17: The Generalized Method of Moments
ECON 837

Brian Krauth (adapted from notes by Simon Woodcock), Spring 2010

Generalized Method of Moments (GMM) estimators are extremely popular among applied researchers. GMM is a very flexible estimation framework that yields consistent and asymptotically normal parameter estimates under minimal assumptions, and that can be applied to a wide range of models. It does not require specifying a parametric joint distribution for the data (as is the case for maximum likelihood), and this is one of the main reasons for its popularity. However nothing in life is free, and the price of making fewer assumptions than ML is that you lose asymptotic efficiency.

As the name implies, GMM generalizes a conceptually simple estimation framework known as the method of moments (MM). The simple idea behind MM estimation is to set sample moments equal to their population counterparts, and solve for parameters of interest. The simplest example is estimating the mean of a random sample $Y_1, Y_2, ..., Y_n$. If we assume $E[Y_i] = \mu$ for all i, then our **population moment condition** is

$$E[Y_i - \mu] = 0.$$

The sample analog of this is the **sample moment condition**:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0. \tag{1}$$

The MM estimator $\hat{\mu}$ solves (1). Since (1) is just a sample mean, we know by Khinchine's WLLN and the Slutsky Theorem that $\hat{\mu}$ is consistent for μ. Likewise, by the CLT and delta-method, it has an asymptotically normal distribution.

Another simple example is given by least squares regression. Here, we have a convenient set of population moment conditions (sometimes called **orthogonality conditions** in this case, for obvious reasons):

$$E[\mathbf{x}_i\varepsilon_i] = E[\mathbf{x}_i(y_i - \mathbf{x}_i'\beta)] = \mathbf{0}.$$

The sample analog is, of course,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left(y_i - \mathbf{x}_i'\hat{\beta}_{GMM}\right) = \mathbf{0}. \tag{2}$$

Notice that the k equations in (2) are just the least squares normal equations. Hence the least squares estimator is a MM estimator. The IV estimator will also turn out to be an MM estimator.

GMM generalizes the method of moments to the case where the parameters of interest are overidentified. The over-identified case is when we have more moment conditions than parameters to estimate. We have seen overidentification before in the context of 2SLS estimation, and in fact we can show (if time permits) that 2SLS is also a GMM estimator.

The most important practical application of GMM is in estimating dynamic models.

Hansen-Singleton

The classic example of a GMM application is Hansen and Singleton's (1982) estimation of the textbook representative agent asset pricing model.

The setup is a standard macroeconomic model in which a representative agent chooses consumption and asset holdings to maximize utility:

$$U = \sum_{t=0}^{\infty} \beta^t u(C_t)$$

subject to a sequence of budget constraints:

$$\begin{aligned} C_t + \sum_{a=1}^{A} Q_{a,t+1} &\leq \sum_{a=1}^{A} R_{a,t} Q_{a,t} + Y_t \\ Q_{a,t+1} &\geq 0 \qquad \forall i, t \\ Q_{a,0} & \quad \text{given} \end{aligned}$$

where $Q_{a,t}$ is the agent's holding of asset a at the beginning of time period t, Y_t is the agent's labor income, and $R_{a,t}$ is the gross return on asset a. Both Y_t and $R_{a,t}$ are stochastic.

The Euler equations for this model are:

$$E\left(\beta \frac{u'(C_{t+1})}{u'(C_t)} R_{a,t+1} \middle| I_t \right) = 1 \qquad \forall a$$

where I_t is the set of all information available to the agent at time t (i.e., when choosing $Q_{a,t+1}$) The complete past history of consumption and asset returns are included in I_t, along with just about anything else (rainfall patterns in Cleveland, etc.).

Now in order to have something to estimate we parameterize the utility function to the CRRA form: $u(c) = \frac{c^{1-\sigma}}{1-\sigma}$. Then $u'(c) = c^{-\sigma}$, and

$$E\left(\beta \frac{C_{t+1}^{-\sigma}}{C_t^{-\sigma}} R_{at+1} \middle| I_t \right) = 1 \qquad \forall i$$

Let $W_t = W(I_t)$ be a w-vector of instruments based on I_t, then:

$$E\left(W_t \left(\beta \frac{C_{t+1}^{-\sigma}}{C_t^{-\sigma}} R_{a,t+1} - 1 \right) \right) = 0 \qquad \forall i$$

So we can estimate the model parameters β and σ by applying GMM to the $L = wA$ moment conditions

$$\frac{1}{T} \sum_{t=1}^{T} W_t \left(\beta \frac{C_{t+1}^{-\sigma}}{C_t^{-\sigma}} R_{a,t+1} - 1 \right) = 0 \qquad \forall i$$

Since there are only 2 model parameters, the model is just identified if $L = 2$ and overidentified if $L > 2$.

GMM Estimation

More generally, suppose we have a model with k parameters θ, and theory suggests $L \geq k$ functionally independent population moment conditions:

$$E\left[m_{il}\left(\theta\right)\right] = 0$$

for $i = 1, 2, ..., n$ observations and where $l = 1, 2, ..., L$. We define the sample moment conditions as

$$\bar{m}_l\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n} m_{il}\left(\theta\right) = 0.$$

When $L = k$, we have as many equations as unknowns and there is a unique solution for the GMM estimator, $\hat{\theta}_{GMM}$: the unique solution to the sample moment conditions. In general, when $L > k$, there is no value of θ that solves all L conditions. We could choose any k of the L conditions and compute the (unique) parameter vector that solves this set of moment conditions. But the choice of conditions would be completely arbitrary and there are

$${}_LC_k = \binom{L}{k} = \frac{L!}{k!\left(L-k\right)!}$$

different estimates we could compute in this way. An alternative is to choose a parameter estimate that comes “closest” to satisfying all the moment conditions in some sense. In fact, GMM estimators minimize a weighted sum of squares of the sample moment conditions

$$Q_n(\theta) = \bar{\mathbf{m}}_n\left(\theta\right)' \mathbf{W}_n \bar{\mathbf{m}}_n\left(\theta\right) \tag{3}$$

where $\bar{\mathbf{m}}_n\left(\theta\right)$ is the $L \times 1$ vector of sample moment conditions, and $\mathbf{W}_n$ is any $L \times L$ positive definite weight matrix. For example, we could choose $\mathbf{W}_n = \mathbf{I}_L$ and minimize the equally-weighted sum of squares

$$Q_n(\theta) = \bar{\mathbf{m}}_n\left(\theta\right)' \bar{\mathbf{m}}_n\left(\theta\right).$$

Sometimes, the weighted sum of squares Q in eq. (3) is called a **distance function**, and the estimator that minimizes it is called the **minimum distance estimator**. Usually, we use this terminology when the $\bar{\mathbf{m}}\left(\theta\right)$ are not sample moments, but some other kind of criterion (e.g., orthogonality conditions).

As you might imagine, there is an optimal choice of weighting matrix $\mathbf{W}_n$. We will discuss the optimal GMM weights shortly. For now just take $\mathbf{W}_n$ as given.

Properties of the GMM estimator

We'll begin by sketching a proof of consistency. Let θ_0 denote the true parameter vector. Stacking up the L population moment conditions, GMM estimation is based on the vector of population moment conditions:

$$E\left[\mathbf{m}_i\left(\theta_0\right)\right] = \mathbf{0} \tag{4}$$

and the vector of sample moment conditions:

$$\bar{\mathbf{m}}_n(\theta_0) = \mathbf{0} \tag{5}$$

where n indexes the sample size. In the overidentified case, the sample moment conditions cannot all be satisfied exactly unless they are functionally dependent. However, when the expectation (4) exists, we know by Khinchine's WLLN that

$$\text{plim } \bar{\mathbf{m}}_n(\theta_0) = \text{plim } \frac{1}{n}\sum_{i=1}^{n} \mathbf{m}_i(\theta_0) = E[\mathbf{m}_i(\theta_0)] = \mathbf{0}$$

since $\bar{\mathbf{m}}_n(\theta_0)$ is just a sample mean. When the sample size is n, the GMM estimator minimizes the weighted sum of squares:

$$Q_n(\theta) = \bar{\mathbf{m}}_n(\theta)' \mathbf{W}_n \bar{\mathbf{m}}_n(\theta).$$

Assuming that $\mathbf{W}_n$ has a finite probability limit $\mathbf{W}$, it follows immediately that

$$\text{plim } Q_n(\theta_0) = 0.$$

However, since $Q_n(\theta) \geq 0$ for every n and θ (since $\mathbf{W}_n$ is positive definite), and since the GMM estimator $\hat{\theta}_{GMM}$ is a minimizer, we also know that

$$0 \leq Q_n\left(\hat{\theta}_{GMM}\right) \leq Q_n(\theta_0)$$

for every n. Therefore plim $Q_n\left(\hat{\theta}_{GMM}\right) =$ plim $Q_n(\theta_0) = 0$, and the Slutsky Theorem tells us that

$$\text{plim } \bar{\mathbf{m}}_n\left(\hat{\theta}_{GMM}\right) = \text{plim } \bar{\mathbf{m}}_n(\theta_0).$$

If the moment conditions are one-to-one (i.e., $\theta_1 \neq \theta_2 \Rightarrow \bar{\mathbf{m}}_n(\theta_1) \neq \bar{\mathbf{m}}_n(\theta_2)$) then plim$\hat{\theta}_{GMM} = \theta_0$ and the GMM estimator is consistent.

The GMM estimator is also asymptotically normal. Let $\bar{\mathbf{G}}_n(\theta)$ denote the gradient matrix:

$$\bar{\mathbf{G}}_n(\theta) = \frac{\partial \bar{\mathbf{m}}_n(\theta)}{\partial \theta'} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \bar{\mathbf{m}}_i(\theta)}{\partial \theta'}.$$

We will assume that the gradient matrix evaluated at θ_0 has a finite probability limit

$$\text{plim } \bar{\mathbf{G}}_n(\theta_0) = \bar{\mathbf{G}}(\theta_0).$$

The GMM estimator solves the first order conditions

$$\frac{\partial Q_n\left(\hat{\theta}_{GMM}\right)}{\partial \hat{\theta}_{GMM}} = 2\bar{\mathbf{G}}_n\left(\hat{\theta}_{GMM}\right)' \mathbf{W}_n \bar{\mathbf{m}}_n\left(\hat{\theta}_{GMM}\right) = \mathbf{0}.$$

If we take a first-order Taylor expansion of $\bar{\mathbf{m}}_n\left(\hat{\theta}_{GMM}\right)$ around θ_0, we get

$$\bar{\mathbf{m}}_n\left(\hat{\theta}_{GMM}\right) = \bar{\mathbf{m}}_n(\theta_0) + \bar{\mathbf{G}}_n\left(\bar{\theta}\right)\left(\hat{\theta}_{GMM} - \theta_0\right)$$

for $\bar{\theta} = \lambda \hat{\theta}_{GMM} + (1-\lambda)\,\theta_0$. Substituting this into the FOC, we see that

$$\bar{\mathbf{G}}_n \left(\hat{\theta}_{GMM}\right)' \mathbf{W}_n \bar{\mathbf{m}}_n (\theta_0) + \bar{\mathbf{G}}_n \left(\hat{\theta}_{GMM}\right)' \mathbf{W}_n \bar{\mathbf{G}}_n \left(\bar{\theta}\right) \left(\hat{\theta}_{GMM} - \theta_0\right) = \mathbf{0}.$$

Therefore,

$$\sqrt{n}\left(\hat{\theta}_{GMM} - \theta_0\right) = -\left[\bar{\mathbf{G}}_n \left(\hat{\theta}_{GMM}\right)' \mathbf{W}_n \bar{\mathbf{G}}_n \left(\bar{\theta}\right)\right]^{-1} \bar{\mathbf{G}}_n \left(\hat{\theta}_{GMM}\right)' \mathbf{W}_n \sqrt{n} \bar{\mathbf{m}}_n (\theta_0)\,.$$

Assuming the gradients are continuous, consistency of $\hat{\theta}_{GMM}$ and the Slutsky theorem imply

$$\text{plim} \bar{\mathbf{G}}_n \left(\hat{\theta}_{GMM}\right) = \text{plim} \bar{\mathbf{G}}_n \left(\bar{\theta}\right) = \text{plim} \bar{\mathbf{G}}_n (\theta_0) = \bar{\mathbf{G}} (\theta_0)$$

and hence

$$\sqrt{n}\left(\hat{\theta}_{GMM} - \theta_0\right) \xrightarrow{d} -\left[\bar{\mathbf{G}}(\theta_0)' \mathbf{W} \bar{\mathbf{G}}(\theta_0)\right]^{-1} \bar{\mathbf{G}}(\theta_0)' \mathbf{W} \sqrt{n} \bar{\mathbf{m}}_n (\theta_0)\,. \tag{6}$$

Now, the CLT tells us that under appropriate conditions

$$\sqrt{n} \bar{\mathbf{m}}_n (\theta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma})$$

where $\mathbf{\Sigma} = E\left[\mathbf{m}_i(\theta_0)\, \mathbf{m}_i(\theta_0)'\right]$. Since all terms other than $\bar{\mathbf{m}}_n(\theta_0)$ on the right hand side of (6) are constants, it follows that

$$\sqrt{n}\left(\hat{\theta}_{GMM} - \theta_0\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{GMM})$$

where

$$\mathbf{V}_{GMM} = \left[\bar{\mathbf{G}}(\theta_0)' \mathbf{W} \bar{\mathbf{G}}(\theta_0)\right]^{-1} \bar{\mathbf{G}}(\theta_0)' \mathbf{W} \mathbf{\Sigma} \mathbf{W} \bar{\mathbf{G}}(\theta_0) \left[\bar{\mathbf{G}}(\theta_0)' \mathbf{W} \bar{\mathbf{G}}(\theta_0)\right]^{-1}$$

so that

$$\hat{\theta}_{GMM} \overset{a}{\sim} N\left(\theta_0, \frac{1}{n}\mathbf{V}_{GMM}\right).$$

Choosing the Weight Matrix

Notice that $\mathbf{V}_{GMM}$ depends on the choice of the weight matrix $\mathbf{W}_n$ (asymptotically, it depends on $\text{plim}\mathbf{W}_n = \mathbf{W}$). This suggests that there may be an optimal choice of weight matrix that minimizes the variance of the GMM estimator. In fact, Hansen (1982) shows that asymptotically, the optimal weight matrix is $\mathbf{W} = \mathbf{\Sigma}^{-1}$. In this case the variance of the limiting distribution of the GMM estimator is

$$\begin{aligned}
\mathbf{V}^*_{GMM} &= \left[\bar{\mathbf{G}}(\theta_0)' \mathbf{\Sigma}^{-1} \bar{\mathbf{G}}(\theta_0)\right]^{-1} \bar{\mathbf{G}}(\theta_0)' \mathbf{\Sigma}^{-1} \mathbf{\Sigma} \mathbf{\Sigma}^{-\mathbf{1}} \bar{\mathbf{G}}(\theta_0) \left[\bar{\mathbf{G}}(\theta_0)' \mathbf{\Sigma}^{-\mathbf{1}} \bar{\mathbf{G}}(\theta_0)\right]^{-1} \\
&= \left[\bar{\mathbf{G}}(\theta_0)' \mathbf{\Sigma}^{-1} \bar{\mathbf{G}}(\theta_0)\right]^{-1} \bar{\mathbf{G}}(\theta_0)' \mathbf{\Sigma}^{-1} \bar{\mathbf{G}}(\theta_0) \left[\bar{\mathbf{G}}(\theta_0)' \mathbf{\Sigma}^{-\mathbf{1}} \bar{\mathbf{G}}(\theta_0)\right]^{-1} \\
&= \left[\bar{\mathbf{G}}(\theta_0)' \mathbf{\Sigma}^{-1} \bar{\mathbf{G}}(\theta_0)\right]^{-1}.
\end{aligned}$$

Feasible Estimation With Optimal Weights

A problem with using the optimal weight matrix is that it requires knowledge of the variance of the sample moment conditions, which we can only estimate once we have $\hat{\theta}_{GMM}$. How to proceed? Easy. All we need is a consistent estimate of $\mathbf{\Sigma}^{-1}$, which we can obtain from any consistent estimate of $\bar{\mathbf{m}}_n(\theta_0)$. Since GMM estimation using **any** positive definite weight matrix will yield a consistent estimate of θ_0, we can resort to two-step GMM estimation

1. Set $\mathbf{W} = \mathbf{I}_L$ (or any other convenient positive definite matrix) and compute the GMM estimator, $\hat{\theta}_{GMM}$, for those weights. Use these estimates to compute the sample variance of $\sqrt{n}\bar{\mathbf{m}}\left(\hat{\theta}_{GMM}\right)$, denoted $\hat{\mathbf{\Sigma}}$.

2. Redo the GMM estimation with weights $\mathbf{W} = \hat{\mathbf{\Sigma}}^{-1}$.

You could iterate on this process indefinitely (or until convergence). But the result would be asymptotically equivalent to the two-step estimator. We can estimate the asymptotic variance of the optimal feasible GMM estimator using:

$$\hat{\mathbf{V}}^*_{GMM} = \left[\bar{\mathbf{G}}\left(\hat{\theta}_{GMM}\right)' \hat{\mathbf{\Sigma}}^{-1} \bar{\mathbf{G}}\left(\hat{\theta}_{GMM}\right)\right]^{-1}.$$

Testing Overidentifying Restrictions

As always, we can use Wald and LM tests to test additional restrictions on the model parameters themselves. As with 2SLS or SEM, if our model is overidentified, we can also test the model specification itself.

Notice that in the exactly identified case, the distance function

$$\hat{Q} = \bar{\mathbf{m}}\left(\hat{\theta}_{GMM}\right)' \mathbf{W} \bar{\mathbf{m}}\left(\hat{\theta}_{GMM}\right) = 0$$

since the GMM estimates solve $\bar{\mathbf{m}}\left(\hat{\theta}_{GMM}\right) = \mathbf{0}$ exactly. In the overidentified case the sample moment conditions will not be satisfied exactly, and hence $\hat{Q} \neq 0$. However, notice that if we use the optimal weight matrix we have

$$\begin{aligned} n\hat{Q} &= \left[\sqrt{n}\bar{\mathbf{m}}\left(\hat{\theta}_{GMM}\right)'\right] \hat{\mathbf{\Sigma}}^{-1} \left[\sqrt{n}\bar{\mathbf{m}}\left(\hat{\theta}_{GMM}\right)\right] \\ &= \left[\sqrt{n}\bar{\mathbf{m}}\left(\hat{\theta}_{GMM}\right)'\right] \left\{Est.Asy.Var\left[\sqrt{n}\bar{\mathbf{m}}\left(\hat{\theta}_{GMM}\right)\right]\right\}^{-1} \left[\sqrt{n}\bar{\mathbf{m}}\left(\hat{\theta}_{GMM}\right)\right] \end{aligned}$$

which is just a Wald statistic of the null hypothesis that there is a solution to the moment conditions. Under this null,

$$n\hat{Q} \overset{a}{\sim} \chi^2_{L-k}.$$

Intuitively, if the model parameters are overidentified by the moment conditions, the sample moment conditions are imposing $L - k$ restrictions on the parameter vector. This test (with different notation) is commonly called Hansen's J test.

Appendix: Some more examples

GMM and 2SLS

Consider the IV estimator from Lecture 17. We used $\mathbf{w}_i$ to denote the vector of instruments, $\mathbf{z}_i$ to denote the vector of explanatory variables (endogenous and exogenous), and δ to denote the coefficient vector. Consider the population moment conditions:

$$E\left[\mathbf{w}_i\left(y_i - \mathbf{z}_i'\delta\right)\right] = \mathbf{0} \tag{7}$$

(these are orthogonality conditions again). The sample analog of (7) is

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i\left(y_i - \mathbf{z}_i'\hat{\delta}_{GMM}\right) = \mathbf{0}. \tag{8}$$

When the number of instruments (M) equals the number of parameters in δ (k), i.e., the exactly identified case, then the sample moment conditions define k equations in k unknowns. You can verify that in the exactly identified case, the solution to (8) is the IV estimator. In the overidentified case ($M > k$) we have more equations than unknowns and we cannot solve the sample moment conditions exactly for unique estimates of δ. This is the GMM case, to which we now turn.

We've already seen an equivalence between IV estimation and GMM in the exactly identified case. What about in the overidentified case? Consider equation j of an M equation SEM:

$$y_{jt} = \mathbf{z}_{jt}'\delta_j + \varepsilon_{jt}.$$

Let $\mathbf{x}_t$ denote the $k \times 1$ vector of all exogenous variables in the system for observation t. Suppose we define the population moment (orthogonality) condition:

$$E\left[\mathbf{x}_t\varepsilon_{jt}\right] = E\left[\mathbf{x}_t\left(y_{jt} - \mathbf{z}_{jt}'\delta_j\right)\right] = \mathbf{0}$$

and corresponding sample moment conditions:

$$\bar{\mathbf{m}}_T\left(\delta_j\right) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\left(y_{jt} - \mathbf{z}_{jt}'\delta_j\right) = \mathbf{0}.$$

When $M > k$, we saw previously that the IV estimator was the 2SLS estimator. When we use GMM weights $\mathbf{W}_T = \left(\mathbf{X}'\mathbf{X}\right)^{-1}$, then the GMM estimator is the 2SLS estimator also.

If we define population and sample moment conditions in this way for each of the M equations in the system and stack them up, so that

$$\bar{\mathbf{m}}_T\left(\delta\right) = \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\left(y_{1t} - \mathbf{z}_{1t}'\delta_1\right) \\ \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\left(y_{2t} - \mathbf{z}_{2t}'\delta_2\right) \\ \vdots \\ \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\left(y_{Mt} - \mathbf{z}_{Mt}'\delta_M\right) \end{bmatrix} = \mathbf{0}$$

then the GMM estimator is the 3SLS estimator when our weight matrix is

$$\mathbf{W}_T = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \cdots & \mathbf{W}_{1M} \\ \mathbf{W}_{21} & \mathbf{W}_{22} & \cdots & \mathbf{W}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{M1} & \mathbf{W}_{2M} & \cdots & \mathbf{W}_{MM} \end{bmatrix}^{-1}$$

with elements $\mathbf{W}_{jl} = \hat{\sigma}_{jl} \left(\mathbf{X}'\mathbf{X}\right)$ for $\hat{\sigma}_{jl} = T^{-1} \left(\mathbf{y}_j - \mathbf{Z}_j \delta_j^{2SLS}\right)' \left(\mathbf{y}_l - \mathbf{Z}_l \delta_l^{2SLS}\right)$. Compare these weights to the GLS weights we used to derive the 3SLS estimator in the first place ...