GENERALIZED METHOD OF MOMENTS I

Whitney K. Newey

MIT

October 2007

THE GMM ESTIMATOR: The idea is to choose estimates of the parameters by setting sample moments to be close to population counterparts. To describe the underlying moment model and the GMM estimator, let β denote a $p \times 1$ parameter vector, w_i a data observation with $i = 1, ..., n$, where n is the sample size. Let $g_i(\beta) = g(w_i, \beta)$ be a $m \times 1$ vector of functions of the data and parameters. The GMM estimator is based on a model where, for the true parameter value β_0 the moment conditions

$$E[g_i(\beta_0)] = 0$$

are satisfied.

The estimator is formed by choosing β so that the sample average of $g_i(\beta)$ is close to its zero population value. Let

$$\hat{g}(\beta) \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} g_i(\beta)$$

denote the sample average of $g_i(\beta)$. Let $\hat{A}$ denote an $m \times m$ positive semi-definite matrix. The GMM estimator is given by

$$\hat{\beta} = \arg\min_{\beta} \hat{g}(\beta)' \hat{A} \hat{g}(\beta).$$

That is $\hat{\beta}$ is the parameter vector that minimizes the quadratic form $\hat{g}(\beta)' \hat{A} \hat{g}(\beta)$.

The GMM estimator chooses $\hat{\beta}$ so the sample average $\hat{g}(\beta)$ is close to zero. To see this let $\|g\|_{\hat{A}} = \sqrt{g' \hat{A} g}$, which is a well defined norm as long as $\hat{A}$ is positive definite. Then since taking the square root is a strictly monotonic transformation, and since the minimand of a function does not change after it is transformed, we also have

$$\hat{\beta} = \arg\min_{\beta} \|\hat{g}(\beta) - 0\|_{\hat{A}}.$$

Thus, in a norm corresponding to $\hat{A}$ the estimator $\hat{\beta}$ is being chosen so that the distance between $\hat{g}(\beta)$ and 0 is as small as possible. As we discuss further below, when $m = p$, so there are the same number of parameters as moment functions, $\hat{\beta}$ will be invariant to $\hat{A}$ asymptotically. When $m > p$ the choice of $\hat{A}$ will affect $\hat{\beta}$.

The acronym GMM is an abreviation for "generalized method of moments," refering to GMM being a generalization of the classical method moments. The method of moments is based on knowing the form of up to p moments of a variable y as functions of the parameters, i.e. on

$$E[y^j] = h_j(\beta_0), (1 \leq j \leq p).$$

The method of moments estimator $\hat{\beta}$ of β_0 is obtained by replacing the population moments by sample moments and solving for $\hat{\beta}$, i.e. by solving

$$\frac{1}{n}\sum_{i=1}^{n}(y_i)^j = h_j(\hat{\beta}),\ (1 \leq j \leq p).$$

Alternatively, for

$$g_i(\beta) = (y_i - h_1(\beta), ..., y_i^p - h_p(\beta))',$$

method of moments solves $\hat{g}(\hat{\beta}) = 0$. This also means that $\hat{\beta}$ minimizes $\hat{g}(\beta)'\hat{A}\hat{g}(\beta)$ for any $\hat{A}$, so that it is a GMM estimator. GMM is more general in allowing moment functions of different form than $y_i^j - h_j(\beta)$ and in allowing for more moment functions than parameters.

One important setting where GMM applies is instrumental variables (IV) estimation. Here the model is

$$y_i = X_i'\beta_0 + \varepsilon_i, E[Z_i\varepsilon_i] = 0,$$

where Z_i is an $m \times 1$ vector of instrumental variables and X_i a $p \times 1$ vector of right-hand side variables. The condition $E[Z_i\varepsilon_i] = 0$ is often called a population "orthogonality condition" or "moment condition. "Orthogonality" refers to the elements of Z_i and ε_i being orthogonal in the expectation sense. The moment condition refers to the fact that the product of Z_i and $y_i - X_i'\beta$ has expectation zero at the true parameter. This moment condition motivates a GMM estimator where the moment functions are the vector of

products of instrumental variables and residuals, as in

$$g_i(\beta) = Z_i(y_i - X_i'\beta).$$

The GMM estimator can then be obtained by minimizing $\hat{g}(\beta)'\hat{A}\hat{g}(\beta)$.

Because the moment function is linear in parameters there is an explicit, closed form for the estimator. To describe it let $Z = [Z_1, ..., Z_n]'$, $X = [X_1, ..., X_n]'$, and $y = (y_1, ..., y_n)'$. In this example the sample moments are given by

$$\hat{g}(\beta) = \sum_{i=1}^{n} Z_i(y_i - X_i'\beta)/n = Z'(y - X\beta)/n.$$

The first-order conditions for minimization of $\hat{g}(\beta)'\hat{A}\hat{g}(\beta)$ can be written as

$$0 = X'Z\hat{A}Z'(y - X\hat{\beta}) = X'Z\hat{A}Z'y - X'Z\hat{A}Z'X\hat{\beta}.$$

These assuming that $X'Z\hat{A}Z'X$ is nonsingular, this equation can be solved to obtain

$$\hat{\beta} = (X'Z\hat{A}Z'X)^{-1}X'Z\hat{A}Z'y.$$

This is sometimes referred to as a generalized IV estimator. It generalizes the usual two stage least squares estimator, where $\hat{A} = (Z'Z)^{-1}$.

Another example is provided by the intertemporal CAPM. Let c_i be consumption at time i, R_i is asset return between i and $i+1$, α_0 is time discount factor, $u(c, \gamma_0)$ utility function, Z_i observations on variables available at time i. First-order conditions for utility maximization imply that moment restrictions satisfied for

$$g_i(\beta) = Z_i\{R_i \cdot \alpha \cdot u_c(c_{i+1}, \gamma)/u_c(c_i, \gamma) - 1\}.$$

Here GMM is nonlinear IV; residual is term in brackets. No autocorrelation because of one-step ahead decisions (c_{i+1} and R_i known at time $i+1$). Empirical Example: Hansen and Singleton (1982, Econometrica), $u(c, \gamma) = c^\gamma/\gamma$ (constant relative risk aversion), c_i monthly, seasonally adjusted nondurables (or plus services), R_i from stock returns. Instrumental variables are 1, 2, 4, 6 lags of c_{i+1} and R_i. Find γ not significantly different

than one, marginal rejection from overidentification test. Stock and Wright (2001) find weak identification.

Another example is dynamic panel data. It is a simple model that is important starting point for microeconomic (e.g. firm investment) and macroeconomic (e.g. cross-country growth) applications is

$$E^*(y_{it}|y_{i,t-1}, y_{i,t-2}, ..., y_{i0}, \alpha_i) = \beta_0 y_{i,t-1} + \alpha_i,$$

where α_i is unobserved individual effect and $E^*(\cdot)$ denotes a population regression. Let $\eta_{it} = y_{it} - E^*(y_{it}|y_{i,t-1}, ..., y_{i0}, \alpha_i)$. By orthogonality of residuals and regressors,

$$\begin{aligned} E[y_{i,t-j}\eta_{it}] &= 0, (1 \leq j \leq t, t = 1, ..., T), \\ E[\alpha_i \eta_{it}] &= 0, (t = 1, ..., T). \end{aligned}$$

Let Δ denote the first difference, i.e. $\Delta y_{it} = y_{it} - y_{i,t-1}$. Note that $\Delta y_{it} = \beta_0 \Delta y_{i,t-1} + \Delta \eta_{it}$. Then, by orthogonality of lagged y with current η we have

$$E[y_{i,t-j}(\Delta y_{it} - \beta_0 \Delta y_{i,t-1})] = 0, (2 \leq j \leq t, t = 1, ..., T).$$

These are instrumental variable type moment conditions. Levels of y_{it} lagged at least two period can be used as instruments for the differences. Note that there are different instruments for different residuals. There are also additional moment conditions that come from orthogonality of α_i and η_{it}. They are

$$E[(y_{iT} - \beta_0 y_{i,T-1})(\Delta y_{it} - \beta_0 \Delta y_{i,t-1})] = 0, (t = 2, ..., T-1).$$

These are nonlinear. Both sets of moment conditions can be combined. To form big moment vector by "stacking". Let

$$\begin{aligned} g_i^t(\beta) &= \begin{pmatrix} y_{i0} \\ \vdots \\ y_{i,t-2} \end{pmatrix} (\Delta y_{it} - \beta \Delta y_{i,t-1}), (t = 2, ..., T), \\ g_i^\alpha(\beta) &= \begin{pmatrix} \Delta y_{i2} - \beta \Delta y_{i1} \\ \vdots \\ \Delta y_{i,T-1} - \beta \Delta y_{i,T-2} \end{pmatrix} (y_{iT} - \beta y_{i,T-1}). \end{aligned}$$

These moment functions can be combined as

$$g_i(\beta) = (g_i^2(\beta)', ..., g_i^T(\beta)', g_i^\alpha(\beta)')'.$$

Here there are $T(T-1)/2+(T-2)$ moment restrictions. Ahn and Schmidt (1995, Journal of Econometrics) show that the addition of the nonlinear moment condition $g_i^\alpha(\beta)$ to the IV ones often gives substantial asymptotic efficiency improvements.

Arellano and Bond approach:

$$g_i(\beta) = \begin{pmatrix} \Delta y_{i,1} \\ \Delta y_{i,2} \\ \vdots \\ \Delta y_{i,t-1} \end{pmatrix} (y_{it} - \beta y_{i,t-1})$$

Assumes that have representation

$$y_{it} = \sum_{j=1}^{\infty} a_{tj} y_{i,t-j} + b\alpha_i.$$

Hahn, Hausman, Kuersteiner approach: Long differences

$$g_i(\beta) = \begin{pmatrix} y_{i0} \\ y_{i2} - \beta y_{i1} \\ \vdots \\ y_{i,T-1} - \beta y_{i,T-2} \end{pmatrix} (y_{iT} - y_{i1} - \beta(y_{i,T-1} - y_{i0})]$$

Has better small sample properties by getting most of the information with fewer moment conditions.

IDENTIFICATION: Identification is essential for understanding any estimator. Unless parameters are identified, no consistent estimator will exist. Here, since GMM estimators are based on moment conditions, we focus on identification based on the moment functions. The parameter value β_0 will be identified if there is a unique solution to

$$\bar{g}(\beta) = 0, \bar{g}(\beta) = E[g_i(\beta)].$$

If there is more than one solution to these moment conditions then the parameter is not identified from the moment conditions.

One important necessary order condition for identification is that $m \geq p$. When $m < p$, i.e. there are fewer equations to solve than parameters. there will typically be multiple solutions to the moment conditions, so that β_0 is not identified from the moment conditions. In the instrumental variables case, this is the well known order condition that there be more instrumental variables than right hand side variables.

When the moments are linear in the parameters then there is a simple rank condition that is necessary and sufficient for identification. Suppose that $g_i(\beta)$ is linear in β and let $G_i = \partial g_i(\beta)/\partial\beta$ (which does not depend on β by linearity in β). Note that by linearity $g_i(\beta) = g_i(\beta_0) + G_i(\beta - \beta_0)$. The moment condition is

$$0 = \bar{g}(\beta) = G(\beta - \beta_0), G = E[G_i]$$

The solution to this moment condtion occurs only at β_0 if and only if

$$rank(G) = p.$$

If $rank(G) = p$ then the only solution to this equation is $\beta - \beta_0 = 0$, i.e. $\beta = \beta_0$. If $rank(G) < p$ then there is $c \neq 0$ such that $Gc = 0$, so that for $\beta = \beta_0 + c \neq \beta_0$,

$$\bar{g}(\beta) = Gc = 0.$$

For IV $G = -E[Z_i X_i']$ so that $rank(G) = p$ is one form of the usual rank condition for identification in the linear IV seeting, that the expected cross-product matrix of instrumental variables and right-hand side variables have rank equal to the number of right-hand side variables.

In the general nonlinear case it is difficult to specify conditions for uniqueness of the solution to $\bar{g}(\beta) = 0$. Global conditions for unique solutions to nonlinear equations are not well developed, although there has been some progress recently. Conditions for local identification are more straightforward. In general let $G = E[\partial g_i(\beta_0)/\partial\beta]$. Then, assuming $\bar{g}(\beta)$ is continuously differentiable in a neighborhood of β_0 the condition $rank(G) = p$ will be sufficient for local identification. That is, $rank(G) = p$ implies that there exists a neighborhood of β_0 such that β_0 is the unique solution to $\bar{g}(\beta)$ for all β in that neighborhood.

Exact identification refers the case where there are exactly as many moment conditions as parameters, i.e. $m = p$. For IV there would be exactly as many instruments as right-hand side variables. Here the GMM estimator will satisfy $\hat{g}(\hat{\beta}) = 0$ asymptotically. When there is the same number of equations as unknowns, one can generally solve the equations, so a solution to $\hat{g}(\beta) = 0$ will exist asymptotically. The proof of this statement (due to McFadden) makes use of the first-order conditions for GMM, which are

$$0 = \left[\partial \hat{g}(\hat{\beta})/\partial\beta\right]' \hat{A}\hat{g}(\hat{\beta}).$$

The regularity conditions will require that both $\partial\hat{g}(\hat{\beta})/\partial\beta$ and $\hat{A}$ are nonsingular with probability approaching one (w.p.a.1), so the first-order conditions imply $\hat{g}(\hat{\beta}) = 0$ w.p.a.1. This will be true whatever the weight matrix, so that $\hat{\beta}$ will be invariant to the form of $\hat{A}$.

Overidentification refers to the case where there are more moment conditions than parameters, i.e. $m > p$. For IV this will mean more instruments than right-hand side variables. Here a solution to $\hat{g}(\beta) = 0$ generally will not exist, because this would solve more equations than parameters. Also, it can be shown that $\sqrt{n}\hat{g}(\hat{\beta})$ has a nondegenerate asymptotically normal distribution, so that the probabability of $\hat{g}(\hat{\beta}) = 0$ goes to zero. When $m > p$ all that can be done is set sample moments close to zero. Here the choice of $\hat{A}$ matters for the estimator, affecting its limiting distribution.

TWO STEP OPTIMAL GMM ESTIMATOR: When $m > p$ the GMM estimator will depend on the choice of weighting matrix $\hat{A}$. An important question is how to choose $\hat{A}$ optimally, to minimize the asymptotic variance of the GMM estimator. It turns out that an optimal choice of $\hat{A}$ is any such that $\hat{A} \xrightarrow{p} \Omega^{-1}$, where Ω is the asymptotic variance of $\sqrt{n}\hat{g}(\beta_0) = \sum_{i=1}^n g_i(\beta_0)/\sqrt{n}$. Choosing $\hat{A} = \hat{\Omega}^{-1}$ to be the inverse of a consistent estimator $\hat{\Omega}$ of Ω will minimize the asymptotic variance of the GMM estimator. This leads to a two-step optimal GMM estimator, where the first step is construction of $\hat{\Omega}$ and the second step is GMM with $\hat{A} = \hat{\Omega}^{-1}$.

The optimal $\hat{A}$ depends on the form of Ω. In general a central limit theorem will lead

to

$$\Omega = \lim_{n \longrightarrow \infty} E[n\hat{g}(\beta_0)\hat{g}(\beta_0)'],$$

when the limit exists. Throughout these notes we will focus on the stationary case where $E[g_i(\beta_0)g_{i+\ell}(\beta_0)']$ does not depend on i. We begin by assuming that $E[g_i(\beta_0)g_{i+\ell}(\beta_0)'] = 0$ for all positive integers ℓ. Then

$$\Omega = E[g_i(\beta_0)g_i(\beta_0)'].$$

In this case Ω can be estimated by replacing the expectation by a sample average and β_0 by an estimator $\tilde{\beta}$, leading to

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g_i(\tilde{\beta})g_i(\tilde{\beta})'.$$

The $\tilde{\beta}$ could be obtained by GMM estimator by using a choice of $\hat{A}$ that does not depend on parameter estimates. For example, for IV $\tilde{\beta}$ could be the 2SLS estimator where $\hat{A} = (Z'Z)^{-1}$.

In the IV setting this $\hat{\Omega}$ has a heteroskedasticity consistent form. Note that for $\tilde{\varepsilon}_i = y_i - X_i'\tilde{\beta}$,

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' \tilde{\varepsilon}_i^2.$$

The optimal two step GMM (or generalized IV) estimator is then

$$\hat{\beta} = (X'Z\hat{\Omega}^{-1}Z'X)^{-1}X'Z\hat{\Omega}^{-1}Z'y.$$

Because the 2SLS corresponds to a non optimal weighting matrix this estimator will generally have smaller asymptotic variance than 2SLS (when $m > p$). However, when homoskedasticity prevails, $\hat{\Omega} = \hat{\sigma}_\varepsilon^2 Z'Z/n$ is a consistent estimator of Ω, and the 2SLS estimator will be optimal. The 2SLS estimator appears to have better small sample properties also, as shown by a number of Monte Carlo studies, which may occur because using a heteroskedasticity consistent $\hat{\Omega}$ adds noise to the estimator.

When moment conditions are correlated across observations, an autocorrelation consistent variance estimator estmator can be used, as in

$$\hat{\Omega} = \hat{\Lambda}_0 + \sum_{\ell=1}^{L} w_{\ell L}(\hat{\Lambda}_\ell + \hat{\Lambda}_\ell'), \hat{\Lambda}_\ell = \sum_{i=1}^{n-\ell} g_i(\tilde{\beta})g_{i+\ell}(\tilde{\beta})'/n.$$

where L is the number of lags that are included and the weights $w_{\ell L}$ are used to ensure $\hat{\Omega}$ is positive semi-definite. A common example is Bartlett weights $w_{\ell L} = 1 - \ell/(L+1)$, as in Newey and West (1987). It is beyond the scope of these notes to suggest choices of L.

A consistent estimator $\hat{V}$ of the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_0)$ is needed for asymptotic inference. For the optimal $\hat{A} = \hat{\Omega}^{-1}$ a consistent estimator is given by

$$\hat{V} = (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}, \hat{G} = \partial \hat{g}(\hat{\beta})/\partial \beta.$$

One could also update the $\hat{\Omega}$ by using the two step optimal GMM estimator in place of $\tilde{\beta}$ in its computation. The value of this updating is not clear. One could also update the $\hat{A}$ in the GMM estimator and calculate a new GMM estimator based on the update. This iteration on $\hat{\Omega}$ appears to not improve the properties of the GMM estimator very much.

A related idea that is important is to simultaneously minimize over β in $\hat{\Omega}$ and in the moment functions. This is called the continuously updated GMM estimator (CUE). For example, when there is no autocorrelation, for $\hat{\Omega}(\beta) = \sum_{i=1}^n g_i(\beta)g_i(\beta)'/n$ the CUE is

$$\hat{\beta} = \arg\min_{\beta} \hat{g}(\beta)'\hat{\Omega}(\beta)^{-1}\hat{g}(\beta).$$

The asymptotic distribution of this estimator is the same as the two step optimal GMM estimator but it tends to have smaller bias in the IV setting, as will be discussed below. It is generally harder to compute than the two-step optimal GMM.

ASYMPTOTIC THEORY FOR GMM: We mention precise results for the i.i.d. case and give intuition for the general case. We begin with a consistency result:

If the data are i.i.d. and i) $E[g_i(\beta)] = 0$ if and only if $\beta = \beta_0$ (identification); ii) the GMM minimization takes place over a compact set B containing β_0; iii) $g_i(\beta)$ is continuous at each β with probability one and $E[\sup_{\beta \in \mathcal{B}} \|g_i(\beta)\|]$ is finite; iv) $\hat{A} \xrightarrow{p} A$ positive definite; then $\hat{\beta} \xrightarrow{p} \beta_0$.

See Newey and McFadden (1994) for the proof. The idea is that, for $g(\beta) = E[g_i(\beta)]$, by the identification hypothesis and the continuity conditions $g(\beta)'Ag(\beta)$ will be bounded

away from zero outside any neighborhood $\mathcal{N}$ of β_0. Then by the law of large numbers and iv), so will $\hat{g}(\beta)'\hat{A}\hat{g}(\beta)$. But, $\hat{g}(\hat{\beta})'\hat{A}\hat{g}(\hat{\beta}) \leq \hat{g}(\beta_0)'\hat{A}g(\beta_0) \xrightarrow{p} 0$ from the definition of $\hat{\beta}$ and the law of large numbers, so $\hat{\beta}$ must be inside $\mathcal{N}$ with probability approaching one.

The compact parameter set is not needed if $g_i(\beta)$ is linear, like for IV.

Next we give an asymptotic normality result:

If the data are i.i.d., $\hat{\beta} \xrightarrow{p} \beta_0$ and i) β_0 is in the interior of the parameter set over which minimization occurs; ii) $g_i(\beta)$ is continuously differentiable on a neighborhood N of β_0 iii) $E[\sup_{\beta\in\mathcal{N}} \|\partial g_i(\beta)/\partial\beta\|]$ is finite; iv) $\hat{A} \xrightarrow{p} A$ and $G'AG$ is nonsingular, for $G = E[\partial g_i(\beta_0)/\partial\beta]$; v) $\Omega = E[g_i(\beta_0)g_i(\beta_0)']$ exists, then

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V), V = (G'AG)^{-1}G'A\Omega AG(G'AG)^{-1}.$$

See Newey and McFadden (1994) for the proof. Here we give a derivation of the asymptotic variance that is correct even if the data are not i.i.d..

By consistency of $\hat{\beta}$ and β_0 in the interior of the parameter set, with probability approaching (w.p.a.1) the first order condition

$$0 = \hat{G}'\hat{A}\hat{g}(\hat{\beta}),$$

is satisfied, where $\hat{G} = \partial\hat{g}(\hat{\beta})/\partial\beta$. Expand $\hat{g}(\hat{\beta})$ around β_0 to obtain

$$0 = \hat{G}'\hat{A}\hat{g}(\beta_0) + \hat{G}'\hat{A}\bar{G}(\hat{\beta} - \beta_0),$$

where $\bar{G} = \partial\hat{g}(\bar{\beta})/\partial\beta$ and $\bar{\beta}$ lies on the line joining $\hat{\beta}$ and β_0, and actually differs from row to row of $\bar{G}$. Under regularity conditions like those above $\hat{G}'\hat{A}\bar{G}$ will be nonsingular w.p.a.1. Then multiplying through by $\sqrt{n}$ and solving gives

$$\sqrt{n}(\hat{\beta} - \beta_0) = -\left(\hat{G}'\hat{A}\bar{G}\right)^{-1}\hat{G}'\hat{A}\sqrt{n}\hat{g}(\beta_0).$$

By an appropriate central limit theorem $\sqrt{n}\hat{g}(\beta_0) \xrightarrow{d} N(0, \Omega)$. Also we have $\hat{A} \xrightarrow{p} A$, $\hat{G} \xrightarrow{p} G$, $\bar{G} \xrightarrow{p} G$, so by the continuous mapping theorem, $\left(\hat{G}'\hat{A}\bar{G}\right)^{-1}\hat{G}'\hat{A} \xrightarrow{p} (G'AG)^{-1}G'A$. Then by the Slutzky lemma,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} -(G'AG)^{-1}G'AN(0, \Omega) = N(0, V).$$

The fact that $A = \Omega^{-1}$ minimizes the asymptotic varince follows from the Gauss Markov Theorem. Consider a linear model.

$$E[Y] = G\delta, Var(Y) = \Omega.$$

The asymptotic variance of the GMM estimator with $A = \Omega^{-1}$ is $(G'\Omega^{-1}G)^{-1}$. This is also the variance of generalized least squares (GLS) in this model. Consider an estmator $\hat{\delta} = (G'AG)^{-1}G'AY$. It is linear and unbiased and has variance V. Then by the Gauss-Markov Theorem,

$$V - (G'\Omega^{-1}G)^{-1} \text{ is p.s.d..}$$

We can also derive a condition for A to be efficient. The Gauss-Markov theorem says that GLS is the the unique minimum variance estimator, so that A is efficient if and only if

$$(G'AG)^{-1}G'A = (G'\Omega^{-1}G)^{-1}G'\Omega^{-1}.$$

Transposing and multiplying gives

$$\Omega AG = GB,$$

where B is a nonsingular matrix. This is the condition for A to be optimal.

TESTING IN GMM: An important test statistic for GMM is the test of overidentifying restrictions that is given by

$$T = n\hat{g}(\hat{\beta})'\hat{\Omega}^{-1}\hat{g}(\hat{\beta}).$$

When the moment conditions $E[g_i(\beta_0)] = 0$ are satisfied then as the sample size grows we will have

$$T \xrightarrow{d} \chi^2(m-p).$$

Thus, a test with asymptotic level α consists of rejecting if $T \geq q$ where q is the the $1-\alpha$ quantile of a $\chi^2(m-p)$ distribution.

Under the conditions for asymptotic normality, Ω nonsinglar, and $\hat{\Omega} \xrightarrow{p} \Omega$, for an efficient GMM estimator it follows that $T \xrightarrow{d} \chi^2(m-p)$.

See Newey and McFadden (1994) for a precise proof. Here we outline the proof. Let R denote a symmetric square root of the matrix Ω (i.e. $RR = \Omega$) and let $H = R^{-1}G$. Using expansions similarly to the proof of asymptotic normality, we have

$$\begin{aligned}\sqrt{n}\hat{g}(\hat{\beta}) &= \sqrt{n}\hat{g}(\beta_0) + \bar{G}\sqrt{n}(\hat{\beta} - \beta_0) = [I - \bar{G}(\hat{G}'\hat{\Omega}^{-1}\bar{G})^{-1}\hat{G}'\hat{\Omega}^{-1}]\sqrt{n}\hat{g}(\beta_0) \\ &= [I - G(G'\Omega^{-1}G)^{-1}G'\Omega^{-1}]\sqrt{n}\hat{g}(\beta_0) + o_p(1), \\ &= R[I - H(H'H)^{-1}H']R^{-1}\sqrt{n}\hat{g}(\beta_0) + o_p(1),\end{aligned}$$

where $\bar{G}$ is given above in the discussion of asymptotic normality and $o_p(1)$ is a random variable that converges in probability to zero. Then by $R^{-1}\Omega R^{-1} = I$ we have $R^{-1}\sqrt{n}\hat{g}(\beta_0) \xrightarrow{d} N(0, I)$ so that

$$\sqrt{n}\hat{g}(\hat{\beta}) \xrightarrow{d} R[I - H(H'H)^{-1}H']U, U \sim N(0, I).$$

By consistency of $\hat{\Omega}$ and nonsingularity of Ω it follows that $\hat{\Omega}^{-1} \xrightarrow{p} \Omega^{-1}$. Then by the continuous mapping theorem $R\hat{\Omega}^{-1}R \xrightarrow{p} I$ so that by $I - H(H'H)^{-1}H'$ idempotent with rank $m - p$,

$$T \xrightarrow{d} U'(I - H(H'H)^{-1}H')U.$$

By a standard result in multivariate normal theory, $U'(I - H(H'H)^{-1}H')U$ is distributed as $\chi^2(m - p)$, giving the result.

This statistic will not detect all misspecification. It is only a test of overidentifying restrictions. Intuitively, p moments are "used up" in estimating β. In the exactly identified case where $m = p$ note that $\hat{g}(\hat{\beta}) = 0$, so there is no way to test any moment conditions.

An example is in IV estimation, where $g_i(\beta) = Z_i(y - X_i'\beta)$. There the test stastistic is

$$T = \hat{\varepsilon}'Z\hat{\Omega}^{-1}Z'\hat{\varepsilon}/n, \hat{\varepsilon} = y - X\hat{\beta}.$$

One could also use an updated $\hat{\Omega}$, based on $\hat{\varepsilon}$. This takes particular form in the independent observations case. When homoskedasticity holds and $\hat{\Omega} = \hat{\varepsilon}'\hat{\varepsilon}/T$ the test statistic is

$$T = n \cdot \hat{\varepsilon}'Z(Z'Z)^{-1}Z'\hat{\varepsilon}/\hat{\varepsilon}'\hat{\varepsilon},$$

that is nR^2 from regression of $\hat{\varepsilon}$ on Z. When $\hat{\beta}$ is the generalized IV estimator and the heteroskedasticity consistent $\hat{\Omega} = \sum_{i=1}^n Z_i Z_i' \hat{\varepsilon}_i^2 / n$ is used in forming the test statistic it is

$$T = e'\hat{r}(\hat{r}'\hat{r})^{-1}\hat{r}'e,$$

the nR^2 from regressing $e = (1, ..., 1)'$ on $\hat{r} = [\hat{\varepsilon}_1 Z_1, ..., \hat{\varepsilon}_n Z_n]'$.

It is also possible to test subsets of moment restrictions. Partition $g_i(\beta) = (g_i^1(\beta)', g_i^2(\beta)')'$ and Ω conformably. One simple test is

$$\hat{T}_1 = \min_\beta n\hat{g}(\beta)'\hat{\Omega}^{-1}\hat{g}(\beta) - \min_\beta n\hat{g}^2(\beta)'\hat{\Omega}_{22}^{-1}\hat{g}^2(\beta).$$

The asymptotic distribution of this is $\chi^2(m_1)$ where m_1 is the dimension of $\hat{g}^1(\beta)$.

Another version can be formed as

$$n\tilde{g}^{1\prime}\tilde{\Omega}^{-1}\tilde{g}^1,$$

where $\tilde{g}^1 = \hat{g}^1(\hat{\beta}) - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{g}^2(\hat{\beta})$ and $\tilde{\Omega}$ is an estimator of the asymptotic variance of $\sqrt{n}\tilde{g}^1$. If $m_1 \leq p$ then, except in any degenerate cases, a Hausman test based on the difference of the optimal two-step GMM estimator using all the moment conditions and using just $\hat{g}^2(\beta)$ will be asymptotically equivalent to this test, for any m_1 parameters. See Newey (1984, GMM Specification Testing, Journal of Econometrics.)

We can also consider tests of a null hypothesis of the form

$$H_0 : s(\beta_0) = 0,$$

where $s(\beta)$ is a $q \times 1$ vector of functions.

Let $s(\beta)$ by a $q \times 1$ vector of functions with $q < p$ and $rank(\partial s(\beta_0)/\partial\beta) = q$, $\tilde{\beta}$ be a restricted GMM estimator $\tilde{\beta} = \arg\min_{s(\beta)=0} \hat{g}(\beta)'\hat{\Omega}^{-1}\hat{g}(\beta)$, and $\tilde{G} = \partial\hat{g}(\tilde{\beta})/\partial\beta$. Under the null hypotheses $H_0 : s(\beta) = 0$ and the same conditions as for the overidentification test,

$$\begin{aligned}
Wald &: W = ns(\hat{\beta})'[\partial s(\hat{\beta})/\partial\beta(\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}\partial s(\hat{\beta})/\partial\beta]^{-1}s(\hat{\beta}) \xrightarrow{d} \chi^2(q), \\
SSR &: n\hat{g}(\tilde{\beta})'\hat{\Omega}^{-1}\hat{g}(\tilde{\beta}) - n\hat{g}(\hat{\beta})'\hat{\Omega}^{-1}\hat{g}(\hat{\beta}) - W \xrightarrow{p} 0, \\
LM &: n\hat{g}(\tilde{\beta})'\hat{\Omega}^{-1}\tilde{G}(\tilde{G}'\hat{\Omega}^{-1}\tilde{G})^{-1}\tilde{G}'\hat{\Omega}^{-1}\hat{g}(\tilde{\beta}) - W \xrightarrow{p} 0.
\end{aligned}$$

Here SSR is like sum of squared residuals, LM is Lagrange Multiplier, and W is a Wald statistic. The asymptotic approximation often more accurate for SSR and LM than W. Equivalence also holds under sequence of local alternatives.

ADDING MOMENT CONDITIONS: The optimality of the two step GMM estimator has interesting implications. One simple but useful implication is that adding moment conditions will also decrease (or at least not decrease) the asymptotic variance of the optimal GMM estimator. This occurs because the optimal weighting matrix for fewer moment conditions is not optimal for all the moment conditions. To explain further, suppose that $g_i(\beta) = (g_i^1(\beta)', g_i^2(\beta)')'$. Then the optimal GMM estimator for just the first set of moment conditions $g_i^1(\beta)$ is uses

$$\hat{A} = \begin{pmatrix} (\hat{\Omega}^1)^{-1} & 0 \\ 0 & 0 \end{pmatrix},$$

where $\hat{\Omega}^1$ is a consistent estimator of the asymptotic variance of $\sum_{i=1}^n g_i^1(\beta_0)/\sqrt{n}$. This $\hat{A}$ is not generally optimal for the entire moment function vector $g_i(\beta)$.

For example, consider the linear regression model

$$E[y_i|X_i] = X_i'\beta_0.$$

The least squares estimator is a GMM estimator with moment functions $g_i^1(\beta) = X_i(y_i - X_i'\beta)$. The conditional moment restriction implies that $E[\varepsilon_i|X_i] = 0$ for $\varepsilon_i = y_i - X_i'\beta_0$. We can add to these moment conditions by using nonlinear functions of X_i as additional "instrumental variables." Let $g_i^2(\beta) = a(X_i)(y_i - X_i'\beta)$ for some $(m-p) \times 1$ vector of functions of X_i. Then the optimal two-step estimator based on

$$g_i(\beta) = \begin{pmatrix} X_i \\ a(X_i) \end{pmatrix} (y_i - X_i'\beta)$$

will be more efficient than least squares when there is heteroskedasticity. This estimator has the form of the generalized IV estimator described above where $Z_i = (X_i', a(X_i)')'$. It will provide no efficiency gain when homoskedasticity prevails. Also, the asymptotic variance estimator $\hat{V} = (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}$ tends to provide a poor approximation to the variance of $\hat{\beta}$. See Cragg (1982, Econometrica). Interesting questions here are what and how

many functions to include in $a(X)$ and how to improve the variance estimator. Some of these issues will be further discussed below.

Another example is provided by missing data. Consider again the linear regression model, but now just assume that $E[X_i\varepsilon_i] = 0$, i.e. $X_i'\beta_0$ may not be the conditional mean. Suppose that some of the variables are sometimes missing and W_i denote the variables that are always observed. Let Δ_i denote a complete data indicator, equal to 1 if (y_i, X_i) are observed and equal to 0 if only W_i is observed. Suppose that the data is missing completely at random, so that Δ_i is independent of W_i. Then there are two types of moment conditions available. One is $E[\Delta_i X_i\varepsilon_i] = 0$, leading to a moment function of the form

$$g_i^1(\beta) = \Delta_i X_i(y_i - X_i'\beta).$$

GMM for this moment condition is just least squares on the complete data. The other type of moment condition is based on $Cov(\Delta_i, a(W_i)) = 0$ for any vector of functions $a(W)$, leading to a moment function of the form

$$g_i^2(\eta) = (\Delta_i - \eta)a(W_i).$$

One can form a GMM estimator by combining these two moment conditions. This will generally be asymptotically more efficient than least squares on the complete data when Y_i is included in W_i. Also, it turns out to be an approximately efficient estimator in the presence of missing data. As in the previous example, the choice of $a(W)$ is an interesting question.

Although adding moment conditions often lowers the asymptotic variance it may not improve the small sample properties of estimators. When endogeneity is present adding moment conditions generally increases bias. Also, it can raise the small sample variance. Below we discuss criteria that can be used to evaluate these tradeoffs.

One setting where adding moment conditions does not lower asymptotic efficiency is when those the same number of additional parameters are also added. That is, if the second vector of moment functions takes the form $g_i^2(\beta, \gamma)$ where γ has the same dimension as $g_i^2(\beta, \gamma)$ then there will be no efficiency gain for the estimator of β. This

situation is analogous to that in the linear simultaneous equations model where adding exactly identified equations does not improve efficiency of IV estimates. Here adding exactly identified moment functions does not improve efficiency of GMM.

Another thing GMM can be used for is derive the variance of two step estimators. Consider a two step estimator $\hat{\beta}$ that is formed by solving

$$\frac{1}{n}\sum_{i=1}^{n} g_i^2(\beta, \hat{\gamma}) = 0,$$

where $\hat{\gamma}$ is some first step estimator. If $\hat{\gamma}$ is a GMM estimator solving $\sum_{i=1}^{n} g_i^1(\gamma)/n = 0$ then $(\hat{\beta}, \hat{\gamma})$ is a (joint) GMM estimator for the triangular moment conditions

$$g_i(\beta, \gamma) = \begin{pmatrix} g_i^1(\gamma) \\ g_i^2(\beta, \gamma) \end{pmatrix}.$$

The asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_0)$ can be calculated by applying the general GMM formula to this triangular moment condition.

When $E[\partial g_i^2(\beta_0, \gamma_0)/\partial\gamma] = 0$ the asymptotic variance of $\hat{\beta}$ will not depend on estimation of γ, i.e. will the same as for GMM based on $g_i(\beta) = g_i^2(\beta, \gamma_0)$. A sufficient condition for this is that

$$E[g_i^2(\beta_0, \gamma)] = 0$$

for all γ in some neighborhood of γ_0. Differentiating this identity with respect to γ, and assuming that differentiation inside the expectation is allowed, gives $E[\partial g_i^2(\beta_0, \gamma_0)/\partial\gamma] = 0$. The interpretation of this is that if consistency of the first step estimator does not affect consistency of the second step estimator, the second step asymptotic variance does not need to account for the first step.

CONDITIONAL MOMENT RESTRICTIONS: Often times the moment restrictions on which GMM is based arise from conditional moment restrictions. Let $\rho_i(\beta) = \rho(w_i, \beta)$ be a $r \times 1$ residual vector. Suppose that there are some instruments z_i such that the conditional moment restrictions

$$E[\rho_i(\beta_0)|z_i] = 0$$

are satisfied. Let $F(z_i)$ be an $m \times r$ matrix of instrumental variables that are functions of z_i. Let $g_i(\beta) = F(z_i)\rho_i(\beta)$. Then by iterated expectations,

$$E[g_i(\beta_0)] = E[F(z_i)E[\rho_i(\beta_0)|z_{\beta i}]] = 0.$$

Thus $g_i(\beta)$ satisfies the GMM moment restrictions, so that one can form a GMM estimator as described above. For moment functions of the form $g_i(\beta) = F(z_i)\rho_i(\beta)$ we can think of GMM as a nonlinear instrumental variables estimator.

The optimal choice of $F(z)$ can be described as follows. Let $D(z) = E[\partial\rho_i(\beta_0)/\partial\beta|z_i = z]$ and $\Sigma(z) = E[\rho_i(\beta_0)\rho_i(\beta_0)'|z_i = z]$. The optimal choice of instrumental variables $F(z)$ is

$$F^*(z) = D(z)'\Sigma(z)^{-1}.$$

This $F^*(z)$ is optimal in the sense that it minimizes the asymptotic variance of a GMM estimator with moment functions $g_i(\beta) = F(z_i)\rho_i(\beta)$ and a weighting matrix A. To show this optimality let $F_i = F(z_i), F_i^* = F^*(z_i)$, and $\rho_i = \rho_i(\beta_0)$. Then by iterated expectations, for a GMM estimator with moment conditions $g_i(\beta) = F(z_i)\rho_i(\beta)$,

$$G = E[F_i\partial\rho_i(\beta_0)/\partial\beta] = E[F_iD(z_i)] = E[F_i\Sigma(z_i)F_i^{*\prime}] = E[F_i\rho_i\rho_i'F_i^{*\prime}].$$

Let $h_i = G'AF_i\rho_i$ and $h_i^* = F_i^*\rho_i$, so that

$$G'AG = G'AE[F_i\rho_ih_i^{*\prime}] = E[h_ih_i^{*\prime}], G'A\Omega AG = E[h_ih_i'].$$

Note that for $F_i = F_i^*$ we have $G = \Omega = E[h_i^*h_i^{*\prime}]$. Then the difference of the asymptotic variance for $g_i(\beta) = F_i\rho_i(\beta)$ and some A and the asymptotic variance for $g_i(\beta) = F_i^*\rho_i(\beta)$ is

$$\begin{aligned} & (G'AG)^{-1}G'A\Omega AG(G'AG)^{-1} - (E[h_i^*h_i^{*\prime}])^{-1} \\ = \; & (E[h_ih_i^{*\prime}])^{-1}\left\{E[h_ih_i'] - E[h_ih_i^{*\prime}]\left(E[h_i^*h_i^{*\prime}]\right)^{-1}E[h_i^*h_i']\right\}\left(E[h_i^*h_i']\right)^{-1}. \end{aligned}$$

The matrix in brackets is the second moment matrix of the population least squares projection of h_i on h_i^* and is thus positive semidefinite, so the whole matrix is positive semi-definite.

Some examples help explain the form of the optimal instruments. Consider the linear regression model $E[y_i|X_i] = X_i'\beta_0$ and let $\rho_i(\beta) = y_i - X_i'\beta$, $\varepsilon_i = \rho_i(\beta_0)$, and $\sigma_i^2 = E[\varepsilon_i^2|X_i] = \Sigma(z_i)$. Here the instruments $z_i = X_i$. A GMM estimator with moment conditions $F(z_i)\rho_i(\beta) = F(X_i)(y_i - X_i'\beta)$ is the estimator described above that will be asymptotically more efficient than least squares when $F(X_i)$ includes X_i. Here $\partial\rho_i(\beta)/\partial\beta = -X_i'$, so that the optimal instruments are

$$F_i^* = \frac{-X_i}{\sigma_i^2}.$$

Here the GMM estimator with the optimal instruments in the heteroskedasticity corrected generalized least squares.

Another example is a homoskedastic linear structural equation. Here again $\rho_i(\beta) = y_i - X_i'\beta$ but now z_i is not X_i and $E[\varepsilon_i^2|z_i] = \sigma^2$ is constant. Here $D(z_i) = -E[X_i|z_i]$ is the reduced form for the right-hand side variables. The optimal instruments in this example are

$$F_i^* = \frac{-D(z_i)}{\sigma^2}.$$

Here the reduced form may be linear in z_i or nonlinear.

For a given $F(z)$ the GMM estimator with optimal $A = \Omega^{-1}$ corresponds to an approximation to the optimal estimator. For simplicity we describe this interpretation for $r = p = 1$. Note that for $g_i = F_i\rho_i$ it follows similarly to above that $G = E[g_i h_i^{*\prime}]$, so that

$$G'\Omega^{-1} = E[h_i^* g_i'](E[g_i g_i'])^{-1}.$$

That is $G'\Omega^{-1}$ are the coefficients of the population projection of h_i^* on g_i. Thus we can interpret the first order conditions for GMM

$$0 = \hat{G}'\hat{\Omega}^{-1}\hat{g}(\hat{\beta}) = \hat{G}'\hat{\Omega}^{-1}\sum_{i=1}^n F_i\rho_i(\beta)/n,$$

can be interpreted as an estimated mean square approximation to the first order conditions for the optimal estmator

$$0 = \sum_{i=1}^n F_i^*\rho_i(\beta)/n.$$

(This holds for GMM in other models too).

One implication of this interpretation is that if the number and variety of the elements of F increases in such a way that linear combinations of F can approximate any function arbitrarily well then the asymptotic variance for GMM with optimal A will approach the optimal asymptotic variance. To show this, recall that m is the dimension of F_i and let the notation F_i^m indicate dependence on m. Suppose that for any $a(z)$ with $E[\Sigma(z_i)a(z_i)^2]$ finite there exists $m \times 1$ vectors π^m such that as $m \longrightarrow \infty$

$$E[\Sigma(z_i)\{a(z_i) - \pi^{m\prime}F_i^m\}^2] \longrightarrow 0.$$

For example, when z_i is a scalar the nonnegative integer powers of a bounded monotonic transformation of z_i will have this property. Then it follows that for $h_i^m = \rho_i F_i^{m\prime}\Omega^{-1}G$

$$\begin{aligned} E[\{h_i^* - h_i^m\}^2] &\leq E[\{h_i^* - \rho_i F_i^{m\prime}\pi^m\}^2] = E[\rho_i^2\{F_i^* - F_i^{m\prime}\pi^m\}^2] \\ &= E[\Sigma(z_i)\{F_i^* - F_i^{m\prime}\pi^m\}^2] \longrightarrow 0. \end{aligned}$$

Since h_i^m converges in mean square to h_i^*, $E[h_i^m h_i^{m\prime}] \longrightarrow E[h_i^* h_i^{*\prime}]$, and hence

$$(G'\Omega^{-1}G)^{-1} = (G'\Omega^{-1}E[g_i g_i']\Omega^{-1}G)^{-1} = (E[h_i^m h_i^{m\prime}])^{-1} \longrightarrow (E[h_i^* h_i^{*\prime}])^{-1}.$$

Because the asymptotic variance is minimzed at h_i^* the asymptotic variance will approach the lower bound more rapidly as m grows than h_i^m approaches h_i^*. In practice this may mean that it is possible to obtain quite low asymptotic variance with relatively few approximating functions in F_i^m.

An important issue for practice is the choice of m.