Notes On Method-of-Moments Estimation

James L. Powell
Department of Economics
University of California, Berkeley

Unconditional Moment Restrictions and Optimal GMM

Most estimation methods in econometrics can be recast as *method-of-moments* estimators, where the p-dimensional parameter of interest θ_0 is assumed to satisfy an *unconditional moment restriction*

$$E[m(z_i, \theta_0)] \equiv \mu(\theta) = 0 \qquad (*)$$

for some r-dimensional vector of functions $m(z_i, \theta)$ of the observable data vector z_i and possible parameter value θ in some parameter space Θ. Assuming that θ_0 is the *unique* solution of this population moment equation (equivalent to identification when only (*) is imposed), a method-of-moments estimator $\hat{\theta}$ is defined as a solution (or near-solution) of a sample analogue to (*), replacing the population expectation by a sample average.

Generally, for θ_0 to uniquely solve (*), the number of components r of the moment function $m(\cdot)$ must be at least as large as the number of components p in θ – that is, $r \geq p$, known as the "order condition" for identification. When θ_0 is identified and $r = p$ – termed "just identification" – a natural analogue of the population moment equation for θ_0 defines the method-of-moment estimator as the solution to the p-dimensional sample moment equation

$$\begin{aligned} \bar{m}(\hat{\theta}) &\equiv \frac{1}{n}\sum_{\iota=1}^{n} m(z_i, \hat{\theta}) \\ &= 0, \end{aligned} \qquad (**)$$

where $z_1, ..., z_n$ are all assumed to satisfy (*). The simplest setting, assumed hereafter, is that $\{z_i\}$ is a random sample (i.e., z_i is i.i.d), but this is hardly necessary; the $\{z_i\}$ can be dependent and/or have heterogeneous distributions, provided an "ergodicity" result $\bar{m}\,(\theta) - E[\bar{m}(\theta)]$ can be established.

Examples of estimators in this class include the maximum likelihood estimator (with $m(z_i, \theta)$ the "score function," i.e., derivative of the log density of z_i with respect to θ for an i.i.d. sample) and the classical least squares estimator (with $z_i \equiv (y_i, x_i')'$ and $m(z_i, \theta) = (y_i - x_i'\theta)x_i$, the product of the residuals and regressors). Another example is the *instrumental variables* estimator for the linear model

$$y_i = w_i'\theta_0 + \varepsilon_i,$$

where y_i and $w_i \in R^p$ are subvectors of z_i and the error term ε_i is assumed to be orthogonal to some other subvector $x_i \in R^r$ of z_i, i.e.,

$$E[\varepsilon_i x_i] = E[(y_i - w_i'\theta_0)x_i] = 0.$$

When $r = p$ – i.e., the number of "instrumental variables" x_i equals the number of right-hand-side regressors w_i – then the *instrumental variables estimator*

$$\hat{\theta} = \left[\frac{1}{n}\sum_{\iota=1}^{n} x_i w_i'\right]^{-1} \frac{1}{n}\sum_{\iota=1}^{n} x_i y_i$$

is the solution to (**) when $m(z_i, \theta) = (y_i - w_i'\theta)x_i$.

Returning to the general moment condition (*), if $r > p$ – termed "overidentification" of θ_0 – the system of equations $\bar{m}(\theta) = 0$ is overdetermined, and in general no solution of this sample analogue to (*) will exist. In this case, an analogue estimator can be defined to make $\bar{m}(\theta)$ "close to zero," by defining

$$\hat{\theta} = \arg\min_{\Theta} S_n(\theta),$$

where $S_n(\theta)$ is a quadratic form in the sample moment function $\bar{m}(\theta)$,

$$S_n(\theta) \equiv [\bar{m}(\theta)]' A_n \bar{m}(\theta),$$

and A_n some non-negative definite, symmetric "weight matrix," assumed to converge in probability to some limiting value A_0, i.e.,

$$A_n \rightarrow^p A_0.$$

Here $\hat{\theta}$ is called a *generalized method of moments (GMM)* estimator, with large-sample properties that will depend upon the limiting weight matrix A_0. Examples of possible (nonstochastic) weight matrices are $A_n = I_r$, an $r \times r$ identity matrix – which yields $S_n(\theta) = ||\bar{m}(\theta)||^2$ – or

$$A_n = \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix},$$

for which the estimator $\hat{\theta}$ sets the first p components of $\bar{m}(\hat{\theta})$ equal to zero. More generally, A_n will have estimated components; once the asymptotic (normal) distribution of $\hat{\theta}$ is derived for a given value of A_0, the optimal choice of A_0 (to minimize the asymptotic variance) can be determined, and a feasible efficient estimator can be constructed if this optimal weight matrix can be consistently estimated.

The consistency theory for $\hat{\theta}$ is standard for extremum estimators: the first step is to demonstrate uniform consistency of $S_n(\theta)$ to its probability limit

$$S(\theta) \equiv [\mu(\theta)]' A_0 \mu(\theta),$$

that is,

$$\sup_{\Theta} |S_n(\theta) - S(\theta)| \rightarrow^p 0,$$

and then to establish that the limiting minimand $S(\theta)$ is uniquely minimized at $\theta = \theta_0$, which follows if

$$A_0^{1/2}\mu(\theta) \neq 0 \qquad if \qquad \theta \neq \theta_0,$$

where $A_0^{1/2}$ is any square root of the weight matrix A_0. Establishing both the uniform convergence of the minimand S_n to its limit S and uniqueness of θ_0 as the minimizer of S will require primitive assumptions on the distribution of z_i, the form of the moment function $m(\cdot)$, and the limiting weight matrix A_0 which vary with the particular problem.

Among the standard "regularity conditions" on the moment function $m(\cdot)$ is an assumption that it is "smooth" (i.e., continuously differentiable) in θ; then, if θ_0 is assumed to be in the interior of the parameter space Θ, then with probability approaching one the consistent GMM estimator $\hat{\theta}$ will satisfy a first-order condition for minimization of S,

$$\begin{aligned} 0 &= \frac{\partial S_n(\hat{\theta})}{\partial \theta} \\ &= 2\left[\frac{\partial \bar{m}(\hat{\theta})}{\partial \theta'}\right]' A_n \bar{m}(\hat{\theta}). \end{aligned}$$

If the derivative of the average moment function $\bar{m}(\theta)$ converges uniformly in probability to its expectation in a neighborhood of θ_0 (which must be established in the usual way), then consistency of $\hat{\theta}$ implies that

$$\left[\frac{\partial \bar{m}(\hat{\theta})}{\partial \theta'}\right] \rightarrow^p M_0 \equiv \left[\frac{\partial \mu(\theta_0)}{\partial \theta'}\right].$$

This, plus convergence in probability of A_n to A_0, means that the first-order condition can be rewritten as

$$0 = M_0' A_0 \bar{m}(\hat{\theta}) + o_p(\bar{m}(\hat{\theta})).$$

Inserting the usual Taylor's series expansion of $\bar{m}(\hat{\theta})$ around the true value θ_0,

$$\bar{m}(\hat{\theta}) = \bar{m}(\theta_0) + \left[\frac{\partial \bar{m}(\hat{\theta})}{\partial \theta'}\right](\hat{\theta} - \theta_0) + o_p(||\hat{\theta} - \theta_0||),$$

yields

$$\begin{aligned} 0 &= M_0' A_0 \left[\bar{m}(\theta_0) + \left[\frac{\partial \bar{m}(\hat{\theta})}{\partial \theta'}\right](\hat{\theta} - \theta_0) + o_p(||\hat{\theta} - \theta_0||)\right] + o_p(\bar{m}(\hat{\theta})) \\ &\equiv M_0' A_0 \bar{m}(\theta_0) + M_0' A_0 M_0 (\hat{\theta} - \theta_0) + r_n, \end{aligned}$$

where r_n is a generic remainder term. Assuming it can be verified that

$$r_n = o_p\left(\frac{1}{\sqrt{n}}\right)$$

by the usual methods, the normalized difference between the estimator $\hat{\theta}$ and the true value θ_0 has the asymptotically-linear representation

$$\sqrt{n}(\hat{\theta} - \theta_0) = [M_0' A_0 M_0]^{-1} M_0' A_0 \cdot \sqrt{n}\bar{m}(\theta_0) + o_p(1).$$

But $\sqrt{n}\bar{m}(\theta_0)$ is a normalized sample average of mean-zero, i.i.d. random vectors $m(z_i, \theta_0)$, so by the Lindeberg-Levy central limit theorem,

$$\sqrt{n}\bar{m}(\theta_0) \rightarrow^d \mathcal{N}(0, V_0),$$

where

$$\begin{aligned} V_0 &\equiv Var[m(z_i, \theta_0)] \\ &= E[m(z_i, \theta_0) m(z_i, \theta_0)'], \end{aligned}$$

and thus

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d \mathcal{N}(0, [M_0' A_0 M_0]^{-1} M_0' A_0 V_0 A_0 M_0 [M_0' A_0 M_0]^{-1}),$$

which has a rather ungainly looking expression for the asymptotic covariance matrix.

By definition, an efficient choice of limiting weight matrix A_0 will minimize the asymptotic covariance matrix of $\hat{\theta}$ (in a positive semi-definite sense). The same proof as for the Gauss-Markov theorem can be used to show that this product of matrices will be minimized by choosing A_0 to make the "middle matrix" $M_0' A_0 V_0 A_0 M_0$ equal to an "outside matrix" $M_0' A_0 M_0$ being inverted. That is,

$$[M_0' A_0 M_0]^{-1} M_0' A_0 V_0 A_0 M_0 [M_0' A_0 M_0]^{-1} \geq [M_0' V_0^{-1} M_0]^{-1},$$

where the inequality means the difference in the two matrices is positive semi-definite; equality is obviously achieved if A_0 is chosen as

$$A_0^* \equiv V_0^{-1} = [Var[m(z_i, \theta_0)]]^{-1},$$

up to a (positive) constant of proportionality.

A feasible version of the optimal GMM estimator requires a consistent estimator of the covariance matrix V_0. This can be obtained in two steps: first, by calculation of a non-optimal estimator $\hat{\theta}$ using an arbitrary sequence A_n for which $\hat{\theta}$ is consistent (e.g., $A_n = I_r$), and then by construction of a sample analogue to V_0,

$$\hat{V} \equiv \frac{1}{n}\sum_{\iota=1}^{n} m(z_i, \hat{\theta}) \left[m(z_i, \hat{\theta})\right]'.$$

The resulting *optimal GMM estimator* $\hat{\theta}^*$ will have asymptotic distribution

$$\sqrt{n}(\hat{\theta}^* - \theta_0) \rightarrow^d \mathcal{N}(0, [M_0' V_0^{-1} M_0]^{-1}),$$

and its asymptotic covariance matrix is consistently estimated by $[\hat{M}'\hat{V}^{-1}\hat{M}]^{-1}$, where

$$\hat{M} \equiv \frac{1}{n}\sum_{\iota=1}^{n} \frac{\partial m(z_i, \hat{\theta}^*)}{\partial \theta'}.$$

Inference on θ_0 can then be based upon the usual large-sample normal theory.

For the example of the linear model with endogenous regressors,

$$\begin{aligned} y_i &= w_i'\theta_0 + \varepsilon_i, \\ 0 &= E[\varepsilon_i x_i] = E[(y_i - w_i'\theta_0)x_i], \end{aligned}$$

the relevant matrices for the asymptotic distribution of $\hat{\theta}^*$ are

$$\begin{aligned} M_0 &= E\left[\frac{\partial[(y_i - w_i'\theta_0)x_i]}{\partial \theta'}\right] \\ &= E\left[x_i w_i'\right] \end{aligned}$$

and

$$\begin{aligned} V_0 &= Var[(y_i - w_i'\theta_0)x_i] \\ &= E[(y_i - w_i'\theta_0)^2 x_i x_i']. \end{aligned}$$

The first step in efficient estimation of θ_0 might be based upon the (inefficient) *two-stage least squares (2SLS)* estimator

$$\begin{aligned} \hat{\theta} &= \left(\left[\frac{1}{n}\sum_{\iota=1}^{n} w_i x_i'\right]\left[\frac{1}{n}\sum_{\iota=1}^{n} x_i x_i'\right]^{-1}\left[\frac{1}{n}\sum_{\iota=1}^{n} x_i w_i'\right]\right)^{-1} \\ &\quad \cdot \left[\frac{1}{n}\sum_{\iota=1}^{n} w_i x_i'\right]\left[\frac{1}{n}\sum_{\iota=1}^{n} x_i x_i'\right]^{-1}\left[\frac{1}{n}\sum_{\iota=1}^{n} x_i y_i\right], \\ &\equiv (\hat{M}' A_n \hat{M})^{-1}\hat{M}' A_n \left[\frac{1}{n}\sum_{\iota=1}^{n} x_i y_i\right] \end{aligned}$$

which is a GMM estimator using $m(z_i, \theta) \equiv (y_i - w_i'\theta)x_i$,

$$\hat{M} \equiv \left[\frac{1}{n}\sum_{\iota=1}^{n} x_i w_i'\right]$$

and

$$A_n \equiv \left[\frac{1}{n}\sum_{\iota=1}^{n} x_i x_i'\right]^{-1}.$$

With this preliminary, $\sqrt{n}$-consistent estimator of θ_0, the efficient weight matrix is consistently estimated as

$$\hat{V}^{-1} \equiv \left[\frac{1}{n}\sum_{\iota=1}^{n}(y_i - w_i'\hat{\theta})^2 x_i x_i'\right]^{-1},$$

and the efficient GMM estimator is

$$\hat{\theta}^* \equiv (\hat{M}'\hat{V}^{-1}\hat{M})^{-1}\hat{M}'\hat{V}^{-1}\left[\frac{1}{n}\sum_{\iota=1}^{n} x_i y_i\right],$$

which has the approximate normal distribution

$$\hat{\theta}^* \overset{A}{\sim} \mathcal{N}\left(\theta_0, \frac{1}{n}(\hat{M}'\hat{V}^{-1}\hat{M})^{-1}\right).$$

If the error terms $\varepsilon_i \equiv y_i - w_i'\theta_0$ happen to be homoskedastic,

$$\begin{aligned} Var[\varepsilon_i|x_i] &\equiv \sigma^2(x_i) \\ &= \sigma_0^2, \end{aligned}$$

then

$$\begin{aligned} V_0 &\equiv E[\varepsilon_i^2 x_i x_i'] \\ &= \sigma_0^2 E[x_i x_i'] \\ &= \sigma_0^2 \text{ plim } A_n, \end{aligned}$$

and the 2SLS estimator $\hat{\theta}$ would be asymptotically efficient, and asymptotically equivalent to the efficient GMM estimator $\hat{\theta}^*$.

Conditional Moment Restrictions and Efficient Instrumental Variables

Now consider the case when a stronger *conditional moment restriction*

$$0 = E[u(z_i, \theta_0)|x_i] \equiv E[u_i|x_i],$$

where $u(z_i, \theta)$ is some q-dimensional vector of known functions of the (i.i.d.) random vector z_i and $\theta \in \Theta \subset R^p$. (Since $E[u_i|x_i]$ is a random variable, we interpret such equalities as holding with probability one, here and throughout.) Such moment restrictions can sometimes be derived as consequences of expected utility maximization; more generally, they are often imposed on additive error terms in structural models. For instance, for the linear equation

$$y_i = w_i'\theta_0 + \varepsilon_i,$$

a common assumption is that the error terms ε_i have conditional mean zero given the instrumental variables x_i,

$$E[\varepsilon_i|x_i] = 0,$$

in which case the moment function $u(\cdot)$ is just the residual function $u(z_i, \theta) = y_i - w_i'\theta$, with $u(z_i, \theta_0) \equiv \varepsilon_i$. Here $q = 1$, which is generally less than p, the number of components of θ_0 to be estimated.

Assuming the function $u(\cdot)$ is bounded above (on Θ) by some square-integrable function, i.e.,

$$\sup_{\Theta} ||u(z_i, \theta)|| \leq b(z_i), \qquad E[b(z_i)]^2 < \infty,$$

it follows (by iterated expectations) that an unconditional moment restriction

$$\begin{aligned} 0 &= E[h(x_i)u(z_i, \theta_0)] \\ &\equiv E[m(z_i, \theta_0)] \end{aligned} \tag{***}$$

holds, where $h(.)$ is any $r \times q$ matrix of functions of x_i with

$$E[||h(x_i)||^2] \equiv E[tr\{h(x_i)[h(x_i)]'\}] < \infty.$$

We can think of each column of $h(x_i)$ as a vector of "instrumental variables" for the corresponding component of $u(z_i, \theta_0)$, whose products are added together to obtain the (unconditional) moment function $m(\cdot)$. While the dimension q of the conditional moment function $u(\cdot)$ needs not be as large as the number of parameters p, the number of rows r of the matrix of instrumental variables $h(x_i)$ must be no smaller than p if estimation of θ_0 is to be based upon the implied unconditional moment restriction $0 \equiv E[m(z_i, \theta_0)] = E[h(x_i)u(z_i, \theta_0)]$.

For a *given* choice of instrument matrix $h(x_i)$, the theory for unconditional moment restrictions above can be applied to determine the form and asymptotic distribution of the optimal GMM estimator $\hat{\theta}^* = \hat{\theta}^*(h)$; that is, the optimal estimator is

$$\begin{aligned} \hat{\theta}^* &= \arg\min_{\Theta} [\bar{m}(\theta)]'\hat{V}^{-1}\bar{m}(\theta) \\ &= \arg\min_{\Theta} [\bar{m}(\theta)]'\hat{V}^{-1}\bar{m}(\theta), \end{aligned}$$

where now

$$\bar{m}(\theta) \equiv \frac{1}{n}\sum_{\iota=1}^{n} h(x_i)u(z_i, \theta)$$

and

$$\begin{aligned} \text{plim}\, \hat{V} &\equiv V_0 \\ &\equiv Var[h(x_i)u(z_i, \theta_0)] \\ &= E[h(x_i)u(z_i, \theta_0)[u(z_i, \theta_0)]'[h(x_i)]'] \\ &= E[h(x_i)\Sigma(x_i)[h(x_i)]'], \end{aligned}$$

for

$$\begin{aligned} \Sigma(x_i) &\equiv Var(u(z_i, \theta_0)|x_i) \\ &= E[u(z_i, \theta_0)[u(z_i, \theta_0)]'|x_i]. \end{aligned}$$

The asymptotic distribution of $\hat{\theta}^*$ is thus

$$\sqrt{n}(\hat{\theta}^* - \theta_0) \rightarrow^d \mathcal{N}(0, [M_0' V_0^{-1} M_0]^{-1}),$$

where

$$\begin{aligned} M_0 &\equiv E\left[\frac{\partial m(z_i, \theta_0)}{\partial \theta'}\right] \\ &= E\left[h(x_i)\frac{\partial u(z_i, \theta_0)}{\partial \theta'}\right]. \end{aligned}$$

In terms of the function $h(x_i)$, the asymptotic covariance matrix of $\hat{\theta}^*$ is

$$[M_0' V_0^{-1} M_0]^{-1} = \left(E\left[h(x_i)\frac{\partial u(z_i, \theta_0)}{\partial \theta'}\right]' \cdot \left[E[h(x_i)\Sigma(x_i)[h(x_i)]']\right]^{-1} \cdot E\left[h(x_i)\frac{\partial u(z_i, \theta_0)}{\partial \theta'}\right]\right)^{-1}.$$

To find the *best* choice of instrumental variable matrix $h(x_i)$ across all possible square-integrable functions of the conditioning variables x_i, we would minimize this matrix over $h(x_i)$. By the same Gauss-Markov-type argument as for the optimal GMM estimator, the best choice $h^*(x_i)$ will equate the "inner matrix" $E[h(x_i)\Sigma(x_i)[h(x_i)]']$ with the "outer matrix" $E\left[h(x_i)\partial u(z_i, \theta_0)/\partial \theta'\right]$ (and its transpose). By inspection, this happens when

$$\begin{aligned} h^*(x_i) &= E\left[\frac{\partial u(z_i, \theta_0)}{\partial \theta'}|x_i\right]' \cdot [\Sigma(x_i)]^{-1} \\ &\equiv D(x_i)' \cdot [\Sigma(x_i)]^{-1}. \end{aligned}$$

So in this case the asymptotic covariance matrix reduces to

$$\begin{aligned} &\left[E\left(E\left[h^*(x_i)\frac{\partial u(z_i, \theta_0)}{\partial \theta'}\right]' \cdot \left[E[h^*(x_i)\Sigma(x_i)[h^*(x_i)]']\right]^{-1} \cdot E\left[h^*(x_i)\frac{\partial u(z_i, \theta_0)}{\partial \theta'}\right]\right)\right]^{-1} \\ = &\left[E\left(D(x_i)'[\Sigma(x_i)]^{-1}D(x_i)\right)\right]^{-1} \\ = &\left[E\left(E\left[\frac{\partial u(z_i, \theta_0)}{\partial \theta'}|x_i\right]' \cdot [\Sigma(x_i)]^{-1} E\left[\frac{\partial u(z_i, \theta_0)}{\partial \theta'}|x_i\right]\right)\right]^{-1}. \end{aligned}$$

This formula looks very similar to the form of the asymptotic covariance matrix $[M_0' V_0^{-1} M_0]^{-1}$ for GMM estimation with unconditional moment restrictions, except that the expected derivative and variance matrices M_0 and V_0 are replaced by their "conditional" analogues $D(x_i)$ and $\Sigma(x_i)$, and the product $D(x_i)'[\Sigma(x_i)]^{-1}D(x_i)$ is averaged over x_i before being inverted.

Again returning to the example of the linear model with endogenous regressors,

$$\begin{aligned} y_i &= w_i'\theta_0 + \varepsilon_i, \\ 0 &= E[\varepsilon_i x_i] = E[(y_i - w_i'\theta_0)x_i], \end{aligned}$$

here $q = 1$,

$$\begin{aligned} D(x_i) &\equiv E\left[\frac{\partial u(z_i, \theta_0)}{\partial \theta'}|x_i\right] \\ &= E\left[\frac{\partial (y_i - w_i'\theta_0)}{\partial \theta'}|x_i\right] \\ &= -E[w_i'|x_i] \end{aligned}$$

and

$$\begin{aligned}\Sigma(x_i) &\equiv \sigma^2(x_i) \\ &= Var((y_i - w_i'\theta_0)|x_i) \\ &\equiv Var(\varepsilon_i|x_i).\end{aligned}$$

In the special case with $w_i = x_i$ (i.e., all regressors are exogenous), $D(x_i) = x_i'$, and the optimal sample moment condition for the restriction $E[\varepsilon_i|x_i] = 0$ is the first-order condition for weighted LS estimation, with weights $1/\sigma^2(x_i)$ inversely proportional to the conditional variance of the errors.

Global Optimality of GMM