AGEC 652 - Lecture 5.2

Review of nonlinear estimation: MLE and GMM

Diego S. Cardoso

Spring 2022

Course roadmap

1. Intro to Scientific Computing
2. Numerical operations and representations
3. Systems of equations
4. Optimization
5. **Structural estimation**
    1. Introduction
    2. **Review of estimation methods** $\leftarrow$ You are here
    3. Estimation of single and multiple agent models

*These slides draw from Wooldridge (2010) and course materials by Matt Woerman (UMass) and Chris Conlon (NYU Stern).

Why do we need MLE and GMM?

Sometimes, OLS is all we need

- The parameters we need to estimate are linear in the model
- Or they can be made linear with some clever transformation (like logs)

But, in many many cases, we need to estimate parameters that enter nonlinearly in the model

- We need to resort to nonlinear estimators

MLE and GMM are the most used estimators for structural modeling and estimation

Maximum Likelihood in one slide

1. We observe some data $(y_i, x_i), i = 1, \ldots, N$ and assume it comes from a joint distribution described by parameter vector θ

2. For any given θ we can calculate the joint probability of our data

   - If observations are i.i.d., this joint probability is a product of individual probabilities of drawing (y_i, x_i)

3. Using Bayes' rule, we can calculate the probability of θ given (y_i, x_i): the **likelihood function**

4. The MLE estimate is the value of θ that maximizes the likelihood: *we pick the parameters that make it most likely to generate the observed data*

Maximum Likelihood intuition

Suppose we draw five numbers from a normal distribution:

$$y = 47.3, 51.2, 50.5, 44.9, 53.1$$

And we consider two candidate distributions: $N(0, 1)$ or $N(50, 1)$

- What is the likelihood of $N(0, 1)$ generating y? Virtually zero
- What is the likelihood of $N(50, 1)$ generating y? Definitely greater

So, between these two, we pick $\mu = 50$, since it's *more likely* to generate y than $\mu = 0$

Maximum Likelihood example

Linear regression: we have $Y_i = X_i\beta + \epsilon_i$ and assume $\epsilon_i|X_i \sim N(0, \sigma^2)$. This implies $Y_i|X_i \sim N(X_i\beta, \sigma^2)$

Given the parameters and i.i.d. observations, the data come from a joint distribution (ϕ is the standard normal PDF)

$$Pr(Y_1, \ldots, Y_N|X_1, \ldots, X_N; \beta, \sigma^2) = \prod_{i=1}^{N} \Pr(Y_i|X_i, \beta, \sigma^2) = \prod_{i=1}^{N} \phi(Y_i - X_i\beta; 0, \sigma^2)$$

By Bayes' rule

$$L(\beta, \sigma^2|X, Y) = \prod_{i=1}^{N} \Pr(\beta, \sigma^2|Y_i, X_i) \propto \prod_{i=1}^{N} \Pr(Y_i|X_i, \beta, \sigma^2)$$

Maximum Likelihood example

Then, we use optimization methods to find

$$(\hat{\beta}_{MLE}, \hat{\sigma}^2_{MLE}) = \arg\max_{\beta,\sigma^2} L(\beta, \sigma^2 | X, Y)$$

- We can show that this solution is analytically equivalent to OLS
- As we saw it in the optimization tutorial, in practice we take logs to transform that product inside L into a sum
  - We maximize the *log-likelihood function* $l(\beta, \sigma^2 | X, Y)$

Generalized Method of Moments in one slide

1. Our economic model defines the following population moment conditions: at the true parameter θ_0, $g(x;\theta)$ are on average equal to zero

$$E[g(x;\theta_0)] = 0$$

2. We observe some data $x_i, i = 1, \ldots, N$ and calculate sample analogue

$$E[g(x;\theta)] \approx \frac{1}{N}\sum_{i=1}^{N} g(x_i;\theta) \equiv g_N(\theta)$$

3. The GMM estimate is given by (W_N is a weighting matrix)

$$\hat{\theta}_{GMM} = \arg\min_{\theta} g_N(\theta)' \, W_N \; g_N(\theta)$$

(Generalized) Method of Moments intuition

Suppose we draw five numbers from an unknown distribution

$$y = 47.3, 51.2, 50.5, 44.9, 53.1$$

Suppose this unknown distribution has mean μ, giving a population moment condition

$$E[y_i] = \mu \Rightarrow E[y_i - \mu] = 0$$

We expect the population moment condition to also hold in the sample analogue

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu) = 0$$

(Generalized) Method of Moments intuition

Forget the weighting matrix for now. We have one condition and one parameter, so this is effectively a special case: just *Method of Moments*

For our estimate, we pick $\hat{\mu}$ that minimizes

$$\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu})\right)^2$$

In this simple case, this is just solving for $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu}) = 0$

Generalized Method of Moments example

Linear regression: again, we have $Y_{it} = X_i\beta + \epsilon_i$. But now, instead of normality, we assume ϵ_i is orthogonal to data X_i (a $1 \times K$ vector)

$$E[x_{ki}\epsilon_i] = 0 \Rightarrow E[x_{ki}(Y_i - X_i\beta)] = 0$$

- This actually gives K moment conditions: one for each variable $x_{ki} \in X_i$

We replace these K conditions with their respective sample analogues and solve for

$$\frac{1}{N}\sum_{i=1}^{N} x_{ki}(Y_i - X_i\hat{\beta}) = 0$$

- We can show that this solution is analytically equivalent to OLS, too

MLE vs GMM

It's the same old **bias (or robustness) vs. efficiency trade-off**

- With MLE, we need to make assumption on distributions of unobservables
  - When our assumptions are correct, MLE is more efficient $\rightarrow$ lower variance
  - Has good small sample properties (less bias, more efficiency with small data)
  - If our assumptions are inadequate, estimates are more biased

- With GMM, we don't need to assume distributions and can rely only on moment conditions from the theoretical and statistical model
  - This is more robust $=$ less bias
  - Has good large sample properties (less bias, more efficiency with large data)
  - But it's in general less efficient than MLE $\rightarrow$ higher variance

Choosing between MLE and GMM

- *How much data is available?*
  - Large data sets favor GMM: good large sample properties require fewer assumptions. Smaller data sets might require stronger distributional assumptions $\rightarrow$ MLE
- *How complex is the model?*
  - MLE is better suited for linear and quadratic models, but technically difficult to compute with highly nonlinear models. For the latter case, GMM might be better
- *How comfortable are you making distributional assumptions?*
  - MLE requires you to fully specify distributions. If there is good theoretical grounding for these assumptions, MLE is a good idea. Otherwise, GMM is the more attractive option

Up next

We are going to review MLE and GMM with a focus on application: properties of these estimators and how to implement them in practice

We won't cover

- Proofs of asymptotic properties
- Small sample properties
- Hypothesis testing and model selection statistics
- Specialized numerical methods for their estimation

Some good textbooks to learn about these details: Wooldridge (PhD-level), Hayashi, Hansen, Greene

Maximum Likelihood Estimation

MLE: General case

1. Start with the **joint density of the data** $z_1, \ldots, z_N$ given by $f_Z(Z;\theta)$

2. Assuming an i.i.d. sample, construct the **log likelihood function**[1]

$$l(\theta|Z) = \log(\prod_{i=1}^{N} f_Z(z_i;\theta)) = \sum_{i=1}^{N} \log f_Z(z_i;\theta)$$

3. Compute $\hat{\theta}_{MLE} = \arg\max_\theta l(\theta|Z)$

4. Compute $Var(\hat{\theta}_{MLE})$, the variance-covariance matrix of the estimates

[1] We take logs to simplify computation. Log is a positive monotonic transformation, so it preserves the max.

Properties of MLE

Under some *regularity conditions*, MLE has the following properties

1. Consistency
2. Asymptotic normality
3. Asymptotic efficiency
4. Invariance

Properties of MLE: Consistency

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$$

As sample size grows to infinity, $\hat{\theta}_{MLE}$ gets arbitrarily close to the true parameter value, θ_0

Properties of MLE: Asymptotic normality

$$\sqrt{N}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} \mathrm{N}\left(0, I(\theta_0)^{-1}\right)$$

where $I(\theta_0)$ is the Fisher Information Matrix, given by

$$I(\theta_0) = -E\left[\frac{\partial^2 l(\theta_0)}{\partial\theta_0 \partial\theta_0'}\right]$$

As the sample size grows to infinity, the distribution of $\hat{\theta}_{MLE}$ converges to a normal distribution with mean as the true parameter value and a particular Variance-Covariance structure

Properties of MLE: Asymptotic normality

Note that $I(\theta_0)$ is the expectation of the Hessian of l evaluated at the true parameter

- This has a meaningful intuition: we are more certain of MLE estimates when the (log-) likelihood function has more curvature!

The asymptotic variance-covariance matrix is then given by

$$Var(\hat{\theta}_{MLE}) = \{-E\,[\frac{\partial^2 l(\theta_0)}{\partial\theta_0 \partial\theta_0'}]\}^{-1}$$

Properties of MLE: Asymptotic efficiency

$\hat{\theta}_{MLE}$ achieves the Cramér-Rao lower bound

$$Var(\hat{\theta}_{MLE}) = I(\theta_0)^{-1}$$

- No consistent estimator has lower asymptotic variance than the MLE

Properties of MLE: Invariance

Let $f(\theta_0)$ be a continuous and continuously differentiable function. Then

$$\hat{f}(\theta_0)_{MLE} = f(\hat{\theta}_{MLE})$$

- The MLE of a function of θ is the function applied to the $\hat{\theta}_{MLE}$

MLE variance estimator

The variance-covariance matrix of the MLE can be estimated using

$$Var(\hat{\theta}_{MLE}) = -\{ \frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\Big|_{\theta=\hat{\theta}_{MLE}} \}^{-1}$$

So we calculate the Hessian of l at the estimated parameter values.

- This is the simplest variance estimator. More robust estimators exist but are beyond our scope here

Computing $\hat{\theta}_{MLE}$

We can use any of the maximization methods we've seen so far to calculate $\hat{\theta}_{MLE}$

- Unconstrained optimization with `Optim` (you can use the log/exp transformation to avoid domain problems with negative σ^2)
- Constrained optimization with `JuMP` (can set a constraint for $\sigma^2 \geq 0$)
- If closed-form derivatives of l are easy to obtain, you can use nonlinear rootfinding methods

There is a specialized Quasi-Newton method for MLE called *BHHH (Berndt-Hall-Hall-Hausman)*

- It has good properties approximating the Hessian of log-likelihoods and is faster than computing the actual Hessian every iteration

Computing $Var(\hat{\theta}_{MLE})$

To estimate the variance-covariance matrix, you can

- Derive the analytic Hessian (usually hard)
- Calculate it numerically using, for example, `ForwardDiff.hessian`

Once you've calculated the variance-covariance matrix, standard errors can be easily calculated as the square root of its diagonal elements

$$SE(\hat{\theta}_{MLE}) = \sqrt{diag(Var(\hat{\theta}_{MLE}))}$$

For a step-by-step example of MLE, please review the optimization tutorial from unit 4

Generalized Method of Moments

GMM: General case

1. Start with data $z_1, \ldots, z_N$ drawn from a population with M moment conditions that are functions of vector θ with $K \leq M$ parameters[1]

$$E[g(Z;\theta)] = 0$$

Where do moment conditions come from?

- Economic model conditions: first-order optimality, market clearing, zero arbitrage, etc
- Statistical assumptions: error orthogonality $(E[x\epsilon] = 0)$
- Instruments orthogonality $(E[z\epsilon] = 0)$
- Model fit: predicted market shares are equal to realized market shares

[1] The "generalized" in GMM comes from allowing more moment conditions than parameters; the "standard" Method of Moments requires M=K

GMM: General case

2. Construct empirical (sample analogue) moment conditions

$$\frac{1}{N}\sum_{i=1}^{N} g(z_i;\theta) = 0$$

GMM: General case

3. Compute the GMM estimate

$$\hat{\theta}_{GMM} = \arg\min_{\theta} Q_N(\theta), \ Q_N(\theta) = [\frac{1}{N}\sum_{i=1}^{N} g(z_i;\theta)]' W [\frac{1}{N}\sum_{i=1}^{N} g(z_i;\theta)]$$

- If $M = K$ and the problem is well-conditioned, then $\frac{1}{N}\sum_{i=1}^{N} g(z_i;\theta) = 0$ is $K \times K$ (non)linear system and we can find the $\hat{\theta}$ that solves it
- But if $M > K$, we almost certainly can't find K parameters that satisfy more that K conditions simultaneously
  - So we look for parameters that get as close as possible to satisfying all moment conditions $\rightarrow$ we minimize deviations from zero, weighted by a $M \times M$ matrix W

GMM: General case

4. Compute $Var(\hat{\theta}_{GMM})$, the variance-covariance matrix of the estimates

- More on that soon

Properties of GMM

Under some *regularity conditions*, GMM has the following properties

1. Consistency
2. Asymptotic normality

- Note that, unlike MLE, GMM is not asymptotically efficient

These properties require some assumptions on the empirical moments

Properties of GMM: empirical moments assumption

We assume the following about empirical moments at the true parameter value, θ_0

1. Empirical moments obey the law of large numbers

$$\frac{1}{N}\sum_{i=1}^{N} g(z_i;\theta_0) \xrightarrow{p} 0$$

2. The derivatives of the empirical moments converge to the $M \times K$ Jacobian matrix

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial g(z_i;\theta)}{\partial \theta'}\bigg|_{\theta=\theta_0} \xrightarrow{p} D_0 \equiv D(\theta_0) = E\,[\frac{\partial g(z_i;\theta_0)}{\partial \theta_0'}]$$

Properties of GMM: empirical moments assumption

3. Empirical moments obey the central limit theorem

$$\sqrt{N}\frac{1}{N}\sum_{i=1}^{N} g(z_i;\theta_0) \xrightarrow{d} \mathrm{N}\left(0, S_0\right)$$

where $S_0 = E[g(z_i;\theta_0)g(z_i;\theta_0)']$ is the variance-covariance matrix of moments (an $M \times M$ matrix)

- We also need to assume that the weighting matrix converges to W_0, a finite symmetric positive definite matrix

$$W \xrightarrow{p} W_0$$

Properties of GMM: consistency

$$\hat{\theta}_{GMM} \xrightarrow{p} \theta_0$$

As sample size grows to infinity, $\hat{\theta}_{GMM}$ gets arbitrarily close to the true parameter value, θ_0

Properties of GMM: Asymptotic normality

$$\sqrt{N}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{d} \mathrm{N}\left(0, V_0\right)$$

where V_0 has a typical *sandwich form*

$$V_0 = \underbrace{(D_0' W_0 D_0)^{-1}}_{\text{bread}} \underbrace{(D_0' W_0 S_0 W_0 D_0)}_{\text{filling}} \underbrace{(D_0' W_0 D_0)^{-1}}_{\text{bread}}$$

As the sample size grows to infinity, the distribution of $\hat{\theta}_{GMM}$ converges to a normal distribution with mean as the true parameter value and a particular Variance-Covariance structure

GMM variance estimator

Any valid weighting matrix W yields a consistent GMM estimator

But the choice of W affects variance, so we want to use some optimal W that minimizes the variance of the estimator

It can be shown that the **optimal weigthing matrix** is given by

$$W_0 = S_0^{-1} = \{E[g(z_i;\theta_0)g(z_i;\theta_0)']\}^{-1}$$

which yields

$$Var(\hat{\theta}_{GMM}) = (D_0' S_0^{-1} D_0)^{-1}$$

This is the "non-robust" VCOV matrix, i.e., assuming homoskedasticity and no clustering/residual correlation structure. Check references to see how to construct robust versions.

Computing $\hat{\theta}_{GMM}$

This is the objective function for GMM

$$\hat{\theta}_{GMM} = \arg\min_{\theta} Q_N(\theta), \; Q_N(\theta) = [\frac{1}{N}\sum_{i=1}^{N} g(z_i;\theta)]' \, W \, [\frac{1}{N}\sum_{i=1}^{N} g(z_i;\theta)]$$

Once again, we can use numerical optimization to calculate $\hat{\theta}_{GMM}$

But what W do we use?

Computing $\hat{\theta}_{GMM}$

Technically, the GMM estimator is consistent for any symmetric positive definite matrix

But different matrices give different variances $\rightarrow$ we want to pick W that minimizes variance

This is, the sample analogue of the **optimal weigthing matrix**

$$\hat{W} = \hat{S}^{-1} = \{E[g(z_i; \hat{\theta})g(z_i; \hat{\theta})']\}^{-1}$$

But we have a chicken and egg problem here: we need $\hat{W}$ to estimate $\hat{\theta}$ but we need $\hat{\theta}$ to get $\hat{W}$!

Computing $\hat{\theta}_{GMM}$: 2-step GMM

One way to solve this problem is to apply a widely-used algorithm: the **2-step GMM**

Step 1

- Using $W = I$ (identity matrix), estimate $\hat{\theta}_1$
  - Alternatives exist, but we are going to stick with the simplest here
- With $\hat{\theta}_1$, calculate $\hat{W} = \{E[g(z_i; \hat{\theta})g(z_i; \hat{\theta})']\}^{-1}$

Computing $\hat{\theta}_{GMM}$: 2-step GMM

Step 2

- Using $\hat{W}$ from step 1, estimate $\hat{\theta}_{GMM}$
- Recalculate $\hat{S}^{-1} = \{E[g(z_i;\hat{\theta})g(z_i;\hat{\theta})']\}^{-1}$
- Calculate $\hat{D} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial g(z_i;\hat{\theta}_{GMM})}{\partial \theta'}$
- Then, calculate the asymptotic Variance-Covariance matrix $Var(\hat{\theta}_{GMM}) = (\hat{D}'\hat{S}^{-1}\hat{D})^{-1}$

Extra: Bootstrap

Issues with asymptotic variance estimators

- We might be interested in the standard errors or confidence intervals of some function $f(\hat{\theta})$
  - One solution is to use the *Delta method*: basically, a first-order Taylor expansion of the asymptotic variance of $\hat{\theta}$
  - Another solution is to resample B times from the data and estimate $\hat{\theta}_b$ for each resample $\rightarrow$ sample from the distribution of $\hat{\theta}$. This the **bootstrap** method
- Consistent estimators might still have large bias in finite samples
  - Bootstrapping is also useful to adjust for this type of bias (provided that the conditions for its correctness are satisfied)

Bootstrap: basic algorithm

- Observations $(z_1, \ldots, z_N)$ are drawn from some measure P, so we can form a nonparametric estimate $\hat{P}$ by assuming that each observation has weight $1/N$

Basic bootstrap algorithm:

1. Simulate a new sample $Z^* = (z_1^*, \ldots, z_N^*) \sim \hat{P}$. This is, draw n values **with replacement** from our data
2. Compute any statistic of $f(Z^*)$ you would like
   - Could be something simple, like an OLS coefficient, or complicated, like Nash equilibrium parameters
3. Repeat 1 and 2 B times and calculate $Var(f_b)$ or $CI(f_1, \ldots, f_B)$

Bootstrap: bias correction

Key idea: $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ approximates the sampling distribution of $\hat{\theta}$

- We can then calculate

$$E[\hat{\theta}^*] = \bar{\theta}^* = \frac{1}{B-1} \sum_{b=1}^{B} \hat{\theta}^*_b$$

Bootstrap: bias correction

- We can use $\bar{\theta}^*$ to bias correct our estimates
  - Recall $\theta = E[\hat{\theta}] - Bias(\hat{\theta})$
  - From bootstrap: $Bias_{bs}(\hat{\theta}) = \bar{\theta}^* - \hat{\theta}$

Then,

$$\hat{\theta} - Bias_{bs}(\hat{\theta}) = \hat{\theta} - (\bar{\theta}^* - \hat{\theta}) = 2\hat{\theta} - \bar{\theta}^*$$

- Most nonlinear models are *consistent but biased*, especially in small samples
  - But correcting bias is not for free: there's always the bias-variance trade-off

Bootstrap: variance

We can also use the sampled values $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ to calculate the **bootstrapped variance** of the estimator

$$Var(\hat{\theta}^*) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_b - \bar{\theta}^*)^2$$

Bootstrap: confidence intervals

We can also calculate **bootstrapped confidence intervals**. There are two basic ways

1. Empirical quantiles (preferred way)
    - Sort values $\hat{\theta}^*_B$ and take

$$CI : [\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2}]$$

2. Asymptotically normal (relies on CLT)

$$CI : \hat{\theta} \pm 1.96\sqrt{Var(\hat{\theta}^*)}$$

Bootstrap isn't magic

Bootstrapped statistics are easy to program

- But for complicated models, it can take a lot of time to resample and estimate multiple times
  - Good thing though: this is highly parallelizable

But bootstrapping isn't magic: it depends on asymptotic theory and will fail if you use it incorrectly

- If you are constructing standard errors for something that isn't asymptotically normal, it won't work
- It samples with replacement $=$ i.i.d. But if i.i.d. does not hold in your data, it might fail (but it can be fixed in certain cases)

Final words

Here we conclude the "theoretical" part of this unit

We will close this course with two interactive examples/tutorials (time permitting) to give you some hands on experience with these methods

- A single-agent model of labor supply with taxes
- A multiple-agent model of Nash-Bertrand competition with discrete choice consumers

Stay tuned: tutorial slides and Jupyter notebooks will be uploaded to the course's repository on GitHub, with links on Brightspace

Thank you!