CHAPTER 3. GENERALIZED METHOD OF MOMENTS

1. INTRODUCTION

This chapter outlines the large-sample theory of Generalized Method of Moments (GMM) estimation and hypothesis testing. The properties of consistency and asymptotic normality (CAN) of GMM estimates hold under regularity conditions much like those under which maximum likelihood estimates are CAN, and these properties are established in essentially the same way. Further, the trinity of Wald, Lagrange Multiplier, and Likelihood Ratio test statistics from maximum likelihood estimation extend virtually unchanged to this more general setting. Our treatment provides a unified framework that specializes to both classical maximum likelihood methods and traditional linear models estimated on the basis of orthogonality restrictions.

Suppose data z are generated by a process that is parameterized by a k×1 vector θ. Let $l(z,\theta)$ denote the log likelihood of z, and let θ_o denote the true value of θ in the population. Suppose there is an m×1 vector of functions of z and θ, denoted $g(z,\theta)$, that have zero expectation in the population if and only if θ equals θ_o:

$$\text{(1)} \qquad \mathbf{E}g(z,\theta) \equiv \int g(z,\theta)\cdot e^{l(z,\theta o)}dz = 0 \text{ iff } \theta = \theta_o.$$

The $\mathbf{E}\, g(z,\theta)$ are *generalized moments*, and the analogy principle suggests that an estimator of θ_o can be obtained by solving for θ that makes the sample analogs of the population moments small. Assume that linear dependancies among the moments are eliminated, so that $g(z,\theta_o)$ has a positive definite m×m covariance matrix. We say that the problem is *under-identified* if m < k, *just-identified* if m = k, and *over-identified* if m > k. If m > k, there are *over-identifying* moments that can be used to improve estimation efficiency and/or test the internal consistency of the model.

In this setup, there are several alternative interpretations of z. It may be the case that z is a complete description of the data and $l(z,\theta)$ is the "full information" likelihood. Alternately, some components of observations may be margined out, and $l(z,\theta)$ may be a marginal "limited information" likelihood. Examples are the likelihood for one equation in a simultaneous equations system, or the likelihood for continuous observations that are classified into discrete categories. Also, there may be "exogenous" variables (covariates), and the full or limited information likelihood above may be written conditioned on the values of these covariates. From the standpoint of statistical analysis, variables that are conditioned out behave like constants. Then, it does not matter for the discussion of estimation and hypothesis testing that follows which interpretation above applies, except that when regularity conditions are stated it should be understood that they hold almost surely with respect to the distribution of covariates.

Suppose an i.i.d. sample $z_1,...,z_n$ from the data generation process. A *GMM estimator* of θ_o is a vector T_n that minimizes the generalized distance of the sample moments from zero, where this generalized distance is defined by the quadratic form

(2) $$Q_n(\theta) = \tfrac{1}{2} g_n(\theta)' W_n(\tau_n) g_n(\theta), \quad \text{with} \quad g_n(\theta) \equiv \frac{1}{n} \sum_{t=1}^{n} g(z_t,\theta) \ ,$$

where $W_n(\theta)$ is a m×m positive definite symmetric matrix, in general depending on θ, that is evaluated at some sequence of "preliminary estimates" τ_n. The $W_n(\tau_n)$ define a "distance metric". For brevity, we will let W_n denote $W_n(\tau_n)$. We will assume that $W_n(\theta)$ converges in probability uniformly in θ to a continuous positive definite limit $W(\theta)$ and that W_n converges to a positive definite limit W. This will usually be the result of having preliminary estimates τ_n that converge in probability to θ_o, so that the rules for probability limits imply that $W_n(\tau_n)$ converges in probability to $W(\theta_o)$. Note that it is unnecessary to know the form of the log likelihood function $l(z,\theta)$ in order to calculate the GMM estimator, and in fact GMM estimation is particularly useful when $l(z,\theta)$ is not completely specified and only the moment condition $E\, g(z,\theta_o) = 0$ can be assumed. However, some statistical properties of GMM estimators (e.g., asymptotic efficiency) will depend on the interplay of $g(z,\theta)$ and $l(z,\theta)$.

For the GMM estimator to have good statistical properties, we will require either that $Q_n(\theta)$ have a unique global minimum with probability approaching one as sample size increases, or that we have some method that with probability approaching one can pick out the "true" global minimum from among contending candidates. In the just-identified case m = k, the matrix W_n does not enter the first-order-conditions for T_n (Verify), and could be chosen by default be the m×m identity matrix. However, even if the estimation problem is just identified for unconstrained estimation, the distance metric will matter in hypothesis testing when $Q_n(\theta)$ is minimized subject to constraint.

In the over-identified case m > k, not all the components of $g_n(T_n)$ can be made zero simultaneously, and the matrix W_n influences the estimator by determining how deviations from zero are weighted. Define the m×m covariance matrix of the moments, $\Omega(\theta) \equiv \mathbf{E}\, g(z,\theta)g(z,\theta)'$. Efficient weighting of a given set of m moments requires that W_n converge to $\Omega(\theta_o)^{-1}$ as $n \to \infty$. The reason is essentially the same as the reason underlying generalized least squares: when observations are correlated or have different variances, it is efficient to give less weight to observations that have high variances or are highly correlated. We shall term a GMM estimator that has W_n converging to $\Omega(\theta_o)^{-1}$ a *best* GMM estimator. A good candidate for W_n is $\Omega_n(\tau_n)^{-1}$, where

(3) $$\Omega_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} g(z_t,\theta)g(z_t,\theta)' ,$$

and τ_n is a consistent preliminary estimate of θ_o. One good way to get a consistent preliminary estimator τ_n is to minimize a GMM criterion that uses the identity matrix $\mathbf{I}_m$ for W_n.

Define the m×k Jacobean matrix $G(\theta) \equiv -\mathbf{E}\, \nabla_\theta g(z,\theta)$, and let

(4) $$G_n(\theta) = \frac{-1}{n} \sum_{t=1}^{n} \nabla_\theta g(z_t,\theta).$$

Then the array $G_n(\tau_n)$ evaluated at a consistent preliminary estimate τ_n of θ_o has probability limit $G(\theta_o)$. Hereafter, Ω_n and G_n will be used as shorthand for $\Omega_n(\tau_n)$ and $G_n(\tau_n)$, respectively, and Ω and G will be used as shorthand for $\Omega(\theta_o)$ and $G(\theta_o)$.

Under the regularity conditions given later in Theorem 1, we will show that a GMM estimator with a distance metric W_n that converges in probability to a positive definite matrix W will be CAN with an asymptotic covariance matrix $(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$, and a best GMM estimator with a distance metric W_n that converges in probability to $\Omega(\theta_o)^{-1}$ will be CAN with an asymptotic covariance matrix $(G'\Omega^{-1}G)^{-1}$. The following lemma justifies the sorbeque "best":

Lemma 3.1. $(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1} - (G'\Omega^{-1}G)^{-1}$ *is positive semidefinite*.

Proof: Consider the matrix $I - \Omega^{-1/2}G\,(G'\Omega^{-1}G)^{-1}\,G'\Omega^{-1/2}$. Multiply this matrix by itself and note that you get the same matrix back, so it is idempotent, and therefore positive semidefinite. Postmultiply this matrix by $\Omega^{1/2}WG(G'WG)^{-1}$ and premultiply it by the transpose of this matrix. The result, which is $(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1} - (G'\Omega^{-1}G)^{-1}$, must again be positive semidefinite. □

Exercise 1. Prove Lemma 3.1 by constructing a regression model $y = \Omega^{-1/2}G\beta + \nu$ with m observations and k parameters that satisfies Gauss-Markov assumptions. Then the OLS covariance matrix is smaller than the one for the transformed regression $W^{1/2}\Omega^{1/2}y = W^{1/2}G\beta + W^{1/2}\Omega^{1/2}\nu$.

Exercise 2. Show in Lemma 1 that if m = k, so that all the matrices in $(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$ are square and non-singular, then one can collect terms and the expression reduces to $(G')^{-1}\Omega(G)^{-1} = (G'\Omega^{-1}G)^{-1}$. This confirms that in the just-identified case W does not matter.

Several special cases of the general GMM setup occur frequently in applications: First, if $f(z,\theta)$ is a scalar function with the property that $\mathbf{E}\,f(z,\theta_o) \le \mathbf{E}\,f(z,\theta)$, then one estimation criterion is to minimize the sample analog $f_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} f(z_t,\theta)$; this is called an *extremum estimator*. A leading example of an extremum criterion function is $f(z,\theta) = -\,l(z,\theta)$, the negative of a full or limited information log likelihood function. Then, full or limited information maximum likelihood estimators are extremum estimators. A GMM estimator with moments $g(z,\theta) = \nabla_\theta f(z,\theta)$ and any distance metric has the property that the GMM criterion is minimized at the extremum estimator. When one can guarantee that the GMM criterion has no roots other than the extremum estimator, then one can treat the extremum estimator in its equivalent GMM form. More generally, we can use the GMM apparatus if we have some method of excluding "bad" roots from the analysis. We show in Section 3.6 that an asymptotic equivalence continues to hold between an extremem estimator and a GMM estimator with moments $g(z,\theta) = \nabla_\theta f(z,\theta)$ and an appropriate distance metric when estimation is carried out under the constraints imposed by a null hypothesis.

A second special case is $z = (y,x,w)$ and $g(z,\theta) = w'(y-x\theta)$, so that the moment conditions assert orthogonality in the population between *instruments* w and regression *disturbances* $\varepsilon = y - x\theta_o$. For this problem, GMM specializes to two-stage least squares (2SLS), or if w = x, to OLS. We show in Section 3.6 that these linear regression setups generalize directly to nonlinear regression orthogonality conditions based on the form $g(z,\theta) = w'(y-h(x,\theta))$, where h is a function that is known up to the parameter θ and by assumption a vector of m exogenous variables w are orthogonal to the

regression disturbances $y - h(x,\theta_o)$. This is an important application of GMM, and as an exercise the reader should translate all general statements about GMM estimators into statements for this model.

In discussing the statistical properties of GMM estimators, we will denote *convergence in probability* by $\rightarrow_p$, and *convergence in distribution* by $\rightarrow_d$. If a sequence of events occur with probability approaching one, we say that they occur *in probability limit*. A sequence of random variables Y_n is *stochastically bounded* if for each $\varepsilon > 0$ there exists a constant M such that for all n, $\text{Prob}(|Y_n| > M) < \varepsilon$. We will sometimes use the notation $Y_n = Y_o + o_p$ for $Y_n \rightarrow_p Y_o$ and $Y_n = O_p(1)$ for a stochastically bounded sequence.

We will need some definitions for random functions on a subset Θ of a Euclidean space $\mathbb{R}^k$. Let (S,F,P) denote a probability space. Define a *random function* as a mapping Y from $\Theta \times S$ into $\mathbb{R}$ with the property that for each $\theta \in \Theta$, $Y(\theta,\cdot)$ is measurable with respect to (S,F,P). Note that $Y(\theta,\cdot)$ is simply a random variable, and that $Y(\cdot,s)$ is simply a function of $\theta \in \Theta$. Usually, the dependence of Y on the state of nature is suppressed, and we simply write $Y(\theta)$. A random function is also called a *stochastic process*, and $Y(\cdot,s)$ is termed a *realization* of this process. A random function $Y(\theta,\cdot)$ is *almost surely continuous* at $\theta_o \in \Theta$ if for s in a set that occurs with probability one, $Y(\cdot,s)$ is continuous in θ at θ_o. It is useful to state this definition in more detail. For each $\varepsilon > 0$, define

$$A_k(\varepsilon,\theta_o) = \left\{ s \in S \,\middle|\, \sup_{|\theta-\theta_o| \le 1/k} |Y(\theta,s) - Y(\theta_o,s)| > \varepsilon \right\}.$$

Almost sure continuity states that these sets converge monotonically as $k \rightarrow \infty$ to a set $A_o(\varepsilon,\theta_o)$ that has probability zero.

The condition of almost sure continuity allows the modulus of continuity to vary with s, so there is not necessarily a fixed neighborhood of θ_o independent of s on which the function varies by less than ε. For example, the function $Y(\theta,s) = \theta^s$ for $\theta \in [0,1]$ and s uniform on [0,1] is continuous at $\theta = 0$ for every s, but $A_k(\varepsilon,0) = [0, \frac{-\log \varepsilon}{\log k})$ has positive probability for all k. The exceptional sets $A_k(\varepsilon,\theta)$ can vary with θ, and there is no requirement that there be a set of s with probability one, or for that matter with positive probability, where $Y(\theta,s)$ is continuous for all θ. For example, assuming $\theta \in [0,1]$ and s uniform on [0,1], and defining $Y(\theta,s) = 1$ if $\theta \geq s$ and $Y(\theta,s) = 0$ otherwise gives a function that is almost surely continuous everywhere and always has a discontinuity.

Several results on stochastic limits that will be needed for the analysis of GMM estimators; see McFadden, "Limit Theorems in Statistics", 240A lecture notes:

Lemma 3.2. *For sequences of random vectors* Y_n *and* Z_n, *(1) for* c *a constant,* $Y_n \rightarrow_p c$ *if and only if* $Y_n \rightarrow_d c$; *(2) if* $Y_n \rightarrow_d Y_o$ *and* $Z_n - Y_n \rightarrow_p 0$, *then* $Z_n \rightarrow_d Y_o$; *and (3) if* $Y_n \rightarrow_d Y_o$ *and* f *is a continuous function on an open set containing the support of* Y_o, *then* $f(Y_n) \rightarrow_d f(Y_o)$.

Lemma 3.3 (Uniform WLLN). *Assume* $Y_i(\theta)$ *are independent identically distributed random functions with a finite mean* $\psi(\theta)$ *for* θ *in a closed bounded set* $\Theta \subseteq \mathbb{R}^k$. *Assume* $Y_i(\cdot)$ *is almost surely continuous at each* $\theta \in \Theta$. *Assume that* $Y_i(\cdot)$ *is dominated; i.e., there exists a random variable* Z *with a finite mean that satisfies* $Z \geq \sup_{\theta\in\Theta}|Y_1(\theta)|$. *Then* $\psi(\theta)$ *is continuous in* θ *and*

$X_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\theta)$ *satisfies* $\sup_{\theta\in\Theta}|X_n(\theta) - \psi(\theta)| \rightarrow_p 0$.

Lemma 3.4 (Continuous Mapping). *If* $Y_n(\theta) \rightarrow_p Y_o(\theta)$ *uniformly for* θ *in* $\Theta \subseteq \mathbb{R}^k$, *random vectors* $\tau_o,\tau_n \in \Theta$ *satisfy* $\tau_n \rightarrow_p \tau_o$, *and* $Y_o(\theta)$ *is almost surely continuous at* τ_o, *then* $Y_n(\tau_n) \rightarrow_p Y_o(\tau_o)$.

The following result gives regularity conditions under which GMM estimators have good asymptotic properties.

Theorem 3.1. (Newey and McFadden (1994, Thm. 2.6 and Thm. 3.4)) *Consider an i.i.d. sample* z_t, *for* $t = 1,...,n$*; the GMM criterion* $Q_n(\theta) = \frac{1}{2}g_n(\theta)'W_n g_n(\theta)$ *given by* (2)*, with* $W_n = W_n(\tau_n)$ *and* τ_n *a sequence of "preliminary estimates" converging in probability to a limit* τ_o*; the arrays* $\Omega_n(\theta)$ *given by* (3) *and* $G_n(\theta)$ *given by* (4)*; and the GMM estimator* $T_n = \operatorname{argmin}_{\theta\in\Theta} Q_n(\theta)$. *Assume*:

(i) *The domain* Θ *of* θ *is a compact subset of* $\mathbb{R}^k$ *and* θ_o *is in its interior*.
(ii) *The log likelihood function* $l(z,\theta)$ *is measurable in* z *for each* θ*, and almost surely (with respect to* z*) twice continuously differentiable with respect to* θ *in a neighborhood of* θ_o.
(iii) *The function* g *is measurable in* z *for each* θ, *and almost surely (with respect to* z*) is continuous on* Θ *and on a neighborhood of* θ_o *continuously differentiable in* θ, *with the derivative Lipschitz; i.e., there is a function* $\alpha(z)$ *with finite expectation such that for* θ,θ' *in the neighborhood of* θ_o, $|\nabla_\theta g(z,\theta) - \nabla_\theta g(z,\theta')| \leq \alpha(z)|\theta - \theta'|$.
(iv) $\mathbf{E}g(z,\theta) = 0$ *if and only if* $\theta = \theta_o$.
(v) $\Omega(\theta_o)$ *is a positive definite* $m\times m$ *matrix and* $G(\theta_o)$ *is an* $m\times k$ *matrix of rank* k.
(vi) $W(\theta)$ *is a positive definite* $m\times m$ *matrix that is continuous in* θ, $W_n(\theta) \rightarrow_p W(\theta)$ *uniformly in* θ, *and* $W_n \rightarrow_p W$.
(vii) *There exists a function* $\alpha(z)$, *with finite expectation, that dominates* $g(z,\theta)g(z,\theta)'$ *and* $\nabla_\theta g(z,\theta)$; i.e., $+\infty > \mathbf{E}\alpha(z)$, $|g(z,\theta)g(z,\theta)'| \leq \alpha(z)$, *and* $|\nabla_\theta g(z,\theta)| \leq \alpha(z)$.

If an estimator T_n^* *satisfies* $Q_n(T_n^*) \rightarrow_p 0$, *then* $T_n^* \rightarrow_p 0$, *and if* $n\cdot Q_n(T_n^*)$ *is stochastically bounded, then* $n^{1/2}\cdot g_n(T_n^*)$ *and* $n^{1/2}\cdot(T_n^* - \theta_o)$ *are stochastically bounded. The unconstrained GMM estimator* T_n *satisfies these conditions and is consistent and asymptotically normal* (CAN), *with*

$$n^{1/2}(T_n - \theta_o) \rightarrow_d N(0,(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}). \tag{5}$$

If in addition either $W_n \rightarrow_p \Omega^{-1}$, *or else just-identification (i.e.,* $m = k$*) with* W_n *an arbitrary non-singular matrix, then* T_n *is a best GMM estimtor that is CAN with* $B \equiv G'\Omega^{-1}G$ *and*

$$n^{1/2}(T_n - \theta_o) \rightarrow_d N(0,B^{-1}). \tag{6}$$

Before proving this result, it is useful to comment on the meaning and role of the regularity conditions (i)-(vii). Assumption (i) restricts the parameters to a closed and bounded subset of Euclidean space. This is not a substantive restriction in applications, as Θ can be very large; e.g., the set of real vectors that can be represented as floating point numbers on a computer. The condition requires that Θ contain an open neighborhood around θ_o. This restricts some applications where the true parameter is on the boundary of a feasible range, and where CAN breaks down. For example, in a regression where a coefficient is restricted to be non-negative and is truely zero, its asymptotic distribution will be a mixture of a truncated normal and a point probability. When (i) holds, the estimator can be characterized in terms of its first-order condition. Assumptions (ii), (iii), and (vii) are mathematical regularity conditions that guarantee that the moment functions are continuous and have finite variances, and that in a neighborhood of θ_o they can be differentiated. Condition (vii) is called a *dominance* condition, and guarantees that one can interchange the order of taking expectations and differentiating. These assumptions can be weakened, at the price of making the proof of CAN much more difficult, and at some point the CAN result will fail. Most applications will satisfy (ii), (iii), and (vii); exceptions are problems involving thresholds where CAN is problematic and special treatment is required. Assumption (iv) is a key identification condition that rules out both local identification failures (e.g., an interval of parameter values that explain the data equally well) and global failures (e.g., multiple roots in the limit). It is possible in applications that this assumption fails, and that the GMM procedure could pick out a "wrong" inconsistent root. However, if there is some method of sorting out multiple roots of the GMM criterion and settling on the "right" root with a probability approaching one as sample size increases, then consistency can be proved with a weaker version of (iv) that holds on some open neighborhood of θ_o. This situation may arise when the wings of the GMM criterion function contain the first-order condition for optimization of a sample function whose population expectation is optimized at θ_o, since then the heigth of the sample function can be used to sort multiple roots and pick the "right" one closest to a global optimum. Whether assumption (iv) holds in an application is a substantive issue that should be resolved by analysis of the economic model.

Assumptions (v) and (vi) are essential for the CAN result. One can show using (ii), (iii), and (vii) that $\Omega(\theta)$ is positive semidefinite, and positive definite at points in every neighborhood of θ_o, and that $G(\theta)$ is of rank k at points in every neighborhood of θ_o. Then, the definiteness of $\Omega(\theta_o)$ and the full rank of $G(\theta_o)$ are technical strengthenings of these conditions that exclude primarily pathological cases. (There are a few testing problems, discussed later, where some derivatives are identically zero under a null hypothesis and it is necessary to carry out the analysis in terms of higher-order derivatives. For example, tests for the presence of mixing will often encounter this problem.) If $\Omega_n(\theta)$ is given by (3) and $G_n(\theta)$ is given by (4), then (i), (iii), and (vii) satisfy the hypothesis of Lemma 3, implying that $\Omega_n(\theta) \to_p \Omega(\theta)$ and $G_n(\theta) \to_p G(\theta)$ uniformly in θ. Assumption (vi) holds trivially if $W_n = W$ is a positive definite array of constants, such as $\mathbf{I}_m$. The condition $W_n(\theta) \to_p W(\theta)$ uniformly in θ holds by Lemma 3 if $W_n(\theta)$ is an array of almost surely continuous dominated functions that converges pointwise to a postive definite matrix $W(\theta)$. This will be true in particular if $W_n(\theta) = \Omega_n(\theta)^{-1}$. If τ_n is a sequence converging in probability to τ_o, then $W_n = W_n(\tau_n) \to_p W(\tau_o)$ by Lemma 4. In most applications, either $W_n(\theta)$ does not depend on θ, or $W_n(\theta)$ is evaluated at a sequence of preliminary estimators τ_n that converge in probability to θ_o. Summarizing

the discussion of (i)-(vii), all the regularity conditions require checking in each application, but the one that requires the most careful examination is the identification condition (iv).

Proof of Theorem 1. A preliminary step shows that $n^{1/2}\, g_n(\theta_o)$ is asymptotically normal, that $G_n(\theta)$, $\Omega_n(\theta)$, and $W_n(\theta)$ converge in probability uniformly in θ to $G(\theta)$, $\Omega(\theta)$, and $W(\theta)$, respectively, and that $n \cdot Q_n(\theta_o)$ is stochastically bounded. The first step in the proof shows for T_n^* satisfying $Q_n(T_n^*) \to_p 0$ that $T_n^* \to_p \theta_o$. The second step shows for T_n^* satisfying $n \cdot Q_n(T_n^*)$ stochastically bounded that $n^{1/2} \cdot (T_n^* - \theta_o)$ is stochastically bounded. These two steps imply that a preliminary estimator τ_n that uses an easily calculated distance metric such as $\mathbf{I}_m$ is consistent, and hence that $\Omega_n(\tau_n) \to_p \Omega$ and $G_n(\tau_n) \to_p G$. They also imply that T_n is consistent and stochastically bounded. The third step applies the mean value theorem to the first-order condition for T_n and uses rules for asymptotic limits to show that $n^{1/2}(T_n - \theta_o)$ is asymptotically normal.

Preliminary Step: The expression $g_n(\theta_o)$ is a sample average of i.i.d. random vectors with mean zero and finite covariance matrix Ω. Then the Lindeberg-Levy central limit theorem implies

$$\Omega^{-1/2} n^{1/2} g_n(\theta_o) \equiv U_n \to_d U \sim N(0, \mathbf{I}_m). \tag{7}$$

The expressions $g_n(\theta)$, $G_n(\theta)$, and $\Omega_n(\theta)$ are sample averages that converge in probability for each fixed θ to $\mathbf{E}g(\theta)$, $G(\theta)$, and $\Omega(\theta)$, respectively, by Kinchine's law of large numbers. Conditions (i), (iii), and (vii) establish that these functions are dominated and almost surely continuous on the compact set Θ. Then the hypotheses of Lemma 3 are satisfied, so the convergence is uniform in θ. Condition (vi) gives $W_n(\theta) \to_p W(\theta)$ uniformly in θ. This condition plus (7) implies by Lemma 2 that $n \cdot Q_n(\theta_o)$ is stochastically bounded.

Step 1: Consider any estimator T_n^* that satisfies $Q_n(T_n^*) \to_p 0$. For each fixed θ, the Kinchine law of large numbers implies that $g_n(\theta) \to_p \mathbf{E}g(\theta)$. We have established that the convergence in probability of $g_n(\theta)$ to $\mathbf{E}g(\theta)$ is uniform in θ. Combined with the condition $W_n \to_p W$ from (vi), this implies $Q_n(\theta) \to_p \tfrac{1}{2}(\mathbf{E}g(\theta))' W (\mathbf{E}g(\theta))$ uniformly in θ. Outside each small neighborhood of θ_o, the probability limit of $Q_n(\theta)$ is uniformly bounded away from zero by (iv). Therefore, T_n^* is, with probability approaching one, within each small neighborhood. This establishes consistency of T_n^*.

Step 2: Consider any estimator T_n^* that satisfies $n \cdot Q_n(T_n^*)$ stochastically bounded. This condition implies $Q_n(T_n^*) \to_p 0$, and thus $T_n^* \to_p \theta_o$ by Step 1. The mean value theorem and (7) give

$$n^{1/2} g_n(T_n^*) = n^{1/2} g_n(\theta_o) - G_n\, n^{1/2}(T_n^* - \theta_o) = \Omega^{1/2} U_n - G_n\, n^{1/2}(T_n^* - \theta_o), \tag{8}$$

with G_n evaluated at points between T_n^* and θ_o. Apply the triangle inequality for the GMM distance metric to the vector $G_n\, n^{1/2}(T_n^* - \theta_o) = \Omega^{1/2} U_n - n^{1/2} g_n(T_n^*)$ to obtain

$$\tfrac{1}{2} n^{1/2}(T_n^* - \theta_o)' G_n' W_n G_n\, n^{1/2}(T_n^* - \theta_o) \le \tfrac{1}{2} U_n' \Omega^{1/2} W_n\, \Omega^{1/2} U_n + n \cdot Q_n(T_n^*). \tag{9}$$

The first term on the right-hand-side of (9) converges in distribution by Lemma 2, and hence is stochastically bounded. Together with the hypothesis that $n \cdot Q_n(T_n^*)$ is stochastically bounded, this implies that $n^{1/2}(T_n^* - \theta_o)' G_n' W_n G_n\, n^{1/2}(T_n^* - \theta_o)$ is stochastically bounded. The uniform convergence

of $G_n(\theta)$ and Lemma 4 imply $G_n'W_nG_n \to_p G'WG$ positive definite. Let $\lambda > 0$ be the smallest characteristic root of $G'WG$. Then in probability limit

$$(10) \qquad (\lambda/2)\cdot n^{1/2} \cdot |T_n^*-\theta_o|^2 \le n^{1/2}(T_n^*-\theta_o)'G_n'W_nG_n\, n^{1/2}(T_n^*-\theta_o) = O_p(1),$$

establishing that $n^{1/2}(T_n^*-\theta_o)$ is stochastically bounded. In (8), this implies that $n^{1/2}g_n(T_n^*)$ is stochastically bounded.

Step 3: Consider the GMM estimator $T_n = \text{argmin}_{\theta\in\Theta} Q_n(\theta)$. Then $Q_n(T_n) \le Q_n(\theta_o)$, and the condition that $n\cdot Q_n(\theta_o)$ is stochastically bounded implies by Steps 1 and 2 that T_n is consistent and $n^{1/2}(T_n-\theta_o)$ is stochastically bounded. The first-order condition for T_n is $0 = G(T_n)'W_n\, n^{1/2}g_n(T_n)$. Substituting the mean value expansion (7) in this first-order condition gives

$$(11) \qquad 0 = -G(T_n)'W_n\Omega^{1/2}U_n + G(T_n)'W_nG_n\, n^{1/2}(T_n-\theta_o).$$

We established in Step 2 that in probability limit, $G(T_n)'W_nG_n$ is non-singular and $(G(T_n)'W_nG_n)^{-1} \to_p (G'WG)^{-1}$. Then, $n^{1/2}(T_n-\theta_o) = (G(T_n)'W_nG_n)^{-1}\, G(T_n)'W_n\Omega^{1/2}U_n$ exists in probability limit. The array $(G(T_n)'W_nG_n)^{-1}$ converges in probability, and hence in distribution, to $(G'WG)^{-1}$; the array $G(T_n)'W_n\Omega^{1/2}$ converges in probability, and hence in distibution, to $G'W\Omega^{1/2}$; and U_n converges in distribution to U. Then Lemma 2 implies that the continuous function that is the product of these terms converges in distribution to the product of the limits; i.e., $n^{1/2}(T_n-\theta_o) \to_d (G'WG)^{-1}G'W\Omega^{1/2}U$, which is normal with covariance matrix $(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$. This establishes (5). When $W = \Omega^{-1}$ or $m = k$, (6) follows. □

In the GMM criterion (2), $W_n(\tau_n)$ is treated as an array of constants that does not vary with θ. Then the first-order condition for minimization of $Q_n(\theta)$ is

$$(12) \qquad 0 = n^{1/2}\nabla_\theta Q_n(T_n) = G_n(T_n)'W_n(\tau_n)\, n^{1/2}g_n(T_n).$$

Slightly different variants of the GMM estimator are obtained if (1) $G_n(T_n)$ in this formula is replaced by $G_n(\tau_n)$, where τ_n is a consistent preliminary estimate of θ_o, by $G_n(\theta_o)$, or by $G(\theta_o)$; and/or (2) $W_n(\tau_n)$ is replaced by $W_n(T_n)$, by $W_n(\theta_o)$, or by $W(\theta_o)$. Additional variants arise if $W_n(\theta)$ is treated as a function of θ, leading to the modified first-order condition

$$(13) \qquad 0 = G_n(T_n)'W_n(T_n)\, n^{1/2}g_n(T_n) + \text{vec}\,[\text{tr}\{[\partial W_n(T_n)/\partial\theta_r]\, n^{1/2}g_n(T_n)g_n(T_n)'\}],$$

where $\partial W_n/\partial\theta_r$ is the array of derivatives of W_n with respect to component θ_r of θ, "tr" denotes the trace of a matrix, and "vec" denotes a vector made from the components $r = 1,...,k$. One variant, commonly used for the iterative computation of GMM estimators, solves $0 = G_n(\tau_n)'W_n(\tau_n)\, n^{1/2}g_n(T_n)$, with τ_n an earlier iterate. We will show that while these variants may differ in finite samples, they are all asymptotically equivalent.

Corollary 3.1. *Suppose conditions* (i)-(vii). *Suppose* $W(\theta)$ *is continuously differentiable in a neighborhood of* θ_o, *and that the derivatives of* $W_n(\theta)$ *converge uniformly in probability limit to the derivatives of* $W(\theta)$ *on a neighborhood of* θ_o. *Then* $T_n^* = \text{argmin}_{\theta\in\Theta}\ \frac{1}{2}g_n(\theta)'W_n(\theta)g_n(\theta)$ *with* $W_n(\theta)$ *treated as a function of* θ *is asymptotically equivalent to the GMM estimator* T_n *that satisfies* (12); *i.e.,* $n^{1/2}(T_n^* - T_n) \to_p 0$, *implying* T_n^* *is CAN with the limiting distribution* (5). *Also, variants of GMM estimators that solve* (12) *or* (13)*, obtained by replacing terms with terms that have the same probability limit, are also asymptotically equivalent to* T_n, *and to a limiting GMM estimator that satisfies*

$$0 = -G'W\Omega^{1/2}U_n + G'WG\, n^{1/2}(T_n-\theta_o). \tag{14}$$

Proof: Letting $Q_n(\theta) = \frac{1}{2}g_n(\theta)'W_n(\theta)g_n(\theta)$ now denote the GMM criterion with the distance metric treated as a function of θ, the estimator T_n^* satisfies $n\cdot Q_n(T_n^*) \le n\cdot Q_n(\theta_o) = O_p(1)$, implying by Theorem 1 that T_n^* is consistent and $n^{1/2}g_n(T_n^*)$ and $n^{1/2}\cdot(T_n^* - \theta_o)$ are stochastically bounded. The final term in the first-order condition (13), $\text{vec}\,[\text{tr}\{[\partial W_n(T_n^*)/\partial\theta_r]\, n^{1/2}g_n(T_n^*)g_n(T_n^*)'\}]$, contains the product of an array $[\partial W_n(T_n^*)/\partial\theta_r]$ that converges in probability to a finite array $[\partial W(\theta_o)/\partial\theta_r]$ by Lemma 4, the stochastically bounded term $n^{1/2}g_n(T_n^*)$, and the term $g_n(T_n^*)$ that converges in probability to zero. By Lemma 2, the product of these terms converges in probability to zero. Substituting (8) into (13) then gives

$$0 = G_n(T_n^*)'W_n(T_n^*)\,\Omega^{1/2}U_n - G_n(T_n^*)'W_n(T_n^*)\,G_n\, n^{1/2}(T_n^*-\theta_o) + o_p.$$

Using the consistency and stochastic boundedness of T_n^* and Lemmas 2 and 4, this expression can be written

$$0 = G'W\Omega^{1/2}U_n - G'WG\, n^{1/2}(T_n^*-\theta_o) + o_p,$$

implying that $n^{1/2}(T_n^*-T_n) = o_p$. Further, this argument can be applied with any of the terms in (12) or (13) replaced by expressions with the same probability limit, establishing that all such variants are asymptotically equivalent to the T_n that solves (14). □

The asymptotic covariance matrices $(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$ or $B^{-1} = (G'\Omega^{-1}G)^{-1}$ can be estimated using $G_n(\tau_n)$ and $\Omega_n(\tau_n)$, where τ_n is any consistent (preliminary) estimator of θ_o, by Lemmas 3 and 4. A practical procedure for estimation is to first estimate θ_o using the GMM criterion with an arbitrary W_n, such as the m×m identity matrix $\mathbf{I}_m$. This produces an initial CAN estimator τ_n. Then use the formulas above to estimate the asymptotically efficient $W_n = \Omega_n(\tau_n)^{-1}$, and use the GMM criterion with this distance metric to obtain the final estimator T_n.

Differentiating the identity $0 \equiv \int g(z,\theta)e^{l(z,\theta)}dz$ with respect to θ, and evaluating the result at θ_o yields the condition

$$\Gamma \equiv \mathbf{E}g(z,\theta_o)\nabla_\theta l(z,\theta_o)' \equiv -\mathbf{E}\nabla_\theta g(z,\theta_o) \equiv G. \tag{15}$$

It will sometimes be convenient to estimate G by

(16) $$\Gamma_n = \frac{1}{n}\sum_{t=1}^{n} g(z_t,\tau_n)\nabla_\theta l(z_t,\tau_n)'.$$

In the maximum likelihood case $g = \nabla_\theta l$, one has $\Omega = \Gamma = \mathbf{E}[\nabla_\theta l(z_t,\theta_o)]'[\nabla_\theta l(z_t,\theta_o)]'$ and by the information equality, $G = -E\,\nabla_{\theta\theta} l(z_t,\theta_o) = \mathbf{E}[\nabla_\theta l(z_t,\theta_o)]'[\nabla_\theta l(z_t,\theta_o)]' = \Omega$, so that the asymptotic covariance matrix of the unconstrained estimator simplifies to Ω^{-1}.

Using (16), one has $\Gamma_n'\Omega_n^{-1} = \left[\sum_{t=1}^{n} \nabla_\theta l(z_t,\tau_n) g(z_t,\tau_n)'\right]\left[\sum_{t=1}^{n} g(z_t,\tau_n) g(z_t,\tau_n)'\right]^{-1}$. But each row of this array can be interpreted as the coefficients obtained from an OLS regression of the corresponding component of $\nabla_\theta l(z_t,\tau_n)$ on $g(z_t,\tau_n)$. Then the right-hand side of the first-order condition for a best GMM estimator, $0 = \Gamma_n'\Omega_n^{-1} g_n(T_n)$, can be usefully interpreted as the projection of $\nabla_\theta l(z_t,\tau_n)$ onto the subspace spanned by $g(z_t,\tau_n)$. This is then the linear combination of $g(z_t,\tau_n)$ that most closely approximates $\nabla_\theta l(z_t,\tau_n)$. The GMM estimator T_n sets this approximate score to zero. One implication of this result is that if $g(z_t,\tau_n) = \nabla_\theta l(z_t,\tau_n)$, then the projection returns this vector and $\Gamma_n'\Omega_n^{-1}$ is the identity matrix. Another implication is that if $g(z_t,\tau_n)$ contains $\nabla_\theta l(z_t,\tau_n)$ plus other moments, then $\Gamma_n'\Omega_n^{-1}$ will be the horizonal concatination of an identity matrix and a matrix of zeros, so that the GMM first-order condition coincides with the condition for MLE, and the added moments are given zero weight. Then, the added moments add no information and cannot improve asymptotic efficiency.

2. THE NULL HYPOTHESIS AND THE CONSTRAINED GMM ESTIMATOR

Suppose there is an r-dimensional null hypothesis on the data generation process,

(17) $$H_o: a(\theta_o) = 0,$$

where $a(\cdot)$ is a r×1 vector of continuously differentiable functions and the r×k matrix $A \equiv \nabla_\theta a(\theta_o)$ has rank r. The null hypothesis may be linear or nonlinear. A particularly simple case is H_o: $\theta = \theta^o$, or $a(\theta) \equiv \theta - \theta^o$, so the parameter vector θ is completely specified under the null. Other examples are $a(\theta_o) = \theta_{1o}$, a linear hypothesis that the first parameter is zero, and $a(\theta_o) = (\theta_{10}/\theta_{20} - \theta_{30}/\theta_{40})$, a non-linear hypothesis that two ratios of parameters are equal. In general there will be k-r parameters to be estimated when one imposes the null.

We will consider alternatives to the null of the form

(18) $$H_1: a(\theta_o) \neq 0,$$

or *asymptotically local* alternatives of the form

(19) $$H_{1n}: a(\theta_o) = \delta n^{-1/2} \neq 0.$$

More precisely, for local alternatives we consider the sequence of problems where $l(z,\theta)$ is the log likelihood of an observation, $\theta_{no} = \theta_o - A(A'A)^{-1}\delta n^{-1/2}$ is the sequence of true parameter values, and $a_n(\theta) = \delta n^{-1/2} + A(\theta-\theta_o)$ is the sequence of (locally linear) constraints. These problems then satisfy $a_n(\theta_{no}) = 0$ and $a_n(\theta_o) = \delta n^{-1/2}$. In econometric analysis, interesting alternatives are often sufficiently "local" in large samples so that asymptotic distributions under local alternatives give good estimates of power.

One can define a *constrained* GMM estimator by optimizing the GMM criterion subject to the null hypothesis:

$$T_{an} = \operatorname{argmin}_{\theta\in\Theta} Q_n(\theta) \quad \text{subject to} \quad a(\theta) = 0. \tag{20}$$

For local alternatives, the constraints become $a_n(\theta) = \delta n^{-1/2} + A(\theta-\theta_o)$. The following result establishes consistency of T_{an} under the null hypothesis or local alternatives:

Lemma 3.5. *Assume conditions* (i)-(vii) *in Theorem 1. Assume that under the null hypothesis the true parameter vector* θ_o *satisfies the constraints* $a(\theta_o) = 0$, *and that in the sequence of local alternative problems the true parameter vectors* $\theta_{no} = \theta_o - A(A'A)^{-1}\delta n^{-1/2}$ *satisfy the sequence of constraints* $a_n(\theta) = \delta n^{-1/2} + A(\theta-\theta_o) = 0$. *Then* $T_{an} \to_p \theta_o$ *and* $n^{1/2}\cdot(T_{an} - \theta_o)$ *is stochastically bounded.*

Proof: Under the null hypothesis, $a(\theta_o) = 0$ implies $n\cdot Q_n(T_{an}) \le n\cdot Q_n(\theta_o)$. From the preliminary step in the proof of Theorem 1, $n\cdot Q_n(\theta_o)$ is stochastically bounded. Then, Theorem 1 establishes that T_{an} is consistent and $n^{1/2}\cdot(T_{an} - \theta_o)$ is stochastically bounded. Under the sequence of local alternatives, $a_n(\theta_{no}) = 0$, implying that

$$\begin{aligned} n\cdot Q_n(T_{an}) \le n\cdot Q_n(\theta_{no}) &= [n^{1/2}\cdot g_n(\theta_{no})]' W_n [n^{1/2}\cdot g_n(\theta_{no})] \\ &= [n^{1/2}\cdot g_n(\theta_o) + G_n A'(AA')^{-1}\delta]' W_n [n^{1/2}\cdot g_n(\theta_o) + G_n A'(AA')^{-1}\delta], \end{aligned}$$

where G_n is evaluated at points between θ_{no} and θ_o. Theorem 1 established that $n^{1/2}\cdot g_n(\theta_o)$ is stochastically bounded. The continuity of $G(\theta)$ established in the proof of Theorem 1 and the compactness of Θ imply that $G_n A'(AA')^{-1}\delta$ is stochastically bounded. Together, these results imply that $n\cdot Q_n(T_{an})$ is stochastically bounded, and hence by Theorem 1 that $T_{an} - \theta_{no} \to_p 0$ and $n^{1/2}\cdot(T_{an} - \theta_{no})$ is stochastically bounded. Then, $n^{1/2}(\theta_{no} - \theta_o) = -A(A'A)^{-1}\delta$ implies $T_{an} \to_p \theta_o$ and $n^{1/2}\cdot(T_{an} - \theta_o)$ stochastically bounded. □

Next consider asymptotic normality of the constrained estimator under the null or local alternatives. Define a Lagrangian for T_{an}: $L_n(\theta,\gamma) = Q_n(\theta) - a(\theta)'\gamma$. In this expression, γ is the $r\times 1$ vector of undetermined Lagrangian multipliers; these will be non-zero when the constraints are binding. The first-order conditions for solution of the constrained optimization problem are

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} n^{1/2}\,\nabla_\theta Q_n(T_{an}) - \nabla_\theta a(T_{an})' n^{1/2}\,\gamma_{an} \\ -n^{1/2}\,a(T_{an}) \end{bmatrix}. \tag{21}$$

The Lagrangian multipliers γ_{an} are random variables. Lemma 5, and when applicable the argument given in the proof of Corollary 1, imply $\nabla_\theta Q_n(T_{an}) \to_p -G'W\mathbf{E}g(z,\theta_o) = 0$. Further, $\nabla_\theta a(T_{an}) \to_p A$, implying $A'\gamma_{an} = -\nabla_\theta Q_n(T_{an}) + o_p \to_p 0$, and since A is of full rank, $\gamma_{an} \to_p 0$.

We next outline the argument for asymptotic normality, which parallels the argument given in Theorem 1 for the unconstrained estimator, and relate the asymptotic distributions of T_n, T_{an}, and γ_{an}. Noting that T_{an} satisfies (8), and then approximating G_n by G and W_n by W, one gets

$$n^{1/2}g_n(T_{an}) = n^{1/2}g_n(\theta_o) - G_n\, n^{1/2}(T_{an} - \theta_o) = \Omega^{1/2}U_n - G\, n^{1/2}(T_{an} - \theta_o) + o_p$$

and $n^{1/2}\nabla_\theta Q_n(T_{an}) = G'W\, n^{1/2}g_n(T_{an}) + o_p$. Under local alternatives (or the null when $\delta = 0$),

$$n^{1/2}a(T_{an}) = n^{1/2}a(\theta_o) + A\, n^{1/2}(T_{an} - \theta_o) + o_p \equiv \delta + A\, n^{1/2}(T_{an} - \theta_o) + o_p.$$

Substituting these in the first-order conditions and letting $C = G'WG$ yields

$$\text{(22)} \qquad \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} G'W\Omega^{1/2}U_n \\ -\delta \end{bmatrix} - \begin{bmatrix} C & A' \\ A & 0 \end{bmatrix} \begin{bmatrix} n^{1/2}(T_{an} - \theta_o) \\ n^{1/2}\gamma_{an} \end{bmatrix} + o_p.$$

As a shorthand, write $C = G'WG$. From the formulas for partitioned inverses,

$$\begin{bmatrix} C & A' \\ A & 0 \end{bmatrix}^{-1} = \begin{bmatrix} C^{-1} - C^{-1}A'(AC^{-1}A')^{-1}AC^{-1} & C^{-1}A'(AC^{-1}A')^{-1} \\ (AC^{-1}A')^{-1}AC^{-1} & -(AC^{-1}A')^{-1} \end{bmatrix},$$

Applying this to (22) yields

$$\text{(23)} \qquad \begin{bmatrix} n^{1/2}(T_{an} - \theta_o) \\ n^{1/2}\gamma_{an} \end{bmatrix} = \begin{bmatrix} -C^{-1}A'(AC^{-1}A')^{-1} \\ (AC^{-1}A')^{-1} \end{bmatrix}\delta + \begin{bmatrix} C^{-1} - C^{-1}A'(AC^{-1}A')^{-1}AC^{-1} \\ (AC^{-1}A')^{-1}AC^{-1} \end{bmatrix} G'W\Omega^{1/2}U_n + o_p.$$

From Corollary 1, $n^{1/2}(T_n - \theta_o) = C^{-1}G'W\Omega^{1/2}U_n + o_p$. Substitute this in (26) to conclude that

$$\text{(24)} \qquad n^{1/2}(T_n - T_{an}) = C^{-1}A'(AC^{-1}A')^{-1}AC^{-1}G'W\Omega^{1/2}U_n + C^{-1}A'(AC^{-1}A')^{-1}\delta + o_p.$$

Note that $An^{1/2}(T_n - T_{an}) = AC^{-1}G'W\Omega^{1/2}U_n + \delta + o_p$, and that $n^{1/2}(T_n - T_{an})$ can be represented as the linear transformation $C^{-1}A'(AC^{-1}A')^{-1}$ of $An^{1/2}(T_n - T_{an})$. We also have

$$\text{(25)} \qquad n^{1/2}a(T_n) = n^{1/2}a(\theta_o) + A\, n^{1/2}(T_n - \theta_o) + o_p = AC^{-1}G'W\Omega^{1/2}U_n + \delta + o_p.$$

The expansion $n^{1/2}g_n(T_{an}) = G'W\Omega^{1/2}U_n - G'WG\, n^{1/2}(T_{an} - \theta_o) + o_p$ combined with (23) implies $n^{1/2}g_n(T_{an}) = (\mathbf{I}_m - GC^{-1}G'W + GC^{-1}A'(AC^{-1}A')^{-1}AC^{-1}G'W)\Omega^{1/2}U_n + GC^{-1}A'(AC^{-1}A')^{-1}\delta + o_p$, and $n^{1/2}\nabla_\theta Q_n(T_{an}) = G'W\, n^{1/2}g_n(T_{an}) = A'(AC^{-1}A')^{-1}AC^{-1}G'W\Omega^{1/2}U_n + A'(AC^{-1}A')^{-1}\delta + o_p$. Then,

$$\text{(26)} \qquad AC^{-1}n^{1/2}\nabla_\theta Q_n(T_{an}) = AC^{-1}G'Wn^{1/2}g_n(T_{an}) + o_p = AC^{-1}G'W\Omega^{1/2}U_n + \delta + o_p.$$

Table 1 summarizes these results. The table shows that the r×1 vectors $An^{1/2}(T_n-T_{an})$, $n^{1/2}a(T_n)$, $(AC^{-1}A')n^{1/2}\gamma_{an}$, and $AC^{-1}n^{1/2}\nabla_\theta Q_n(T_{an})$ all equal $AC^{-1}G'W\Omega^{1/2}U_n + \delta + o_p$. Consequently, they are asymptotically equivalent and asymptotically normal with mean δ and non-singular covariance matrix $A(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}A'$. This table shows that all the statistics can be expressed as linear transformations of $n^{1/2}(T_n-\theta_o)$. This makes it simple to determine the asymptotic distributions of tests that use these statistics.

The asymptotic covariance matrices for the Table 1 statistics follow from their formulas and the result that U_n is asymptotically standard normal, and are given in Table 2. For a best GMM estimator, with $W = \Omega^{-1}$ implying that $H \equiv G'W\Omega WG = G'\Omega^{-1}G = C = B$, the asymptotic covariance matrices simplify considerably. The asymptotic covariances matrices always satisfy

$$\text{acov}(T_n-T_{an}) = \text{acov}(T_n) + \text{acov}(T_{an}) - \text{acov}(T_n, T_{an}) - \text{acov}(T_{an}, T_n),$$

but for a best GMM estimator one has $\text{acov}(T_n, T_{an}) = \text{acov}(T_{an})$, giving the simplification

$$\text{acov}(T_n-T_{an}) = \text{acov}(T_n) - \text{acov}(T_{an}) \tag{27}$$

or *the variance of the difference equals the difference of the variances*. This proposition is familiar in a maximum likelihood context where the variance in the deviation between an efficient estimator and any other estimator equals the difference of the variances. We see here that it also applies to *relatively* efficient GMM estimators that use available moments and constraints optimally.

Table 1. The Statistics and their Relationships

	Statistic	Formula (with $C = G'WG$)	Transformations of Other Statistics
1	$n^{1/2}g_n(\theta_o)$	$\Omega^{1/2}U_n + o_p$	---
2	$n^{1/2}(T_n-\theta_o)$	$C^{-1}G'W\Omega^{1/2}U_n + o_p$	$C^{-1}G'Wn^{1/2}g_n(\theta_o)$
3	$n^{1/2}(T_{an}-\theta_o)$	$-C^{-1}A'(AC^{-1}A')^{-1}\delta + [C^{-1}-C^{-1}A'(AC^{-1}A')^{-1}AC^{-1}]G'W\Omega^{1/2}U_n + o_p$	$n^{1/2}(T_n-\theta_o) - C^{-1}A'(AC^{-1}A')^{-1}\, n^{1/2}a(T_n)$
4	$n^{1/2}(T_n-T_{an})$	$C^{-1}A'(AC^{-1}A')^{-1}\delta + C^{-1}A'(AC^{-1}A')^{-1}AC^{-1}G'W\Omega^{1/2}U_n + o_p$	$C^{-1}A'(AC^{-1}A')^{-1}\, n^{1/2}a(T_n)$
5	$A\, n^{1/2}(T_n-T_{an})$	$\delta + AC^{-1}G'W\Omega^{1/2}U_n + o_p$	$n^{1/2}a(T_n)$
6	$n^{1/2}\gamma_{an}$	$(AC^{-1}A')^{-1}\delta + (AC^{-1}A')^{-1}AC^{-1}G'W\Omega^{1/2}U_n + o_p$	$(AC^{-1}A')^{-1}\, n^{1/2}a(T_n)$
7	$AC^{-1}A'n^{1/2}\gamma_{an}$	$\delta + AC^{-1}G'W\Omega^{1/2}U_n + o_p$	$n^{1/2}a(T_n)$
8	$n^{1/2}a(T_n)$	$\delta + AC^{-1}G'W\Omega^{1/2}U_n + o_p$	$\delta + A\, n^{1/2}(T_n-\theta_o)$
9	$n^{1/2}\nabla_\theta Q_n(T_{an})$	$A'(AC^{-1}A')^{-1}\delta + A'(AC^{-1}A')^{-1}AC^{-1}G'W\Omega^{1/2}U_n + o_p$	$A'(AC^{-1}A')^{-1}\, n^{1/2}a(T_n)$
10	$AC^{-1}n^{1/2}\nabla_\theta Q_n(T_{an})$	$\delta + AC^{-1}G'W\Omega^{1/2}U_n + o_p$	$n^{1/2}a(T_n)$

Table 2. Asymptotic Covariance Matrices

(Note: $B = G'\Omega^{-1}G$, $C = G'WG$, $H = G'W\Omega WG$)

	Statistic	Asymptotic Covariance Matrix	Asymptotic Covariance Matrix if $W = \Omega^{-1}$
1	$n^{1/2}g_n(\theta_o)$	Ω	Ω
2	$n^{1/2}(T_n-\theta_o)$	$C^{-1}HC^{-1}$	B^{-1}
3	$n^{1/2}(T_{an}-\theta_o)$	$[C^{-1}-C^{-1}A'(AC^{-1}A')^{-1}AC^{-1}]H[C^{-1}-C^{-1}A'(AC^{-1}A')^{-1}AC^{-1}]$	$B^{-1} - B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}$
4	$n^{1/2}(T_n-T_{an})$	$C^{-1}A'(AC^{-1}A')^{-1}AC^{-1}HC^{-1}A'(AC^{-1}A')^{-1}AC^{-1}$	$B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}$
5	$A\, n^{1/2}(T_n-T_{an})$	$AC^{-1}HC^{-1}A'$	$AB^{-1}A'$
6	$n^{1/2}\gamma_{an}$	$(AC^{-1}A')^{-1}AC^{-1}HC^{-1}A'(AC^{-1}A')^{-1}$	$(AB^{-1}A')^{-1}$
7	$AC^{-1}A'n^{1/2}\gamma_{an}$	$AC^{-1}HC^{-1}A'$	$AB^{-1}A'$
8	$n^{1/2}a(T_n)$	$AC^{-1}HC^{-1}A'$	$AB^{-1}A'$
9	$n^{1/2}\nabla_\theta Q_n(T_{an})$	$A'(AC^{-1}A')^{-1}AC^{-1}HC^{-1}A'(AC^{-1}A')^{-1}A$	$A'(AB^{-1}A')^{-1}A$
10	$AC^{-1}n^{1/2}\nabla_\theta Q_n(T_{an})$	$AC^{-1}HC^{-1}A'$	$AB^{-1}A'$

3. THE TEST STATISTICS

The test statistics for the null hypothesis fall into three major classes, sometimes called the *trinity*. *Wald statistics* are based on deviations of the unconstrained estimates from values consistent with the null. *Lagrange Multiplier* (LM) or *Score statistics* are based on deviations of the constrained estimates from values solving the unconstrained problem. *Distance metric statistics* for best GMM estimators are based on differences in the GMM criterion between the unconstrained and constrained estimators. In the case of maximum likelihood estimation, the distance metric statistic is asymptotically equivalent to the *likelihood ratio statistic*. There are several variants for Wald statistics in the case of the general non-linear hypothesis; these reduce to the same expression in the simple case where the parameter vector is completely determined under the null. The same is true for LM statistics. There are often significant computational advantages to using one member or variant of the trinity rather than another. On the other hand, the Wald and LM statistics are all *asymptotically equivalent*, and for best GMM estimators the distance metric statistic is also asymptotically equivalent Thus, at least to first-order asymptotic approximation, there is no statistical reason to choose between them. This pattern of first-order asymptotic equivalence for GMM estimates is exactly the same as for maximum likelihood estimates.

Table 3 gives the test statistics that can be used for the hypothesis $a(\theta_o) = 0$. For best GMM estimators with $W = \Omega^{-1}$, the full trinity of tests are available. Some of the test statistics that are available for best GMM estimators do not have versions that are asymptotically equivalent for general GMM estimators, and the corresponding cells are omitted from the table. In Section 6, we consider important special cases, including maximum likelihood and nonlinear least squares. In particular, in these special cases, or when the hypothesis is that a subset of the parameters are constants, there are some simplifications of the test statistics, and some versions are indistinguishable.

The central result is that all of the test statistics in each column are asymptotically equivalent under the null hypothesis or a local alternative to the null. Under the null, they have a common limiting chi-square distribution with degrees of freedom r equal to the dimension of the null hypothesis. Under a local alternative, they have a common limiting non-central chi-square distribution with r degrees of freedom and non-centrality parameter $\delta'[AC^{-1}HC^{-1}A']^{-1}\delta$ in the general case and $\delta'(AB^{-1}A')^{-1}\delta$ in the best estimator case. It is useful to relate the expression for the non-centrality parameter to outputs from econometric estimation packages. Typically, a package that does GMM estimation, or one of its specializations such as maximum likelihood or non-linear least squares, will automatically estimate Ω_n^{-1} and use it as the distance metric, and will supply an estimate V of the covariance matrix of the estimates; namely $V = (G_n'\Omega_n^{-1}G_n)^{-1}/n$, where G_n and Ω_n are estimates of G and Ω respectively. If the alternative to the null is H_1: $a(\theta_o) = c$, then $\delta = cn^{1/2}$, and the non-centrality parameter written in terms of V and c is $\delta'(AB^{-1}A')^{-1}\delta = c'(AVA')^{-1}c$. These result will be stated formally and proved following some general observations on the various test statistics.

Table 3. Test Statistics for GMM Estimators

(Note: $B = G'\Omega^{-1}G$, $C = G'WG$, $H = G'W\Omega WG$)

	General Estimators with $W \neq \Omega^{-1}$	**Best Estimators with $W = \Omega^{-1}$**
Wald Statistics		
W_{1n}	$n a(T_n)'[AC^{-1}HC^{-1}A']^{-1} a(T_n)$	$n a(T_n)'[AB^{-1}A']^{-1} a(T_n)$
W_{2n}, flavor 1	$n(T_n - T_{an})'\text{acov}(T_n - T_{An})^{-}(T_n - T_{an})$	$n(T_n - T_{an})'\{\text{acov}(T_n) - \text{acov}(T_{An})\}^{-}(T_n - T_{an})$
W_{2n}, flavor 2	$n(T_n - T_{an})'A'[AC^{-1}HC^{-1}A']^{-1}A(T_n - T_{an})$	$n(T_n - T_{an})'A'(AB^{-1}A')^{-1}A(T_n - T_{an})$
W_{3n}	– – \–	$n(T_n - T_{an})'B(T_n - T_{an})$
Lagrange Multiplier Statistics		
LM_{1n}	$n\gamma_{an}'AC^{-1}A'[AC^{-1}HC^{-1}A']^{-1}AC^{-1}A'\,\gamma_{an}$	$n\gamma_{an}'AB^{-1}A'\gamma_{an}$
LM_{2n}, flavor 1	$n\nabla_\theta Q_n(T_{an})'[A'(AC^{-1}A')^{-1}AC^{-1}HC^{-1}A'(AC^{-1}A')^{-1}A]^{-}\nabla_\theta Q_n(T_{an})$	$n\nabla_\theta Q_n(T_{an})'\{A'(AB^{-1}A')^{-1}A'\}^{-}\nabla_\theta Q_n(T_{an})$
LM_{2n}, flavor 2	$n\nabla_\theta Q_n(T_{an})'A'[AC^{-1}HC^{-1}A']^{-1}A\nabla_\theta Q_n(T_{an})$	$n\nabla_\theta Q_n(T_{an})'B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}\nabla_\theta Q_n(T_{an})$
LM_{3n}	– – –	$n\nabla_\theta Q_n(T_{an})'B^{-1}\nabla_\theta Q_n(T_{an})$
Distance Metric Statistic		
DM_n	– – –	$2n[Q_n(T_{an}) - Q_n(T_n)]$
Asymptotic Distribution Under the Null:	$\chi^2(r)$	$\chi^2(r)$
Asymptotic Distribution Under Local Alternatives	$\chi^2(r,nc)$	$\chi^2(r,nc)$
Non-centrality Parameter (nc)	$\delta'(AC^{-1}HC^{-1}A')^{-1}\delta$	$\delta'(AB^{-1}A')^{-1}\delta$

FIGURE 1. GMM TESTS

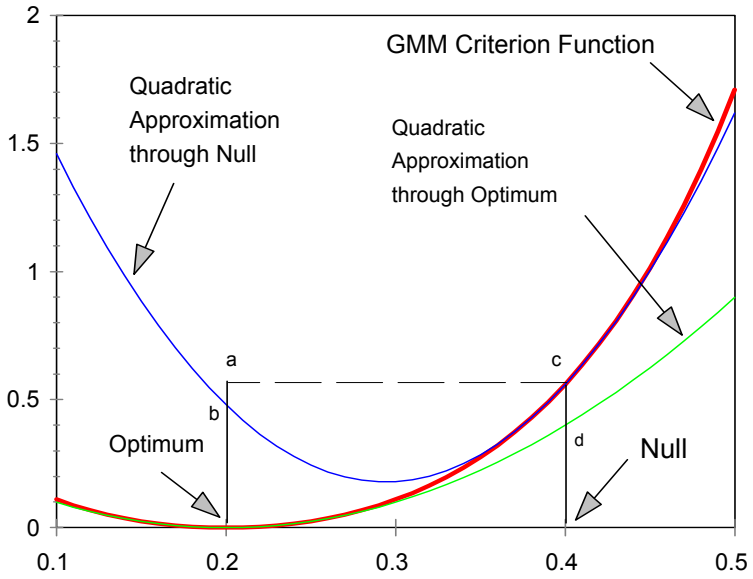

Figure 1 illustrates the relationship between distance metric (DM), Wald (W), and Score (LM) tests for a best GMM estimator. In the case of maximum likelihood estimation, this figure is inverted, the criterion is log likelihood rather than the distance metric, and the DM test is replaced by the likelihood ratio test. The "Optimum" and "Null" points on the θ axis give the unconstrained (T_n) and constrained (T_{an}) estimators, respectively. The GMM criterion function is plotted, along with quadratic approximations to this function through the respective arguments T_n and T_{an}. The Wald statistic (W) can be interpreted as twice the difference in the height at T_n and T_{an} of the quadratic approximation through the optimum; the height d in the figure. The Lagrange Multiplier (LM) statistic can be interpreted as twice the difference in the height at T_n and T_{an} of the quadratic approximation through the null; the difference a - b in the figure. The Distance Metric (DM) statistic is twice the difference in the height at T_n and T_{an} of the GMM criterion, the height c in the figure. Note that if the criterion function were exactly quadratic, then the three statistics would be identical.

The Wald statistic W_{1n} asks how close are the unconstrained estimators to satisfying the constraints; i.e., how close to zero is $a(T_n)$? This variety of the test is particularly useful when the unconstrained estimator is available and the matrix A is easy to compute. For example, when the null is that a subvector of parameters equal constants, then A is a selection matrix that picks out the corresponding rows and columns of $\text{acov}(T_n) = C^{-1}HC^{-1}$ (which reduces to B^{-1} for a best estimator), and this test reduces to a quadratic form with the deviations of the estimators from their hypothesized values in the wings, and the inverse of their asymptotic covariance matrix in the center. In the special case H_o: $\theta = \theta^o$, one has $A = \mathbf{I}_k$.

The Wald test W_{2n} is useful if both the unconstrained and constrained estimators are available. For best GMM estimation, its first version requires only the readily available asymptotic covariance matrices of the two estimators, but for $r < k$ requires calculation of a generalized inverse. Algorithms for this are available, but are often not as numerically stable as classical inversion

algorithms because near zero and exact zero characteristic roots are treated very differently. The second version of W_{2n}, available for either general or best GMM estimators, involves only ordinary inverses, and is potentially quite useful for computation in applications.

The Wald statistic W_{3n}, which is only available for best GMM estimators, treats the constrained estimators *as if they were constants with a zero asymptotic covariance matrix*. This statistic is particularly simple to compute when the unconstrained and constrained estimators are available, as no matrix differences or generalized inverses are involved, and the matrix A need not be computed. The statistic W_{2n} is at least as large as W_{3n} in finite samples, since the center of the second quadratic form is $\text{acov}(T_n)^{-1}$ and the center of the first quadratic form is $\{\text{acov}(T_n) - \text{acov}(T_{an})\}^{-}$, while the tails are the same. Nevertheless, the two statistics are asymptotically equivalent.

The approach of Lagrange multiplier or score tests is to calculate the constrained estimator T_{an}, and then to base a statistic on the discrepancy from zero at this argument of a condition that would be zero if the constraint were not binding. The statistic LM_{1n} asks how close the Lagrangian multipliers γ_{an}, measuring the degree to which the hypothesized constraints are binding, are to zero. This statistic is easy to compute if the constrained estimation problem is actually solved by Lagrangian methods, and the multipliers are obtained as part of the calculation. The statistic LM_{2n} asks how close to zero is the gradient of the distance criterion, evaluated at the constrained estimator. This statistic is useful when the constrained estimator is available and it is easy to compute the gradient of the distance criterion, say using the algorithm to seek minimum distance estimates. The second version of LM_{2n} avoids computation of a generalized inverse.

The statistic LM_{3n}, available for best GMM estimators, bears the same relationship to LM_{2n} that W_{3n} bears to W_{2n}. This flavor of the test statistic is particularly convenient to calculate when the gradient of the likelihood function is available, as it can be obtained by two auxiliary regressions starting from the constrained estimator T_{an}:

a. *Regress* $\nabla_\theta l(z_t,T_{an})'$ *on* $g(z_t,T_{an})$, *and retrieve fitted values* $\nabla_\theta l^*(z_t,T_{an})'$.

b. *Regress* 1 *on* $\nabla_\theta l^*(z_t,T_{an})$, *and retrieve fitted values* $\hat{y}_t$. *Then* $LM_{3n} = \frac{1}{n}\sum_{t=1}^{n} \hat{y}_t^2$.

For MLE, $g = \nabla_\theta l$ and the first regression is redundant, so that this procedure reduces to OLS.

Another form of the auxiliary regression for computing LM_{3n} is available in the case of non-linear instrumental variable regression. Consider the model $y_t = h(x_t,\theta_o) + \varepsilon_t$ with $\mathbf{E}(\varepsilon_t|w_t) = 0$ and $\mathbf{E}(\varepsilon_t^2|w_t) = \sigma^2$, where w_t is a vector of instruments. Define $z_t = (y_t,x_t,w_t)$ and $g(z_t,\theta) = w_t[y_t-h(x_t,\theta)]$. Then $Eg(z,\theta_o) = 0$ and $\mathbf{E}g(z,\theta_o)g(z,\theta_o)' = \sigma^2\mathbf{E}ww'$. The GMM criterion $Q_n(\theta)$ for this model is

$$(28) \qquad \left(\frac{1}{n}\sum_{t=1}^{n} w_t(y_t - h(x_t,\theta))'\right)\left(\frac{1}{n}\sum_{t=1}^{n} w_t w_t'\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} w_t(y_t-h(x_t,\theta))\right)/2\sigma^2.$$

Optimization is not affected by the scalar σ^2. Consider the hypothesis $a(\theta_o) = 0$, and let T_{an} be the constrained GMM estimator. One can compute LM_{3n} by the following method:

a. Regress $\nabla_\theta h(x_t, T_{an})$ on w_t, and retrieve the fitted values $\nabla_\theta \hat{h}_t$.
b. Regress the residual $u_t = y_t - h(x_t, T_{an})$ on $\nabla_\theta \hat{h}_t$, and retrieve the fitted values $\hat{u}_t$.

Then $LM_{3n} = n \sum_{t=1}^{n} \hat{u}_t^2 / \sum_{t=1}^{n} u_t^2 \equiv nR^2$, with R^2 the *uncentered* multiple correlation coefficient.

Note that this is not in general the same as the standard R^2 produced by OLS programs, since the denominator of that definition is the sum of squared deviations of the dependent variable about its mean. When the dependent variable has mean zero, the centered and uncentered definitions coincide.

The approach of the distance metric test is based on the difference between the values of the distance metric at the constrained and unconstrained estimates. It has a limiting chi-square distribution and is asymptotically equivalent to the other members of the trinity only for best GMM estimators. This estimator is particularly convenient when both the unconstrained and constrained estimators can be computed, and the estimation algorithm returns the goodness-of-fit statistics. In the case of linear or non-linear least squares, this is the familiar test statistic based on the sum of squared residuals from the constrained and unconstrained regressions.

The statistical properties of the trinity are summarized in the following theorem:

Theorem 3.2. *Assume the regularity conditions (i)-(vii). For general GMM estimation with* $W \neq \Omega^{-1}$*, the statistics in the middle column of Table 3 are asymptotically equivalent, and are asymptotically distributed central chi-square with r degrees of freedom under the null hypothesis, and non-central chi-square with r degrees of freedom and a non-centrality parameter* $\delta'(AC^{-1}HC^{-1}A')^{-1}\delta$ *under local alternatives. For best GMM estimation with* $W = \Omega^{-1}$*, the statistics in the last column of Table 3 are asymptotically equivalent, and are asymptotically distributed chi-square with r degrees of freedom under the null hypothesis, and non-central chi-square with r degrees of freedom and a non-centrality parameter* $\delta'(AB^{-1}A')^{-1}\delta$ *under local alternatives.*

Proof: Define $V_n = (AC^{-1}HC^{-1}A')^{-1/2} n^{1/2} a(T_n) = (AC^{-1}HC^{-1}A')^{-1/2}\{\delta + AC^{-1}G'W\Omega^{1/2}U_n\} + o_p$. This vector has mean $(AC^{-1}HC^{-1}A')^{-1/2}\delta$ and covariance matrix $\mathbf{I}_r$. But the sum of squares of a normal random vector with an identity covariance matrix is non-central chi-square with degrees of freedom equal to its dimension and non-centrality parameter equal to the sum of squares of its mean. This implies that $V_n'V_n = n^{1/2}a(T_n)'(AC^{-1}HC^{-1}A')^{-1}n^{1/2}a(T_n)$ has this asymptotic distribution with degrees of freedom r and non-centrality parameter $\delta'(AC^{-1}HC^{-1}A')^{-1}\delta$. This establishes the asymptotic distribution of W_{1n}. From Table 1, the statistics $An^{1/2}(T_n - T_{an})$, $(AC^{-1}A')n^{1/2}\gamma_{an}$, and $AC^{-1}n^{1/2}\nabla_\theta Q_n(T_{an})$ all equal $n^{1/2}a(T_n)$ up to order o_p. Hence, quadratic forms in these statistics, with the center $(AC^{-1}HC^{-1}A')^{-1}$, will all be asymptotically equivalent to $V_n'V_n$. This establishes the asymptotic equivalence of W_{1n}, W_{2n}, LM_{1n}, and LM_{2n}. These results for W_{2n} and LM_{2n} establish the Moore-Penrose generalized inverse formulas $\text{acov}(T_n - T_{an})^- = A'[AC^{-1}HC^{-1}A']^{-1}A$ for general GMM estimators and $\{\text{acov}(T_n) - \text{acov}(T_{an})\}^- = A'(AB^{-1}A')^{-1}A$ for best GMM estimators, and show that the alternative flavors of W_{2n} and LM_{2n} are asymptotically equivalent. This equivaence could also have been established by application of Lemma 4 in the appendix to this chapter.

The asymptotic equivalence of W_{2n} and W_{3n} for best GMM estimators is established from the formula $n^{1/2}(T_n - T_{an}) = B^{-1}A'(AB^{-1}A')^{-1}\delta + B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}U_n + o_p$. Premultiplying by

$(AB^{-1}A')^{-1/2}A$ gives $V_n = (AB^{-1}A')^{-1/2}\{\delta + AB^{-1}G'\Omega^{-1/2}U_n\} + o_p$, and $W_{2n} = V_n'V_n$. Premultiplying by $B^{1/2}$ gives $V_n^* = B^{-1/2}A'(AB^{-1}A')^{-1}\{\delta + AB^{-1}G'\Omega^{-1/2}U_n\} + o_p = B^{-1/2}A'(AB^{-1}A')^{-1/2}V_n. + o_p.$ and $W_{3n} = V_n^{*\prime}V_n^* = V_n'V_n + o_p = W_{2n} + o_p$. This result could also have been obtained using Appendix Lemma 4 by observing that the asymptotic covariance matrix $B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}$ of $n^{1/2}(T_n - T_{an})$ is $A'(AB^{-1}A')^{-1}A$, and that B also satisfies the Appendix condition (i) for a generalized inverse. A similar argument establishes the asymptotic equivalence of LM_{2n} and LM_{3n}: premultiply the expression $n^{1/2}\nabla_\theta Q_n(T_{an}) = A'(AB^{-1}A')^{-1}\delta + A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}U_n + o_p$ by, respectively, $(AB^{-1}A')^{-1/2}AB^{-1}$ and $B^{-1/2}$, and observe that the inner products of two vectors that result are to order o_p equal to LM_{2n} and LM_{3n} and equal to each other.

Make a Taylor's expansion of $n^{1/2}g_n(T_{an})$ about T_n: $n^{1/2}g_n(T_{an}) = n^{1/2}g_n(T_n) + G_n\, n^{1/2}(T_{an} - T_n) + o_p$. Substitute this in the expression for DM_n and use the fact that $G_n'W_n\, n^{1/2}g_n(T_n) = 0$ to obtain

$$
\begin{aligned}
(29)\qquad DM_n = 2n\{Q_n(T_{an}) - Q_n(T_n)\} &= n^{1/2}g_n(T_n)'G_n'W_n\, n^{1/2}g_n(T_n) \\
&\quad + 2\, n^{1/2}(T_{an} - T_n)'G_n'W_n\, n^{1/2}g_n(T_n) + n^{1/2}(T_{an} - T_n)'G_n'W_nG_n\, n^{1/2}(T_{an} - T_n) + o_p \\
&= n(T_{an}-T_n)'G'WG(T_{an}-T_n) + o_p.
\end{aligned}
$$

Then, for best GMM estimators, $G'WG = B$ and $DM_n = W_{3n} + o_p$.

For general GMM estimators with $W \neq \Omega^{-1}$, the quadratic form $n(T_n-T_{an})'acov(T_n)^{-1}(T_n-T_{an})$ that would define W_{3n}, and the quadratic form $n\nabla_\theta Q_n(T_{an})acov(T_n)\nabla_\theta Q_n(T_{an})$ that would define LM_{3n} fail to have representations as inner products of asymptotically normal vectors with idempotent covariance matrices, and hence fail to have limiting chi-square distributions. From (29), DM_n is asymptotically equivalent to $n(T_{an}-T_n)'C(T_{an}-T_n)$, which also fails to have a representation as the inner product of a vector with an idempotent covariance matrix. This shows that the statistics W_{3n}, LM_{3n}, and DM_n are not available for general GMM estimators where $W \neq \Omega^{-1}$. □

4. TWO-STAGE GMM ESTIMATION

A common econometric problem is to do estimation when some parameters have already been estimated from a previous stage, often on the same data. One common case is where the problem contains constructed variables whose construction depended on parameters estimated in a previous round. In general, the use of consistent estimates from a previous round will not cause a problem with consistency in later stages, but it will add noise to the problem that appears in the asymptotic covariance matrix of the later-stage estimators.

There are a few cases, such as feasible GLS with normal disturbances, where no correction of the asymptotic covariance matrix is needed. This is due in the GLS case to a block diagonality in the information matrix between regression coefficients and parameters in the covariance matrix. There is a simple rule, due to Whitney Newey, for determining whether previous stage estimation will add something to the asymptotic covariance matrix in the current stage: *There will be a contribution if and only if consistency in the first stage is necessary for consistency in the second stage.*

When a correction is required, the following generic GMM framework can be used to establish the form of this correction. Suppose one observes variables (x,y,z), where x is exogenous, and (y,z) are variables whose behavior is being modeled. Let $f(y,z|x,\alpha,\beta)$ be the joint density of the observations, conditioned on x, with parameter vectors α and β. Assume that it can be written

(30) $$f(y,z|x,\alpha,\beta) = f^c(z|x,y,\alpha)f^m(y|x,\alpha,\beta)$$

or

(31) $$f(y,z|x,\alpha,\beta) = f^c(z|x,y,\alpha,\beta)f^m(y|x,\alpha).$$

This is the standard decomposition of a joint density into a conditional density times a marginal density, and the only restriction we are imposing is that we can parameterize (or reparameterize) the problem so that either the conditional density or the marginal density does not depend on the parameter β. This corresponds to the usual situation in two-stage methods, where at the first stage one looks at limited information that involves a subset of the full parameter vector.

One concrete example of this setup is sequential estimation of the parameters in a two-level nested logit model, in which f^c is the likelihood of choice at the lower level, conditioned on choice of a upper level branch, and f^m is the likelihood of choice among the upper level branches. In this application, the model can be parameterized so that upper branch parameters do not appear in f^c. A second concrete example is two-step estimation of the Tobit model, in which y is an indicator for whether the response is zero or positive, z is the quantitative level of the response, f^c is the likelihood of the quantitative response conditioned on whether it is zero or not, and f^m is the likelihood of the indicator. In this example, the problem can be parameterized so that parameters that enter the quantitative response likelihood do not enter the likelihood for the indicator.

Suppose in the first stage one estimates the parameter vector α using moments

(32) $$0 = \mathbf{E}_n h(a_n;x,y,z),$$

where $\mathbf{E}_n$ denotes empirical expectation (or sample average). If there are over-identifying moments, assume that they are already weighted by the GMM criterion so that the dimension of h is the dimension of α. A necessary condition for consistency is $\mathbf{E}h(\alpha;x,y,z) = 0$ if and only if $\alpha = \alpha_o$. An important case is limited information maximum likelihood: $h(\alpha;x,y,z) = \nabla_\alpha l^c(z|x,y,\alpha)$, where $l^c = \log f^c$; or $h(\alpha;x,y,z) = \nabla_\alpha l^m(y|x,\alpha)$, where $l^m = \log f^m$.

Suppose in the second stage one estimates a parameter vector β using moments

(33) $$0 = \mathbf{E}_n g(b_n,a_n;x,y,z),$$

where a_n is inserted from the previous stage. Assume that g is defined, by GMM weighting if necessary, so that its dimension equals the dimension of β. Again, important cases are maximum likelihood: $g(\beta,\alpha;x,y,z) = \nabla_\beta l^m(y|x,\alpha,\beta)$ or $g(\beta,\alpha;x,y,z) = \nabla_\beta l^c(y|z,x,\alpha,\beta)$, with α treated as if it were known. In the first of these cases, the moments g do not depend on z. Whether or not g depends on z turns out to make a substantial difference in the final covariance formula. The case of constructed variables is handled by writing them as functions of the parameters α that enter their construction.

The original parameters of the problem may be estimated, perhaps in combination with other parameters, in both the first and second stages. The classification into α and β may require reparameterization. The following rules may help: If first-stage estimates of original parameters are used solely as starting values for second-stage estimation of the same parameters, then classify these as β parameters, as these first-stage estimates are only a computational device and have no influence on the final solution of the second-stage moments. If first stage estimates of original parameters are used for other purposes, such as construction of estimated variables, and are then reestimated in the second stage, then they should appear in both α and β as *separate* parameters. Of course, original parameters estimated only at the first stage go into α, and original parameters estimated only at the second stage go into β.

Make a Taylor's expansion of both the first-stage and the second-stage moment conditions around the true β_o and α_o, and suppress the x,y,z arguments to simplify notation:

$$(34) \qquad \begin{bmatrix} 0 \\ 0 \end{bmatrix} = n^{1/2} \begin{bmatrix} \boldsymbol{E_n} h(\alpha_o) \\ \boldsymbol{E_n} g(\beta_o,\alpha_o) \end{bmatrix} - \begin{bmatrix} A \\ B \end{bmatrix} n^{1/2}(a_n - \alpha_o) - \begin{bmatrix} 0 \\ C \end{bmatrix} n^{1/2}(b_n - \beta_o) + o_p,$$

where A = -plim $\mathbf{E}_n \nabla_\alpha h(\alpha_o)$, B = -plim $\mathbf{E}_n \nabla_\alpha g(\beta_o,\alpha_o)$, and C = -plim $\mathbf{E}_n \nabla_\beta g(\beta_o,\alpha_o)$.

The term $n^{1/2} \begin{bmatrix} \boldsymbol{E_n} h(\alpha_o) \\ \boldsymbol{E_n} g(\beta_o,\alpha_o) \end{bmatrix}$ is asymptotically normal, by a central limit theorem, with a covariance matrix $\begin{bmatrix} \Omega_{hh} & \Omega_{hg} \\ \Omega_{gh} & \Omega_{»} \end{bmatrix}$. Solve the first block of equations and substitute them into the second block to obtain

$$(35) \qquad 0 = n^{1/2}\{\mathbf{E}_n g(\beta_o,\alpha_o) + BA^{-1}\mathbf{E}_n h(\alpha_o)\} - Cn^{1/2}(b_n - \beta_o) + o_p.$$

The term in braces on the right-hand-side of this expression has an asymptotic covariance matrix

$$(36) \qquad \Omega_{gg} - BA^{-1}\Omega_{hg} - \Omega_{gh}A'^{-1}B' + BA^{-1}\Omega_{hh}A'^{-1}B'.$$

Then, solving for $n^{1/2}(b_n - \beta_o)$, one obtains the result that its asymptotic covariance matrix is

$$(37) \qquad C^{-1}\{\Omega_{gg} - BA^{-1}\Omega_{hg} - \Omega_{gh}A'^{-1}B' + BA^{-1}\Omega_{hh}A'^{-1}B'\}C'^{-1}$$

All the terms of this covariance matrix could be estimated from sample analogs, computed at the consistent estimates. The following table summarizes consistent estimators for the various covariance terms; recall that $\mathbf{E}_n$ denotes empirical expectation (sample average):

Matrix	Estimator
C	$-\mathbf{E}_n\nabla_\beta g(b_n,a_n)$
B	$-\mathbf{E}_n\nabla_\alpha g(b_n,a_n)$
A	$-\mathbf{E}_n\nabla_\alpha h(a_n)$
Ω_{hh}	$\mathbf{E}_n h(a_n)h(a_n)'$
Ω_{gh}	$\mathbf{E}_n g(b_n,a_n)h(a_n)'$
Ω_{gg}	$\mathbf{E}_n g(b_n,a_n)g(b_n,a_n)'$

The terms Ω_{gh} and Ω_{hh} add to the asymptotic covariance matrix, relative to the case of α_o known. If B = 0, there is no correction; this is the "block diagonality" case where β can be estimated consistently even if the estimator of α is not consistent. If α is estimated from an *independent* data set, then $\Omega_{gh} = 0$, but one will still need a correction due to the contribution from Ω_{hh}. Also, if g does not depend on z, then $\Omega_{gh} = \mathbf{E}_{y|x}\{g\cdot\mathbf{E}_{z|x,y}h\} = 0$. This is true, in particular, in the case that the second stage estimator is marginal maximum likelihood in which z does not appear and α is treated as given.

The identities $0 \equiv \int\int h\exp(l)dzdy$ and $0 \equiv \int\int g\exp(l)dzdy$ can be differentiated to obtain the conditions

$$(38) \qquad A \equiv -\mathbf{E}\nabla_\alpha h = \mathbf{E}h\cdot\nabla_\alpha l\ ,\ B \equiv -\mathbf{E}\nabla_\alpha g = \mathbf{E}g\cdot\nabla_\alpha l\ ,\ C \equiv -\mathbf{E}\nabla_\beta g = \mathbf{E}g\cdot\nabla_\beta l\ .$$

If g does not depend on z, then $\mathbf{E}g\cdot\nabla_\alpha l^c = \mathbf{E}_{y|x}(g\cdot\mathbf{E}_{z|y,x}\nabla_\alpha l^c) = 0$, implying $B = \mathbf{E}g\cdot(\nabla_\alpha l^m)'$. Sample averages of these outer products estimate the corresponding matrices consistently.

Simplification occurs when the first stage is conditional maximum likelihood that does not depend on β, and the second stage is marginal maximum likelihood that treats the first stage parameter estimates as fixed. Then, $A = \mathbf{E}\nabla_\alpha l^c\cdot(\nabla_\alpha l^c)' = \Omega_{hh}$, $B = \mathbf{E}\nabla_\beta l^m\cdot(\nabla_\alpha l^m)'$, $C = \mathbf{E}\nabla_\beta l^m\cdot(\nabla_\beta l^m)' = \Omega_{gg}$, and $\Omega_{hg} = \mathbf{E}\nabla_\alpha l^c(\nabla_\beta l^m)' = 0$, so that the covariance matrix is $C^{-1} + C^{-1}BA^{-1}B'C^{-1}$.

Similarly, when the first stage is marginal maximum likelihood that does not depend on β, and the second stage is conditional maximum likelihood treating α as fixed, one has $A = \mathbf{E}\nabla_\alpha l^m\cdot(\nabla_\alpha l^m)' = \Omega_{hh}$, $B = \mathbf{E}\nabla_\beta l^c\cdot(\nabla_\alpha l^c)'$, $C = \mathbf{E}\nabla_\beta l^c\cdot(\nabla_\beta l^c)' = \Omega_{gg}$, and $\Omega_{hg} = \mathbf{E}\nabla_\alpha l^c(\nabla_\beta l^m)' = 0$, and the covariance matrix $C^{-1} + C^{-1}BA^{-1}B'C^{-1}$.

The terms in these covariance matrix expressions involve sample averages of squares and cross-products of scores (gradients) of first and second stage log likelihoods. These should all be obtainable as intermediate output from a maximum likelihood program, except for terms involving the gradient of the second-stage likelihood with respect to α. The latter would be simple to obtain in a program like TSP, which does automatic analytic differentiation, or could be obtained by numerical differentiation.

Exercise 2: Consider the problem of Heckman two-stage estimation of a Tobit model, $y = x\theta + \sigma\varphi(x\theta/\sigma)/\Phi(x\theta/\sigma) + \zeta$ for $y > 0$, where $\mathbf{E}(\zeta|y > 0\ \&\ x) = 0$, and where the inverse Mills ratio is calculated from a first-stage probit on the same data. Reparameterize $\alpha = \theta/\sigma$ and $\beta = (\theta,\sigma)$. In this case, h in the generic notation is the score of the marginal log likelihood for the probit, which is influenced only by α, and g is the set of OLS orthogonality conditions, which depend on both

α and β through the condition $y = x\theta + \sigma\varphi(x\alpha)/\Phi(x\alpha)$. Work out the corrected asymptotic covariance matrix for θ and σ.

Exercise 3: Consider the two-level nested multinomial logit model, with first stage estimation applied to the lower level of the choice tree, and used to compute summary variables ("inclusive values") that are then treated as variables in the second stage estimation.

5. ONE-STEP THEOREMS

Under standard regularity conditions, GMM estimators are *locally linear*, which means that within a suitable neighborhood of the estimator, the first-order conditions for these estimators are in large samples approximately linear, with higher-order terms being asymptotically negligible. This has an important practical implication: if one can get an initial estimator τ_n that is within the suitable neighborhood, then one can get to the full GMM estimator, or at least an asyptotically equivalent flavor of it, in one linear step. This has the computational advantage that at this stage no iterative computation is required, and the step can usually be carried out by a simple least squares regression. This also has a useful statistical advantage: the asymptotic covariance matrix of the one-step estimator will be the same as that of the GMM estimator, with its attendant efficiency properties, rather than the possibly much more complex covariance matrix of the initial estimator. For example, the initial estimator might be the result of multiple-stage estimation, as described in the previous section, with a covariance matrix of the form given in that section. However, one linear step starting from that estimator gives a result that is asymptotically equivalent to solving the full joint GMM problem. Alternately, one might start from initial GMM estimators, and in one step obtain a result that is asymptotically equivalent to full maximum likelihood estimation. Within the context of hypothesis testing with GMM estimates, it is possible to go in one linear step from any suitable initially consistent estimator to estimators that are asymptotically equivalent to either the unconstrained or constrained GMM estimators.

The first result based on these ideas is estimation of an expectation that depends on estimated parameters. Suppose one wishes to estimate $\mathbf{E}_z m(z,\theta_o)$, where m is a vector of functions of random variables z and a parameter vector θ that has true value θ_o. If τ_n is any consistent estimator of θ_o, the sample average of $m(z_t,\theta)$ converges in probability to $\mathbf{E}_z m(z,\theta)$ *uniformly* in θ, and $\mathbf{E}_z m(z,\theta)$ is continuous in θ, then

$$\frac{1}{n}\sum_{t=1}^{n} m(z_t,\tau_n) \rightarrow_p \mathbf{E}_z m(z,\theta_o). \tag{39}$$

This works because

$$Prob\left(\left| \frac{1}{n}\sum_{t=1}^{n} m(z_t,\tau_n) - E_z m(z,\tau_n) \right| > \varepsilon \right) \le \frac{1}{n}\sum_{t=1}^{n} Prob\left(sup_\theta \left| m(z_t,\theta) - E_z m(z,\theta) \right| > \varepsilon \right) \rightarrow 0 \tag{40}$$

and $\mathbf{E}_z m(z,\tau_n) \rightarrow \mathbf{E}_z m(z,\theta_o)$. Suppose one strengthens the requirement on τ_n to the condition that it be $n^{1/2}$-consistent, meaning that $n^{1/2}(\tau_n - \theta_o)$ is stochastically bounded, or for each $\varepsilon > 0$ there exists $M > 0$ such that

(41) $\text{Prob}(|n^{1/2}(\tau_n - \theta_o)| > M) < \varepsilon$ for all n.

Suppose that $m(z,\theta)$ satisfies a Lipschitz condition at θ_o; i.e., there exists a function $L(z)$ with a finite expectation such that $|m(z,\theta) - m(z,\theta_o)| \leq L(z)\cdot|\theta - \theta_o|$. Then the result holds without requiring uniform convergence in probability for sample averages of $m(z,\theta)$.

The preceding result is useful for calculation of Wald or Lagrange Multiplier test statistics, which require estimation of $G(\theta_o)$, $\Omega(\theta_o)$, and/or $A(\theta_o)$. The arrays $G_n(\theta)$, $\Omega_n(\theta)$, and $A_n(\theta)$ are uniformly convergent, and the result establishes for any initial consistent estimator τ_n that $G_n(\tau_n) \to_p G(\theta_o)$, $\Omega_n(\tau_n) \to_p \Omega(\theta_o)$, and $A_n(\tau_n) \to_p A(\theta_o)$. Then, using these estimates preserves the asymptotic equivalence of the tests under the null and local alternatives. In particular, one can evaluate terms entering the definitions of these arrays at T_n, T_{an}, or any other consistent estimator of θ_o. In sample analogs that converge to these arrays by the law of large numbers, one can freely substitute sample and population terms that leave the probability limits unchanged. For example, if $z_t = (y_t,x_t)$ and τ_n is any consistent estimator of θ_o, then Ω can be estimated by (1) an analytic expression for $\mathbf{E}g(z,\theta)g(z,\theta)'$, evaluated at τ_n, (2) a sample average $\frac{1}{n}\sum_{t=1}^{n} g(z_t,\tau_n)g(z_t,\tau_n)'$, or (3) a sample average of conditional expectations $\frac{1}{n}\sum_{t=1}^{n} E_{y|x}\, g(y,x_t,\theta)g(y,x_t,\theta)'$ evaluated at $\theta = \tau_n$. It should be noted however that these first-order equivalences do *not* hold in finite samples, or even to higher orders of $n^{1/2}$. Thus, there may be clear choices between these when higher orders of approximation are taken into account.

The second result, called the *one-step theorem*, considers the first-order condition associated with a GMM criterion function, $0 = G_n'\Omega_n^{-1} g_n(\theta)$. Suppose one has an initial $n^{1/2}$-consistent estimator τ_n for θ_o. A Taylor's expansion of the first-order condition about τ_n yields

$$G_n'\Omega_n^{-1} g_n(\theta) = G_n'\Omega_n^{-1}g_n(\tau_n) + G_n'\Omega_n^{-1}G_n(\theta - \tau_n) + O((\theta - \tau_n)^2).$$

Then, a one-step approximation to the unconstrained GMM estimator is

(42) $$T_{on} = \tau_n - (G_n'\Omega_n^{-1}G_n)^{-1}G_n'\Omega_n^{-1}g_n(\tau_n).$$

A Taylor's expansion around θ_o of the GMM first-order condition, evaluated at τ_n, yields

$$n^{1/2}G_n'\Omega_n^{-1}g_n(\tau_n) = n^{1/2}G_n'\Omega_n^{-1}g_n(\theta_o) + G_n'\Omega_n^{-1}G_n\cdot n^{1/2}(\tau_n - \theta_o) + o_p.$$

Combine this with the condition $-G_n'\Omega_n^{-1}g_n(\tau_n) = G_n'\Omega_n^{-1}G_n n^{1/2}(T_{on} - \tau_n)$ to conclude that

$$-n^{1/2}G_n'\Omega_n^{-1}g_n(\theta_o) = G_n'\Omega_n^{-1}G_n n^{1/2}(T_{on} - \theta_o) + o_p,$$

and the condition

$$-n^{1/2}G_n'\Omega_n^{-1}g_n(\theta_o) = G_n'\Omega_n^{-1}G_n n^{1/2}(T_n - \theta_o) + o_p$$

to conclude that

$$0 = G_n'\Omega_n^{-1}G_n n^{1/2}(T_{on} - T_n) + o_p, \tag{43}$$

so that T_{on} and T_n are asymptotically equivalent.

The one-step theorem can also be applied to the constrained GMM estimator. Suppose the null hypothesis, or a local alternative, $a(\theta_o) = \delta \cdot n^{-1/2}$, is true. Define one-step constrained estimators from the Lagrangian first-order conditions:

$$\begin{bmatrix} T_{oan} \\ \gamma_{oan} \end{bmatrix} = \begin{bmatrix} \tau_n \\ 0 \end{bmatrix} - \begin{bmatrix} B & A' \\ A & 0 \end{bmatrix}^{-1} \begin{bmatrix} \nabla_\theta Q_n(\tau_n) \\ -a(\tau_n) \end{bmatrix} . \tag{44}$$

Note in this definition that $\gamma = 0$ is a trivial initially consistent estimator of the Lagrangian multipliers under the null or local alternatives, and that the arrays B and A can be estimated at τ_n. The one-step theorem again applies, yielding $n^{-1/2}(T_{oan}-T_{an}) \to_p 0$ and $n^{-1/2}(\gamma_{oan}-\gamma_{an}) \to_p 0$. Then, these one-step equivalents can be substituted in any of the test statistics of the trinity without changing their asymptotic distribution.

A regression procedure for calculating the one-step expressions is often useful for computation. The adjustment from τ_n yielding the one-step unconstrained estimator is obtained by a two-stage least squares regression of the constant one on $\nabla_\theta l(z_t,\tau_n)$, with $g(z_t,\tau_n)$ as instruments; i.e.,

a. Regress each component of $\nabla_\theta l(z_t,\tau_n)$ on $g(z_t,\tau_n)$ in the sample t = 1,...,n, and retrieve fitted values $\nabla_\theta l^*(z_t,\tau_n)$;
b. Regress 1 on $\nabla_\theta l^*(z_t,\tau_n)$; and adjust τ_n by the amounts of the fitted coefficients.

Step (a) yields $\nabla_\theta l^*(z_t,\tau_n)' = g(z_t,\tau_n)\Omega_n^{-1}\Gamma_n$, and step (b) yields coefficients

$$\Delta = \left[\sum_{t=1}^{n} [\nabla_\theta l^*(z_t,\tau_n)][\nabla_\theta l^*(z_t,\tau_n)]'\right]^{-1} \sum_{t=1}^{n} \nabla_\theta l^*(z_t,\tau_n)$$
$$= (\Gamma_n'\Omega_n\Gamma_n)^{-1}\Gamma_n'\Omega_n g_n(\tau_n).$$

This is the adjustment indicated by the one-step theorem.

Computation of one-step constrained estimators is conveniently done using the formulas

$$T_{oan} = T_{on} - B^{-1}A'(AB^{-1}A')^{-1}a(T_{on}) \equiv \tau_n + \Delta - B^{-1}A'(AB^{-1}A')^{-1}[a(\tau_n) + A\Delta] \tag{45}$$
$$\gamma_{oan} = -(AB^{-1}A')^{-1}a(T_{on}) \equiv -(AB^{-1}A')^{-1}[a(\tau_n) + A\Delta]$$

with A and B evaluated at τ_n. To derive these formulas from the first-order conditions for the Lagrangian problem, replace $\nabla_\theta Q_n(\tau_n)$ by the expression $-(\Gamma_n'\Omega_n^{-1}\Gamma_n')(T_{on} - \tau_n)$ from the one-step definition of the unconstrained estimator, replace $a(\tau_n)$ by $a(T_{on}) + A(T_{on} - \tau_n)$, and use the formula for a partitioned inverse.

6. SPECIAL CASES

Extremum Estimators. Consider data z with a log likelihood function $l(z,\theta_o)$, where θ_o is the true value of θ in the population. Suppose $f(z,\theta)$ is a scalar function whose expectation is minimized at θ_o; i.e., $\mathbf{E}f(z,\theta) \geq \mathbf{E}f(z,\theta_o)$, with equality if and only if $\theta = \theta_o$. For a random sample z_i, i = 1,...,n, consider the extremum estimator

$$(46) \qquad T_n = \operatorname{argmin}_\theta f_n(\theta) \quad \text{where} \quad f_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} f(z_i,\theta).$$

For the example $f(z,\theta) = -l(z,\theta)$, the negative of the log likelihood function, the extremum estimator is the maximum likelihood estimator. Another example that is common in econometrics is the non-linear least squares criterion with z = (x,y) and $f(z,\theta) = (y_i - h(x_i,\theta))^2/2$, yielding the non-linear least squares (NLLS) estimator.

Suppose that the function $f(z,\theta)$ is three times continuously differentiable in θ on an open neighborhood of θ_o, almost surely in z. Then, the population condition $\mathbf{E}f(z,\theta) \geq \mathbf{E}f(z,\theta_o)$ implies the moment condition $\mathbf{E}\,\nabla_\theta f(z,\theta_o) = 0$, and the extremum estimator T_n satisfies the first-order condition $0 = \nabla_\theta f_n(T_n)$. Differentiating the identity $\int_z [\nabla_\theta f(z,\theta)]\cdot e^{l(z,\theta)}dz \equiv 0$ yields the equality

$$\mathbf{(47)} \qquad \mathbf{E}[\nabla_\theta f(z,\theta)][\nabla_\theta l(z,\theta)]' + \mathbf{E}\,\nabla_{\theta\theta} f(z,\theta) \equiv 0,$$

called the *generalized information equality*. In the maximum likelihood case $f(z,\theta) = -l(z,\theta)$, this implies $\mathbf{E}[\nabla_\theta f(z,\theta)][\nabla_\theta f(z,\theta)]' \equiv \mathbf{E}\nabla_{\theta\theta} f(z,\theta)$. However, the last equality is not true in general for extremum criteria, only for those that produce estimators that are asymptotically efficient (i.e., asymptotically equivalent to maximum likelihood). Newey and McFadden (1994, Sect. 5.3) use this observation to develop a general criterion for asymptotic efficiency of estimators.

The population moment condition can be used to define a GMM criterion,

$$(48) \qquad Q_n(\theta) = [\nabla_\theta f_n(\theta)]'\cdot G_n(\theta)^{-1}\cdot[\nabla_\theta f_n(\theta)],$$

where

$$(49) \qquad G_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta\theta} f(z_i,\theta) \to_p G(\theta) \equiv \mathbf{E}\nabla_{\theta\theta} f(z.\theta).$$

The second-order condition for a locally unique extremum estimator is that $G(\theta_o)$ is positive semi-definite, and definite at points in each neighborhood of θ_o. Rule out pathological cases by making the technical assumption that $G(\theta_o)$ is positive definite. Then $G_n(\theta)$, evaluated at a preliminary estimator τ_n that converges in probability to θ_o, is eventually positive definite, so that it defines a legal distance metric. The extremum estimator T_n satisfies $Q_n(T_n) = 0$, so that it is also a GMM estimator. Obviously, this result does not depend on the choice of the distance metric $G_n(\theta)^{-1}$, or on whether $G_n(\theta)$ is treated as a constant array or as a function of θ in the process of optimization. However, for estimation of θ_o subject to constraints and the development of test statistics, the GMM criterion based on a consistent approximation to the distance metric $G(\theta_o)^{-1}$ is needed.

Because the unconstrained extremum estimator can be interpreted as an unconstrained GMM estimator, its large sample statistical properties can be stated as a corollary of the statistical theory of unconstrained GMM estimators. In the following paragraphs, we show how these results extend to estimators obtained under constraint, and how asymptotically equivalent test statistics can be developed using the extremum and the GMM criteria. As a consequence, it is unnecessary for most problems to develop an asymptotic theory for extremum estimators separate from the asymptotic theory for GMM estimators. There are however several practical reasons to introduce and treat extremum estimators separately from GMM estimators. First, while an extremum estimator is a GMM estimator, there may be other roots to the equation $\mathbf{E}\,\nabla_\theta f(z,\theta) = 0$, corresponding to other local extrema of $\mathbf{E}f(z,\theta)$. To make a full equivalence between extremum estimators and GMM estimators, one needs to either have an extremum criterion for which $\mathbf{E}\,\nabla_\theta f(z,\theta) = 0$ has a unique root, with other local extrema ruled out, or one needs to augment the GMM criterion with a procedure that picks out the "correct" root in probability limit. An example of the first situation is a criterion for which $\mathbf{E}$ $f(z,\theta)$ is a globally convex function of θ. An example of the second situation is a procedure that in probability limit finds all the roots of $Q_n(\theta)$, and picks from among them the one that minimizes the extremal criterion in the sample. Second, it is usually computationally simpler to maximize a scalar function than to find roots of a vector of functions, because the heigth of the extremum criterion can be used to verify movement toward a solution and to test for convergence.

To examine more closely the relationship of extremum estimators and GMM estimators based on the first-order conditions from the extremum problem, consider the respective estimators when they are obtained subject to an $r\times 1$ vector of constraints $a(\theta) = 0$. The constrained extremum problem has a Lagrangian $L(\theta,\gamma) = f_n(\theta) - \gamma' a(\theta)$, where γ is a vector of Lagrange multipliers, and the estimator T_{an} satisfies the first-order condition

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} n^{1/2}\cdot\nabla_\theta f_n(T_{an}) - [\nabla_\theta a(T_{an})]' n^{1/2}\gamma_{an} \\ n^{1/2}\cdot a(T_{an}) \end{bmatrix}. \tag{55}$$

Correspondingly, the constrained GMM estimator has a Lagrangian $L(\theta,\gamma) = Q_n(\theta) - \gamma' a(\theta)$, and the first-order condition

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} n^{1/2}\cdot\nabla_\theta Q_n(T_{an}) - [\nabla_\theta a(T_{an})]' n^{1/2}\gamma_{an} \\ n^{1/2}\cdot a(T_{an}) \end{bmatrix}. \tag{56}$$

If the distance metric $G_n(\theta)^{-1}$ is treated as an array of constants when the first-order conditions are calculated, then

$$n^{1/2}\cdot\nabla_\theta Q_n(\theta) = G_n(\theta)' G_n(\theta)^{-1} n^{1/2}\cdot\nabla_\theta f_n(\theta) = n^{1/2}\cdot\nabla_\theta f_n(\theta), \tag{57}$$

so that the first-order condition for the constrained GMM problem coincides with the first-order condition for the constrained extremum problem, and the constrained extremum estimators (T_{an},γ_{an}) are also constrained GMM estimators. Under the regularity conditions of Theorem 1, T_{an} is CAN under the null hypothesis or under local alternatives; see Section 2. Alternately, suppose $G_n(\theta)^{-1}$ is treated as a function of θ in forming the first-order conditions, so that one has

(58) $$n^{1/2}\cdot\nabla_\theta Q_n(\theta) = n^{1/2}\cdot\nabla_\theta f_n(\theta) + \text{vec}\{[n^{1/2}\cdot\nabla_\theta f_n(\theta)]' \quad \frac{\partial G_n(\theta)^{-1}}{\partial \theta_r} \quad [\nabla_\theta f_n(\theta)]\},$$

with the last term denoting a vector with elements corresponding to the components θ_r of θ for r = 1,...,k. But the contribution of the last term is asymptotically negligible, so that the constrained extremum estimator and this form of the constrained GMM estimator, while not necessarily identical, are asymptotically equivalent.

Maximum Likelihood. We have noted that maximum likelihood estimation, with $l(z,\theta)$ the log likelihood of an observation, can be treated as GMM estimation with moments equal to the score, $g = \nabla_\theta l$. The statistics in Table 2 remain the same, with the previously derived simplification that $B = \Omega = G = \Gamma$. The likelihood ratio statistic $2n[L_n(T_n) - L_n(T_{an})]$, where $L_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} l(z_t,\theta)$, is shown by a Taylor's expansion about T_n to be asymptotically equivalent to the Wald statistic W_{3n}, and hence to all the statistics in Table 2. Note that LR and DM occupy comparable places in the trinity for maximum likelihood and GMM estimation respectively.

Suppose one sets up an estimation problem in terms of a maximum likelihood criterion, but that one does not in fact have the true likelihood function. Suppose that in spite of this misspecification, optimization of the selected criterion yields consistent estimates. One place this commonly arises is when panel data observations are serially correlated, but one writes down the *marginal* likelihoods of the observations ignoring serial correlation. These are sometimes called *pseudo-likelihood* criteria. The resulting estimators can be interpreted as GMM estimators, so that hypotheses can be tested using the statistics in Table 2. Note however that now $G \neq -\Omega$, so that $B = G'\Omega^{-1}G$ must be estimated in full, and one cannot do tests using a likelihood ratio of the pseudo-likelihood function.

Least Squares. Consider the nonlinear regression model $y = h(x,\theta) + \varepsilon$, and suppose $E(y|x) = h(x,\theta)$ and $E((y-h(x,\theta))^2|x) = \sigma^2$. The least squares criterion $Q_n(\theta) = \frac{1}{2n}\sum_{t=1}^{n} (y_t - h(z_t,\theta))^2$ is asymptotically equivalent to GMM estimation with $g(z,\theta) = (y-h(x,\theta))\nabla_\theta h(x,\theta)$ and a distance metric $\Omega_n = \frac{\sigma^2}{2n}\sum_{t=1}^{n} [\nabla_\theta h(x,\theta_o)][\nabla_\theta h(x,\theta_o)]'$. For this problem, $B = \Omega = G$. If $h(z_t,\theta) = z_t'\theta$ is linear, one has $g(z_t,\theta) = u_t(\theta)z_t$, where $u_t(\theta) = y_t - z_t'\theta$ is the regression residual, and $\Omega_n = \frac{1}{n}\sum_{t=1}^{n} z_t z_t'$.

Instrumental Variables. Consider the regression model $y_t = h(z_t,\theta_o) + \varepsilon_t$ where ε_t may be correlated with $\nabla_\theta h(z_t,\theta_o)$. Suppose there are *instruments* w such that $\mathbf{E}(\varepsilon_t|w_t) = 0$. For this problem,

one has the moment conditions $g(y_t,z_t,w_t,\theta) = (y_t - h(z_t,\theta))f(w_t)$ satisfying $\mathbf{E}g(y_t,z_t,w_t,\theta_o) = 0$ for any vector of functions f(w) of the instruments, so the GMM criterion becomes

$$Q_n(\theta) = \left[\frac{1}{n}\sum_{t=1}^{n}(y_t - h(z_t,\theta))f(w_t)\right]\left[\frac{1}{n}\sum_{t=1}^{n}f(w_t)f(w_t)'\right]^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}(y_t - h(z_t,\theta))f(w_t)\right] .$$

Suppose that it were feasible to construct the conditional expectation of the gradient of the regression function conditioned on w, $q_t = \mathbf{E}(\nabla_\theta h(z_t,\theta_o)|w_t)$. This is the optimal vector of functions of the instruments, in the sense that the GMM estimator based on f(w) = q will yield estimators with an asymptotic covariance matrix that is smaller in the positive definite sense than any other distinct vector of functions of w. A feasible GMM estimator with good efficiency properties may then be obtained by first obtaining a preliminary consistent estimator τ_n employing a simple practical distance metric, second regressing $\nabla_\theta h(z_t,\tau_n)$ on a flexible family of functions of w_t, such as low-order polynomials in w, and third using fitted values from this regression as the vector of functions $f(w_t)$ in a final GMM estimation. Simplifications of this problem result when $h(z,\theta) = z'\theta$ is linear in θ; in this case, the feasible procedure above is simply 2SLS, and no iteration is needed.

Simple hypotheses. An important practical case of the general nonlinear hypothesis $a(\theta_o) = 0$ is that a subset of the parameters are zero. (A hypothesis that parameters equal constants other than zero can be reduced to this case by reparameterization.) Assume $\theta' = (\alpha',\beta')$ where β is of dimension r and α is of dimension k-r, and H_o: β = 0. The first-order conditions for solution of this problem are $0 = \nabla_\alpha Q_n(T_{an})$, $0 = \nabla_\beta Q_n(T_{an}) + \gamma_{an}$, implying $\gamma_{an} = -\nabla_\beta Q_n(T_{an})$, and $A = [0\ I_r]$ is a r×k matrix whose first k-r columns are zero. Let $C \equiv B^{-1}$ be the asymptotic covariance matrix of $n^{1/2}(T_n - \theta_o)$, and $AB^{-1}A' = C_{\beta\beta}$ the submatrix of C for β. Taylor's expansions about T_n of the first-order conditions imply $n^{1/2}(T_{1,n}-T_{1,an}) = -B_{\alpha\alpha}B_{\alpha\beta}\, n^{1/2}T_{2,n} + o_p$ and $n^{1/2}\gamma_{an} = [B_{\beta\beta}-B_{\beta\alpha}B_{\alpha\alpha}^{-1}B_{\alpha\beta}]n^{1/2}T_{2,n} + o_p = \beta|_n'C_{\beta\beta}^{-1}T_{2,n} + o_p$. Then the Wald statistics are

$$W_{1n} = nT_{2,n}'C_{\beta\beta}^{-1}T_{2,n},\ W_{2n} = n\begin{bmatrix}T_{1,n}-T_{1,an}\\ T_{2,n}\end{bmatrix}'\begin{bmatrix}B_{\alpha\beta}\\ B_{\beta\beta}\end{bmatrix}C_{\beta\beta}^{-1}\begin{bmatrix}B_{\beta\alpha} & B_{\beta\beta}\end{bmatrix}\begin{bmatrix}T_{1,n}-T_{1,an}\\ T_{2,n}\end{bmatrix},$$

$$W_{3n} = n\begin{bmatrix}T_{1,n}-T_{1,an}\\ T_{2,n}\end{bmatrix}'B\begin{bmatrix}T_{1,n}-T_{1,an}\\ T_{2,n}\end{bmatrix}.$$

You can check the asymptotic equivalence of these statistics by substituting the expression for $n^{1/2}(T_{1,n}-T_{1,an})$. The LM statistic, in any version, becomes $LM_n = n\nabla_\beta Q_n(T_{an})'C_{\beta\beta}\nabla_\beta Q_n(T_{an})$. Recall that B, hence C, can be evaluated at any consistent estimator of θ_o. In particular, the constrained estimator is consistent under the null or under local alternatives. The LM testing procedure for this case is then to (a) compute the constrained estimator $T_{1,an}$ subject to the condition β = 0, (b) calculate the gradient and hessian of Q_n with respect to the full parameter vector, evaluated at $T_{1,an}$ and β = 0, and (c) form the quadratic form above for LM_n from the β part of the gradient and the β submatrix of the inverse of the hessian. Note that this does not require any iteration of the GMM criterion with respect to the full parameter vector.

It is also possible to carry out the calculation of the LM_n test statistic using auxiliary regressions. This could be done using the auxiliary regression technique introduced earlier for the

calculation of LM_{3n} in the case of any nonlinear hypothesis, but a variant is available for this case that reduces the size of the regressions required. The steps are as follows:

a. Regress $\nabla_\alpha l(z_t,T_{an})'$ and $\nabla_\beta l(z_t,T_{an})'$ on $g(z_t,T_{an})$, and retrieve the fitted values $\nabla_\alpha l^*(z_t,T_{an})'$ and $\nabla_\beta l^*(z_t,T_{an})'$.
b. Regress $\nabla_\beta l^*(z_t,T_{an})$ on $\nabla_\alpha l^*(z_t,T_{an})$, and retrieve the *residual* $u(z_t,T_{an})$.
c. Regress the constant 1 on the residual $u(z_t,T_{an})$, and calculate the sum of squares of the *fitted* values of 1. This quantity is LM_n.

In the case of maximum likelihood estimation, Step (a) is redundant and can be omitted.

7. TESTS FOR OVER-IDENTIFYING RESTRICTIONS

Consider the GMM estimator based on moments $g(z_t,\theta)$, where g is m×1, θ is k×1, and m > k, so there are *over-identifying moments*. The criterion

$$Q_n(\theta) = (1/2)g_n(\theta)'\Omega_n^{-1}g_n(\theta),$$

evaluated at its minimizing argument T_n for any $\Omega_n \to_p \Omega$, has the property that $2nQ_n \equiv 2nQ_n(T_n) \to_d \chi^2(m\text{-}k)$ under the null hypothesis that $\mathbf{E}g(z,\theta_o) = 0$. This statistic then provides a specification test for the over-identifying moments in g. It can also be used as an indicator for convergence in numerical search for T_n.

To demonstrate this result, recall that $-\Omega^{-1/2} n^{1/2}g_n(\theta_o) = U_n \to_d U \sim N(0,I)$ and $n^{1/2}(T_n - \theta_o) = B^{-1}G'\Omega^{-1/2}U_n + o_p$. Then, a Taylor's expansion yields

$$\Omega^{-1/2} n^{1/2}g_n(T_n) = -U_n + \Omega^{-1/2}GB^{-1}G'\Omega^{-1/2}U_n + o_p = -R_nU_n + o_p,$$

where $R_n = I - \Omega^{-1/2}G(G'\Omega^{-1}G)^{-1}G'\Omega^{-1/2}$ is idempotent of rank m - k. Then

$$2nQ_n(T_n) = U_n'R_nU_n + o_p \to_d \chi^2(m\text{-}k) .$$

Suppose that instead of estimating θ using the full list of moments, one uses a linear combination $Lg(z,\theta)$, where L is r×m with $k \le r < m$. In particular, L may select a subset of the moments. Let T_{an} denote the GMM estimator obtained from these moment combinations, and assume the identification conditions are satisfied so T_{an} is $n^{1/2}$-consistent. Then the statistic $S = ng_n(T_{an})'\Omega_n^{-1/2}R_n\Omega_n^{-1/2}g_n(T_{an}) \to_d \chi^2(m\text{-}k)$ under H_o, and this statistic is asymptotically equivalent to the statistic $2nQ_n(T_n)$. This result holds for any $n^{1/2}$-consistent estimator τ_n of θ_o, not necessarily the optimal GMM estimator for the moments $Lg(z,\theta)$, or even an initially consistent estimator based on only these moments. The distance metric in the center of the quadratic form S does not depend on L, so that the formula for the statistic is invariant with respect to the choice of the initially consistent estimator. This implies in particular that the test statistics S for over-identifying restrictions, starting from different subsets of the moment conditions, are all asymptotically equivalent. However, the presence of the

idempotent matrix R_n in the center of the quadratic form S is critical to its statistical properties. Only the GMM distance metric criterion using all moments, evaluated at T_n, is asymptotically equivalent to S. Substitution of another consistent estimator τ_n in place of T_n yields an asymptotically equivalent version of S, but $2nQ_n(\tau_n)$ is not asymptotically chi-square distributed.

The test for overidentifying restrictions can be recast as a LM test by artificially embedding the original model in a richer model. Partition the moments

$$g(z,\theta) = \begin{bmatrix} g^1(z,\theta) \\ g^2(z,\theta) \end{bmatrix},$$

where g^1 is kx1 with $G_1 = \mathbf{E}\nabla_\theta g^1(z,\theta_o)$ of rank k, and g^2 is (m-k)x1 with $G_2 = E\nabla_\theta g^2(z,\theta_o)$. Embed this in the model

$$g^*(z,\theta,\psi) = \begin{bmatrix} g^1(z,\theta) \\ g^2(z,\theta)+\psi \end{bmatrix}$$

where ψ is a (m-k) vector of additional parameters. The first-order-condition for GMM estimation of this expanded model is

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} G_{1n} & G_{2n} \\ 0 & I_{m-k} \end{bmatrix} \begin{bmatrix} \Omega_n & 0 \\ 0 & I_{m-k} \end{bmatrix} \begin{bmatrix} g_n(T_{an}) \\ g_n(T_{an}) - \psi_n \end{bmatrix}$$

The second block of conditions are satisfied by $\psi_n = g_n(T_{an})$, no matter what T_{an}, so T_{an} is determined by $O = G_n\Omega_n g_n(T_{an})$. This is simply the estimator obtained from the first block of moments, and coincides with the earlier definition of T_{an}. Thus, *unconstrained* estimation of the *expanded* model coincides with *restricted* estimation of the original model. Next consider GMM estimation of the expanded model subject to H_o:$\psi = O$. This constrained estimation obviously coincides with GMM estimation using all moments in the original model, and yields T_n. Thus, *constrained* estimation of the *expanded* model coincides with *unrestricted* estimation of the original model.

The Distance Metric test statistic for the constraint $\psi = 0$ in the expanded model is $DM_n = 2n[Q_n(T_n,0) - Q_n(T_n,\psi_n)] \equiv 2nQ_n(T_n)$, where Q_n denotes the criterion as a function of the expanded parameter list. One has $Q_n(T_n,0) \equiv Q_n(T_n)$ from the coincidence of the constrained expanded model estimator and the unrestricted original model estimator, and one has $Q_n(T_{an},\psi_n) = 0$ since the number of moments equals the number of parameters. Then, the test statistic $2nQ_n(T_n)$ for overidentifying restrictions is identical to a distance metric test in the expanded model, and hence asymptotically equivalent to any of the trinity of tests for H_o: $\psi = O$ in the expanded model.

We give four examples of econometric problems that can be formulated as tests for over-identifying restrictions:

Example 1. If $y = x\beta+\varepsilon$ with $E(\varepsilon|x) = 0$, $E(\varepsilon^2|x) = \sigma^2$, then the moments

$$g^1(z,\beta) = \begin{bmatrix} x(y-x\beta) \\ (y-x\beta)^2 - \sigma^2 \end{bmatrix}$$

can be used to estimate β and σ^2. If ε is normal, then GMM estimators based on g^1 are MLE. Normality can be tested via the additional moments that give skewness and kurtosis,

$$g^2(x,\beta) = \begin{bmatrix} (y-x\beta)^3/\sigma^3 \\ (y-x\beta)^4/\sigma^4 - 3 \end{bmatrix}.$$

GMM estimators based on all the moments g are again MLE

Example 2. In the linear model $y = xb+\varepsilon$ with $E(\varepsilon|x) = 0$ and $E(\varepsilon_t\varepsilon_s|x) = 0$ for $t \neq s$, but with possible heteroskedasticity of unknown form, one gets the OLS estimates b of β and $V(b) = s^2(X'X)^{-1}$ under the null hypothesis of homoskedasticity. A test for homoskedasticity can be based on the population moments $0 = \mathbf{E}\ \text{vecu}[x'x(\varepsilon^2 - \sigma^2)]$, where "vecu" means the vector formed from the upper triangle of the array. The sample value of this moment vector is

$$\text{vecu}\left[\frac{1}{n}\sum_{t=1}^{n} x_t'x_t\left((y_t-x_t\beta)^2 - s^2\right)\right],$$

the difference between the White robust estimator and the standard OLS estimator of $\text{vecu}[X'\Omega X]$.

Example 3. If $l(z,\theta)$ is the log likelihood of an observation, and T_n is the MLE, then an additional moment condition that should hold if the model is specified correctly is the information matrix equality

$$0 = E\,\nabla_{\theta\theta} l(z,\theta_o) + E\nabla_\theta l(z,\theta_o)\nabla_\theta l(z,\theta_o)'.$$

The sample analog is White's information matrix test, which then can be interpreted as a GMM test for over-identifying restrictions.

Example 4. In the nonlinear model $y = h(x,\theta) + \varepsilon$ with $E(\varepsilon|x) = 0$, and T_n a GMM estimator based on moments $w(x)(y-h(x,\theta))$, where $w(x)$ is some vector of functions of x, suppose one is interested in testing the stronger assumption that ε is *independent* of x. A necessary and sufficient condition for independence is $\mathbf{E}[w(x) - \mathbf{E}w(x)]f(y - h(x,\theta_o)) = 0$ for every function f and vector of functions w for which the moments exist. A specification test can be based on a selection of such moments.

8. SPECIFICATION TESTS IN LINEAR MODELS[1]

GMM tests for over-identifying restrictions have particularly convenient forms in linear models. Three standard specification tests will be shown to have this interpretation. We will use

[1] Paul Ruud contributed substantially to this section.

projections and a few of their properties in the following discussion; a more detailed discussion of projections is given in the Appendix to this chapter. Let $P_X = X(X'X)^{-}X$ denote the *projection matrix* from $\mathbb{R}^n$ onto the linear subspace $\mathbf{X}$ *spanned* by a n×p array X; note that it is idempotent. (We use a Moore-Penrose generalized inverse in the definition of P_X to handle the possibility that X is less than full rank; see the Appendix.) Let $Q_X = I - P_X$ denote the projection matrix onto the linear subspace orthogonal to $\mathbf{X}$. If $\mathbf{X}$ is a subspace generated by an array X and $\mathbf{W}$ is a subspace generated by an array W = [X Z] that contains $\mathbf{X}$, then $P_X P_W = P_W P_X = P_X$ and $Q_X P_W = P_W - P_X$.

Omitted Variables Test: Consider the regression model $y = X\beta + \varepsilon$, where y is n×1, X is n×k, $\mathbf{E}(\varepsilon|X) = 0$, and $\mathbf{E}(\varepsilon\varepsilon'|X) = \sigma^2 I$. Suppose one has the hypothesis H_o: $\beta_1 = 0$, where β_1 is a p×1 subvector of β, and let X* denote the n×(k-p) array of variables whose coefficients are not constrained under the null hypothesis. Define u = y - Xb to be the residual associated with an estimator b of β. The GMM criterion is then $2nQ = u'X(X'X)^{-1}X'u/\sigma^2$. The *projection matrix* $P_X \equiv X(X'X)^{-1}X'$ that appears in the center of this criterion can obviously be decomposed as $P_X \equiv P_{X^*} + (P_X - P_{X^*})$. Under H_o, $u = y - X_2 b_2$ and $X'u$ can be interpreted as k = p + q over-identifying moments for the q parameters β_2. Then, the GMM test statistic for over-identifying restrictions is the minimum value $2nQ_n^*$ in b_2 of $u'P_X u/\sigma^2$. But $P_X u = P_{X^*} u + (P_X - P_{X^*})y$ and $\min_{b_2} u' P_{X^*} u = 0$ (at the OLS estimator under H_o that makes u orthogonal to X_2). Then $2nQ_n = y'(P_X - P_{X^*})y/\sigma^2$. The unknown variance σ^2 in this formula can be replaced by any consistent estimator s^2, in particular, the estimated variance of the disturbance from either the restricted or the unrestricted regression, without altering the asymptotic distribution, which is $\chi^2(q)$ under the null hypothesis.

The statistic $2nQ_n$ has three alternative interpretations. First,

$$2nQ_n = y'P_X y/\sigma^2 - y'P_{X^*} y/\sigma^2 = \frac{SSR_{X_2} - SSR_X}{\sigma^2},$$

which is the difference of the sum of squared residuals from the restricted regression under H_o and from the unrestricted regression, normalized by σ^2. This is a large-sample version of the usual finite-sample F-test for H_o. Second, note that the fitted value of the dependent variable from the restricted regression is $\hat{y}_o = P_{X^*} y$, and from the unrestricted regression is $\hat{y}_u = P_X y$, so that

$$2nQ_n = (\hat{y}_o'\hat{y}_o - \hat{y}_u'\hat{y}_u)/\sigma^2 = (\hat{y}_o - \hat{y}_u)'(\hat{y}_o - \hat{y}_u)/\sigma^2 = \|\hat{y}_o - \hat{y}_u\|^2/\sigma^2.$$

Then, the statistic is calculated from the distance between the fitted values of the dependent variable with and without H_o imposed. Note that it can be computed from fitted values without any covariance matrix calculation. Third, let b_o denote the GMM estimator restricted by H_o and b_u denote the unrestricted GMM estimator. Then, b_o consists of the OLS estimator for β_2 and the hypothesized value 0 for β_1, while b_u is the OLS estimator for the full parameter vector. Note that $\hat{y}_o = Xb_o$ and $\hat{y}_u = Xb_u$, so that $\hat{y}_o - \hat{y}_u = X(b_o - b_u)$. Then

$$2nQ_n = (b_o - b_u)'(X'X/\sigma^2)(b_o - b_u) = (b_o - b_u)'V(b_u)^{-1}(b_o - b_u).$$

This is the Wald statistic W_{3n}. From the equivalent form W_{2n} of the Wald statistic, this can also be written as a quadratic form $2nQ_n = b_{1,u}'V(b_{1,u})^{-1}b_{1,u}$, where $b_{1,u}$ is the subvector of unrestricted estimates for the parameters that are zero under the null hypothesis.

Two other important cases of specification tests in linear models are discussed in the following chapters. *Endogeneity tests* are discussed in the chapter on instrumental variables, and *tests for over-identifying restrictions* are discussed in the chapter on simultaneous equations.

APPENDIX

Projections: Consider a Euclidean space $\mathbb{R}^n$ of dimension n, and suppose X is a n×p array with columns that are vectors in this space. Let **X** denote the linear subspace of $\mathbb{R}^n$ that is *spanned* or *generated* by X.; and i.e., the space formed by all linear combinations of the vectors in X. Every linear subspace can be identified with an array such as X. The dimension of the subspace is the rank of X. (The array X need not be of full rank, although if it is not, then a subarray of linearly independent columns also generates **X**.) A given X determines a unique subspace, so that X characterizes the subspace. However, any set of vectors contained in the subspace that form an array with the rank of the subspace, in particular any array XA with rank equal to the dimension of X, also generates **X**. Then, X is not a unique characterization of the subspace it generates.

The *projection* of a vector y in $\mathbb{R}^n$ into the subspace **X** is defined as the point v in **X** that is the minimum Euclidean distance from y. Since each vector v in **X** can be represented as a linear combination Xα of an array X that generates **X**, the projection is characterized by the value of α that minimizes $(y-X\alpha)'(y-X\alpha)$. The solution to this problem is the OLS estimator $\hat{\alpha} = (X'X)^{-}X'y$ and $v = X\hat{\alpha} = X(X'X)^{-}X'y$. In these formulas, we use $(X'X)^{-}$ rather than $(X'X)^{-1}$; the former denotes the Moore-Penrose *generalized* inverse, and is defined even if X is not of full rank (see below). The array $P_X = X(X'X)^{-}X'$ is termed the *projection matrix* for the subspace **X**; it is the linear transformation in $\mathbb{R}^n$ that maps any vector in the space into its projection v in **X**. The matrix P_X is *idempotent* (i.e., $P_XP_X = P_X$ and $P_X = P_X'$), and every idempotent matrix can be interpreted as a projection matrix. These observations have two important implications: First, the projection matrix is uniquely determined by X, so that starting from a different array that generates **X**, say an array S = XA, implies $P_X = P_S$. (One could use the notation $P_{\mathbf{X}}$ rather than P_X to emphasize that the projection matrix depends only on the subspace, and not on any particular set of vectors that generate **X**.) Second, if a vector y is contained in **X**, then the projection into **X** leaves it unchanged, $P_Xy = y$.

Define $Q_X = I - P_X = I - X(X'X)^{-1}X'$; it is the projection to the subspace orthogonal to that spanned by X. Every vector y in $\mathbb{R}^n$ is uniquely decomposed into the sum of its projection P_Xy onto **X** and its projection Q_Xy onto the subspace orthogonal to **X**. Note that $P_XQ_X = 0$, a property that holds in general for two projections onto orthogonal subspaces.

If **X** is a subspace generated by an array X and **W** is a subspace generated by an array W = [X Z] that contains X, then $\mathbf{X} \subseteq \mathbf{W}$. This implies that $P_XP_W = P_WP_X = P_X$; i.e., a projection onto a subspace is left invariant by a further projection onto a larger subspace, and a two-stage projection onto a large subspace followed by a projection onto a smaller one is the same as projecting directly

onto the smaller one. The subspace of **W** that is orthogonal to **X** is generated by $Q_X W$; i.e., it is the set of linear combinations of the residuals, orthogonal to X, obtained by regressing W on X. Note that any y in $\mathbb{R}^n$ has a unique decomposition $P_X y + Q_X P_W y + Q_W y$ into the sum of projections onto three mutually orthogonal subspaces, **X**, the subspace of **W** orthogonal to **X**, and the subspace orthogonal to **W**. The projection $Q_X P_W$ can be rewritten $Q_X P_W = P_W - P_X = P_W Q_X = Q_X P_W Q_X$, or since $Q_X W = Q_X[X \;\; Z] = [0 \;\; Q_X Z]$, $Q_X P_W = P_{Q_X W} = P_{Q_X Z} = Q_X Z(Z'Q_X Z)^{-} Z'Q_X$. This establishes that P_W and Q_X commute. This condition is necessary and sufficient for the product of two projections to be a projection; equivalently, it implies that $Q_X P_W$ is idempotent since $(Q_X P_W)(Q_X P_W) = Q_X(P_W Q_X)P_W = Q_X(Q_X P_W)P_W = Q_X P_W$.

Generalized Inverses: Some test statistics are conveniently defined using generalized inverses. This section gives a constructive definition of a generalized inverse, and lists some of its properties. A k×m matrix A^- is a *Moore-Penrose generalized inverse* of a m×k matrix A if it has three properties:

(i) $AA^-A = A$,
(ii) $A^-AA^- = A^-$
(iii) AA^- and A^-A are symmetric

There are other generalized inverse definitions that have some, but not all, of these properties; in particular A^+ will denote any matrix that satisfies (i), or $AA^+A = A$.

First, a method for constructing the generalized inverse is described, and then some of the implications of the definition are developed. The construction is called the *singular value decomposition* (SVD) of a matrix, and is of independent interest as a tool for finding the eigenvalues and eigenvectors of a symmetric matrix, and for calculation of inverses of moment matrices of data with high multicollinearity; see Press *et al* (1986) for computational algorithms and programs.

Lemma 1. Every real m×k matrix A *of rank r can be decomposed into a product* $A = UDV'$ *where* D *is a r×r diagonal matrix with positive non-increasing elements down the diagonal, and* U *and* V *are column-orthonormal matrices of respective dimension m×r and k×r; i.e.,* $U'U = I_r = V'V$.

Proof: The m×m matrix AA' is symmetric and positive semidefinite. Then, there exists a m×m orthonormal matrix W, partitioned $W = [W_1 \; W_2]$ with W_1 of dimension m×r, such that $W_1'(AA')W_1 = G$ is diagonal with positive, non-increasing diagonal elements, and $W_2'(AA')W_2 = 0$, implying $A'W_2 = 0$. Define D from G by replacing the diagonal elements of G by their positive square roots. Note that $W'W = I = WW' \equiv W_1 W_1' + W_2 W_2'$. Define $U = W_1$ and $V' = D^{-1}U'A$. Then, $U'U = I_r$ and $V'V = D^{-1}U'AA'UD^{-1} = D^{-1}GD^{-1} = I_r$. Further, $A = (I_m - W_2 W_2')A = UU'A = UDV'$. This establishes the decomposition. □

Note that if A is symmetric, then U is the array of eigenvectors of A corresponding to the non-zero roots, so that $A'U = UD_1$, with D_1 the r×r diagonal matrix with the non-zero eigenvalues in descending magnitude down the diagonal. In this case, $V = A'UD^{-1} = UD_1D^{-1}$. Since the elements of D_1 and D are identical except possibly for sign, the columns of U and V are either equal (for positive roots) or reversed in sign (for negative roots). Thus, if A is positive semidefinite, it has a SVD decomposition $A = UDU'$ with U column-orthonormal and D positive diagonal.

Lemma 2. The *Moore-Penrose generalized inverse of a* m×k *matrix* A *(which has a SVD A = UDV′) is the matrix* $A^- = VD^{-1}U$, *where* V is k×r, D is r×r, *and* U *is* r×m. *Let* A^+ *denote any matrix, including* A^-, *that satisfies* $AA^+A = A$. *These matrices satisfy:*

(1) $A^+ = A^{-1}$ *if* A *is square and non-singular*.

(2) *The system of equations* Ax = y *has a solution if and only if* $y = AA^+y$, *and the linear subspace of all solutions is the set of vectors* $x = A^+y + [I - A^+A]z$ *for all* $z \in \mathbb{R}^k$.

(3) AA^+ *and* A^+A *are idempotent*.

(4) *If* A *is idempotent, then* $A = A^-$.

(5) *If* A = BCD *with* B *and* D *nonsingular, then* $A^- = D^{-1}C^-B^{-1}$, *and any matrix* $A^+ = D^{-1}C^+B^{-1}$ *satisfies* $AA^+A = A$.

(6) $(A')^- = (A^-)'$

(7) $(A'A)^- = A^-(A^-)'$

(8) $(A^-)^- = A = AA'(A^-)' = (A^-)'A'A$.

(9) *If* $A = \sum_i A_i$ *with* $A_i'A_j = 0$ *and* $A_iA_j' = 0$ *for* $i \neq j$, *then* $A^- = \sum_i A_i^-$.

Lemma 3. If A *is* m×m, *symmetric, and positive semidefinite of rank r, then*

(1) There exist Q *positive definite and* R *idempotent of rank r such that* A = QRQ *and* $A^- = Q^{-1}RQ^{-1}$.

(2) There exists an m×r *column-orthonormal matrix* U *such that* U′AU = D *is positive diagonal,* A = UDU′, $A^- = UD^{-1}U' = U(U'AU)^{-1}U'$, *and any matrix* A^+ *satisfying condition (i) for a generalized inverse,* $AA^+A = A$, *has* $U'A^+U = D^{-1}$.

(3) A *has a symmetric square root* $B = A^{1/2}$, *and* $A^- = B^-B^-$.

Proof: Let U be an m×r column-orthonormal matrix of eigenvectors of A corresponding to the positive characteristic roots, and W be a m×(m-r) column-orthonormal matrix of eigenvectors corresponding to the zero characteristic roots. Then [U W] is an orthonormal matrix diagonalizing A, with $\begin{bmatrix} U' \\ W' \end{bmatrix} A \begin{bmatrix} U & W \end{bmatrix} = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$ and D positive diagonal. Define $Q = \begin{bmatrix} U & W \end{bmatrix} \begin{bmatrix} D^{1/2} & 0 \\ 0 & I_{m-r} \end{bmatrix} \begin{bmatrix} U' \\ W' \end{bmatrix}$, and R = UU′. The diagonalizing transformation implies U′AU = D and AW = 0. One has U′U = I_r, W′W = I_{m-r}, and UU′ + WW′ = I_m. Since AW = 0, A = A[UU′ + WW′] = AUU′. Then D = U′AU = U′AA⁺AU = UAUU′A⁺UU′AU = DU′A⁺UD, implying $U'A^+U = D^{-1}$. Define $B = UD^{1/2}U'$. □

Lemma 4. Suppose $y \sim N(C\mu, CC')$, *with* C *a* m×r *matrix of rank r. Let* A = CC′ *and* λ = Cμ. *Then for any matrix* A^+ *satisfying condition (i) for a generalized inverse,* $AA^+A = A$, *one has* $y'A^+y = y'A^-y$ *distributed noncentral chi-square with r degrees of freedom and noncentrality parameter* $\lambda'A^-\lambda$.

Proof: Use the orthonormal matrix [U W] from the proof of Lemma 3, so that U′CC′U = D, a positive diagonal r×r matrix, and C′W = 0. Then, the nonsingular transformation

$$z = \begin{bmatrix} D^{-1/2} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} U' \\ W' \end{bmatrix} y$$

has mean $\begin{bmatrix} D^{-1/2}U'C\mu \\ 0 \end{bmatrix}$ and covariance matrix $\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$, so that $z_1 = D^{-1/2}U'y$ is distributed $N(D^{-1/2}U'C\mu, I_r)$ and $z_2 = W'y = 0$. It is standard that $z'z$ has a non-central chi-square distribution with r degrees of freedom and non-centrality parameter $\mu'C'UD^{-1}U'C\mu = \lambda'A^{-}\lambda$. From result (2) of Lemma 3, $U'A^{+}U = D^{-1}$. Then

$$y'A^{+}y = y'[UU' + WW']A^{+}[UU' + WW']y = y'UD^{-1}U'y = y'A^{-}y$$

and

$$y'A^{-}y = y'UD^{-1}U'y = y'UD^{-1/2}D^{-1/2}U'y = z_1'z_1. \ \square$$