Overview [1]

Jasmine(Yu) Hao

Faculty of Business and Economics
Hong Kong University

August 30, 2021

[1] This section is based on Stock and Watson (2020), Chapter 1.

What is econometrics I

- the science of **testing economic theories**
- the set of tools used for **forecasting future values of economic variables**
- process of **fitting mathematical economic models** to real-world data
- it is the science and art of using historical data to make numerical, or quantitative, policy recommendations in government and business.

What is econometrics II

use economic theory and statistical techniques to analyze economic data.

- ▷ widely applied: finance, labor economics, macroeconomics, microeconomics, marketing, and economic policy.
- ▷ commonly used in other social sciences.
- ▷ This course:
  - ◇ introduce to core set of methods
  - ◇ answer quantitative questions about business and government policy.
  - ◇ decisions rely on understanding relationships among variables in the world around us.
  - ◇ require quantitative answers to quantitative questions.

Example Question #1 I

Does Reducing Class Size Improve Elementary School Education?

Debate for reform of the U.S. public education system: reduce class sizes at elementary schools increase students grades?

- **costs money**: It requires hiring more teachers and (potentially) building more classrooms.
- must weigh these costs against the benefits, **precise quantitative understanding** of the likely benefits.
  - How beneficial is the effect on basic learning of smaller classes? Is it possible that smaller class size actually has no effect on basic learning?

Example Question #1 II

- Common sense: more learning occurs when there are fewer students.
- Examine empirical evidence(evidence based on data) relating class size to basic learning in elementary schools.
- Dataset: 420 California school districts in 1999.
- Data suggests: Students in districts with small class sizes tend to perform better on standardized tests than students in districts with larger classes.

Causal Relationship I

Correlation does not suggest causal relationship

- Fact: coincide with the idea that smaller classes produce better test scores,
- Other advantages that students in districts with small classes have
  - districts with small class sizes tend to have wealthier residents, could have more opportunities for learning outside the classroom.
  - It could be **these extra learning opportunities** that lead to higher test scores, not smaller class sizes.

Example Question #2 I

Is There Racial Discrimination in the Market for Home Loans?

- By law, U.S. lending institutions **cannot take race into account** when deciding to grant or deny a request for a mortgage:
  - Applicants who are identical in all ways except their race should be equally likely to have their mortgage applications approved.
- In theory, then, there should be no racial bias in mortgage lending.
- Researchers at the Federal Reserve Bank of Boston found (using data from the early 1990s)
  - that 28% of black applicants are denied mortgages,
  - while only 9% of white applicants are denied.

Example Question #2 II

- Fact: more black than white applicants are denied in the Boston Fed data
  - **does not** by itself provide evidence of discrimination by mortgage lenders
  - because the black and white applicants **differ in many ways other than their race**.
  - Need to examine: is there a difference in the probability of being denied for **otherwise identical applicants** and, how big is this difference?
- Method to quantify the effect of race on the chance of obtaining a mortgage, **holding constant other applicant characteristics**(their ability to repay the loan.)

Example Question #3 I

Does Healthcare Spending Improve Health Outcomes?

- Healthcare: reduce avoidable deaths, improvement of the health-related quality of life of individuals
- How to measure the benefit and cost of health care?
  - absolute and per capita terms, variations in health outcomes across countries, life expectancy at birth.

Example Question #3 II

Basic economics says that **more expenditure on healthcare should generally reduce avoidable mortality**.

- **spending elasticity for mortality**. If the amount spent on healthcare increases by 1%, by what percentage will avoidable mortality decrease?
- policy objectives are based on avoidable mortality;
  - e.g., one of the United Nations Development Programme's sustainable development goals is that all countries should aim to reduce "under-5 mortality to at least as low as 25 per 1,000 live births."
- But how should the goal be met: from expanding healthcare services or other services? if increasing healthcare spending is to form part of the mix of policies, by how much will it need to increase?

To answer: estimates of the spending elasticity for mortality. examine empirical evidence about the returns to healthcare spending—either based on variations in spending across countries or within countries over time.

Example Question #3 III

Challenges: the **extensive heterogeneity** across countries.

- **observable heterogeneity**: factors that affect countries' mortality rates that may also be associated with healthcare expenditure,
  - the income per capita of each country. This can be controlled for using **multiple regression analysis**.
- **unobservable heterogeneity**: Unobserved factors determining both health expenditure and health outcome.
- **Simultaneously causality**: causality running in both directions.

Quantitative Questions, Quantitative Answers I

Each of these questions requires a **numerical answer**.
Economic theory provides clues about that answer, but the actual value of the number must be learned empirically, that is, by analyzing data.
Uncertainties in the answers:

- A different set of data would produce a different numerical answer.
- Conceptual framework to measure of how precise the answer is.

Causal Effects and Idealized Experiments I

- Causality means that a specific action (applying fertilizer) leads to a specific, measurable consequence (more tomatoes).
    - an action is said to cause an outcome if the outcome is the direct result of that action.
    - eg. Touching a hot stove causes you to get burned, drinking water causes you to be less thirsty, putting air in your tires causes them to inflate, putting fertilizer on your tomato plants causes them to produce more tomatoes.

Estimation of Causal Effects I

- One way to measure the causal effect on tomato yield (measured in kilograms) of applying a certain amount of fertilizer is to conduct an experiment.
- plants many plots of identical tomatoes.
    - Some plots get 100 grams of fertilizer per square meter.
    - At the end of the growing season, the weighs the harvest from each plot.
    - The difference between the average yield per square meter is the effect of the fertilizer treatment.

Estimation of Causal Effects II

- This is an example of a **randomized controlled experiment**.
  - a **control group** that receives no treatment (no fertilizer)
  - a **treatment group** that receives the treatment (100 g/m^2 of fertilizer). It is **randomized** in the sense that the treatment is assigned randomly.
- Eliminates the possibility of a systematic relationship between,
  - for example, how sunny the plot is and whether it receives fertilizer.
- If this experiment is properly implemented on a large enough scale, then it will yield an estimate of the **causal effect** on the outcome of interest (tomato production) of the treatment (applying 100 g/m^2 of fertilizer).

Estimation of Causal Effects III

Causal effect: effect on an outcome of a given action or treatment, as measured in an ideal randomized controlled experiment.

- only systematic reason for differences in outcomes is the treatment itself.
- In social science: not an option because they are unethical, impossible to execute satisfactorily, too time-consuming, or prohibitively expensive.
- Even with nonexperimental data, important to think ideal randomized controlled experiment is important - it provides a definition of a causal effect.

Experimental versus Observational Data I

Experimental data come from experiments designed to evaluate a treatment or policy or to investigate a causal effect.

- the state of Tennessee financed a large randomized controlled experiment examining class size in the 1980s.(S.W. Ch. 13)
- thousands of students were randomly assigned to classes of different sizes for several years and were given standardized tests annually.
- The Tennessee class size experiment cost millions of dollars and required the ongoing cooperation of many administrators, parents, and teachers over several years.

Experimental versus Observational Data II

Why cannot experiment?

- human subjects are difficult to administer and to control,
- expensive and difficult to administer, also unethical.
- Most economic data are obtained by observing real-world behavior.(**observational data**).
- Data source:
  - Surveys: e.g. telephone surveys of consumers,
  - administrative records, e.g. such as historical records on mortgage applications maintained by lending institutions.

Experimental versus Observational Data III

Challenges to estimate causal effects

- In the real world, levels of "treatment" (the amount of fertilizer in the tomato example, the student–teacher ratio in the class size example) are **not random**, so it is difficult to sort out the effect of the "treatment" from other relevant factors.
- Much of econometrics devote to methods for challenges to estimate causal effects.
- Data Types: cross-sectional data, time series data, and panel data.

Cross-Sectional Data I

- Data on different entities—workers, consumers, firms, governmental units, and so forth— for a single time period are called **cross-sectional** data.
- For example, the data on test scores in California school districts are cross sectional. Those data are for 420 entities (school districts) for a single time period (1999).
- In general, the number of entities on which we have observations is denoted n; for example, in the California data set, $n = 420$.

Cross-Sectional Data II

TABLE 1.1 Selected Observations on Test Scores and Other Variables for California School Districts in 1999

Observation (District) Number	District Average Test Score (fifth grade)	Student–Teacher Ratio	Expenditure per Pupil ($)	Percentage of Students Learning English
1	690.8	17.89	$6385	0.0%
2	661.2	21.52	5099	4.6
3	643.6	18.70	5502	30.0
4	647.7	17.36	7102	0.0
5	640.8	18.67	5236	13.9
⋮	⋮	⋮	⋮	⋮
418	645.0	21.89	4403	24.3
419	672.2	20.20	4776	3.0
420	655.8	19.04	5993	5.0

Note: The California test score data set is described in Appendix 4.1.

Link to download the data:`http://fmwww.bc.edu/ec-p/data/stockwatson/caschool.des`.

Cross-Sectional Data III

- The California test score data set contains measurements of several different variables for each district.
- Each row lists data for a different district.
  - The average student–teacher ratio in that district is 17.89.
  - Average expenditure per pupil in district 1 is $6385.
  - The remaining rows present data for other districts. The order of the rows is arbitrary, and the number of the district, which is called the **observation number**, is an arbitrarily assigned number that organizes the data.

Time Series Data I

Time series data are data for a single entity (person, firm, country) collected at multiple time periods.

- Our data set on the growth rate of GDP and the term spread in the United States is an example of a time series data set.
  - the growth rate of GDP and the term spread
  - for a single entity (the United States)
  - for 232 time periods.
- The number of observations (that is, time periods) in a time series data set is denoted T(1960:Q1 to 2017:Q4), this data set contains $\mathrm{T} = 232$ observations.

Time Series Data II

TABLE 1.2 Selected Observations on the Growth Rate of GDP and the Term Spread in the United States: Quarterly Data, 1960:Q1–2017:Q4

Observation Number	Date (year: quarter)	GDP Growth Rate (% at an annual rate)	Term Spread (percentage points)
1	1960:Q1	8.8%	0.6
2	1960:Q2	−1.5	1.3
3	1960:Q3	1.0	1.5
4	1960:Q4	−4.9	1.6
5	1961:Q1	2.7	1.4
⋮	⋮	⋮	⋮
230	2017:Q2	3.0	1.4
231	2017:Q3	3.1	1.2
232	2017:Q4	2.5	1.2

Note: The United States GDP and term spread data set is described in Appendix 15.1.

Panel Data I

- Panel data, also called **longitudinal data**, are data for multiple entities in which each entity is observed at two or more time periods.
- Our data on cigarette consumption and prices are an example of a panel data set.

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 12.1.

Panel Data II

- The **number of entities** in a panel data set is denoted n,
- and the **number of time periods** is denoted T.
- In the cigarette data set, we have observations on $n = 48$ continental U.S. states (entities) for $T = 11$ years (time periods) from 1985 to 1995.
- Thus, there is a total of $n * T = 48 * 11 = 528$ observations.

References I

Stock, J. H. and Watson, M. W. (2020). *Introduction to econometrics*, volume 4. Pearson New York.