This version: November 16, 2016

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 moments. It is particularly useful in estimating structural models in which moment conditions can be derived from economic theory. GMM emerges as one of the most popular estimators in modern econometrics, and it includes conventional methods like the two-stage least squares (2SLS) and the three-stage least square as special cases.

1.1 Examples of Endogeneity

As econometricians mostly work with non-experimental data, we cannot overstate the importance of the endogeneity problem. We go over a few examples.

Example 1 (Dynamic Panel Model)**.** We know that the first-difference (FD) estimator is consistent for (static) panel data model. Nevertheless, the FD estimator encounters difficulty in a dynamic panel model

$$y_{it} = \beta_1 + \beta_2 y_{it-1} + \beta_3 x_{it} + \alpha_i + \epsilon_{it},$$

even if we assume

$$\mathbb{E}\left[\epsilon_{it} | \alpha_i, x_{i1}, \ldots, x_{iT}, y_{it-1}, y_{it-2}, \ldots, y_{i0}\right] = 0. \tag{1}$$

When taking difference of the above equation for periods t and $t-1$, we have

$$(y_{it} - y_{it-1}) = \beta_2 (y_{it-1} - y_{it-2}) + \beta_3 (x_{it} - x_{it-1}) + (\epsilon_{it} - \epsilon_{it-1}).$$

Under (1), $\mathbb{E}[(x_{it} - x_{it-1})(\epsilon_{it} - \epsilon_{it-1})] = 0$, but

$$\mathbb{E}[(y_{it-1} - y_{it-2})(\epsilon_{it} - \epsilon_{it-1})] = -\beta_2 \mathbb{E}[y_{it-1}\epsilon_{it-1}] = -\beta_2 \mathbb{E}\left[\epsilon_{it-1}^2\right] \neq 0. \quad \square$$

Example 2 (Keynesian-Type Macro Equations)**.** This is a model borrowed from Hayashi (2000, p.193) but originated from Haavelmo (1943). An econometrician is interested in learning β_2, the marginal propensity of consumption, in the Keynesian-type equation

$$C_i = \beta_1 + \beta_2 Y_i + u_i \tag{2}$$

where C_i is household consumption, Y_i is the GNP, and u_i is the unobservable error. However, Y_i and C_i are connected by an accounting equality (with no error)

$$Y_i = C_i + I_i,$$

where I_i is investment. We assume $\mathbb{E}[u_i|I_i] = 0$ as investment is determined in advance. OLS (2) will be inconsistent because in the reduced-form $Y_i = \frac{1}{1-\beta_2}(\beta_1 + u_i + I_i)$ implies $\mathbb{E}[Y_i u_i] = \mathbb{E}\left[u_i^2\right]/(1-\beta_2) \neq 0$. $\square$

Example 3 (Classical Measurement Error)**.** Endogeneity also emerges when an explanatory variables is not directly observable but is replaced by a mea-

surement with error. Suppose the true linear model is

$$y_i = \beta_1 + \beta_2 x_i^* + u_i, \tag{3}$$

with $\mathbb{E}[u_i|x_i^*] = 0$. We cannot observe x_i^* but we observe x_i, a measurement of x_i^*, and they are linked by

$$x_i = x_i^* + v_i$$

with $\mathbb{E}[v_i|x_i^*, u_i] = 0$. Such a formulation of the measurement error is called the *classical measurement error*. When we substitute out the unobservable x_i^* in (3), we have

$$y_i = \beta_1 + \beta_2 (x_i - v_i) + u_i = \beta_1 + \beta_2 x_i + e_i \tag{4}$$

where $e_i = u_i - \beta_2 v_i$. The correlation

$$\mathbb{E}[x_i e_i] = \mathbb{E}[(x_i^* + v_i)(u_i - \beta_2 v_i)] = -\beta_2 \mathbb{E}\left[v_i^2\right] \neq 0.$$

OLS (4) would not deliver a consistent estimator. □

Example 4 (Demand-Supply System)**.** See Hansen's Chapter 15.

1.2 GMM in Linear Model

In this section we discuss GMM in a linear single structural equation. A structural equation is a model of economic interest. For example, (2) is a

structural equation in which β_2 can be interpreted as the marginal propensity of consumption. Consider the following linear structural model

$$y_i = x_{1i}\beta_1 + z_{1i}\beta_2 + \epsilon_i, \tag{5}$$

where x_{1i} is a k_1-dimensional endogenous explanatory variables, z_{1i} is a k_2-dimensional exogenous explanatory variables with the intercept included. In addition, we have z_{2i}, a k_3-dimensional excluded exogenous variables. Let $K = k_1 + k_2$ and $L = k_2 + k_3$. Denote $x_i = (x_{1i}, z_{1i})$ as a K-dimensional explanatory variable, and $z_i = (z_{1i}, z_{2i})$ as an L-dimensional exogenous vector. In the context of endogeneity, we can call the exogenous variable instrument variables, or simply instruments. Let $\beta = (\beta_1', \beta_2')'$ be a K-dimensional parameter of interest. From now on, we rewrite (5) as

$$y_i = x_i\beta + \epsilon_i, \tag{6}$$

and we have a vector of instruments z_i.

Before estimating any structural econometric model, we must check identification. A model is *identified* if there is a one-to-one mapping between the distribution of the observed variables and the parameters. In other words, in an identified model any two parameter values β and $\tilde{\beta}$, $\beta \neq \tilde{\beta}$, cannot generate exactly the same distribution for the observable data. In the context of (6), identification requires that the true value β_0 is the only value on the parameters space that satisfies the moment condition

$$\mathbb{E}\left[z_i'\left(y_i - x_i\beta\right)\right] = 0_L. \tag{7}$$

The rank condition is sufficient and necessary for identification.

Assumption (Rank condition)**.** $\operatorname{rank}(\mathbb{E}[x_i' z_i]) = K$.

Note that $\mathbb{E}[x_i' z_i]$ is a $K \times L$ matrix. The rank condition implies the *order condition* $L \geq K$, which says that the number of excluded instruments must be no fewer than the number of endogenous variables.

Theorem. *The parameter in (7) is identified if and only if the rank condition holds.*

Proof. (The "if" direction). For any $\tilde{\beta}$ such that $\tilde{\beta} \neq \beta_0$,

$$\mathbb{E}\left[z_i'\left(y_i - x_i\tilde{\beta}\right)\right] = \mathbb{E}\left[z_i'\left(y_i - x_i\beta_0\right)\right] + \mathbb{E}\left[z_i' x_i\right]\left(\beta_0 - \tilde{\beta}\right) = 0_K + \mathbb{E}\left[z_i' x_i\right]\left(\beta_0 - \tilde{\beta}\right).$$

Because $\operatorname{rank}(\mathbb{E}[x_i' z_i]) = K$, we would have $\mathbb{E}[z_i' x_i]\left(\beta_0 - \tilde{\beta}\right) = 0_L$ if and only if $\beta_0 - \tilde{\beta} = 0_K$, which violates $\tilde{\beta} \neq \beta_0$. Therefore β_0 is the unique value that satisfies (7).

(The "only if" direction is left as an exercise. Hint: By contrapositiveness, if the rank condition fails, then the model is not identified. We can easily prove the claim by making an example.) □

Because identification is a prerequisite for structural estimation, from now on we always assume that the model is identified. When it is just-identified ($L = K$), by (7) we can express the parameter as

$$\beta = \left(\mathbb{E}\left[z_i' x_i\right]\right)^{-1} \mathbb{E}\left[z_i' y_i\right]. \tag{8}$$

It follows by the principle of method of moments that

$$\hat{\beta} = \left(\frac{Z'X}{n} \right)^{-1} \frac{Z'y}{n} = \left(Z'X \right)^{-1} Z'y,$$

which is exactly the 2SLS when $L = K$. In the rest of this section, we focus on the over-identified case ($L > K$). When $L > K$, (7) involves more equations than the number of parameters, directly taking the inverse as in (8) is inapplicable.

In order to express β explicitly, we define a criterion function

$$Q(\beta) = \mathbb{E}\left[z_i' (y_i - x_i \beta) \right]' W \mathbb{E}\left[z_i' (y_i - x_i \beta) \right],$$

where W is an arbitrary $L \times L$ positive-definite symmetric matrix. Because of the quadratic form, $Q(\beta) \geq 0$ for all β. Identification indicates that $Q(\beta) = 0$ if and only if $\beta = \beta_0$. Therefore we conclude

$$\beta_0 = \arg\min_{\beta} Q(\beta).$$

Since $Q(\beta)$ is a smooth function of β, the minimizer β_0 can be characterized by the first-order condition

$$0_K = \frac{\partial}{\partial \beta} Q(\beta_0) = -\mathbb{E}\left[z_i' x_i \right]' W \mathbb{E}\left[z_i' (y_i - x_i \beta_0) \right] = -\mathbb{E}\left[x_i' z_i \right] W \mathbb{E}\left[z_i' (y_i - x_i \beta_0) \right]$$

Rearranging the above equation, we have

$$\mathbb{E}\left[x_i' z_i \right] W \mathbb{E}\left[z_i' x_i \right] \beta_0 = \mathbb{E}\left[x_i' z_i \right] W \mathbb{E}\left[z_i' y_i \right].$$

Denote $\Sigma = \mathbb{E}[z_i' x_i]$. Under the rank condition, $\Sigma' W \Sigma$ is invertible so that

we can solve

$$\beta_0 = \left(\Sigma' W \Sigma\right)^{-1} \Sigma' W \mathbb{E}\left[z_i' y_i\right].$$

In practice, we use the sample moments to replace the corresponding population moments. The GMM estimator mimics its population formula.

$$\begin{aligned}
\widehat{\beta} &= \left(\frac{1}{n}\sum x_i' z_i W \frac{1}{n}\sum z_i' x_i\right)^{-1} \frac{1}{n}\sum x_i' z_i W \frac{1}{n}\sum z_i' y_i \\
&= \left(\frac{X'Z}{n} W \frac{Z'X}{n}\right)^{-1} \frac{X'Z}{n} W \frac{Z'y}{n} \\
&= \left(X'ZWZ'X\right)^{-1} X'ZWZ'y.
\end{aligned}$$

Exercise. The same GMM estimator $\hat{\beta}$ can be obtained by minimizing

$$\left[\frac{1}{n}\sum_{i=1}^{n} z_i'\left(y_i - x_i\beta\right)\right]' W \left[\frac{1}{n}\sum_{i=1}^{n} z_i'\left(y_i - x_i\beta\right)\right] = \frac{\left(y - X\beta\right)' Z}{n} W \frac{Z'\left(y - X\beta\right)}{n},$$

or more concisely,

$$\hat{\beta} = \arg\min_{\beta} \left(y - X\beta\right)' ZWZ' \left(y - X\beta\right).$$

Now we check the asymptotic properties of $\widehat{\beta}$. A few assumptions are in order.

Assumption (A.1). *$Z'X/n \xrightarrow{\text{p}} \Sigma$ and $Z'\epsilon/n \xrightarrow{\text{p}} 0_L$.*

A.1 assumes that we can apply a law of large numbers, so that that the sample moments $Z'X/n$ and $Z'\epsilon/n$ converge in probability to their population counterparts.

Theorem. *Under A.1, $\widehat{\beta}$ is consistent.*

Proof. The step is similar to the consistency proof of OLS.

$$\begin{aligned}\widehat{\beta} &= \left(X'ZWZ'X\right)^{-1} X'ZWZ'\left(X'\beta_0 + \epsilon\right) \\ &= \beta_0 + \left(\frac{X'Z}{n} W \frac{Z'X}{n}\right)^{-1} \frac{X'Z}{n} W \frac{Z'\epsilon}{n} \xrightarrow{\text{p}} \beta_0.\end{aligned}$$ □

To check asymptotic normality, we assume that a central limit theorem can be applied.

Assumption (A.2). $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i'\epsilon_i \Rightarrow N\left(0_L, \Omega\right)$, *where* $\Omega = \mathbb{E}\left[z_i' z_i \epsilon_i^2\right]$.

Theorem (Asymptotic Normality). *Under A.1 and A.2,*

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \Rightarrow N\left(0_K, \left(\Sigma'W\Sigma\right)^{-1} \Sigma'W\Omega W\Sigma \left(\Sigma'W\Sigma\right)^{-1}\right). \tag{9}$$

Proof. Multiply $\widehat{\beta} - \beta_0$ by the scaling factor $\sqrt{n}$,

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta} - \beta_0\right) &= \left(\frac{X'Z}{n} W \frac{Z'X}{n}\right)^{-1} \frac{X'Z}{n} W \frac{Z'\epsilon}{\sqrt{n}} \\ &= \left(\frac{X'Z}{n} W \frac{Z'X}{n}\right)^{-1} \frac{X'Z}{n} W \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i'\epsilon_i.\end{aligned}$$

The conclusion follows as $\frac{X'Z}{n} W \frac{Z'X}{n} \xrightarrow{\text{p}} \Sigma'W\Sigma$ and $\frac{X'Z}{n} W \frac{1}{\sqrt{n}} \sum z_i'\epsilon_i \Rightarrow \Sigma'W \times N\left(0, \Omega\right)$. □

It is clear from (9) that the GMM estimator's asymptotic variance depends on the choice of W. A natural question follows: can we optimally choose a W to make the asymptotic variance as small as possible? Here we claim the result without a proof.

Claim. The choice $W = \Omega^{-1}$ makes $\widehat{\beta}$ an asymptotically efficient estimator, under which the asymptotic variance is

$$\left(\Sigma'\Omega^{-1}\Sigma\right)^{-1}\Sigma'\Omega^{-1}\Omega\Omega^{-1}\Sigma\left(\Sigma'\Omega^{-1}\Sigma\right)^{-1} = \left(\Sigma'\Omega^{-1}\Sigma\right)^{-1}.$$

In practice, Ω is unknown but can be estimated. Hansen (1982) suggests the following procedure, which is known as the *two-step GMM*.

1. Choose any valid W, say $W = I_L$, to get a consistent (but inefficient in general) estimator $\hat{\beta}$. Save the residual $\widehat{\epsilon}_i = y_i - x_i\widehat{\beta}$ and estimate the variance matrix $\widehat{\Omega} = \frac{1}{n}\sum z_i'z_i\widehat{\epsilon}_i^2$.

2. Set $W = \widehat{\Omega}^{-1}$ and obtain a second estimator

$$\widehat{\beta} = \left(X'Z\widehat{\Omega}^{-1}Z'X\right)^{-1}X'Z\widehat{\Omega}^{-1}Z'y.$$

   This second estimator is asymptotic efficient.

If we further assume conditional homoskedasticity, then $\Omega = \mathbb{E}\left[z_i'z_i\epsilon_i^2\right] = \mathbb{E}\left[z_i'z_i\mathbb{E}\left[\epsilon_i^2|z_i\right]\right] = \sigma^2\mathbb{E}\left[z_i'z_i\right]$. Therefore in the first-step of the two-step GMM we can estimate the variance of the error term by $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{\epsilon}_i^2$ and the variance matrix by $\widehat{\Omega} = \widehat{\sigma}^2\frac{1}{n}\sum_{i=1}^n z_i'z_i = \widehat{\sigma}^2 Z'Z/n$. When we plug this $W = \widehat{\Omega}^{-1}$ into the GMM estimator,

$$\begin{aligned}\widehat{\beta} &= \left(X'Z\left(\widehat{\sigma}^2\frac{Z'Z}{n}\right)^{-1}Z'X\right)^{-1}X'Z\left(\widehat{\sigma}^2\frac{Z'Z}{n}\right)^{-1}Z'y \\ &= \left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1}X'Z\left(Z'Z\right)^{-1}Z'y.\end{aligned}$$

This is exactly the same expression of 2SLS for $L > K$. Therefore, 2SLS can be viewed as a special case of GMM with $W = (Z'Z/n)^{-1}$. Under conditional homoskedasticity, 2SLS is the efficient estimator; otherwise 2SLS is inefficient.

1.3 GMM in Nonlinear Model

The principle of GMM can be used in models where the parameter enters the moment conditions nonlinearly. Let $g_i(\beta) = g(w_i, \beta) \mapsto \mathbb{R}^L$ be a function of the data w_i and the parameter β. If economic theory implies $\mathbb{E}[g_i(\beta)] = 0$, we can write the GMM population criterion function as

$$Q(\beta) = \mathbb{E}[g_i(\beta)]' W \mathbb{E}[g_i(\beta)]$$

Example. Nonlinear models nest the linear model as a special case. For the linear IV model in the previous section, the data is $w_i = (y_i, x_i, z_i)$, and the moment function is $g(w_i, \beta) = z_i'(y_i - x_i\beta)$.

In practice we use the sample moments to mimic the population moments in the criterion function

$$Q_n(\beta) = \left(\frac{1}{n}\sum_{i=1}^n g_i(\beta)\right)' W \left(\frac{1}{n}\sum_{i=1}^n g_i(\beta)\right).$$

The GMM estimator is defined as

$$\hat{\beta} = \arg\min_{\beta} Q_n(\beta).$$

In these nonlinear models, a closed-form solution is in general unavailable,

while the asymptotic properties can still be established. We state these asymptotic properties without proofs.

Theorem. *If the model is identified, and*

$$\mathbb{P}\left[\sup_{\beta}\left|\frac{1}{n}\sum_{i=1}^{n} g_i(\beta) - \mathbb{E}\left[g_i(\beta)\right]\right| > \varepsilon\right] \to 0$$

for any constant $\varepsilon > 0$*, then* $\hat{\beta} \xrightarrow{\text{p}} \beta$*. If in addition* $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\beta_0) \Rightarrow N(0,\Omega)$*, then*

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \Rightarrow N\left(0, \left(\Sigma' W \Sigma\right)^{-1}\left(\Sigma' W \Omega W \Sigma\right)\left(\Sigma' W \Sigma\right)^{-1}\right)$$

where $\Sigma = \mathbb{E}\left[\frac{\partial}{\partial \beta'} g_i(\beta_0)\right]$ *and* $\Omega = \mathbb{E}\left[g_i(\beta_0)\, g_i(\beta_0)'\right]$*. If we choose* $W = \Omega^{-1}$*, then the GMM estimator is efficient, and the asymptotic variance becomes* $\left(\Sigma'\Omega^{-1}\Sigma\right)^{-1}$*.*

Remark. The list of assumptions in the above statement is incomplete. We only lay out the key conditions but neglect some technical details.

$Q_n(\beta)$ measures how close are the moments to zeros. It can serve as a test statistic with proper formulation. Under the null hypothesis $\mathbb{E}\left[g_i(\beta)\right] = 0_L$, this so-called "$J$-test" checks whether a moment condition is violated. The test statistic is

$$\begin{aligned} J\left(\widehat{\beta}\right) &= n\left(\frac{1}{n}\sum_{i=1}^{n} g_i\left(\widehat{\beta}\right)\right)' \widehat{\Omega}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g_i\left(\widehat{\beta}\right)\right) \\ &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i\left(\widehat{\beta}\right)\right)' \widehat{\Omega}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i\left(\widehat{\beta}\right)\right) \end{aligned}$$

where $\widehat{\Omega}$ is a consistent estimator of Ω, and $\widehat{\beta}$ is an efficient estimator, for example, the second $\widehat{\beta}$ from the two-step GMM. This statistics converges in distribution to a chi-square random variable with degree of freedom $L-K$. That is, under the null,

$$J\left(\widehat{\beta}\right) \Rightarrow \chi^2\left(L-K\right)$$

If the null hypothesis is false, then the test statistic tends to be large, and it is more likely to reject the null.