is an r by r nonsingular matrix, h_0 will have a unique zero at β_0 if, and only if, $Ef(x_1, \cdot)$ has a unique zero at β_0. When a_0 has fewer rows than columns ($s < r$), condition (iii) imposes the more stringent requirement that s linear combinations of the population orthogonality conditions are satisfied only at the true parameter vector. For this reason, it may be judicious to choose a_0 to be an r by r nonsingular matrix.[11]

The compactness condition (ii) of Theorem 2.1 can be weakened if a special structure is imposed on the function f. Consider the following assumption.

ASSUMPTION 2.6: $f(x_1, \beta) = c_0(x_1) + c_1(x_1)\lambda(\beta)$ where $c_0(x_1)$ is an r dimensional column vector, $c_1(x_1)$ is an r by m matrix, and $\lambda(\beta)$ is an m dimensional vector.

[10] At the recommendation of the editor, detailed versions of proofs in this section are not included in the paper but are available from the author on request.

[11] Sargan [**29**] and Amemiya [**2**] note that from the standpoint of obtaining desirable small sample properties, one should try to conserve on the number of orthogonal conditions used in the estimation.

This assumption accommodates models that are linear in the variables but not necessarily in the parameters. Our next theorem establishes consistency for models with orthogonality conditions that satisfy Assumption 2.6.

THEOREM 2.2: *Suppose Assumptions* 2.1–2.6 *are satisfied. If* (i) (S,σ) *is locally compact*; (ii) λ *is continuous on* S, *and for any* $\phi > 0$ *there exists a* $\rho > 0$ *such that* $|\lambda(\alpha) - \lambda(\beta)| < \rho$, $\sigma(\alpha, \beta) < \phi$; (iii) *for any* $\rho > 0$,

$$\inf\left\{ \frac{1}{1+|\lambda(\beta)|} |h(\beta)| : \beta \in S, |\lambda(\beta) - \lambda(\beta_0)| > \rho \right\} > 0;$$

then the GMM *estimator* $\{b_N : N \geq 1\}$ *exists and converges almost surely to* β_0.

In examining Theorem 2.2, let us first consider the case in which $\lambda(\beta) = \beta$. Condition (i) is easily verified for $(S,\sigma) = (R^q, |\ |)$. The function h_0 is given by

$$h_0(\beta) = a_0 E[c_0(x_1)] + a_0 E[c_1(x_1)]\beta.$$

Suppose that a_0 and $Ec_1(x_1)$ are both of full rank. Furthermore, we assume that β_0 is a zero of h_0. This is sufficient to imply that for any $\rho > 0$

$$\inf\left\{ \frac{1}{1+|\beta|} |h_0(\beta)| : \beta \in R^q, |\beta - \beta_0| > \rho \right\} > 0,$$

and consequently condition (iii) is met. Thus, for models that are linear in parameters, Theorem 2.2 requires no additional assumptions on the parameter space in order to achieve consistency. Of course, this result could be demonstrated easily by explicitly solving for the estimator. Nonetheless, a consistency result for linear models in which the underlying stochastic process is stationary and ergodic is embedded in Theorem 2.2.

The more interesting aspect of Theorem 2.2 is that it provides a consistency result for models that are nonlinear in the parameters and does not explicitly employ a compactness requirement. For this reason, we will examine conditions (ii) and (iii) in more depth. Condition (ii) requires that the mapping defined by the inverse of λ be continuous at the true parameter vector. In interpreting assumption (iii), it is fruitful to view R^m as the unrestricted parameter space. The function λ is used to indicate the elements of that space which satisfy the restrictions generated by the model. In particular, we let

$$P = \{\theta \in R^m : \lambda(\beta) = 0 \text{ for some } \beta \in S\},$$

i.e., P is the image of λ over the set S, and the set S indexes elements of P that satisfy the restrictions. We define another set Q as

$$Q = \{\theta \in R^m : a_0 E c_0(x_1) + a_0 E c_1(x_1)\theta = 0\}.$$

The hyperplane Q consists of all the elements of R^m that satisfy the population orthogonality conditions used in estimation. From conditions (ii) and (iii), we are guaranteed that $Q \cap P = \{\theta_0\}$ where $\theta_0 = \lambda(\beta_0)$.

We now endeavor to construct sufficient conditions for condition (iii) to hold.

We define a function ξ by

$$\xi(\rho) = \inf\{|\theta - \pi| : \pi \in Q, \theta \in P, |\pi - \theta_0| \geq \rho, |\theta - \theta_0| \geq \rho\}.$$

The following lemma supplies some sufficient conditions for (iii) of Theorem 2.2 to hold.

LEMMA 2.3: *Suppose Assumptions* 2.4 *and* 2.5 *are satisfied. If* (i) *for any* $\rho > 0$, $\xi(\rho) > 0$; (ii) $\lim_{\rho \to \infty} \inf \xi(\rho)/\rho > 0$; *then condition* (*iii*) *of Theorem* 2.2 *is satisfied.*

Condition (i) of Lemma 2.3 says that it is not possible for elements in P to get arbitrarily close to elements in Q outside of the neighborhood of θ_0. Condition (ii) of Lemma 2.3 says that the distance between P and Q outside a neighborhood of radius ρ of θ_0 eventually grows at least proportionately with ρ.[12] A condition like (ii) is needed because although the set P is specified *a priori*, the set Q is not known and a_0, $E[c_0(x_1)]$, and $E[c_1(x_1)]$ have to be estimated. Using estimators of these matrices, define the random set

$$Q_N = \left\{ \theta \in R^m : a_N \frac{1}{N} \sum_{n=1}^{N} c_0(x_n) + a_N \frac{1}{N} \sum_{n=1}^{N} c_1(x_n)\theta = 0 \right\}.$$

Even very small errors in estimating a_0, $E[c_0(x_1)]$, and $E[c_1(x_1)]$ get magnified in terms of the distance between Q_N and $\theta \in Q$ as θ becomes large in absolute value. To insure that the GMM estimator is consistent, we have to rule out the possibility that $Q_N \cap P$ contains an element far away from θ_0 for sufficiently large sample size N.

If S is compact and $P \cap Q = \{\theta_0\}$, then assumptions (i) and (ii) of Lemma 2.3 are trivially met. However, Lemma 2.3 and Theorem 2.2 can be applied in situations in which S is not compact. In fact, an important special case occurs when $Q = \{\theta_0\}$. This means that the unrestricted parameter vector θ_0 is uniquely determined by the population orthogonality conditions used in estimation. When $Q = \{\theta_0\}$ assumptions (i) and (ii) of Lemma 2.3 are easily verified.[13]

The consistency Theorems 2.1 and 2.2 illustrate the potential tradeoff between assumptions on the function f and assumptions on the parameter space S in order to obtain strong consistency of the GMM estimator. Theorem 2.1 most closely resembles other consistency theorems in the literature for nonlinear instrumental variables, where the parameter space is assumed to be compact [**1**, **24**]. In contrast to those theorems, Theorem 2.1 does not assume that disturbances are serially independent. Theorem 2.2 relaxes the compactness assumption at the cost of being more restrictive about the specification of f.

[12] A requirement equivalent to condition (ii) of Lemma 2.3 can be formulated in terms of asymptotic cones. If we let $As(P)$ and $As(Q)$ denote the asymptotic cones of P and Q, respectively, then condition (ii) is equivalent to requiring that $As(P) \cap As(Q) = \{0\}$.

[13] In the example considered in Hansen and Sargent [**18**, pp. 33–36], $Q \neq \{\theta_0\}$. Malinvaud [**27**, p. 350] has proved a theorem for minimum distance estimators similar to Theorem 2.2 in cases in which $Q = \{\theta_0\}$. Huber [**23**] has a general treatment of consistency in cases in which the observation vector is independent and identically distributed.

Before concluding this discussion on consistency, one additional theorem is considered. Suppose elements in R^q are partitioned into two subvectors, i.e., $\beta' = (\beta_1', \beta_2')$. Furthermore, suppose the metric σ is

$$\sigma(\beta, \gamma) = \max\{\sigma_1(\beta_1, \gamma_1), \sigma_2(\beta_2, \gamma_2)\}$$

where σ_1 is a metric defined on

$$S_1 = \{\beta_1 : (\beta_1', \beta_2')' \in S \text{ for some } \beta_2\}$$

and σ_2 is a metric defined on

$$S_2 = \{\beta_2 : (\beta_1', \beta_2') \in S \text{ for some } \beta_1\}.$$

In some circumstances it may be computationally convenient to construct a strongly consistent estimator $\{b_{1,N} : N \geq 1\}$ of $\beta_{1,0}$ by using a subset of the orthogonality conditions provided by the model. In particular, Theorems 3.1 or 3.2 could be used to establish this consistency. After obtaining this estimator of $\beta_{1,0}$, we can construct an estimator of $\beta_{2,0}$ by minimizing

$$|h_N[\omega, b_{1,N}(\omega), \beta_2]|^2$$

with respect to β_2 such that $(\beta_1', \beta_2')' \in S$, where $\beta_1 = b_{1,N}(\omega)$. Theorem 2.3 establishes the consistency of this recursive estimator.

THEOREM 2.3: *Suppose Assumptions* 2.1–2.5 *are satisfied. If* (i) *the conditions of Theorem* 3.1 *are satisfied*; (ii) $\{b_{1,N} : N \geq 1\}$ *converges almost surely to* $\beta_{1,0}$; (iii) *for any sequence* $\{\gamma_j : j \geq 1\}$ *in* S *such that* $\{\gamma_{1,j} : j \geq 1\}$ *converges to* $\beta_{1,0}$, *there exists a sequence* $\{\beta_{2,j} : j \geq 1\}$ *such that* $\{(\gamma_{1,j}, \beta_{2,j}) : j \geq 1\}$ *is a sequence in* S *that converges to* β_0; *then a* GMM *estimator* $\{b_N : N \geq 1\}$ *exists and converges almost surely to* β_0.[14]

A couple of comments are in order about Theorem 2.3. First, condition (iii) imposes an extra requirement on the parameter space S. If σ_1 and σ_2 are defined by the absolute value norm, then a sufficient condition for (iii) to hold is that S be convex. However, condition (iii) can be satisfied in the absence of convexity. Second, some of the coordinate functions of $h_N[\omega, b_{1,N}(\omega), \cdot]$ may not actually depend on β_2. If this is the case, computation of the criterion function for the second step of this recursive procedure can be simplifed by ignoring these coordinate functions.

3. THE ASYMPTOTIC DISTRIBUTION OF THE GMM ESTIMATOR

In this section we establish the asymptotic normality of a generic GMM estimator. Our discussion adopts a different but closely related formulation of

[14] A version of Theorem 2.3 also can be established using the assumptions of Theorem 2.2 and conditions (ii) and (iii) of Theorem 2.3.

GMM estimation to that in Section 2. The first-order conditions of the minimization problem used to define a GMM estimator in Section 2 have the interpretation of setting q linear combinations of the r sample orthogonality conditions to zero where q is the dimensionality of the parameter space. It turns out that estimators obtained by minimizing or maximizing other criterion functions, e.g., quasi-maximum likelihood or least squares estimators, oftentimes can be interpreted in the same manner by examining the corresponding first-order conditions.[15] Our approach in this section is to adopt a generic form of the first-order conditions and to assume that consistency has already been established. For estimators not included in the discussion of Section 2, consistency might be established by appealing to other treatments of those estimators or by appropriately modifying the proof strategy employed in Section 2.

We begin our asymptotic distribution discussion by describing the underlying assumptions which we make. We extend the index set of the stochastic process containing the observable variables from the nonnegative integers to include all of the integers. For studying probabilistic properties, Doob [7, p. 456] argues that this extension is innocuous.

ASSUMPTION 3.1: $\{x_n : -\infty < n < +\infty\}$ is stationary and ergodic.

We modify the specification and role of the set S in the analysis.

ASSUMPTION 3.2: S is an open subset of R^q that contains β_0.

We use the metric implied by the absolute value norm to define our notion of convergence on S. We place additional requirements on the function f and the process $\{x_n : -\infty < n < +\infty\}$.

ASSUMPTION 3.3: $f(\cdot, \beta)$ and $\partial f/\partial\beta(\cdot, \beta)$ are Borel measurable for each $\beta \in S$ and $\partial f/\partial\beta(x, \cdot)$ is continuous on S for each $x \in R^p$.

ASSUMPTION 3.4: $\partial f_1/\partial\beta$ is first moment continuous at β_0, and $E[\partial f/\partial\beta(x_1, \beta_0)]$ exists, is finite, and has full rank.

We adopt the notation $d_0 = E[\partial f/\partial\beta(x_1, \beta_0]$.

Our first consistency theorem (Theorem 2.1) relies on the condition that f_1 is first moment continuous. A link between this condition and Assumptions 3.3 and 3.4 is provided by Lemma 3.1.

[15] Hausman [21], among others, has provided an instrumental variables interpretation of maximum likelihood estimators by examining the first-order conditions of the maximization problem solved in obtaining the estimators. Avery, Hansen, and Holtz [3] illustrate how to apply results of this section to consistent, quasi-maximum likelihood estimators of multiperiod probit models. Hansen and Hodrick [17] apply results from this section to least squares estimators.

LEMMA 3.1: *Suppose Assumptions* 3.3 *and* 3.4 *are satisfied. If* $E[f(x_1, \beta_0)]$ *exists and is finite, then* f_1 *is first moment continuous at* β_0.[16]

When f takes the special form given by Assumption 2.6, λ is differentiable with $\partial\lambda/\partial\beta$ continuous on S, and $c_1(x_1)$ has a finite expectation, then Assumptions 3.3 and 3.4 are satisfied as long as $E[c_1(x_1)]\partial\lambda/\partial\beta(\beta_0)$ has full rank.

As in Section 2, we will think of the function f as defining the orthogonality conditions that we consider using in estimation. Let

$$w_n = f(x_n, \beta_0) \qquad \text{for} \quad -\infty < n < +\infty$$

and

$$v_j = E\left[w_0 \mid w_{-j}, w_{-j-1}, \ldots\right] - E\left[w_0 \mid w_{-j-1}, w_{-j-2}, \ldots\right] \qquad \text{for } j \geq 0.$$

Assumptions 3.1 and 3.3 imply that $\{w_n : -\infty < n < +\infty\}$ is stationary and ergodic. An iterated expectations argument can be employed to establish that $\{v_j : j \geq 0\}$ is a martingale difference sequence.

ASSUMPTION 3.5: $E[w_0 w_0']$ exists and is finite, $E[w_0 \mid w_{-j}, w_{-j-1}, \ldots]$ converges in mean square to zero, and $\sum_{j=0}^{\infty} E[v_j' v_j]^{1/2}$ is finite.

Among other things, Assumption 3.5 implies that

$$E\left[f(x_n, \beta_0)\right] = 0 \qquad \text{for} \quad -\infty < n < +\infty,^{17}$$

and provides sufficient conditions suggested by Hannan [**16**] for applying a central limit theorem for stationary, ergodic processes proved by Gordin [**13**].

We could conceive of estimating β_0 by selecting a value of β that satisfies the r equations:

$$g_N(\beta) = 0. \tag{5}$$

This may not be possible since (5) involves only q unknowns and r can exceed q. Instead, we follow Sargan [**29**, **30**] and reduce the number of equations to q by using linear combinations of the r equations. To accomplish this, we introduce a sequence of q by r random matrices $\{a_N^* : N \geq 1\}$ and make the following assumption:

ASSUMPTION 3.6: $\{a_N^* : N \geq 1\}$ converges in probability to a constant matrix a_0^* which has full rank.

[16] Abbreviated versions of the proofs of some of the results in this section and in Section 4 are provided in the Appendix. More detailed versions of the proofs can be obtained from the author on request.

[17] This implication can be seen by employing an iterated expectations argument and noting that $E[w_0] = E[f(x_n, \beta_0)]$.

We require that a GMM estimator $\{b_N^* : N \geq 1\}$ asymptotically satisfy the set equations

$$a_0^* Ef(x_n, \beta) = 0$$

in the sense of Definition 3.1.

DEFINITION 3.1: The GMM estimator $\{b_N^* : N \geq 1\}$ is a sequence of random vectors that converges in probability to β_0 for which $\{\sqrt{N} a_N^* g_N(b_N^*) : N \geq 1\}$ converges in probability to zero.

Before showing that this GMM estimator is asymptotically normal and displaying the dependence of its asymptotic covariance matrix on the limiting weighting matrix a_0^*, we discuss the link between this estimator and the GMM estimator of Definition 2.1. Note that

$$|h_N(\beta)|^2 = |a_N g_N(\beta)|^2 = g_N(\beta)' a_N' a_N g_N(\beta).$$

Assuming that the first-order conditions for the problem of minimizing $|h_N|^2$ are satisfied by b_N, then

$$\frac{\partial g_N}{\partial \beta}(b_N)' a_N' a_N g_N(b_N) = 0. \tag{6}$$

Let a_N^* be the q by r matrix

$$a_N^* = \frac{\partial g_N}{\partial \beta}(b_N)' a_N' a_N. \tag{7}$$

Substituting (6) into (7), we obtain $a_N^* g_N(b_N) = 0$ which trivially satisfies one of the key requirements of Definition 2.1. Once we establish the strong consistency of the estimator of Definition 2.1, require that the first-order conditions (6) be satisifed, and demonstrate that the sequence $\{a_N^* : N \geq 1\}$ converges in probability to a constant matrix, then we obtain a GMM estimator of Definition 3.1. Lemma 3.2 supplies sufficient conditions for $\{a_N^* : N \geq 1\}$ as defined by (7) to converge in probability to a constant matrix.

LEMMA 3.2: *Suppose Assumptions* 3.1–3.4 *are satisfied. If* (i) $\{b_N : N \geq 1\}$ *converges in probability to* β_0; (ii) $\{a_N : N \geq 1\}$ *converges in probability to* a_0; *then* $\{(\partial g_N/\partial \beta)(b_N) : N \geq 1\}$ *converges in probability to* d_0 *and* $\{a_N^* : N \geq 1\}$ *given by* (7) *converges in probability to* $a_0^* = d_0' a_0' a_0$.

While the above discussion shows how the estimators of Definition 2.1 can be viewed as GMM estimators under Definition 3.1, our asymptotic distribution is not limited to estimators of this form. Any consistent estimators which minimize

or maximize criterion functions with first-order conditions that can be represented as

$$a_N^* g_N(b_N^*) + H_N(b_N^*) = 0$$

where $\{\sqrt{N}H_N(b_N^*): N \geq 1\}$ converges in probability to zero for an appropriate choice of a_N^*, f, and H_N, can be viewed as special cases of the generic GMM estimator of Definition 3.1.[18] Thus, various forms of least squares and quasi-maximum likelihood along with nonlinear instrumental variables estimators are included in our asymptotic distribution discussion.

In preparation for our asymptotic distribution theorem, we let

$$R_w(j) = E\left[w_0 w'_{-j}\right].$$

Assumptions (3.1) and (3.5) insure that $R_w(j)$ is finite and that the matrix

$$S_w = \sum_{j=-\infty}^{+\infty} R_w(j)$$

is well defined and finite.[19] Theorem 3.1 displays the asymptotic distribution of the GMM estimator.

THEOREM 3.1: *Suppose Assumptions* 3.1–3.6 *are satisfied. Then* $\{\sqrt{N}(b_N^* - \beta_0): N \geq 1\}$ *converges in distribution to a normally distributed random vector with mean zero and covariance matrix* $(a_0^* d_0)^{-1} a_0^* S_w a_0^{*\prime} (a_0^* d_0)^{-1\prime}$.[20]

Since S_w plays a prominent role in the expression for the asymptotic covariance matrix, we shall examine Assumption (3.5) in conjunction with the computation of S_W. We focus on situations in which

$$f(x_n, \beta_0) = u_n \otimes z_n \tag{10}$$

where we view z_n as a vector of the instrumental variables and u_n is a vector of the disturbance terms from the econometric model. Let

$$R_u(j) = E\left[u_n u'_{n-j}\right]$$

and

$$R_z(j) = E\left[z_n z'_{n-j}\right]$$

and assume that $R_u(0)$ and $R_z(0)$ exist and are finite. It is instructive for us to examine five special cases.

[18] The minimax estimator of Sargan [**30**] can be interpreted as a GMM estimator with a nontrivial H_N.

[19] Under Assumptions 3.1 and 3.5 it can be shown that the elements in the autocovariance function for $\{w_n: -\infty < n < +\infty\}$ are absolutely summable.

[20] If S_w is singular, then it may be the case that the asymptotic covariance matrix for the GMM estimator is singular. If this happens, the GMM estimator has a degenerate normal asymptotic distribution.

CASE (i):

$$E[\, u_n \mid z_n, u_{n-1}, z_{n-1}, u_{n-2}, \ldots \,] = 0,$$

$$E[\, u_n u_n' \mid z_n, u_{n-1}, z_{n-1}, u_{n-2}, \ldots \,] = R_u(0).$$

This case includes as a special subcase models in which $\{u_n : -\infty < n < +\infty\}$ is a sequence of independent, random vectors and u_n is independent of $\{z_j : -\infty < j \leq n\}$. It is straightforward to verify that

$$E[\, w_0 \mid w_{-j}, w_{-j-1}, \ldots \,] = 0 \qquad \text{for} \quad j \geq 1,$$

$$v_0 = w_0,$$

and

$$v_j = 0 \qquad \text{for} \quad j \geq 1.$$

This shows that Assumption 3.5 is satisfied. Also, it can be demonstrated that

$$R_w(j) = 0 \qquad \text{for} \quad j \neq 0$$

and

$$S_w = R_w(0) = R_u(0) \otimes R_z(0).$$

Thus, S_w can be computed from the second moments of z_n and u_n.

CASE (ii):

$$E[\, u_n \mid z_n, u_{n-1}, z_{n-1}, u_{n-2}, \ldots \,] = 0.$$

This case differs from Case (i) in that we no longer assume that the conditional covariance matrix for u_n is independent of the conditioning set. This allows for a particular form of heteroskedasticity. The stationarity assumption, however, restricts us to circumstances in which the unconditional variances of $\{u_n : -\infty < n < +\infty\}$ are constant. As in Case (i), it can be verified that Assumption 3.5 of Theorem 3.1 is met and that

$$R_w(j) = 0 \qquad \text{for} \quad j \neq 0.$$

In contrast with Case (i), we can no longer compute $R_w(0)$, and consequently S_w, from the second moments of u_n and z_n. More specifically, we have

$$S_w = R_w(0) = E[\, u_n u_n' \otimes z_n z_n' \,].$$

This form of S_w arises in the multiperiod probit estimators proposed by Avery, Hansen, and Hotz **[3]**.

CASE (iii):

$$E[\, u_n \mid z_n, u_{n-k}, z_{n-1}, u_{n-k-1}, \ldots \,] = 0.$$

Models in which the disturbance term is orthogonal to an extensive information set shifted back k time periods, such as the nonlinear rational expectations models studied by Hansen and Singleton [**20**] and the linear (in the variables) rational expectations models studied by Cumby, Huizinga, and Obstfeld [**5**] are included in this case. It can be verified that

$$E[\, w_0 \,|\, w_{-j}, w_{-j-1}, \ldots \,] = 0 \qquad \text{for} \quad j \geq k$$

and

$$v_j = 0 \qquad \text{for} \quad j \geq k.$$

This means that Assumption 3.5 is satisfied. Also,

$$R_w(j) = 0 \qquad \text{for} \quad j \geq k$$

and

$$S_w = \sum_{j=k+1}^{k-1} R_w(j).$$

Computation of S_w entails only the determination of a finite number of the autocovariances of $\{w_n : -\infty < n < +\infty\}$.

CASE (iv):

$$E[\, u_n \,|\, z_n, u_{n-k}, z_{n-1}, u_{n-k-1}, \ldots \,] = 0,$$

$$E[\, u_n u_{n-j} \,|\, z_n, u_{n-k}, z_{n-1}, u_{n-k-1}, \ldots \,] = R_u(j) \qquad \text{for} \quad 0 \leq j < k.$$

This case is embedded in Case (iii), and thus we know that assumptions (ii) and (iii) of Theorem 3.1 are satisfied. Since the conditional autocovariances of $\{u_n : -\infty < n < +\infty\}$ are assumed constant, it follows that

$$R_w(j) = R_u(j) \otimes R_z(j) \qquad \text{for} \quad -k+1 \leq j \leq k-1$$

and

$$S_w = \sum_{j=k+1}^{k-1} R_u(j) \otimes R_z(j).$$

Thus S_w can be computed from a finite number of the autocovariances of $\{z_n : -\infty < n < +\infty\}$ and $\{u_n : -\infty < n < +\infty\}$. Brown and Maital [**4**] and Hansen and Hodrick [**18**] use these assumptions to study k-step ahead forecasting equations. McCallum [**28**] employs these assumptions in proposing instrumental variables procedures for linear rational expectations models.

One set of sufficient conditions for the conditional autocovariances of $\{u_n : -\infty < n < +\infty\}$ to be constant is as follows. Suppose $y_n' = (u_n', z_{n+k}')$ and that the conditional expectation

$$E[\, y_n \,|\, y_{n-1}, y_{n-2}, \ldots \,]$$

is equal to the corresponding best linear predictor. Also, let

$$y_n - E[\, y_n \mid y_{n-1}, y_{n-2}, \cdots \,] = u_n^*$$

and suppose that

$$E[\, u_n^* u_n^* \mid y_{n-1}, y_{n-2}, \cdots \,]$$

is constant and hence independent of elements in the conditioning set. These are sufficient to imply that

$$E[\, u_n u_{n-j}' \mid z_n, u_{n-k}, z_{n-1}, u_{n-k-1}, \cdots \,] = R_u(j).$$

CASE (v): Suppose that the $\{(u_n', z_n') : -\infty < n < +\infty\}$ process is linearly regular and has fourth order cumulants that are zero. For simplicity we assume that

$$E[\, z_n \,] = 0,$$

$$E[\, u_n \,] = 0.$$

Let

$$E[\, u_n z_{n-j}' \,] = \begin{bmatrix} R_{uz}^1(j) \\ R_{uz}^2(j) \\ \vdots \\ R_{uz}^L(j) \end{bmatrix}$$

where L is the number of elements in the disturbance vector and where $R_{uz}^k(j)$ is an M dimensional row vector with the same number of elements as are in the instrument vector. Define

$$\bar{R}_w(j) = \begin{bmatrix} R_{uz}^1(-j)' R_{uz}^1(j) & R_{uz}^2(-j)' R_{uz}^1(j) & \cdots & R_{uz}^L(-j)' R_{uz}^1(j) \\ R_{uz}^1(-j)' R_{uz}^2(j) & R_{uz}^2(-j)' R_{uz}^2(j) & \cdots & R_{uz}^L(-j)' R_{uz}^2(j) \\ \vdots & \vdots & & \vdots \\ R_{uz}^1(-j)' R_{uz}^L(j) & R_{uz}^2(-j)' R_{uz}^L(j) & \cdots & R_{uz}^L(-j)' R_{uz}^L(j) \end{bmatrix}.$$

Then it can be shown that

$$R_w(j) = \bar{R}_w(j) + R_u(j) \otimes R_z(j),$$

$$S_w = \sum_{j=-\infty}^{+\infty} \left[\bar{R}_w(j) + R_u(j) \otimes R_z(j) \right]. \tag{8}$$

An alternative representation for S_w can be obtained using spectral density

matrices. Let

$$S_u(\omega) = \sum_{j=-\infty}^{+\infty} e^{-i\omega j} R_u(j),$$

$$S_z(\omega) = \sum_{j=-\infty}^{+\infty} e^{-i\omega j} R_z(j),$$

$$S_{uz}^k(\omega) = \sum_{j=-\infty}^{+\infty} e^{-i\omega j} R_{uz}^k(j).$$

The following relationships hold:

$$\sum_{j=-\infty}^{+\infty} R_u(j) \otimes R_z(j) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[S_u(\omega) \otimes S_z(-\omega) \right] d\omega,$$

$$\sum_{j=-\infty}^{+\infty} R_{uz}^k(-j)' R_{uz}^l(j) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[S_{uz}^k(\omega)' S_{uz}^l(\omega) \right] d\omega. \tag{9}$$

Substituting relations (9) into (8) yields an equivalent representation for S_w. As noted above, Assumption 3.5 implies that $R_{uz}(0) = 0$. In the rational expectations models studied by Hansen and Sargent [**19**], it is assumed that $R_{uz}(j) = 0$ for $j \geq 0$. This additional assumption can be used to simplify the expressions obtained in (8) and (9).

The five special cases discussed above illustrate how auxiliary assumptions imply alternative formulas for calculating S_w. These auxiliary assumptions also can be used to obtain formulas for models with orthogonality conditions that have representations other than (10). Assumption 3.5, however, accommodates models that do not necessarily satisfy the defining assumptions of any of the five special cases discussed above. Some of the models examined by Hansen and Sargent [**19**] are not included in these cases as well as models whose orthogonality conditions emerge because certain equations define best linear predictors but not conditional expectations. Theorem 3.1 can be applied to these models as well.

In order to make asymptotically valid inferences and construct asymptotically correct confidence regions, it is necessary to have consistent estimators of a_0^*, d_0, S_w. Since $\{a_N^* : N \geq 1\}$ is assumed to converge in probability to a_0, we can use a_N^* as our estimator of a_0^*. A natural candidate for estimating d_0 is $d_N = (1/N)\, \partial g_N / \partial \beta(b_N^*)$. From Lemma 3.2 and Assumptions (3.1)–(3.4) of Theorem 3.1, d_N is consistent. Consistent estimation of S_w is a little more involved. Let

$$w_n^N = f(x_n, b_N^*),$$

$$R_w^N(j) = \frac{1}{N} \sum_{n=1+j}^{N} w_n^N w_{n-j}^{N'}.$$

Lemma 3.3 provides conditions that are sufficient to guarantee that $R_w(j)$ is a consistent estimator of $R_w(j)$.

LEMMA 3.3: *Suppose Assumptions* 3.1–3.5 *are satisfied. If* f_1 *is second moment continuous at* β_0, *then* $\{R_w^N(j): N \geq 1\}$ *converges to* $R_w(j)$ *in probability.*

In situations where S_w depends on a finite number of autocovariances, i.e.,

$$S_w = \sum_{j=-k+1}^{k-1} R_w(j),$$

we can use Lemma 3.2 to argue that:

$$S_w^N = \sum_{j=-k+1}^{k-1} R_w^N(j)$$

is a consistent estimator of S_w. Furthermore, if the conditional covariance assumptions of Case (i) or (iv) are met, the special structure of S_w can be exploited even further. Lemma 3.3 can be used to establish consistency of the sample autocovariances of the estimated disturbances and the instruments. In the general case, S_w cannot necessarily be computed from a finite number of autocovariances which complicates its consistent estimation. However, S_w is the spectral density matrix of $\{w_n: -\infty < n < +\infty\}$ at frequency zero, and a consistent estimator of S_w can be obtained by using procedures appropriate for estimating spectral density matrices.[21]

Up until now we have said relatively little about selection of the matrices $\{a_N^*: N \geq 1\}$. As is clear from the conclusion of Theorem 3.1, different choices of these weighting matrices give rise to GMM estimators with different asymptotic covariance matrices. In fact, Theorem 3.1 provides a convenient scheme for comparing the asymptotic distributions of elements of a whole family of econometric estimators formed by taking different weighted averages of the orthogonality conditions that emerge from the model. One could conceive of determining an "optimal" estimator from this class, where an optimal estimator is one that has an asymptotic covariance matrix at least as small as any other element in the class. This approach can be viewed as an extension of Sargan's [**29**, **30**] discussion of how to obtain the optimal linear combinations of instruments to use in estimation given a finite set of instruments are specified *a priori*. Given a finite set of orthogonality conditions we show the optimal linear combinations (the a_0^* matrix) to use in estimating β_0. In describing the solution to this optimization problem, it is convenient to introduce some definitions and notation.

For a given function f and a given stochastic process $\{x_n: -\infty < n < +\infty\}$, we maintain Assumptions 3.1–3.5 and we assume that S_w is nonsingular. Associated with f and $\{x_n: -\infty < n < +\infty\}$, we consider a family A of GMM

[21] See Hannan [**15**] for a discussion of alternative strategies for estimating spectral density matrices. In this paper we do not formally establish consistency of these spectral estimators of S_w. At the very least, it appears we would want to make the additional assumption that $\partial f_1/\partial\beta$ be second-moment continuous at β_0 in proving consistency. Hannan [**16**] provides some comments about consistent estimation of spectral density matrices under Assumptions 3.1 and 3.5. An advantage of spectral estimators of S_w over truncated autocovariance estimators is that spectral estimators are constrained to be positive semidefinite.

estimators of β_0. To each element of A, we assume there corresponds a sequence of q by r weighting matrices that converge in probability to a constant matrix of full rank such that the element of A satisfies Definition 3.1 of a GMM estimator using this particular sequence of weighting matrices. Two elements of A are said to be asymptotically equivalent if they have the same asymptotic covariance matrix. Using Theorem 3.1, it is obvious that if two elements of A have the same limiting weighting matrix, then they are asymptotically equivalent. We now consider a theorem that provides us with characterization of an optimal estimator.

THEOREM 3.2: *Suppose* $\{b_N^* : N \geq 1\} \in A$ *and that the limiting weighting matrix associated with* $\{b_N^* : N \geq 1\}$ *satisfies*

$$(10) \qquad (a_0^* d_0)^{-1} a_0^* S_w a_0^* (a_0^* d_0)^{-1\prime} = \left(d_0' S_w^{-1} d_0\right)^{-1}.$$

Then $\{b_N^* : N \geq 1\}$ *is optimal with asymptotic covariance matrix* $(d_0' S_w^{-1} d_0)^{-1}$. *Furthermore, all optimal estimators in A will have a limiting weighting matrix that satisfies* (10), *and*

$$(11) \qquad a_0^* = e d_0' S_w^{-1}$$

for some q by q nonsingular matrix e.

In order to determine an optimal choice of a_0^* of the form specified in (11), it is necessary to have a consistent estimator of d_0 and S_w. This can be accomplished by initially employing a not necessarily optimal GMM estimator and using one of the estimation strategies mentioned earlier for S_w and d_0.

In considering the GMM estimators of Section 2, we indicated that from the standpoint of consistency it may be desirable to employ a square nonsingular matrix a_0 as the limiting weighting matrix. If we choose a_0 such that $a_0' a_0 = S_w^{-1}$ and a_N is some consistent estimator of a_0, then Lemma 3.2 informs us that the corresponding a_0^* is equal to $d_0 S_w^{-1}$. Thus the resulting estimator is optimal. Under the assumptions defining Case (i) above, the choice of $a_0' a_0 = S_w^{-1}$ yields the nonlinear instrumental variables estimators discussed by Amemiya [**1**], Jorgenson and Laffont [**24**], and Gallant [**11**].

Our optimality result in Theorem 3.2 is limited in that it takes the specification of the orthogonality conditions as given and does not discuss how to construct optimally orthogonality conditions. Amemiya [**2**] describes how to accomplish this latter task in environments with serially uncorrelated disturbances. A drawback of his approach is that in many circumstances his construction is not possible to implement in practice. A related limitation of our optimality result is that it only allows a finite number of orthogonality conditions to be considered. Hayashi and Sims [**22**] and Hansen and Sargent [**19**] discuss optimality in linear environments in circumstances where orthogonality conditions

$$E\left[u_n \otimes z_{n-m}\right] = 0$$

for $m \geq 0$ can be used in estimation. Such a specification admits infinitely many orthogonality conditions in constructing instrumental variables estimators. These authors allow the disturbances to be serially correlated, but they rule out conditional heteroskedasticity.

4. TESTING OVER-IDENTIFYING RESTRICTIONS

When the number of orthogonality conditions, r, exceeds the number of parameters to be estimated, q, tests of the restrictions implied by the econometric model are available. As was noted in Section 3, estimation of the model parameters sets q linear combinations of the r sample orthogonality conditions equal to zero, at least asymptotically. Thus, when the model is true, there are $r - q$ linearly independent combinations of the orthogonality conditions that ought to be close to zero but are not actually set to zero. This provides us with a scheme for testing the over-identifying restrictions of the model which is elaborated upon below.

Using the results of Section 3, we can obtain the asymptotic distribution of $\{\sqrt{N} g_N(b_N^*) : N \geq 1\}$. Recall that $g_N(b_N^*)$ is an expression for the sample orthogonality conditions evaluated at the parameter estimator b_N^*. Lemma 4.1 provides the desired asymptotic result.

LEMMA 4.1: *Suppose Assumptions* 3.1–3.6 *are satisfied. Then* $\{\sqrt{N} g_N(b_N^*) : N \geq 1\}$ *converges in distribution to a normal random vector with mean zero and covariance matrix* $\zeta_0 = [I - d_0(a_0^* d_0)^{-1} a_0^*] S_w [I - d_0(a_0^* d_0)^{-1} a_0^*]'$.

Since we have assumed that $\{\sqrt{N} a_N^* g_N(b_N^*)] : N \geq 1\}$ converges to zero in probability, it is reasonable to suspect that the asymptotic covariance matrix ζ_0 given in Lemma 4.1 is singular. We can verify this singularity by premultiplying ζ_0 by a_0^* and obtaining a matrix of zeroes. Although ζ_0 is singular, if S_w is nonsingular and r exceeds q, then ζ_0 is not zero. Hence there are linear combinations of the sample orthogonality conditions that have a nondegenerate asymptotic distribution. These linear combinations of sample orthogonality conditions can be used to obtain asymptotically valid test statistics of the model restrictions.

We wish to examine a particularly convenient test statistic of this form. This test can be viewed as an extension of a specification test proposed by Sargan [**29**, **30**] and of the specification test associated with minimum chi-square estimators (see Ferguson [**10**]). Let

$$\tau_N = g_N(b_N^*)' \left(S_w^N\right)^{-1} g_N(b_N^*),$$

where $\{S_w^N : N \geq 1\}$ is a consistent estimator of S_w. Its asymptotic distribution is given in Lemma 4.2 assuming $\{b_N^* : N \geq 1\}$ is an optimal estimator as defined in Section 3.

LEMMA 4.2: *Suppose Assumptions* 3.1–3.6 *of Theorem* 4.1 *are satisfied and that* $a_0 = e d_0' S_w^{-1}$ *for some* q *by* q *nonsingular matrix* e. *Then* $N\tau_N$ *converges in*

distribution to a chi-square distributed random variable with $r - q$ degrees of freedom.

Recall from Section 3 that choosing a_N such that $a_N' a_N = (S_w^N)^{-1}$ gives rise to a GMM estimator that is optimal and has a nonsingular limiting weighting matrix. Lemma 4.2 provides us with the asymptotic distribution of the minimized value of the criterion function $|a_N g_N(\cdot)|^2$. This can be used as a statistic to test the over-identifying restrictions of the econometric model.[22]

5. CONCLUSION

This paper has provided a discussion of the large sample properties of a class of econometric estimators that are defined in terms of orthogonality conditions. Viewing estimation in this way is convenient for comparing estimators that exploit, at least implicitly, the same set of orthogonality conditions and is suggestive of computationally practical estimators in situations in which asymptotically efficient estimation is computationally burdensome. The contribution of this paper is to provide a discussion of consistency and asymptotic normality of estimators under conditions not previously examined by other researchers.

In our discussion we exploited the assumption that the underlying stochastic process of observables is stationary and ergodic. Assumptions of this nature oftentimes play a role in model specification. Lucas [**25**] and Lucas and Sargent [**26**] have emphasized that in time series modeling based on dynamic theory, the stochastic properties of the forcing variables play a critical role in model specification. Characterizing the forcing variables as a stationary process is clearly convenient in deriving the dynamic decision rules of economic agents because stationary processes have time invariant probability laws. Furthermore, the stationary assumption and the theorems in this paper accommodate potentially complicated conditional covariance structures for the disturbance terms and the observable variables. On the other hand, there exist many situations in which it would be useful to relax the stationarity assumption. It seems likely, however, that such extensions will either employ more obscure regularity conditions, or will employ regularity conditions that are not uniformly weaker than those used in this paper. Nonetheless, it would be a useful exercise to examine the extent to which results like those obtained here remain intact or could easily be modified when a subset of forcing variables are not stationary (even though they may have a time invariant representation). Along this vein, extensions of the cross-sectional results of White [**32**, **33**, **34**] and cross-sectional and time series

[22]This result can be viewed as an extension of Sargan's [**29**, **30**] derivation of the asymptotic distribution of what he refers to as the smallest characteristic root. Gallant and Jorgenson [**12**] propose a test for restrictions that is the nonlinear three-stage least-squares analogue of the likelihood ratio test. A derivation of the asymptotic distribution of their test statistic could be obtained in the estimation environment considered here. Avery, Hansen, and Hotz [**3**] use Lemma 4.1 directly to obtain some alternative specification tests.

results of Eicker [8] to the class of time series estimators considered here would be of interest.

Carnegie-Mellon University

Manuscript received March, 1979; final revision received July, 1981.

APPENDIX

In this Appendix we provide a brief sketch of some of the results in Sections 3 and 4. A more detailed version of the proofs to all of the lemmas and theorems presented in this paper is available from the author on request.

PROOF OF THEOREM 3.1: We write $\partial g_N/\partial\beta$ in terms of its r row functions:

$$\frac{\partial g_N}{\partial \beta} = \begin{bmatrix} \frac{\partial g^1}{\partial \beta} \\ \vdots \\ \frac{\partial g_N^r}{\partial \beta} \end{bmatrix}$$

and we let

$$Dg_N(\beta^1, \ldots, \beta^r) = \begin{bmatrix} \frac{\partial g_N^1}{\partial \beta}(\beta^1) \\ \vdots \\ \frac{\partial g_N^r}{\partial \beta}(\beta^r) \end{bmatrix}.$$

Using Taylor's theorem and Assumptions 3.2–3.4, with probability arbitrarily close to one for sufficiently large N we can write

$$g_N(b_N^*) = g_N(\beta_0) + Dg_N\left(\bar{b}_N^1, \ldots, \bar{b}_N^r\right)(b_N^* - \beta_0) \tag{12}$$

where b_N^i is between β_0 and b_N^* for $i = 1, \ldots, r$. Premultiplying by a_N^*, we obtain

$$a_N^* g_N(b_N^*) = a_N^* g_N(\beta_0) + a_N^* Dg_N\left(\bar{b}_N^1, \ldots, \bar{b}_N^r\right)(b_N^* - \beta_0).$$

Since $\{b_N^* : N \geq 1\}$ converges in probability to β_0, it follows that $\{\bar{b}_n^i : N \geq 1\}$ converges in probability to β_0 for $i = 1, \ldots, r$. Thus Lemma 4.2 implies that $\{Dg_N(\bar{b}_N^1, \ldots, \bar{b}_N^r) : N \geq 1\}$ converges in probability to d_0. Using Assumptions (3.4) and (3.6) we know that for sufficiently large N with probability arbitrarily close to one we can write

$$\begin{aligned} b_N^* - \beta_0 = &-\left[a_N^* Dg_N\left(\bar{b}_N^1, \ldots, \bar{b}_N^r\right)\right]^{-1} a_N^* g_N(\beta_0) \\ &+\left[a_N^* Dg_N\left(\bar{b}_N^1, \ldots, \bar{b}_N^r\right)\right]^{-1} a_N^* g_N(b_N^*). \end{aligned} \tag{13}$$

Using Assumptions 3.1, 3.5, and Theorem 1 in Hannan [17], it can be shown that $\{\sqrt{N} g_N(\beta_0) : N \geq 1\}$ converges in distribution to a normally distributed random vector with mean zero and covariance matrix S_w. We use Assumption (3.6) to conclude that $\{\sqrt{N}(b_N^* - \beta_0) : N \geq 1\}$ converges in

distribution to a normally distributed random vector with mean zero and covariance matrix

$$(a_0^* d_0)^{-1} a_0^* S_w a_0^{*\prime} (a_0^* d_0)^{-1\prime}.$$

PROOF OF THEOREM 3.2: First, we factor $S_w = CC'$ where C is r by r and nonsingular. Second, we let

$$D = (a_0^* d_0)^{-1} a_0^* C - \left(d_0' S_w^{-1} d_0\right)^{-1} d_0' C^{-1\prime}.$$

Third, we note that

$$DC^{-1} d_0 = 0.$$

Fourth, we verify that

$$(a_0^* d_0)^{-1} a_0^* S_w a_0^{*\prime} (a_0^* d_0)^{-1\prime} = DD' + \left(d_0' S_w^{-1} d_0\right)^{-1}.$$

Thus, $(d_0' S_w^{-1} d_0)^{-1}$ is a lower bound for the asymptotic covariance matrix of elements in A. This lower bound is attained if and only if $D = 0$.

We can premultiply D by $a_0^* d_0$, postmultiply D by C^{-1}, and claim that if $D = 0$, then

$$a_0^* - a_0^* d_0 \left(d_0' S_w^{-1} d_0\right)^{-1} d_0' S_w^{-1} = 0$$

or

$$a_0^* = e d_0' S_w^{-1}$$

where

$$e = a_0^* d_0 \left(d_0' S_w^{-1} d_0\right)^{-1}.$$

On the other hand, if we let

$$a_0^* = e d_0' S_w^{-1}$$

for some q by q nonsingular matrix e, we can verify that $D = 0$.

PROOF OF LEMMA 4.1: We substitute (12) into (13) and obtain

$$\sqrt{N} g_N(b_N^*) = \left[I - Dg_N\left(\bar{b}_N^1, \ldots, \bar{b}_N^r\right) \left\{ a_N^* Dg_N\left(\bar{b}_N^1, \ldots, \bar{b}_N^r\right) \right\}^{-1} a_N^* \right] \sqrt{N} g_N(\beta_0)$$
$$+ Dg_N\left(\bar{b}_N^1, \ldots, \bar{b}_N^r\right) \left[a_N^* Dg_N\left(\bar{b}_N^1, \ldots, \bar{b}_N^r\right) \right]^{-1} a_N^* g_N(b_N^*).$$

Recall from the Proof of Theorem 3.1,

$$Dg_N\left(\bar{b}_N^1, \ldots, \bar{b}_n^r\right) \to d_0,$$

$$a_N^* \to a_0^*$$

in probability and $\{\sqrt{N} g_N(\beta_0) : N \geq 1\}$ converges to a normally distributed random vector with mean zero and covariance matrix S_w. The conclusion of Lemma 4.1 follows immediately.

PROOF OF LEMMA 4.2: First, we factor $S_w^N = C_N C_N'$ and adopt some normalization so that

$$C_N \to C \text{ in probability}$$

where C is nonsingular. Second, we determine the asymptotic covariance matrix of $\{\sqrt{N}(C_N)^{-1}g_N(b_N^*): N \geq 1\}$. Using Lemma 4.1 and the fact that a_N^* was chosen optimally we conclude that this covariance matrix is

$$I - C^{-1}d_0(d_0'S_w^{-1}d_0)^{-1}d_0'C^{-1\prime}$$

which is idempotent and has rank $r - q$. It follows that $\{Ng_N(b_N^*)(S_N)^{-1}g_N(b_N^*): N \geq 1\}$ is asymptotically chi-square distributed with $r - q$ degrees of freedom.

REFERENCES

[1] AMEMIYA, T: "The Nonlinear Two-Stage Least-Squares Estimator," *Journal of Econometrics*, 2(1974), 105–110.

[2] ———: "The Maximum Likelihood and Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equations Model," *Econometrica*, 45(1977), 955–968.

[3] AVERY, R. B., L. P. HANSEN, AND V. J. HOTZ: "Multiperiod Probit Models and Orthogonality Condition Estimation," Carnegie-Mellon University, Pittsburgh, Pennsylvania, 1981.

[4] BROWN, B. W., AND S. MAITAL: "What Do Economists Know? An Empirical Study of Experts' Expectations," *Econometrica*, 49(1981), 491–504.

[5] CUMBY, R., J. HUIZINGA, AND M. OBSTFELD: "Two-Step, Two-Stage Least Squares Estimation in Models with Rational Expectations," National Bureau of Economic Research, Technical Paper 11, 1981.

[6] DEGROOT, M. H.: *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

[7] DOOB, J. L.: *Stochastic Processes*. New York: John Wiley and Sons, 1953.

[8] EICKER, F.: "Limit Theorems for Regressions with Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability I*, ed. by L. M. LeCam and J. Neyman. Berkeley: University of California Press, 1967.

[9] ENGLE, R. F.: "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of Inflationary Expectations," University of California, San Diego, Department of Economics Discussion Paper 79-39, 1979.

[10] FERGUSON, T. S.: "A Method of Generating Best Asymptotically Normal Estimates with Application to the Estimation of Bacterial Densities," *Annals of Mathematical Statistics*, 29(1958), 1046–1062.

[11] GALLANT, A. R.: "Three-Stage Least-Squares Estimation for a System of Simultaneous, Nonlinear, Implicit Equations," *Journal of Econometrics*, 5(1977), 71–88.

[12] GALLANT, A. R. AND D. W. JORGENSON: "Statistical Inference for a System of Nonlinear, Implicit Equations in the Context of Instrumental Variable Estimation," *Journal of Econometrics*, 11(1979), 275–302.

[13] GORDIN, M. I.: "The Central Limit Theorem for Stationary Processes," *Soviet Mathematics-Doklady*, 10(1969), 1174–1176.

[14] HAKKIO, C. A.: "Expectations and the Forward Exchange Rate," National Bureau of Economic Research Working Paper No. 439, 1980.

[15] HANNAN, E. J.: *Multiple Time Series*. New York: John Wiley and Sons, 1970.

[16] ———: "Central Limit Theorems for Time Series Regression," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 26(1973), 157–170.

[17] HANSEN, L. P., AND R. J. HODRICK: "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis," *Journal of Political Economy*, 88(1980), 829–853.

[18] HANSEN, L. P., AND T. J. SARGENT: "Formulating and Estimating Dynamic Linear Rational Expectations Models," *Journal of Economic Dynamics and Control*, 2(1980), 7–46.

[19] ———: "Instrumental Variables Procedures for Linear Rational Expectations Models," forthcoming in *Journal of Monetary Economics*.

[20] HANSEN, L. P., AND K. J. SINGLETON: "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," forthcoming in *Econometrica*.

[21] HAUSMAN, J. A.: "An Instrumental Variable Approach to Full Information Estimators for Linear and Certain Nonlinear Econometric Models," *Econometrica*, 43(1975), 727–738.

[22] HAYASHI, F., AND C. A. SIMS: "Nearly Efficient Estimation of Time Series Models with Predetermined, But Not Exogenous, Instruments," University of Minnesota, 1981.

[23] HUBER, P. J.: "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability I*, ed. by L. M. LeCam and J. Neyman. Berkeley: University of California Press, 1967.
[24] JORGENSON, D. W., AND J. LAFFONT: "Efficient Estimation of Nonlinear Simultaneous Equations with Additive Disturbances," *Annals of Economic and Social Measurement*, 3(1974), 615–640.
[25] LUCAS, R. E., JR.: "Econometric Policy Evaluation: A Critique," in *The Phillips Curve and Labor Markets, Carnegie-Rochester Conferences on Public Policy*, ed. by K. Brunner and A. H. Meltzer. Amsterdam: North Holland, 1976.
[26] LUCAS, R. E., JR., AND T. J. SARGENT: "After Keynesian Macroeconomics," in *After the Phillips Curve: Persistence of High Inflation and High Unemployment*. Boston: Federal Reserve Bank of Boston, 1978.
[27] MALINVAUD, E.: *Statistical Method of Econometrics*. Amsterdam: North Holland, 1970.
[28] MCCALLUM, B. T.: "Topics Concerning the Formulation, Estimation, and Use of Macroeconometric Models with Rational Expectations," *American Statistical Association Proceedings of the Business and Economic Statistics Section*, (1979), 65–72.
[29] SARGAN, J. D.: "The Estimation of Economic Relationships Using Instrumental Variables," *Econometrica*, 26(1958), 393–415.
[30] ———: "The Estimation of Relationships with Autocorrelated Residuals by the Use of Instrumental Variables," *Journal of the Royal Statistical Society B*, 21(1959), 91–105.
[31] SIMS, C. A.: "Are There Exogenous Variables in Short-Run Production Relationships?" *Annals of Economic and Social Measurement*, 1(1972), 17–36.
[32] WHITE, H.: "Nonlinear Regression on Cross-Section Data," *Econometrica*, 48(1980), 721–746.
[33] ———: "Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48(1980), 817–838.
[34] ———: "Instrumental Variables Regression on Cross-Section Data," San Diego: University of California Press, Department of Economics Discussion Paper 80-7, 1980.