GENERALIZED METHOD OF MOMENTS

z **data generated by a process parameterized by a $k\times 1$ vector θ.**

$l(z,\theta)$ **log likelihood of z,**

θ_o **true value of θ in the population.**

$g(z,\theta)$ **$m\times 1$ vector of functions of z and θ that have zero expectation in the population if and only if θ equals θ_o:**

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

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.

Example. $z = (x,y)$, $y = f(x,\theta_0) + \epsilon$, $x \perp\!\!\!\perp \epsilon$

$$g(z,\theta) = P(x)'(y-f(x,\theta))$$

with P(x) a vector of polynomials in x.

Assume $g(z,\theta_o)$ has a positive definite m×m covariance matrix.
The GMM 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.

Suppose an i.i.d. sample $z_1,\ldots,z_n$ 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,

(2) $$Q_n(\theta) = \tfrac{1}{2} g_n(\theta)' W_n(\tau_n) g_n(\theta), \quad \text{with}$$

$$g_n(\theta) \equiv \frac{1}{n}\sum_{t=1}^{n} g(z_t,\theta) ,$$

$W_n(\theta)$ is a m×m positive definite symmetric matrix, in general depending on θ, evaluated at some sequence of "preliminary estimates" τ_n.

The $W_n(\tau_n)$ define a "distance metric". Let W_n = $W_n(\tau_n)$. Assume that $W_n(\theta)$ converges in probability uniformly in θ to $W(\theta)$, a continuous positive definite limit. Let W = plim W_n. If plim $\tau_n = \theta_o$, then plim $W_n(\tau_n) = W(\theta_o) = W$.

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., possibly asymptotic efficiency) will depend on the interplay of $g(z,\theta)$ and $l(z,\theta)$.

$\Omega(\theta) \equiv E\, g(z,\theta)g(z,\theta)'$ m×m covariance matrix of the moments.

Efficient weighting requires plim $W_n = \Omega(\theta_o)^{-1}$. Call a GMM estimator that has plim $W_n = \Omega(\theta_o)^{-1}$ 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 first minimize a GMM criterion using the identity matrix I_m for W_n.

$G(\theta) \equiv -E\, \nabla_\theta g(z,\theta)$ m×k Jacobean matrix

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

$G_n(\tau_n)$ 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 $G(\theta_o)$.

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.

Special cases

- $f(z,\theta)$ **is a scalar function with the property that E** $f(z,\theta_o) \le$ **E** $f(z,\theta)$**. 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 is** $f(z,\theta) = - l(z,\theta)$**, the negative of a full or limited information log likelihood function. 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 <u>other</u> than the extremum estimator, then one can treat the extremum estimator in its equivalent GMM form.**

- **$z = (y,x,w)$ 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.**

- **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 $y - h(x,\theta_o)$.**

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. In detail, for each $\varepsilon > 0$, define

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

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

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.

Example: 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. If $\theta \in [0,1]$ and s is uniform on [0,1], $Y(\theta,s) = 1$ if $\theta \geq s$ and $Y(\theta,s) = 0$ otherwise is almost surely continuous everywhere but has a discontinuity.

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

Lemma 3.3 (Uniform WLLN). ***Assume*** $\mathbf{Y_i(\theta)}$ ***are independent identically distributed random functions with a finite mean*** $\boldsymbol{\psi(\theta)}$ ***for*** $\boldsymbol{\theta}$ ***in a closed bounded set*** $\boldsymbol{\Theta \subseteq \mathbb{R}^k}$. ***Assume*** $\mathbf{Y_i(\cdot)}$ ***is almost surely continuous at each*** $\boldsymbol{\theta \in \Theta}$***. Assume that*** $\mathbf{Y_i(\cdot)}$ ***is dominated; i.e., there exists a random variable Z with a finite mean that satisfies*** $\mathbf{Z \geq \sup_{\theta\in\Theta}|Y_1(\theta)|}$. ***Then*** $\boldsymbol{\psi(\theta)}$ ***is continuous in*** $\boldsymbol{\theta}$ ***and*** $\mathbf{X_n(\theta)} = \frac{1}{n}\sum_{i=1}^{n}$ ***satisfies***

$$\mathbf{\sup_{\theta\in\Theta}|X_n(\theta) - \psi(\theta)| \to_p 0.}$$

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

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*** $T_n = \text{argmin}_{\theta\in\Theta}\, Q_n(\theta)$. ***Assume*** **(i) to (vii):**

> **(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, $|\nabla_\theta g(z,\theta) - \nabla_\theta g(z,\theta')| \le \alpha(z)|\theta - \theta'|$.**

(iv) $Eg(z,\theta) = 0$ ***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, $W_n(\theta) \to_p W(\theta)$** ***uniformly in*** **$\theta$*, and*** **$W_n \to_p W$.**

(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., $+\infty > E\alpha(z)$, $|g(z,\theta)g(z,\theta)'| \le \alpha(z)$,** ***and*** **$|\nabla_\theta g(z,\theta)| \le \alpha(z)$.**

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

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

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

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

Proof of Theorem 1. Step 0 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. Step 1 shows for T_n* satisfying $Q_n(T_n^*) \to_p 0$ that $T_n^* \to_p \theta_o$. Step 2 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 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. Step 3 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.

Step 0: 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

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

The expressions $g_n(\theta)$, $G_n(\theta)$, and $\Omega_n(\theta)$ are sample averages that converge in probability for each fixed θ to E*g*(θ), $G(\theta)$, 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 $g_n(\theta) \to_p Eg(\theta)$. We have established that the convergence in probability of $g_n(\theta)$ to $Eg(\theta)$ is uniform in θ. Combined with the condition $W_n \to_p W$ from (vi), this implies $Q_n(\theta) \to_p ½(Eg(\theta))'W(Eg(\theta))$ 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

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

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

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

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

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. □

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 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

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

It will sometimes be convenient to estimate G by

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

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

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

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.

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 $a(\theta_o) = (\theta_{1o}/\theta_{2o} - \theta_{3o}/\theta_{4o})$, 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}: $a(\theta_o) = \delta n^{-1/2} \neq 0$.

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:

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

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.***

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×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

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

The Lagrangian multipliers γ_{an} are random variables. Lemma 5, and when applicable the argument given in the proof of Corollary 1, imply $\nabla_\theta Q_n(T_{an}) \to_p -G'WEg(z,\theta_o) = 0$. 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, $\gamma_{an} \to_p 0$.

The argument for asymptotic normality parallels the argument given in Theorem 1 for the unconstrained estimator, and relates 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

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

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 $\delta = 0$),

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

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

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

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 (23):

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

From Corollary 1, $n^{1/2}(T_n - \theta_o) = C^{-1}G'W\Omega^{1/2}U_n + o_p$. Substitute this in (23) to conclude that

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

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

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

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) and
$K = (I_m - GC^{-1}G'W + GC^{-1}A'(AC^{-1}A')^{-1}AC^{-1}G'W)$
implies

$$n^{1/2}g_n(T_{an}) = K\Omega^{1/2}U_n + GC^{-1}A'(AC^{-1}A')^{-1}\delta + o_p,$$

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,

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

Table 1 summarizes the results. The table shows that the r×1 vectors $An^{1/2}(T_n\text{-}T_{an})$, $n^{1/2}a(T_n)$, $(AC^{-1}A')n^{1/2}\gamma_{an}$, 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 $n^{1/2}(T_n\text{-}\theta_o)$. 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

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

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

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

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.

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 $\delta'(AB^{-1}A')^{-1}\delta = c'(AVA')^{-1}c$.

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}	$na(T_n)'[AC^{-1}HC^{-1}A']^{-1}a(T_n)$	$na(T_n)'[AB^{-1}A']^{-1}a(T_n)$
W_{2n}, flavor 1	$n(T_n-T_{an})'acov(T_n - T_{An})^{-}(T_n - T_{an})$	$n(T_n-T_{an})'\{acov(T_n) - acov(T_{An})\}^{-}(T_n - T_{an})$
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

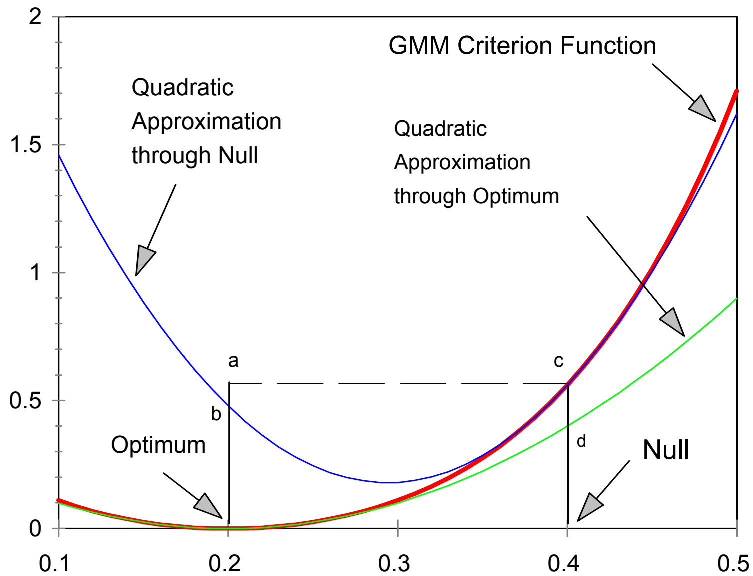


Figure 1 illustrates the relationship between distance metric (DM), Wald (W), and Score (LM) tests for a best GMM estimator. In the case of maximum likelihood estimation, this figure is inverted, the criterion is log likelihood rather than the distance metric, and the DM test is replaced by the likelihood 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 $acov(T_n) = C^{-1}HC^{-1}$ (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 $A = I_k$.

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} 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 $E(\varepsilon_t|w_t) = 0$ and $E(\varepsilon_t^2|w_t) = \sigma^2$, 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 $Eg(z,\theta_o)g(z,\theta_o)' = \sigma^2 Eww'$. The GMM criterion $Q_n(\theta)$ for this model is (28)

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

Optimization is not affected by the scalar σ^2.

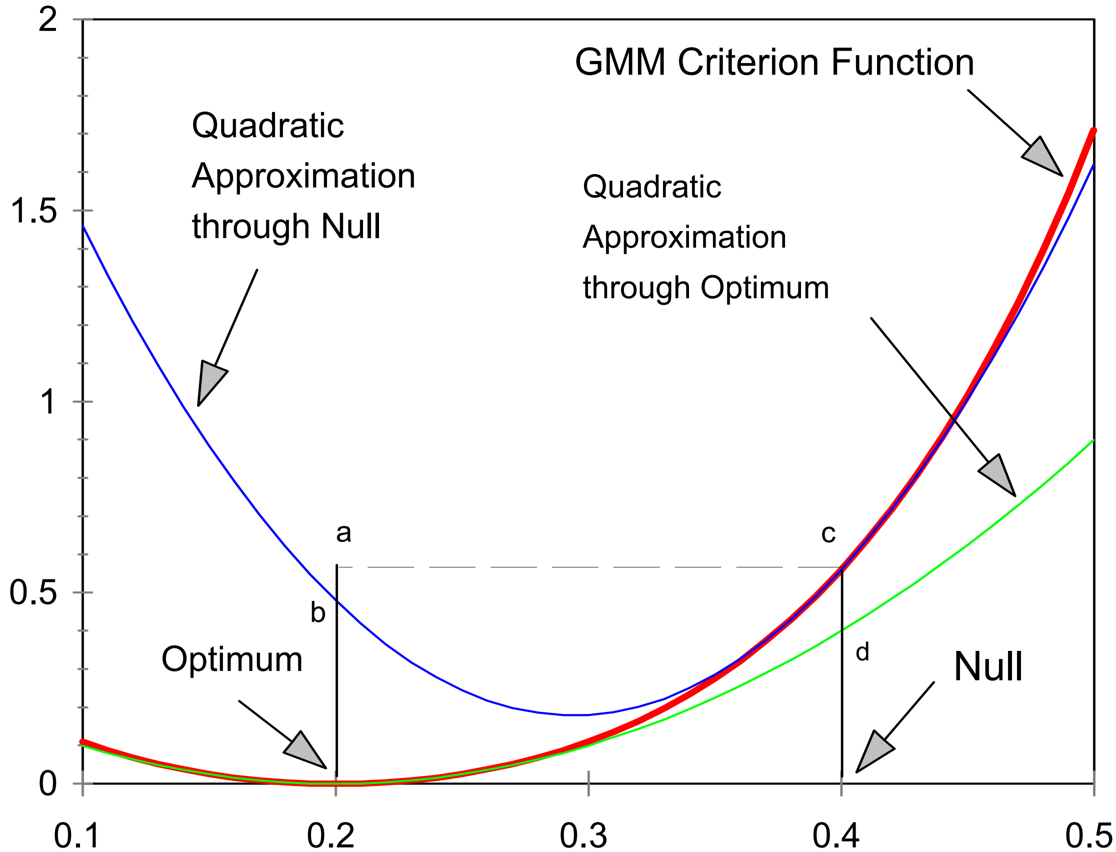
Consider the hypothesis $a(\theta_o) = 0$, and let T_{an} be the constrained GMM estimator. One can compute LM_{3n} by the following method:

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.



TWO-STAGE GMM ESTIMATION

ONE-STEP THEOREMS

SPECIAL CASES

Extremum estimators
Ordinary Least Squares
Simple hypotheses

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 $Eg(z,\theta_o) = 0$.

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 $G_1 = E\nabla_\theta g^1(z,\theta_o)$ 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:ψ = 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 ψ = 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: ψ = 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 = xβ+ε with E(ε|x) = 0, E(ε^2|x) = σ^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+ε with E(ε|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 = E vecu[x$'$x(ε^2- σ^2)], where "vecu" means the vector formed from the upper triangle of the array. The sample value of this moment vector is

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

the difference between the White robust estimator and the standard OLS estimator of 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

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

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,θ) + ε with E(ε|x) = 0, and T_n a GMM estimator based on moments w(x)(y-h(x,θ)), 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 E[w(x) - Ew(x)]f(y- h(x,θ_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.

SPECIFICATION TESTS IN LINEAR MODELS

GMM tests for over-identifying restrictions have particularly convenient forms in linear models. Three standard specification tests will be shown to have this interpretation. Let $P_X = X(X'X)^{-}X$ denote the *projection matrix* from $\mathbb{R}^n$ onto the linear subspace 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.) Let $Q_X = I - P_X$ denote the projection matrix onto the linear subspace orthogonal to 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 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, $E(\varepsilon|X) = 0$, and $E(\varepsilon\varepsilon'|X) = \sigma^2 I$. 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, $u = y - X_2b_2$ 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 $u'P_Xu/\sigma^2$. But $P_Xu = P_{X*}u + (P_X - P_{X*})y$ 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,

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

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 $\hat{\mathbf{y}}_o = \mathbf{P}_{X*}\,\mathbf{y}$, and from the unrestricted regression is $\hat{\mathbf{y}}_u = \mathbf{P}_X\mathbf{y}$, 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

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

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.