Advanced Econometrics

Lecture 0: Introduction (Hansen Chapter 1)

Instructor: Ma, Jun

Renmin University of China

Fall 2018

What is Econometrics?

- Econometrics is the unified study of economic models, mathematical statistics, and economic data.
- **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.

The Probability Approach to Econometrics

- Economic models should be explicitly designed to incorporate randomness; stochastic errors should not be simply added to deterministic models to make them random.
- 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.

Structural Approach

- **Structural approach**. A probabilistic economic model is specified, and the quantitative analysis performed under the assumption that the economic model is correctly specified.
- This should not be mixed up with "structural" equations covered in Hansen Chapter 11.
- The structural approach typically leads to likelihood-based analysis, including maximum likelihood and Bayesian estimation.

Quasi-Structural Approach

- A criticism of the structural approach is that it is misleading to treat an economic model as correctly specified.
- The **quasi-structural** approach to inference views a structural economic model as an approximation rather than the truth.
- This theory has led to the concepts of the pseudo-true value (the parameter value defined by the estimation problem), the quasi-likelihood function, quasi-MLE, and quasi-likelihood inference.

Non-Structural (Reduced-Form) Approach

- An economist uses intuition and background knowledge to specify an equation describing hypothesized relationships between a dependent variable of interest and some explanatory variables.
- Non-structural analyses have a disadvantage in that very often they provide limited insight into the economic activities of interest. The conclusions are very often qualitative.
- In structural analysis, an economist formulates an economic model to explain some observed behavior, chooses specific functional forms for the model's components (e.g., consumers' utility function or firms' cost function).
- Structural analysis is usually more difficult to implement. However, it enables studying counterfactual scenarios which are very often of much more interest.

Two Examples

- Non-Structural Econometric Analysis: D. Angrist and A. Krueger. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics.*
- Structural Econometric Analysis: Y. Luo, I. Perrigne and Q. Vuong. Structural analysis of nonlinear pricing. *Journal of Political Economy*.

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.

Econometric Terms and Notation

- 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.

Econometric Terms and Notation

- 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.

Econometric Terms and Notation

- 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 and Experimental Data

实验数据 & 观测数据
↓
可以控制变量

- 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.
- Ideally, we would use **experimental** data to answer these questions. To measure the returns to schooling, an experiment might randomly divide children into groups, mandate different levels of education to the different groups, and then follow the children's wage path after they mature and enter the labor force. The differences between the groups would be direct measurements of the effects of different levels of education.

Observational Data and Experimental Data

- Instead, most economic data is **observational**. To continue the above example, through data collection we can record the level of a person's education and their wage.
- But from observational data it is difficult to infer **causality**, as we are not able to manipulate one variable to see the direct effect on the other. For example, a person's level of education is (at least partially) determined by that person's choices. These factors are likely to be affected by their personal abilities and attitudes towards work. The fact that a person is highly educated suggests a high level of ability, which suggests a high relative wage.

Standard Data Structures

- There are five major types of economic data sets which are distinguished by the dependence structure across observations.
- **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. It is conventional to assume that cross-sectional observations are mutually independent.
- **Time-series** data are indexed by time. Typical examples include macroeconomic aggregates, prices and interest rates. This type of data is characterized by serial dependence. Most aggregate economic data is only available at a low frequency (annual, quarterly or perhaps monthly) so the sample size is typically much smaller. An exception is financial data where data are available at a high frequency (weekly, daily, hourly, or by transaction) so sample sizes can be quite large.

Standard Data Structures

- **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.

Standard Data Structures

- 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.
- Sometimes the label "independent" is misconstructed. 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 $\boldsymbol{z}_i$.
- Furthermore, if the data is randomly gathered, 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.

Standard Data Structures

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 $i = 1, \ldots, n$.

Standard Data Structures

- 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.
- The random sampling framework was a major intellectual breakthrough of the late 19th century, allowing the application of mathematical statistics to the social sciences. Before this conceptual development, methods from mathematical statistics had not been applied to economic data as the latter was viewed as non-random.

Sources for Economic Data

- Looking for Chinese datasets? Check the Datasets page provided by New York University Shanghai: https://datascience.shanghai.nyu.edu/datasets.
- Another good source of data is from authors of published empirical studies. Most journals in economics require authors of published papers to make their datasets generally available.
- If you are interested in using the data from a published paper, first check the journal's website and then check the website(s) of the paper's author(s). Most academic economists maintain webpages, and some make available replication files complete with data and programs.

Econometric Software

- **Stata** is a powerful statistical program continuously being updated with new methods with a user-friendly interface. It is limited when you want to use new or less-common econometric methods which have not yet been programed.
- **R** and **MATLAB** are high-level matrix programming languages with a wide variety of built-in statistical functions. Many econometric methods have been programed in these languages and are available on the web. **R** is open-source and free. **Octave** is compatible with **MATLAB** scripts and is free.
- For highly-intensive computational tasks, standard low-level programming language such as **Fortran** or **C** is advantageous and can lead to major gains in computational speed, at the cost of increased time in programming and debugging.
- **Julia** is the new-generation open-source high-performance but user-friendly programming language.

Common Symbols

y	scalar 标量
$\boldsymbol{x}$	vector 向量
$\boldsymbol{X}$	matrix 矩阵
$\mathbb{R}^k$	Euclidean k space 欧几里得空间
$\mathbb{E}(y)$	mathematical expectation 数学期望
$\text{var}(y)$	variance 方差
$\text{cov}(x, y)$	covariance 协方差
$\text{var}(\boldsymbol{x})$	covariance matrix 协方差矩阵
$\text{corr}(x, y)$	correlation 相关系数
$\Pr$	probability
$\longrightarrow$	limit
$\xrightarrow{p}\left(\xrightarrow{d}\right)$	convergence in probability (distribution) 依概率/分布收敛
$\text{plim}_{n\to\infty}$	probability limit 概率极限

Common Symbols

$\mathrm{N}(\mu, \sigma^2)$	normal distribution with mean $\mu$and variance σ^2 正态分布(μ,σ^2)
χ^2_k	chi-square distribution with k degrees of freedom
$\boldsymbol{I}_n$	$n \times n$ identity matrix $n\times n$单位矩阵
tr$\boldsymbol{A}$	trace 矩阵的迹
$\boldsymbol{A}'$	matrix transpose 转秩
$\boldsymbol{A}^{-1}$	matrix inverse 矩阵的逆
$\boldsymbol{A} > \boldsymbol{0}$	positive definite 正定
$\boldsymbol{A} \geq \boldsymbol{0}$	positive semi-definite 半正定
$\| \boldsymbol{a} \|$	Euclidean norm 向量的模
$\| \boldsymbol{A} \|$	matrix (Frobinius or spectral) norm 矩阵的模
$\approx$	approximate equality
$\sim$	is distributed as
log	natural logarithm 自然对数.即 ln.