
Advanced Mathematical Methods

WS 2022/23

5 Parameter Estimation

Dr. Julie Schnaitmann

*Department of Statistics, Econometrics and Empirical
Economics*

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



WIRTSCHAFTS- UND
SOZIALWISSENSCHAFTLICHE
FAKULTÄT

Online References

Applied Econometrics Lecture (by Prof. Grammig, available on TIMMS)

- Lecture 37: Law of Large Numbers
- Lecture 38: Central Limit Theorem

Implications of a random sample

$\{X_1, X_2, \dots, X_n\}$ is a random sample if all draws of the random variable are **independently** and **identically distributed (iid)**.

Implications:

$$F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) \stackrel{\text{independent}}{=} F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot \dots \cdot F_{X_n}(x_n)$$

$$\stackrel{\text{identical}}{=} F_X(x_1) \cdot F_X(x_2) \cdot \dots \cdot F_X(x_n)$$

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) \stackrel{\text{independent}}{=} f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot \dots \cdot f_{X_n}(x_n)$$

$$\stackrel{\text{identical}}{=} f_X(x_1) \cdot f_X(x_2) \cdot \dots \cdot f_X(x_n)$$

Asymptotic ($n \rightarrow \infty$) results

For a random sample $\{X_1, X_2, \dots, X_n\}$ with finite $\mathbb{E}(X_i)$ and $Var(X_i)$ and an appropriately large n , following concepts apply:

- 1 Law of Large Numbers (LLN)
- 2 Central Limit Theorem (CLT)

Law of Large Numbers

Law of Large Numbers

$$\lim_{n \rightarrow \infty} P \left[\left| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}(X_i) \right| > \epsilon \right] = 0 \quad \text{for any } \epsilon > 0.$$

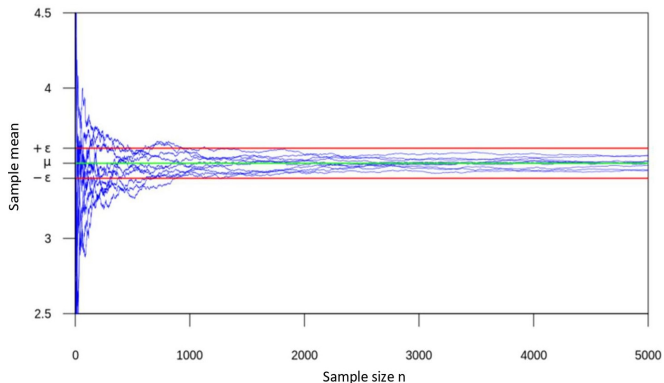
When the LLN holds, moments of the distribution of a population can be consistently estimated by moments of a random sample.

I.e., we can write:

- $\text{plim} \frac{1}{n} \sum_{i=1}^n X_i = \mathbb{E}(X)$
- $\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}(X)$

LLN and convergence in probability

Illustration of convergence in probability using LLN:



\Rightarrow A sequence $\{X_i\}$ converges in probability to a constant μ

Central Limit Theorem

Central Limit Theorem

$$\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}(X_i) \right] \stackrel{a}{\sim} \mathcal{N}(0, \text{Var}(X_i))$$

E.g. if $\{z_i\}$ is iid with $\mathbb{E}(z_i) = \mu$ and $\text{Var}(z_i) = \sigma^2$ and

$\bar{z}_n = \frac{1}{n} \sum_{i=1}^n z_i \xrightarrow{p} \mu$ then,

$$\sqrt{n}(\bar{z}_n - \mu) \xrightarrow{d} y \sim \mathcal{N}(0, \sigma^2)$$

$$\text{or } \bar{z}_n - \mu \stackrel{a}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$

$$\text{or } \bar{z}_n \stackrel{a}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Parameter estimation

Underlying principle:

- 1 Assumption about the distribution of a random variable in the population: $\mathbb{E}(X) = \mu$, $\text{Var}(X) = \sigma^2 < \infty$
- 2 Draw of a random sample
- 3 Use estimation functions to estimate the unknown parameters of the distribution: μ, σ^2
- 4 Estimation functions (short: estimators) are measurable functions of random variables $\Rightarrow \hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$

Quality of estimators

Finite sample properties of estimators:

- Bias
- Variance
- Efficiency (MSE)

Asymptotic concepts ($n \rightarrow \infty$):

- Consistency
- Asymptotic normality

Quality of estimators

Bias of $\hat{\theta}_n$:

- If $\mathbb{E}(\hat{\theta}_n) = \theta \Rightarrow$ unbiased estimator
- If $\mathbb{E}(\hat{\theta}_n) \neq \theta \Rightarrow$ biased estimator

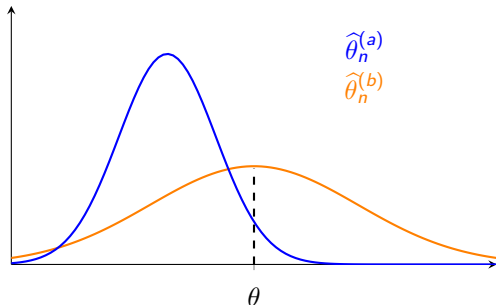
Variance of $\hat{\theta}_n$:

- Preferable: small variance of the estimation function \Rightarrow small $\text{Var}[\hat{\theta}_n(X_1, X_2, \dots, X_n)]$

Quality of estimators

Mean Squared Error (MSE) of $\hat{\theta}_n$:

- Preferable: an unbiased estimator with a small variance
⇒ Efficiency
- $MSE(\hat{\theta}_n) \equiv \mathbb{E}[(\hat{\theta}_n - \theta_n)^2] = Var(\hat{\theta}_n) + Bias(\hat{\theta}_n)^2$
- Trade-off: variance vs. bias of an estimator



Quality of estimators

Consistency of $\hat{\theta}_n$:

- The bigger the sample size gets, the smaller the sampling error $|\hat{\theta}_n - \theta|$ should become
- Formally $\hat{\theta}_n$ is consistent if

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

- Short-hand notation: $\hat{\theta}_n \xrightarrow{P} \theta$ or $\text{plim}_{n \rightarrow \infty} \hat{\theta}_n = \theta$
- Thus, for an infinitely large sample the estimator must be equal to the population parameter

Quality of estimators

Convergence in distribution and asymptotic normality of $\hat{\theta}_n$

Convergence in distribution: $z_n \xrightarrow{d} z$

If the c.d.f. of z_n converges to the c.d.f. of z at each point of continuity, then z_n converges in distribution to z .

Asymptotic normality of $\hat{\theta}_n$:

The distribution of an estimator $\hat{\theta}_n$ converges to the distribution of another random variable Z , which is normally distributed.

Generally:

$$x_N \xrightarrow{d} Z \sim \mathcal{N}(0, \sigma^2)$$

Applied to the Central Limit Theorem:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Z \sim \mathcal{N}(0, \text{Var}(\hat{\theta}_n))$$

Estimation techniques

- 1 Method of Moments
- 2 Maximum Likelihood Method
- 3 Ordinary Least Squares (OLS)

Method of Moments

Idea:

- Use the known relationship between parameters and theoretical moments
- Replace theoretical by empirical moments
- Method of Moment estimators are consistent

Method of Moments

For example, Exponential distribution $f_X(x; \lambda) = \lambda e^{-\lambda x}$:

- theoretical moments:

$$\textcircled{1} \mathbb{E}(X) = \frac{1}{\lambda} \implies \lambda = \frac{1}{\mathbb{E}(X)}$$

$$\textcircled{2} \text{Var}(X) = \frac{1}{\lambda^2} \implies \lambda = \frac{1}{\sqrt{\text{Var}(X)}} = \frac{1}{\sqrt{\mathbb{E}(X^2) - \mathbb{E}(X)^2}}$$

- replace by empirical moments:

$$\textcircled{1} \hat{\lambda}_1 = \frac{1}{\frac{1}{n} \sum_{i=1}^n x_i}$$

$$\textcircled{2} \hat{\lambda}_2 = \frac{1}{\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2}}$$

Maximum Likelihood Method

Idea:

- Chose $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_K)'$ such that the likelihood function $\mathcal{L}(\hat{\theta})$ is maximized
- The likelihood function $\mathcal{L}(\tilde{\theta})$ is the joint density function of X_1, \dots, X_n with parameters $\tilde{\theta}$

$$\mathcal{L}(\tilde{\theta}) = f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \tilde{\theta})$$

Intuition:

- 1 We decide on the distribution of $X \sim (\theta)$ in the population
- 2 Draw a random sample of X
- 3 Keep the sample fixed and choose the parameter such that $\mathcal{L}(\tilde{\theta})$ is maximized \Rightarrow "Maximize the probability to observe the realization of our sample"

Maximum Likelihood Method

Recipe:

- 1 Set up the likelihood function and exploit the **iid** sample implications
- 2 Take the \ln of the likelihood function
- 3 Maximize the log-likelihood function w.r.t. the parameter $\tilde{\theta}$

ML - 1. Set up the likelihood function

Use the iid property of a random sample:

$$\mathcal{L}(\tilde{\theta}) = f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \tilde{\theta})$$

$$\stackrel{\text{independent}}{=} f_{X_1}(x_1; \tilde{\theta}) \cdot f_{X_2}(x_2; \tilde{\theta}) \cdot \dots \cdot f_{X_n}(x_n; \tilde{\theta})$$

$$\stackrel{\text{identical}}{=} f_X(x_1; \tilde{\theta}) \cdot f_X(x_2; \tilde{\theta}) \cdot \dots \cdot f_X(x_n; \tilde{\theta})$$

$$= \prod_{i=1}^n f_X(x_i; \tilde{\theta})$$

ML - 2. Log-Transformation

Logarithmic rules

① $\ln(a \cdot b) \Leftrightarrow \ln(a) + \ln(b)$

② $\ln\left(\frac{a}{b}\right) \Leftrightarrow \ln(a) - \ln(b)$

③ $\ln(a)^b \Leftrightarrow b \cdot \ln(a)$

④ $\ln(e^x) \Leftrightarrow e^{\ln(x)} \Leftrightarrow x$

$$\mathcal{L}(\tilde{\theta}) = \prod_{i=1}^n f_X(x_i; \tilde{\theta}) \quad | \ln$$

$$\ln \mathcal{L}(\tilde{\theta}) = \sum_{i=1}^n \ln f_X(x_i; \tilde{\theta})$$

ML - 3. Log-likelihood maximization

Maximize the log-likelihood function:

$$\hat{\theta}_{ML} = \underset{\tilde{\theta}}{\operatorname{argmax}} \ln \mathcal{L}(\tilde{\theta}) = \sum_{i=1}^n \ln f_X(x_i; \tilde{\theta})$$

Determine $\hat{\theta}_{ML}$ via the F.O.C. $\frac{\partial \ln \mathcal{L}(\tilde{\theta})}{\partial \tilde{\theta}} \stackrel{!}{=} 0$:

$$\sum_{i=1}^n \frac{\partial \ln f_X(x_i; \hat{\theta})}{\partial \tilde{\theta}_1} \stackrel{!}{=} 0$$

\vdots

$$\sum_{i=1}^n \frac{\partial \ln f_X(x_i; \hat{\theta})}{\partial \tilde{\theta}_K} \stackrel{!}{=} 0$$

Properties of the ML estimator

If the likelihood function is correctly specified, the ML estimator has following properties:

- Consistency
- Asymptotic efficiency (for large n this estimator has the smallest MSE)
- Asymptotic normality

Conditional Maximum Likelihood (CML)

Procedure:

- 1 Specification of the conditional distribution $Y|X = x$
- 2 Specification of conditional moments (functions of X)
- 3 Insert conditional moments into the likelihood function
- 4 Maximize the (log-) likelihood function w.r.t. the unknown parameter

The marginal, joint and conditional distribution

Relationship of the marginal, joint, and conditional distribution:

$$f_{X_1|X_2} = \frac{f_{X_1, X_2}}{f_{X_2}}$$

$$\Leftrightarrow f_{X_1, X_2} = f_{X_1|X_2} \cdot f_{X_2}$$

$$\Leftrightarrow f_{X_2} = \frac{f_{X_1, X_2}}{f_{X_1|X_2}}$$

The marginal, joint and conditional distribution

Example: Exploiting this relationship allows us to write the joint density f_{X_1, \dots, X_5} as product of four conditional and one marginal density:

$$f_{X_1, X_2} = f_{X_2|X_1} \cdot f_{X_1} \quad (1)$$

$$\underline{f_{X_1, X_2, X_3}} = \underline{f_{X_3|X_1, X_2}} \cdot \underline{f_{X_1, X_2}} \quad (2)$$

$$f_{X_1, X_2, X_3, X_4} = f_{X_4|X_1, X_2, X_3} \cdot \underline{f_{X_1, X_2, X_3}} \quad (3)$$

$$f_{X_1, X_2, X_3, X_4, X_5} = f_{X_5|X_1, X_2, X_3, X_4} \cdot f_{X_1, X_2, X_3, X_4} \quad (4)$$

Plugging in (1)-(4) f_{X_1, \dots, X_5} can be expressed as:

$$= f_{X_5|X_1, X_2, X_3, X_4} \cdot f_{X_4|X_1, X_2, X_3} \cdot f_{X_3|X_1, X_2} \cdot f_{X_2|X_1} \cdot f_{X_1}$$

Example CML: Probit Model

- A binary response model \rightarrow outcome y is either 0 or 1
- $Y|X = x \sim Be(p(x))$
- Thus, not a linear probability model \rightarrow probability for a certain event doesn't change linear in x

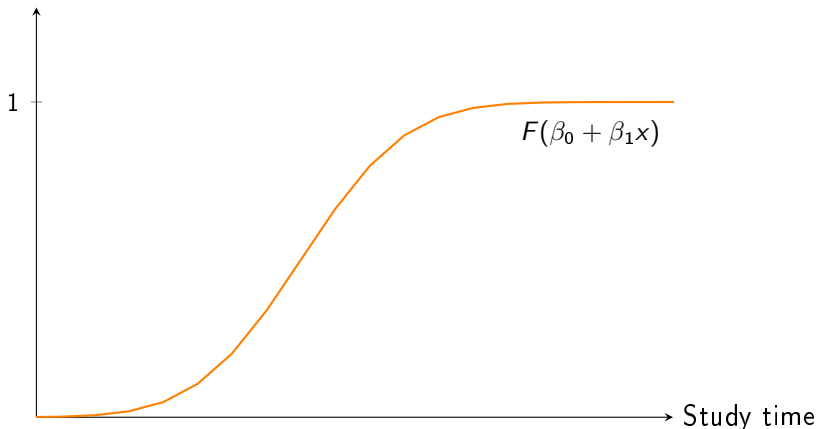
$$p(x) = E[Y|X = x] = F(\beta_0 + \beta_1 x) = \underbrace{\int_{-\infty}^{\beta_0 + \beta_1 x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz}_{\text{c.d.f. of a standard normal distribution}}$$

Example CML: Probit Model

- Pass or fail in an exam given the amount of study time
- X : Study time
 $y = 1$: Pass
 $y = 0$: Fail
- Probability to pass the exam doesn't change linearly with increasing study time
- $F(\beta_0 + \beta_1 x) = P(Y = 1|X) = p(x) \rightarrow$ probability Pass
 $1 - F(\beta_0 + \beta_1 x) = P(Y = 0|X) = 1 - p(x) \rightarrow$ probability Fail
 \Rightarrow Bernoulli distribution

Example CML: Probit Model

Probability to pass the exam as a function of study time:



Example CML: Probit Model

Estimating the parameters β_0 and β_1 :

- 1 Set up conditional likelihood function using a random sample

$$\mathcal{L}(\tilde{\beta}_0, \tilde{\beta}_1) = \prod_{i=1}^n \underbrace{p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}}_{\text{Bernoulli prob func}}$$

- 2 Maximize the log-likelihood w.r.t. β_0 and β_1

$$\frac{\partial \ln \mathcal{L}}{\partial \tilde{\beta}_0} \stackrel{!}{=} 0 \quad \frac{\partial \ln \mathcal{L}}{\partial \tilde{\beta}_1} \stackrel{!}{=} 0$$

- 3 Solve the F.O.C. for $\hat{\beta}_0$ and $\hat{\beta}_1$

Ordinary Least Squares (OLS)

Basic idea:

- Explain a variable Y (dependent variable) as linear combination of other variables X_1, X_2, \dots, X_K (explanatory variables/regressors) and the residual

$$Y = \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_2 + \dots + \tilde{\beta}_K X_K + \tilde{u}$$

- While Y and X are directly observable, \tilde{u} is not $\Rightarrow \tilde{u}$ depends on how we choose $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_K$

$$\tilde{u}_i = \tilde{u}_i(\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_K) = y_i - \tilde{\beta}_1 X_{i1} - \tilde{\beta}_2 X_{i2} - \dots - \tilde{\beta}_K X_{iK}$$

- However, it is clearly defined how $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_K$ must be chosen

OLS - Choosing $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_K$

Choosing $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_K$ either by:

- 1 Minimization of the least squares function:

$$\hat{\beta} = \underset{\tilde{\beta}}{\operatorname{argmin}} = \sum_{i=1}^n (y_i - \tilde{\beta}'x_i)^2$$

- 2 Moment restrictions: $\hat{u} = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_K x_{iK}$

$$\frac{1}{n} \sum_{i=1}^n \hat{u}_i x_{i1} \stackrel{!}{=} 0, \quad \frac{1}{n} \sum_{i=1}^n \hat{u}_i x_{i2} \stackrel{!}{=} 0, \quad \dots, \quad \frac{1}{n} \sum_{i=1}^n \hat{u}_i x_{iK} \stackrel{!}{=} 0$$

⇒ Both approaches lead to the same result

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i' y_i \right) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$$

OLS - classical assumptions

1 Linearity

$$Y = \beta_1 + \beta_2 X_2 + \dots + \beta_K X_K$$

2 Strict exogeneity

$$\mathbb{E}(u|X_1, X_2, \dots, X_K) = 0$$

3 Conditional homoscedasticity

$$\text{Var}(u|X_1, X_2, \dots, X_K) = \sigma^2$$

4 Distribution assumption

$$u|X_1, X_2, \dots, X_K \sim \mathcal{N}(0, \sigma^2)$$

OLS Example

We want to estimate the effect of more education on wage
(β_2 : return to schooling)

$$wage = \beta_1 + \beta_2 education + \beta_3 experience + \beta_4 experience^2 + u$$

Estimate $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$

\Rightarrow Under the above mentioned assumptions $\hat{\beta}_2$ gives the estimated marginal effect (ceteris paribus) of schooling on wage