

Lecture 1

Introduction

(Chapter 1 in the textbook)

Econ 322 Econometrics

Spring 2025

- 1 Introduction to Econometrics
- 2 Steps in an Empirical Analysis
- 3 Causality and the Notion of Ceteris Paribus
- 4 The Structure of Economic Data

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, and evaluate government and business policy.
- The study of *econometrics* is completely intertwined with the study of *economics*. The two subjects cannot exist without the other.

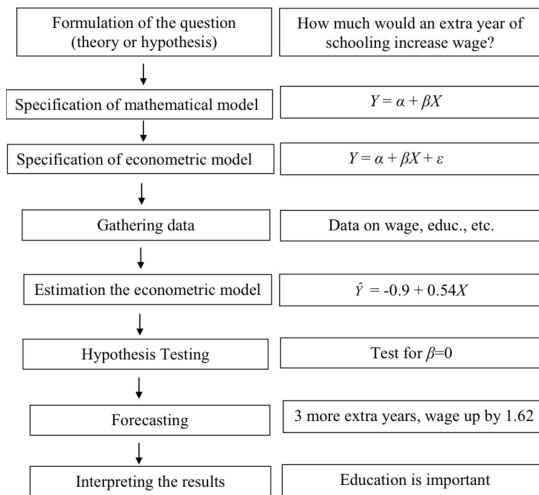
What is Econometrics?

- Econometrics is the application of statistical methods to economic data. As we will see during this course, economic theory helps us develop models of economic behavior while econometrics help us take these models to data and test our theories.
- Econometrics also help us find empirical regularities that need to be explained.
- Econometrics is its own discipline (separate from statistics) because it focuses on problems inherent in analyzing data generated by individuals, firms, and other entities acting strategically, and interacting with one another.

What is Econometrics?

- The purpose of this course is to introduce you to the tools needed to look at data and answer interesting economic questions. As we will see during this course, the nature of the subject and the data that we observe leads to interesting problems not commonly faced in other statistical problems.

Steps in Economic Analysis



Causality versus Correlation

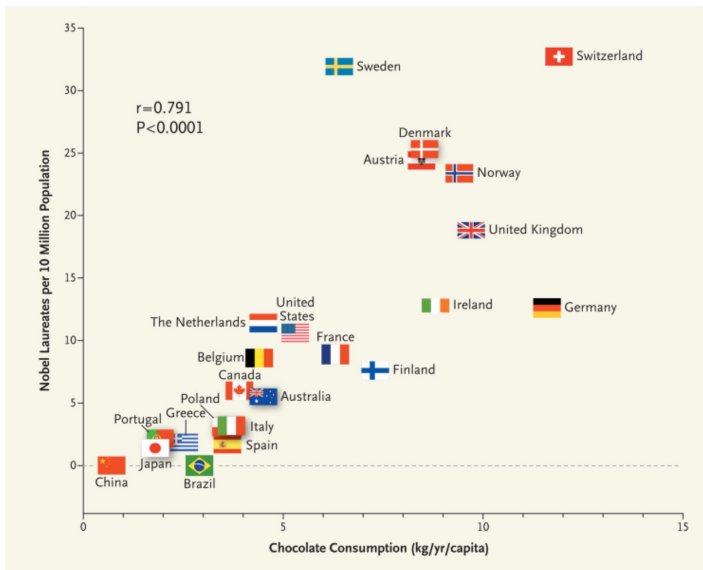
- We will typically be interested in measuring the causal impact of one variable (e.g., a job training program) on an outcome variable of interest (e.g., wages)
- It is important to distinguish this from simply finding a relationship (correlation) between two variables
- What we want to measure is the impact of one variable on another, **ceteris paribus; i.e., holding everything else constant**
 - It only makes sense when we say X causes Y when everything else is held constant

Examples of Bad Causal Influence

- The more firemen sent to a fire, the more damage that is done
 - Children who get tutored get worse grades than children who do not get tutored
 - Mortality rates are negatively correlated with the numbers of televisions per capita
- Both are impacted by GDP

Examples of Bad Causal Influence (cont)

- Serious crime is strongly correlated with the amount of ice cream sold by street vendors
Both increase with summer heat
- The size of your palm is negatively correlated with how long you live
Woman have smaller palms and live longer
- The number of Nobel prizes won by a country (adjusting for population) correlates well with per capita chocolate consumption



Other Examples

- Effects of fertilizer on crop yield
- The returns to education
- The effect of a change in minimum wages on employment
- The effect of law enforcement on city crime levels
- The impact of hospital visits on health outcomes
- The impact of a voluntary time-of-use rate on a household's pattern of electricity usage

Defining Causality

- We can define what we mean by causality using Rubin's notion of potential outcomes
- Suppose we are interested in the causal impact of a change (e.g., an additional year of education) on an outcome of interest (e.g., the individual's wages)
- For an individual, let
 $Y_i(1)$ denote the outcome with the change and
 $Y_i(0)$ denote the outcome without the change

Defining Causality

- The causal impact of the change for that individual is then $Y_i(1) - Y_i(0)$
- The problem is that we only observe one of these potential outcomes
- The practical problem is finding a way to fill-in the missing information
- Randomized experiments provide one approach
- Natural experiments provide another

Experiments

- The question that we face in this course is how best to answer causal questions given the data we have. In a perfect world we are able to formulate an experiment that we can use to generate data.

Experiments and Econometrics

In the study of economics, we usually cannot do controlled randomized experiments for a number of reasons:

- cost – usually very expensive to put such a trial together
- hard to do – very hard for a government (or authority) to experiment with the economy
- ethics – it would not be ethical in some cases to do an experiment:
e.g. suppose we wanted to know the economic impact of asbestos exposure to individuals. We couldn't form an experimental group and expose them to asbestos in order to find out!

Using Observational Data in Econometrics

- Unfortunately the use of experiments is very difficult in economics.
- We usually have access to *nonexperimental data*, which are sometimes called *observational* or *retrospective* data. These are data sets collected in a passive manner, after we observe outcomes on individuals, firms, schools, and so on. We simply act as “observers” of what has happened and then try to learn from what we observe.
- Using data of this type to make statistical inference has many problems. By the end of this course we will have enough tools to answer all but the most complicated questions.

Data

There are four main types of data we see in economics:

- 1 cross-sectional data
- 2 time series data
- 3 pooled cross-sections
- 4 panel data

Cross-Sectional Data

- Cross-sectional data involves looking at a group of individuals, firms, states, countries at one particular point in time. We are essentially using the cross-sectional variation across the observations in our sample for statistical analysis. (Or, at least, time does not play a crucial role. Interview dates for surveys may vary somewhat.)
- An example of cross-sectional data is depicted in Table 1.1. In this data we see various information about 420 California school districts in 1998.

Cross-Sectional Data

TABLE 1.1 Selected Observations on Test Scores and Other Variables for California School Districts in 1998

Observation (district) number	District average test score (fifth grade)	Student-teacher ratio	Expenditure per pupil (\$)	% of students learning English
1	690.8	17.89	\$6,385	0.0%
2	661.2	21.52	5,099	4.6
3	643.6	18.70	5,502	30.0
4	647.7	17.36	7,102	0.0
5	640.8	18.67	5,236	13.9
⋮	⋮	⋮	⋮	⋮
418	645.0	21.89	4,403	24.3
419	672.2	20.20	4,776	3.0
420	655.8	19.04	5,993	5.0
<i>Note:</i> The California test score data set is described in Appendix 4.1.				

Cross-Sectional Data

- It is useful to see how cross-sectional data sets are stored in common statistical packages. The following is a listing of data from the data set BEAUTY.DTA. It contains hourly wage information and a measure of “looks” for 871 married people. It also includes years of schooling (highest grade completed), gender, and workforce experience.
- The variable *female* is a binary (zero-one variable) indicating gender. The variables *belavg* and *abvavg* are also binary and are derived from the variable *looks*. We discuss binary variables later, along with why the natural log of wage and the square of experience appear in this data set.

```
. use beauty
```

```
. des
```

Contains data from beauty.dta

```
obs:      871
vars:      9
size:      20,904 (99.9% of memory free)
```

variable name	storage type	display format	value label	variable label
wage	float	%9.0g		hourly wage
female	byte	%8.0g		=1 if female
looks	byte	%8.0g		from 1 to 5
educ	byte	%8.0g		years of schooling
exper	byte	%8.0g		years of workforce experience
lwage	float	%9.0g		log(wage)
belavg	byte	%8.0g		=1 if looks <= 2
abvavg	byte	%8.0g		=1 if looks >=4
expersq	int	%8.0g		exper^2

Sorted by:

```
. list in 1/15
```

	wage	female	looks	educ	exper	lwage	belavg	abvavg	expersq
1.	5.73	1	4	14	30	1.745715	0	1	900
2.	4.28	1	3	12	28	1.453953	0	0	784
3.	11.57	0	3	16	38	2.448416	0	0	1444
4.	11.42	0	3	16	27	2.435366	0	0	729
5.	3.91	1	3	12	20	1.363537	0	0	400
6.	8.76	0	3	16	12	2.170196	0	0	144
7.	4.03	0	4	16	6	1.393766	0	1	36
8.	5.14	0	2	17	19	1.637053	1	0	361
9.	7.99	0	4	16	12	2.078191	0	1	144
10.	6.01	0	4	16	17	1.793425	0	1	289
11.	5.16	0	3	17	7	1.640936	0	0	49
12.	11.54	0	4	17	12	2.445819	0	1	144
13.	7.69	1	3	16	7	2.039921	0	0	49
14.	6.79	0	3	14	19	1.915451	0	0	361
15.	6.87	0	3	12	33	1.927164	0	0	1089

Cross-Sectional Data

- The previous “list” command shows us the first 15 people in the data set.
- Each row in the file represents one person. We read across the row to see the values for each variable. For example, person 1 earns \$5.73 an hour, is a female, has a “looks” rating of 4 (out of 5), has 14 years of education, and has 30 years of workforce experience.
- A feature of this data set, and other cross-sectional data sets that we use, is that the ordering of the individuals is arbitrary and unimportant. It is irrelevant whom we label observation 1, or observation 2, and so on. Nothing would be lost by randomly rearranging the order of the individuals.

Time Series Data

- The second type of data we use in economics is **time series** data. An example of time series data can be found in Table 1.2.
- In this table we see quarterly data on unemployment and inflation for the USA from 1959 until 2000. Thus time series data is the observation of one individual (person, firm, state, country) across many points in time.
- A key feature of time series data is that the order is (usually) important. We need to know, for example, that the outcome on unemployment in 2008 precedes that for 2009.

Time Series Data

TABLE 1.2 Selected Observations on the Rates of Consumer Price Index (CPI) Inflation and Unemployment in the United States: Quarterly Data, 1959–2000.

Observation number	Date (Year:quarter)	CPI inflation rate (% per year at an annual rate)	Unemployment rate (%)
1	1959:II	0.7%	5.1%
2	1959:III	2.1	5.3
3	1959:IV	2.4	5.6
4	1960:I	0.4	5.1
5	1960:II	2.4	5.2
⋮	⋮	⋮	⋮
165	2000:II	3.0	4.0
166	2000:III	3.5	4.0
167	2000:IV	2.8	4.0
Note: The U.S. inflation and unemployment data set is described in Appendix 12.1.			

Pooled Cross-Sections

- Data having both time series and cross-sectional features
- Combines cross-sections gathered in distinct time periods, but not for the same agents (except by chance)
- The advantage over a single cross-section is that one can examine changes over time
- This can be particularly useful in studying the impact of policy changes

Panel Data (Longitudinal Data)

- The final type of data is a mix of the first two. Superficially, a **panel** data set has a structure similar to a pooled cross section. The key difference is that with a panel data set the same units (people, houses, schools, and so on) are followed over time.
- For example, there are a couple of thousand elementary schools in New Jersey. We can collect information on test pass rates, spending, and some socioeconomic variables annually over, say, the last 10 years.
- Following the same units over time has advantages when trying to infer causality.
- An example of this data is given in Table 1.3. Here we can use both cross-sectional and time variation to answer any economic question we have about the data.

Panel Data

TABLE 1.3 Selected Observations on Cigarette Sales, Prices, and Taxes, by State and Year for U.S. States, 1985–1995

Observation number	State	Year	Cigarette sales (packs per capita)	Average price per pack (including taxes)	Total taxes (cigarette excise tax + sales tax)
1	Alabama	1985	116.5	\$1.022	\$0.333
2	Arkansas	1985	128.5	1.015	0.370
3	Arizona	1985	104.5	1.086	0.362
⋮	⋮	⋮	⋮	⋮	⋮
47	West Virginia	1985	112.8	1.089	0.382
48	Wyoming	1985	129.4	0.935	0.240
49	Alabama	1986	117.2	1.080	0.334
⋮	⋮	⋮	⋮	⋮	⋮
96	Wyoming	1986	127.8	1.007	0.240
97	Alabama	1987	115.8	1.135	0.335
⋮	⋮	⋮	⋮	⋮	⋮
528	Wyoming	1995	112.2	1.585	0.360
Note: The cigarette consumption data set is described in Appendix 10.1.					