

Econometrics 1

Introduction

Christelle Dumas

U. of Cergy-Pontoise

October 9, 2009

- Econometrics 1: C. Dumas - Linear econometrics model, instrumental variables - → 1 hour test during the last lecture (15/100)
- Econometrics 2: M. Gurgand - Evaluation, advanced IV -
 - 2 hours test on Econometrics 1 & 2 (30/100)
- Econometrics 3: L. Behaghel - Selection problems, qualitative models, panel data - → 1 hour test during the last lecture (15/100)
- Econometrics 4: M. Bensalem - Time series -
 - 2 hours test on Econometrics 3 & 4 (30/100)
- Classes: M. Mahjoub (10/100)

- Wooldridge, "Introductory Econometrics: A Modern Approach", Cincinnati, 2000.
- Wooldridge, "Econometric Analysis of Cross Section and Panel Data", MIT Press, 2002.
- Angrist & Pischke, "Mostly harmless econometrics: an empiricist's companion", Princeton University Press, 2009.
- Cameron & Trivedi, "Microeconometrics, methods and applications", Cambridge University Press, 2005.
- Behaghel, "Lire l'économétrie", Repères.
- papers → download from the website

Some tutorials:

- http://rlab.lse.ac.uk/it/it_docs/Introduction_to_stata.pdf
- http://www.ats.ucla.edu/stat/stata/notes_old/default.htm
- <http://stataproject.blogspot.com/2007/12/project-1-getting-data.html>
- http://www.masterape.ens.fr/wdocument/master/stata_bozio.pdf
(in French)

4 questions to be asked (and answered)

- relationship of interest
- ideal experiment
- the identification strategy
- the mode of inference

What is the causal relationship of interest?

- Cause and effects
- making predictions about the consequences of changing circumstances/policies
- example: returns to schooling;
 - causal impact of schooling on wages useful for predicting earning consequences of a change in compulsory attendance laws, or changing costs of attending college...
- unit of observation: individual human being, firms, countries
 - Acemoglu, Johnson, Robinson (2001): effect of colonial institutions on economic growth (more democracy implies more economic growth?)

What experiment could ideally be used to capture the causal effect of interest?

- returns to schooling → imagine offering potential dropouts a reward for finishing school and then studying the consequences
- political institutions → randomly assign different government structures to former colonies

Why do we want to run experiments? we'll see that in a minute

Fundamentally unidentified questions?

- effect of race or gender → change the chromosomes? no → make sb believe you're black... asian... woman... fake applications
 - effect of start age on 1st grade test scores
 - randomly select some kids to start kindergarten at age 6 instead of 5 - do they learn more?
 - but group that started school at age 7 is older (maturation effect)
 - test those who started at age 6 in second grade and those who started at age 7 in 1st grade so everybody is tested at 7.
 - but 1st group has spent more time in school.
- cannot disentangle start-age effect from maturation and time-in-school effects as long as kids are enrolled → adult's outcomes

What is your identification strategy?

Identification strategy = the manner in which a researcher uses observational data (not generated by a randomized trial) to approximate a real experiment.

Example: Angrist & Krueger (1991) for returns to schooling

- interaction between compulsory attendance laws and students' season of birth to estimate the effects of finishing high school on wages

What is your mode of statistical inference?

Describe:

- population
- sample
- assumptions made for standard errors

The selection problem

Example: "Do hospitals make people healthier?"

- Natural approach: compare the health status of those who have been to the hospital to the health of those who have not.
- NHIS: "During the past 12 months, was the respondent a patient in a hospital overnight?", "Would you say your health in general is excellent (1), very good, good, fair, poor (5)?"

Group	Sample size	Mean health status	Std. error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003
Difference		0.71	0.012

Formalization (1)

- Going to the hospital makes people sicker?
- People who seek medical care are probably less healthy to begin with
- hospital treatment $D_i = \{0, 1\}$, binary random variable
- a measure of health status Y_i
- is Y_i affected by hospital care?

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

$$= Y_{0i} + (Y_{1i} - Y_{0i})D_i \quad (2)$$

Formalization (2)

- $Y_{1i} - Y_{0i}$ is the causal effect of hospitalization for an individual (treatment effect can be different from 1 person to the other)
- but we never observe Y_{1i} AND Y_{0i}
- naive comparison:

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \\ \text{observed difference in average health} & \quad \text{average treatment effect on the treated} \\ & + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ & \quad \text{selection bias} \end{aligned}$$

- $E[Y_{1i} - Y_{0i}|D_i = 1]$ is the average causal effect of hospitalization on those who were hospitalized
- selection bias: difference in average Y_{0i} between those who were and were not hospitalized

Random assignment solves the selection problem

- Random assignment of D_i solves the selection problem because random assignment makes D_i independent of potential outcomes

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

- random assignments of D_i eliminates the selection bias
- experiments often reveal things that are not what they seem on the basis of naive comparisons alone
 - on-the-job training for groups of disadvantaged workers in order to increase employment and earnings
- STAR experiment

The STAR experiment (1)

Estimate the effects of smaller classes in primary school
(cost/benefit analysis)

- non-experimental data suggest little or no link between class size and student learning
 - weaker students often deliberately grouped into smaller classes
- experiment tries to compare apples to apples
- study implemented for a cohort of kindergartners in 85-86, lasted 4 years
 - average class size in regular Tennessee classes: 22.3
 - 3 treatments:
 - small classes with 13-17 children
 - regular classes with 22-25 children and a part-time teacher's aide
 - regular classes with a full time teacher's aide

The STAR experiment (2)

1st question: has randomization successfully balanced children's characteristics across the different treatment groups?

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

Variable	Students who entered STAR in kindergarten			Joint <i>P</i> -value
	Small	Regular	Regular/Aide	
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

Notes: Adapted from Krueger (1999), Table 1. The table shows means of variables by treatment status. The *P*-value in the last column is for the *F*-test of equality of variable means across all three groups. All variables except attrition are for the first year a student is observed. The free lunch variable is the fraction receiving a free lunch. The percentile score is the average percentile score on three Stanford Achievement Tests. The attrition rate is the proportion lost to follow up before completing third grade.

Results to come...

Regression analysis of experiments (1)

If treatment effect is the same for everybody, $Y_{1i} - Y_{0i} = \rho$,

$$\begin{aligned} Y_i &= E(Y_{0i}) + \rho D_i + (Y_{0i} - E(Y_{0i})) \\ &= \alpha + \rho D_i + \eta_i \end{aligned}$$

Conditional expectation of the equation:

$$\begin{aligned} E[Y_i | D_i = 1] &= \alpha + \rho + E[\eta_i | D_i = 1] \\ E[Y_i | D_i = 0] &= \alpha + E[\eta_i | D_i = 0] \end{aligned}$$

so that:

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

Regression analysis of experiments (2)

- Selection bias amounts to correlation between the regression error term η_i and the regressor D_i
- this correlation reflects the difference in (pre-treatment) potential outcomes between those who get treated and the others since

$$E[\eta_i | D_i = 1] - E[\eta_i | D_i = 0] = E[Y_i | D_i = 1] - E[Y_i | D_i = 0]$$

- in an experiment where D_i is randomly assigned
 - selection term disappears
 - regression of Y_i on D_i estimates the causal effect of interest ρ

STAR experiment (3)

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 (2.19)	5.37 (1.26)	5.36 (1.21)	5.37 (1.19)
Regular/side class	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)
White/Asian (1 = yes)	-	-	8.35 (1.35)	8.44 (1.36)
Girl (1 = yes)	-	-	4.48 (.63)	4.39 (.63)
Free lunch (1 = yes)	-	-	-13.15 (.77)	-13.07 (.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	.26	.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.

$$Y_i = \alpha + \rho D_i + X_i' \gamma + \eta_i$$

Introduction of control variables X_i important:

- to increase explained share of Y 's variance and get more precise estimates
- to control for systematic differences between individuals in the different class types

- Regressions are well-suited to the analysis of experimental data
- in some cases, can also be used to approximate experiments in the absence of random assignment