Chapter 1

Generalized Method of Moments

Generalized method of moments (GMM) is an estimation principle that extends **method of moments**. It seeks the parameter value that minimizes a quadratic form of the sample moments. It is particularly useful in estimating structural models in which moment conditions can be (hopefully) derived from economic theory. GMM emerges as one of the most popular estimators in modern econometrics.

1.1 Estimating Equations

We are interested in a moment function $g_i(\theta) = g(z_i, \theta) \in \mathbb{R}^m$, where θ is a p-dimensional parameter. Motivated by some economic theory, we look for the true value $\theta_0 \in \Theta$ as the solution to the following m moment equations

$$\mathbb{E}\left[g_i(\theta)\right] = \mathbf{0}_m.$$

These are the population **estimating equations**.

The true value θ_0 is identified if it uniquely satisfies $\mathbb{E}\left[g_i(\theta_0)\right] = 0$. To measure the distance of $\mathbb{E}\left[g_i(\theta)\right]$ from $\mathbf{0}_m$, define

$$S(\theta; W) = \mathbb{E}\left[g_i(\theta)\right]' W \mathbb{E}\left[g_i(\theta)\right],$$

where W is an $m \times m$ symmetric positive-definite matrix W, which is called the **weighting matrix**. For identification and consistent estimation, the choice of W is arbitrary, and one valid candidate is the identity matrix I_m. Formally, **identification** means that for any $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that

$$\inf_{\theta \in \Theta \setminus \mathcal{N}_\varepsilon(\theta_0)} S(\theta; W) > S(\theta_0; W) + \delta.$$

Next, we provide more conditions under which we can verify identification. Suppose $g(\cdot, \theta)$ is twice differentiable with respect to θ over the support of z_i, and the differentiation and expectation is exchangeable. Denote the $m \times p$ matrix $D_i(\theta) = \frac{\partial}{\partial \theta'} g_i(\theta)$, and $D_0 = \mathbb{E}\left[D_i(\theta_0)\right]$. Define

$$\underset{(p\times 1)}{\psi(\theta; W)} = \frac{\partial}{\partial \theta} S(\theta; W) = 2\mathbb{E}\left[D_i'(\theta)\right] W \mathbb{E}\left[g_i(\theta)\right]$$

$$\underset{(p\times p)}{H(\theta; W)} = \frac{\partial^2}{\partial \theta \partial \theta'} S(\theta; W) = 2\sum_{k=1}^{m} \left(W\mathbb{E}\left[g_i(\theta)\right]\right)_k \mathbb{E}\left[\frac{\partial^2 g_k(z_i, \theta)}{\partial \theta \partial \theta'}\right] + 2\mathbb{E}\left[D_i'(\theta)\right] W \mathbb{E}\left[D_i(\theta)\right].$$

(Be careful about the way to stack the second derivative of g.) Take a Taylor expansion of $S(\theta; W)$ around θ_0:

$$S(\theta; W) = S(\theta_0; W) + \psi(\theta_0; W)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0) H(\theta_0; W)(\theta - \theta_0) + h.o.t,$$

where the higher-order term (h.o.t.) is $o(\|\theta - \theta_0\|^2)$. Evaluated at $\theta = \theta_0$, since $\mathbb{E}[g_i(\theta_0)] = 0$ we have

$$\begin{aligned} S(\theta_0; W) &= 0 \\ \psi(\theta_0; W) &= \mathbf{0}_m \\ H(\theta_0; W) &= 2D_0' W D_0. \end{aligned}$$

If we take a heuristic approach by ignoring the h.o.t, for any θ such that $\|\theta - \theta_0\| > \delta$, we obtain

$$S(\theta; W) = \frac{1}{2}(\theta - \theta_0)' H(\theta_0; W)(\theta - \theta_0) = (\theta - \theta_0)' D_0' W D_0 (\theta - \theta_0).$$

We need a full rank D_0 to guarantee that there exists a constant c such that

$$S(\theta; W) > \|\theta - \theta_0\|^2 c \cdot \lambda_{\min}(W) \geq \delta^2 c \cdot \lambda_{\min}(W).$$

Otherwise, there exists a $\theta^\dagger$ such that $\theta^\dagger \neq \theta_0$ but $D_0(\theta - \theta_0) = \mathbf{0}_m$; we cannot identify θ_0 because $S(\theta^\dagger; W) = 0$ as well.

Exercise 1.1. Argue that a necessary condition for identification is $p \leq m$. In other words, the number of moment conditions must be no smaller than the number of parameters.

Example 1.1. Consider the linear IV model $y_i = x_i'\theta + \varepsilon_i$, where x_i is a p-dimensional endogenous variable and w_i is an m-dimensional instrumental variable. The moment function $g_i(\theta) = w_i(y_i - x_i'\theta)$. The orthogonality between z_i and ε_i ensures

$$\mathbb{E}[g_i(\theta_0)] = \mathbb{E}\left[w_i\left(y_i - x_i'\theta_0\right)\right] = \mathbb{E}[w_i \varepsilon_i] = \mathbf{0}_m.$$

The necessary condition for identification is a full rank $D_0 = -\mathbb{E}[w_i x_i']$, which is the relevant condition between w_i and x_i.

The above heuristic argument for identification only ensures **local identification**, because we neglect the h.o.t. in a small neighborhood around θ_0. Sufficient conditions for **global identification** is challenging and under active research.

1.2 GMM Estimator

Given an iid sample z_i, $i = 1, \ldots, n$, we expect the sample moments

$$\frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0) \approx \mathbf{0}_m.$$

Therefore, it is reasonable to estimate θ by finding some value in Θ to set $\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$ close to $\mathbf{0}_m$.

Example 1.2. Consider again the linear IV model. If $p = m$ (just identified), then we can solve

$$\frac{1}{n}\sum_{i=1}^{n} g_i(\theta) = \frac{1}{n}\sum_{i=1}^{n} w_i\left(y_i - x_i'\theta\right) = \frac{1}{n}\sum_{i=1}^{n} w_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} w_i x_i'\right)\theta = \mathbf{0}_m$$

with $\widehat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n} w_i x_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} w_i y_i = \left(W'X\right)^{-1}W'Y$. If $m > p$ (over-identified), the m-equation system

$$\left(\frac{1}{n}\sum_{i=1}^{n} w_i x_i'\right)\theta = \frac{1}{n}\sum_{i=1}^{n} w_i y_i$$

has no solution because there are few than m free parameters in θ.

To work with just identified cases and the over-identified cases in a unified framework, we use the sample moment $\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$ to mimic the population moment $\mathbb{E}\left[g_i(\theta)\right]$, and define the sample criterion function

$$S_n(\theta) = \bar{g}_n'(\theta)\, W \bar{g}_n(\theta)$$

as the counterpart of the population distance $S(\theta; W)$. The **GMM estimator**

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} S_n(\theta)$$

is the minimizer of the sample criterion function $S_n(\theta)$. Notice that GMM is not an M-function as it is not a sample average of the loss function of each individual observation. Instead, it is a quadratic form of a sample average of the moment function $g_i(\theta)$.

The proof of consistency of the GMM under identification and ULLN is identical to that of the M-estimator. We leave it as an exercise.

Exercise 1.2. Show that $\widehat{\theta} \xrightarrow{p} \theta_0$ if (i) θ_0 is identified, (ii) $\sup_{\theta\in\Theta_0}\left|S_n(\theta) - S(\theta)\right| \xrightarrow{p} 0$.

Next, given consistency we further check the asymptotic distribution of the GMM estimator under conditions of differentiability. Denote

$$\bar{\psi}_n(\theta; W) = \frac{\partial}{\partial\theta} S_n(\theta; W) = 2\bar{D}_n'(\theta)\, W \bar{g}_n(\theta),$$

where $\bar{D}_n(\theta) = \frac{1}{n}\sum D_i(\theta)$, and

$$\bar{H}_n(\theta; W) = \frac{\partial^2}{\partial\theta\partial\theta'} S_n(\theta; W) = 2\sum_{k=1}^{m}\left(W\bar{g}_n(\theta)\right)_k \frac{\partial^2 \bar{g}_{n,k}(z_i,\theta)}{\partial\theta\partial\theta'} + 2\bar{D}_n'(\theta)\, W \bar{D}_n(\theta).$$

The first-order condition of the optimality of GMM yields

$$\mathbf{0}_m = \bar{\psi}_n(\widehat{\theta}; W) = \bar{\psi}_n(\theta_0; W) + \bar{H}_n(\theta_0; W)\left(\widehat{\theta} - \theta_0\right) + h.o.t,$$

where the right-hand side comes from a Taylor expansion of $\bar{\psi}_n(\widehat{\theta}; W)$ around θ_0. We again ignore the higher-order term and rearrange:

$$\sqrt{n}(\widehat{\theta} - \theta_0) = -\bar{H}_n^{-1}(\theta_0; W) \times \sqrt{n}\bar{\psi}_n(\theta_0; W). \tag{1.1}$$

We analyze the numerator $\sqrt{n}\bar{\psi}_n(\theta_0;W) = 2\bar{D}_n'(\theta_0)W\frac{1}{\sqrt{n}}\sum_i g_i(\theta_0)$. Since $\mathbb{E}[g_i(\theta_0)] = 0$, if the variance is finite we can apply CLT:

$$\frac{1}{\sqrt{n}}\sum_i g_i(\theta_0) \xrightarrow{d} N(0,\Omega_0)$$

where $\Omega_0 = \mathrm{var}[g_i(\theta_0)] = \mathbb{E}[g_i(\theta_0)g_i'(\theta_0)]$. If in addition $\bar{D}_n(\theta_0) \xrightarrow{p} D_0$ by some LLN, we have

$$\sqrt{n}\bar{\psi}_n(\theta_0;W) \xrightarrow{d} N\left(0, 4D_0'W\Omega_0WD_0\right).$$

The denominator

$$\bar{H}_n(\theta_0;W) = 2\sum_{k=1}^{m}(W\bar{g}_n(\theta_0))_k\frac{\partial^2\bar{g}_{n,k}(z_i,\theta_0)}{\partial\theta\partial\theta'} + 2\bar{D}_n'(\theta_0)W\bar{D}_n(\theta_0) \xrightarrow{p} 2D_0WD_0,$$

where the first term vanishes as $\bar{g}_n(\theta_0) \xrightarrow{p} \mathbf{0}_m$. Putting the numerator and the denominator together, we conclude

$$\sqrt{n}(\widehat{\theta}-\theta_0) \xrightarrow{d} N\left(0, \left(D_0'WD_0\right)^{-1}D_0'W\Omega_0WD_0\left(D_0'WD_0\right)^{-1}\right).$$

It is obvious that the asymptotic variance of the GMM estimator depends on the choice of the weight matrix W. If we choose $W = \Omega_0^{-1}$, the asymptotic variance is simplified as

$$\sqrt{n}(\widehat{\theta}-\theta_0) \xrightarrow{d} N\left(0, \left(D_0'\Omega_0^{-1}D_0\right)^{-1}\right).$$

In turns out that this choice of W achieves efficiency.

1.3 Over-identification Test

The definition of $\widehat{\theta}$ entails that $S_n(\widehat{\theta};W)$ is the minimum of $S_n(\theta;W)$. It turns out that the scaled (by n) criterion function $nS_n\left(\theta;\Omega_0^{-1}\right)$ can serve as a test statistic for model specification.

The scaling factor n is essential because $S_n(\widehat{\theta};W) \leq S_n(\theta_0;W) \xrightarrow{p} 0$ due to $\bar{g}_n(\theta_0) \xrightarrow{p} \mathbf{0}_m$, whereas

$$nS_n(\theta_0;W) = \left(\sqrt{n}\bar{g}_n'(\theta_0)\right)W\left(\sqrt{n}\bar{g}_n(\theta_0)\right)$$

is a quadratic form of an asymptotically normal random vector $\sqrt{n}\bar{g}_n'(\theta_0)$. In particular, the choice $W = \Omega_0^{-1}$ gives

$$nS_n\left(\theta_0;\Omega_0^{-1}\right) \xrightarrow{d} \chi^2(p), \tag{1.2}$$

but this is different from the asymptotic distribution of

$$nS_n\left(\widehat{\theta};\Omega_0^{-1}\right) = n\bar{g}_n'\left(\widehat{\theta}\right)\Omega_0^{-1}\bar{g}_n\left(\widehat{\theta}\right) = \left\|\Omega_0^{-1/2}\sqrt{n}\bar{g}_n\left(\widehat{\theta}\right)\right\|. \tag{1.3}$$

Decompose

$$\sqrt{n}\bar{g}_n\left(\widehat{\theta}\right) = \sqrt{n}\bar{g}_n(\theta_0) + \sqrt{n}\left(\bar{g}_n\left(\widehat{\theta}\right) - \bar{g}_n(\theta_0)\right) \tag{1.4}$$

The second term, by a Taylor expansion around θ_0 and putting the h.o.t. into $o_p(1)$, becomes

$$\begin{aligned}\sqrt{n}\left(\bar{g}_n\left(\widehat{\theta}\right) - \bar{g}_n(\theta_0)\right) &= \sqrt{n}\bar{D}_n'(\theta_0)\times\left(\widehat{\theta}-\theta_0\right) + o_p(1)\\ &= -\bar{D}_n'(\theta_0)\times\bar{H}_n^{-1}\left(\theta_0;\Omega_0^{-1}\right)\times\sqrt{n}\bar{\psi}_n\left(\theta_0;\Omega_0^{-1}\right) + o_p(1)\\ &= -\bar{D}_n'(\theta_0)\bar{H}_n^{-1}\left(\theta_0;\Omega_0^{-1}\right)2\bar{D}_n'(\theta_0)\Omega_0^{-1}\sqrt{n}\bar{g}_n(\theta_0) + o_p(1)\\ &= -D_0\left(D_0'\Omega_0^{-1}D_0\right)^{-1}D_0'\Omega_0^{-1}\sqrt{n}\bar{g}_n(\theta_0) + o_p(1).\end{aligned} \tag{1.5}$$

where the second line follows by (1.1), the third line by the definition of $\bar{\psi}_n$, and the last line by citing the probabilistic limit of the matrices $\bar{H}_n$ and $\bar{D}_n$. (1.4) and (1.5) imply

$$\begin{aligned} \Omega_0^{-1/2}\sqrt{n}\bar{g}_n\left(\widehat{\theta}\right) &= \Omega_0^{-1/2}\left(I_m - D_0\left(D_0'\Omega_0^{-1}D_0\right)^{-1}D_0'\Omega_0^{-1}\right)\sqrt{n}\bar{g}_n\left(\theta_0\right) + o_p(1) \\ &\xrightarrow{d} \Omega_0^{-1/2}\left(I_m - D_0\left(D_0'\Omega_0^{-1}D_0\right)^{-1}D_0'\Omega_0^{-1}\right)\times N\left(0,\Omega_0\right) \\ &\sim \Omega_0^{-1/2}\left(I_m - D_0\left(D_0'\Omega_0^{-1}D_0\right)^{-1}D_0'\Omega_0^{-1}\right)\Omega^{1/2}\times N\left(0,I_m\right) \\ &\sim \left(I_m - \Omega_0^{-1/2}D_0\left(D_0'\Omega_0^{-1}D_0\right)^{-1}D_0'\Omega_0^{-1/2}\right)\times N\left(0,I_m\right). \end{aligned}$$

Since $\left(I_m - \Omega^{-1/2}D_0\left(D_0'\Omega_0^{-1}D_0\right)^{-1}D_0'\Omega^{-1/2}\right)$ is idempotent with rank $(p-m)$, we have

$$nS_n\left(\widehat{\theta};\Omega_0^{-1}\right) = \left\|\Omega_0^{-1/2}\sqrt{n}\bar{g}_n\left(\widehat{\theta}\right)\right\| \xrightarrow{d} \chi^2\left(m-p\right).$$

Compared with (1.2), the estimated p-dimensional parameter $\widehat{\theta}$ deducts p degrees of freedom.

To make the test statistic feasible, we can use any consistent estimator $\widehat{\Omega}$ for Ω_0, for example,

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g_i\left(\widehat{\theta}\right)g_i'\left(\widehat{\theta}\right) - \bar{g}_n\left(\widehat{\theta}\right)\bar{g}_n'\left(\widehat{\theta}\right).$$

The feasible statistic $nS_n\left(\widehat{\theta};\widehat{\Omega}^{-1}\right)$ is called the J-statistic.