Econometrics 2 – Part 1

Arieda Muço

Central European University

Winter 2023

Contact

- Arieda Muço: MucoA@ceu.edu
- Sergey Lychagin: LychaginS@ceu.edu
- Ramzi Chariag (TA): chariag_ramzi@phd.ceu.edu

Textbooks

- Introductory Econometrics: A Modern Approach by Wooldridge
- Mostly Harmless Econometrics by Angrist and Pischke
- Mastering Metrics by Angrist and Pischke
- Casual Inference: The Mixtape by Cunningham
- Reading list of applied papers

Why Applied Econometrics?

What do Economists really do?

Della Vigna and Card

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

Applied Microeconometrics Papers

30
25
20
15
10
5
0
1970
money
people

35
30
25
20
15
10
5
0
2016
money
people

Applied Microeconometrics

- Investigate micro-econometric methods and research designs
- Discuss applied research papers - learn from good examples
- Perform own empirical exercises based on data (real and simulated)
- Topics: Labor, Population, Health, Public, Development, and Political Economics

TA Sessions

- Six TA sessions
- You are encouraged to work in groups, but every student hands in individual solutions
- Solutions have to be turned in at specified dates (via Moodle)
- Problem sets mostly replicate and discuss estimation results
  - Provided with datasets

Course Goals

Guidance for conducting sensible applied research projects

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

Grading

- **Quizzes in Class** (10% of final grade)
- **Problem Sets** (40% of final grade)
- **Final Exam** (50% of final grade)
  - Cover the material presented in class and seminars
  - Questions that require short answers

Outline: Tentative Schedule

Part 1

- Economic research questions: causality
- The experimental ideal
- Instrumental Variables, Regression Discontinuity

Part 2

- Panel Data, Fixed Effects, Differences-in-Differences
- Matching
- Program Evaluation: Nonparametric Methods

Research Questions

1. What is the relationship of interest
   - Descriptive research
   - 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. What is the identification strategy?
4. What is the mode of statistical inference?

Experimental data

Hard and costly to obtain

- Controlled laboratory experiment
- Randomized controlled trials

Give one extra year of education to a randomly selected group and compare means in the treatment and control group. Due to randomization the extra year is unrelated to parents education, neighbourhood income, etc

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 time series for each cross-sectional unit

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

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

Method of Shoe and Leather

- Snow surveyed houses in large parts of London
- Water company
- Cholera victims
- 300,000 households involved
- Reward: clear result

GREAT MARLBOROUGH STREET
WORK HOUSE
PORTLAND STREET
PUMP
PUMP
PUMP
PUMP
GOLDEN SQUARE
GREAT PULTENEY STREET
BRIDLE STREET
PETER STREET
ARCHER STREET
BRUNSWICK STREET
STREET
ST JOHN ST
BURLINGTON STREET

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

The Experimental Ideal

Social experiment is the most influential research design. Why?

Solves Selection Problem

Cameron et al (2016): RCT in Development Economics

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)
- Selection: who is hospitalized?

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

Potential Outcome Model

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

$$\begin{array}{ll} Y_{i1} & \text{outcome with treatment} \\ Y_{i0} & \text{outcome without treatment} \end{array}$$

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

$$Y_i = \begin{cases} Y_{i1} & \text{if } D_i = 1 \\ Y_{i0} & \text{if } D_i = 0 \end{cases} = Y_{i0} + (Y_{i1} - Y_{i0})D_i$$

In the population distribution of Y_{i1} and Y_{i1}, we can compare the average health of *treated* and *non-treated*

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

$$\underbrace{E[Y_{i1}|D_i = 1] - E[Y_{i0}|D_i = 1]}_{\text{average treatment effect}} + \underbrace{E[Y_{i0}|D_i = 1] - E[Y_{i0}|D_i = 0]}_{\text{selection bias}}$$

Observed outcome

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

$$E[\underbrace{Y_{i0} + (Y_{i1} - Y_{i0})D_i}_{Y_i}|D_i = 1] - E[\underbrace{Y_{i0} + (Y_{i1} - Y_{i0})D_i}_{Y_i}|D_i = 0]$$

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

$$E[Y_{i0}|D_i = 1] + E[(Y_{i1} - Y_{i0})D_i|D_i = 1] - E[Y_{i0}|D_i = 0] - E[(Y_{i1} - Y_{i0})D_i|D_i = 0]$$

Which can be rewritten as:

$$E[Y_{i0}|D_i = 1] + E[(Y_{i1} - Y_{i0})|D_i = 1] - E[Y_{i0}|D_i = 0]$$

And is equivalent to:

$$E[Y_{i0}|D_i = 1] + E[Y_{i1}|D_i = 1] - E[Y_{i0}|D_i = 1] - E[Y_{i0}|D_i = 0]$$

Rearranging we get:

$$\underbrace{E[Y_{i1}|D_i = 1] - E[Y_{i0}|D_i = 1]}_{\text{average treatment effect}} + \underbrace{E[Y_{i0}|D_i = 1] - E[Y_{i0}|D_i = 0]}_{\text{selection bias}}$$

Example-Mastering Metrics

	Khuzdar Khalat	Maria Moreño
Potential outcome without insurance: Y_{0i}	$\boxed{3}$	5
Potential outcome with insurance: Y_{1i}	4	$\boxed{5}$
Treatment (insurance status chosen): D_i	1	0
Actual health outcome: Y_i	4	5
Treatment effect: $Y_{1i} - Y_{0i}$	1	0

Random assignment as a solution

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

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{i1}|D_i = 1] - E[Y_{i0}|D_i = 0] \\ &= E[Y_{i1}|D_i = 1] - E[Y_{i0}|D_i = 1] \\ &= E[Y_{i1} - Y_{i0}|D_i = 1] = E[Y_{i1} - Y_{i0}] \end{aligned}$$

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

Women as Policy Makers

Reservation legislation, of council seats, affects the types of public goods provided.

Treatment:

- In India, 1/3 of council seats are reserved to women
- Gram Panchayat (GP), is composed of several villages. Randomization on GP numbers predating the reform.

Make sure randomization is successful by comparing pre-determined village characteristics. We are looking for balance across treatment groups.

VILLAGE CHARACTERISTICS IN RESERVED AND UNSERVED GP, 1991 CENSUS

	West Bengal			Rajasthan		
Dependent Variables	Mean, Reserved GP (1)	Mean, Unreserved GP (2)	Difference (3)	Mean, Reserved GP (4)	Mean, Unreserved GP (5)	Difference (6)
Total Population	974	1022	−49	1249	1564	−315
	(60)	(46)	(75)	(123)	(157)	(212)
Female Literacy Rate	.35	.34	.01	.05	.05	.00
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
Male Literacy Rate	.57	.58	−.01	.28	.26	.03
	(.01)	(.01)	(.01)	(.02)	(.02)	(.03)
% Cultivated Land that Is Irrigated	.45	.43	.02	.05	.07	−.02
	(.03)	(.02)	(.04)	(.01)	(.01)	(.02)
Dirt Road	.92	.91	.01	.40	.52	−.11
	(.02)	(.01)	(.02)	(.08)	(.07)	(.10)
Metal Road	.18	.15	.03	.31	.34	−.04
	(.03)	(.02)	(.03)	(.07)	(.06)	(.10)
Bus Stop or Train Station	.31	.26	.05	.40	.43	−.03
	(.04)	(.02)	(.04)	(.08)	(.07)	(.10)
Number of Public Health Facilities	.06	.08	−.02	.29	.19	.10
	(.01)	(.01)	(.02)	(.08)	(.06)	(.10)
Tube Well Is Available	.05	.07	−.02	.02	.03	−.01
	(.03)	(.02)	(.07)	(.02)	(.02)	(.03)
Handpump Is Available	.84	.88	−.04	.90	.97	−.06
	(.04)	(.03)	(.05)	(.05)	(.02)	(.05)
Wells	.44	.47	−.02	.93	.91	.01
	(.07)	(.04)	(.08)	(.04)	(.04)	(.06)
Tap Water	.05	.03	.01	.12	.09	.03
	(.03)	(.02)	(.03)	(.05)	(.04)	(.06)
Number of Primary Schools	.95	.91	.04	.93	1.16	−.23
	(.07)	(.03)	(.08)	(.09)	(.10)	(.15)
Number of Middle Schools	.05	.05	.00	.43	.33	.10
	(.01)	(.01)	(.01)	(.08)	(.07)	(.10)
Number of High Schools	.09	.10	−.01	.14	.07	.07
	(.01)	(.01)	(.02)	(.06)	(.04)	(.07)
F-Statistics: Difference Jointly Significant			.93			1.54
(p-value)			(.53)			(.11)

EFFECT OF WOMEN'S RESERVATION ON WOMEN'S POLITICAL PARTICIPATION

Dependent Variables	Mean, Reserved GP (1)	Mean, Unreserved GP (2)	Difference (3)
West Bengal			
Fraction of Women Among Participants	9.80	6.88	2.92
in the Gram Samsad (in percentage)	(1.33)	(.79)	(1.44)
Have Women Filed a Complaint to	.20	.11	.09
the GP in the Last 6 Months	(.04)	(.03)	(.05)
Have Men Filed a Complaint to the GP	.94	1.00	.06
in the Last 6 Months	(.06)		(.06)
Observations	54	107	
Rajasthan			
Fraction of Women Among Participants	20.41	24.49	−4.08
in the Gram Samsad (in percentage)	(2.42)	(3.05)	(4.03)
Have Women Filed a Complaint to	.64	.62	.02
the GP in the Last 6 Months	(.07)	(.06)	(.10)
Have Men Filed a Complaint to the GP	.95	.88	.073
in the Last 6 Months	(.03)	(.04)	(.058)
Observations	40	60	

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

$$E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0] = E[Y_{i0}|D_i = 1] - E[Y_{i0}|D_i = 0]$$

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.

STAR Experiment

Large randomized experiment involving 11,600 children in 1985-86, who were followed over 4 years. Students were assigned into one of the following treatments.

Treatment:

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

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.