The Nature of Econometrics and Economic Data

Caio Vigo

The University of Kansas
Department of Economics

Summer 2019

These slides were based on *Introductory Econometrics* by Jeffrey M. Wooldridge (2015)

Topics

What is Econometrics?

• **Econometrics** is the set of tools by which economists, and others in the social sciences, analyze data. We can use econometrics to:

- Estimate economic relationships;
- Test economic theories;
- Evaluate government and business policy.

What is Econometrics?

Econometrics focuses on problems inherent in analyzing data generated by individuals, firms, and other entitities acting strategically, and interacting with one another.

- For example:
  - Does higher school spending is related with higher SAT/ACT scores?
  - What are the effects to the US economy if we raise our tariffs?
  - What is the effect of a 20-week training program on worker's hourly wage?
  - What are the effects on crime rates if we legalize marijuana?

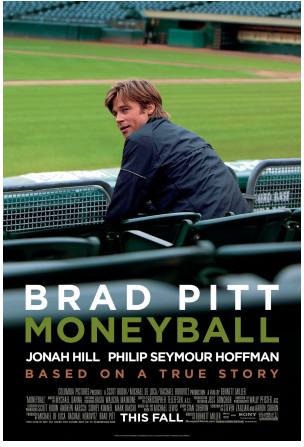
What is Econometrics?

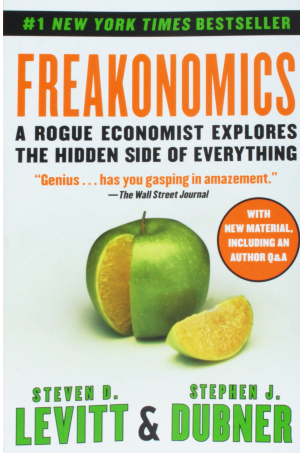
- A simple correlation analysis might not be sufficient because causality can be difficult to infer.

- In economics, theory and empirical analysis are both important.

- An **empirical analysis** uses data to test a theory, estimate an economic relationship, or determine the effects of a policy or intervention. Econometrics allows us to analyze data using formal statistical methods.

What is Econometrics?





Is Econometrics = Statistics?

• Econometrics is its own discipline (separate from statistics) mainly because there exists the following difference:

Experimental Data $\neq$ Nonexperimental Data

Is Econometrics = Statistics?

- **Experimental data:**
  - Data from controlled experiments;
  - Common in the natural sciences (physics, chemistry, ...) and in the biomedical fields;
  - However, harder to find or generate in the social sciences (although some exist).

Is Econometrics = Statistics?

- **Nonexperimental data**
  - These are data sets collected in a passive manner, after we observe outcomes on individuals, firms, schools, and so on.
  - We just "observe" the data without having any control over it,i.e., we simply act as "observers"of what has happened and then try to learn from what we observe.
  - Other names for nonexperimental data: **observational** or **retrospective** data.

Why Study Econometrics?

- Important to be able to apply economic theory to real world data.

- Theory may be ambiguous as to the effect of some policy change, and in any case theory rarely tells us how large the effect might be.

- Forecasting economic variables (inflation, interest rates, housing starts, and so on) is important, too.

Topics

1 What is Econometrics?
Why Study Econometrics?
2 Steps in an Empirical Analysis
3 The Structure of Economic Data
4 Causality and the Notion of Ceteris Paribus

Steps for a Successful Empirical Study

Steps for a Successful Empirical Study

Step 1: Carefully pose a question.

Step 2: Specify an economic or conceptual model.

Step 3: Turn the economic model into an econometric model.

Step 4: Collect data on the variables and use statistical methods to estimate the parameters, construct confidence intervals for the parameters, and test hypotheses.

Step 2. Specify an economic model

Step 2. Specify an **economic model**, or at least a conceptual model, to study the phenomenon of interest. Formal economic modeling (such as utility maximizaton) is often used, but one can get by with careful economic reasoning that is less formal.

Example

To study the effects of job training on worker productivity, where productivity is measured by observed hourly wage, we can start with an equation such as

$$wage = f(educ, exper, training)$$

where $educ$ is a measure of schooling, *exper* is a measure of workforce experience, and $training$ is a measure of time spent in job training (the variable of most interest).

Step 3: Turn the economic model into an **econometric** model

Example

We can specify an econometric model for the wage/job training example as

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u$$

- The constants β_0, β_1, β_2, and β_3 ("the betas") are the **parameters** of the model, and it is these (especially β_3 in this example) that we hope to estimate.

- Ideally we will be able to collect data on $wage$, $educ$, $exper$, and $training$ from a large group of working people.

Step 3: Turn the economic model into an **econometric** model

- The last term in the equation

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u$$

is u. It's called the **error term** or **disturbance**.

- It plays a very important role in econometrics.

- It represents all other factors that can affects someone's wage, including native intelligence, motivation, and so on.

- The error term can also capture measurement problems in one or more of the variables.

Step 3: Turn the economic model into an **econometric** model

- We will want to use statistical methods, and data, to estimate and test hypotheses about the **parameters**.
- For example, the hypothesis that job training has no effect on wage is $\beta_3 = 0$.
- The hypothesis that one year of experience is worth one year of education is $\beta_1 = \beta_2$.

Steps for a Successful Empirical Study

Step 1: Carefully pose a question.

Step 2: Specify an economic or conceptual model.

Step 3: Turn the economic model into an econometric model.

Step 4: Collect data on the variables and use statistical methods to estimate the parameters, construct confidence intervals for the parameters, and test hypotheses.

Topics

1 What is Econometrics?
Why Study Econometrics?
2 Steps in an Empirical Analysis
3 The Structure of Economic Data
4 Causality and the Notion of Ceteris Paribus

The Structure of Economic Data

- Economic data sets come in a variety of types.

Types of Economic Data

- Cross-Sectional Data
- Time Series Data
- Pooled Cross Sections
- Panel or Longitudinal Data

Cross-Sectional Data

- Data are collected on individuals, families, firms, schools, or some other units at a given point in time.
- Time is not important. It does not play a crucial role.
- We will assume that a cross-sectional data set represents a **random sample**.

What is random sample again? Example

Cross-Sectional Data

The Importance of Random Sampling

- Random sampling (with replacement) generates observations that are i.i.d.

- Intuitively, a random sample is **representative of the population of interest**, and gives us the best chance of learning about the population.

Cross-Sectional Data

- **Ordering of the individuals is arbitrary and unimportant.**

- Does not matter whom we label observation 2, or observation 5, or observation 136 and so on.

- What happens if I shuffle the dataset?

Nothing would be lost be if we randomly rearrange the order of the individuals.

Time Series Data

• Consists of observations on variables observed over a stretch of time. Examples include interest rates, unemployment rates, and crime rates.

• A key feature of time series data is that the **order is important**. We need to know, for example, that the outcome on unemployment in 2008 precedes that for 2009.

Time Series = correlated observations

Time Series Data

• Another important difference with cross-sectional data is that we cannot assume outcomes are independent across observation (that is, across time).

For example, knowing what gross domestic product is in 2009 tells us a lot about its likely range in 2010.

• When we apply econometric methods to time series data, we will have to recognize that the observations are correlated across time.

Time Series Data

- Many time series exhibit clear **trends**. While real GDP sometimes rises and sometimes falls, on average it has grown over time.
- The notion of trend is not relevant for cross-sectional data.
- The **frequency** with which time series data are recorded can also be important.
  - Are the data observed once a month?
  - Once a quarter?
  - Annually?
  - Daily, such as closing prices for the stock market.?

Independently Pooled Cross Sections

• A data set consisting of independently pooled cross sections means that we have collect cross-sectional data at different points in time and pool them together.

Example

We may randomly sample from the working U.S. population in 1990, 2000, and 2010.

Our goal may be to see how the importance of attending college on salaries has changed over time.

• If we obtain a **random sample in each year** it would be very small compared to the entire population.

• It would be **very rare that the same person would appear twice**; if someone appears twice nothing is harmed by ignoring that fact.

Independently Pooled Cross Sections

- Pooled cross sections are very useful for policy analysis - to study an intervention.
- The idea is to collect data from the years before and after a key policy change.

Example

How does a new incinerator affect the sales price of homes?

- Because the observations are independent (both within and across time periods), pooled cross sections can be analyzed much like a single cross section.

Independently Pooled Cross Sections

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85,500	42	1600	3	2.0
2	1993	67,300	36	1440	3	2.5
3	1993	134,000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243,600	41	2600	4	3.0
251	1995	65,000	16	1250	2	1.0
252	1995	182,400	20	2200	4	2.0
253	1995	97,500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57,200	16	1100	2	1.5

Source: Wooldridge, Jeffrey M. (2015). Introductory Econometrics: A Modern Approach.

Panel Data

- Panel data set has a structure similar to a pooled cross section.
- **Main difference:** the same units (people, houses, schools, and so on) are followed over time.
- Following the same units over time has advantages when trying to **infer causality.**
- Panel data analysis is a more advanced topic.

Panel Data

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350,000	8.7	440
2	1	1990	8	359,200	7.2	471
3	2	1986	2	64,300	5.4	75
4	2	1990	1	65,100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260,700	9.6	286
298	149	1990	6	245,000	9.8	334
299	150	1986	25	543,000	4.3	520
300	150	1990	32	546,200	5.2	493

Source: Wooldridge, Jeffrey M. (2015). Introductory Econometrics: A Modern Approach.

Topics

1. What is Econometrics?
   - Why Study Econometrics?
2. Steps in an Empirical Analysis
3. The Structure of Economic Data
4. Causality and the Notion of Ceteris Paribus

Causality

- The concept of **causality** is key in econometrics.

- Does bigger high schools *causes* students to have a higher GPA? Does bigger cities *causes* a higher number of crimes?

Correlation $\not\Rightarrow$ Causality

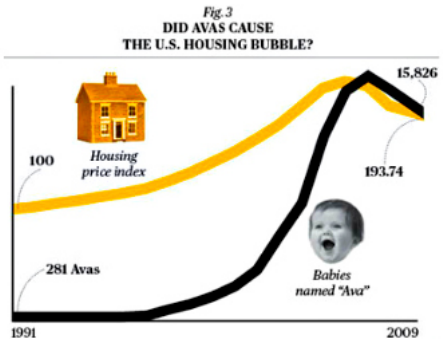
Finding correlations in data might be **suggestive** but is **RARELY conclusive**.

Causality

• Crucial to establishing causality is the notion of **ceteris paribus**: "all (relevant) factors equal."

• If we succeed, via statistical methods, in "holding fixed" all other relevant factors, then **SOMETIMES** we establish that changes in one variable (say, education) in fact "cause" changes in another variable (wage).

Causality



Causality

Example: What is the value of another year of education on one's earnings?

- We can imagine the type of experiment we would have to run to obtain experimental data.
- At birth, each child is randomly given a highest grade that he/she must complete – no more, no less.
- Then, we eventually record, hourly or monthly or annual earnings.

Causality

- The experiment is not feasible, and would be morally repugnant, anyway.

- For problems such as measuring the value of education, we must usually rely on observational data. We can, for very large random samples of people, collect information on education and earnings.

Causality

• The problem for inferring causality from, say, a simple correlation analysis is that individuals and their parents largely determine the amount of schooling.

• Probably on average people who are smarter or more capable choose to become better educated. But more capable people would earn more, on average, than less capable individuals.

• Seeing a positive correlation between earnings and schooling need not imply that it is due to schooling.

• Other **confounding factors** (such as intelligence and past experience) could explain most of the difference in earnings.

Causality

• The problem of individuals influencing their education levels is an example of **self-selection**.

Example

• Suppose we want to study the effects of attending college lectures on performance in a course.

• If the better students, on average, also attend lectures more frequently, a simple correlation analysis can be **misleading**.

The individuals (i.e.,students) self-select into the variable (i.e., how much they attend lecture).

• Self-selection is often a **serious concern in the social sciences.**