Topics in Applied Econometrics for Public Policy

Master in Economics of Public Policy, BSE

Handout 0: Overview of the Course & Review of Basic Concepts

Laura Mayoral

IAE and BSE

Barcelona, Spring 2025

Welcome!

This course

- Two sections

- Section 1. Advanced topics in Econometrics: Non Parametric Statistics and Quantile Regression

- Section 2. Methods for Panel data (Lidia Farré)

Today's Goal

1. Overview of the contents of the course

2. Description of the overall logistics of the course.

3. Review of basic concepts (mostly for individual reading).

1. Overview of the course

But, first: A quick summary of your econometrics sequence

- So far, in your econometrics sequence:

- Introduction to Econometric Analysis:

Linear models for the conditional mean. Estimation methods: OLS and IV

- Advanced Estimation methods:

Alternative estimation methods (ML, GMM, etc).

- Two common characteristics in previous courses:

- Interest is typically placed on estimating the conditional mean: $E(y|X)$

- It's typically assumed a particular DGP (=data generating process)/model for the data. This is called Parametric Estimation.

Why the conditional mean, why parametric estimation?

- Why so much interest is placed on estimating the conditional mean $E(y|X)$?

- Response: Conditional expectation $E(y|X)$ is the optimal* predictor of Y given X .

- Optimal*: best under specified conditions (=a particular loss function)

- Conditional expectation as the optimal predictor:

- Consider the problem: Given data on y and X, what is the best forecast of y=g(X)? (i.e., what's the best way of combining information on X to produce the best predictor for y)

- first, what is best?

  - best=" lowest mean squared error (MSE)"

  - MSE=mean squared error: average of the squared prediction errors e, where $e = \widehat{y} - y$

- Conditional expectation as the optimal predictor:

- Consider the problem: Given data on y and X, what is the best forecast of y=g(X)? (i.e., what's the best way of combining information on X to produce the best predictor for y)

- first, what is best?

  - best=" lowest mean squared error (MSE)"

  - MSE=mean squared error: average of the squared prediction errors e, where $e = \widehat{y} - y$

- solution: $g(X) = E(y|X)$

- About $E(y|X)$:

- In general, it's a non-linear function of X.

- But one "magical" case: (y, X) jointly normal.

- About $E(y|X)$:

- In general, it's a non-linear function of X.

- But one "magical" case: (y, X) jointly normal.

  - In this case: $E(y|X) = X\beta$

  - Then we estimate $y = X\beta + \epsilon$ under the key assumption that $E(\epsilon|X) = 0$

- This simple derivation gives us

1) a "good" function of X to predict y: conditional expectation (birth of our interest in this function)

2) the shape of this function (if normality): linear function (birth of parametric modelling)

Summarizing

- These are the first bricks in the wall of regression analysis:

- The best predictor of y given X is the conditional expectation

  - But "best" depends on the loss function employed. Other loss functions will give other solutions different from the conditional expectation.

- If normality holds: the best prediction of y given X is just a linear function of X.

  - But normality is a strong assumption in many scenarios. Without normality, linearity has to be interpreted as an approximation to the true (nonlinear) conditional expectation, clearly sometimes this can be a stretch.

This course, I

- We will depart from this framework in two directions

Direction I: Interest in estimation methods that are generally valid under mild assumptions:

- Imposing linearity and/or a specific distribution on the data are strong assumptions

This course, I

- We will depart from this framework in two directions

Direction I: Interest in estimation methods that are generally valid under mild assumptions:

- Imposing linearity and/or a specific distribution on the data are strong assumptions

- Tradeoff between efficiency and validity:
  - Imposing assumptions that are correct leads to more efficient estimators
  - Imposing assumptions that are not true leads to inconsistent estimators

Non parametric estimation

- Departure point: in the vast majority of cases we don't know the "true" model or the "true" distribution of the data.

- Approach: We will look at methods that are valid under mild assumptions (we will impose mild restrictions on the DGP)

$\rightarrow$ Non-parametric (or semi-parametric) estimators.

Direction II: Interest in other aspects of the distribution of the data

- We will estimate other "quantities" different from the conditional expectation
  - for instance, conditional median (rather than conditional mean)
  - more generally: conditional quantiles
- Why?

- Why? At least three reasons

1. In some situations we can't estimate the conditional mean. For instance, if data are censored. However, we can estimate the conditional median.

- An example: you're studying the effectiveness of a new drug on extending the survival time of patients. 30% of the patients are still alive when the study ends (right censoring). Average survival time cannot be computed (you would need data on the whole distribution), but median survival could.

- Why? At least three reasons

1. In some situations we can't estimate the conditional mean. For instance, if data are censored. However, we can estimate the conditional median.

- An example: you're studying the effectiveness of a new drug on extending the survival time of patients. 30% of the patients are still alive when the study ends (right censoring). Average survival time cannot be computed (you would need data on the whole distribution), but median survival could.

2. The conditional mean ceases to be the "best" predictor if one changes the loss function.

- An example: if instead of a quadratic loss function, an absolute value loss function is employed, the conditional median becomes the optimal predictor.

3. The type of problem you're interested in is key!

- There are problems in economics that require to look at other moments of the data: poverty, inequality, etc.

- Examples:

- Income Inequality and Economic Mobility: Quantile regression can be used to explore how different factors influence the income distribution at various levels. For example, it can assess how education, demographic characteristics, or geographic location impact not just the average income (mean) but the lower and upper tails of the income distribution.

- Economic Policy Evaluation: Quantile regression is particularly useful in evaluating the effectiveness of policy interventions across different sectors of the economy. For example, it can help assess whether tax breaks or subsidies have different impacts on low-income versus high-income households, thereby informing more equitable policy designs.

- Housing Economics: When studying factors affecting housing prices, quantile regression can help policymakers understand how different variables, like location, size, or proximity to amenities, affect various segments of the housing market. This method can pinpoint if certain factors are driving up prices particularly in the upper quartiles (luxury housing market) versus the lower quartiles (affordable housing).

- Impact of Education on Earnings: Traditional regression methods focus on the average effect of education on earnings. However, quantile regression can show how this effect varies across different income levels. For instance, it might reveal that obtaining a college degree has a higher impact on earnings at the 90th percentile (high earners) compared to the 10th percentile (low earners), suggesting different returns to education across the income distribution.

2. About logistics

- The topics outlined above will keep us busy for the first 20 hours of this term.

- 5 hours a week: 4h with me and 1 h with the RA

- Materials will be posted in Classroom

- Website of the course:

http://mayoral.iae-csic.org/econometrics2025b/econometrics_2025.htm

- Check the syllabus for information about grading, references, etc.

- Please check it regularly for updates.

3. Review of Basic Concepts

The remaining of this document reviews three basic points that will be used during this course.

- This is very elementary material that by now you most likely already master (otherwise, make sure you do by the end of this week!).

- A. Some basic probability concepts

- B. Converge of random variables

- C. Estimators and basic properties.

All of this is already known, but please read the notes carefully to refresh these concepts, if needed.

A. Some probability background

- See: Greene (Appendix);

Definition of Probability

- Consider an experiment that has various possible outcomes.

Each possible outcome is represented as a point in a set. Each of these points are elementary events.

- Other events can be formed by combining elementary events.

- Sample space, Ω: the set that contains all elementary events.

Definition of probability

- Probabilities will be assigned to the elementary events according to certain axioms.

- Let Ω be the sample space, A be an event and P(.) is a probability assignment. The three axioms that define a probability are:

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- If A_1, A_2, ... are disjoint events, then $P(\cup_j A_j) = \sum_j P(A_j)$

Example

- Consider the random experiment of throwing a dice.
- $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Each of these elements are the elementary events.
- Other events can be defined by combining elementary events. A=$\{1,2\}$;
- The probability of each of the elementary events is 1/6.
- $P(1 \cup 2 \cup 3 \cup 4 \cup 5 \cup 6) = 1$; $P(A) = 2/6$;

Random variables

- Definition: A random variable is a function from Ω to the real numbers such that every element of Ω gets one real value.

- Example

- You are interested in the color of the eyes in a population. The set of possible colors are $\{black, brown, green, blue\}$. A random variable is the function that maps this set of events to numbers.

- Let X be your random variable: "color of the eyes".

- $X = \{1, 2, 3, 4\}$. This means $X = 1$ if eyes are black, $= 2$ if brown, etc.

- The values of a random variable can be arbitrary (we could define as well: $X = \{10, 20, 30, 40\}$.)

Random variables and their realizations

- A random variable is a function;

- it represents all the possible outcomes of your random experiment.

- We can associate probabilities to each of these outcomes.

- Realisation of a random variable: Once the random experiment has taken place, we observe its realisation. This is not random anymore.

- We usually use capital letters to denote random variables (r.v.) and small letters to denote particular realisations of these variables.

X=eye color; $x = blue$.

Example

- You are about to toss a coin. $\Omega = \{heads, tails\}$; $X = \{0, 1\}$
- Each of the values of X has an associated probability (if the coin is balanced, 0.5)
- Now you toss the coin, you get heads (X=0): this is a realisation of X.
- This realisation is not random anymore.

Types of random variables

Discrete random variables

- X is discrete if the number of distinct possible outcomes is either finite or countably infinite.

- For instance $X =$ outcome after tossing a coin; $Y =$ number of times that one should toss a coin until the first tails appears.

The assignment of probabilities in this case is done via a function, $f(x) = P(X = x)$, called the *probability mass function* (pmf), that has the properties:

- $f(x) \geq 0$
- $\sum_i f(x_i) = 1$

Continuous random variables

- X takes values in an interval.
- Examples: X: height of this class, unemployment rate, inflation rate, etc.
- The assignment of probabilities is done via the *probability density function* (pdf), $f(x)$, that has the properties:
- $f(x) \geq 0$
- $\int_i f(x_i) = 1$

- If X is continuous, then $P(X = x) = 0$ for all x.

- $\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\, dx$.

Distribution Function

- The cumulative distribution function is defined as:

$$F(x) = P(X \leq x).$$

If X is discrete, then

$$F(x_k) = \sum_{i=x_1}^{x_k} P(X = x_i),$$

where $x_1 \leq ... \leq x_k$.

If X is continuous,

$$F(x) = \int_{-\infty}^{x} f(x)\, dx.$$

Moments of a univariate distribution

- The shape of a probability distribution can be described with the help of its moments

There are two types of moments:

$$\begin{aligned} \mu_r &= E\left(X^r\right), \text{ is the rth raw moment} \\ \mu_r^* &= E\left(X-\mu_X\right)^r \text{ is the } rth \text{ central moment} \end{aligned}$$

Each of the moments provides some information about the distribution of X. For instance μ_1 is the mean, μ_2^* is the variance.

- Exercise

- Find out what are the names of μ_3^* and μ_4^* and what aspects of the distribution of X describe.

Some important moments.

Expectations

- The expected value of a random variable X, denoted as μ, is the first uncentered moment.

- It provides an idea of the central values of the distribution of X.

- Calculation.

$$E(X) = \mu_X = \begin{cases} \sum_i x_i P(X = x_i), \ if \ X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x)\, dx, \text{ if } X \text{ is continuous} \end{cases}.$$

Expected value of a function of X

- In situations, we are interested in obtaining the mean of a function of X. Let $Z = g(X)$, be a function of X, then

$$E(Z) = \mu_Z = \begin{cases} \sum_i g(x_i) P(X = x_i), \ if \ X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f(x)\, dx, \text{ if } X \text{ is continuous} \end{cases}.$$

- Why is this?
- Particular case: Expectation of a Linear transformation
- If $Z = a + bX$, then

$$E(Z) = a + bE(X). \tag{1}$$

- Because of the expression above, E(.) is called a linear operator

Variance.

- The variance of a distribution is a measure of dispersion with respect to the expected value.

$$Var\left(X\right) = \sigma_X = E\left(X - E\left(X\right)\right)^2.$$

- The variance is always larger than (or equal) to zero.
- If $Var(X) = 0$, then X is a number (has no variation).

Standard deviation

- Notice that the mean and the variance are not measured in the same units.
- Standard deviation: square root of the variance.

$$\sigma = \left(E\left(X^2\right) - E\left(X\right)^2 \right)^{1/2}$$

- The standard deviation is measured in the same units as the expected value.

Variance of a linear transformation of X

- The variance is not a linear operator.

- The variance of a linear function of X. is given by:

Let $Z = a + bX$, then

$$Var\left(Z\right) = b^2 Var\left(X\right).$$

Joint distributions

- Assume you have 2 different random experiments. It is possible to assign probabilities to the outcomes of two experiments at the same time.

- Example: for a given population, we define two variables X=age; $Y = height$. What is the probability that a person chosen at random from this population is older than 20 and taller than 1:70m? i.e., $P(X > 20, Y > 1.7)$?

- The joint distribution of X and Y will allow us to compute the probability above.

- We can also define the joint distribution of any group of variables. Let $X = (X_1, ..., X_n)'$ denote a group of random variables.

Joint distributions.

- The joint distribution completely characterizes the vector of random variables X.

- If X is discrete, then $f(x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n)$ is the joint probability mass function. It has to verify similar conditions as the univariate pmf.

- If X is continuous, then we can assign probability through $f(x_1, ..., x_n)$, the joint probability density function. It has to satisfy the conditions $f(x_1, ..., x_n) \geq 0$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dydx = 1.$$

Covariance and correlation

- The covariance between a pair of r.v. measures the degree of linear association between them. It is defined as

$$Cov(X,Y) = E\left((X - E(X))(Y - E(Y))\right)$$

- Interpretation of this measure: we can only interpret the sign of the covariance.

$$\begin{aligned} Cov(X,Y) &> 0: \text{ there is a positive linear relation btw } X, Y \\ Cov(X,Y) &= 0: \text{ there is not a linear relation btw } X, Y \\ Cov(X,Y) &< 0: \text{ there is a negative relationship btw } X, Y. \end{aligned}$$

If $Cov(X,Y) = 0$ it is said that X and Y and uncorrelated.

Correlation

- The covariance depends on the units of measurement of X and Y and therefore the magnitude of the covariance is NOT informative about the strength of the linear association between X and Y.

- The correlation is a standardized version of the covariance. It is bounded between [-1,1] and therefore not only the sign but also the strength of the relationship can be assessed with it.

$$Corr(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}.$$

- Interpretation:

$$
\begin{aligned}
Corr(X,Y) &= 1: \text{ the relation btw X, Y is positive} \\
&\quad \text{and perfectly linear} \\
1 &< Corr(X,Y) < 0: \text{ there a positive linear relation,} \\
&\quad \text{that is higher the higher corr is to 1} \\
Cov(X,Y) &= 0: \ X, Y \text{ are uncorrelated} \\
0 &< Corr(X,Y) < -1: \text{ there a negative linear relation,} \\
&\quad \text{that is higher the higher corr is to -1} \\
Corr(X,Y) &= -1: \text{ the relation btw X, Y is negative} \\
&\quad \text{and perfectly linear}
\end{aligned}
$$

Covariance and correlation of linear transformations of X and Y

- If $Z = a + bX$, $V = c + dY$, then,

$$Cov\left(Z, V\right) = bdCov\left(X, Y\right).$$

- If $Z = a + bX$, $V = c + dY$, then

$$Corr\left(Z, V\right) = Corr\left(X, Y\right).$$

More properties of expectations, variances and correlations

- The following relationships are very important, you should remember them!

- The expected value of a sum is the sum of expectations

$$E(\alpha_1 X_1 + \alpha_2 X_2 +, \ldots, \alpha_n X_n + c) = \alpha_1 E(X_1) + \alpha_2 E(X_2) +, \ldots, \alpha_n E(X_n) + c$$

- The variance of a sum is the sum of the variances if ONLY if the variables are uncorrelated. General case:

$$Var(\alpha_1 X_1 + \alpha_2 X_2 + c) = \alpha_1^2 Var(X_1) + \alpha_2^2 Var(X_2) + 2\alpha_1 \alpha 2 Cov(X_1, X_2)$$

- Covariance of a sum:

$$Cov(\alpha_1 X_1 + \alpha_2 X_2 + c, \alpha_3 X_3 + d) = \alpha_1 \alpha_3 Cov(X_1, X_3) + \alpha_2 \alpha_3 Cov(X_2, X_3)$$

- Combining the last two expressions, you can find out more expressions, for instance:

- Variance of n variables

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} Cov(X_i, X_j)$$

Marginal distributions

- Consider a bivariate distribution, (X, Y) with probability function $f(x, y)$.
- From $f(x, y)$, it is possible to recover the distributions of X and Y alone, i.e., distributions that do not depend on the other variable.
- These distributions are called marginal distributions.
- If (X, Y) are discrete: $f(x) = \sum_y f(x, y)$; $f(y) = \sum_x f(x, y)$;
- If (X, Y) are continuous: $f(x) = \int_y f(x, y)\, dy$; $f(y) = \int_x f((x, y)\, dx$;

Conditional distributions

- Conditional distributions play a crucial role in econometrics.

- Assume that the variables (X, Y) are related and we have some information about the variable X. Assume further that we have observed that $X = x$. We would like to update the probability of Y given the information available of X, that is, $X = x$.

- Example: Suppose that we are studying Y=height, X=weight of a population and that we have observed that the weight of a person chosen at random is 55kg. Clearly, the probability of having a particular height, say 1.90, given that we know that the person's weight is 55kg, would be different that the unconditional probability of being 1.9.

Conditional distributions

- The distribution of Y conditional to $X = x$ is defined as:

$$f(y|X = x) = \frac{f(x, y)}{f(x)}.$$

- The conditional distribution $f(y|X = x)$ is a probability function and therefore, has to verify the same conditions as any p.d.f or p.m.f (that is, is positive and has to add up –integrate– to 1).

- $f(y|X = x)$ is a function of X. That is, as X takes different values, we would obtain different $f(y|X = x)$

Independence

- In this course we will be interested in how one variable responds to the changes of a related variable.
- However, it can be the case that a variable does not react to the changes of some other variables because they are not related.
- This lack of relationship is called independence.
- The variables (X, Y) are stochastically independent iff (if and only if)

$$f(x, y) = f(x) f(y).$$

- Exercise: show that the latter result is equivalent to:

$$f(x|y) = f(x) \text{ and } f(y|x) = f(y).$$

Independence vs uncorrelation

- X, Y independent $\longrightarrow$ X, Y uncorrelated
- X, Y uncorrelated $\longrightarrow$ X, Y not necessarily independent

- Independence implies the lack of any relationship between X and Y and is a very strong condition.

- It is a much stronger condition than lack of correlation.

- Lack of correlation only means lack of linear relationship between X and Y. It can be the case that $corr\,(X,Y)=0$ but that X and Y are not independent.

- There is an important exception: if $(X,Y)'$ follow a normal bivariate distribution and are uncorrelated, then they are also independent.

The Normal distribution.

Univariate Normal distributions

- Let X be a continuous r.v.

- It follows a $N\left(\mu, \sigma^2\right)$ distribution (a Normal distribution with mean=μ and variance σ^2 if its pdf is given by

$$f\left(x\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(x-\mu\right)^2/2\sigma^2}.$$

- Standard Normal Distribution: N(0,1)

- The Normal distribution is symmetric

- To compute probabilities from a normal distribution: we use the tables (corresponding to a standard normal distribution).

- This distribution is very important in econometrics/statistics.

- Many techniques assume normality

- Many tests to check normality.

- STATA hint: use qnorm to compare the quantiles of any variable with those of a normal distribution.

The Multivariate Normal distributions

- Let $(X, Y)'$ be a pair of random variables.

- They follow a bivariate normal distribution, denoted as,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}\right),$$

that is, a Normal distribution with vector of means μ and variance-covariance matrix Σ if for any real vector λ,

$$\lambda' \begin{pmatrix} X \\ Y \end{pmatrix}$$

is a univariate normal distribution.

- Normal distributions are very important in statistics and econometrics for two reasons:

- They are very common (Central Limit Theorem).

- They have very good properties and then it is very convenient to work under the normality assumption.

- In particular one of this good properties is if X and Y are multivariate normal

$$Y|X \sim N(E(Y|X), var(Y|X))$$

and $E(Y|X) = a + bX$.

See Goldberger Chapter 7 for a description of the properties of these distributions.

B. Convergence of Random variables: A quick review

Key elements of asymptotic theory:

- The meaning of convergence of random variables.
- The most important convergence results
- The Law of Large Numbers
- The Central Limit Theorem.

Convergence of random sequences

- Consider a sequence of random variables: $X_1, X_2, \ldots, X_n$

- A sequence of r.v. converges to a limit if for large values of n the sequence and the limit are "close".

- But what does "close" mean when considering random variables?

- Defining 'closeness' in random variables is a bit more complicated than in the deterministic case.

- There are several ways to define "closeness". We will look at two: convergence in probability and convergence in distribution.

Convergence in probability

- Consider a sequence of random variables $X_1, \ldots, X_n$ or $\{X_i\}_{i=1}^n$.
- X_n converges in probability to X, written $X_n \xrightarrow{p} X$, if for every $\varepsilon > 0$

$$P(|X_n - X| > \varepsilon) \to 0 \text{ as } n \to \infty$$

- Convergence in probability looks at the values of the variables: the probability that the distance between X_n and its limit is "large" tends to zero.

Convergence in distribution

- Consider a sequence of random variables $X_1, \ldots, X_n$ or $\{X_i\}_{i=1}^n$.

- X_n converges in distribution to X, written $X_n \xrightarrow{d} X$, or $X_n \Rightarrow X$, if for all $x \in C$, where C is the set of continuity points of the distribution function $\mathrm{F}_X(.)$ of $X, then$

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

Convergence in distribution

- Consider a sequence of random variables $X_1, \ldots, X_n$ or $\{X_i\}_{i=1}^n$.

- X_n converges in distribution to X, written $X_n \xrightarrow{d} X$, or $X_n \Rightarrow X$, if for all $x \in C$, where C is the set of continuity points of the distribution function $\mathrm{F}_X(.)$ of $X, then$

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

- Convergence in distribution looks at the distribution of the variables, which should be very "close" (in the standard –deterministic– sense, as they are not random) as n gets large.

- Notes:

- The limit X can be a random variable or a constant.

- if X is a constant, we say that the limit has a degenerate distribution (as all the probability mass is concentrated in one point)

- The two modes of convergence are related. If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.

- The opposite result is not true in general (unless X is a constant).

Order in probability: $o_p(.)$ and Op(.)

Definition 1 *(Convergence in probability to zero)* X_n *converges in probability to zero, written* $X_n = o_p(1)$ *or* $X_n \xrightarrow{p} 0$, *if for every* $\varepsilon > 0$

$$P(|X_n| > \varepsilon) \to 0 \text{ as } n \to \infty$$

Definition 2 *(Boundedness in probability) The sequence* $\{X_n\}$ *is bounded in probability, denoted as* $X_n = O_p(1)$, *if for every* $\varepsilon > 0$ *there exists* $\delta(\varepsilon) \in (0, \infty)$ *such that*

$$P(|X_n| > \delta(\varepsilon)) < \varepsilon \text{ for all } n$$

Clearly if $X_t = o_p(1)$, then $X_t = O_p(1)$.

Some interesting results about probability orders

Definition 3 *(Algebra of probability orders)*

i) $X_n = o_p(a_n)$ *iff* $a_n^{-1} X_n = o_p(1)$,

ii) $X_n = O_p(a_n)$ *iff* $a_n^{-1} X_n = O_p(1)$.

■ Exercise: True of false. $\hat{\beta}_n$ is the OLS estimator of β under the usual hypothesis.

a) $\hat{\beta} = o_p(1)$

Some interesting results about probability orders

Definition 3 *(Algebra of probability orders)*

i) $X_n = o_p(a_n)$ *iff* $a_n^{-1} X_n = o_p(1)$,

ii) $X_n = O_p(a_n)$ *iff* $a_n^{-1} X_n = O_p(1)$.

■ Exercise: True of false. $\hat{\beta}_n$ is the OLS estimator of β under the usual hypothesis.

a) $\hat{\beta} = o_p(1)$

b) $\hat{\beta} - \beta = o_p(1)$

Some interesting results about probability orders

Definition 3 *(Algebra of probability orders)*

i) $X_n = o_p(a_n)$ *iff* $a_n^{-1} X_n = o_p(1)$,

ii) $X_n = O_p(a_n)$ *iff* $a_n^{-1} X_n = O_p(1)$.

■ Exercise: True of false. $\hat{\beta}_n$ is the OLS estimator of β under the usual hypothesis.

a) $\hat{\beta} = o_p(1)$

b) $\hat{\beta} - \beta = o_p(1)$

c) $\hat{\beta} - \beta = O_p(n^{.5})$

Some interesting results about probability orders

Definition 3 *(Algebra of probability orders)*

i) $X_n = o_p(a_n)$ *iff* $a_n^{-1} X_n = o_p(1)$,

ii) $X_n = O_p(a_n)$ *iff* $a_n^{-1} X_n = O_p(1)$.

■ Exercise: True of false. $\hat{\beta}_n$ is the OLS estimator of β under the usual hypothesis.

a) $\hat{\beta} = o_p(1)$

b) $\hat{\beta} - \beta = o_p(1)$

c) $\hat{\beta} - \beta = O_p(n^{.5})$

d) $\hat{\beta} - \beta = o_p(n^{-.25})$

Some interesting results about probability orders

Definition 3 *(Algebra of probability orders)*

i) $X_n = o_p(a_n)$ *iff* $a_n^{-1} X_n = o_p(1)$,

ii) $X_n = O_p(a_n)$ *iff* $a_n^{-1} X_n = O_p(1)$.

■ Exercise: True of false. $\hat{\beta}_n$ is the OLS estimator of β under the usual hypothesis.

a) $\hat{\beta} = o_p(1)$

b) $\hat{\beta} - \beta = o_p(1)$

c) $\hat{\beta} - \beta = O_p(n^{.5})$

d) $\hat{\beta} - \beta = o_p(n^{-.25})$

e) $\hat{\beta} - \beta = o_p(n^{-.5})$

Some interesting results about probability orders

Definition 3 *(Algebra of probability orders)*

i) $X_n = o_p(a_n)$ *iff* $a_n^{-1} X_n = o_p(1)$,

ii) $X_n = O_p(a_n)$ *iff* $a_n^{-1} X_n = O_p(1)$.

■ Exercise: True of false. $\hat{\beta}_n$ is the OLS estimator of β under the usual hypothesis.

a) $\hat{\beta} = o_p(1)$

b) $\hat{\beta} - \beta = o_p(1)$

c) $\hat{\beta} - \beta = O_p(n^{.5})$

d) $\hat{\beta} - \beta = o_p(n^{-.25})$

e) $\hat{\beta} - \beta = o_p(n^{-.5})$

f) $\hat{\beta} - \beta = O_p(n^{-.5})$

Proposition 1 *If* X_n *and* Y_n, *n=1, 2, ... are random variables defined on the same probability space and* $a_n > 0$, $b_n > 0$, $n = 1, 2, ...$, *then*

i) $if\ X_n = o_p(a_n)$ *and* $Y_n = o_p(b_n)$, *then* $X_n Y_n = o_p(a_n b_n)$; $X_n + Y_n = o_p(\max(a_n, b_n))$; $|X_t|^r = o_p(a_n^r)$, *for* $r > 0$

ii) $if\ X_n = o_p(a_n)$ *and* $Y_n = O_p(b_n)$, *then* $X_n Y_n = o_p(a_n b_n)$

iii) the statement (i) is valid if $o_p(.)$ *is replaced everywhere by* $O_p(.)$

Proposition 2 *If* $\{X_n\}$ *and* $\{Y_n\}$ *are two sequences of random* $k-vectors$ *such that* $X_n - Y_n = o_p(1)$ *and* $X_n \xrightarrow{d} X$, *then* $Y_n \xrightarrow{d} X$.

Limit Theorems

"Winwood Reade is good upon the subject," said Holmes. "He remarks that, while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant.")

Sir Arthur Conan Doyle, "The sign of the four"

Limit Theorems

- The Law of Large Numbers and the Central Limit Theorem are the most important results for computing the limits of sequences of random variables.

- There are many versions of LLN and CLT that differ on the assumptions about the dependence of the variables.

- Since we are assuming random sampling (=our data is i.i.d), then we have enough with their simplest versions: LLN and CLT for i.i.d. random variables.

Law of Large Numbers for iid sequences

Let $\{X_i\}_{i=1}^n$ be an $i.i.d$ sequence of random variables with finite mean μ then

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$

Proof. A very simple proof of this result can be provided if we further assume that $\text{var}(X_i) = \sigma^2 < \infty$. Then, by Chebychev's inequality:

$$\begin{aligned} P\left(\left| n^{-1} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon) \right) &\leq var(n^{-1} \sum_{i=1}^{n} X_i)/\varepsilon^2 \\ &= n^{-2} \sum_{i=1}^{n} var(X_i)/\varepsilon^2 \\ &= \frac{n\sigma^2}{n^2\varepsilon^2} \to 0. \end{aligned}$$

What does this mean?

- The STATA file handout2_LLN.do computes a small Monte Carlo simulation that shows you that this theorem is actually true. Run it so you can start experimenting with random numbers!

- The file does the following:

1. Fix n=100. Generate n random numbers using a χ^2 distribution with one degree of freedom. Notice that $E(X_i) = 1$. Compute $\bar{X}_n$ and store this value.

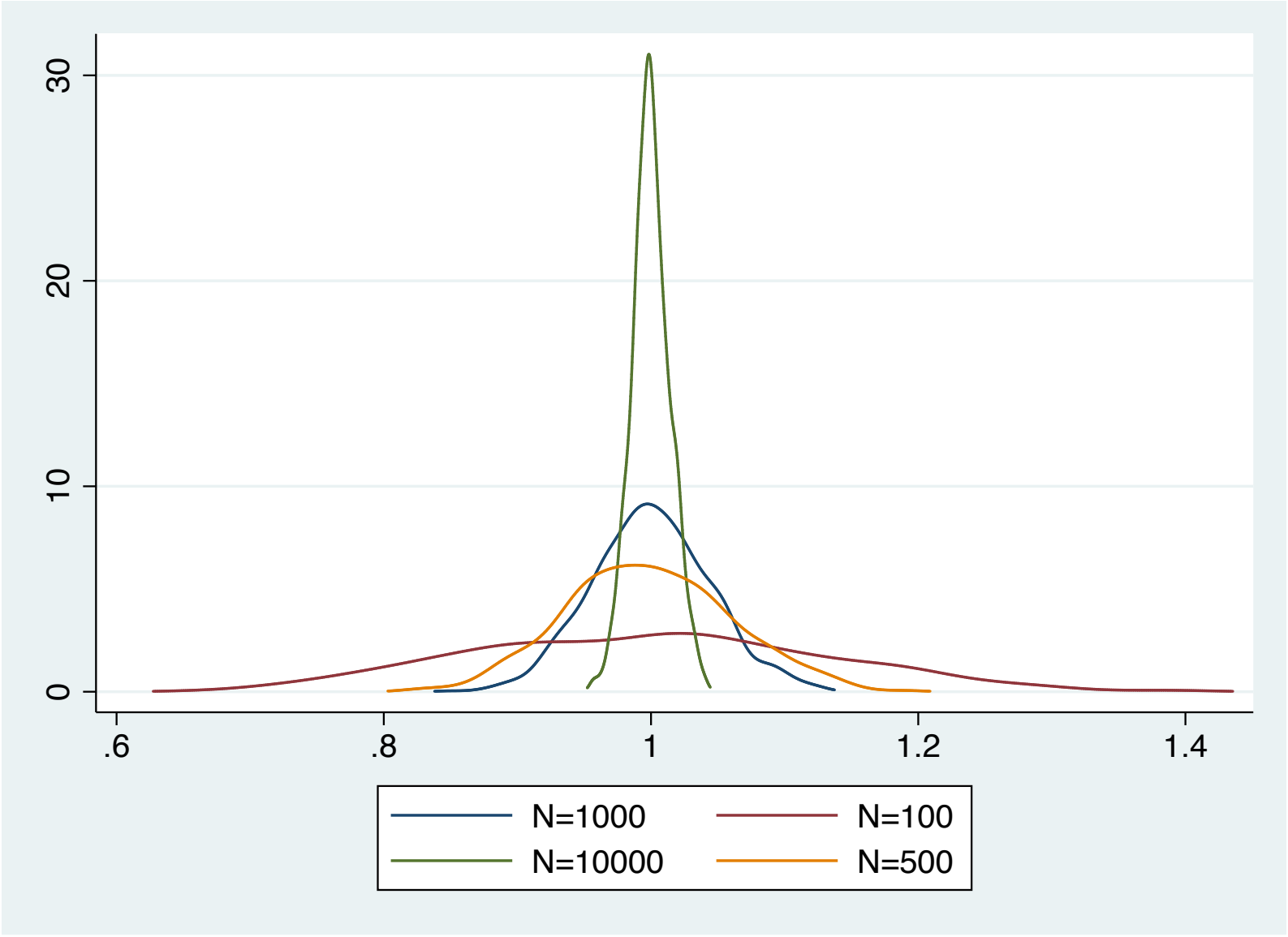
2. Repeat this R=1000 times. This allows us to see the distribution of $\bar{X}_{100}$

3. Repeat 1. and 2. for different values of $n = \{500, 1000, 10000\}$.

4. Plot the obtained distributions corresponding to $\bar{X}_{100}$, $\bar{X}_{500}$, $\bar{X}_{1000}$ and $\bar{X}_{10000}$.

The LLN in practice

This is what you get . . . what do you observe?



Central Limit theorem for $i.i.d.$ sequences

Let $\{X_i\}_{i=1}^n$ be a sequence of $i.i.d(\mu, \sigma^2)$ random variables.

Then

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} N(0,1)$$

The CLT in practice

Go to the following link to see an illustration of the CLT

https://demonstrations.wolfram.com/IllustratingTheCentralLimitTheoremWithSumsOfBernoulliRandomV/

Takeaway

- Asymptotic theory: tools to approximate the distribution of (functions of) random variables.
- Why? because in most cases we won't be able to determine the exact distribution of those variables.
- We would compute limits of a sequence of random variables X_n as gets approaches infinity
- Two modes of convergence: in probability and in distribution
- Two key results: CLT and LLN

3. Estimators and basic properties

3. Estimators and basic properties

Estimators and basic properties

- Remember that our goal is to be establish relationships among variables.

- Often, this relationship is captured by parameter(s) relating those variables.

- Example. Y=wages; X=years of education. Suppose that these variables are related linearly, then

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- β is our parameter of interest.

- If the parameter is identified, we can gather data to obtain an estimate of it and of its standard error (=a measure of the uncertainty of our estimator).

- An estimator is a function of the observable data that is used to estimate an unknown population parameter.

- An estimate is the result from the actual application of the function to a particular dataset,

- In general, many different estimators are possible for any given parameter.

- Estimators are random variables, thus we can (should) compute their associated distribution

- An estimate is a realization of the corresponding estimator. Thus, it is not random.

- Let n be the size of the sample used in the computation of an estimator of the parameter θ. Thus, for each sample size we can define an estimator: $\bar{\theta}_n$.

- Let $\hat{\theta}_n$ be an estimator of the population parameter θ computed with a sample of size n. Then, $\hat{\theta}_n$ is a function that maps each sample S to its sample estimate $\hat{\theta}_n(S)$. The sequence $\{\hat{\theta}_n\}$ is an example of a sequence of random variables, so the concepts introduced above are applicable to $\{\hat{\theta}_n\}$.

- Different tools to obtain the limit of a sequence of estimators: the LLN, the CLT; non-parametric estimation often requires additional tools.

Properties of estimators

Some desirable properties of $\hat{\theta}_n$ are the following.

- Consistency: $\hat{\theta}_n$ is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$.

- Related: The rate of convergence of $\hat{\theta}_n$ is n^b if

$$n^b(\hat{\theta}_n - \theta) = O_p(1)$$

- In parametric estimators the rate of convergence is $n^{1/2}$
- In non-parametric estimators it's typically smaller.

- Unbiasedness: $\hat{\theta}_n$ is unbiased if $\mathrm{E}\left(\hat{\theta}_n\right) = \theta$ and is asymptotically unbiased if $\lim_{n\to\infty} E\left(\hat{\theta}_n\right) = \theta$.

- Asymptotic Normality. A consistent estimator $\hat{\theta}_n$ is asymptotically normal around the true parameter θ if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V)$, where V is called the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$.

- Efficiency: An unbiased estimator $\hat{\theta}_n$ is efficient if it has the lowest possible variance among all unbiased estimators.

Takeaway

- We are interested in the value of unknown parameters
- We would use data and econometric techniques to figure out the values of those parameters
- Estimators (random variables) and estimates (the particular value that an estimator gets when a particular dataset is employed).
- Many possible estimators are available, How do we choose among them?
- We want estimators with good properties.
- consistency, asymptotic normality, efficiency, unbiasedness. . .