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*

$$E[m(w_i, \theta_0)] \equiv \mu(\theta) = 0 \tag{*}$$

for some q-dimensional vector of functions $m(w_i, \theta)$ of the observable data vector w_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 q of the moment function $m(\cdot)$ must be at least as large as the number of components p in θ – that is, $q \geq p$, known as the "order condition" for identification. When θ_0 is identified and $q = p$ – 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

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

where $w_1, ..., w_n$ are all assumed to satisfy (*). The simplest setting, assumed hereafter, is that $\{w_i\}$ is a random sample (i.e., w_i is i.i.d), but this is hardly necessary; the $\{w_i\}$ can be dependent and/or have heterogeneous distributions, provided an "ergodicity" result $\bar{m}\,(\theta) - E[\bar{m}(\theta)] \xrightarrow{p} 0$ can be established.

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

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

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

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

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

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

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

Returning to the general moment condition (*), if $q > p$ – 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.,

$$A_n \to^p A_0.$$

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 $A_n = I_q$, an $q \times q$ 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 w_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

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

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

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

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 $m(w_i, \theta_0)$, so by the Lindeberg-Levy central limit theorem,

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

where

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

and thus

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

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

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

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., $A_n = I_q$), and then by construction of a sample analogue to V_0,

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

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

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

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

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

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,

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

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

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

and

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

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

(2SLS) estimator

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

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

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

and

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

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

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

and the efficient GMM estimator is

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

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 $\varepsilon_i \equiv y_i - x_i'\theta_0$ happen to be homoskedastic,

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

then

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

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