Advanced Econometrics

Lecture 1: Introduction

Instructor: Ma, Jun

Renmin University of China

September 18, 2021

What is econometrics?

- Econometrics is the unified study of economic models, mathematical statistics, and economic data.
- Econometrics is concerned with the development of statistical methods for:
  - Estimation of economic relationships/causal effects.
  - Testing of economic theories.
  - Forecasting of important economic variables.
  - Evaluation of government and business policy.
- Econometric theory concerns the development of tools and methods, and the study of the properties of econometric methods.
- Applied econometrics is a term describing the development of quantitative economic models and the application of econometric methods to these models using economic data.

Examples

- Estimation of demand and supply functions. Elasticity of demand/supply can be used to evaluate the effect of taxation.
- Testing the efficient market hypothesis (asset returns cannot be predicted from their own past).
- Mincer, J., *Schooling, Experience, and Earnings*, 1974. Estimation of return to schooling and experience using individual census data.
  - Used to determine the optimal amount of schooling.
  - Study education in developing countries.
  - Study gender and race discrimination.
  - Study the impact of immigration on labour markets.
- Paarsch, H. J., *Journal of Econometrics*, 1997. Estimation of optimal reserve price for BC timber auctions.
- Sun, A. and Zhao, Y., *Journal of Development Economics*, 2016. Divorce, abortion and the child sex ratio: The impact of divorce reform in China

The probabilistic approach to econometrics

- Economic models are only approximations .
- A model can take into account a number of important factors, but there will be many factors left out that also affect outcomes. Economic models should be explicitly designed to incorporate randomness.
- Once we acknowledge that an economic model is a probability model, it follows naturally that an appropriate tool way to quantify, estimate, and conduct inferences about the economy is through the powerful theory of mathematical statistics.
- The appropriate method for a quantitative economic analysis follows from the probabilistic construction of the economic model.

Econometric Terms and Notation

- In a typical application, an econometrician has a set of repeated measurements on a set of variables. E.g., in a labor application the variables could include weekly earnings, educational attainment, age, and other descriptive characteristics. We call this information the data, dataset, or sample.
- We use the term observations to refer to the distinct repeated measurements on the variables. An individual observation often corresponds to a specific economic unit, such as a person, household, corporation, firm, organization, country, state, city or other geographical region.

- Economists typically denote variables by the italicized roman characters y, x, and/or z. The convention in econometrics is to use the character y to denote the variable to be explained, while the characters x and z are used to denote the conditioning (explaining) variables.
- Following mathematical convention, real numbers (elements of the real line $\mathbb{R}$, also called scalars) are written using lower case italics such as y, and vectors (elements of $\mathbb{R}^k$) by lower case bold italics such as $\boldsymbol{x}$, e.g.

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}.$$

Upper case bold italics such as $\boldsymbol{X}$ are used for matrices.

- We denote the number of observations by the natural number n and subscript the variables by the index i to denote the individual observation, e.g. y_i, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$.

Definition
The i^{th} **observation** is the set $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$. The **sample** is the set $\{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i) : i = 1, \ldots, n\}$.

- In some contexts we use indices other than i, such as in time-series applications where the index t is common and T is used to denote the number of observations.
- In panel studies we typically use the double index it to refer to individual i at a time period t.

- We typically use Greek letters such as β, θ and σ^2 to denote unknown parameters of an econometric model, and will use boldface, e.g. $\boldsymbol{\beta}$ or $\boldsymbol{\theta}$, when these are vector-valued.
- Estimates are typically denoted by putting a hat$\widehat{}$, tilde$\widetilde{}$ or bar$\bar{}$ over the corresponding letter, e.g. $\widehat{\beta}$ and $\widetilde{\beta}$ are estimates of β.
- The covariance matrix of an econometric estimator will typically be written using the capital boldface $\boldsymbol{V}$, often with a subscript to denote the estimator, e.g. $\boldsymbol{V}_{\widehat{\beta}} = \text{var}\left(\widehat{\beta}\right)$ as the covariance matrix for $\widehat{\beta}$.
- Hopefully without causing confusion, we will use the notation $\boldsymbol{V}_{\beta} = \text{avar}\left(\widehat{\beta}\right)$ to denote the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ (the variance of the asymptotic distribution). $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ denotes an estimate of $\boldsymbol{V}_{\beta}$.

Observational data, experimental data and causality

- A common econometric question is to quantify the impact of one set of variables on another variable. E.g. a concern in labor economics is the returns to schooling — the change in earnings induced by increasing a worker's education, holding other variables constant.
- In order to say that one variable has a causal effect on another, other factors affecting the outcome must be held fixed (controlled for). If the outcome changes as the variable changes with other factors held constant, we say that the variable has a causal effect.
- Ideally, we would use experimental data to answer these questions. Natural sciences use controlled lab experiments. Experiment are often impossible in economics (too costly and/or for ethical reasons).

- Most economic data is observational. Econometrics encompasses a wide range of statistical tools that allow us to estimate causal effects using observational data, which is more challenging.
- The causal effect is individual-specific and unobserved. E.g., the causal effect of schooling on wages for an individual worker is the difference in wages he/she would receive if we could change his/her level of education holding all other factors constant. The counterfactual wage under a different level of education is unobserved.

Correlation is not causation

- While we are interested in causal relations, statistics allows us to establish correlations (associations) in the data.
- "Dog owners are much happier than cat owners" (reported in *Washington Post*, Apr. 5, 2019)
  - The correlation between reported happiness and dog ownership not hard to believe.
  - Is there a causal effect? In other words, letting everybody own a dog makes the whole population happier?
- Going from correlations to causation requires making untestable assumptions on the structural model that generates the data.

Structural models

- Suppose Y is an economic outcome variable of interest (e.g., wage rate of individual workers, academic achievement of individual students, rate of return of some asset...), X is a vector of observed explanatory variables.
- There are factors in a vector ϵ that affect the outcome and are unobserved to the researcher.
- The fact that (X, ϵ) determines Y can be formulated as a functional relationship $Y = g(X, \epsilon)$. The causal effect of some variable in X on Y is given by the partial derivative of g with respect to that variable.
- This structural model (the relation g and the distribution of (X, ϵ)) characterizes the data generating mechanism of Y. We observe a sample $\{Y_i, X_i\}_{i=1}^n$ from the model, i.e., for some unobserved ϵ_i, $Y_i = g(X_i, \epsilon_i)$.
- We wish to recover the structural relation g, but there is no hope if we do not put any restriction on the model.

- We often use economic theory to justify the assumptions: what variables are in (X, ϵ) and what is the form of g.
- Two approaches:
  - Structural approach: an economic model (an agent maximizing utility subject to constraints) provides a list of variables (X, ϵ), specifies how (X, ϵ) determines Y and the researcher chooses specific functional forms for the model's components (e.g., consumers' utility function or firms' cost function). This approach is usually more difficult to implement.
  - Non-structural (statistical) approach: the restriction on g originates from statistical concerns rather than an economic model and the list of variables (X, ϵ) comes from understanding of the decision process that determines Y and background knowledge. E.g., specify a linear model $g(X, \epsilon) = \alpha + \beta X + \epsilon$ with unknown (α, β) which can be estimated by least squares.

Standard data structures

- Cross-sectional data sets have one observation per individual. Surveys and administrative records are a typical source for cross-sectional data. In typical applications, the individuals surveyed are persons, households, firms or other economic agents. In many cases the sample size is quite large.
- Example: A cross-sectional data set on wages and other individual characteristics (Table 1.1, Page 7):

obs number	wage	education	experience	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
⋮	⋮	⋮	⋮	⋮	⋮

- The order of observations is not important.
- It is usually natural to assume that the observations are statistically independent.

- Time-series data are indexed by time. Typical examples include macroeconomic aggregates, prices and interest rates.
- A time series data set consists of observation on several variables over time.
- Example: Minimum wage, unemployment, and related data for Puerto Rico (Table 1.3, Page 9):

obs number	year	minimum wage	unemployment	gnp
1	1950	0.20	15.4	878.7
2	1951	0.21	16.0	925.0
3	1952	0.23	14.8	1015.9
⋮	⋮	⋮	⋮	⋮

- The frequency at which the data is collected can be daily, weekly, monthly, quarterly, and annually. In Finance, high frequency trade data.
- The order of observations is important.
- Observations are often correlated.

- Panel data combines elements of cross-section and time-series. These data sets consist of a set of n individuals (typically persons, households, or corporations) measured repeatedly over T periods. In some panel data contexts, $n \gg T$. In other panel data contexts (for example when countries are taken as the unit of measurement), $T \gg n$.
- Clustered samples are related to panel data. In clustered sampling, the observations are grouped into "clusters" which are treated as mutually independent, yet allowed to be dependent within the cluster.
- Spatial dependence is another model of interdependence. The observations are treated as mutually dependent according to a spatial measure (for example, geographic proximity). Spatial dependence can also be viewed as a generalization of time series dependence.

- Most of this text will be devoted to cross-sectional data under the assumption of mutually independent observations.
- By mutual independence we mean that the i^{th} observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ is independent of the j^{th} observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ for $i \neq j$. In this case we say that the data are independently distributed.
- It is a statement about the relationship between observations i and j, not a statement about the relationship between y_i and $\boldsymbol{x}_i$ and/or $\boldsymbol{z}_i$.
- It is reasonable to model each observation as a draw from the same probability distribution. In this case we say that the data are identically distributed. If the observations are mutually independent and identically distributed, we say that the observations are independent and identically distributed, iid, or a random sample.

Definition (1.5.1, Hansen)

The observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$are a **sample** from the distribution F if they are identically distributed across $i = 1, \ldots, n$ with joint distribution F.

Definition (1.5.2, Hansen)

The observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$are a **random sample** if they are mutually independent and identically distributed (***iid***) across $i = 1, \ldots, n$.

- In the random sampling framework, we think of an individual observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ as a realization from a joint probability distribution $F(y, \boldsymbol{x}, \boldsymbol{z})$ which we can call the population. This "population" is infinitely large.
- The goal of statistical inference is to learn about features of from the sample. The assumption of random sampling provides the mathematical foundation for treating economic statistics with the tools of mathematical statistics.

Common Symbols

y	scalar
$\boldsymbol{x}$	vector
$\boldsymbol{X}$	matrix
$\mathbb{R}^k$	Euclidean k space
$\mathbb{E}(y)$	mathematical expectation
$\text{var}(y)$	variance
$\text{cov}(x, y)$	covariance
$\text{var}(\boldsymbol{x})$	covariance matrix
$\text{corr}(x, y)$	correlation
$\Pr$	probability
$\longrightarrow$	limit
$\xrightarrow{p}\left(\xrightarrow{d}\right)$	convergence in probability (distribution)
$\text{plim}_{\text{n}\to\infty}$	probability limit

Common Symbols

$\mathrm{N}(\mu, \sigma^2)$	normal distribution with mean $\mu$and variance σ^2
χ^2_k	chi-square distribution with k degrees of freedom
$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\mathrm{tr}\boldsymbol{A}$	trace
$\boldsymbol{A}'$	matrix transpose
$\boldsymbol{A}^{-1}$	matrix inverse
$\boldsymbol{A} > \boldsymbol{0}$	positive definite
$\boldsymbol{A} \geq \boldsymbol{0}$	positive semi-definite
$\| \boldsymbol{a} \|$	Euclidean norm
$\| \boldsymbol{A} \|$	matrix (Frobinius or spectral) norm
$\approx$	approximate equality
$\sim$	is distributed as
log	natural logarithm