# Introductory Econometrics

## Lecture 1: Introduction

Instructor: Ma, Jun

Renmin University of China

September 7, 2021

# What is Econometrics?

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.

# 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

# Why statistics?

- Economic theory is used to construct models characterizing relationships between variables of interest.
- However, 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.
- We therefore replace the exact (deterministic) model with a probabilistic model.

# Causality

- Natural sciences use controlled lab experiments. Experiment are often impossible in economics (too costly and/or for ethical reasons).
- Econometrics encompasses a wide range of statistical tools that allow us to estimate causal effects using observational data, which is more challenging.
- 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.
- 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 $\epsilon$ that affect the outcome and are unobserved to the researcher.
- The fact that $(X, \epsilon)$ 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 $(X, \epsilon)$) 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 $\epsilon_i$, $Y_i = g(X_i, \epsilon_i)$.
- 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 $(X, \epsilon)$ 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 $(X, \epsilon)$, specifies how $(X, \epsilon)$ 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 $(X, \epsilon)$ 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 $(\alpha, \beta)$ which can be estimated by least squares.

# Examples of linear models

- **Education:**

$$\log(\text{Wage}) = \alpha + \beta \times \text{Years of Schooling} + U,$$

$U$ = other factors, for example, ability. Since it is very hard to control for ability, one can overestimate the return to education by relying on usual correlations.

- **Size of the police force and crime:**

$$\text{Number of Crimes} = \alpha + \beta \times \text{Size of the Police Force} + U.$$

Usually, cities with a lot of criminal activity have a bigger police force. Simple correlations can spuriously indicate that the size of the police force has a positive effect on the crime rates. This is an example of simultaneous equations model.

# Types of data: cross-section

- A cross-sectional data set consists of observations on individuals such as workers or firms collected in a single period of time.

- 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.

# Types of data: time series

- 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; trends.

# Types of data: panel

- A panel data set consists of a time series for each cross-sectional member.
- Example: A two-year panel data set on city crime statistics (Table 1.5, Page 11):

| obs numb | city | year | murders | population | unempl | police |
|---|---|---|---|---|---|---|
| 1 | 1 | 1986 | 5 | 350000 | 8.7 | 440 |
| 2 | 1 | 1990 | 8 | 359200 | 7.2 | 471 |
| 3 | 2 | 1986 | 2 | 64300 | 5.4 | 75 |
| 4 | 2 | 1990 | 1 | 65100 | 5.5 | 75 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |