Econometrics 2 (Part 1)

Sergey Lychagin

Central European University

Winter 2020

Contact

- Sergey Lychagin: LychaginS@ceu.edu
  Office: Budapest campus, Nador 13, 5
- Arieda Muço: MucoA@ceu.edu
  Office: Budapest campus, Nador 13, 507
- Boldizsár Juhász (TA): Juhasz_Boldizsar@phd.ceu.edu

Textbooks

- Introductory Econometrics: A Modern Approach by Wooldridge
- Mostly Harmless Econometrics by Angrist and Pischke
- Reading list of applied papers

Grading

- **Quizzes in Class** (5% of the final grade)
- **Problem Sets** (10%)
- **Term paper** (25%)
  - Mandatory for passing the course
- **Final Exam** (60%)
  - The exam will be closed book, closed notes
  - Cover the material presented in class and seminars

Problem Sets

- Problem sets:
  - Replication of estimates from published papers
  - Practice questions from the past exams
- Work in groups, submit one solution per group.
- Solutions have to be turned in at specified dates (via Moodle)

Stata and TA Sessions

- CEU lab
- Problem sets will use Stata
- You are supposed to have good knowledge of Stata from Econometrics 1
- First TA session will brush up your Stata and data wrangling skills
- Six TA sessions in total

Term Paper

- An individual term paper is required at the end of the course
- The paper should consist of a simple empirical analysis
  - It should be at most 10 pages long (including tables and figures), text 1.5 or double spaced, font size 12
- There will be three deadlines
  - Submit a research proposal via Moodle (deadline TBA)
  - Hand in the first draft for review at the Center for Academic Writing (**May 17**)
  - Upload in Moodle the final version in **.pdf** format (**May 22**)

Talk to us about your topic before submitting the proposal.

Course Goals

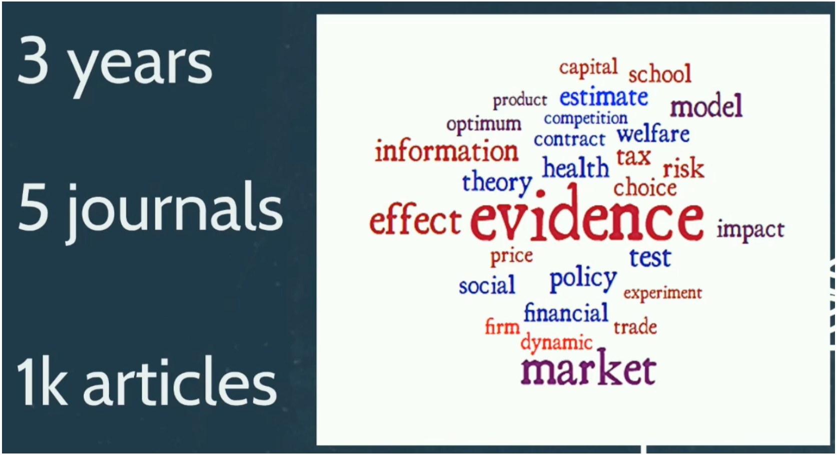
Guidance for conducting sensible applied research projects

1. Formulate research question
2. Find data
3. Econometric method, statistical analysis
4. Interpret estimation results
5. Policy recommendations

Outline: Tentative Schedule

1. Economic research questions: causality
2. The experimental ideal
3. Linear regression
4. Instrumental Variables
5. Panel Data, Fixed Effects, Differences-in-Differences
6. Matching
7. Program Evaluation: Nonparametric Methods
8. Regression Discontinuity

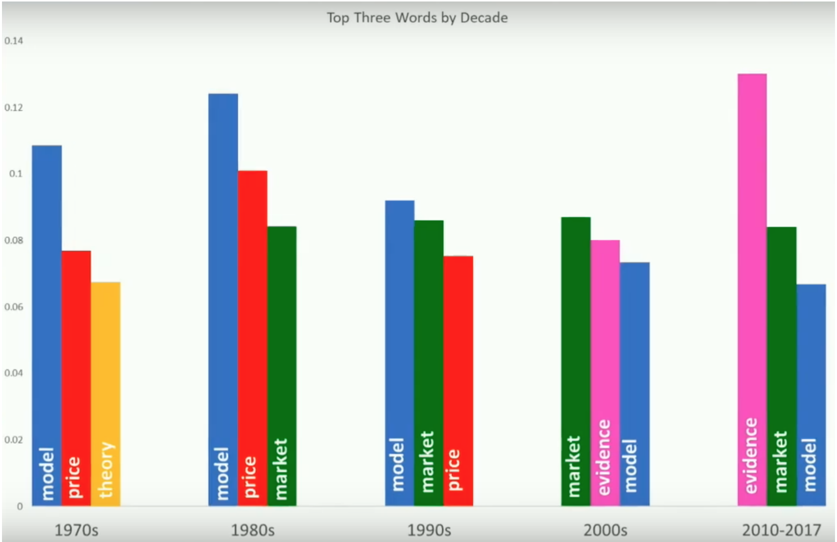
Della Vigna and Card



3 years

5 journals

1k articles


Top Three Words by Decade
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
model
price
theory
model
price
market
model
market
price
market
evidence
model
evidence
market
model
1970s
1980s
1990s
2000s
2010-2017

Research Strategy

1. What is the relationship of interest
   - Descriptive (*"Are women more risk averse than men?"*)
   - Prediction in a fixed environment (*"If a consumer likes movie A, would she watch movie B?"*)
   - **Causal inference**
     - ⋆ A causal relationship is useful for making predictions about the consequences of changing circumstances
     - ⋆ i.e Effect of education on wages, effect of class size on student achievement, deworming on education and health, institutions on growth ...
2. Which experiment could be ideally used?
   - Often hypothetical but helps formulating the question
3. Can we find the effect if we have infinite data ("identification strategy")?
4. How can we use finite sample to get a good estimate and test hypotheses (estimation/inference)?

Experimental data

Hard and costly to obtain

- Controlled laboratory experiment
- Randomized controlled trials (RCTs)

Give an extra year of education to a random group and compare means in the treatment and control group.
Randomization ⇒ the extra year is unrelated to parental education, income, etc.

Why using experiments?

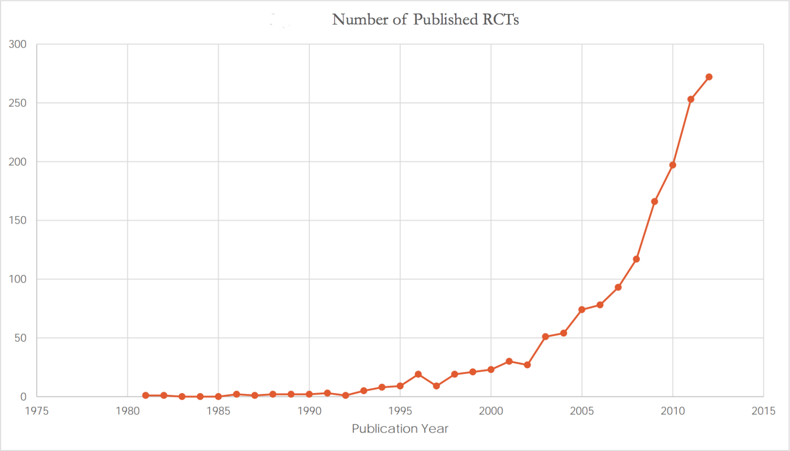
Example: Do hospitals make people healthier?

- Compare health status of people who were treated in hospital with those not treated in the average population
- Get data from National Health Interview Survey (NHIS)
- "During the last month, did you stay in hospital overnight?" Yes/No
- "Your health is": Excellent (1), very good (2), good (3), fair (4), poor (5)

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

Selection: who is hospitalized? **An experiment would take care of this selection issue**

Cameron et al (2016): RCT in Development Economics



Potential Outcome Model

$D_i = \{0, 1\}$ treatment variable (hospital care)
For each population unit i we consider two *potential outcomes* (health status)

$$\begin{array}{ll} Y_{1i} & \text{outcome with treatment} \\ Y_{0i} & \text{outcome without treatment} \end{array}$$

The gain from treatment or *causal effect* for unit i is

$$Y_{i1} - Y_{i0}$$

Problem: For each i, only one of Y_{i1} or Y_{i0} is observed.

Observed outcome

We observe

$$Y_i \;=\; \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

In the population distribution of Y_{1i} and Y_{0i}, we can compare the average health of *treated* and *non-treated*

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{\text{observed difference}} \;=$$

$$\underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{average treatment effect}} \;+\; \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{selection bias}}$$

Observed outcome

Recall that the observed difference is $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$
Which can be rewritten as:

$$E[\underbrace{Y_{0i} + (Y_{1i} - Y_{0i})D_i}_{Y_i}|D_i = 1] - E[\underbrace{Y_{0i} + (Y_{1i} - Y_{0i})D_i}_{Y_i}|D_i = 0]$$

From the properties of the conditional expectation we can rearrange the above equation as:

$$E[Y_{0i}|D_i = 1] + E[(Y_{1i} - Y_{0i})D_i|D_i = 1] - E[Y_{0i}|D_i = 0] - E[(Y_{1i} - Y_{0i})D_i|D_i = 0]$$

Which can be rewritten as:

$$E[Y_{0i}|D_i = 1] + E[(Y_{1i} - Y_{0i})|D_i = 1] - E[Y_{0i}|D_i = 0]$$

And is equivalent to:

$$E[Y_{0i}|D_i = 1] + E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

Rearranging we get:

$$\underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{average treatment effect}} \quad + \quad \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{selection bias}}$$

Random assignment as a solution

Random assignment makes D_i independent of the potential outcome. If D_i is independent of Y_i then $E[Y_i|D_i] = E[Y_i|D_i = 1] = E[Y_i|D_i = 0] = E[Y_i]$

$$\begin{aligned}
E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\
&= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \\
&= E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}]
\end{aligned}$$

The observed difference in mean outcomes equals the *average treatment effect*. Examples:

- health treatments
- government sponsored training programs
- education production: effect of class size, teacher quality, etc. on student achievement

STAR Experiment

Large randomized experiment involving 11,600 children in 1985-86, who were followed over 4 years

Treatment:

- small classes: 13-17 kids
- regular classes: 22-25 kids
- regular classes with teacher aid

Successful randomization: Subjects' socioeconomic background characteristics balanced across treatment groups

Table 2.2.1: Comparison of treatment and control characteristics in the Tennessee STAR experiment

		Students who entered STAR in kindergarten			
	Variable	Small	Regular	Regular/Aide	Joint *P*-value
1.	Free lunch	.47	.48	.50	.09
2.	White/Asian	.68	.67	.66	.26
3.	Age in 1985	5.44	5.43	5.42	.32
4.	Attrition rate	.49	.52	.53	.02
5.	Class size in kindergarten	15.10	22.40	22.80	.00
6.	Percentile score in kindergarten	54.70	48.90	50.00	.00

Table 2.2.2: Experimental estimates of the effect of class-size assignment on test scores

Explanatory variable	(1)	(2)	(3)	(4)
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes)	–	–	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	–	–	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	–	–	-13.15	-13.07
			(.77)	(.77)
White teacher	–	–	–	-.57
				(2.10)
Teacher experience	–	–	–	.26
				(.10)
Master's degree	–	–	–	-0.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Note: Adapted from Krueger (1999), Table 5. The dependent variable is the Stanford Achievement Test percentile score. Robust standard errors that allow for correlated residuals within classes are shown in parentheses. The sample size is 5681.

Regression Analysis of Experiments

Assume $Y_{i1} - Y_{i0} = \rho$ *constant* treatment effect

$$Y_i = \alpha + \rho D_i + \eta_i$$

$$\begin{aligned} E[Y_i|D_i=1] &= \alpha + \rho + E[\eta_i|D_i=1] \\ E[Y_i|D_i=0] &= \alpha + E[\eta_i|D_i=0] \\ E[Y_i|D_i=1] - E[Y_i|D_i=0] &= \rho + \underbrace{E[\eta_i|D_i=1] - E[\eta_i|D_i=0]}_{\text{selection bias}} \end{aligned}$$

Selection bias amounts to correlation between regression error η_i and D_i.

Regression Analysis of Experiments

We know about the selection bias

$$E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0] = E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

If D_i is randomly assigned, the selection bias is equal to zero. Thus estimating the regression model results in the *causal effect* ρ.

Regression model with covariates

$$Y_i = \alpha + \rho D_i + \beta X_i + \eta_i$$

If X_i uncorrelated with D_i, including them will not affect estimate of ρ, but increase precision.

Table 2.2.2: Experimental estimates of the effect of class-size assignment on test scores

Explanatory variable	(1)	(2)	(3)	(4)
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes)	–	–	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	–	–	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	–	–	-13.15	-13.07
			(.77)	(.77)
White teacher	–	–	–	-.57
				(2.10)
Teacher experience	–	–	–	.26
				(.10)
Master's degree	–	–	–	-0.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Note: Adapted from Krueger (1999), Table 5. The dependent variable is the Stanford Achievement Test percentile score. Robust standard errors that allow for correlated residuals within classes are shown in parentheses. The sample size is 5681.

Non-experimental data - Observational data

- Cross sectional data - sample of units taken at a given point in time, random sampling
- Time series data - one variable observed over time, observations in general not independent over time
- Pooled cross sections
- Panel data - a cross section of individual time series

People choose to go to university. Choice depends on other factors affecting wage. Difficult to unbundle effects. Try to find variation of years of schooling which is random. Try to account for other factors: controls.

Statistical Models and Shoe Leather

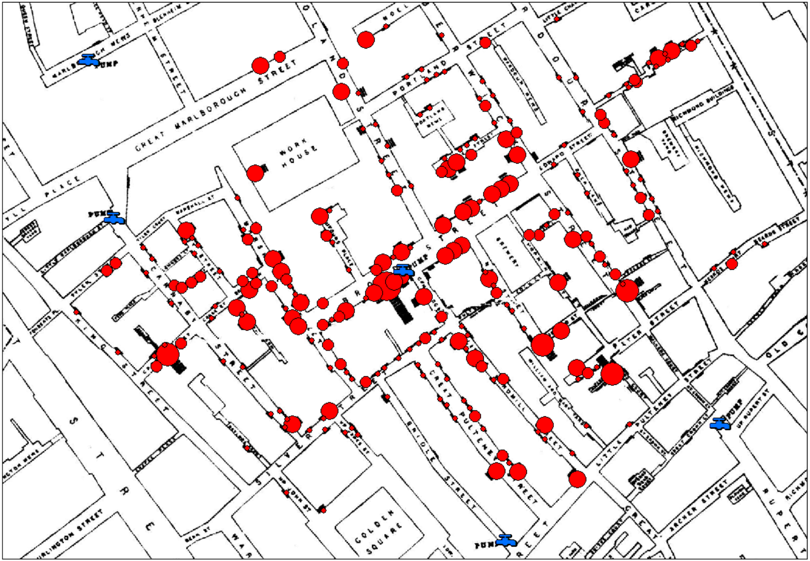
Freedman, David A. (1991) "Statistical Models and Shoe Leather", Sociological Methodology, 21, 291-313

John Snow studies of the cholera epidemics in Europe in the 19th century and proves that cholera is a waterborne infectious disease

- In the 19^{th} century no microbiology, limited microscopes
- Theory: diseases result from "poison in the air" - miasma
- Cholera shows up in Europe in epidemic waves
- Snow studied spatial pattern of epidemics along tracks of human commerce
- Influence of water supply on incidence of cholera?

Detective work, dealing with circumstancial evidence

- Who was the first case in the early London epidemic? Any links to the second case?
- One building is affected, a building next to it is not. Why?
- A brewery in the epicenter is not affected. Why?
- Snow surveyed houses in large parts of London
- Water company
- Cholera victims
- 300,000 households involved


GREAT MARLBOROUGH STREET
WORK HOUSE
PUMP
PUMP
PUMP
PUMP
PUMP
PORTLAND STREET
BROAD STREET
BREWERY
GOLDEN SQUARE
REGENT STREET
BRIDLE STREET
GREAT PULTENEY STREET
PETER STREET
BERWICK STREET
DEAN STREET
RUPERT STREET
ARCHER STREET
LITTLE PULTENEY STREET
RICHMOND BUILDINGS
RICHMOND MEWS
NOEL STREET
POLAND STREET
SILVER STREET
KING STREET
MARSHALL ST
CAMBRIDGE STREET
CROSS STREET
PORTLAND MEWS
HOPKINS STREET
MARLBOROUGH MEWS
BURLINGTON STREET

Is cholera a waterborne or an airborne disease?

London in the 1800's: different water companies serve different areas

- Some companies take water from the Thames polluted by sewage
- 2 companies
  - Southwark & Vauxhall: downstream from sewage discharges
  - Lambeth: intake point upstream
- Both companies served the same parts of London during the 1853-54 cholera epidemic
- Sometimes houses next to each other in the same street were served by the 2 different companies
  - Each company supplies rich and poor, large and small houses, no difference in condition or occupation
- Idea: compare number of cholera victims

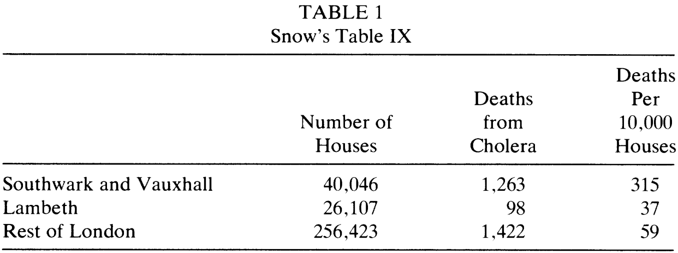


TABLE 1
Snow's Table IX

	Number of Houses	Deaths from Cholera	Deaths Per 10,000 Houses
Southwark and Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

MICROCOSM dedicated to the London Water Companies
Brought forth all monstrous, all prodigious things,
Hydras, and Gorgons, and Chimeras dire. Vide Milton
MONSTER SOUP commonly called THAMES WATER. Being a correct representation of that precious stuff doled out to us !!!