Introductory Econometrics
Ch1 An Introduction

LIU Chenyuan

Spring 2024

Outline

Course Information

What is Econometrics?

The Structure of Economic Data

Instructors

- ► LIU Chenyuan (before the Midterm)
  Office: Lihua B629
  Email:liuchy3@sem.tsinghua.edu.cn
  Office hours: Thursday 4 - 5 pm

- ► FENG Yingjie (after the Midterm)
  Office: Lihua B628
  Email:fengyj@sem.tsinghua.edu.cn
  Office hours: Thursday 4 - 5 pm

Course Information

- ▶ Two parallel sessions:
    1. Monday morning at 9:50 - 12:15
    2. Monday evening at 19:20 - 21:45
- ▶ Feel free to go to any session.
- ▶ Prerequisite: undergraduate-level probability theory and mathematical statistics.
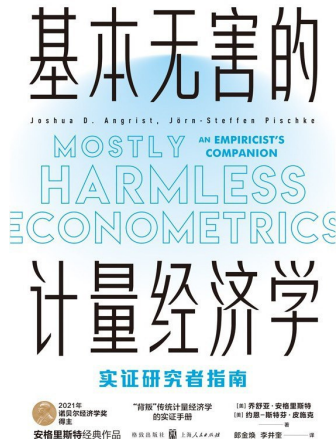- ▶ Course Website: Web-Learning System on Info

Grading Policy

1. Lecture
2. Homework
   - ▶ Four assignments, 20%
   - ▶ Please contact TAs with any questions about the HW.
   - ▶ Office hours: Thursday before HW due date at 2 - 3 pm
   - ▶ Late submissions get partial scores (see syllabus)
3. Exam
   - ▶ Midterm exam: April 14, Sunday evening, 40%
   - ▶ Final exam: TBD, 40%

Teaching Assistants

In the order of HW grading:

- ▶ LIN Pengsheng, lps22@mails.tsinghua.edu.cn
- ▶ LIU Jiayue, jy-liu21@mails.tsinghua.edu.cn
- ▶ LOU Jing, louj21@mails.tsinghua.edu.cn
- ▶ PENG Lu, pengl.18@sem.tsinghua.edu.cn

Programming Exercises: Stata

We will have Stata sessions in the lecture.

Data Processing with Stata — Cheat Sheet

For more info, *see* Stata's reference manual (stata.com)

Useful Shortcuts

F2 — keyboard buttons	**Ctrl** + **9**
describe data	open a new do-file
Ctrl + **8**	**Ctrl** + **D**
open the data editor	highlight text in do-file, then ctrl + d executes it in the command line
clear	
delete data in memory	

AT COMMAND PROMPT

PgUp **PgDn**	scroll through previous commands
Tab	autocompletes variable name after typing part
cls	clear the console (where results are displayed)

Set up

pwd
print current (working) directory

cd *'C:\Program Files\Stata16'*
change working directory

dir
display filenames in working directory

dir *.dta*
list all Stata data in working directory — underlined parts are shortcuts; use "capture" for "cap"

capture log close
close the log on any existing do-files

log using "myDoFile.txt", **replace**
create a new log file to record your work and results

search mdesc

ssc install mdesc — packages contain extra commands that expand Stata's toolkit
install the package mdesc; needs to be done once

Import Data

sysuse auto, clear — for many examples, we load current data (auto data)

use 'yourStataFile.dta', **clear**
load a dataset from the current directory

import excel "yourSpreadsheet.xlsx", /* — frequently used options are shortcuts: use "clear", */ sheet("Sheet1") cellrange(A2:H11) firstrow

import delimited "yourFile.csv", /*
*/ rowrange(2:11) colrange(1:8) varnames(2)

Basic Syntax

All Stata commands have the same format (syntax):

[**by** varlist:]	*command*	[varlist2]	[=exp]	[if exp]	[in range]	[weight]	[using filename]	[,options]
apply the command across each unique combination of variables in *varlist*	function: what are you going to do to *varlist*?		to save output as a new variable	condition: only apply the function if something is true	apply to specific rows	apply weights	pull data from a file (if not loaded)	special options for *command*

bysort rep78 : summarize price if foreign == 0 & price >= 9000, detail

To find out more about any command–like what options it takes–type **help** *command*

Basic Data Operations

Arithmetic
+ add (numbers) combine (strings)
+ ! or ~ not
- subtract
* multiply
/ divide
^ raise to a power

Logic
& and
| or
== tests if something is equal
= assigns a value to a variable
< less than
<= less than or equal to
> greater than
>= greater or equal to
!= ~= not equal

if foreign == 1 & price <= 10000

	Chevy Cet.	4	3,984
	Chev. Chev.	4	3,299
	Buick Rivera	6	10,372
	Honda Civic	4	4,499
	Volvo 260	6	11,995

if foreign == 1 | price >= 10000

	Chevy Cet.	4	3,984
	Chev. Chev.	4	3,299
	Buick Rivera	6	10,372
	Honda Civic	4	4,499
	Volvo 260	6	11,995

Change Data Types

Stata has 6 data types, and data can also be missing:
no data   true/false   words   numbers
missing   byte   string   int long float double
To convert between numbers & strings:

1   **gen** foreignString = string(foreign)
     tostring foreign, **gen**(foreignString)
     decode foreign, **gen**(foreignString)   "1"
                                                       "1"
                                                       "foreign"

1   **gen** foreignNumeric = real(foreign)   "1"
     destring foreignString, **gen**(foreignNumeric)   "1"
     encode foreignString, **gen**(foreignNumeric)   "foreign"

recast double mpg
generic way to convert between types

Explore Data

VIEW DATA ORGANIZATION

codebook make price
describe data type, format, and any value/variable labels

count
number of rows (observations)

count if price > 5000
can be combined with logic

ds, has(type string)
search for variable types, variable name, or variable label

isid mpg
check if mpg uniquely identifies the data

SEE DATA DISTRIBUTION

codebook make price
overview of variable type, stats, number of missing/unique values

summarize make price mpg
print summary statistics (mean, stdev, min, max) for variables

inspect mpg
show histogram of data and number of missing or zero observations

histogram mpg, **frequency**
plot a histogram of the distribution of a variable

BROWSE OBSERVATIONS WITHIN THE DATA

browse or **Ctrl** + **8** — Missing values are treated as the largest positive number. To exclude missing values, ask whether the value is less than *
open the data editor

list make price **if** price > 10000 & !missing(price)   **clist**... (compact form)
list the make and price for observations with price > $10,000

display price[4]
display the 4th observation in price; only works on single values

gsort price mpg   (ascending)   **gsort** –price –mpg   (descending)
sort in order, first by price then miles per gallon

Change Data Types (Summarize Data)

Summarize Data

include missing values : create binary variable for every rep78

tabulate rep78, **mi gen**(repairRecord)
one-way table: number of rows with each value of rep78

tabulate rep78 foreign, **mi**
two-way table: cross-tabulate number of observations for each combination of rep78 and foreign

bysort rep78 : **tabulate** foreign
for each value of rep78, apply the command tabulate foreign

tabstat price with mpg, **by**(foreign) **stat**(mean sd n)
create compact table of summary statistics   displays stats formats numbers for all data

table row, **contents**(freq mean sd price) f(%9.2fc) **row**
create a flexible table of summary statistics

collapse (mean) price (max) mpg, **by**(foreign) — replaces data
calculate mean price & max mpg by car type (foreign)

Create New Variables

generate mpgSq = mpg^2   **gen byte** lowPr = price < 4000
create a new variable. Useful also for creating binary variables based on a condition (**generate byte**)

generate n = _n
display the 4th observation in price

gen n = _n   **bysort** rep78: **gen** repairIdx = _n
_n creates a running index of observations in a group

generate totRows = _N   **bysort** rep78: **gen** repairTot = _N

7

Textbooks

Course Outline

- ▶ All you need to follow is the lecture notes.
- ▶ The textbook and references are recommended but not required.
- ▶ We will adjust the pace as the class goes on and will try to cover all topics listed.
- ▶ Once you finish this course, Introductory Econometrics II is highly recommended!

Outline

What is Econometrics?

▶ Econometrics is an application of statistical methods to economic data in order to give empirical content to economic relationships.

▶ What do econometrics do?
  ▶ Combine statistical techniques with economic theory
  ▶ Estimating economic relationships
  ▶ Testing economic theories
  ▶ Evaluating and implementing government and business policy

Three Types of Econometrics Questions

There are three basic types of questions that economists are interested in:

► Descriptive

► Forecasting

► Causal (or structural)

Descriptive Questions

We are interested in describing the pattern in the economic data. Examples:

- ► How much do men and women earn in 2023 on average in China?
- ► Do interest rates and the stock market move together?
- ► Do institutional investors earn more than individual investors?
- ► What was the price level of consumer goods last quarter?

Descriptive Questions

Typically, if we have data, we would know the answers to these questions. There are two challenges:

1. Sampling
   We typically do not observe the full population but rather a sample. We need to draw conclusions about the population based on the sample.

2. Summary statistics
   Some data are complicated, so we need to come up with a nice way to summarize them.

Example: iCPI

Internet-based consumer price index:
http://www.bdecon.com/
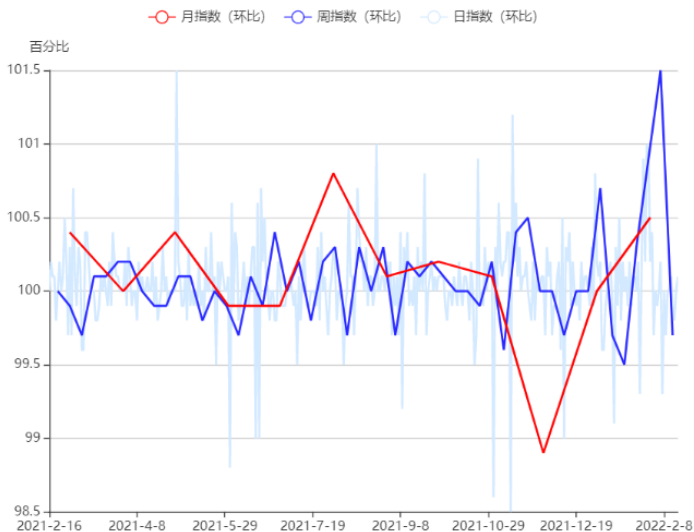
Forecasting Questions

We want to forecast something that has not yet happened.

▶ What will the GDP growth rate be next year?

▶ What will the Shanghai Stock Exchange Composite Index be tomorrow?

▶ How much revenue will my firm earn next month?

▶ What is the likelihood of another pandemic around the world in the next three years?

▶ What will my salary be when I graduate?

Forecasting Questions

- ▶ We never know the answers exactly until they happen.
- ▶ Sometimes, knowing what is likely to happen can be important.

There are two challenges:

1. Under-fitting: model does not explain the current data well
2. Over-fitting: model explains the current data too well

Under-fitting vs Over-fitting
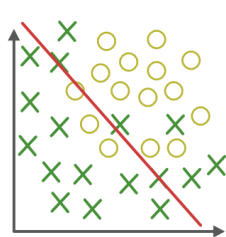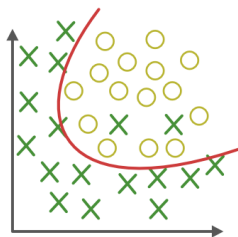
Under-fitting · Appropirate-fitting · Over-fitting
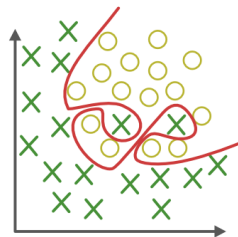
Causal Questions (Structural Questions)

We are interested in understanding causal relations to guide our decisions: "What will happen if something else happens?"

Examples:

- ▶ If the central bank lowers interest rates today, what will happen to inflation tomorrow?

- ▶ If the government raises the cigarette tax by 1 pp, how many people will quit smoking?

- ▶ If I exercise one hour every day from now on, how many more years will I live?

- ▶ How much more money will you earn as a result of taking this course?

Correlation vs Causation

▶ Correlation: how two random variables move together
▶ The difference between causation and correlation is a key concept in econometrics.
▶ Suppose we observe that two variables, X and Y are correlated.
▶ What does this suggest?

Case 1: $X \to Y$ or $Y \to X$

- Money supply $\to$ inflation
- Increase in minimum wage $\to$ increase in wages
- Retirement $\to$ decline in income

Case 2: $X \to Y$ and $Y \to X$

Does police reduce crime?[1]

Source: FBI Uniform Crime Reports, 1999
Police and Crime Per Capita from 2494 U.S. Towns and Cities with Population Greater than 10,000

---
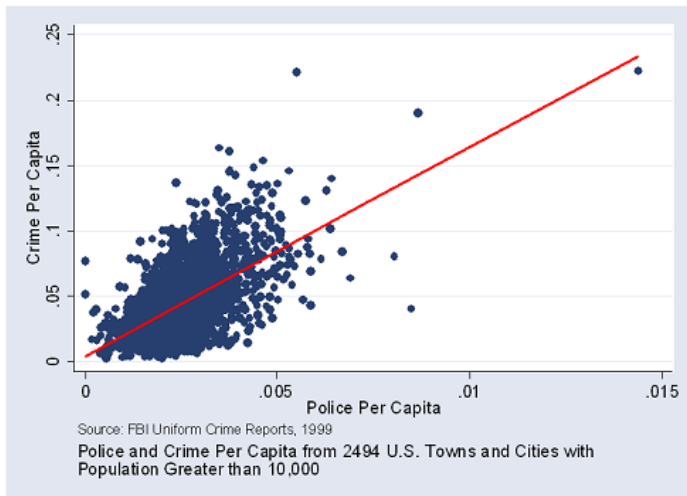
[1] S.D. Levitt, "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime," *American Economic Review*, 87(3), (June 1997), pp. 270-90.

Case 3: $Z \to X$ and $Z \to Y$

Drowning Deaths and Ice Cream Consumption by Month in Spain (2018)

Legend: ● Drowning Deaths ● Ice Cream Consumption In Kilograms
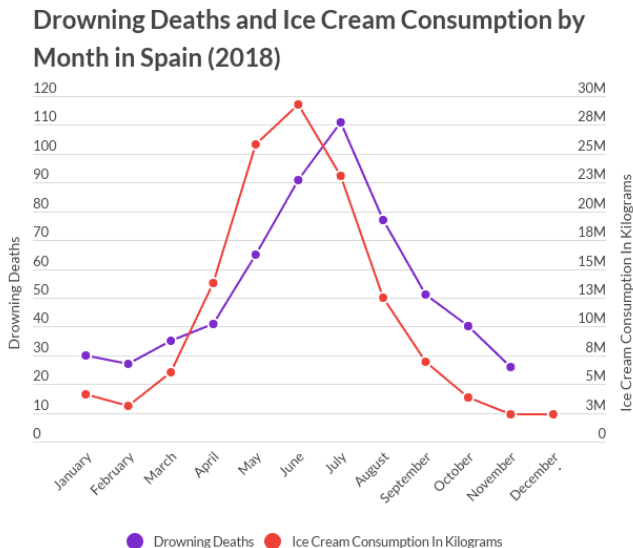
Correlation vs Causation

▶ The difference between causation and correlation is a key concept in econometrics.
▶ We would like to
  ▶ **identify** causal effects
  ▶ **estimate** their magnitude.
▶ It is generally agreed that this is very difficult to do; having an economic model is often essential in establishing the causal interpretation.

Causal Questions

▶ In natural sciences, we can answer them by running experiments.

▶ Economists sometimes do experiments.
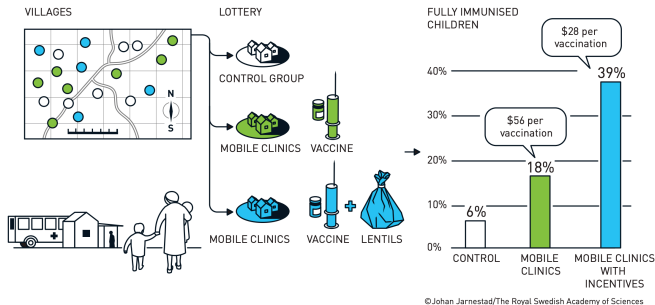
Figure: 2019 Nobel Prize:"experimental approach to alleviating global poverty"

Challenges of Running Experiments
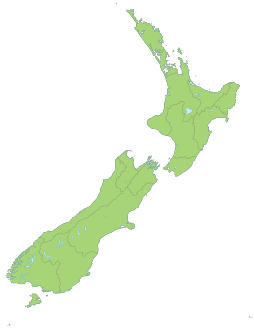
Challenges of Running Experiments
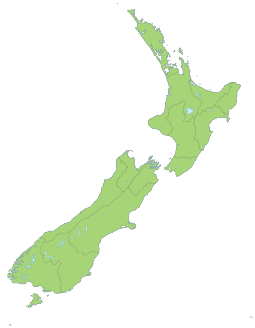
Sometimes...

► Experiments are impossible

► Experiments are expensive

► Experiments are morally repugnant

Experimental vs Observational Data

- ▶ Experimental data: collected in controlled experiments
- ▶ **Observational data**: researcher is a passive collector of the data
  - ▶ Also called nonexperimental data or retrospective data
- ▶ Econometrics focuses on causal problems inherent in collecting and analyzing observational economic data.

Econometrics and Observational Data

▶ In this course, we mainly focus on causal problems.

▶ We generally will never know the answer to these questions without running an experiment.

▶ Econometrics rarely "solves" causal problems. However, it is very useful for helping us understand them.

Summary of Three Types of Analysis

▶ Descriptive: if we had data, we would know the answer.

▶ Forecasting: if we wait long enough and have data, we will know the answer.

▶ Causal: unless we can run the perfect experiment, we will never know the answer for sure.

Outline

The Structure of A Dataset[2]

Imagine a two-dimensional table

- ▶ Each column represents a **variable**
- ▶ Each row represents an **observation**
- ▶ Each cell represents a **value**

---

[2]Grolemund Garrett and Wickham Hadley. Ch12 Tidy data, R for Data Science

Cross-Sectional Data

▶ Each observation is uniquely determined by an **unit**

▶ A cross-sectional data set consists of a sample of units taken at a given point in time.

▶ A unit could be an individual, a household, a firm, a city, a province, a country, etc.

ID	Province	GDP
1	Jiangxi	x
2	Hebei	y
3	Shandong	z

▶ We typically assume the sample is drawn from the underlying population randomly.

Time Series Data

▶ Each observation is uniquely determined by **time**

▶ A time series data set consists of observations on a variable or several variables over time.

ID	Year	GDP	avg income
1	2015	x	a
2	2016	y	b
3	2017	z	x

▶ Time is an important dimension in a time series data set.
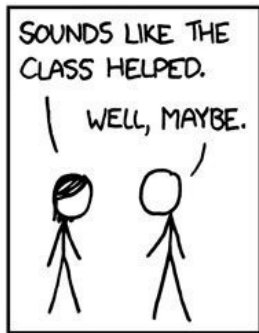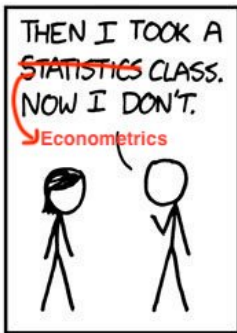
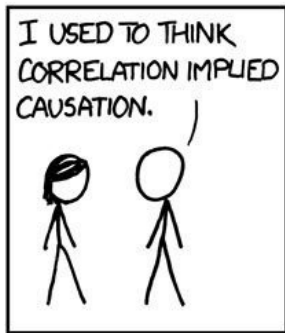▶ Typically, observations across time are correlated.

Pooled Cross Sections and Panel Data

- ► Pooled cross-sections include cross-sectional data in multiple years.
- ► A panel (longitudinal) data set consists of a time series for each cross-sectional member in the data set.
- ► Panel data: the same units over time; pooled cross sections: different units, different times.
- ► Each observation is uniquely determined by the unit and the time.

ID	Year	Province	GDP
1	2015	Jiangxi	x
2	2016	Jiangxi	y
3	2017	Jiangxi	z
4	2015	Hebei	s
5	2016	Hebei	t
6	2017	Hebei	u

Recap

- Three types of analyses
  - Descriptive
  - Forecasting
  - **Causal**
- Three types of data structure
  - **Cross-sectional**
  - Time series
  - Pooled cross section and panel