

CHAPTER 1: INTRODUCTION AND A BRIEF REVIEW OF RELEVANT TOOLS

Joan Llull

Microeconometrics.
IDEA PhD Program. Fall 2017
joan.llull [at] movebarcelona [dot] eu

OVERVIEW

In this course, we review several **econometric methods** used in the analysis of **micro-data**.

Our focus:

Cross-Sectional Data { Discrete Outcomes
Other Limited Dependent Variables

Panel Data { Continuous Outcomes
Discrete Outcomes

Chapter 2: Linear Panel Data.

- Static Models.
- Dynamic Models (with feedback from lagged dependent variable).

Chapter 3: Discrete Outcome Models: *Probit*, *Logit*, and alternatives.

- Dichotomic outcomes (Yes or No).
- Multinomial outcomes (Red, Yellow, Green, or Blue).

Chapter 4: Other limited dependent variables.

- Censoring (top-coded wages).
- Truncation (deductible in insurance).

Chapter 5: Duration analysis.

- (probability of exiting from unemployment conditional on having been unemployed for A months).

Chapter 6: Dynamic Discrete Choice Structural Models.

- (replace a bus engine today vs leave it for tomorrow).

Chapter 1: Review of MLE, GMM, and Numerical Methods.

MAXIMUM LIKELIHOOD

The Likelihood Principle

Likelihood principle: our estimate of θ is the one that maximizes the likelihood of our sample $(\mathbf{y}, X) = ((y_1, \mathbf{x}'_1)', \dots, (y_N, \mathbf{x}'_N)')$.

Likelihood is the “**probability**” of observing the sample, i.e. $\Pr[\mathbf{y}, X; \theta]$ for discrete data; $f(\mathbf{y}, X; \theta)$ for continuous.

The **likelihood function** is $L_N^*(\theta) \equiv f(\mathbf{y}, X; \theta) = f(\mathbf{y}|X; \theta)f(X; \theta)$, which for *i.i.d.* data is $f(\mathbf{y}, X; \theta) = \prod_{i=1}^N f(y_i, \mathbf{x}_i; \theta)$.

We assume $f(X; \theta) = f(X)$ so we can focus on $f(\mathbf{y}|X; \theta)$.

Hence, our object of interest is the (conditional) **log-likelihood function**:

$$\mathcal{L}_N(\theta) \equiv \sum_{i=1}^N \ln f(y_i | \mathbf{x}_i; \theta).$$

The Maximum Likelihood Estimator (MLE)

The MLE is defined by the following optimization problem:

$$\hat{\boldsymbol{\theta}}_{ML} \equiv \arg \max_{\boldsymbol{\theta} \in \Theta} \frac{1}{N} \mathcal{L}_N(\boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta} \in \Theta} \frac{1}{N} \sum_{i=1}^N \ln f(y_i | \mathbf{x}_i; \boldsymbol{\theta}).$$

This estimator is:

- Fully parametric
- An extremum estimator
- An m-estimator

The FOC of the problem is:

$$\frac{1}{N} \sum_{i=1}^N \frac{\partial \ln f(y_i | \mathbf{x}_i; \hat{\boldsymbol{\theta}}_{ML})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

Identification

The *true* parameter vector $\boldsymbol{\theta}_0$ is identified if there are *no observationally equivalent* parameters.

More formally, $\boldsymbol{\theta}_0$ is identified if the *Kullback-Leibler inequality* is satisfied:

$$\Pr[f(y|\mathbf{x}; \boldsymbol{\theta}) \neq f(y|\mathbf{x}; \boldsymbol{\theta}_0)] > 0 \quad \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Regularity conditions

If the following two assumptions hold,

- i. The specified density $f(y|\mathbf{x}; \boldsymbol{\theta})$ is the **data generating process** (dgp)
- ii. The **support** of y does not depend on $\boldsymbol{\theta}$

then the **regularity conditions** are satisfied:

$$\mathbb{E}_f \left[\frac{\partial \ln f(y|\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] = \mathbf{0}$$

$$-\mathbb{E}_f \left[\frac{\partial^2 \ln f(y|\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] = \mathbb{E}_f \left[\frac{\partial \ln f(y|\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(y|\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right].$$

The latter condition is a.k.a. **information matrix equality**.

Consistency

Using identification and the first regularity condition:

$$\mathbb{E}[\ln f(y|\mathbf{x}; \boldsymbol{\theta})] < \mathbb{E}[\ln f(y|\mathbf{x}; \boldsymbol{\theta}_0)] \quad \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

By the Law of Large Numbers (LLN):

$$\frac{1}{N} \sum_{i=1}^N \ln f(y_i|\mathbf{x}_i; \boldsymbol{\theta}) \xrightarrow{p} \mathbb{E}[\ln f(y|\mathbf{x}; \boldsymbol{\theta})].$$

Then, as $N \rightarrow \infty$, $\hat{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta}} \mathcal{L}_N(\boldsymbol{\theta}) \xrightarrow{p} \arg \max_{\boldsymbol{\theta}} \mathcal{L}_0(\boldsymbol{\theta}) = \boldsymbol{\theta}_0$,

whenever:

- i. The parameter space Θ is **compact**
- ii. $\mathcal{L}_N(\boldsymbol{\theta})$ is **measurable** for all $\boldsymbol{\theta}$

Asymptotic distribution

Using a first order *Taylor Expansion* of the FOC around θ_0 :

$$\sqrt{N}(\hat{\theta} - \theta_0) = - \left(\frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \ell_i(\theta^*)}{\partial \theta \partial \theta'} \right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\partial \ell_i(\theta_0)}{\partial \theta},$$

where $\ell_i(\theta) \equiv \ln f(y_i | \mathbf{x}_i; \theta)$, and θ^* is between $\hat{\theta}$ and θ_0 .

Assuming i.i.d. observations and regularity+consistency conditions, by LLN:

$$- \left(\frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \ell_i(\theta^*)}{\partial \theta \partial \theta'} \right)^{-1} \xrightarrow{p} - \mathbb{E} \left[\frac{\partial^2 \ell_i(\theta_0)}{\partial \theta \partial \theta'} \right]^{-1}.$$

By CLT:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\partial \ell_i(\theta_0)}{\partial \theta} \xrightarrow{d} \mathcal{N} \left(0, \mathbb{E} \left[\frac{\partial \ell_i(\theta_0)}{\partial \theta} \frac{\partial \ell_i(\theta_0)}{\partial \theta'} \right] \right).$$

Finally, using Cramer theorem and IM equality:

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N} (0, \Omega_0), \quad \Omega_0 = - \mathbb{E} \left[\frac{\partial^2 \ell_i(\theta_0)}{\partial \theta \partial \theta'} \right]^{-1} = \mathbb{E} \left[\frac{\partial \ell_i(\theta_0)}{\partial \theta} \frac{\partial \ell_i(\theta_0)}{\partial \theta'} \right]^{-1}.$$

Since $\Omega_0 = IM^{-1}$, it is the **Cramer-Rao** lower bound (efficient estimator).

GENERALIZED METHOD OF MOMENTS

General formulation

Let $\boldsymbol{\theta}$ be the parameter vector of interest, defined by the set of moments (or orthogonality conditions):

$$\mathbb{E}[\boldsymbol{\psi}(\boldsymbol{w}; \boldsymbol{\theta})] = \mathbf{0},$$

where

- \boldsymbol{w} is a (vector) random variable,
- and $\boldsymbol{\psi}(\cdot)$ is a vector function such that $\dim(\boldsymbol{\psi}) \geq \dim(\boldsymbol{\theta})$.

Estimation

Consider a random **sample** with N observations $\{\mathbf{w}_i\}_{i=1}^N$.

GMM estimation is based on the **sample moment conditions**:

$$\mathbf{b}_N(\boldsymbol{\theta}) \equiv \frac{1}{N} \sum_{i=1}^N \psi(\mathbf{w}_i; \boldsymbol{\theta}).$$

The **GMM estimator** minimizes the quadratic distance of $\mathbf{b}_N(\boldsymbol{\theta})$ from zero:

$$\hat{\boldsymbol{\theta}}_{GMM} \equiv \arg \min_{\boldsymbol{\theta} \in \Theta} \mathbf{b}_N(\boldsymbol{\theta})' W_N \mathbf{b}_N(\boldsymbol{\theta}),$$

where W_N is semi-positive definite, and $\text{rank}(W_N) \geq \dim(\boldsymbol{\theta})$.

If the problem is **just-identified** ($\dim(\psi) = \dim(\boldsymbol{\theta})$):

$$\mathbf{b}_N(\hat{\boldsymbol{\theta}}_{GMM}) = \mathbf{0}.$$

Consistency

GMM is an **extremum estimator**, so general consistency results hold, as in MLE.

Conditions and intuition are similar to MLE:

- Parameter space $\Theta \in \mathbb{R}^K$ is **compact**.
- $W_N \mathbf{b}_N(\boldsymbol{\theta}) \xrightarrow{p} W_0 \mathbb{E}[\psi(\mathbf{w}; \boldsymbol{\theta})]$.
- **Identification:** $W_0 \mathbb{E}[\psi(\mathbf{w}; \boldsymbol{\theta})] = 0 \Leftrightarrow \boldsymbol{\theta} = \boldsymbol{\theta}_0$.

If these conditions hold, $\hat{\boldsymbol{\theta}}_{GMM} \xrightarrow{p} \boldsymbol{\theta}_0$.

Asymptotic distribution

Following similar steps as in the MLE case, if

- $\hat{\boldsymbol{\theta}}_{GMM}$ is a **consistent estimator** of $\boldsymbol{\theta}_0$,
- $\boldsymbol{\theta}$ is in the **interior** of Θ ,
- $\psi(\mathbf{w}; \boldsymbol{\theta})$ is **once differentiable** with respect to $\boldsymbol{\theta}$,
- $D_N(\boldsymbol{\theta}) \equiv \partial \mathbf{b}_N(\boldsymbol{\theta}) / \boldsymbol{\theta}' \xrightarrow{p} D_0(\boldsymbol{\theta})$, $D_0(\boldsymbol{\theta})$ continuous at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,
- For $D_0 \equiv D_0(\boldsymbol{\theta}_0)$, the matrix $D_0' W_0 D_0$ is **non-singular**,
- $\sqrt{N} \mathbf{b}_N(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, V_0)$, with $V_0 = \mathbb{E}[\psi(\mathbf{w}; \boldsymbol{\theta}_0) \psi(\mathbf{w}; \boldsymbol{\theta}_0)']$,

then $\sqrt{N}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0)$, with:

$$\boldsymbol{\Omega}_0 = (D_0' W_0 D_0)^{-1} D_0' W_0 V_0 W_0 D_0 (D_0' W_0 D_0)^{-1}.$$

Optimal weighting matrix

Efficiency is achieved with any W_N that delivers $W_0 = \kappa V_0^{-1}$.

This includes $W_N = V_0^{-1}$ (**unfeasible**), but also $W_N = \hat{V}_N^{-1}$, where \hat{V}_N is any consistent estimator of V_0 .

Optimal GMM estimator is implemented in two steps:

1. Obtain $\hat{\theta}_{GMM}(W_N^0)$ for an initial guess W_N^0 .
2. Re-estimate using

$$\widehat{W}_{opt} \equiv \left(\sum_{i=1}^N \psi(\mathbf{w}_i; \hat{\theta}_{GMM}(W_N^0)) \psi(\mathbf{w}_i; \hat{\theta}_{GMM}(W_N^0))' \right)^{-1}$$

as the new weighting matrix.

NUMERICAL METHODS

Differentiation

We use the **definition of a derivative**:

$$f'(x) = \lim_{\epsilon \rightarrow 0} \frac{f(x + \epsilon) - f(x)}{\epsilon} \Rightarrow f'(x) \approx \frac{f(x + h) - f(x)}{h},$$

for a small h (e.g. 10^{-6}).

More accurate (and costly) is the **two-sided** differential:

$$f'(x) \approx \frac{f(x + h) - f(x - h)}{2h}.$$

To compute a **gradient** $\nabla_f(\mathbf{x})$, one element is moved at a time.

Newton-Raphson optimization

Originally conceived for **finding roots**.

Approximates the function by the **tangent** line and finds the **intercept** (iterative procedure):

$$\frac{f(x_n) - 0}{x_n - x_{n+1}} = f'(x_n) \Rightarrow x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Extension to **optimization** is natural:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}.$$

In the **multivariate** case:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - [H_f(\mathbf{x}_n)]^{-1} \nabla_f(\mathbf{x}_n).$$

Integration

Numerical integration (quadrature) consists of a **weighted sum** of a finite set of **evaluations** of the integrand.

Integration weights depend on the **method** and on **precision**.

Deterministic methods: midpoint rule, trapezoidal rule, Simpson's rule, Gaussian,...

Alternative: **Monte Carlo integration**.