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 \( \theta_0 \) is assumed to satisfy an unconditional moment restriction

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

for some \( r \)-dimensional vector of functions \( m(z_i, \theta) \) of the observable data vector \( z_i \) and possible parameter value \( \theta \) in some parameter space \( \Theta \). Assuming that \( \theta_0 \) 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 \( \theta_0 \) 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 \( \theta \) — that is, \( r \geq p \), known as the “order condition” for identification. When \( \theta_0 \) is identified and \( r = p \) — termed “just identification” — a natural analogue of the population moment equation for \( \theta_0 \) defines the method-of-moment estimator as the solution to the \( p \)-dimensional sample moment equation

\[
\bar{m}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} m(z_i, \hat{\theta}) \tag{**})
\]

where \( z_1, \ldots, z_n \) 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 \( \bar{m}(\theta) - E[\bar{m}(\theta)] \xrightarrow{P} 0 \) 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 \( \theta \) 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 $w_i \in \mathbb{R}^p$ are subvectors of $z_i$ and the error term $\varepsilon_i$ is assumed to be orthogonal to some other subvector $x_i \in \mathbb{R}^p$ 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

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

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 $\theta_0$ – 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 $A_n = I_r$, 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 $\bar{m}(\theta)$ 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

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

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 $\theta_0$ 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 $\theta$; then, if $\theta_0$ is assumed to be in the interior of the parameter space $\Theta$, then with probability approaching one the consistent GMM estimator $\hat{\theta}$ will satisfy a first-order condition for minimization of $S$.

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

If the derivative of the average moment function $\hat{m}(\theta)$ converges uniformly in probability to its expectation in a neighborhood of $\theta_0$ (which must be established in the usual way), then consistency of $\hat{\theta}$ implies that

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

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

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

Inserting the usual Taylor’s series expansion of $\hat{m}(\hat{\theta})$ around the true value $\theta_0$,

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

3
yields
\[ 0 = M'_0 A_0 \left[ \bar{m}(\theta_0) + \left[ \frac{\partial \bar{m}(\hat{\theta})}{\partial \theta'} \right] (\hat{\theta} - \theta_0) + o_p(||\hat{\theta} - \theta_0||) \right] + o_p(\bar{m}(\hat{\theta})) \]
\[ \equiv M'_0 A_0 \bar{m}(\theta_0) + M'_0 A_0 M_0 (\hat{\theta} - \theta_0) + r_n, \]
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 \( \theta_0 \) 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,
\[ \sqrt{n} \bar{m}(\theta_0) \rightarrow^d N(0, V_0), \]
where
\[ V_0 \equiv \text{Var}[m(z_i, \theta_0)] = E[m(z_i, \theta_0)m(z_i, \theta_0)^\prime], \]
and thus
\[ \sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d N(0, [M'_0 A_0 M_0]^{-1} M'_0 A_0 V_0 A_0 M_0 [M'_0 A_0 M_0]^{-1}), \]
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
\[ A_0^* = V_0^{-1} = [\text{Var}[m(z_i, \theta_0)]]^{-1}, \]
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 $V_0$,

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

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

$$\sqrt{n}(\hat{\theta}^* - \theta_0) \overset{d}{\to} N(0, [M'_0 V_0^{-1} M_0]^{-1}) ,$$

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

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

Inference on $\theta_0$ can then be based upon the usual large-sample normal theory.

For the example of the linear model with endogenous regressors,

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

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

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

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

$$= E \left[ x_i w_i' \right]$$

and

$$V_0 = Var[(y_i - w_i' \theta_0) x_i]$$

$$= E[(y_i - w_i' \theta_0)^2 x_i x_i'] .$$

The first step in efficient estimation of $\theta_0$ might be based upon the (inefficient) two-stage least squares
\[(2SLS)\] estimator

\[
\hat{\theta} = \left( \left[ \frac{1}{n} \sum_{i=1}^{n} w_i x_i' \right] \left[ \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right]^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} x_i w_i' \right] \right)^{-1}
\cdot \left[ \frac{1}{n} \sum_{i=1}^{n} w_i x_i' \right] \left[ \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right]^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} x_i y_i \right],
\]

\[
\equiv (\hat{M}' A_n \hat{M})^{-1} \hat{M}' A_n \left[ \frac{1}{n} \sum_{i=1}^{n} x_i y_i \right]
\]

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

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

and

\[
A_0 \equiv \left[ \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right]^{-1}.
\]

With this preliminary, \(\sqrt{n}\)-consistent estimator of \(\theta_0\), the efficient weight matrix is consistently estimated as

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

and the efficient GMM estimator is

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

which has the approximate normal distribution

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

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

\[
\text{Var}[\varepsilon_i | x_i] \equiv \sigma^2(x_i)
\]

\[
= \sigma^2_0,
\]

then

\[
V_0 \equiv E[\varepsilon_i^2 x_i x_i']
\]

\[
= \sigma^2_0 E[x_i x_i']
\]

\[
= \sigma^2_0 \text{ plim } A_n,
\]
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

$$y_i = w_i^t \theta_0 + \varepsilon_i,$$

a common assumption is that the error terms $\varepsilon_i$ 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 $u(z_i, \theta) = y_i - w_i^t \theta$, with $u(z_i, \theta_0) \equiv \varepsilon_i$. Here $q = 1$, which is generally less than $p$, the number of components of $\theta_0$ to be estimated.

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

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

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

$$0 = E[h(x_i)u(z_i, \theta_0)] \equiv E[m(z_i, \theta_0)]$$

holds, where $h(\cdot)$ 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.
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 $\theta_0$ 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

$$
\hat{\theta}^* = \arg \min_{\Theta} \left[ \hat{\mu}(\theta)' \hat{V}^{-1} \hat{\mu}(\theta) \right]
$$

where now

$$
\hat{\mu}(\theta) = \frac{1}{n} \sum_{i=1}^{n} h(x_i)u(z_i, \theta)
$$

and

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

for

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

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

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

where

$$
M_0 \equiv E \left[ \frac{\partial m(z_i, \theta_0)}{\partial \theta} \right] = E \left[ h(x_i) \frac{\partial u(z_i, \theta_0)}{\partial \theta} \right].
$$
In terms of the function \( h(x_i) \), the asymptotic covariance matrix of \( \hat{\theta}^* \) is

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

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

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

\[
\equiv D(x_i)' \cdot \Sigma(x_i)^{-1}.
\]

So in this case the asymptotic covariance matrix reduces to

\[
\left[ E \left( E \left[ h^*(x_i) \frac{\partial u(z_i, \theta_0)}{\partial \theta'} \right] \right)' \cdot \left[ E[h^*(x_i) \Sigma(x_i)|h^*(x_i)] \right]^{-1} \cdot E \left[ h^*(x_i) \frac{\partial u(z_i, \theta_0)}{\partial \theta'} \right] \right]^{-1}
\]

\[
= \left[ E \left( D(x_i)' \Sigma(x_i)^{-1} D(x_i) \right) \right]^{-1}
\]

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

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,

\[
y_i = w'_i \theta_0 + \varepsilon_i,
\]

\[
0 = E[\varepsilon_i x_i] = E[(y_i - w'_i \theta_0) x_i],
\]

here \( q = 1 \),

\[
D(x_i) = E \left[ \frac{\partial u(z_i, \theta_0)}{\partial \theta'} | x_i \right]
\]

\[
= E \left[ \frac{\partial (y_i - w'_i \theta_0)}{\partial \theta'} | x_i \right]
\]

\[
= -E[w'_i | x_i]
\]
and

\[ \Sigma(x_i) \equiv \sigma^2(x_i) \]
\[ = Var((y_i - w_i\theta_0)|x_i) \]
\[ \equiv Var(\varepsilon_i|x_i). \]

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**