Econ 2120: Section 8

Nonlinear GMM and MLE

Ashesh Rambachan

Fall 2018

Outline

Extremum Estimators

An estimator $\hat{\theta}$ is an **extremum estimator** if it is defined as

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

where $\Theta \subset \mathbb{R}^K$. $Q_n(\cdot)$ is some function of our sample data.

You'll spend a lot of 2140 studying the properties of extremum estimators. GMM is a special case of this class of estimators.

There is a lot to be said about extremum estimators.

References: Newey & McFadden (1994), Hayashi Ch. 8.

Outline

Set up

As before, the data are $W_1, \ldots, W_n$ i.i.d. for $i = 1, \ldots, n$.

We are given a moment function $\psi(\cdot, \cdot)$ that satisfies

$$E[\psi(W_i, \gamma)] = 0$$

for some *unique* γ in the parameter space, where as a function of γ $\psi : \mathbb{R}^K \to \mathbb{R}^L$ with $L \geq K$.

This is our moment condition.

Assuming γ is unique is assuming the model is **identified** – there are more primitive assumptions that can be made to justify this.

Example: Euler Equations

Recall:

$$E[\delta R_{t+1} \frac{u'(c_{t+1})}{u'(c_t)} - 1 | I_t] = 0,$$

where R_{t+1} = rate of return on savings instrument, δ = discount rate, I_t = information up to date t. If $u(c) = \frac{c^{1-\sigma}}{1-\sigma}$, then

$$E[\delta R_{t+1} (\frac{c_{t+1}}{c_t})^{-\gamma} - 1 | I_t] = 0.$$

So, let x_t be a vector of variables that are known at date t (i.e. $x_t \in I_t$). This implies that

$$E[x_t(\delta R_{t+1} (\frac{c_{t+1}}{c_t})^{-\gamma} - 1) | I_t] = 0 \implies E[x_t(\delta R_{t+1} (\frac{c_{t+1}}{c_t})^{-\gamma} - 1)] = 0.$$

This is a moment equation with

$$\gamma = (\delta, \gamma), \quad \psi((x_t, R_{t+1}), \gamma) = x_t(\delta R_{t+1} (\frac{c_{t+1}}{c_t})^{-\gamma} - 1).$$

GMM Estimator

As before, we define the **sample moment function**:

$$g_n(a) = \frac{1}{n}\sum_{i=1}^{n}\psi(W_i, a).$$

By the LLN, $g_n(\gamma) \xrightarrow{p} E[\psi(W_i, \gamma)] = 0$. So, it's natural to consider an estimator $\hat{\gamma}$ that sets

$$g_n(\hat{\gamma}) \approx 0.$$

Just vs. Over-identified

If $L = K$, we are **just-identified**. Try to directly solve the non-linear system of equation (e.g. apply some root-finding algorithm).

If $L > K$, we are **over-identified**. As before, we can introduce a $K \times L$ weight matrix $\hat{D}$ and solve for

$$\hat{D} g_n(\hat{\gamma}) = 0.$$

Or, we can consider the minimum norm problem

$$\hat{\gamma} = \arg\min_a g(a)' \hat{C} g(a),$$

where $\hat{C}$ is $L \times L$ and conv. in prob. to a positive definite, symmetric matrix C. From the FOC, we see there's a one-to-one mapping between $\hat{C}$ and $\hat{D}$ with

$$\hat{D} = (\underset{K \times L}{\frac{\partial g_n(\hat{\gamma})}{\partial a}})' \hat{C}.$$

Outline

Consistency

For our purposes, we will assume that

$$\hat{\gamma} \xrightarrow{p} \gamma.$$

There are primitive conditions that will ensure this holds – (identification + uniform convergence; See Newey-McFadden (1994), Hayashi Ch 8, Wooldridge Ch 14.)

Given consistency, we will provide a heuristic sketch of the proof of asymptotic normality.

Key tool: The delta method.

Asymptotics: Just identified case

Our estimator satisfies

$$g_n(\hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\psi(W_i, \hat{\gamma}) = 0$$

and consider

$$\sqrt{n}g_n(\hat{\gamma}) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\psi(W_i, \hat{\gamma}) = 0.$$

Apply the mean-value theorem to get

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\psi(W_i, \gamma) + (\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi(W_i, \gamma^*)}{\partial a})\sqrt{n}(\hat{\gamma} - \gamma) = 0$$

$$\sqrt{n}g_n(\gamma) + \frac{\partial g_n(\gamma^*)}{\partial a}\sqrt{n}(\hat{\gamma} -) = 0,$$

where γ^* is somewhere on the segment connecting $\hat{\gamma}$ and γ.

Asymptotics: Just identified case

Re-arrange and we get that

$$\sqrt{n}(\hat{\gamma} - \gamma) = -(\frac{\partial g_n(\gamma^*)}{\partial a})^{-1}(\sqrt{n} g_n(\gamma)).$$

We have that $g_n(\gamma) \xrightarrow{p} E[\psi(W_i, \gamma) = 0$. So, apply a CLT to $\sqrt{n} g_n(\gamma)$ and get

$$\sqrt{n} g_n(\gamma) \xrightarrow{d} N(0, E[\psi(W_i, \gamma)\psi(W_i, \gamma)']).$$

Then, $\gamma^* \xrightarrow{p} \gamma$ because it is sandwiched between $\hat{\gamma}$ and γ. So, we have that

$$(\frac{\partial g_n(\gamma^*)}{\partial a})^{-1} \xrightarrow{p} E[\frac{\partial \psi(W_i, \gamma)}{\partial a}]^{-1}$$

Here I used CMT and that if $\theta_n \xrightarrow{p} \theta$ and $Q_n(\theta) \xrightarrow{p} Q(\theta)$, then

$$Q_n(\theta_n) \xrightarrow{p} Q(\theta).$$

We're also assuming that $E[\frac{\partial \psi(W_i, \gamma)}{\partial a}]$ is invertible.

Asymptotics: Just identified case

So, we have that

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, \alpha\Sigma\alpha'),$$

where

$$\alpha' = \alpha = E[\underset{K\times K}{\frac{\partial\psi(W_i, \gamma)}{\partial a}}]^{-1},$$

$$\Sigma = E[\psi(W_i, \gamma)\psi(W_i, \gamma)'].$$

Asymptotics: Over-identified case

Lecture notes walk through a similar argument to derive the asymptotic distribution in the over-identified case. I restate it below.

Claim: In the over-identified case,

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, \alpha\Sigma\alpha'),$$

where

$$\underset{K\times L}{\alpha} = [\underset{K\times L}{D} E[\frac{\partial\psi(W_i,\gamma)}{\underset{L\times K}{\partial a}}]]^{-1} \underset{K\times L}{D},$$

$$\underset{L\times L}{\Sigma} = E[\psi(W_i,\gamma)\psi(W_i,\gamma)'].$$

Outline

Optimal Weight Matrix

For the over-identified case, we had to select a weight matrix $\hat{D} \xrightarrow{p} D$. Importantly, the limiting distribution depend on D.

Different choices of D led to different asymptotic covariance matrices.

Natural Q: What weight matrix should we select?

One idea: Select the weight matrix to "minimize" asymptotic variance

Let $Avar_D$ be the asymptotic covariance matrix of $\hat{\gamma}$ with weight matrix D.

Minimizing asymptotic variance means finding the D^* such that

$$Avar_{D^*} - Avar_D$$

is positive semi-definite for all other choices of D. This implies that the asymptotic variance of any linear combination of $\hat{\gamma}$ is minimized at D^*.

Optimal Weight Matrix

Claim:

$$\underset{K\times L}{D^*} = E[\underset{K\times L}{\frac{\partial\psi(W_i,\gamma)}{\partial a}}]' E[\underset{L\times L}{\psi(W_i,\gamma)\psi(W_i,\gamma)'}]^{-1}.$$

Then,

$$Avar_{D^*} = (E[\frac{\partial\psi(W_i,\gamma)}{\partial a}]' E[\psi(W_i,\gamma)\psi(W_i,\gamma)']^{-1} E[\frac{\partial\psi(W_i,\gamma)}{\partial a}])^{-1}.$$

Outline

Two-step GMM

We have an optimal weight matrix – that's great. But we need to construct a sample analogue of it. Here's way to get it:

(1) Pick some weight matrix, e.g. $\hat{C} = I$, and compute $\hat{\delta}(I)$ by solving the GMM objective for this choice of weight matrix.

(2) Use $\hat{\delta}$ to plug-in and construct an estimate of $\hat{D}^*, \hat{C}^*$ – just use the sample analogue.

(3) Estimate $\hat{\delta}(D^*)$ by minimizing the GMM objective for this choice of weight matrix, $\hat{D}^*, \hat{C}^*$.

Natural Question: Do we have to adjust our standard errors since we estimate the optimal weight matrix? No. It will still deliver the efficient estimator.

Two-step GMM – simple example

Suppose we want to estimate $\sigma^2 = V(Y_i) = E[(Y_i - \mu)^2]$, where Y_i is 1×1. The sample analogue is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Let's use GMM to derive the asymptotic variance of $\hat{\sigma}^2$.

Keep in mind we want to know: Do we need to correct for the fact that we estimated μ?

To do so, we'll construct the following moment function

$$\psi(Y_i, (\mu, \sigma)) = \begin{pmatrix} \psi_1(Y_i, \mu) \\ \psi_2(Y_i, (\mu, \sigma)) \end{pmatrix} = \begin{pmatrix} Y_i - \mu \\ (Y_i - \mu)^2 - \sigma^2 \end{pmatrix},$$

where $\sigma = vec(\Sigma)$.

Two-step GMM – simple example

The moment condition is

$$E[\psi(Y_i, (\mu, \sigma))] = 0.$$

We can solve this in *two steps*:

(1) $E[\psi_1(Y_i, \mu)] = 0 \implies \mu = E[Y_i]$.

(2) Plug $\mu = E[Y_i]$ into $E[\psi_2(Y_i, (\mu, \sigma))] = 0$ and its immediate that $\sigma^2 = V(Y_i)$.

The sample counterpart is

$$\frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, (\hat{\mu}, \hat{\sigma})) = 0.$$

Again, solve this in two steps:

(1) $\frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\mu} = 0 \implies \hat{\mu} = \bar{Y}$.

(2) Plug $\hat{\mu} = \bar{Y}$ into $\frac{1}{n}\sum_{i=1}^{n} \psi_2(Y_i, (\hat{\mu}, \hat{\sigma})) = 0$ to get that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y})^2$.

Two-step GMM – simple example

Now, we can apply our GMM results from earlier to derive the limiting distribution of $\hat{\mu}, \hat{\sigma}$. In particular, let $\gamma = (\mu', \sigma')'$. Then, we have that

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, \Lambda),$$

where

$$\Lambda = \alpha \Sigma \alpha', \quad \alpha = E[\frac{\partial \psi(W_i, \gamma)}{\partial a}]^{-1}, \quad \Sigma = E[\psi(W_i, \gamma)\psi(W_i, \gamma)'].$$

We're only interested in Λ_{22}. Can we write what that is? Note that

$$\frac{\partial \psi(W_i, \gamma)}{\partial a} = \begin{pmatrix} E[\frac{\partial \psi_1(Y_i, \mu)}{\partial \mu}] & 0 \\ 0 & E[\frac{\partial \psi_2(Y_i, (\mu, \sigma^2))}{\partial \sigma^2}] \end{pmatrix}$$

So, if you multiply through $\Lambda_{22} = E[\frac{\partial \psi_2(Y_i, (\mu, \sigma^2))}{\partial \sigma^2}]^{-1} E[\psi_2(Y_i, (\mu, \sigma^2))\psi_2(Y_i, (\mu, \sigma^2))'] E[\frac{\partial \psi_2(Y_i, (\mu, \sigma^2))}{\partial \sigma^2}]^{-1}$.

Two-step GMM

The asymptotic variance is unaffected by the fact that we estimated μ and simply plugged in our estimate directly.

This turns out to be general! Suppose our moment function has the form

$$\psi(W_i, (a_1, a_2)) = \begin{pmatrix} \psi_1(W_i, a_1) \\ \psi_2(W_i, (a_1, a_2)) \end{pmatrix},$$

where we partitioned our parameters into $\gamma = (\gamma_1, \gamma_2)$. Assume $dim(\psi_1) = dim(a_1), dim(\psi_2) = dim(a_2)$ – i.e. just identified case. We're only interested in γ_2.

The sample analogue is

$$\frac{1}{n} \sum_{i=1}^{n} \psi(W_i, (\hat{\gamma}_1, \hat{\gamma}_2)) = 0.$$

We solve this in two steps:

(1) Solve $\frac{1}{n} \sum_{i=1}^{n} \psi_1(W_i, \hat{\gamma}_1) = 0$ to get $\hat{\gamma}_1$.
(2) Plug this in and solve $\frac{1}{n} \sum_{i=1}^{n} \psi_2(W_i, (\hat{\gamma}_1, \hat{\gamma}_2))$ to get $\hat{\gamma}_2$.

Two-step GMM

If

$$E[\frac{\partial \psi_2(W_i, (\gamma_1, \gamma_2))}{\partial a_1}] = 0,$$

then

$$\sqrt{n}(\hat{\gamma}_2 - \gamma_2) \xrightarrow{d} N(0, \alpha_{22}\Sigma_{22}\alpha_{22}'),$$

where

$$\alpha_{22} = E[\frac{\partial \psi_2(W_i, (\gamma_1, \gamma_2))}{\partial a_2}]^{-1}, \Sigma_{22} = E[\psi_2(W_i, (\gamma_1, \gamma_2))\psi_2(W_i, (\gamma_1, \gamma_2))].$$

Maximum Likelihood Estimation

We'll now transition to another estimation strategy – maximum likelihood estimation. This is an example of an **M-estimator**, which is another class of **extremum estimators**.

Can think of maximum likelihood estimation as a maximum a posteori estimator with a flat prior.

If you believe that the likelihood is well-specified, then MLE has some extremely attractive properties.

Today: Just want to set-up the MLE estimator and next time, we'll discuss its properties.

Set Up

As always, assume random sampling $W_i \sim F$ i.i.d for $i = 1, \ldots, n$. We specify a set of distributions

$$\{P_\theta : \theta \in \Theta\}.$$

Each distribution has a density that is well-defined. We denote it $f(w|\theta)$. For a given θ, $f(w|\theta)$ is the **likelihood function**.

We say the model is **well-specified** if for some $\theta^* \in \Theta$,

$$F = P_{\theta^*}.$$

If it is possible that there does not exist such a θ^*, we refer to f as a **pseudo-likelihood** or a **quasi-likelihood**.

Set Up

We could also set this up in terms of **conditional likelihoods.** That is, divide W_i into Y_i, Z_i and then we model the distribution of $Y_i|Z_i$. Our family of distributions

$$\{P_\theta : \theta \in \Theta\}$$

is now a set of conditional likelihood functions.

> **Example**: Normal linear model – we modeled $Y_i|X_i \sim N(X_i'\beta, \sigma^2)$. So, the set of distributions was the normal family with mean parameterized by β and variance σ^2.

MLE: Definition

We observe $W_1, \ldots, W_n$. Fix a particular value of θ. The joint density of $W_1, \ldots, W_n|\theta$ is given by

$$f(W_1, \ldots, W_n|\theta) = \Pi_{i=1}^n f(W_i|\theta).$$

The **maximum likelihood estimator** is the value of the parameter that maximizes the likelihood of the observed data.

$$\hat{\theta}^{MLE} = \arg\max_{\theta \in \Theta} \Pi_{i=1}^n f(W_i|\theta).$$

Equivalently, we can take the log of the objective function because it is a monotone function. And define the **log-likelihood**

$$L_n(W_1, \ldots, W_n|\theta) = \sum_{i=1}^n \log(f(W_i|\theta))$$

Then,

$$\hat{\theta}^{MLE} = \arg\max_{\theta \in \Theta} L_n(W_1, \ldots, W_n|\theta).$$

Probit and Logit applications

Not going to cover this in section – I assume this is material you have seen before. If not, make sure to read these sections of Note 10.

Next time: What are the properties of MLE? How can we derive its limit distribution?