ECON818: Advanced Econometrics II

Chapter 9: Generalized Method of Moments

The Maximum Likelihood Estimator (MLE) is fully efficient among consistent and asymptotically normally distributed estimators with one big disadvantage: it requires that the density of the observed random variable(s) be fully specified. The Generalized Method of Moments (GMM) method needs fewer assumptions about the distribution. Extending the classic method of moments idea, the GMM estimator encompasses almost all the methods-based estimators that are commonly used.

The basic idea behind method of moments estimators is as follows. Say we have moment functions $m(\theta)$ and we know that they equal zero in the population for the true value of the parameter. We can estimate $\hat{m}(\theta)$ and look for $\hat{\theta}$ that makes $\hat{m} = 0$. We look for the $\hat{\theta}$ that minimizes the moment function's squared sum using a normal function $||m||$. If the solution is overidentified, there are different norm functions each associated with a weighting matrix W, which is defined by the data.

Definition

Consider the following population moment condition:

$$E[g(x_i, \theta)] = 0$$

Where $g(x, \theta)$ is the population moment and x_i is an l x 1 random vector. Its sample equivalent is:

$$\overline{g}(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(x_i; \theta)$$

The empiricial moment condition $\overline{g}(\theta) = 0$ is the sample counterpart of the population moment condition $E[g(x_i, \theta)] = 0$. If we have m moment conditions and k parameters then the method for solving depends on the ratio between the two.

1. Underidentified: $m < k$
   If $m < k$, there are more parameters than moment conditions, and it is impossible to find a unique solution to $\overline{g}(\theta) = 0$ without more information such as restrictions.

2. Exactly identified (Method of Moments): $m = k$
   If $m = k$, there exists a unique solution to $\overline{g}(\theta) = 0$. The Method of Moments estimator $\hat{\theta}$ can be identified by minimizing the criterion function S_n, which is equal to the squared sum of $\overline{g}(\theta)$:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} S_n(\theta) = \overline{g}(\theta)' \overline{g}(\theta)$$

3. Overidentified (Generalized Method of Moments): $m > k$
   If $m > k$, there is no unique solution to $\overline{g}(\theta) = 0$. In this situation, we estimate by minimizing a criterion function S_n which is equal to the weighted square sum of $\overline{g}(\theta)$. That is,

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \overline{g}(\theta)' W_n \overline{g}(\theta)$$

If we have case (1), there is no solution. We now proceed to examine solution methods 2 and 3 respectively.

1 Method of Moments

If we have k moment equations and k parameters, we can use the Method of Moments estimator to try to find a unique solution. Consider the following moment equations:

$$\overline{g}(\theta) = \begin{pmatrix} \overline{g_1}(\theta_1, ..., \theta_k) \\ \vdots \\ \overline{g_k}(\theta_1, ..., \theta_k) \end{pmatrix} = \begin{pmatrix} m_1 - \mu_1(\theta_1, ..., \theta_k) \\ \vdots \\ m_k - \mu_k(\theta_1, ..., \theta_k) \end{pmatrix} = 0$$

This says that the sample moments are defined as the difference between the sample statistic for the moment m and the true statistic μ. Under some regularity conditions, the method of moment estimator can be obtained by inverting this condition to solve:

$$\hat{\theta} = [\theta_1, ..., \theta_k]' = \hat{\theta}(m_1, ..., m_k)$$

Example 1 Suppose that X_i is an i.i.d. random sample where the PDF is:

$$f(x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp\left[-\frac{1}{2\sigma^2}(X_i - \mu)^2\right]$$

Consider the first 2 central moments:

$$E[X_i] - \mu = 0$$
$$E[(X_i - \mu)^2] - \sigma^2 = 0$$

and with

$$\overline{g_1} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$
$$\overline{g_2} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

their sample counterparts are:

$$\overline{g_1} - \mu = 0$$
$$\overline{g_2} - \sigma^2 = 0$$

Because $\overline{X}$ converges in probability to $E[X_i] = \mu$, $\overline{g_2}$ is also a consistent estimator of $\sigma^2 = E[(X_i - \mu)^2]$. Now these two moment conditions comprise a system of 2 equations with 2 unknown parameters.

How do we solve such a system? Since the statistic $E(x_i) - \mu = 0$ that we use here will never be exactly equal to zero, we must use the squared sum of differences to solve:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} S_n(\theta) = \overline{g}(\theta)' \overline{g}(\theta)$$

where S_n is based on the sample data. If we take the derivative of this quadratic, we can find our $\hat{\theta}$. This is:

$$\frac{\partial S_n(\theta)}{\partial \theta}\Big|_{\hat{\theta}} = 0$$

which should give us $\hat{\theta} = \hat{\mu}$ and $\hat{\sigma}$.

Although we have exact identification, with two moment equations and two unknowns, we might not be able to solve using this method. The method of moments estimator will not give us our solutions if the different moment equations collapse into a single equation when we take the derivative. This can be shown by returning to our example with the first two moment equations and two unknowns, μ and σ. We solve by differentiating the criterion equation S_n with respect first to μ and then to σ:

$$\begin{aligned} S_n(\mu, \sigma) &= \begin{bmatrix} (\overline{X} - \mu) \\ \frac{1}{n}\sum (X_i - \overline{X})^2 - \sigma^2 \end{bmatrix}' \begin{bmatrix} (\overline{X} - \mu) \\ \frac{1}{n}\sum (X_i - \overline{X})^2 - \sigma^2 \end{bmatrix} \\ &= (\overline{X} - \mu)^2 + 2(\overline{X} - \mu)(\frac{1}{n}\sum (X - \overline{X})^2 - \sigma^2) + \frac{1}{n^2}\left[\sum (X_i - \overline{X})^2 - \sigma^2\right]^2 \end{aligned}$$

Solving for $\hat{\mu}$,

$$\begin{aligned} \frac{\partial S_n(\theta)}{\partial \mu}\Big|_{\hat{\mu}} = 0 \rightarrow 0 &= 2(\overline{X} - \hat{\mu})(-1) + 2(-1)(\frac{1}{n}\sum (X_i - \overline{X})^2 - \hat{\sigma}^2) \\ 0 &= -2(\overline{X} - \hat{\mu}) - 2(\frac{1}{n}\sum (X_i - \overline{X})^2 - \hat{\sigma}^2) \\ 0 &= (\overline{X} - \hat{\mu}) + (\frac{1}{n}\sum (X_i - \overline{X})^2 - \hat{\sigma}^2) \\ \hat{\mu} - \overline{X} &= \frac{1}{n}\sum (X_i - \overline{X})^2 - \hat{\sigma}^2 \end{aligned}$$

Solving for $\hat{\sigma}$,

$$\frac{\partial S_n(\theta)}{\partial \sigma^2} = 0 \rightarrow 0 = -2(\overline{X} - \hat{\mu}) + \frac{2}{n^2}(\sum (X_i - \overline{X}) - \hat{\sigma^2})(-1)$$

$$0 = (\overline{X} - \hat{\mu}) + \frac{1}{n^2}(\sum (X_i - \overline{X}) - \hat{\sigma^2})$$

$$\hat{\mu} - \overline{X} = \frac{1}{n^2}(\sum (X_i - \overline{X}) - \hat{\sigma^2})$$

Both $\frac{\partial S}{\partial \mu}$ and $\frac{\partial S}{\partial \sigma^2}$ tell us the same thing: $\frac{1}{n}\sum (X_i - \overline{X})^2 - \hat{\sigma}^2 = \hat{\mu} - \overline{X}$. Our two equations have collapsed to one equation, which cannot solve for two unknowns. We need more moments to solve, even though we have exact identification. In order to employ more moment equations, we need to turn to the next method at hand: GMM.

2 Generalized Method of Moments (GMM)

If $m > k$, with more moment conditions than parameters, there is no unique solution to $\overline{g}(\theta) = 0$. In this situation, we solve for $\hat{\theta}$ by minimizing a weighted squared sum of $\overline{g}(\theta)$, where the weighting matrix W_n is used. That is, we minimize the following criterion function:

$$S_n(\theta) = g_n(\theta) W_n g_n(\theta)$$

The procedure to solve using GMM is similar to the Method of Moments, except that we now use a weighting matrix. The weighting matrix allows us to weigh the different moment equations that we have in defining the estimator. We may have equations not only for the first two moments (for mean and variance respectively), but also for the third and fourth moments (for skewness and kurtosis respectively). If we have say three moment equations, but only two parameters, we need to use a weighting matrix.

We proceed to demonstrate the GMM method through example.

Example 2 Let's examine a linear regression model:

$$y_i = x_i' \beta_0 + \epsilon_i$$

with $E(\epsilon_i | x_i) = 0$. This condition implies that the following moment conditions hold:

$$E(\epsilon_i | x_i) = E[(y_i - x_i' \beta_0) | x_i] = 0$$

Where

$$(y_i - x_i' \beta_0) x_i = g(y_i; x_i, \beta_0)$$

The sample counterpart is:

$$\overline{g}(\beta) = \frac{1}{n} \sum_{i=1}^{n} g(y_i; x_i, \beta)$$

$$= \frac{1}{n} X'(y - X\beta)$$

The Method of Moments criterion function $S_n(\theta) = \overline{g}'\overline{g}$ won't identify $\hat{\beta}$. We can identify $\hat{\beta}$ using GMM criterion function, though. We take the derivation of that function, defined as the weighted square sum of the sample moment function, where the weighting matrix is defined as $W_n = (X'X)^{-1}$:

$$\begin{aligned} S_n(\beta) &= \overline{g}(\beta) W_n \overline{g}(\beta) \\ &= \frac{1}{n^2}(y - X\beta)' X (X'X)^{-1} X'(y - X\beta) \end{aligned}$$

We solve by taking the derivative of this criterion function with respect to β:

$$\begin{aligned} \frac{\partial S_n(\theta)}{\partial \beta}\Big|_{\hat{\beta}} &= -\frac{2}{n^2} X'(X'X)(X'X)^{-1}(Y - X'\beta) = 0 \\ X'Y &= X'X\hat{\beta} \\ (X'X)^{-1}X'Y &= (X'X)^{-1}X'X\hat{\beta} \end{aligned}$$

Which implies that the GMM estimator is the OLS estimator:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Example 3 Suppose a representative consumer (i) maximizes

$$E\left[\sum_{\tau=1}^{T}(1+\delta)^{-\tau}U(C_{t+\tau})|I_t\right]$$

subject to

$$A_t = \sum_{\tau=1}^{T}(1+r)^{\tau}(C_{t+\tau} - E_{t+\tau})$$

where

$$\begin{aligned} U(.) &: \text{ utility function} \\ I_t &: \text{ information set at time } t \\ \delta &: \text{ subjective time preference rate} \\ r &: \text{ interest rate} \\ T &: \text{ length of economic life} \\ C_t &: \text{ consumption at time } t \\ E_t &: \text{ earnings at time } t \\ A_t &: \text{ assets at time } t \end{aligned}$$

The Euler equation is given by

$$E[U'(C_{t+1})|I_t] = \frac{1+\delta}{1+r}U'(C_t)$$

Consider

$$U(c) = \frac{c^{1-\gamma}}{1-\gamma},$$

where γ denotes the coefficient of relative risk aversion. Now the Euler equation becomes

$$E[C_{t+1}^{-\gamma}|I_t] = \frac{1+\gamma}{1+r}C_t^{-\gamma}$$

or equivalently

$$E\left[1 - \frac{1+\delta}{1+r}\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}|I_t\right] = 0$$

Now let Q_t be an instrument which is a $m \times 1$ random vector included in I_t. Then, by the law of iterated expectations, we have:

$$E\left[Q_t\left\{1 - \frac{1+\delta}{1+r}\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\right\}\right] = 0$$

where $X_t = (Q_t', C_{t+1}, C_t)'$ and $\theta_0 = \gamma$. Define:

$$\overline{g}(\gamma) = \frac{1}{n}\sum_{i=1}^{n} g_i = \frac{1}{n}\sum_{i=1}^{n} Q_{i,t}\left\{1 - \frac{1+\delta}{1+r}\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\right\}$$

and employ the GMM criterion function with this definition of $\overline{g}(\gamma)$:

$$S_n(\gamma) = \overline{g}(\gamma)' W_n \overline{g}(\gamma)$$

3 Asymptotic Properties

We need a weighting matrix W_n that makes the estimator not only identified but also consistent. By adding more moments to your estimator, you lose on the variance, but you win by getting more precise estimates. The weighting matrix enables us to weigh the moments that are more important to generating precise estimates and that minimize the variance.

We want to find the optimal W_n that minimizes the variance ($\hat{\theta}_{GMM}$) and maximizes the precision. This is the weighting matrix that will be consistent and asymptotically normal. That is, if W_n that converges in probability to W:

$$W_n \xrightarrow{p} W$$

(for some non-random, symmetric, positive definite matrix W), then (under some regularity conditions),

1. $\hat{\theta} \xrightarrow{p} \theta_0$

2. $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \xi)$

This says that if the weighting matrix converges to a fixed matrix, then the estimated parameters will converge to the true ones and the asymptotic distribution will be normal with mean zero and variance ξ. This ξ is defined by the sample data, where

$$\begin{aligned}\xi &= (G'WG)^{-1}G'WVWG(G'WG)^{-1} \\ G &= E\left[\frac{\partial g(X_i,\theta_0)}{\partial \theta'}\right] \\ V &= E[g(X_i,\theta_0)g(X_i,\theta_0)']\end{aligned}$$

Proof

We wish to find ξ, and prove that it is defined as we have stated above. Say you have moment functions $g(X_i,\theta_0)$ and that the X_i are i.i.d.

Step 1

By the Central Limit Theorem, the distribution is converging. If we divide simply by n, the distribution will not converge to the normal, but rather to zero:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta_0) \xrightarrow{d} 0$$

If, however, we divide by the square root of n, the distribution will converge to the normal:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_i,\theta_0) \xrightarrow{d} N(0,V) \qquad (1)$$

Provided that $E||g(X_i,\theta_0)||^2 < \infty$.

By the Weak Law of Large Numbers (WLLN) we also have:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(X_i,\theta_0)}{\partial \theta'} \xrightarrow{p} G \qquad (2)$$

The moment condition does not converge to zero: instead, it converges to a constant, which we call G, provided that $E||\frac{\partial}{\partial \theta'}(X_i,\theta_0)||^2 < \infty$.

Using (1) and (2), we find the first order conditions of the criterion function:

$$\begin{aligned}S_n(\theta_0) &= g(X_i,\theta_0)'W_n g(X_i,\theta_0) \\ S_n(\hat{\theta_0}) &= \left[\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta_0)'\right] W_n \left[\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta_0)\right] \\ \sqrt{n}\frac{\partial S_n(\theta_0)}{\partial \theta} &= 2\frac{1}{\sqrt{n}}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(X_i,\theta_0)}{\partial \theta'}\right]' W_n \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_i,\theta_0)\right] \qquad (3)\end{aligned}$$

where we know that the first bracketed term follows a distribution G and the second term follows a distribution $N(0, V)$. Thus

$$\sqrt{n}\frac{\partial S_n(\theta_0)}{\partial \theta} \xrightarrow{d} 2G'WN(0,V) \xrightarrow{d} N(0, 4G'WVW'G)$$

Since we square the 2 and W in bringing them into the variance. Thus, the distribution of S_n at θ_0 is $N(0, 4G'WVWG)$.

Step 2
We can use this information to find the distribution of the estimator. Recall first that a Taylor series expansion is defined as:

$$f(x) = f(x_0) + f'(x)(x - x_0)$$

Then consider a Taylor series expansion of $\frac{\partial S}{\partial \theta}$:

$$\frac{\partial S_n(\hat{\theta})}{\partial \theta}\Big|_{\hat{\theta}} = \frac{\partial S_n(\theta_0)}{\partial \theta} + \frac{\partial^2 S_n(\theta^*)}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0) = 0$$

$$(\hat{\theta} - \theta_0) = -\left(\frac{\partial S_n(\theta_0)}{\partial \theta}\right)\left(\frac{\partial^2 S_n(\theta^*)}{\partial \theta \partial \theta'}\right)^{-1}$$

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\sqrt{n}\left(\frac{\partial S_n(\theta_0)}{\partial \theta}\right)\left(\frac{\partial^2 S_n(\theta^*)}{\partial \theta \partial \theta'}\right)^{-1}$$

We already know from step one that $\left(\frac{\partial S_n(\theta_0)}{\partial \theta}\right)$ is distributed as $N(0, 4G'WVWG)$. So we just need to find the distribution of $\left(\frac{\partial^2 S_n(\theta^*)}{\partial \theta \partial \theta'}\right)$ to be able to find the distribution of $(\hat{\theta} - \theta_0)$.

$$\frac{\partial^2 S_n(\theta)}{\partial \theta \partial \theta'} = 2\left(\frac{\partial^2 \overline{g}(\theta)}{\partial Y \partial \theta}\right) W_n \overline{g}(\theta) + 2\frac{\partial g'(\theta)}{\partial \theta} W_n \frac{\partial \overline{g}(\theta)}{\partial \theta'}$$

Where $W_n \xrightarrow{p} W$, $\overline{g}(\theta) \xrightarrow{p} 0$, so that the first term goes to zero: $\frac{\partial^2 \overline{g}(\theta)}{\partial Y \partial \theta} W_n \overline{g}(\theta) \to 0$. We know that $\frac{\partial g'(\theta)}{\partial \theta} \to G$, so $2\frac{\partial g'(\theta)}{\partial \theta} W_n \frac{\partial \overline{g}(\theta)}{\partial \theta'} \to 2G'WG$.

Therefore, $\frac{\partial^2 S_n(\theta)}{\partial \theta \partial \theta'} \sim 2G'WG$. Thus,

$$\hat{\theta} - \theta_0 = -\sqrt{n}\Big(\frac{\partial^2 S_n(\theta)}{\partial \theta \partial \theta'}\Big)^{-1}\Big(\frac{\partial S_n(\theta)}{\partial \theta}\Big)\Big|_{\theta=\theta_0}$$

Converges to

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{d_{CLM}} \frac{1}{2}(G'WG)^{-1}N(0, 4G'WVWG)$$

The 2 and 4 will cancel out because the latter is in the variance:

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{d_{CLM}} N(0, (G'WG)^{-1}G'WVWG(G'WG)^{-1})$$

We have found that our variance $\xi = (G'WG)^{-1}G'WVWG(G'WG)^{-1}$, completing the proof.

4 Optimal Weighting Matrix

What is the optimal W which minimizes the asymptotic covariance matrix ξ? Let us first recall the definition of V:

$$\sqrt{n}g(\theta) \xrightarrow{d} N(0, V)$$

such that

$$V = E(g(x, \theta)g(X, \theta))$$

since in the sample,

$$\hat{V} = \overline{g}(\theta)\overline{g}(\theta)'.$$

If $W = V^{-1}$, then the middle part of ξ will cancel out:

$$\begin{aligned} \xi &= (G'WG)^{-1}G'WVWG(G'WG)^{-1} \\ \xi_0(W = V^{-1}) &= (G'V^{-1}G)^{-1}G'V^{-1}VV^{-1}G(G'V^{-1}G)^{-1} \\ &= (G'V^{-1}G)^{-1} \end{aligned}$$

To prove that $W = V^{-1}$ is efficient, we check that $\xi_0 < \xi$, that is, that the variance defined at $W = V^{-1}$ is the smallest amongst all variances.

proof If $\xi_0 < \xi$ then $W = V^{-1}$ is efficient. To prove that $\xi_0 < \xi$, it suffices to show that $\xi_0 - \xi$ is positive semidefinite (psd), or equivalently, that $\frac{1}{\xi_0} - \frac{1}{\xi}$ is positive semidefinite. Note that:

$$\frac{1}{\xi_0} = G'V^{-1}G$$

$$\frac{1}{\xi} = G'WG(G'WVWG)^{-1}G'WG$$

Taking the difference between these two,

$$\begin{aligned} &\frac{1}{\xi_0} - \frac{1}{\xi} = G'V^{-1}G - G'WG(G'WVWG)^{-1}G'WG \\ &= G'V^{-1/2}(I - V^{1/2}WG(GWVWG)^{-1}G'WV^{1/2})V^{-1/2}G \\ &= QPQ' \\ &= (QP)(QP)' \geq 0 \end{aligned}$$

Which is positive semi-definite where

$$\begin{aligned} Q &= G'V^{-\frac{1}{2}} \\ P &= I - V^{\frac{1}{2}}WG(G'WVWG)^{-1}G'WV^{-\frac{1}{2}} \end{aligned}$$

We can also prove positive semidefiniteness by proving that P is positive definite:

$$\begin{aligned} I &= [V^{1/2}WG(G'WVWG)^{-1}G'WV^{1/2}][I - V^{1/2}WG(G'WVWG)^{-1}G'WV^{1/2}] \\ &= 2V^{1/2}WG(G'WVWG)^{-1}G'WV^{1/2} - V^{1/2}WG(G'WVWG)^{-1}(GWVWG)(GW) \\ &= 2V^{1/2}WG(G'WVWG)^{-1}G'WV^{1/2} - V^{1/2}WGGW \end{aligned}$$

This is positive definite, so $\frac{1}{\xi_0} - \frac{1}{\xi}$ is positive definite.

We have thus shown through two methods that $W = V^{-1}$ generates the smallest variance and is thus efficient. Note that if we choose $W = I$, then the variance is $\xi = (G'G)^{-1}G'VG(G'G)^{-1}$, which is still consistent. In fact, we can use this weighting matrix to estimate the efficient W^*; we'll plug it in to obtain V^*, then use this to get an estimate for W^*.

5 Efficient GMM Estimation

The two-stage process to find the efficient GMM estimator is:

1. Minimize $S_n(\theta)$ with $W_n = I_n$ to obtain an estimate $\tilde{\theta}$
   Take the derivative of the moment function's criterion function $S_n(\theta)$ with respect to θ to minimize the square sum, using the identity matrix as the weighting function. The resulting estimate $\tilde{\theta}$ is inefficient but consistent. The underlying distribution here is $\sqrt{n}(\hat{\theta}_{GMM,W=I}, \theta) \xrightarrow{d} N(0, (G'G)^{-1}G'VG(G'G)^{-1})$.

2. Minimize $S_n(\theta)$ using $\tilde{\theta}$ from the first step to obtain an estimate of W_n^*:

$$W_n^* = \hat{V_n}^{*-1} = \left[\frac{1}{n}\sum_{i=1}^{n} \overline{g}(Z_i, \tilde{\theta})\overline{g}(Z_i, \tilde{\theta})'\right]^{-1}$$

   to obtain the efficient two-step estimator $\hat{\theta_E}$.

This estimator has the asymptotic distribution $\sqrt{n}(\hat{\theta_E} - \theta_0) \xrightarrow{d} N(0, \xi_0)$ where $\xi_0 = (G'V^{-1}G)^{-1}$ and where

$$\hat{G} = \frac{1}{n}\sum \frac{\partial g(Z_i, \hat{\theta}_E)}{\partial \theta'}$$
$$\hat{V} = \frac{1}{n}\sum g(Z_i, \hat{\theta}_E)g(Z_i, \hat{\theta}_E)'.$$

6 Endogeneity and Instrumental Variables

Consider a linear regression model

$$y_i = x_i'\beta_0 + u_i$$

Recall from example 2 that if

$$E(u_i|X_i) = 0 \to E(u_iX_i) = 0 \to E(Y_i - X_i'\beta)X_i = 0$$

then $\hat{\beta}_{GMM} = \hat{\beta}_{OLS} = (X'X)^{-1}X'Y$. In this situation, without any endogeneity, OLS is efficient.

If, on the other hand, we still have the linear regression model but $E(u_i X_i) \neq 0$, then we have endogeneity. $E(u_i X_i) = g(\beta)$ is not correct, indeed, we cannot construct moments in this case.

An alternative approach finds instruments z_i such that x and z are correlated and z is not correlated with u, and employs them to estimate. This is called a two-stage instrumental variables estimator.

Formally, consider a linear regression model with a k x 1 vector β_0 and $E(u_i, x_i) \neq 0$. Suppose we can find an $m \times 1$ vector of instruments z_i such that $e(u_i z_i) = 0$. This implies the following moment condition holds:

$$\begin{aligned} E[(y_i - x_i'\beta_0)z_i] &= 0 \\ g(y_i; x_i, z_i, \beta_0) &= (y_i - x_i'\beta_0)z_i \end{aligned}$$

Suppose first that $k = m$, that is, we have the same number of instruments as parameters. Then we can of course throw away $m - k$ instruments but we have found that this will result in a loss of efficiency since we do not exploit all the information available. To use all the information, we consider the linear combinations of the m instruments that yield the most efficient IV estimators. The optimal GMM weighting matrix gives use the optimal combination of instruments that yields the most efficient IV estimator. Let

$$\overline{g}(\beta) = \frac{1}{n}\sum g(y_i; x_i, z_i, \beta_0) = \frac{1}{n}Z'(y - X\beta).$$

We have

$$V = E[g(y_i; x_i, z_i, \beta_0)g(y_i; x_i, z_i, \beta_0)'] = \sigma^2 E[z_i z_i'].$$

Since σ^2 is a constant, the optimal weight matrix is:

$$W_n = \left(\frac{1}{n}\sigma_{i=1}^n z_i z_i'\right)^{-1} = \left(\frac{1}{n}Z'Z\right)^{-1}$$

Since W_n is free of unknown parameters, the efficient GMM procedure in this case does not require the preliminary estimation of β_0 in the first step.

The criterion function with this choice of weight matrix is given by

$$S_n(\beta) = \frac{1}{n}(y - X\beta)'Z(Z'Z)^{-1}Z'(y - X\beta)$$

Taking the derivative, we have the first order condition:

$$X'P_z(y - X\hat{\beta}) = 0$$

where $P_z = Z(Z'Z)^{-1}Z'$. So the optimal IV estimator is:

$$\begin{aligned} \hat{\beta}_{2SLS} &= (X'P_Z X)^{-1}X'P_Z y \\ &= (X^{*\prime}X^*)^{-1}X^{*\prime}y \end{aligned}$$

Practically, we obtain this two-stage least squares IV estimator from the following two steps:

1. Obtain fitted value X^*
   Regress y on Z to get the fitted value $y^* = P_x y$. Then regress X on Z to get the fitted value $X^* = p_z X$

2. estimate $\hat{\beta}_{2SLS}$
   Regress y^* on X^* to obtain

$$\hat{\beta}_{2SLS} = (X' P_Z X)^{-1} X' P_Z y = (X^{*\prime} X^*)^{-1} X^{*\prime} y$$

This is our 2SLS estimator. Note that if there is no endogeneity, it is more efficient to use GMM since the introduction of two stages introduces inefficiency.

7 Testing for Overidentifying Restrictions

Suppose we want to test the validity of the population moment conditions. The null and alternative hypothesis are

$$H_0 : E[g(X_i, \theta_0)] = 0 \text{ for some } \theta_0 \in \Theta$$

$$H_1 : E[g(X_i, \theta_0)] \neq 0 \text{ for any } \theta_0 \in \Theta$$

Assume $m > k$, that is, that the number of moment conditions exceeds the number of parameters to be estimated. Note that, in this case, we have $m - k$ overidentifying restrictions because only k moment conditions are needed to identify k parameters. We use the over-identifying restrictions to test the null hypothesis.

The idea to construct the test statistic is simple: let θ be a GMM estimator. Under H_0, we expect that the sample analogue of the population moment conditions

$$\overline{g}(\hat{\theta}) = \frac{1}{n} \sum g(X_i, \hat{\theta})$$

is close to zero. It will not be identically equal to zero unless $m = k$. So we can take $\overline{g}(\hat{\theta})$ as the basis of our test statistic. The test statistic is:

$$T_n = n * S_n(\hat{\theta}_e) = n * \hat{g}(\hat{\theta})' \hat{V}^{-1} \hat{g}(\hat{\theta})$$

where

$$\hat{V} = \frac{1}{n} \sum g(X_i, \hat{\theta}) g(X_i, \hat{\theta})'$$

Under H_0, we have

$$T_n \xrightarrow{d} \chi^2(m - k)$$

We reject $H - 0$ if $T_n > \chi^2_\alpha(m - k)$ at the significance level α.

$$\sqrt{n}(\hat{\theta}-\theta) \approx -\sqrt{n}\left(\frac{\partial^2 S_n(\theta)}{\partial\theta\,\partial\theta^T}\right)^{-1}\left(\frac{\partial S_n(\theta)}{\partial\theta}\right)$$

$$\frac{\partial^2 S_n(\theta)}{\partial\theta\,\partial\theta^T} \xrightarrow{P} 2G^T W G = 2G^T V^{-1} G \quad \text{for } W = V^{-1}$$

$$\bar{g}(\hat{\theta}) \approx \bar{g}(\theta) + \frac{\partial \bar{g}(\theta)}{\partial\theta}(\hat{\theta}-\theta)$$

$$\sqrt{n}\,\frac{\partial S_n(\theta)}{\partial\theta} = 2\left(\frac{\partial \bar{g}(\theta)}{\partial\theta}\right)^T W_n\,\bar{g}(\theta)$$

$$\approx 2G^T V^{-1}\bar{g}(\theta)$$

$$\frac{\partial \bar{g}(\theta)}{\partial\theta} \xrightarrow{P} G$$

$$\Rightarrow \sqrt{n}(\hat{\theta}-\theta) \approx -\left(2G^T V^{-1} G\right)^{-1} 2G^T V^{-1}\bar{g}(\theta)$$

$$= -(G^T V^{-1} G)^{-1} G^T V^{-1}\bar{g}(\theta)$$

$$\bar{g}(\hat{\theta}) \approx \bar{g}(\theta) - G(G^T V^{-1} G)^{-1} G^T V^{-1}\bar{g}(\theta)$$

$$= \left[I - G(G^T V^{-1} G)^{-1} G^T V^{-1}\right]\bar{g}(\theta)$$

$$A = V^{1/2} W G = V^{-1/2} G$$

$$= \left[I - G(A^T A)^{-1} A^T V^{-1/2}\right]\bar{g}(\theta)$$

$$= \left[I - V^{1/2} V^{-1/2} G (A^T A)^{-1} A^T V^{-1/2}\right)\bar{g}(\theta)$$

$$V^{-1/2}\bar{g}(\hat{\theta}) \approx \left(V^{-1/2} - A(A^TA)^{-1}A^TV^{-1/2}\right)\bar{g}(\theta)$$

$$\approx \left(I - \underline{A(A^TA)^{-1}A^T}\right)V^{-1/2}\bar{g}(\theta)$$

$$\approx M_A V^{-1/2}\bar{g}(\theta) \quad \searrow P_A$$

rank(P_A) $= k$

$$S_n(\hat{\theta}) \approx \bar{g}(\hat{\theta})\, V^{-1/2}\bar{g}(\hat{\theta})$$

~~$= \bar{g}(\theta)^T V^{-1/2} M_A V^{-1/2} \bar{g}(\theta)$~~

$$n\, S_n(\hat{\theta}) = \left[\sqrt{n}\, V^{-1/2}\bar{g}(\theta)\right]^T M_A \left(\sqrt{n}\, V^{-1/2}\bar{g}(\theta)\right)$$

$$\sqrt{n}\, V^{-1/2}\bar{g}(\theta) \xrightarrow{H_0} N_\ell(0, I)$$

$$\rightarrow \chi^2_{r(M_A)},$$

where $r(M_A) = \ell - k$ with ℓ = dimension of $g(\cdot)$ and k = dimension of Θ.