

Lecture Note 10

Generalized Method of Moments

Method of moments:

Let y_1, \dots, y_n be a random sample from a population whose pdf is given by $f(y; \theta_0)$. Suppose that the pdf is such that

$$E(y^r) = m_r(\theta_0). \quad (10.1)$$

The method of moments estimator for θ_0 is defined as the solution to

$$\frac{1}{n} \sum_{i=1}^n y_i^r = m_r(\hat{\theta}_n). \quad (10.2)$$

Note that equation (10.1) holds in the *population*, and equation (10.2) is the corresponding *sample analog*. That is, the *expectation* in (10.1) is replaced with the average in (10.2).

Below we derive the properties of such an estimator.

Note first that if $E(|y^r|) < \infty$, then by the SLLN we have that

$$\frac{1}{n} \sum_{i=1}^n y_i^r \xrightarrow{p} E(y^r),$$

and hence we should hope that $\hat{\theta}_n \xrightarrow{p} \theta_0$, as well.

Example: Exponential distribution

$$f(y; \theta_0) = \begin{cases} \theta_0 e^{-\theta_0 y} & \text{if } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

We have that

$$E(y) = \frac{1}{\theta_0}.$$

Hence, an estimator for θ_0 is a solution to

$$\frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{\hat{\theta}_n},$$

or

$$\hat{\theta}_n = \frac{1}{\bar{y}},$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Generalize method of moments (GMM):

Notation:

y dependent variable

x vector of independent variables (regressors)

θ vector of parameters of interest. We also denote the parameter space by Θ ($\theta \in \Theta$).

θ_0 is the value of the parameter vector θ in the population. Note that it must be that $\theta_0 \in \Theta$.

The population joint density of x and y , conditional on the true parameter vector θ_0 , is given by

$$f(y, x; \theta_0) = f(y|x; \theta_0) g(x),$$

where $f(y|x; \theta_0)$ represents the (parametric) model and $g(x)$ is the density of x , which does not depend on θ_0 .

The data that we have is a random sample (y_i, x_i) , $i = 1, \dots, n$, from the population joint distribution of (y, x) .

The estimator:

The generalized method of moments (GMM) estimator is derived from matching the sample moments with their population counterparts.

In the population there is a set of *moment conditions* that identifies the population parameter vector θ_0 , that is,

$$E_0 [\varphi(y, x; \theta)] = 0 \iff \theta = \theta_0. \tag{10.3}$$

Example: OLS in linear regression model

$$\varphi(y, x; \beta) = (y - x'\beta) x.$$

Then, at the true β , i.e., β_0 , we have

$$\begin{aligned} \varphi(y, x; \beta_0) &= (y - x'\beta_0) x, \\ &= \varepsilon x, \end{aligned}$$

and hence we have (by the law of iterated expectations)

$$E_0 [\varphi(y, x; \beta_0)] = E_0 [\varepsilon x] = 0. \quad (10.4)$$

Note that

$$\begin{aligned} E_0 [\varphi(y, x; \beta)] &= E_0 [(y - x'\beta_0 + x'\beta_0 - x'\beta) x], \\ &= E_0 [(xx') (\beta_0 - \beta)]. \end{aligned}$$

If $E_0 (xx')$ is of full rank, that is, there is no perfect collinearity in the x 's, then $\beta = \beta_0$ is a unique solution to (10.4).

The GMM estimator is defined as the solution to the sample counterpart of (10.3), that is,

$$\frac{1}{n} \sum_{i=1}^n \varphi(y, x; \hat{\theta}_n) = 0. \quad (10.5)$$

There are many cases in economics in which we have a model that implies an equation of the form of (10.3). This is the reason why the GMM is very popular among economists.

For example, consider consumption allocation in two periods:

$$\begin{aligned} \max_{y_1, y_2} u(y_1, y_2, x; \theta_0) \\ \text{s.t. } p_1 y_1 + p_2 y_2 = I, \end{aligned}$$

where y_1 and y_2 are the consumptions in the two periods, p_1 and p_2 are the prices in the two period, and I is the overall income in the two periods.

The first-order condition for the optimal allocation is given by

$$\frac{\frac{\partial}{\partial y_1} u(y_1, y_2, x; \theta_0)}{\frac{\partial}{\partial y_2} u(y_1, y_2, x; \theta_0)} = \frac{p_1}{p_2},$$

so that

$$E \left[\frac{\frac{\partial}{\partial y_1} u(y_1, y_2, x; \theta_0)}{\frac{\partial}{\partial y_2} u(y_1, y_2, x; \theta_0)} - \frac{p_1}{p_2} \right] = 0.$$

Hence, the parameters of the utility function can be estimated by solving:

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{\frac{\partial}{\partial y_1} u(y_{i1}, y_{i2}, x; \hat{\theta}_n)}{\frac{\partial}{\partial y_2} u(y_{i1}, y_{i2}, x; \hat{\theta}_n)} - \frac{p_1}{p_2} \right) = 0.$$

GMM with more equations than unknown parameters:

Note that the GMM starts from the fact that in the population we have

$$E_0 [\varphi(y, x; \theta)] = 0, \quad (10.6)$$

that has a unique solution at $\theta = \theta_0$. This requires that the number of moment conditions be at least as large as the number of parameters. Equation (10.6) represents the *population moment conditions*. The corresponding *sample moment conditions* are provided by

$$m_n(\theta) = \frac{1}{n} \sum_{i=1}^n \varphi(y, x; \theta) = 0. \quad (10.7)$$

But, if there are more equations than unknown parameters we cannot guarantee that there will be an estimator $\hat{\theta}_n$ for which all the moment conditions in (10.7) will hold.

Definition of GMM estimator:

The GMM estimator is defined by

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} Q_n(\theta),$$

where

$$Q_n(\theta) = m_n(\theta)' V_n^{-1} m_n(\theta), \quad (10.8)$$

V_n is a stochastic matrix such that

$$V_n \xrightarrow{p} V,$$

and V is a non-stochastic matrix.

Consistency of GMM estimator:

If

$$E_0 \left[\sup_{\theta \in \Theta} |\varphi(y, x; \theta)| \right] < \infty,$$

then we have that

$$m_n(\theta) \xrightarrow{p} E_0 [\varphi(y, x; \theta)],$$

uniformly over Θ .

Hence,

$$\begin{aligned} Q_n(\theta) &= m_n(\theta)' V_n^{-1} m_n(\theta), \\ &\xrightarrow{p} E_0 [\varphi(y, x; \theta)]' V^{-1} E_0 [\varphi(y, x; \theta)], \end{aligned}$$

uniformly over Θ .

The limit has a unique minimum, equals to 0, at $\theta = \theta_0$.

Using very similar method of proof as we used for the maximum likelihood estimator we can establish that

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

That is, the GMM estimator is consistent estimator for the corresponding population parameter vector.

Asymptotic normality GMM estimator:

In order to obtain the GMM estimator one need to get the first-order conditions and that requires differentiation of $Q_n(\theta)$ in (10.8). This involves use of the chain rule for differentiation of functions of several variables.

For function f and g of one variable the chain rule is that the derivative of $f \circ g$ is $f'(g) \cdot g'$.

The derivative of a function of several variables is the matrix of the linear approximation for that function. This implies that the derivative of

$$f : R^K \longrightarrow R^M$$

is an $M \times K$ matrix. This is somewhat in conflict with what we used earlier, that is

$$\frac{\partial x'Ax}{\partial x} = 2Ax.$$

If we satisfy the rule that the derivative is the matrix of linear approximation, then we should have

$$\frac{\partial x'Ax}{\partial x} = 2x'A.$$

Of course, the difference is just a matter of convention, but for the chain rule the linear approximation convention is (somewhat) easier.

Let M be the number of moment conditions.

Let K be the number of parameters.

Then

$$Q_n(\theta) = h(m_n(\theta)),$$

where

$$h(m) = m'V_n^{-1}m.$$

Hence we have,

$$\frac{\partial h(m)}{\partial m'} 2m'V_n^{-1}.$$

Now, note that $\theta \in R^K$, the function $m_n(\theta) \in R^M$, and $Q_n(\theta) \in R$. Hence, using the chain rule we get

$$\frac{\partial}{\partial \theta'} Q_n(\theta) = \left(\frac{\partial}{\partial m_n(\theta)'} h(m_n(\theta)) \right) \frac{\partial}{\partial \theta'} m_n(\theta).$$

Here, $\partial m_n(\theta)/\partial \theta'$ is an $M \times K$ matrix and $\partial h(m_n(\theta))/\partial m_n(\theta)'$ is a $1 \times M$ vector. Hence, $\partial Q_n(\theta)/\partial \theta'$ is a $1 \times K$ vector, where

$$\begin{aligned} \frac{\partial}{\partial \theta'} m_n(\theta) &= \begin{pmatrix} \frac{\partial}{\partial \theta_1} m_{n,1}(\theta), & \dots, & \frac{\partial}{\partial \theta_K} m_{n,1}(\theta) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \theta_1} m_{n,M}(\theta), & \dots, & \frac{\partial}{\partial \theta_K} m_{n,M}(\theta) \end{pmatrix}, \quad \text{and} \\ \frac{\partial}{\partial m_n(\theta)'} h(m_n(\theta)) &= 2m_n(\theta)' V_n^{-1}. \end{aligned}$$

Hence, the first-order conditions (organized in column vector) are given by:

$$2 \frac{\partial}{\partial \theta} m_n(\hat{\theta}_n) V_n^{-1} m_n(\hat{\theta}_n) = 0. \quad (10.9)$$

A Taylor-series expansion of $\sqrt{n} m_n(\hat{\theta}_n)$ around θ_0 gives:

$$\sqrt{n} m_n(\hat{\theta}_n) = \sqrt{n} m_n(\theta_0) + \frac{\partial}{\partial \theta'} m_n(\theta_n^*) \sqrt{n} (\hat{\theta}_n - \theta_0), \quad (10.10)$$

where θ_n^* is on the line segment connecting $\hat{\theta}_n$ and θ_0 .

Substitution of (10.10) into (10.9) gives:

$$\frac{\partial}{\partial \theta} m_n(\hat{\theta}_n) V_n^{-1} \sqrt{n} m_n(\theta_0) + \frac{\partial}{\partial \theta} m_n(\hat{\theta}_n) V_n^{-1} \frac{\partial}{\partial \theta'} m_n(\theta_n^*) \sqrt{n} (\hat{\theta}_n - \theta_0) = 0. \quad (10.11)$$

Now, if

$$E_0 \left[\sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_k} \varphi_j(y, x; \theta) \right| \right] < \infty,$$

for $j = 1, \dots, M$; $k = 1, \dots, k$, then

$$\begin{aligned} \frac{\partial}{\partial \theta_k} m_{nj}(\theta) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta_k} \varphi_j(y_i, x_i; \theta) \\ &\xrightarrow{a.s.} E_0 \left[\frac{\partial}{\partial \theta_k} \varphi_j(y, x; \theta) \right], \end{aligned}$$

uniformly over Θ .

Hence,

$$\frac{\partial}{\partial \theta} m_n(\hat{\theta}_n) \xrightarrow{p} E_0 \left[\frac{\partial}{\partial \theta} \varphi(y_i, x_i; \hat{\theta}_n) \right] \equiv A(\theta_0),$$

with $A(\theta_0)$ a $K \times M$ matrix.

Consider now the term

$$\sqrt{nm_n}(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(y_i, x_i; \theta_0).$$

This is a sum of i.i.d. random variables which satisfy

$$\begin{aligned} E_0[\varphi(y, x; \theta_0)] &= 0, \quad \text{and} \\ V_0(\varphi(y, x; \theta_0)) &= E_0[\varphi(y, x; \theta_0)\varphi(y, x; \theta_0)'] \equiv W(\theta_0), \end{aligned}$$

with $W(\theta_0)$ an $M \times M$ matrix. Hence we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(y_i, x_i; \theta_0) \xrightarrow{D} N(0, W(\theta_0)).$$

Note now that in the expansion in (10.11) we have

$$\begin{aligned} \frac{\partial}{\partial \theta} m_n(\hat{\theta}_n) &\xrightarrow{p} A(\theta_0), \\ \frac{\partial}{\partial \theta} m_n(\theta_n^*) &\xrightarrow{p} A(\theta_0), \\ V_n^{-1} &\xrightarrow{p} V^{-1}, \quad \text{and} \\ \sqrt{nm_n}(\theta_0) &\xrightarrow{D} N(0, W(\theta_0)). \end{aligned} \tag{10.12}$$

Hence, substitution of the terms in (10.12) into (10.11) gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, \Lambda(\theta_0)),$$

where

$$\Lambda(\theta_0) = \left(A(\theta_0)V^{-1}A(\theta_0)' \right)^{-1} A(\theta_0)V^{-1}W(\theta_0)V^{-1}A(\theta_0)' \left(A(\theta_0)V^{-1}A(\theta_0)' \right)^{-1}. \tag{10.13}$$

Note that if

$$V = W(\theta_0),$$

then (10.13) simplifies to

$$\Lambda(\theta_0) = \left(A(\theta_0)V^{-1}A(\theta_0)' \right)^{-1}. \tag{10.14}$$

Is it meaningful that $V = W(\theta_0)$?

Consider first $\Lambda^{-1}(\theta_0)$:

$$\Lambda^{-1}(\theta_0) = \left(A(\theta_0)V^{-1}A(\theta_0)' \right) \left(A(\theta_0)V^{-1}W(\theta_0)V^{-1}A(\theta_0)' \right)^{-1} \left(A(\theta_0)V^{-1}A(\theta_0)' \right). \tag{10.15}$$

Define

$$\begin{aligned} C &\equiv A(\theta_0)V^{-1}(W(\theta_0))^{1/2}, \quad \text{and} \\ D &\equiv A(\theta_0)V^{-1}(W(\theta_0))^{-1/2}, \end{aligned}$$

with

$$(W(\theta_0))^{1/2}(W(\theta_0))^{1/2} = W(\theta_0).$$

(This is possible because $W(\theta_0)$ is positive definite, and so is $W^{-1}(\theta_0)$.)

Hence, we can rewrite (10.15) as

$$\Lambda^{-1}(\theta_0) = DC'(CC')^{-1}CD.$$

Now, by Cauchy-Schwartz inequality we have

$$\begin{aligned} \Lambda^{-1}(\theta_0) &\leq DD' \\ &= A(\theta_0)W^{-1}(\theta_0)A(\theta_0)', \end{aligned}$$

or

$$\Lambda(\theta_0) \geq \left(A(\theta_0)W^{-1}(\theta_0)A(\theta_0)'\right)^{-1}.$$

Conclusion: $V = W(\theta_0)$ minimizes the variance of the GMM estimator. That is, the optimal *weight* matrix is the inverse of the variance of the moment conditions.

Example: Binary logit model

Define

$$\varphi(y, x; \theta) = \left(y - \frac{e^{x'\theta}}{1 + e^{x'\theta}}\right) h(x),$$

for some function $h(\cdot)$ that depends only on x , but not on y .

Then

$$\begin{aligned} E_0[\varphi(y, x; \theta_0)] &= E_0\left[\left(y - \frac{e^{x'\theta_0}}{1 + e^{x'\theta_0}}\right) h(x)\right], \\ &= 0. \end{aligned}$$

Consequently we have

$$\begin{aligned} W(\theta_0) &= E_0[\varphi(y, x; \theta_0)\varphi(y, x; \theta_0)'] \\ &= E_x\left[\frac{e^{x'\theta_0}}{(1 + e^{x'\theta_0})^2}h(x)h(x)'\right]. \end{aligned}$$

Now, since

$$\frac{\partial}{\partial \theta} \varphi(y, x; \theta) = \frac{e^{x'\theta}}{(1 + e^{x'\theta})^2} xh(x),$$

we have

$$A(\theta_0) = E_x \left[\frac{e^{x'\theta_0}}{(1 + e^{x'\theta_0})^2} xh(x)' \right].$$

Hence, the optimal variance is

$$\Lambda(\theta_0) = \left(E_x \left[\frac{e^{x'\theta_0}}{(1 + e^{x'\theta_0})^2} xh(x)' \right] \left(E_x \left[\frac{e^{x'\theta_0}}{(1 + e^{x'\theta_0})^2} h(x)h(x)' \right] \right)^{-1} E_x \left[\frac{e^{x'\theta_0}}{(1 + e^{x'\theta_0})^2} h(x)x' \right] \right)^{-1}.$$

We specialize now for a scalar θ_0 and a scalar value function $h(\cdot)$. Then

$$\Lambda(\theta_0) = \frac{E_x \left[\frac{e^{x\theta_0}}{(1+e^{x\theta_0})^2} h^2(x) \right]}{E_x \left[\frac{e^{x\theta_0}}{(1+e^{x\theta_0})^2} xh(x) \right]^2}.$$

Consider

$$\Lambda^{-1}(\theta_0) = \frac{E_x \left[\frac{e^{x\theta_0}}{(1+e^{x\theta_0})^2} xh(x) \right]^2}{E_x \left[\frac{e^{x\theta_0}}{(1+e^{x\theta_0})^2} h^2(x) \right]}, \quad (10.16)$$

and define

$$\begin{aligned} a(x) &= \frac{\sqrt{e^{x\theta_0}}}{1 + e^{x\theta_0}} h(x), \quad \text{and} \\ b(x) &= \frac{\sqrt{e^{x\theta_0}}}{1 + e^{x\theta_0}} x. \end{aligned} \quad (10.17)$$

Substitution of $a(x)$ and $b(x)$ from (10.17) into (10.16) gives

$$\Lambda^{-1}(\theta_0) = \frac{E_x [a(x)b(x)]^2}{E_x [a^2(x)]} \leq E_x [b^2(x)] = E_x \left[\frac{e^{x\theta_0}}{(1 + e^{x\theta_0})^2} x^2 \right].$$

Hence, we see that

$$\Lambda(\theta_0) \geq \left(E_x \left[\frac{e^{x\theta_0}}{(1 + e^{x\theta_0})^2} x^2 \right] \right)^{-1},$$

with the equality holds only if $h(x) = x$.

Now note that

$$\frac{1}{n} \sum_{i=1}^n \left(y_i - \frac{e^{x_i \hat{\theta}_n}}{1 + e^{x_i \hat{\theta}_n}} \right) x_i = 0$$

is the first-order condition for the MLE.

Conclusion: The GMM estimator with the smallest variance is the MLE. This is true not only in this example, but in general. The MLE is (*asymptotically*) *efficient*, that is, the asymptotic variance is equal to the Cramer-Rao lower bound.