2019-2020

EMPIRICAL METHODS FOR ECONOMICS AND POLICY EVALUATION

(MICROECONOMETRICS & CAUSAL INFERENCE)

ANDREA IERARDI NOTES

|andrea.ierardi@studenti.unimi.it | andreierardi@gmail.com

SOMMARIO

ALL READINGS

Chapter 1

Compulsory reading:

- Angrist and Pischke [2015]: chapters 2.1-2.2 and 6.1
- Dale and Krueger [2002]

Suggested readings:

- Chetty et al. [2011]
- Hoekstra [2009]
- Krueger [1999]

Chapter 2

Compulsory reading:

- lecture slides

Suggested reading:

- chapter 3 -Gertler et al. (2011) Impact Evaluation in Practice, World
- Bank (available on-line and on Ariel)

Chapter 3

Compulsory readings:

- my lecture slides
- Angrist and Pischke [2015]: chapter 1

Chapter 4

Compulsory readings:

- my lecture slides
- Angrist and Pischke [2015]: chapter 4

Suggested reading:

- Pinotti [2017]

Chapter 5

Compulsory readings

- Angrist and Pischke [2015]: chapter 5

Additional readings:

- Card and Krueger [1994]
- Krueger and Card [2000]

Chapter 5 – Examples

Compulsory readings:

- my lecture slides
- Duflo [2001]
- Draca et al. [2011]

Chapter 6

None

Chapter 7

Compulsory readings:

- my lecture slides
- chapter 3 -Angrist and Pischke [2015]

INTRODUCTION

Table of contents

Practical info

Two theory lectures per week (6 CFU)

One practical lecture (computer–based) per week (compulsory for 9 CFU)

My office: Room 11, 2nd floor, DEMM

Office hours: Wednesday 10am–12:30 pm

Aim of the course

- Analyse main challenges faced by quantitative social scientists in answering empirical questions using microdata.
- Main emphasis: learning how to establish causal relationships between different variables and how to use this evidence to inform policy makers' decisions.
- We will learn about challenges in evaluating public policies and possible solutions

EXAMPLE: SHOULD YOU GO TO MUSEUMS TO LIVE LONGER?


The New York Times

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

From the NYT 22/12/2019: Researchers in London who followed thousands of people 50 and older over a 14-year period discovered that those who went to a museum or attended a concert just once or twice a year were 14 percent less likely to die during that period than those who didn't.

A credible finding?

Stata

- The practical part, taught by Dr Anna Rosso, is compulsory for EPS
- students (9 CFU) only
- You will learn how to handle real–world data, how to use econometric software (Stata) to generate estimates and how to link econometric theory with data work.
- I encourage everyone to familarise with statistical software
- In this course we use Stata, which is available to all Unimi students and staff via a campus licence.
- Stata is available in all computer rooms and labs on campus
- You can download your Unimi licensed copy by sending an email to licenze.campus@unimi.it. PLEASE DO IT!

Syllabus

1. Introduction: the “credibility revolution” in empirical economics
2. Example: Evaluating Education Policies
3. Estimating Causal Policy Effects
4. Randomized Experiments
5. Regression Discontinuity Design
6. Difference in Differences
7. Panel Data
8. Instrumental Variables

Main references:

1. My lecture slides
2. Jousha D. Angrist and Jorn-Steffen Pischke (2015) Mastering Metrics, Princeton
3. Specific papers indicated in each lecture

Additional references:

1. Jousha D. Angrist and Jorn-Steffen Pischke (2008) Mostly Harmless Econometrics: An Empiricist’s Companion. Princeton
2. Wooldridge (20XX) Introductory Econometrics: A Modern Approach, South-Western College Pub.
3. Wooldridge (20XX) Econometric Analysis of Cross Section and Panel Data, MIT Press.
4. Cameron and Trivedi (2010) Microeconometrics Using Stata, Revised Edition, Stata Press
5. Gertler et al. (2011) Impact Evaluation in Practice, World Bank.
6. Khandker et al. (2010) Handbook on Impact Evaluation Quantitative

Exam dates:

- 30 March 2020
- 6 May 2020
- 8 July 2020
- 14 September 2020
- Two more dates to be scheduled between October and December 2020

LECT. 1: EVALUATING EDUCATION POLICIES

Table of contents

Introduction

- Education is considered to be a major determinant of economic growth in both developed and developing countries
- An important question for policy-makers interested in educational policies regards the returns to education.
- In order to decide whether, for instance, compulsory education should be increased by one year, whether a program of college subsidies should be introduced, etc. we would like to know what is the return - in terms of future earnings - of an additional year of education, of going to college, etc.
- How can we estimate the returns to education?

In order to conceive a credible evaluation of a policy, we need to be able to answer four questions:

1) what is the causal relationship of interest?
2) which experiment could ideally be used to capture the causal effect of interest?
3) what is our research design (i.e. identification strategy)?
4) what is our econometric methodology?

- Suppose we are advising the prime minister of country A who considers providing a program of free college education for all
- Such an intervention is extremely costly. But what is the return?
- As a first approximation, we could try and estimate the wage return: what is the increase in wages caused by having college education?
- Our evaluation should then look at all other possible outcomes: employment, health, political participation, crime, innovation, etc.
- But, let's start from the "easy" part: estimating the wage return of education

Estimating returns to schooling

- Suppose we have data on a representative sample of workers in country A: for each individual we observe employment status, wage and education level (plus demographic characteristics, working experience, etc.).
- We create a dummy variable C which is equal to 1 if the individual has a university degree
- We run an OLS regression of wage (Wi) on the dummy Ci (controlling for a vector Xi of individual controls: age, gender, years of experience, nationality, etc.):

$$W_i = \alpha + \beta C_i + X_i\gamma + \epsilon_i \qquad (1)$$

- We will obtain (almost for sure) a positive coefficient: $\widehat{\beta_{OLS}} > 0$

- Indeed, this type of regressions - often with years of schooling S replacing the college dummy C - have been estimated for decades in a very large set of countries
- These equations are called Mincer equations, after Jacob Mincer (1922-2006), one of the founding fathers of modern labor economics
- They (almost) always deliver a positive, significant and relatively large coefficient: education is arguably the best investment one could possibly make
- Montenegro and Patrinos [2014] have estimated the returns to years of schooling using the same specification, estimation procedure, and similar data for 139 economies and 819 harmonized household surveys

FIGURE: AVERAGES RETURNS TO ONE EXTRA YEAR OF SCHOOLING

(Source: Montenegro and Patrinos [2014])

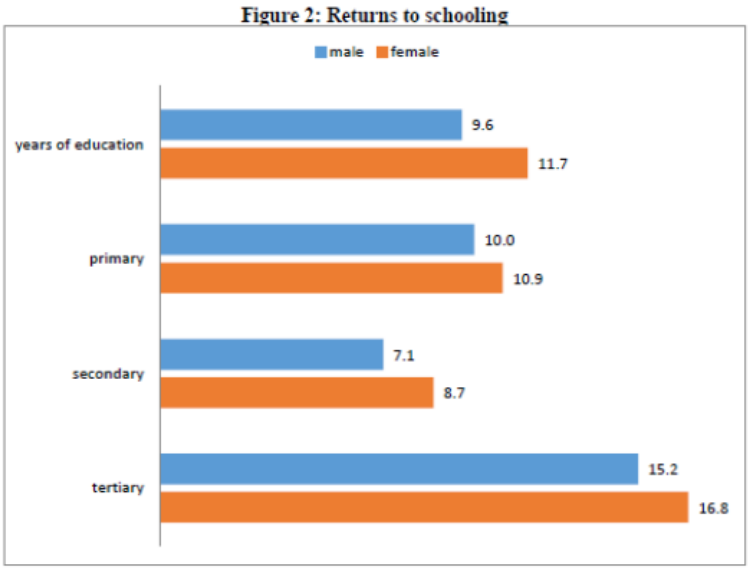
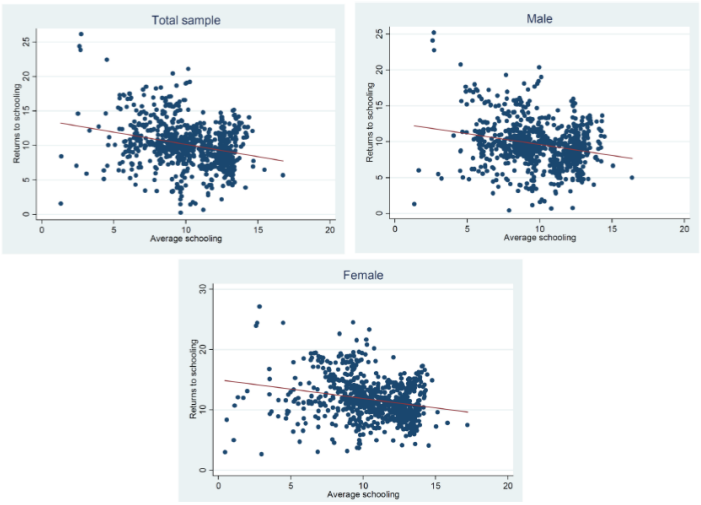


FIGURE: RETURNS TO SCHOOLING AND AVERAGE YEARS OF SCHOOLING

(Source: Montenegro and Patrinos [2014])



- Are all these estimates identifying causal parameters?
- ”In empirical work, the causal relationship between schooling and earnings tells us what people would earn, on average, if we could either change their schooling in a perfectly controlled environment, or change their schooling randomly so that those with different levels of schooling would be otherwise comparable” (Angrist and Pischke [2008]; p.53).
- Crucial questions: is that estimated coefficient identifying the causal impact of education on wage? is there any causal effect at all? are we just capturing the fact that people who are more motivated chose to go to college and higher motivation leads to higher wages? Which fraction of the estimated effect is truly causal?

- If we stop our analysis at the positive coefficient, we have established an empirical fact but we just have some descriptive evidence.
- If we manage to answer all the other questions regarding causality, instead, we have established and quantified a causal relationship.
- Descriptive analyses allow to establish facts (e.g. people more educated earn more). This is an important contribution, but, in order to be able to produce policy implications, we want to understand what explains and causes those facts.

- Suppose the people who chose to obtain college education are different with respect to those who did not (and this is very likely to be the case)
- Then, extending college education to individuals who otherwise would have not chosen to get it will probably imply a different return for them than for those who would have chosen to go to college in any case.
- Policy interventions cannot - and should not - be based on mere descriptive evidence.
- Describing reality is the first step, but understanding causal relationships is the real challenge.

- Why should the estimated coefficient be different from the causal one? What is the problem here?
- We have a problem of “selection into treatment”: the treatment is going to college and people can decide whether to do that or not

- Comparing the outcomes of those who chose to get university education and of those who did not may bear little information regarding the actual effect of university education: we are comparing different individuals
- The ideal experiment would be to randomly assign individuals to college... but it is hard to think that such an experiment could ever be implemented...

- What kind of selection do we expect to have in this case?
- Two examples (there may be more):
  - Smarter people - if not credit constrained - are more likely to go to college (their cost of making the investment in human capital is lower because studying is easier for them); and, if labor market rewards intelligence, they will also earn higher wages
  - Kids from more educated/wealthier families are more likely to go to college; but their family networks may also help them to find better jobs (i.e. earn higher wages)
- This implies that at least part - and possibly all - of the positive effect we have estimated, may be due to individuals with university education being smarter, or coming from more advantaged family backgrounds.

- We run the following regression:

$$W_i = \alpha + \beta C_i + X_i\gamma + \epsilon_i \qquad (2)$$

- Ignoring the vector of individual observable characteristics Xi , the OLS estimate of β is:

$$\widehat{\beta}_{OLS} = \frac{cov(W, C)}{var(C)} = \beta + \frac{cov(\epsilon, C)}{var(C)} \qquad (3)$$

- using OLS, we obtain a consistent estimate of β iff (=if and only if) cov(ε,C) = 0.
- That is, iff the variable C is “exogenous” in our regression.

- If we think that individual’s ability (being smart) and family background may positively influence both the likelihood of going to college and the wage, we have omitted two important variables from the analysis.

The ”true” wage regression should be:

$$W_i = \alpha + \beta C_i + X_i\gamma + \underbrace{\lambda FB_i + \mu A_i + e_i}_{\varepsilon_i} \qquad (4)$$

where FB_i is family background (e.g. parents’ income or education level), A_i is individual ability and e_i is an error term.

- Here we expect both λ and μ to be positive:
  - λ > 0: workers with “better” family background earn more
  - μ > 0: workers with higher “ability” earn more

- The decision of investing in university education, instead, can be written as:

$$C_i = a + J_i b + cFB_i + dA_i + u_i \qquad (5)$$

where J_i are individual controls: some of them, like gender, may be the same we have in Xi , while others, like geographical proximity to a university, may only matter for the schooling decision.

- Here we expect the coefficients c and d to be both positive.

Omitting two relevant variables which are likely to be correlated with the regressor C will cause an **omitted variable bias.** We can sign the OLS bias:

$$\widehat{\beta}_{OLS} = \beta + \frac{cov(\epsilon, C)}{var(C)} = \beta + \lambda \frac{cov(FB, C)}{var(C)} + \mu \frac{cov(A, C)}{var(C)} + \frac{cov(e, C)}{var(C)}$$

We assume cov(e,C) = 0, we have var(C) > 0 (variances are always positive) and we expect: λ > 0, cov(FB,C) > 0, μ > 0 and cov(A,C) > 0.

This implies that our OLS estimate is upward biased (i.e. bias> 0):

$$bias = \lambda \frac{cov(FB, C)}{var(C)} + \mu \frac{cov(A, C)}{var(C)} > 0$$

- Note that the true causal coefficient could even be zero (as in a signaling model; see Spence [1973]) or negative, but we would still find a positive coefficient if the positive bias is large enough.
- ”Despite the overwhelming evidence of a positive correlation between education and labor market status, social scientists have been cautious to draw strong inferences about the causal effect of schooling. In the absence of experimental evidence, it is very difficult to know whether the higher earnings observed for better- educated workers are caused by their higher education, or whether individuals with greater earning capacity have chosen to acquire more schooling.” (Card [1999]).

The most obvious solution to an omitted variable bias is including the ”omitted variable” in the regression as an additional control.

This can be done if two conditions are satisfied:

1. the variable is observable: i.e. it can be measured;
2. the variable is observed: i.e. it has been measured in the survey data used and/or it can be recovered from some other data sources.

In our schooling example, one can control for family background by including parents’ education, occupation and earnings in the regression (and any other variable which measures the social and educational status of the family).

Addressing the ability bias is less straightforward: ability is unobservable.

One can then try to use some observable proxies for ability (e.g. I.Q. scores) or try to eliminate the ability by using repeated observations from the same individual (but schooling is time invariant too...) or observations from individuals with (arguably) same ability (e.g. twin studies).

Throughout the course, we will discuss some of the approaches undertaken by the literature on returns to schooling.

Estimating the effects of school quality

Beyond quantity, quality of schooling may matter

In countries where a sufficient amount of education is guaranteed to everyone through compulsory schooling legislation, the focus of policy-makers may shift on the quality of the education delivered

A relatively smaller literature has addressed the question: does the quality of the school that students attend influence their subsequent earnings?

Schooling is not an homogenous treatment: there are “good” and “bad” schools and spending N years in the former rather than in the latter ones may yield substantially higher returns

But, what determines “school quality”? infrastructures, teachers’ CV, class size, quality of students, location, etc.?

Suppose we want to estimate the impact of school quality on educational achievements of students

Ideal experiment: students are assigned to school whose quality is randomly assigned

Reality: parents/students choose the schools. Endogeneity:

- Parents more concerned about kids’ education will choose better schools; more motivated students will choose better universities
- And kids whose parents are more concerned about their education will (possibly) perform better at school while more motivated students will perform better at university

We can expect to find a positive correlation between school quality and student’s performance. Is it causal?

Or, is it just positive selection?

A randomized experiment: Tennessee Project STAR

Key question in research on the “education production function”: which inputs produce the most learning given their costs?

One of the most expensive input is class size

What is the payoff in terms of higher student achievement of a lower student/teacher ratio?

Two sources of selection:

1. parents from a wealthier background / more concerned about children’s education may choose schools with smaller classes (positive selection)
2. principals and teachers may assign weaker students to smaller classes (negative selection)

Tennessee Project STAR (Student/Teacher Achievement Ratio): designed to estimate the effects of smaller classes in primary school

Experiment characteristics:

- cost: $ 12 million;
- treated group: one cohort of kids in kindergartner in 1985-86; 11,600 children
- duration: four years;

Average class size in Tennesse in early ’80s: 22.3

Three treatments within each school:

1. small classes with 13-17 children
2. regular classes with 22-25 children and one part-time teacher assistant (the usual arrangement)
3. regular classes with 22-25 children and one full-time teacher assistant

Schools with at least three classes in each grade could choose to participate in the experiment

Each school was required to have at least one of each class-size type, and random assignment took place within schools.

Hence, parents could still (endogenously) choose the school, but conditional on that choice, the assignment of their children to small or large classes was decided at random

Within schools, random assignment guarantees that the average unobservables of pupils in small and large classes are not different

Comparing their mean outcomes, therefore, identifies the causal effect of interest

To measure outcomes, students were given a battery of standardized tests at the end of each school year.

Short-run effects

Krueger [1999] analyzes the STAR experiment

He finds a small but positive effect of class size on student test scores

He starts by running some "balance tests" to check that the randomization was properly done

If children have been randomized in different treatments their average observable and unobservable characteristics should not be significantly different

We cannot test unobservable characteristics, but we can look at observable characteritics

Note that pupils could enter the participating schools at different grades (kindergarten; I, II or III grade)

FIGURE: TREATMENT VS CONTROL GROUPS: BALANCE TESTS; KRUEGER [1999]

TABLE I
COMPARISON OF MEAN CHARACTERISTICS OF TREATMENTS AND CONTROLS: UNADJUSTED DATA

A. Students who entered STAR in kindergarten[b]

Variable	Small	Regular	Regular/Aide	Joint *P*-Value[a]
1. Free lunch[c]	.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[d]	.49	.52	.53	.02
5. Class size in kindergarten	15.1	22.4	22.8	.00
6. Percentile score in kindergarten	54.7	49.9	50.0	.00

Reassuringly, the table show that the three groups are very similar to each other.

By experiment design, they differ in class size (see row 5). This is obvious.

Their outcomes (see row 6) are also different (which is not obvious): students in small classes have higher grades than those in large classes and those with full time teaching assistant

If, instead, there were significant differences in pre–treatment characteristics, that would suggest something went wrong with the randomization (e.g. "non-compliance": children are randomly assigned to classes but then parents fight hard to get them reallocated to the classes they prefer)

If there are differences in the observable characteristics it is hard to believe that average unobservable characteristics are instead identical

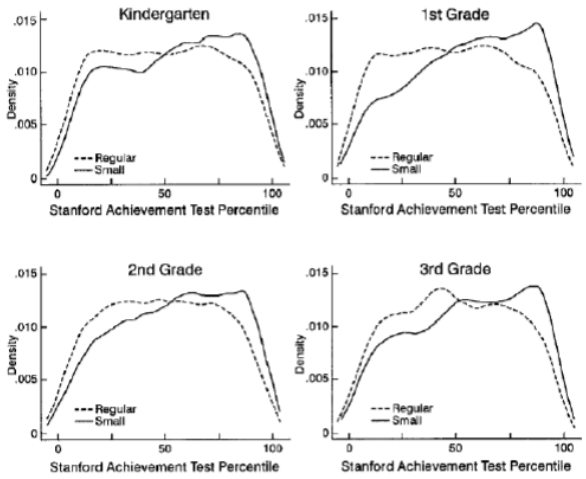


FIGURE I
Distribution of Test Percentile Scores by Class Size and Grade

Figure I displays the density of the average test score distributions for students in small and regular classes at each grade level (K, 1, 2, 3). In all grades, the average student in small classes (continuous line) performed better on this summary test measure than did those in regular or regular/aide classes (dotted line).

Class size matters, and in the expected direction

Krueger [1999] then carries out an econometric analysis to test the robustness of these fidings

Krueger [1999] estimates the following regression:

$$Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 REG/A_{cs} + \beta_3 X_{ics} + \alpha_s + \varepsilon_{ics}$$

where:

- Y_{ics} is the SAT test score of student i in class c at school s;
- $SMALL_{cs}$ is a dummy variable equal to one if the student was assigned to a small class;
- REG/A_{cs} is a dummy variable equal to one if the student was assigned to a regular-size class with an aide;
- X_{ics} is a vector of observed student and teacher covariates
- α_s are school-fixed effects
- ε_{ics} is an error term

The random assignment guarantees that the unobservable characteristics that are concealed in the error term ε_{ics} (e.g. children's ability, parents' commitment and motivation, etc.) are not systematically correlated with the type of classes of the pupils.

That is, pupils in small classes do not have (should not have) in average higher ability or parents more committed to their educational development than those in larger classes.

Hence, we can compare pupils in different class types and obtain causal parameters with a simple OLS regression

The coefficients β1 and β2 identify the effect of the $SMALL_{cs}$ and REG/A_{cs} treatments, respectively, with respect to the control group of pupils assigned to regular classes

Positive and significant coefficients imply that the treatments lead to higher SAT scores

FIGURE: OLS ESTIMATES OF CLASS-SIZE ASSIGNMENT ON SAT SCORES; KRUEGER [1999]

Explanatory variable	OLS: actual class size (1)	(2)	(3)	(4)
			A. Kindergarten	
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	—	—	—	−.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

The Table shows that:

- Students in small classes tend to perform better than those in regular and regular/aide classes (1st row: estimated coefficients are positive and significant)
- Students with a full-time teaching assistant, but in regular sized classes, do not perform better than those in regular classes (2nd row: estimated coefficients are positive but not significant)
- White/Asian students perform better than Black students (row 3); girls perform better than boys (row 4); poor students (free lunch) perform worse (row 5)
- Teachers' characteristics (white and education) have no effect (row 6 and 8); teachers' experience has a small positive effect on pupils' test scores (row 7)

According to these findings, class size seems to matter in the short run: pupils who are allocated to smaller classes tend to have better scores at the end of the year

Does this positive impact have lasting effects in the medium/long run?

If the policy effects vanish after a few years, what is the point of investing public money into that?

This is a crucial question for policy evaluation.

But one needs to wait and observe outcomes of treated pupils as they grow older. Which is not trivial.

Medium-long run effects

Krueger and Whitmore [2001] analyse the effect of the STAR experiment on middle-school test results and on the probability of taking college entrance exams

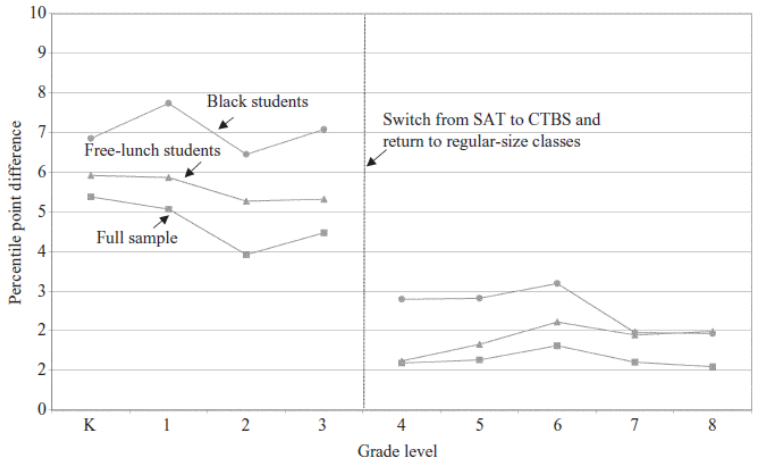
They show that the effect is stronger for minority (black students) and students from a poor background (free-lunch students)

This is a common finding in the education literature: children from disadvantaged family background benefit the most from public educational interventions

Education compensates for what they do not find at home (i.e. educated parents who can assist and enhance their learning process) promoting equality of opportunity

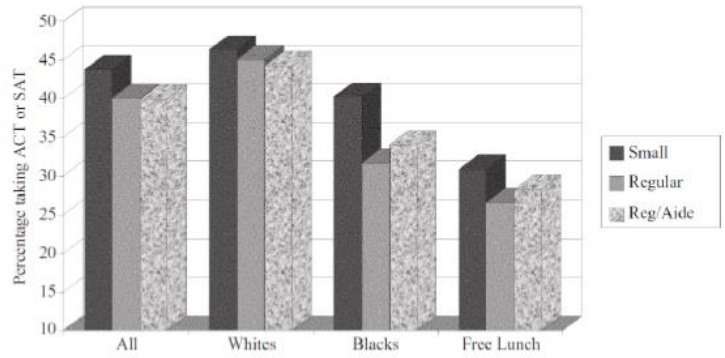
They find that the positive effects on test scores persist after grade 3 (when pupils return to regular classes), although becoming smaller in magnitude

FIGURE: SMALL-CLASS EFFECT ON TEST SCORES; GRADES: K-8 ; KRUEGER AND WHITMORE [2001]



Looking at medium-long term outcomes, they find that attending a small class in the early grade is associated with an increased likelihood of taking a college-entrance exam (ACT or SAT), especially among minority students

FIGURE: SMALL-CLASS EFFECT ON PROBABILITY OF TAKING A COLLEGE-ENTRANCE TEST; KRUEGER AND WHITMORE [2001]



Further, Chetty et al. [2011] manage to track participants in the STAR project up to age 27 and look at a variety of outcomes (earnings, college attendance, college quality, home ownership, savings, etc.)

Crucial policy question: do classroom environments that raise test scores – such as smaller classes and better teachers – cause analogous improvements in adult outcomes?

Again, the duration of the effects clearly matters in computing the policy benefits (to be contrasted with the policy costs)

They find that small class size (see next table):

- increases test scores during the experiment (col. 1; this is Krueger [1999]'s finding);
- increases the probability of attending college (col. 2-3);
- increases the quality of college attended (as measured by average earnings of students; col.4)
- has ambiguous effects on earnings (col. 5);
- improves the summary index (This index combines information on savings behavior, home ownership, marriage rates, mobility rates, and residential neighborhood quality; col 6)
- Some of these effects are not significant

FIGURE: SMALL-CLASS EFFECT ON DIFFERENT ADULT OUTCOMES; CHETTY ET AL. [2011]

TABLE V

EFFECTS OF CLASS SIZE ON ADULT OUTCOMES

Dependent variable	(1) Test score (%)	(2) College in 2000 (%)	(3) College by age 27 (%)	(4) College quality ($)	(5) Wage earnings ($)	(6) Summary index (%of SD)
Small class (no controls)	4.81	2.02	1.91	119	4.09	5.06
	(1.05)	(1.10)	(1.19)	(96.8)	(327)	(2.16)
Small class (with controls)	4.76	1.78	1.57	109	−124	4.61
	(0.99)	(0.95)	(1.07)	(92.6)	(336)	(2.09)
Observations	9,939	10,992	10,992	10,992	10,992	10,992
Mean of dep. var.	48.67	26.44	45.50	27,115	15,912	0.00

FIGURE: SMALL-CLASS EFFECT ON COLLEGE ATTENDANCE; CHETTY ET AL. [2011]

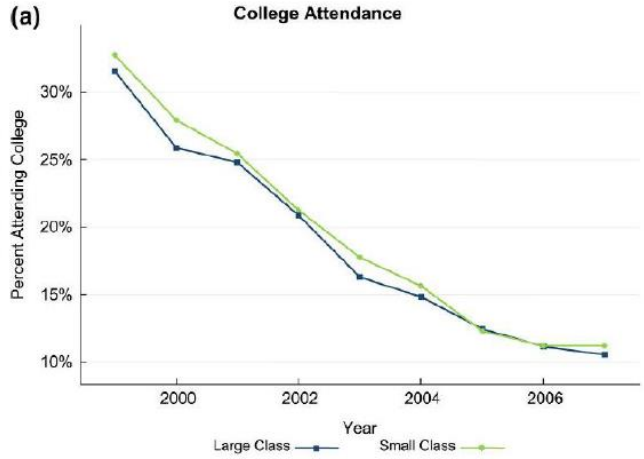
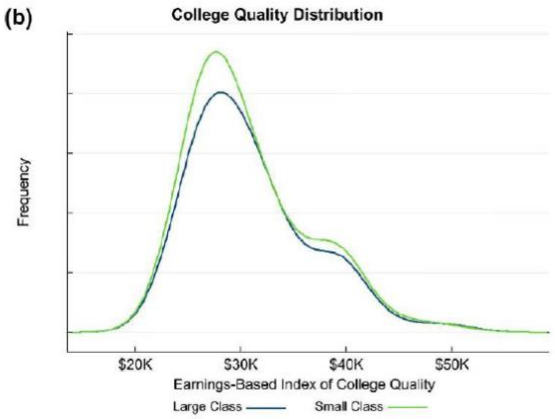


FIGURE: SMALL-CLASS EFFECT ON COLLEGE QUALITY; CHETTY ET AL. [2011]



The impacts of early childhood class assignment on adult outcomes may be particularly surprising because the impacts on test scores fade out rapidly.

The impacts of class size on test scores become statistically insignificant by grade 8 (Krueger and Whitmore [2001]), as do the impacts of class quality on test scores.

Why do the impacts of early childhood education fade out on test scores but reemerge in adulthood?

Chetty et al. [2011] find some suggestive evidence that part of the explanation may be noncognitive skills: they find that kindergarten class quality has significant impacts on noncognitive measures in fourth and eighth grade such as effort, initiative, and lack of disruptive behavior.

These results suggest that high-quality kindergarten classrooms may build noncognitive skills that have returns in the labor market but do not improve performance on standardized tests.

What is the return from going to an elite college? e.g. Harvard vs University of Massachusetts?

Reading: Angrist and Pischke [2015]; chapter 2

Comparison of earnings between those who attend different schools invariably reveal large gaps in favor of elite-college alumni

But, this comparison reflects the fact that Harvard grads typically have better high school grades and higher test scores, are more motivated, and perhaps have other skills and talents

We have a (positive) selection bias: is there any actual Harvard-effect?

We could eliminate it by random assignment, but we should first persuade Harvard to randomize its admission procedures...

In the US, the SAT is a standardized test for college admissions

SAT scores can be used to measure students “quality” and, therefore, college quality, by using college selectivity measured as average SAT scores among admitted students

Most of the descriptive studies have found that students who attended colleges with higher average SAT scores or higher tuition tend to have higher earnings when they are observed in the labor market.

An obvious concern with this conclusion is that students who attend more elite colleges may have greater earnings capacity regardless of where they attend school.

Most past studies have used OLS regression analysis to attempt to control for differences in student attributes that are correlated with earnings and college qualities (e.g. including SAT scores).

But college admission decisions are partially based on student characteristics (motivation, communication and writing skills, etc.) that are generally unobserved by researchers and therefore not held constant in the estimated wage equations

If these unobserved characteristics are positively correlated with wages, then OLS estimates will overstate the payoff to attending a selective school (i.e. it will be upward biased).

And we can clearly expect that the same unobserved characteristics that increase the chances of being accepted at Harvard will also lead to higher wages, better occupations and faster career paths

The returns from an elite college

Dale and Krueger [2002]'s idea: compare earnings among students who were accepted and rejected by a comparable set of colleges and who are comparable in terms of observable variables.

They model college enrollment

College attendance involves three sequential choices:

- a student decides which set of colleges to apply to for admission;
- colleges independently decide whether to admit or reject the student;
- the student and her parents decide which college the student will attend from the subset of colleges that admitted her.

The characteristics that the admissions committee observes and bases admission decisions on can be partitioned into two sets of variables:

- X1i : a set that is subsequently observed by researchers (the students SAT score; high school GPA - grade point average; etc.)
- X2i : a set that is unobserved by researchers (e.g. assessments of the students motivation, ambition, and maturity as reflected in her essay, college interview, and letters of recommendation).

Each college, denoted j , uses the following rule to admit or reject applicant i :

$$Z_{ij} = \gamma_1 X_1 + \gamma_2 X_2 + e_{ij} > C_j$$

$$\text{if} \quad Z_{ij} = \gamma_1 X_{1i} + \gamma_2 X_{2i} + e_{ij} > C_j \quad \text{admit to college j}$$
$$\text{if} \quad Z_{ij} \leq C_j \quad \text{reject application}$$

Where Cj is the cutoff quality level the college uses for admission. More selective colleges have higher cutoffs.

Suppose that the equation linking income to the students attributes is:

$$lnW_i = \beta_0 + \beta_1 SAT_{j*} + \beta_2 X_{1i} + \beta_3 X_{2i} + \varepsilon_i$$

where SAT_{j*} is the average SAT score of matriculants at the college student i attended, and β_1 is the monetary return to attending a more selective college.

Researchers have usually estimated a wage equation that omits X_{2i} :

$$lnW_i = \beta_0 + \beta_1 SAT_{j*} + \beta_2 X_{1i} + u_i$$

where: $u_i = \beta_3 X_{2i} + \varepsilon_i$

Since the labor market rewards X_{2i} , and school-average SAT and X_{2i} are positively correlated, the coefficient on school-average SAT will be biased upward:

$$\begin{aligned} E(lnW_i | SAT_{j*}, X_{1i}) &= \beta_0 + \beta_1 SAT_{j*} + \beta_2 X_{1i} + \\ &+ E(u_i | X_{1i}, \gamma_1 X_{1i} + \gamma_2 X_{2i} + e_{ij*} > C_{j*}) \end{aligned}$$

The expected value of the error term ui is higher for students who were admitted to, and therefore more likely to attend, more selective schools.

In other words, students with high X_{2i} will be more likely to be accepted in selective college where the average SAT_{j*} is higher

If the admission rule used by colleges depended only on X_1, and if X_1 were included in the wage equation, we would have a case of "selection on the observables": i.e. conditioning on X_1 would be enough to consistently identify the impact of college quality on earnings.

In this case, however, we have "selection on the observables and unobservables" since X_{2i} and e_{ij*} are also inputs into admissions decisions.

Dale and Krueger [2002] propose two empirical approaches:

1. the "matched applicant model ": students with the same history of application and rejections should have the same unobservables. Thus, they compare two students who were each accepted by both a highly selective college, such as the University of Pennsylvania, and a moderately selective college, such as Pennsylvania State University, but one student chose to attend Penn and the other Penn State.
2. the "self–revelation model ": students are knowledgeable about their academic potential, and reveal their potential ability by the choice of schools they apply to. Therefore, they compare two students who applied to - but were not necessarily accepted by - both Penn and Penn State.

1) Matched applicant model - Implementation

Include an unrestricted set of dummy variables indicating groups of students who received the same admissions decisions (i.e., the same combination of acceptances and rejections) from the same set of

colleges.

3 alternative matching models:

1. Similar schools: schools with average SAT scores in the same 25 point range. Sample: 6,335 matched applicants.
2. Exact schools: students who applied to and were accepted or rejected by exactly the same schools. Sample: 2,330 matched applicants.

3. Barrons ratings: classify colleges in 4 groups according to their degree of competitiveness. Two colleges are equivalent if they belong to the same category. Sample: 9,202 matched applicants.

1) Matched applicant model - Potential bias

They compare two students who were each accepted by both a highly selective college, such as the University of Pennsylvania, and a moderately selective college, such as Pennsylvania State University, but one student chose to attend Penn and the other Penn State.

But, why were students' choices different?

It is possible that the reason the student chose to attend Penn State over Penn (or vice versa) is also related to that students earnings potential: those who chose to attend a less selective school from their options may have greater or lower earnings potential.

In this case, estimates from the matched-applicant model would be biased upward or downward, depending on whether more talented students chose to matriculate to more or less selective colleges conditional on their options.

2) Self-revelation model

Implementation: this model includes the average SAT score of the schools to which students applied and dummy variables indicating the number of schools to which students applied to control for selection bias.

Potential bias:

- We compare two students who applied to but were not necessarily accepted by both Penn and Penn State.
- The student who attended Penn State is likely to have been rejected by Penn; as a result, the student who attended Penn State is likely to be less promising (as judged by the admissions committee) than the one who attended the University of Pennsylvania.
- If it is generally true that students with higher unobserved ability are more likely to be accepted by (and therefore more likely to attend) the more selective schools, the self-revelation model is likely to overstate the return to school selectivity.

Data:

In order to implement these approaches, one needs data on students characteristics, on their applications, on the outcomes of the screening process performed by colleges and on the final enrolment decisions of students

Moreover, it is necessary for students to be accepted by a diverse set of schools and for some of those students to attend the less selective colleges and others the more selective colleges from their menu of choices.

Empirical results (table III)

Without matching, the coefficient on SATj$*$ is positive and significant (column 1 and 2)

With matching (columns 3-5), it becomes zero (or significantly negative with exact match)

With self-revelation: it becomes zero

In the self-revelation model, log wage are increasing in the average SAT of the college where the students applied and in the number of applications made. Students do know their quality.

TABLE III
LOG EARNINGS REGRESSIONS USING COLLEGE AND BEYOND SURVEY, SAMPLE OF MALE AND FEMALE FULL-TIME WORKERS

	Model					
	Basic model: no selection controls		Matched-applicant model	Alternative matched-applicant models		
	Full sample	Restricted sample	Similar school-SAT matches*	Exact school-SAT matches**	*Barron's* matches***	Self-revelation model
Variable	1	2	3	4	5	6
School-average SAT score/100	0.076	0.082	−0.016	−0.106	0.004	−0.001
	(0.016)	(0.014)	(0.022)	(0.036)	(0.016)	(0.018)
Predicted log(parental income)	0.187	0.190	0.163	0.232	0.154	0.161
	(0.024)	(0.033)	(0.033)	(0.079)	(0.028)	(0.025)
Own SAT score/100	0.018	0.006	−0.011	0.003	−0.005	0.009
	(0.006)	(0.007)	(0.007)	(0.014)	(0.005)	(0.006)
Adjusted R^2	0.107	0.110	0.112	0.142	0.106	0.113
N	14,238	6,335	6,335	2,330	9,202	14,238

Hence, Dale and Krueger [2002] suggest that, once one properly controls for students' characteristics, attending an elite college does not increase future earnings in the labor market

In other words, high ability students would perform well irrespectively of the college attended

If attending an elite college is more expensive than attending a good college, the investment does not seem to be good value for money...

But, in a world with asymmetric information, where potential employers ignore your underlying ability, having attended a selective college sends a clear signal that you are excellent students (see the job market signaling model by Spence [1973])

Returns from enrolling in a flagship university

Hoekstra [2009] examines the economic returns of attending the most selective public state university (flagship university)

He uses a **regression discontinuity design** that compares the earnings of 28 to 33 year olds who were barely admitted to the flagship to those of individuals who were barely rejected.

This paper uses an admission discontinuity to estimate the causal effect of enrollment at the state's flagship university on earnings: in order to be admitted, students needed a SAT (standardized test

widely used for college admissions in the US) above a certain threshold

This design will distinguish the effect of enrollment at the flagship university from other confounding factors so long as the determinants of earnings (e.g., motivation, parental support) are continuous at the admission cutoff.

Under this assumption, any discontinuous jump in earnings at the admission cutoff is properly interpreted as the causal effect of admission to the flagship university on earnings.

The idea of Regression Discontinuity Design is that of looking for some discontinuity that generates a local randomized experiment

Suppose that the test admission threshold to participate in a treatment (e.g. elite college) is 70/100

Candidates who get less than 50 and those who get more than 80 are clearly very different: comparing their labour market outcomes after the treatment would tell us little about the treatment effect

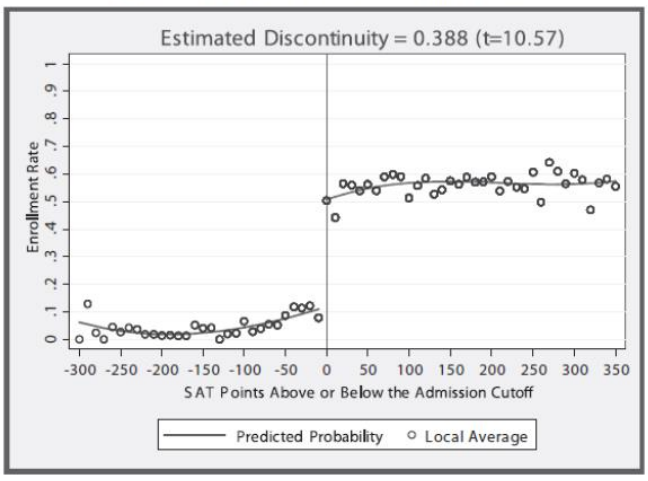
We would capture the fact that the former group of candidates has lower ability than the latter

But, suppose we could compare individuals who got 69 with those who got 70: they are basically identical individuals, but the former group was slightly unlucky during the test

Basically, idiosyncratic shocks in admission test performance make admission close to the cutoff plausibly random (close to the cutoff = local)

DOES THE ADMISSION CUTOFF PREDICT THE ENROLLMENT DECISIONS OF APPLICANTS?

FIGURE I.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



Does the Admission Cutoff Predict the Enrollment Decisions of Applicants?

Yes, there is a substantial jump (38 percentage points) in the probability of being enrolled at the flagship university before and after the cutoff

The probability of being enrolled if a student had a SAT above the threshold was about 50-60 percent

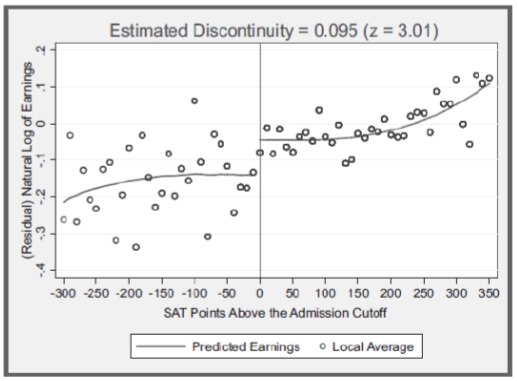
The probability is not 100 percent, because not all students with high SAT would enroll in the flagship university (they may have chosen a different one)

The probability of being enrolled if a student had a SAT below the threshold was around 10 percent for students up to 50 points below the cutoff and basically zero for students with even lower SAT

The probability is not zero for everyone because the admission rules allowed students with SAT slightly below the threshold to be admitted if they had a sufficiently high high school GPA (Grade Point Average)

This is a **fuzzy regression discontinuity design** (more on this later on in the course)

FIGURE 2.—NATURAL LOG OF ANNUAL EARNINGS FOR WHITE MEN TEN TO FIFTEEN YEARS AFTER HIGH SCHOOL GRADUATION (FIT WITH A CUBIC POLYNOMIAL OF ADJUSTED SAT SCORE)



And, is there an earnings discontinuity at the admission cutoff?

Yes, there seems to be a discontinuous increase in earnings at the admission cutoff

There is no reason to expect that individuals who had a SAT slightly below and slightly above the admission cutoff would have different productivity in the labor market: they should command the same wage

The difference we observe is due to the fact that a large fraction (50-60 percent) of those above the threshold attended the best public university in the state

According to Hoekstra [2009]'s estimates, attending the flagship state university increases the earnings of 28- to 33-year-old white men by approximately 20 percent

This is a large return. If it is persistent over the life cycle, it amounts to a lot of money...

Reading list

- Angrist and Pischke [2015]: chapters 2.1-2.2 and 6.1
- Dale and Krueger [2002]

Suggested readings:

- Chetty et al. [2011]
- Hoekstra [2009]
- Krueger [1999]

LECT. 2: ESTIMATING CAUSAL POLICY EFFECTS

Table of contents

In order to conceive a credible policy evaluation, we need to be able to answer four questions:

1) what is the causal relationship of interest?
2) which experiment could ideally be used to capture the causal effect of interest?
3) what is your research design (i.e. “identification strategy”)?
4) what is your econometric methodology?

The potential outcome approach

We now introduce the potential outcome framework which we will use throughout this course. This approach has been given several names, referring to the different authors who have contributed to its early

development: mainly, researchers refer to it as the Fisher, Roy or “causal Rubin” model.

We have a population of individuals: for each individual we observe an outcome variable Y and a treatment variable (a potential cause) D. Suppose we observe that D and Y are correlated.

Does correlation between D and Y imply that D causes Y ? Not necessarily.

We use the following notation:

- i is an index for the individuals in the population;
- Di is the treatment, the potential cause whose effects we want to estimate. D_i is a dummy variable: Di ∈ 0, 1. $D_i = 1$ if individual I has been exposed to treatment;
- Y_i (D_i) is the outcome (which may depend on D_i);
    - Y_i (1) is the outcome if the individual i has received the treatment; alternative notation: Y_i ($D_i = 1$) or Y_{1i} ;
    - Y_i (0) is the outcome if the individual i has not received the treatment; alternative notation: Y_i ($D_i = 0$) or Y_{0i} .

The **potential outcome** for each individual can be written as:

$$Y_i(D_i) = D_i Y_i(1) + (1 - D_i) Y_i(0) = \underbrace{Y_i(0)}_{\text{outcome if not treated}} + D_i \underbrace{(Y_i(1) - Y_i(0))}_{\text{+difference if treated}}$$

This approach requires to think in terms of “counterfactuals”: how reality would have been if something had - or had not - happened.

In this case, the counterfactual is what would have happened to the treated individuals in the absence of treatment. Or, what would have happened to the untreated individuals if they had been given the treatment.

Basically, any empirical question in economics and politics can be thought of in terms of potential treatments and outcomes. And the individual who receives the “treatment” can be a human being as well as a firm, a social group, a region, a country, etc.

- What is the impact of having a kid on female labor supply?
- What is the effect of the presence of immigrants on natives’ wages and employment?
- Are earnings lower for individuals who are female, black, gay, etc.?
- How does divorce change individuals’ wellbeing and mental health?
- Does the presence of natural resources in some areas increase the likelihood of having a conflict (resource curse)?
- How do different electoral campaign technologies affect electoral outcomes?
- Do medical marijuana laws increase hard-drug use?

Fundamental Problem of Causal Inference

Definition of causality.

For every individual i , the event Di = 1 rather than Di = 0 causes the effect $\Delta Yi = Yi(1) - Yi(0)$.

Proposition.

It is impossible to observe for the same individual i both Di = 1 and Di = 0, as well as both Y1(1) and Yi (0). Therefore, it is impossible to observe $\Delta Yi = Yi(1) - Yi(0)$, i.e. the causal effect of the treatment D on the outcome Y (Holland [1986]).

We can see the “fundamental problem of causal inference” as a **missing counterfactual problem**.

We do not – and cannot – observe what would have happened to the treated individuals in the absence of treatment. Or, what would have happened to the untreated individuals if they had been given the treatment.

The existence of this “fundamental problem of causal inference” does not imply that we will not be able to estimate causal effects, but it clearly suggests that it is a challenging enterprise.

Addressing this challenge is one of the main aims of contemporary micro-econometrics (and of this course).

As we will see throughout this course, all the different methods which have been proposed in the literature attempt to provide a reliable substitute for the missing counterfactual.

”At the heart of this kind of policy evaluation is a missing data problem. An individual may either be subject to the policy intervention or she may not, but no one individual can be in both states simultaneously. Indeed, there would be no evaluation problem of the type discussed here if we could observe the counterfactual

outcome for those in the program had they not participated. Constructing this counterfactual in a convincing way is a key ingredient of any serious evaluation method.” (Blundell and Dias [2009], p. 566).

There are two main ways of addressing this problem:

1. produce/use experimental or quasi-experimental data (randomized experiments; natural experiment; regression discontinuity design) and then use straightforward OLS regressions;
2. use more sophisticated econometric methodologies with observational non-experimental data (e.g. instrumental variables, fixed effect estimation, difference-in-differences, etc.)

Homogeneous Vs Heterogeneous treatment effect

Before discussing how to estimate these causal effects, we need to carefully think about the intrinsic characteristics of the treatment effect we want to estimate

Do we expect the treatment effect to be homogenous rather than heterogeneous across the population? Are individual responses to a policy homogeneous or do responses differ across individuals?

If the responses differ, do they differ in a systematic way?

Going back to our college education example: do we expect the "college treatment" to produce an X percent increase in earnings for all those who get their university degree?

Or do we think that some individuals may gain more than others, depending on their observable (gender, age, family income, etc.) and unobservable characteristics (ability, motivation, etc.)?

Recall, for instance, that we saw that the impact on test scores of the Tennessee STAR Project was larger for minority and poor kids

We can write the potential outcomes with and without treatment as:

$$Y_i(1) = \alpha + \beta_i + u_i \qquad (1)$$

$$Y_i(0) = \alpha + u_i \qquad (2)$$

where α is the intercept parameter (i.e. the average outcome without treatment), βi is the effect of treatment on individual i and ui is the unobservable component of Y .

If we replace these expressions into the potential outcome expression:

$$\begin{aligned} Y_i(D_i) &= D_i Y_i(1) + (1 - D_i) Y_i(0) = \\ &= D_i(\alpha + \beta_i + u_i) + (1 - D_i)(\alpha + u_i) = \\ &= \alpha + \beta_i D_i + u_i \end{aligned} \qquad (3)$$

We have:

- **Homogenous returns**: Y1i − Y0i = β ∀i
- Heterogenous returns: Y1i − Y0i = βi

"The distinction between homogenous and heterogeneous responses is central to understand what parameters alternative evaluation methods measure. In the homogeneous linear model, common in elementary econometrics, there is only one impact of the program and it is one that would be common to all participants and nonparticipants alike. In the heterogeneous model, the treated and nontreated may benefit

differently from program participation. In this case, the average treatment effect among the treated will differ from the average value overall or on the untreated individuals. Indeed, we can define a whole distribution of the treatment effects." (Blundell and Dias [2009], p. 569)

After establishing how the returns are expected to be (homogenous Vs heterogenous), one can discuss the (causal) parameter of interest.

Estimation methods typically identify some average impact of treatment over some sub-population.

The four most commonly used parameters are:

1. Average Treatment Effect (ATE);
2. Average Treatment on the Treated (ATT);
3. Average Treatment on the Non-Treated (ATNT);
4. Local Average Treatment Effect (LATE).

1) Average Treatment Effect (ATE):

the average effect in the whole population (whether or not one takes the treatment); this is the expected effect of treating a random individual;

$$ATE = E(Y_{1i} - Y_{0i}) = E(\beta_i)$$

2) Average Treatment on the Treated (ATT):

the average effect among those individuals who took the treatment; this is the expected effect of treating an individual who has chosen to be treated (or has been selected for being treated);

$$ATT = E(Y_{1i} - Y_{0i}|D_i = 1) = E(\beta_i|D_i = 1)$$

3) Average Treatment on the Non-Treated (ATNT):

the average effect among those individuals who did not take the treatment; this is the expected effect of treating an individual who has not chosen to be treated (or has not been selected for being treated);

$$ATNT = E(Y_{1i} - Y_{0i}|D_i = 0) = E(\beta_i|D_i = 0)$$

4) Local Average Treatment Effect (LATE):

the (local) average effect among those individuals who belong to a certain specific sub-group (we will see more on this later in the course)

$$LATE = E(Y_{1i} - Y_{0i}|i \in \{...\}) = E(\beta_i|i \in \{...\})$$

All these parameters will be identical under homogeneous treatment

effects.

Indeed, if $\beta_i = \beta \quad \forall i$:

$$ATE = E(\beta_i) = ATT = E(\beta_i|D_i = 1) = ATNT = E(\beta_i|D_i = 0) = \beta$$

Under heterogeneous treatment effects, instead, a non-random process of selection into treatment may lead to differences between them.

However, whether the impact of treatment is homogeneous or heterogeneous, selection bias may be present (i.e. individuals choosing to be treated are different from those who do not).

Moreover, the problem in estimating these average effects is always the absence of the (average) counterfactual. We cannot observe the same individual with and without treatment. This is precisely the Fundamental Problem of Causal Inference.

The missing counterfactual for each average effect is underlined in the expressions below:

- ATT:

$$E(Y_{1i} - Y_{0i}|D_i = 1) = E(Y_{1i}|D_i = 1) - \underline{E(Y_{0i}|D_i = 1)} \quad (4)$$

- ATNT:

$$E(Y_{1i} - Y_{0i}|D_i = 0) = \underline{E(Y_{1i}|D_i = 0)} - E(Y_{0i}|D_i = 0) \quad (5)$$

- ATE:

$$\begin{aligned}
&E(Y_{1i} - Y_{0i}) = \\
&= ATT \cdot Pr(D = 1) + ATNT \cdot Pr(D = 0) = \\
&= E(Y_{1i} - Y_{0i}|D_i = 1)Pr(D = 1) + \\
&+ E(Y_{1i} - Y_{0i}|D_i = 0)Pr(D = 0) = \\
&= [E(Y_{1i}|D_i = 1) - \underline{E(Y_{0i}|D_i = 1)}]Pr(D = 1) + \\
&+ [\underline{E(Y_{1i}|D_i = 0)} - E(Y_{0i}|D_i = 0)]Pr(D = 0)
\end{aligned}$$

Selection into treatment

Consider our potential outcome expression:

$$\begin{aligned}
Y_i(D_i) &= D_i Y_i(1) + (1 - D_i)Y_i(0) = \\
&= D_i(\alpha + \beta_i + u_i) + (1 - D_i)(\alpha + u_i) = \\
&= \alpha + \beta_i D_i + u_i
\end{aligned}$$

Define with β^{ATE} the Average Treatment Effect parameter and add and subtract it (multiplied by the participation dummy Di) in the expression above.

$$\begin{aligned}
Y_i &= \alpha + \beta_i D_i + u_i + \beta^{ATE} D_i - \beta^{ATE} D_i = \\
&= \alpha + \beta^{ATE} D_i + u_i + (\beta_i - \beta^{ATE}) D_i = \\
&= \alpha + \beta^{ATE} D_i + e_i
\end{aligned}$$

$$\text{where } e_i = u_i + (\beta_i - \beta^{ATE}) D_i. \quad (6)$$

The individual decision to participate into treatment (Di) is likely to be determined by personal characteristics - observed and unobserved - that may also influence the outcome Y .

For instance: are the most motivated individuals those that invest more in their education? Are the individuals with higher returns from education those who obtain more qualifications? Are these groups of people also more likely to have high earnings in the labor market?

We can explicitly model the participation decision

Latent utility from participation: $D_i^* = Z_i\gamma + \nu_i$

Participation decision:

$$D_i = \begin{cases} 1 & \text{if} \quad D_i^* \geq 0 \\ 0 & \text{if} \quad D_i^* < 0 \end{cases}$$

where Zi is a vector of observable characteristics which determine the participation decision.

Note that Di (= participation into treatment) is observed, while D_i^* (= underlying utility from participating into treatment) is not

Now, we can rewrite the ATE, ATT and ATNT parameters as:

- $ATE = E(\beta_i) = \beta^{ATE}$
- $ATT = E(\beta_i|D_i = 1) = E(\beta_i|Z_i\gamma + \nu_i \geq 0)$
- $ATNT = E(\beta_i|D_i = 0) = E(\beta_i|Z_i\gamma + \nu_i < 0);$

Let's rewrite equation (6):

$$Y_i = \alpha + \beta^{ATE} D_i + e_i$$

where $e_i = u_i + (\beta_i - \beta^{ATE})D_i$.

Non-random selection occurs if the unobservable term e_i is correlated with the treatment variable Di .

That is, Di is endogenous in the regression if $Cov(D_i , e_i) \neq 0$

We can identify two main types of selection, depending on whether the correlation between the error term e_i and the treatment variable Di is originated by e_i being correlated with the regressors determining the assignment (Zi) or with the unobservable component *v*i in the selection-into-treatment equation.

In the first case we talk of "**selection on the observables**", while in the second of "**selection on the unobservables**".

If we have selection on the observables, individuals with certain observable characteristics Z (e.g. age, education, marital status, gender, etc.) are systematically more likely to select into treatment (i.e. to choose to undertake the treatment) than individuals with different observables characteristics.

While with selection on the unobservables, the group of treated systematically differs from the group of untreated with respect to unobservable characteristics (e.g. ability, risk aversion, discount rate, motivation, etc.).

Addressing the presence of endogenous selection into treatment is a major challenge for policy evaluation.

We can further distinguish selection by looking at the reason why individuals select into treatment.

- When the correlation between the error term ei and the treatment variable Di is originated by correlation between ui and Di we say there is "**selection on the untreated outcomes**" as individuals with different untreated outcomes are differently likely to become treated.
- If, on the other hand, selection arises due to a relationship between βi and Di we say there is "**selection on the expected gains**", whereby expected gains determine participation.

Summarizing, if the correlation between the unobservable term ei and the treatment variable Di is different from zero (i.e. non random selection) because:

- $cov(e_i, Z_i) \neq 0$, we have **selection on the observables**;
- $cov(e_i, v_i) \neq 0$, we have **selection on the unobservables**;
- $cov(u_i, D_i) \neq 0$, we have **selection on untreated outcomes**;
- $cov(\beta_i, D_i) \neq 0$, we have **selection on the expected gains**.

Note, that we can have several or even all of these types of selection at the same time

Suppose we consider a training program offered to unemployed workers.

Unless the program is offered randomly to the workers, we can expect some endogenous selection to take place.

We can expect younger individuals (selection on the observables) and those who are more motivated in their job search (selection on the unobservables) to be more willing to take up the course.

And the reason why some groups - defined by some observable and/or unobservable characteristics - are more likely to join the treatment is that either their outcomes as untreated would be particularly poor

(selection on untreated outcomes) or they expect particularly high gains from the treatment (or their cost from participating is particularly low) (selection on the expected gains), or both.

A näıve comparison

Even if we can not observe for the same individual the outcome with and without treatment, we can have a sample containing both treated and untreated individuals and we can then compare their average outcomes.

We have:

$$E[Y_i|D_i = 1] = E[D_i Y_{1i} + (1 - D_i)Y_{0i}|D_1 = 1] = E[Y_{1i}|D_i = 1]$$

$$E[Y_i|D_i = 0] = E[D_i Y_{1i} + (1 - D_i)Y_{0i}|D_1 = 0] = E[Y_{0i}|D_i = 0]$$

And both E [Y_{1i} |D_i = 1] and E [Y_{0i} |D_i = 0] are observable and can be estimated from the data.

Hence, we can estimate the difference between the two expected values:

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

Is this the causal effect of D on Y that we want to estimate?

Not necessarily.

Add and subtract the term E [Y0i |Di = 1] on the rhs of the equation:

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

where:

- a: observed difference in average outcome
- b: average treatment effect on the treated (ATT)
- c : selection bias

b is the parameter we would like to estimate (ATT)

But with our naive comparison we obtain a=b+c

By comparing the average outcome of the treated and untreated group we identify the effect of interest (i.e. average treatment effect on the treated) only if the selection bias is zero:

b = a iff c = 0, i.e. iff E [Y0i |Di = 0] = E [Y0i |Di = 1]

The selection bias is zero if the expected outcome if not treated of those who chose to take the treatment (E [Y0i |Di = 1]) is equal to the expected outcome if not treated of those who did not choose to take the treatment (E [Y0i |Di = 0])

The selection bias captures the difference in potential untreated outcomes between the treatment and comparison groups: individuals in the treatment group may have had different outcomes on average even if they had not been treated.

“any difference between the average outcomes of the two groups can be attributed to both the impact of the program or pre-existing differences (the “selection bias”). Without a reliable way to estimate the size of this selection bias, one cannot decompose the overall difference into a treatment effect and a bias term” (Duflo et al. [2008], p.3900).

What if we run a regression to perform our “naive comparison”? We have:

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

- which is equivalent to writing $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$, where:
  - $\alpha = E(Y_{0i})$
  - $\beta = (Y_{1i} - Y_{0i})$
  - $\varepsilon_i = Y_{0i} - E(Y_{0i})$

- Take the conditional expectations of this equation (with treatment status switched on and off):

$$E[Y_i|D_i = 1] = \alpha + \beta + E[\varepsilon_i|D_i = 1]$$

$$E[Y_i|D_i = 0] = \alpha + E[\varepsilon_i|D_i = 0]$$

- Therefore:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \beta + E[\varepsilon_i|D_i = 1] - E[\varepsilon_i|D_i = 0]$$

- where:
  - β: treatment effect
  - $E[\varepsilon_i|D_i = 1] - E[\varepsilon_i|D_i = 0]$: selection bias

In general, we have good reasons to think that individuals who chose to be treated differ with respect to those who did not.

If the treated and untreated individuals are identical (with respect to their unobservables), why have the first ones chosen to get treated while the latter ones did not?

Any difference in the unobservables between the two groups will imply a selection bias different from zero:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \neq \beta \quad \text{if} \quad E[\varepsilon_i|D_i = 1] \neq E[\varepsilon_i|D_i = 0]$$

Solutions to the evaluation problem

At the heart of the development of the different policy evaluation methods is a missing data problem, that is, the so called “Fundamental Problem of Causal Inference”. An individual may be either subject to the intervention or may not, but no one individual can be in both states simultaneously.

“Constructing this counterfactual in a convincing way is a key ingredient of any serious evaluation method.” (Blundell and Dias [2008], p. 2)

The choice of evaluation method will depend on three aspects:

1. the nature of the question to be answered;
2. the type and quality of data available;
3. the mechanism by which individuals are allocated to the program/receive the treatment (i.e. the assignment rule).

One can identify five main approaches to policy evaluation:

1. **Randomized control trials**: this approach is the closest to the “theory-free” method of a clinic trial, relying on the availability of a randomized assignment rule;

2. **Regression Discontinuity Design**: they exploit “natural” discontinuities in the rules used to assign individuals to treatment;

3. **Natural experiments (Difference in Differences)**: they exploit randomization to programs created through some occurring event (determined by Nature or by governments) external to the researcher;

4. **OLS regressions and matching**: these methods attempt to reproduce the treatment group among the non-treated, re-establishing the experimental conditions in a non-experimental setting, but relying on observable variables to account for selection;

5. **Instrumental variables**: they are a step closer to the structural method, relying on exclusion restrictions to achieve identification;

Creating or finding the counterfactual? Experimental, quasi- and non-experimental methods

Experimental, quasi- and non-experimental methods

Main focus of policy evaluation literature:

1. how to implement randomized experiments
2. develop alternatives to randomized experiments, using observational data instead of experimental data (LaLonde [1986]; Blundell and Dias [2008])

With **experimental methods** (i.e. randomized experiments), randomization of treatment is used to avoid any endogenous selection into treatment.

Treated and untreated individuals are chosen randomly and, therefore, they are identical (on average) in both observables and unobservables.

In this case, the missing counterfactual is created by the researcher through the experiment.

Quasi-experimental methods (i.e. regression discontinuity, differences-in-differences, instrumental variables) use observational data and exploit some event (nature, history, policy change, etc.) which created randomness in the assignment to treatment.

Although researchers do not design the experiment, they exploit these natural/social experiments in order to achieve identification.

Non–experimental methods (i.e. OLS and matching) make use of observational data, appropriate identifying assumptions and specific statistical methods to recover the missing counterfactual.

With non–experimental methods, the researcher needs to find a comparison (control) group: the average outcome of the treated group will be compared with the average outcome of the comparison group.

These comparison groups are valid under a set of identifying assumptions, which – by definition – are not testable: the validity of any particular study critically depends on how convincing the assumptions appear.

Internal and external validity

When presenting all the different approaches which have been developed to address the evaluation problem, we will also discuss their internal and external validity.

- **Internal validity** (or internal consistency) refers to whether the approach succeeds in identifying the causal parameter of interest (and under which identifying assumptions).
- **External validity**, instead, refers to the possibility of generalizing the results obtained with that approach to other contexts, samples and variations of the policy (i.e. are the results generalizable and replicable?).

Internal validity is a necessary condition for external validity, but it is not a sufficient one.

In practice, there seems to be some sort of trade–off between internal and external validity of the different methods of estimating causal policy effects: results obtained by perfectly internally consistent approaches may be hardly generalizable, while very general results may be obtained at the price of making several identifying assumptions (which may undermine the internal validity of the approach).

Reading list

Compulsory reading:

- lecture slides

Suggested reading:chapter 3 - Gertler et al. (2011) Impact Evaluation in Practice, World Bank (available on-line and on Ariel)

LECT. 3: RANDOMIZED EXPERIMENTS

Table of contents

“**Experiment**”: all those cases where different treatments are exogenously assigned to different individuals.

The assignment can be decided by the researcher, as in the randomized control trials, or can be (more or less voluntarily) operated by Nature, history or governments.

In the first case, we talk of **social experiments**, while in the second case of natural **experiments** (or quasi-experiments).

Studies which exploit “natural experiments” to achieve identification, examine outcome measures for observations in treatment groups and comparison groups that are not randomly assigned. Good natural experiments are studies in which there is a clearly exogenous source of variation in the explanatory variables that determines the assignment to treatment.

A few examples of natural experiments where Nature “decides” the assignment are: the gender of children (Angrist and Evans [1998]), the German re-unification (Fuchs-Schuendeln and Schuendeln [2005]), terrorist attacks (Tella and Schargrodsky [2004], Draca et al. [2011]), sudden and unexpected inflows of immigrants in one area (Card [1990], Friedberg [2001]), weather shocks, etc.

In other cases, it is the government which “decides” the assignment, for instance, with policy changes which take place in certain areas but not in others, or which affect individuals with certain characteristics and not others (e.g. minimum wage laws, tax reforms, etc.).

As we will see throughout this course, natural experiments allow researchers to identify causal impacts through different econometric techniques such as before-after, difference-in-differences or instrumental variable estimators.

Identification in randomized experiments

Suppose that you can extract two random samples - a treatment T and a control C group - from the population of interest.

By construction, these two samples are statistically identical to each other and to the entire population (as long as the randomization is properly done).

We can write:

$$E(Y_{0i}|i \in C) = E(Y_{0i}|i \in T) = E(Y_{0i})$$

$$E(Y_{1i}|i \in C) = E(Y_{1i}|i \in T) = E(Y_{1i})$$

Suppose all individuals in the treatment group T receive the treatment and all those in the control group C do not (i.e. perfect compliance):

$$Pr(D_i = 1|i \in T) = 1 \qquad and \qquad Pr(D_i = 1|i \in C) = 0$$

- Therefore, we can only observe: $E(Y_{0i}|i \in C)$ and $E(Y_{1i}|i \in T)$. We use the $\widehat{hat}$ to define moments/elements which we can estimate from the data. Hence:

$$E(\widehat{Y_{0i}|i \in C}) = E(Y_{0i}|i \in T) = E(Y_{0i})$$

$$E(Y_{1i}|i \in C) = E(\widehat{Y_{1i}|i \in T}) = E(Y_{1i})$$

Fundamental Problem of Causal Inference is solved!

We can estimate:

$$ATE = E(Y_{1i}) - E(Y_{0i}) = E(\widehat{Y_{1i}|i \in T}) - E(\widehat{Y_{0i}|i \in C})$$

We use the control group C as an image of what would have happened to the treatment group T in the counterfactual situation of no treatment.

If we have data from a randomized experiment, we can run the näıve regression:

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

where Di is now a dummy for assignment to the treatment group.

The randomization ensures that Di is not endogenous (i.e. cov(D_i , ε_i) = 0)

Estimating this equation with OLS we obtain the parameter of interest:

$$\widehat{\beta_{OLS}} = E[\widehat{Y_i|D_i} = 1] - E[\widehat{Y_i|D_i} = 0]$$

“This result tells us that when a randomized evaluation is correctly designed and implemented, it provides an unbiased estimate of the impact of the program in the sample under study. This estimate is **internally valid**.” (Duflo et al. [2008], p.3902).

Examples of “famous” randomized experiments

In developing countries:

- Progresa program in Mexico Paul Schultz [2004], Todd and Wolpin [2006]
- Primary School Deworming Project (PSDP) in Busia (Kenya) Miguel and Kremer [2004].

In the US:

- Tennessee Project STAR (Student/Teacher Achievement Ratio) Krueger [1999].
- “Moving to Opportunities” (MTO) Katz et al. [2001].

Advantages and disadvantages of randomized experiments

Advantages:

- (potentially) the most convincing evaluation method
- completely non-parametric: no assumptions on functional forms and distribution of the error terms; basically, it is just a comparison of two means;
- randomization eliminates selection bias whether it is due to observables or unobservables. Asymptotically, the treatment and the control group have identical distributions of both observables and unobservables;
- the support of the distribution of all relevant variables is the same in both treatment and control.

Disadvantages:

- not always feasible
- not always ethically feasible
- not always politically feasible
- (usually) are quite expensive
- time consuming

Critical issues with randomized experiments

Even when they are feasible, there are still three important issues:

1. Partial (or imperfect) compliance [Conformità parziale]
2. Spillover effects [Effetto di diffusione/ricaduta (in economia)]
3. External validity

1) Partial (or imperfect) compliance

In many cases, only a fraction of the individuals who are offered the treatment actually take it up, and, some members of the comparison group may receive the treatment.

If these "movements" were random, and the researchers did not observe them, the main implication of imperfect compliance would be an attenuation of the actual difference between the outcomes of treated and untreated individuals.

But these "movements" are unlikely to be random: if the more motivated individuals among those who have been "randomized out" manage to get the treatment, and if the less motivated among those "randomized in" drop out, we are back to the problem of selection into treatment (i.e. a selection bias in the effect estimates).

From actual treatment to initial randomization: "To be valid and to prevent the reintroduction of selection bias, **an analysis needs to focus on groups created by the initial randomization**. One must compare all those initially allocated to the treatment group to all those initially randomized to the comparison group, whatever their actual behavior and their actual treatment status. The analysis cannot exclude subjects or cut the sample according to behavior that may have been affected by the random assignment." (Duflo et al. [2008], p.3936)

Z : the variable that is randomly assigned (i.e. Zi = 1 if the individual i was randomly assigned to the treatment).

D: treatment of interest (i.e. Di = 1 if the individual i took the treatment).

Now, denote with Y^z_{0i} the potential outcome for an individual if Zi = 0 (i.e. if she has not been randomly assigned to the treatment), and Y^z_{1i} the potential outcome if Zi = 1 (i.e. if she has been randomly assigned to the treatment).

Given that Z is randomized, we have:

$$E(Y_{0i}^Z|Z_i = 1) = E(Y_{0i}^Z|Z_i = 0) = E(Y_{0i}^Z) \qquad (1)$$

$$E(Y_{1i}^Z|Z_i = 1) = E(Y_{1i}^Z|Z_i = 0) = E(Y_{1i}^Z) \qquad (2)$$

We can now compare the outcomes of those "randomized in" and those "randomized out".

In this way we can estimate a new parameter, the Intention to Treat effect (ITT), which measures the impact of being offered the treatment.

Formal Identification of ITT.

$$\begin{aligned} ITT &= E(Y_{1i}^Z - Y_{0i}^Z|Z_i = 1) = E(Y_{1i}^Z|Z_i = 1) - E(Y_{0i}^Z|Z_i = 1) \\ &= E(Y_{1i}^Z|Z_i = 1) - E(Y_{0i}^Z|Z_i = 0) = E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) \end{aligned}$$

where the randomization of Z allows to replace the unobservable term E (Y^{z}_{0i} |Zi = 1) with the observable term E (Y^{z}_{0i} | Zi = 0). In the last step of the ITT expression, we have returned to the potential outcome notation which refers to the actual treatment D.

Intention to Treat effect (ITT)

Clearly, given that Z is not equal to D (because there is imperfect compliance), the effect of the intention to treat is not the same thing as the effect of the treatment D.

Still, in many contexts, the ITT may actually be the parameter of interest: this is the case when policy-makers are considering the introduction of a policy which offers a treatment but which does not enforce or check the compliance of individuals.

In other cases, instead, the parameter of interest is the actual effect of the treatment. In this case, the randomized variable Z can be used as an instrumental variable for the treatment of interest D

*An **instrumental variable** (sometimes called an "instrument" **variable**) is a third **variable**, Z, used in regression analysis when you have endogenous **variables**—**variables** that are influenced by other **variables** in the model. In other words, you use it to account for unexpected behavior between **variables**.*

We can distinguish three cases of compliance:

1. **Perfect compliance**: all individuals assigned to the treatment group ("randomized in") get the treatment, and all those assigned to the control group ("randomized out") do not get the treatment
2. **One-sided compliance**: all those "randomized out" do not get the treatment; while those "randomized in" can choose not to take the treatment (e.g programs where one can perfectly prevent ineligible individuals from receiving the treatment but one can not force eligible individuals to receive it)
3. **Two sided non-compliance**: some of those "randomized out" manage to get the treatment, while those "randomized in" can choose not to take the treatment.

Depending on the case, we can identify different parameters.

From ITT to ATE/ATT/LATE

Rewrite ITT as:

$$\begin{aligned} ITT &= E(Y_i|Z_i=1) - E(Y_i|Z_i=0) = \\ &= [E(Y_i|Z_i=1, D_i=1)Pr(D_i=1|Z_i=1) + \\ &+ E(Y_i|Z_i=1, D_i=0)Pr(D_i=0|Z_i=1)] + \\ &- [E(Y_i|Z_i=0, D_i=1)Pr(D_i=1|Z_i=0) + \\ &+ E(Y_i|Z_i=0, D_i=0)Pr(D_i=0|Z_i=0)] \end{aligned}$$

1) Perfect compliance

With perfect compliance, we have:

- $Pr(D_i=0|Z_i=1)=0$;
- $Pr(D_i=1|Z_i=0)=0$;
- $Pr(D_i=1|Z_i=1)=Pr(D_i=0|Z_i=0)=1$.

basically, Zi corresponds exactly to Di

This implies that ITT identifies **ATE**:

$$\begin{aligned} ITT &= E(Y_i|Z_i=1, D_i=1) - E(Y_i|Z_i=0, D_i=0) \\ &= E(Y_i|D_i=1) - E(Y_i|D_i=0) = \\ &= E(Y_{1i}) - E(Y_{0i}) = ATE \end{aligned}$$

2) One-sided compliance: ATT

- All those “randomized out” do not get the treatment
- Those “randomized in” can choose not to take the treatment

We have:

- $0 < Pr(D_i = 1|Z_i = 1) < 1;$
- $Pr(D_i = 1|Z_i = 0) = 0;$
- $0 < Pr(D_i = 0|Z_i = 1) < 1;$
- $Pr(D_i = 0|Z_i = 0) = 1.$

We can show that (see Angrist and Pischke [2008] theorem 4.4.2):

$$\underbrace{E(Y_{1i} - Y_{0i}|D_i = 1)}_{ATT} = \frac{\overbrace{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}^{ITT}}{Pr(D_i = 1|Z_i = 1)}$$

The ATT is equal to the ITT divided by the share of individuals who undertook [ha conseguito] the treatment (Di = 1) after having been “randomized in” (Zi = 1).

This is the Wald estimator (IV estimator → Instrumental variable), where the randomized variable Zi is used as an instrument for the actual treatment Di

3) Two sided non-compliance: LATE

- Some of those “randomized out” manage to get the treatment, and, those “randomized in” can choose not to take the treatment.

Again, the dummy Z is used as an instrument for D.

under the two LATE assumptions of independence and monotonicity (Imbens and Angrist [1994]) the IV coefficient can be interpreted as the average treatment effect for a well-defined group of individuals, namely those who are induced by the instrument Z to take advantage of the treatment D (compliers) (more on this later in the course).

Which parameter is identified?

- perfect compliance: ATE
- **imperfect compliance: ITT (Intention to Treat).** Moreover, with specific types of compliance, and under additional assumptions, we can identify:
  - with one-sided compliance: ATT
  - with two sided non-compliance: LATE

“When the randomization only induces imperfect assignment to the treatment and the comparison groups, it is therefore still possible to make meaningful causal statements. However, the average causal effect that is estimated is not necessarily representative of the average causal effect for the entire population. Depending on circumstances, it may or may not be representative of a sub-population of interest.” (Duflo et al. [2008], p.3939).

2) Spillover effects (externalities)

“Spillover = [Effetto di diffusione/ricaduta (in economia)]”

Experimental interventions can create spillover effects such that untreated individuals are (at least partially) affected by the treatment.

Spillovers may be physical, or they may be originated by changes in relative prices, by learning and imitation effect, etc.

Positive spillovers: untreated individuals benefit from the treatment given to the treated individuals

- The estimated ITT will be smaller than the effect one would have observed in the absence of the externality.
- This implies that the entire community - treated and untreated individuals together - is benefiting more from the policy than one would conclude by just looking at the estimated ITT
- The positive externality is not taken into account in the ITT parameter (and it should)

[Per esempio: persone vaccinate non solo beneficiano del vaccino, ma anche coloro che gli stanno attorno (che non sono vaccinati) non possono essere contagiati dalla malattia]

Miguel and Kremer [2004] analyze the effect of a de-worming drug on school performance of pupils in primary school in Kenya:

Intestinal worms affect one in four people worldwide and are particularly prevalent among school-age children in developing countries. While most have light infections, which may be asymptomatic, a minority have heavy infections, which can lead to iron-deficiency anemia, protein-energy malnutrition, abdominal pain, and apathy. Infected kids generally feel weak and tired most of the time: they have lower school attendance and struggle more to benefit from schooling

Low-cost single-dose oral therapies can kill the worms in a very effective way, but reinfection is rapid (unless there is change in the surrounding environment)

Contagion can occur through contact - while bathing and fishing - with infected water or when sharing the same sanitary facilities with infected individuals

Hence, my probability of contagion increases with the number of infected individuals I interact with

An intervention that reduces the infection among a specific population will have a positive impact also on all other untreated individuals who interact with the treated individuals

This is a **positive spillover**, that should be taken into account when evaluating the policy intervention

Suppose we are evaluating the impact of a deworming intervention on students’ grades

- Suppose deworming children increases the average grades of treated pupils by 10 percentage points and of untreated pupils by 2 percentage points (due to positive spillover effects)
- ATT is 10 percentage points.
- However, if we simply compare the grades of treatment and control pupils, we will observe only an 8 percentage point increase.

Miguel and Kremer [2004] evaluate the effect on health, school absenteeism, and test scores of a Kenyan programme where randomization occurred at the school level

Overall 75 schools were randomly allocated to 3 groups and the health intervention was phased in sequentially:

- group 1: received treatment in 1998 and 1999
- group 2: received treatment in 1999
- group 3: received treatment in 2001

Treatment and control group vary over time:

Year	Treatment Group	Control Group
1998	group 1	group 2 and 3
1999	group 1 and 2	group 3

They study:

1. **within-school externalities**: in treated school, some kids were not treated (generally, because they were absent on the day of treatment). Did they benefit from the policy?
2. **across-school externalities**: did untreated schools that were closer to treated school (hence, their pupils lived in the same area) benefit from the treatment?

They find evidence of both types of externalities and a positive policy effect on both health and school attendance (but not on test scores)

With positive spillovers we tend to underestimate the effect of the policy

The opposite is true with negative spillovers (upward bias), which may happen in any case where the treatment produces some kind of “displacement effect” on the untreated (for instance, a cash-transfer programme which induces a rise in food prices by increasing treated individuals’ demand for food)

Conditional cash transfer (CCT) programs aim to reduce poverty by making welfare programs conditional upon the receivers' actions. The government (or a charity) only transfers the money to persons who meet certain criteria.

The externality is part of the policy effect: we want to estimate it

See Duflo et al. [2008] for a brief discussion of approaches to deal with these externalities

3) External validity

Three main issues:

1. Partial and general equilibrium effects.
2. Hawthorne and John Henry effects
3. Generalizing the results (beyond specific programs and samples)

1) Partial and general equilibrium effects

Randomized evaluations always focus on specific areas. They are not able to pick up general equilibrium effect, which may become particularly relevant for assessing the implication of scaling up a program.

Suppose we evaluate a program that randomly assigns vouchers to attend private schools to students in one town and we find that treated pupils (i.e. those receiving the voucher) are more likely to attend a private school

- What would happen if we were giving the voucher to everyone in the city?
- Would they all go to private schools?
- What would happen to public schools?
- Would public school improve their performance to attract students?
- Would they instead shut down (forcing everyone to go to private schools)?

General equilibrium effects can be thought of as another type of externality

2) Hawthorne and John Henry effects

Is the evaluation itself changing the behavior of treatment and/or comparison group?

- **Hawthorne effects**: changes in behavior among the treatment group in response to the fact that they are part of an experiment and not to the treatment itself.
- **John Henry effects**: changes in behavior among the control group who feels "excluded" from the policy.

Main problem: these possible changes in behavior are triggered by the experiment itself and not by the policy.

They will disappear once the experimental phase is concluded and the policy is actually implemented: there may be important differences in the short-run effect of an experimental evaluation and the long-run ones of the real program.

[Insegnanti che sanno di essere in un esperimento e migliorano la loro didattica]

3) Generalizing the results (beyond specific programs and samples)

Three major issues:

1) Is the experimental program implemented with a particularly high level of care that will be then impossible to replicate once the policy is extended to a wider audience? Will the quality of the program substantially deteriorate from the experimental phase to the actual implementation?
2) Can we conclude that because one population responded to a program in one way, another population will respond in the same way to a similar program?
3) Given that a specific version of a program had a given impact, what can we learn about similar, but not identical, programs?

Reading list

Compulsory readings:

- my lecture slides
- Angrist and Pischke [2015]: chapter 1

LECT 4: REGRESSION DISCONTINUITY DESIGN

Table of contents

Regression Discontinuity Design (RDD)

This method applies to cases where the probability of assignment to the treatment group is a discontinuous function of one or more observable variables.

But what does it mean discontinuous function?

For instance: a tuition fees waiver (esenzione) is given to all the students who get a grade of at least 75/100 in a test; unemployed workers are eligible for a training program only if they have not turned 20 yet; etc.

To interpret the results from a RDD one needs to be sure that there is ONLY ONE policy (e.g. tuition fees waiver) changing at that threshold - e.g. turning 16/18 implies a large number of changes in policy for individuals

Suppose the observable variable X determines the assignment, x_0 is a threshold, and all the individuals with values of X above (or below) the threshold are eligible for the treatment.

We have a potential setting to implement a RDD if the probability of receiving the treatment jumps in the neighbourhood of the threshold value x0:

$$Pr(D = 1|X = x_0 + \varepsilon) \neq Pr(D = 1|X = x_0 - \varepsilon) \quad \forall \varepsilon > 0$$

Throughout our discussion of RDD, we will assume - without any loss of generality - that:

$$Pr(D = 1|X = x_0 + \varepsilon) > Pr(D = 1|X = x_0 - \varepsilon) \quad \forall \varepsilon > 0$$

Where $x_0 + \varepsilon$ *is above the threshold and* $x_0 - \varepsilon$ *is below the threshold.*

Although RDD is generally referred to as a method, it would be more precise to define it as a "description of a particular data generating process" (Lee and Lemieux [2010], p. 285).

In order to obtain a RDD one needs to have:

1) a **threshold** value of some observable characteristics which (fully or partially) determines the assignment of individuals to treatment;
2) the individuals cannot be able to manipulate the assignment variable and **precisely sort** above or below the threshold.

The key word here is "**precisely**": on average, students will probably try to get a grade as high as possible, but they cannot have full control of the exact grade they will get at the end (i.e. of their exact position with respect to the threshold).

The important consequence of having imprecise control over the assignment variable is that **the treatment in a neighborhood of the threshold is "as good as randomized**".

One can consider a RDD as a **local randomized experiment**.

For the RDD to be valid we always need to check that the assignment/running variable has a smooth distribution around the threshold.

Suppose we are considering an admission threshold based on a test score:

- We do not want to see that all - or a strangely large number of - students managed to get a grade exactly equal to the admission threshold (perfect control)
- Neither we want to see that everyone is above the threshold or that there is a mass of students just above the threshold

These would all be signs of manipulation.

FIGURE: MANIPULATION AROUND THE THRESHOLD? - LEE AND LEMIEUX [2010]

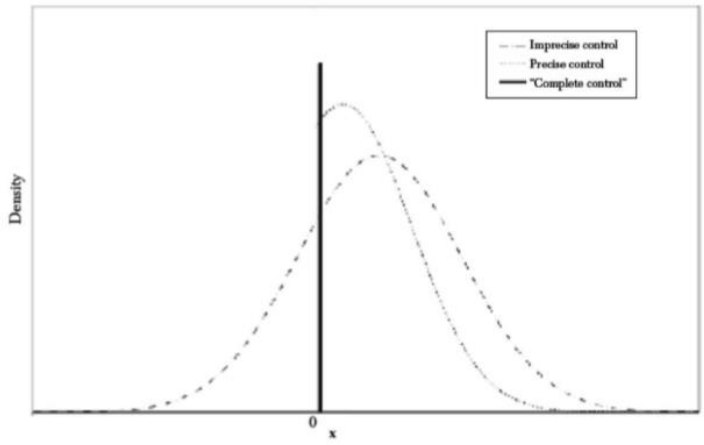


Figure 4. Density of Assignment Variable Conditional on $W = w$, $U = u$

If the impact of any unobservable variable correlated with the variable used to assign treatment is smooth, the following assumption **(continuity of** Y_0) is reasonable for any small $\varepsilon > 0$:

$$E(Y_{0i}|D_i = 1, x_0 - \varepsilon < X_i < x_0 + \varepsilon) = E(Y_{0i}|D_i = 0, x_0 - \varepsilon < X_i < x_0 + \varepsilon)$$

Identifying assumption of the RD: in the absence of the treatment, the outcomes of the individuals just above and just below the threshold value x_0 would have been equal

This implies that there is no selection bias in a sufficiently small neighborhood of x_0.

Idea of RD: under this identifying assumption, one can then estimate the treatment effect by **comparing individuals just above the threshold (treated) with those just below (untreated)**

Depending on the size of the discontinuity in the probability of treatment before and after the threshold x_0, we have two types of RDD:

1. **Sharp RDD**:
   $Pr(D = 1|X = x_0 + \varepsilon) - Pr(D = 1|X = x_0 - \varepsilon) = 1 \quad \forall \varepsilon > 0$
2. **Fuzzy RDD**:
   $0 < Pr(D = 1|X = x_0 + \varepsilon) - Pr(D = 1|X = x_0 - \varepsilon) < 1 \quad \forall \varepsilon > 0$

FIGURE: LINEAR RDD - LEE AND LEMIEUX [2010]

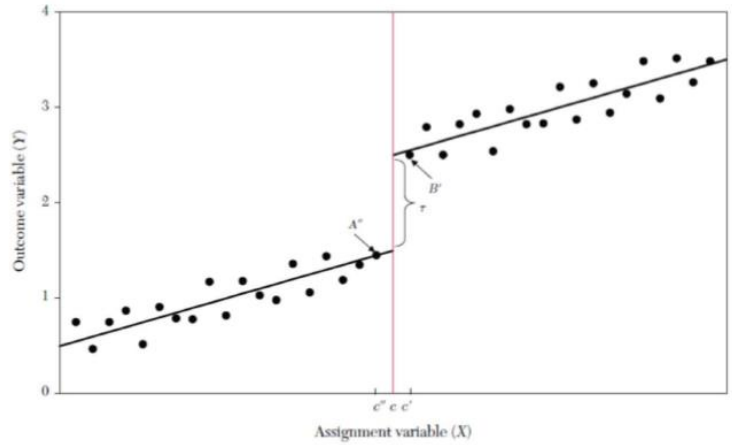


FIGURE: NON-LINEAR RDD - LEE AND LEMIEUX [2010]

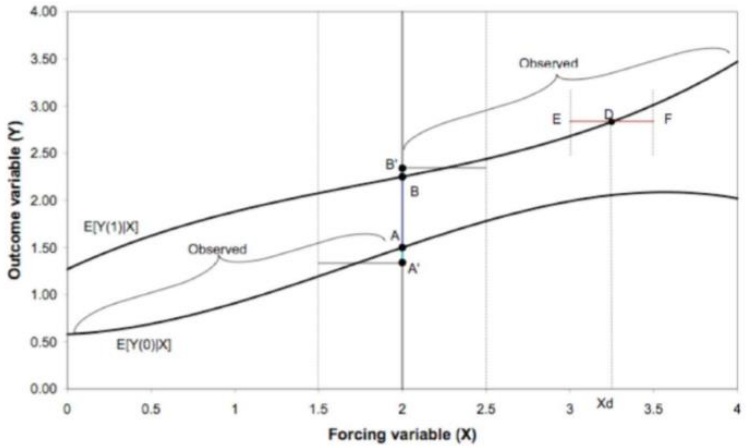
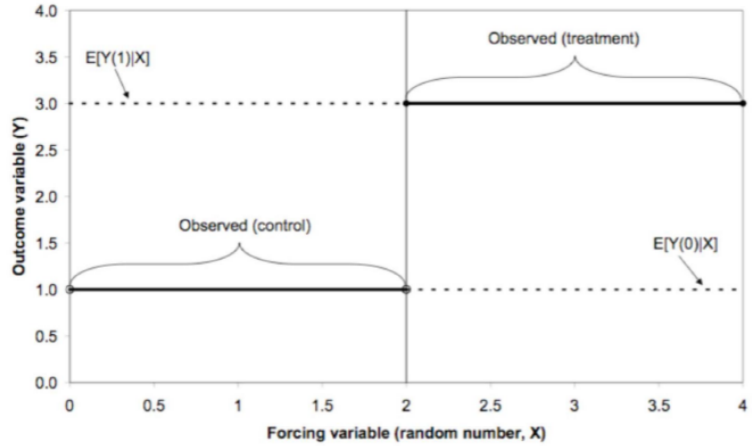


FIGURE: RANDOMIZED EXPERIMENT AS RDD - LEE AND LEMIEUX [2010]



Sharp RDD

The treatment D is a deterministic and discontinuous function of an observed covariate X
The policy is mandatory. Participation into treatment jumps from zero to one as x crosses the threshold x_0:

$$D(x_i) = \begin{cases} 1 & \text{if} \quad X_i \geq x_0 \\ 0 & \text{if} \quad X_i < x_0 \end{cases} \qquad (1)$$

Sharp RDD - Assumptions

1) Sharp RD :

$$Pr(D = 1|X = x_0 - \varepsilon) = 0 \text{ and } Pr(D = 1|X = x_0 + \varepsilon) = 1 \ \forall \varepsilon > 0 \quad (RDD-1)$$

2) Continuity:

$$E(Y_0|X) \text{ is continuous at } X = x_0 \qquad (RDD - 2.1)$$

FIGURE: RDD SHARP DESIGN - IMBENS AND LEMIEUX [2008]

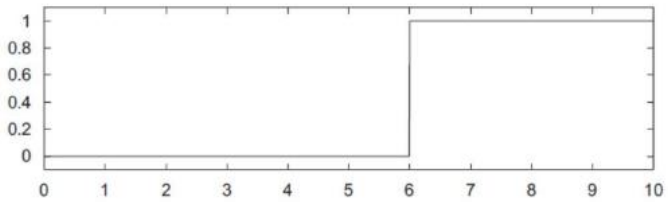


Fig. 1. Assignment probabilities (SRD).

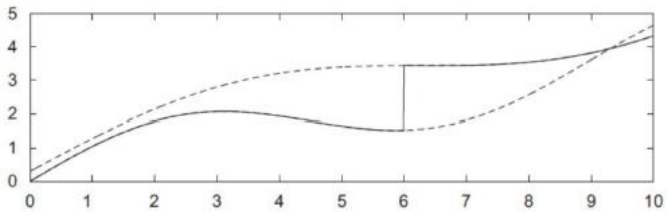


Fig. 2. Potential and observed outcome regression functions.

Sharp RDD - Identification

The treatment effect is estimated by comparing individuals just above the threshold (treated) with those just below (untreated):

$$RDD = \lim_{x_i \to x_0^+} E(Y_i|x_i) - \lim_{x_i \to x_0^-} E(Y_i|x_i) = \lim_{X_i \to x_0^+} E(\beta_i|X_i) = ATT(X_i \to x_0^+)$$

Therefore, the RDD approach identifies the average impact for subjects in the right neighborhood of x_0 (only treated subjects): this is a **Local Average Treatment effect on Treated**.

In order to identify the **Local Average Treatment Effect** at the threshold point x_0, we need to make the additional **assumption of continuity of** $E(Y_1|X)$:

$$E(Y_1|X) \text{ is continuous at } X = x_0 \qquad (RDD - 2.2)$$

Together with assumption RDD-2.1, this implies that $E(\beta \mid X)$ is continuous at $X = x_0$. Therefore:

$$RDD = \lim_{X_i \to x_0^+} E(Y_i|X_i) - \lim_{X_i \to x_0^-} E(Y_i|X_i) =$$
$$= \lim_{X_i \to x_0^+} E(\beta_i|X_i) = E(\beta_i|X_i = x_0) = ATE(X_i = x_0)$$

Moreover, assumption RD-2.2 allows to identify the average impact also for subjects in the left neighbourhood of x_0, the **Local Average Treatment Effect on Non-Treated**:

$$ATNT(X_i \to x_0^-).$$

In general, it is quite difficult to think of practical cases where the continuity condition is satisfied for $E(Y_0|X)$ but not for $E(Y_1|X)$. Therefore, a RDD will generally allow to identify **Local Average Treatment Effect** at the threshold point x_0.

Fuzzy RDD

The fuzzy RDD allows for a **smaller jump in the probability of assignment to the treatment at the threshold.** The difference in probability of treatment does not need to be one as in the sharp RDD, as long as:

$$\lim_{X_i \to x_0^+} Pr(D_i = 1|X_i) \neq \lim_{X_i \to x_0^-} Pr(D_i = 1|X_i)$$

This is the case in any situation where, for instance, the eligibility for the treatment is a discontinuous function of some observable variable X , but the policy is not mandatory.

Since the probability of treatment jumps by less than one at the threshold, the jump in the relationship between Y and X can no longer be interpreted as an average treatment effect.

FIGURE: RDD FUZZY DESIGN - IMBENS AND LEMIEUX [2008]

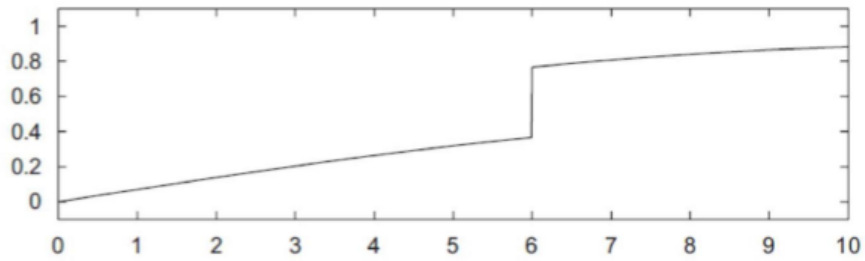
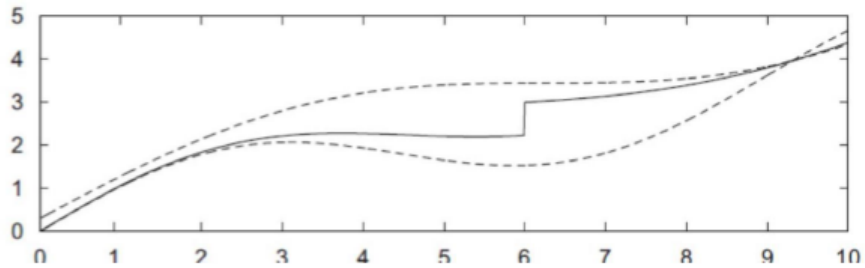


Fig. 3. Assignment probabilities (FRD).



The Average Treatment Effect at the threshold point x_0 can be recovered by dividing the jump in the relationship between Y and X at x_0 by the fraction induced to take-up the treatment at the threshold:

$$E(\beta_i|X_i = x_0) = \frac{\lim_{X_i \to x_0^+} E(Y_i|X_i) - \lim_{X_i \to x_0^-} E(Y_i|X_i)}{\lim_{X_i \to x_0^+} Pr(D(X_i) = 1|X_i) - \lim_{X_i \to x_0^-} Pr(D(X_i) = 1|X_i)}$$

- Notice that the sharp design is just a special case of the fuzzy design: if $\lim_{X_i \to x_0^+} Pr(D = 1|X_i) = 1$ and $\lim_{X_i \to x_0^-} Pr(D = 1|X_i) = 0$, the denominator in the expression above would be equal to 1.

Implementation of RDD

- RDD identification strategy is valid in a small neighborhood of the threshold x_0 but, usually, sample size is too small if one works with observations in a truly small neighborhood of x_0
- Trade-off between internal consistency of the method and feasibility
- Practice: choices of neighborhood; assumptions about the regression curve away from x_0; differential weighting of observations depending on their distance from x_0
- See Lee and Lemieux [2010] for details on the implementation of RDD

Validity of RDD

Three important questions about RDD (Lee and Lemieux [2010])

1) How do I know whether an RDD is appropriate for my context? When are the identification assumptions plausible or implausible?

"When there is a continuously distributed stochastic error component to the assignment variable - which can occur when optimizing agents do not have precise control over the assignment variable - then the variation in the treatment will be as good as randomized in a neighborhood around the discontinuity threshold."

2) Is there any way I can test those assumptions?

"Yes. As in a randomized experiment, the distribution of observed baseline covariates should not change discontinuously at the threshold." In other words, the only "jump" one should observe is in the probability of treatment: plotting all observables characteristics around the threshold is a very informative test.

3) To what extent are results from RDDs generalizable?

"The RD estimand can be interpreted as a weighted average treatment effect, where the weights are the relative ex ante probability that the value of an individuals assignment variable will be in the neighborhood of the threshold."

Strengths and weaknesses of RDD

Strengths:

- the closest approach to a randomized experiment - it is a local randomized experiment - among all other alternatives;
- requires fairly mild assumptions;
- One needs not assume the RDD isolates treatment variation that is as good as randomized: such randomized variation is a consequence of agents' inability to precisely control the assignment variable near the known cutoff (Lee [2008]).

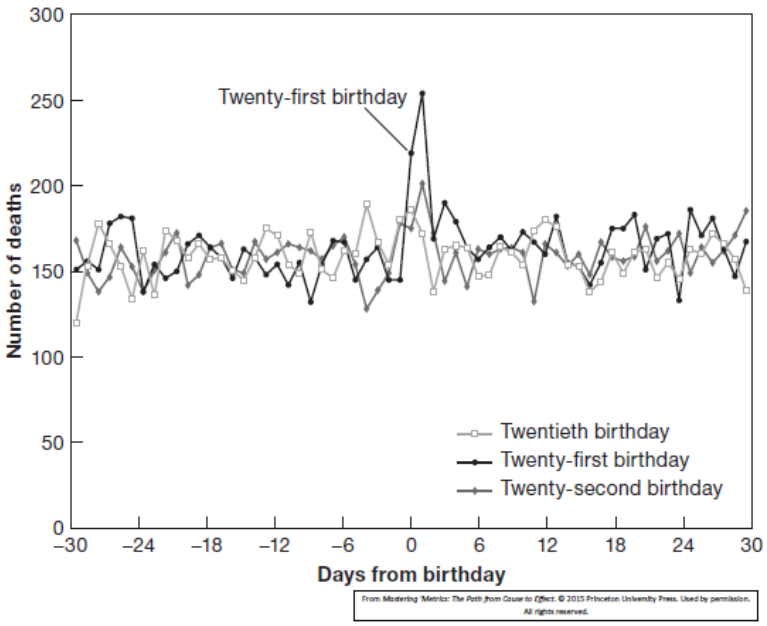
Weaknesses:

- The RD share an important limitation with the randomized trial experiments: "All other things (topic, question, and population of interest) equal, we as researchers might prefer data from a randomized experiment or from an RDD.(...) But in reality, like the randomized experiment (...) an RDD will simply not exist to answer a great number of questions." (Lee and Lemieux [2010], p. 285).
- Given the discontinuity design, the average parameter is only identifiable at a given point x_0 of the distribution of x .

Example: Birthdays and Funerals

- In the US, the minimum legal drinking age (MLDA) is set at age 21
- Would it be a good idea to lower the MLDA to age 18?
- Supporters of this idea say that lowering the MLDA would promote a culture of mature alcohol consumption
- Opponents of this idea answer that keeping the MLDA at age 21
- reduces youth access to alcohol, preventing some harm
- But, do Americans who turn 21 engage in dangerous excessive alcohol consumption?
- We look at Americans aged 20-22 between 1997 and 2003

FIGURE 4.1 BIRTHDAYS AND FUNERALS



1997 and 2003. Deaths here are plotted by day, relative to birthdays, which are labeled as day 0. For example, someone who was born on September 18, 1990, and died on September 19, 2012, is counted among deaths of 22-year-olds occurring on day 1.

Mortality risk shoots up on and immediately following a twenty-first birthday, a fact visible in the pronounced spike in daily deaths on these days. This spike adds about 100 deaths to a baseline level of about 150 per day. The age-21 spike doesn't seem to be a generic party-hardy birthday effect. If this spike reflects birthday partying alone, we should expect to see deaths shoot up after the twentieth and twenty-second birthdays as well, but that doesn't happen. There's something special about the twenty-first birthday. It remains to be seen, however, whether the age-21 effect can be attributed to the MLDA, and whether the elevated mortality risk seen in Figure

4.1 lasts long enough to be worth worrying about.

- Yes, mortality jumps up in the days immediately following the 21st birthday
- And it is not a generic effect of partying hard at each birthday: there is no jump at age 20 and 22
- Is that jump due to the fact that young Americans turning 21 suddenly gain access to alcohol and start heavily using (and abusing) it?

- Carpenter and Dobkin [2009] analyse this question in a **sharp RD design**
- The treatment is having legal access to alcohol

- The probability of having legal access to alcohol is zero until the day before turning 21 and it jumps to 1 on the day of the 21st birthday
- (which does not mean that individuals younger than 21 couldn't get any alcohol in some illegal way)
- The running variable is age

We can estimate the following simple equation:

$$M_a = \alpha + \rho D_a + \gamma a + e_a$$

where:

M_a is the death rate in month a (where month is defined as a 30-day interval counting from the 21st birthday

D_a is the treatment dummy (equal one after turning 21)

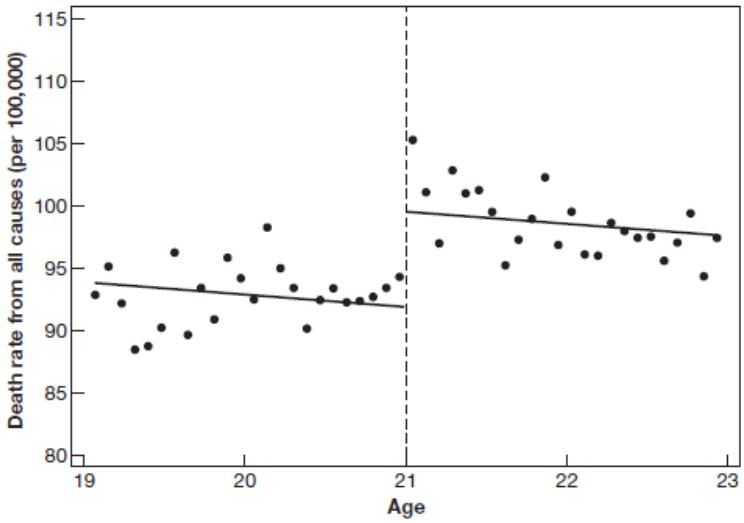
a is a linear control for age in months

If we run this regression we find ρ = 7.7

Average death rates are around 95: hence we find a substantial 8% increase that we can attribute to the MLDA

We also find γ < 0 reflecting the smooth decline in death rate among young people as they mature

FIGURE 4.2 A SHARP RD ESTIMATE OF MLDA MORTALITY EFFECTS



Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the minimum legal drinking age (MLDA) cutoff).

The story linking the MLDA with a sharp and sustained rise in death rates is told in Figure 4.2. This figure plots death rates (measured as deaths per 100,000 persons per year) by month of age (defined as 30-day intervals), centered around the twenty-first birthday. The X-axis extends 2 years in either direction, and each dot in the figure is the death rate in one monthly interval. Death rates fluctuate from month to month, but few rates to the left of the age-21 cutoff are above 95. At ages over 21, however, death rates shift up, and few of those to the right of the age-21 cutoff are below 95. Happily, the odds a young person dies decrease with age, a fact that can be seen in the downward-sloping lines fit to the death rates plotted in Figure 4.2. But extrapolating the trend line drawn to the left of the cutoff, we might have expected an age-21 death rate of about 92; in the language of Chapter 1,

We can estimate a more complex model to allow for non-linearities

For instance, we can include a quadratic running variable control:

$$M_a = \alpha + \rho D_a + \gamma_1 a + \gamma_2 a^2 + e_a$$

Alternatively, we can allow for different running variable coefficients to the left and to the right of the cutoff by interacting a with D_a.

To make the model easier to interpret, we center the running variable around the cutoff a_0 (i.e. age 21):

$$M_a = \alpha + \rho D_a + \gamma(a - a_0) + \delta[(a - a_0)D_a] + e_a$$

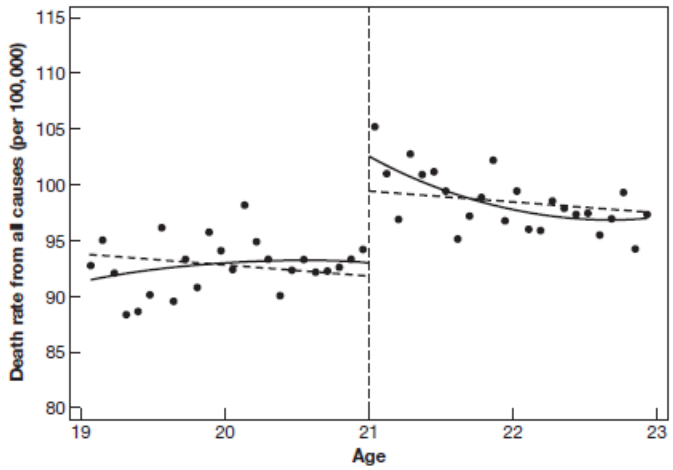
Nonlinear trends and changes in slope at the cutoff can also be combined:

$$M_a = \alpha + \rho D_a + \gamma_1(a - a_0) + \gamma_2(a - a_0)^2 +$$

$$+\delta_1[(a - a_0)D_a] + \delta_2[(a - a_0)^2 D_a] + e_a$$

If we estimate this latter regression equation we find a slightly larger effect at the cutoff than the linear model, equal to about 9.5 deaths per 100,000.

FIGURE 4.4 QUADRATIC CONTROL IN AN RD DESIGN



Notes: This figure plots death rates from all causes against age in months. Dashed lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months. The solid lines plot fitted values from a regression of mortality on an over-21 dummy and a quadratic in age, interacted with the over-21 dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

hand, when the trend relationship between running variable and outcomes is approximately linear, limited extrapolation seems justified. The jump in death rates at the cutoff shows that drinking behavior responds to alcohol access in a manner that is reflected in death rates, an important point of principle, while the MLDA treatment effect extrapolated as far out as age 23 still looks substantial and seems believable, on the order of 5 extra deaths per 100,000. This pattern highlights the value of “visual RD,” that is, careful assessment of plots like Figure 4.4.

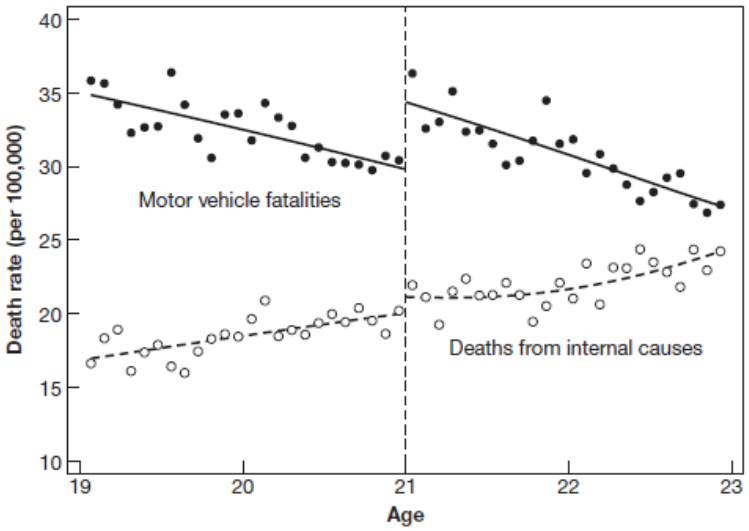
How convincing is the argument that the jump in Figure 4.4 is indeed due to drinking? Data on death rates by cause of death help us make the case. Although alcohol is poisonous, few people die from alcohol poisoning alone, and deaths from

- Which model is the correct one? which estimate should we believe? Plus 7.7. or plus 9.5?

- There is no rule. But we want to see that our results are not too sensitive to the specification we choose
- In this case, the two estimates are slightly different but they lead exactly to the same conclusion
- We should be worried, instead, if the simple liner model delivered a positive significant coefficient and the more complex one delivered a non-significant one
- In this latter case, we should be very cautious in our conclusions (i.e. the effect may or may not be there; it is not robust to alternative specifications)

- How convincing is the argument that the jump in death rate is actually due to drinking?
- It is always very useful to run some **falsification exercises**
- Think about a situation where you would not expect to observe an effect: if you still find it in the data, there may be something wrong
- In this case, data on the cause of death may help us.
- A sudden increase in alcohol consumption may lead to liver disease in the medium-long run, but it does not immediately lead to death
- Drunk driving, instead, can immediately kill
- We should observe an increase in motor vehicle fatalities, but not in deaths due to “internal causes” (e.g. cirrhosis, cancer, etc.)
- If all types of deaths increase at age 21, there may be a genetic Explanation for it (e.g. individuals who turn 21 “expire”...)

FIGURE 4.5 RD ESTIMATES OF MLDA EFFECTS ON MORTALITY BY CAUSE OF DEATH



Notes: This figure plots death rates from motor vehicle accidents and internal causes against age in months. Lines in the figure plot fitted values from regressions of mortality by cause on an over-21 dummy and a quadratic function of age in months, interacted with the dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

“on points close to the cutoff. For the small set of points close to the boundary, nonlinear trends need not concern us at all. This suggests an approach that compares averages in a narrow window just to the left and just to the right of the cutoff. A drawback here is that if the window is very narrow, there are few observations left, meaning the resulting estimates are likely to be too imprecise to be useful. Still, we should be able to trade the reduction in bias near the boundary against the increased variance suffered by throwing data away, generating some kind of optimal window size. The econometric procedure that makes this trade-off is non-

parametric RD. Nonparametric RD amounts to estimating equation (4.2) in a narrow window around the cutoff. That is, we estimate”

TABLE 4.1 SHARP RD ESTIMATES OF MLDA EFFECTS ON MORTALITY

TABLE 4.1
Sharp RD estimates of MLDA effects on mortality

Dependent variable	Ages 19–22		Ages 20–21	
	(1)	(2)	(3)	(4)
All deaths	7.66	9.55	9.75	9.61
	(1.51)	(1.83)	(2.06)	(2.29)
Motor vehicle accidents	4.53	4.66	4.76	5.89
	(.72)	(1.09)	(1.08)	(1.33)
Suicide	1.79	1.81	1.72	1.30
	(.50)	(.78)	(.73)	(1.14)
Homicide	.10	.20	.16	−.45
	(.45)	(.50)	(.59)	(.93)
Other external causes	.84	1.80	1.41	1.63
	(.42)	(.56)	(.59)	(.75)
All internal causes	.39	1.07	1.69	1.25
	(.54)	(.80)	(.74)	(1.01)
Alcohol-related causes	.44	.80	.74	1.03
	(.21)	(.32)	(.33)	(.41)
Controls	age	age, age^2, interacted with over-21	age	age, age^2, interacted with over-21
Sample size	48	48	24	24

Notes: This table reports coefficients on an over-21 dummy from regressions of month-of-age-specific death rates by cause on an over-21 dummy and linear or interacted quadratic age controls. Standard errors are reported in parentheses.

“of potentially misleading nonlinear trends. At the same time, there isn’t much of a jump in deaths due to internal causes, while the standard errors in Table 4.1 suggest that the small jump in internal deaths seen in the figure is likely due to chance. In addition to straightforward regression estimation, an approach that masters refer to as parametric RD, a second RD strategy exploits the fact that the problem of distinguishing jumps from nonlinear trends grows less vexing as we zero in”

Reading list

Compulsory readings:

- my lecture slides
- Angrist and Pischke [2015]: chapter 4

Suggested reading:

- Pinotti [2017]

LECT 5 : DIFFERENCE IN DIFFERENCES

Table of contents I

Natural experiment

Natural experiments can be exploited to achieve identification of treatment effects.

Natural experiment creates (arguably) exogenous changes in the probability of assignment to treatment of individuals.

The researcher can then try to asses whether this exogenous change in treatment probability (or treatment intensity) has caused any statistically significant change in the outcome of interest.

Some “classical” examples

- We start by looking at some “classical” examples of natural experiment used to identify policy effects
- Recall (see lecture 3) that natural example do not need to be “natural”, i.e. decided by Nature (earthquakes, floods, etc.)
- Governments - and even private agents - can “generate” natural experiment
- Let’s discuss a couple of “classical” examples

Cholera in London in 1850

- The father of DID? ?
- John Snow (15 March 1813 - 16 June 1858) was an English physician and a leader in the adoption of anaesthesia and medical hygiene. He is considered to be one of the fathers of modern epidemiology, because of his work in tracing the source of a cholera outbreak in Soho, England, in 1854.
- Snow believed that cholera is transmitted by contaminated drinking water and not by “bad air”, as the prevailing (at the time) miasma theory held
- He was right, and to prove it, he exploited a natural experiment

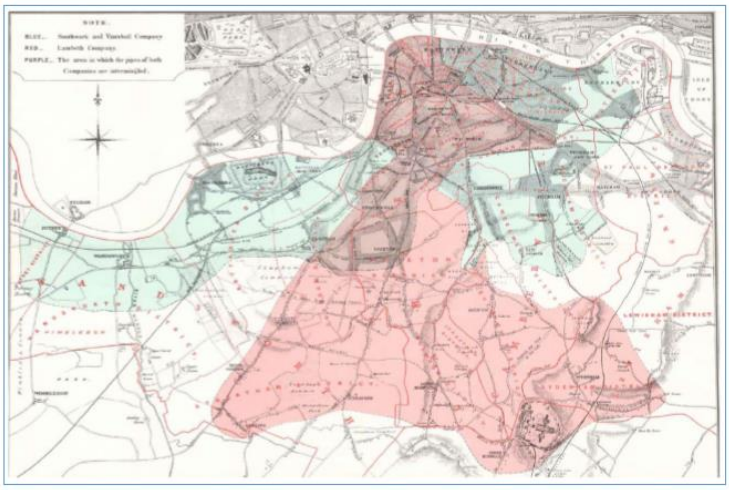
In the 1850s, in London the water was supplied to households by competing private companies

Different companies would supply different areas, but sometimes different companies supplied households in the same street

In south London, there were two main companies: Lambeth Company and Southwark and Vauxhall Company

In 1849, both companies would obtain their water directly from the sewage-contaminated Thames; but in 1852, Lambeth Company moved its water source 22 miles upstream (away from London's sewage)

FIGURE: LONDON COLERA 1850



In 1853/54 cholera outbreak

Death Rates per 10000 people by water company: Lambeth: 10; Southwark and Vauxhall: 150

The difference (+140) may be due to the different quality of the water but also to any other factor which differs in the two areas (e.g. "bad air")

Snow compared death rates in a previous epidemic in 1849: Lambeth: 150; Southwark and Vauxhall: 125

In 1849, there were less deaths in Southwark and Vauxhall than in Lambeth (-25): we can exclude the possibility that Lambeth has always been a better place to live

If anything, before the change in water supply, Lambeth was worse than Southwark and Vauxhall

The difference between the differences in number of deaths in 1849 and in 1850 (-165) can be considered the effect of the change in water supply

The assumption is that - in the absence of the change in water supply - Lambeth would still have had a slightly larger number of deaths

	1849	1853/54	Difference
Lambeth	150	10	-140
Southwark and Vauxhall	125	150	25
Difference	-25	140	-165

The employment effect of minimum wages

- Minimum wages are set to protect workers in low-pay occupations from exploitation and to guarantee them minimum living standards
- The concern, however, is that setting minimum wages above the market wage could induce employers to fire these workers
- Rather than being beneficial for low paid workers, a minimum wage could make them worse off
- Do minimum wages lead to higher employment?
- How can we empirically answer this question?

Minimum wages are usually introduced in an entire state and affect all workers in low-pay occupations

Before the introduction of a minimum wage none of the workers is “treated” and after the introduction they are all “treated”

Who can we compare with whom? That is, which treatment and control groups can we find in this case?

We could compare average wages (for low pay occupations) before and after the policy introduction

This is called a before-after estimator

But, economic conditions may have changed - and for exogenous reasons - before and after the minimum wage reform: for instance, the state may have entered a recession after the policy introduction

The recession is not caused by the policy change, it is driven by international causes (e.g. a financial crisis), but it happens at the same time as the policy change

How can we distinguish the minimum wage effect from the effect of the recession (or of any other event contemporaneous to the policy change)?

Unfortunately, we cannot.

Instead, we could try and compare wages of workers in another state where a minimum wage was not introduced but that was exposed to similar economic conditions (e.g. equally hit by the recession)

Hence, we need to choose a state which is sufficiently similar to the state we are studying (using the US and Luxembourg is probably not a good strategy...)

If we can find a good comparison state (or group of workers) we can look at the change in average wages in that state and compare it to the change observed in the treatment state (where the minimum wage was actually introduced)

? - and ? - analyze the employment effects in the fast-food industry of a substantial increase of minimum wage in New Jersey (NJ)

In 1992, New Jersey’s minimum wage rose from $4.25 to $5.05 per hour

They use eastern Pennsylvania (PA) - where no minimum wage policy change took place in the period considered - as a control group

Alternatively, they use within New Jersey comparisons between initially high-wage fast food stores (those paying more than the new minimum rate prior to its effective date) and low-wage stores.

They ran a telephone survey of fast food stores in New Jersey and Eastern Pennsylvania

The first wave of the survey was run in Feb-Mar 1992, a month before the scheduled increase in minimum wage in New Jersey

The second wave was run in Nov-Dec 1992, about eight months after the minimum wage increase

They chose fast food stores because they tipically and mainly provide low-pay jobs

Later on, they managed to get payroll data from the Bureau of Labor Statistics and they used these data to test the validity of their findings using the telephone survey data, including some additional counties in PA (?)

FIGURE: TREATMENT AND CONTROL AREAS: ?

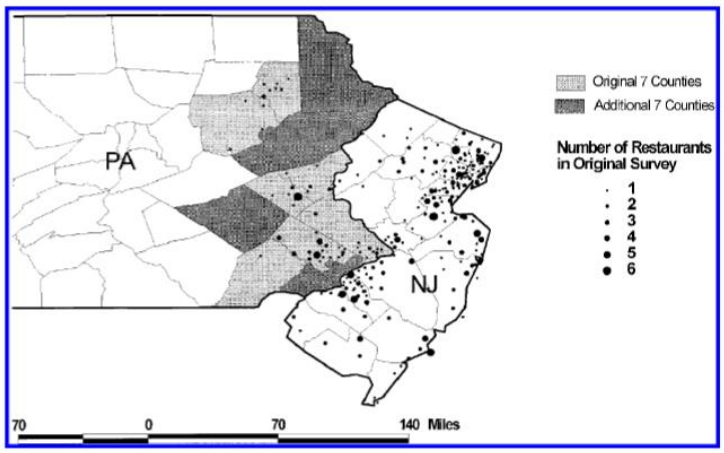


FIGURE 1. AREAS OF NEW JERSEY AND PENNSYLVANIA COVERED BY ORIGINAL SURVEY AND BLS DATA

FIGURE: MEANS OF KEY VARIABLES

TABLE 2—MEANS OF KEY VARIABLES

Variable	Stores in: NJ	PA	t^a
1. *Distribution of Store Types (percentages):*			
a. Burger King	41.1	44.3	−0.5
b. KFC	20.5	15.2	1.2
c. Roy Rogers	24.8	21.5	0.6
d. Wendy's	13.6	19.0	−1.1
e. Company-owned	34.1	35.4	−0.2
2. *Means in Wave 1:*			
a. FTE employment	20.4 (0.51)	23.3 (1.35)	−2.0
b. Percentage full-time employees	32.8 (1.3)	35.0 (2.7)	−0.7
c. Starting wage	4.61 (0.02)	4.63 (0.04)	−0.4
d. Wage = $4.25 (percentage)	30.5 (2.5)	32.9 (5.3)	−0.4
e. Price of full meal	3.35 (0.04)	3.04 (0.07)	4.0
f. Hours open (weekday)	14.4 (0.2)	14.5 (0.3)	−0.3
g. Recruiting bonus	23.6 (2.3)	29.1 (5.1)	−1.0
3. *Means in Wave 2:*			
a. FTE employment	21.0 (0.52)	21.2 (0.94)	−0.2
b. Percentage full-time employees	35.9 (1.4)	30.4 (2.8)	1.8
c. Starting wage	5.08 (0.01)	4.62 (0.04)	10.8
d. Wage = $4.25 (percentage)	0.0	25.3 (4.9)	—
e. Wage = $5.05 (percentage)	85.2 (2.0)	1.3 (1.3)	36.1
f. Price of full meal	3.41 (0.04)	3.03 (0.07)	5.0
g. Hours open (weekday)	14.4 (0.2)	14.7 (0.3)	−0.8
h. Recruiting bonus	20.3 (2.3)	23.4 (4.9)	−0.6

FIGURE: WAGE DISTRIBUTION BEFORE THE REFORM; ?

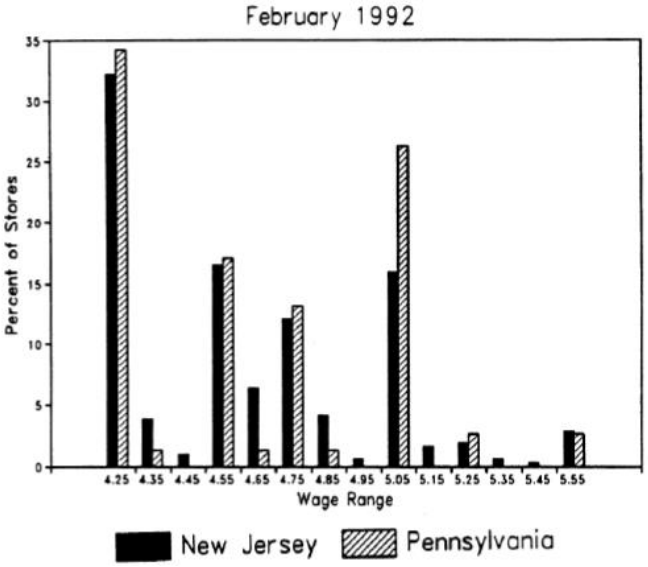
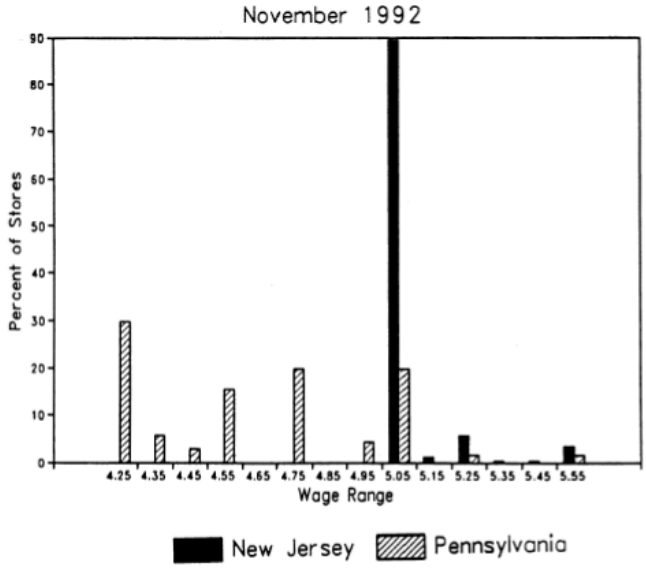


FIGURE: WAGE DISTRIBUTION AFTER THE REFORM; ?



- The table and figures show that the stores in the two states had a very similar wage distribution before the policy change (see “means in wave 1”)
- But the distribution dramatically changed in NJ after the increase in minimum wage (while it did it not change in PA)
- This is the obvious (mechanical) effect of the policy
- It is obvious because there is no doubt that the introduction of a compulsory minimum wage will increase minimum wages paid by firms
- The uncertainty is about the effect on employment

The comparisons of changes in employment in treatment and controls stores (i.e. the difference in differences) are reported in the table below (table 3 in ?)

- Row 1 reports average Full Time Equivalent (FTE) employment before the the policy change: in wave 1, average employment was 23.3 full-time equivalent workers per store in PA, compared with an average of 20.4 in New Jersey.
- Row 2 reports average Full Time Equivalent (FTE) employment after the the policy change: in wave 2, average employment was 21.2 full-time equivalent workers per store in PA, compared with an average of 21 in New Jersey.
- Row 3 reports the difference and the difference-in-differences: FTE employment dropped by 2.2 units in PA while increased by 0.59 units in NJ
- The DiD is 2.76 (plus 13 %): despite the increase in wages, full-time equivalent employment increased in NJ relative to PA

FIGURE: DID ESTIMATES ?

TABLE 3—AVERAGE EMPLOYMENT PER STORE BEFORE AND AFTER THE RISE IN NEW JERSEY MINIMUM WAGE

	Stores by state			Stores in New Jersey[a]			Differences within NJ[b]	
Variable	PA (i)	NJ (ii)	Difference, NJ − PA (iii)	Wage = $4.25 (iv)	Wage = $4.26–$4.99 (v)	Wage ≥ $5.00 (vi)	Low–high (vii)	Midrange–high (viii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	−2.69 (1.37)	−2.17 (1.41)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)	0.75 (1.27)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	−2.04 (1.14)	3.36 (1.48)	2.91 (1.41)
4. Change in mean FTE employment, balanced sample of stores[c]	−2.28 (1.25)	0.47 (0.48)	2.75 (1.34)	1.21 (0.82)	0.71 (0.69)	−2.16 (1.01)	3.36 (1.30)	2.87 (1.22)
5. Change in mean FTE employment, setting FTE at temporarily closed stores to 0[d]	−2.28 (1.25)	0.23 (0.49)	2.51 (1.35)	0.90 (0.87)	0.49 (0.69)	−2.39 (1.02)	3.29 (1.34)	2.88 (1.23)

- Row 4 and 5 are analogous to row 3 but row 4 uses only the subsample of stores that reported valid employment data in both waves, while row 5 treats stores that shut down in wave 2 as stores with employment = 0 (instead of missing data)
- Columns 1-3 use NJ stores as treatment group and PA stores as control group
- Columns 4-8 use only NJ stores and compare stores that - before the reform - had different starting wage levels (low, midrange, high)
- Within NJ comparison lead to very similar conclusions: see row 3 and columns 7 and 8
- In all cases, results from row 4 and 5 confirm results from row 3 (i.e. results are robust to changes in the sample)

Summarizing, following the increase in minimum wage in NJ employment decreased more in untreated stores than in treated ones (where it remained constant / increased slightly)

This may sound strange. What happened?

Economic conditions in this area were negative between 1991 and 1993, and unemployment was on an upward trend everywhere (i.e. employment was decreasing)

Introducing/increasing the minimum wage should reduce employment in a fully competitive labor market

These findings suggest that the labor market in the fast food industry is not competitive

As soon as you allow for some frictions in the labor market (e.g. there are costs associated to finding good employees for the employer and to find jobs for the workers) and/or for employers to have market power (monopsony power) in the labor market, then the introduction of a minimum wage can lead to an increase in employment

Obviously, there are frictions in the labour market and employers do have market power (especially in a low-pay sector such as the fast food industry)

Hence, results are not so surprising (more on this in a Labor Economics course)

Before-After (BA) estimator

Suppose we have a policy change which affects the entire population of interest: all individuals are untreated (treated) before the regime change and are treated (untreated) after that.

Given that treated individuals are in both states (treated and untreated) at different points in time, the idea of a before-after estimator is that of comparing the same individual over time.

BA takes the difference in mean outcomes for the group of treated individuals before and after the treatment occurs.

Taking the within individual difference allows to difference out any fixed individual component which may have determined the selection into treatment.

In other words, suppose “motivated” individuals choose to take the treatment (i.e. select into treatment) and those “not motivated” choose not to take the treatment (i.e. select out)

If we just have cross-sectional data, we can compare treated and untreated individuals.

This is the “*naive comparison*” we discussed before: the problem is that any difference in outcomes we may find may be (entirely) driven by the fact that we are comparing “motivated” and “non-motivated” individuals

- With longitudinal (panel) data, instead, we can just look at treated individuals, before and after the policy change
- We ignore the untreated (“non motivated”) individuals, and compare “motivated” individuals with themselves over time
- But, identification is complicated by the possibility that something else changed before and after the treatment

Identification

BA identifying assumptions:

1 participation depends on observable individual characteristics and on time invariant unobservable characteristics (ass. BA-1)

2 among the treated, the mean outcome in the non-treatment status in time T (after the treatment) and T − 1 (before the treatment) would have been the same:

$$E(Y_{0,T}|D=1) = E(Y_{0,T-1}|D=1) \qquad \text{(ass. BA-2)}$$

If, for instance, we are evaluating the effect of attending an elite college, these assumptions imply:

1) **BA-1**: the decision to select into that college is driven by test scores, city of residence, family income, etc. (= observables) and by the level of commitment/motivation/ability that do not change while in college (= time invariant unobservable characteristics)
2) **BA-2**: if not enrolled in that college, average earnings in the next n years would have been equal (controlling for the additional years of labor market experience) to those in the past n years

Formal identification.

$$BA = E(Y_{1,T} - Y_{0,T-1}|D=1) = E(Y_{1,T} - \underbrace{Y_{0,T} + Y_{0,T}}_{add\ and\ subtract} - Y_{0,T-1}|D=1)$$

$$= E(Y_{1,T} - Y_{0,T}|D=1) + \underbrace{E(Y_{0,T} - Y_{0,T-1}|D=1)}_{=0\ by\ ass.\ BA-2} =$$

$$= E(Y_{1,T} - Y_{0,T}|D=1) = ATT$$

- ATT can be estimated from the data as the difference in mean outcomes for the group of treated individuals before and after the treatment occurs (BA).
- Assumption BA-1 is violated when the unobservables change over time (for instance, while in college, become wiser, more reliable, more organized, etc., independently of the college effect)
- Assumption BA-2 is violated when, in the absence of treatment, the outcomes of the treated individuals would have changed between period T and T − 1.
- This can happen, for instance, if the outcome has some trend, or if some external condition which may influence the outcome (e.g. economic conditions) has changed between period T and T − 1.

These are serious threats to the “credibility” (= internal validity) of a before-after estimator

In particular, the effect of any relevant factor which changes between the before- and after- periods would be confounded with the policy effects

Too many relevant things may change over time (national and international economic conditions, legislation, political variables, innovation, etc.) and we cannot disentangle them from the policy effect we are interested in

Ideally, we would like to have someone who is also observed before and after the policy change, who is also exposed to all these other confounding factors, but that is NOT treated by the policy change

This is the idea of a DID....

LECT 5: DID PART. 2

Table of contents

Introduction

The DID approach makes use of **policy changes** or of **naturally occurring phenomena** that may induce some form of randomization across individuals in the eligibility or the assignment to treatment.

Typically this method is implemented using a before and after comparison across groups.

The DID looks at the period before and at the period after the treatment and compares the change in the average outcomes of the treated group and of the comparison group.

The comparison group is a group that has not been treated in any of the periods.

The **main idea** of this approach is using pre-period (of treatment) differences in outcomes between treatment and control group in order to control for pre-existing differences between the two groups.

It requires the existence of data before and after the treatment for both treatment and comparison group.

A common setting: some policy change which occurred in one region but not in another comparable region.

We have:

	Treated group	Control group
Period before event (T-1)	$D = 0$	$D = 0$
Period after event (T)	$D = 1$	$D = 0$

Formal identification of DID

DID requires two main identifying assumptions:

- Common trend assumption
- Participation into treatment is independent of idiosyncratic shocks

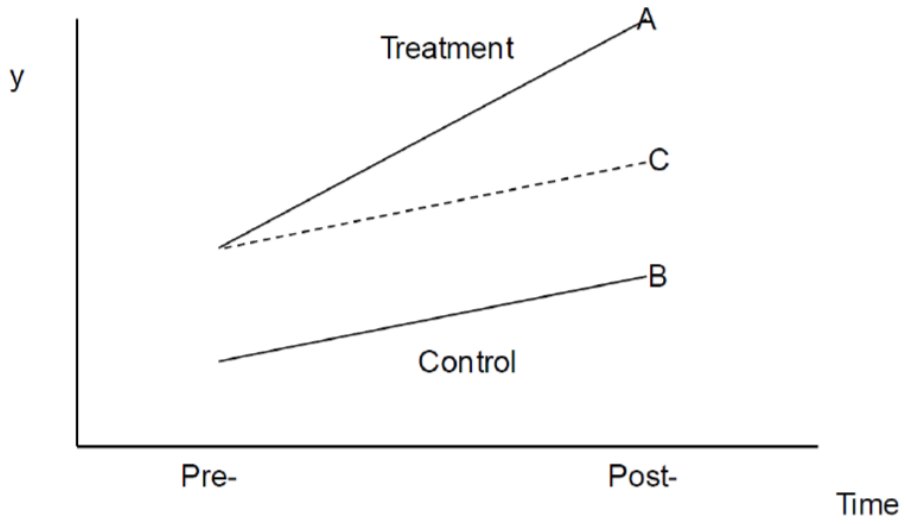
1) Common trend assumption (ass. DID-1)

Define T 1 the period after the treatment occurs and T 0 the period before.
The main identifying assumption of the DID estimator is:

$$E(Y_{0,T1} - Y_{0,T0}|D = 1) = E(Y_{0,T1} - Y_{0,T0}|D = 0) \qquad (1)$$

Absent the treatment, the outcomes in the two groups would have followed parallel trends
In other words, the growth of the outcome in the non-treatment state is independent of treatment allocation.
Crucial assumption: the credibility of a DID estimation hinges on it
Note that it is an assumption because it is something we cannot test: we do not observe E ($Y_{0,T1}$ |D = 1)

FIGURE: COMMON TREND ASSUMPTION



2) Participation into treatment is independent of idiosyncratic shocks (ass. DID-2)

DID approach allows for selection on unobservables but restricts the possible sources of this selection.

Selection into treatment may be determined by the the individual/state/region-specific (unobservable) fixed effect, but it needs to be independent of temporary individual/state/region-specific idiosyncratic shocks.

We will discuss more this assumption later on

Formal identification of DID:

$$
\begin{aligned}
DID &= [E(Y_{1,T_1} - Y_{0,T_0}|D=1)] - [E(Y_{0,T_1} - Y_{0,T_0}|D=0)] = \\
&= [E(Y_{1,T_1} - Y_{0,T_0}|D=1)] - [E(Y_{0,T_1} - Y_{0,T_0}|D=0)] + \\
&- \underbrace{E(Y_{0,T_1} - Y_{0,T_0}|D=1) + E(Y_{0,T_1} - Y_{0,T_0}|D=1)}_{add\ and\ subtract} = \\
&= [E(Y_{1,T_1} - Y_{0,T_0}|D=1)] - E(Y_{0,T_1} - Y_{0,T_0}|D=1) + \\
&+ \underbrace{E(Y_{0,T_1} - Y_{0,T_0}|D=1) - [E(Y_{0,T_1} - Y_{0,T_0}|D=0)]}_{=0\ by\ ass.\ DID-1\ 1} = \\
&= E(Y_{1,T_1}|D=1) - E(Y_{0,T_0}|D=1) + E(Y_{0,T_0}|D=1) - E(Y_{0,T_1}|D=1) = \\
&= E(Y_{1,T_1} - Y_{0,T_1}|D=1) = ATT
\end{aligned}
$$

Under assumption DID-1, the DID estimator identifies the average treatment effect on the treated (ATT).

Additional assumption (or data requirement?)

3) Absence of systematic composition changes within each group

Clearly, the idea DID is comparing the same group over time: if the group is not exactly the same, any observed difference in average outcomes may be simply due to compositional changes.

This is not a concern if one uses longitudinal data, unless there is a substantial attrition in the sample.

When using repeated cross-sections, instead, one needs to make sure that they are representative samples of the same population of interest

Regression DID

DID can be easily estimated with an OLS regression.

Using OLS is a convenient way:

1) to obtain standard errors;
2) to include additional controls.

The heart of the DID setup is an additive structure for potential outcomes.

We discuss it in the conventional setting of two states/areas differently affected by a policy change.

But DID can be used in much more complex settings

We assume that, in the absence of a policy change, the outcome of interest is determined by the sum of an average value across states (α), plus a time-invariant state effect (γs) and a year effect (λt) which is common across regions/states:

$$E(Y_{0,st}|s,t) = \alpha + \gamma_s + \lambda_t$$

where s denotes state and t period.

We assume that the policy effect is a constant δ:

$$E(Y_{1,st}|s,t) = \alpha + \gamma_s + \lambda_t + \delta$$

Therefore:

$$E(Y_{1,st} - Y_{0,st}|s,t) = \delta$$

Define a dummy variable D_{st} which is equal one in ”treated states” in the periods after the policy change has occurred.

- We have:

$$Y_{st} = \alpha + \delta D_{st} + \gamma_s + \lambda_t + \varepsilon_{st} \qquad (2)$$

where $E(\varepsilon_{st}|s,t) = 0$.

Define: TS=treated state; CS= comparison state; B= period before policy change; A= period after policy change

We have:

$$E(Y_{st}|s = CS, t = A) - E(Y_{st}|s = CS, t = B) = \lambda_A - \lambda_B$$

$$E(Y_{st}|s = TS, t = A) - E(Y_{st}|s = TS, t = B) = \lambda_A + \delta - \lambda_B$$

Hence, the population DID is:

$$DID = [E(Y_{st}|s = TS, t = A) - E(Y_{st}|s = TS, t = B)] + \\ - [E(Y_{st}|s = CS, t = A) - E(Y_{st}|s = CS, t = B)] = \delta$$

What is the ”standard” DID regression equation?

We analyze the standard DID case of two states and of a policy change which occurred only in one of the two: for both states we have individual observations for two periods of time (before and after the policy change).

We define a dummy variable TSs which identifies the state where the policy change occurred. And a dummy variable Aftert which is equal to one in the period after the policy change.

Then:

$$Y_{st} = \alpha + \gamma TS_s + \lambda After_t + \delta(TS_s \cdot After_t) + \varepsilon_{st}$$

This equation can be estimated with OLS and the effect of interest is estimated by the coefficient δ on the interaction.

Indeed:

$$E(Y_{st}|TS_s = 0, After_t = 1) - E(Y_{st}|TS_s = 0, After_t = 0) = (\alpha + \lambda) - \alpha = \lambda$$

$$E(Y_{st}|TS_s = 1, After_t = 1) - E(Y_{st}|TS_s = 1, After_t = 0) = \\ (\alpha + \gamma + \lambda + \delta) - (\alpha + \gamma) = \lambda + \delta$$

Therefore:

- Therefore:

$$[E(Y_{st}|TS_s = 1, After_t = 1) - E(Y_{st}|TS_s = 1, After_t = 0)] + \\ -[E(Y_{st}|TS_s = 0, After_t = 1) - E(Y_{st}|TS_s = 0, After_t = 0)] = \delta = ATT$$

The coefficient γ captures any permanent difference in average outcomes between the two states, while the coefficient λ captures any difference in average outcomes between the two periods

Violations of DID assumptions

Rewrite expression (2):

$$Y_{st} = \alpha + \delta D_{st} + \gamma_s + \lambda_t + \varepsilon_{st} = \alpha + \delta D_{st} + u_{st}$$

where D_{st} which is equal one in "treated states" in the periods after the policy change has occurred.

The new error term u_{st} contains a state-specific fixed effect γ_s , a macro time trend λ_t and an idiosyncratic shock ε_{st} . Is the treatment variable endogenous in the regression? We have:

$$E(u_{st}|D_{st}) = \underbrace{E(\gamma_s|D_{st})}_{(a)} + \underbrace{E(\lambda_t|D_{st})}_{(b)} + \underbrace{E(\varepsilon_{st}|D_{st})}_{(c)}$$

(a) $E(\gamma_s|D_{st} = 1) \neq E(\gamma_s|D_{st} = 0)$: selection on unobservable state-specific fixed effect

- states that received the treatment may have different unobservable characteristics with respect to those that were not treated
- the DID approach allows for this type of selection on unobservables (see ass. DID-2)
- this selection term cancels out when taking differences over more period for the same state (as long as there are no compositional changes over time);

(b) $E(\lambda_t|D_{st}= 1) \neq E(\lambda_t|D_{st}= 0)$: different macro trends

- if treated and untreated states have different macro trends, the DID estimates will be biased;
- indeed, part of the observed difference in how outcomes changed over time in treated and comparison groups may due to this difference in trends rather than to the event/policy change;
- the common trend assumption (DID Ass. 1) allows to remove this bias when implementing the DID

(c) $E(\varepsilon_{st}|D_{st}= 1) \neq \varepsilon_{st}|D_{st}= 0)$: selection on unobservable idiosyncratic shocks

- i.e. states that selected into treatment experienced different idiosyncratic shocks with respect to those that were not treated
- by assuming that participation into treatment is independent of idiosyncratic shocks (DID Ass. 2), this term is assumed to be zero.

Hence:

1. the common trend assumption implies: $E(\lambda_t|D_{st}) = \lambda_t \ (\forall s, t)$;
2. the no-selection on idiosyncratic shocks implies: $E(\varepsilon_{st}|D_{st}) = 0$;
3. the absence of compositional changes within groups implies: $E(\gamma_s|D_{sT_1}) = E(\gamma_s|D_{sT_0}) \ (\forall s)$.

The DID estimator identifies the parameter of interest δ if and only if assumptions DD-1, DD-2 and DD-3 are satisfied:

$$\begin{aligned}\widehat{DID} &= E[Y_{TS,T_1} - Y_{TS,T_0}|D_{st}] - E[Y_{CS,T_1} - Y_{CS,T_0}|D_{st}] = \\ &= [(\delta + E(\gamma_{TS}|D_{st}) + E(\lambda_{T_1}|D_{st}) + E(\varepsilon_{TS,T_1}|D_{st}))+ \\ &- (E(\gamma_{TS}|D_{st}) + E(\lambda_{T_0}|D_{st}) + E(\varepsilon_{TS,T_0}|D_{st}))]+ \\ &- [(E(\gamma_{CS}|D_{st}) + E(\lambda_{T_1}|D_{st}) + E(\varepsilon_{CS,T_1}|D_{st}))+ \\ &- (E(\gamma_{CS}|D_{st}) + E(\lambda_{T_0}|D_{st}) + E(\varepsilon_{CS,T_0}|D_{st}))] = \delta\end{aligned}$$

Let's see it in the next slides

Note that in the DID approach the independence of treatment and error term holds in differences.

Main violations of DID assumptions

1) Systematic composition changes within each group
2) Differential macro trends
3) Selection on idiosyncratic shocks (Ashenfelter's dip)

Systematic composition changes within each group

With compositional changes, the state-specific fixed effect does not cancel out when taking differences over time.

This creates an endogeneity issue (assuming common trend and no selection on idiosyncratic shocks):

$$DID = \delta + \underbrace{[(E(\gamma_{TS}|D_{TS,T_1}) - E(\gamma_{TS}|D_{TS,T_0})) - (E(\gamma_{CS}|D_{CS,T_1}) - E(\gamma_{CS}|D_{CS,T_0}))]}_{bias}$$

Not a concern if one uses longitudinal data, unless there is substantial attrition in the sample.

When using repeated cross-sections, instead, one needs to make sure that they are representative samples of the same population of interest

Differential macro trends

Suppose that there is a common macro trend λt , but the two states react differently to it, according to a state specific parameter κs . The observed outcome Y_{st}can be written as:

$$Y_{st} = \alpha + \delta D_{st} + \gamma_s + \kappa_s \lambda_t + \varepsilon_{st}$$

We have (assuming no selection on idiosyncratic shocks and no compositional change within group):

$$DID = \delta + \underbrace{(\kappa_{TS} - \kappa_{CS})(\lambda_{T_1} - \lambda_{T_0})}_{bias \neq 0}$$

- Clearly, the bias is zero if $\kappa_{TS} = \kappa_{CS}$, that is, if the common trend assumption holds.

If one has observation only for two periods - one before and one after the policy change - there is not much one can do to justify the credibility of the common trend assumption.

If, instead, one has data on **multiple periods before the policy change**, several checks can be done on this aspect:

1) Show that control and treatment group had a common trend before the policy (best option!).
2) Placebo DID using, as before and after treatment periods, two periods which are both before the actual treatment.
3) Include state-specific time trends
4) Use a Trend-Adjusted DID:

If there are data for two periods (T_* and T_{**}) previous to the policy change and over which the same macro trend occurred (i.e. $(\lambda_{T_1} - \lambda_{T_0}) = (\lambda_{T_*} - \lambda_{T_{**}})$), one can estimate the Trend-Adjusted DID (TADID):

$$\widehat{TADID} = [DID_{T_1 T_0} - DID_{T_* T_{**}}] =$$
$$= \delta + (\kappa_{TS} - \kappa_{CS})(\lambda_{T_1} - \lambda_{T_0}) - (\kappa_{TS} - \kappa_{CS})(\lambda_{T_*} - \lambda_{T_{**}}) = \delta$$

Selection on idiosyncratic shocks

If we have selection on idiosyncratic shocks (assuming common trend and no compositional changes within groups), we have:

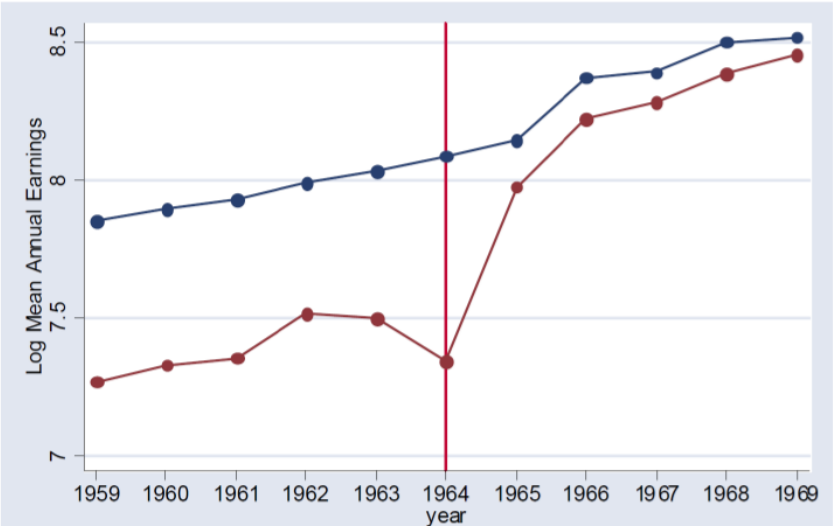
$$\widehat{DID} = \delta + \underbrace{[E(\varepsilon_{TS,T_1} - \varepsilon_{TS,T_0}|D_{st}) - E(\varepsilon_{CS,T_1} - \varepsilon_{CS,T_0}|D_{st})]}_{bias \neq 0}$$

In his 1978 article on the impact of training on earnings (Ashenfelter [1978]), Ashenfelter noticed for the first time what would then came to be called ”Ashenfelters dip”: wages of trainees experienced a fall in the period immediately preceding the training.

Not surprising: workers who experience negative earnings shock have more incentives in joining a training program.

”Nevertheless, this result introduces considerable ambiguity into the empirical analysis for it suggests that some part of the observed earnings increase following training may merely be a return to a permanent path of earnings that was temporarily interrupted by one form of transitory labor market phenomenon or another.” (Ashenfelter [1978], p.51).

FIGURE: ASHENFELTERS DIP



In order to check for an Ashenfelter’s dip, one needs data on at least a few periods before the policy change took place.

One possibility is then estimating the treatment effect with DID using as a pre-treatment period one (or more) which is previous to the actual pre-treatment period (e.g. comparing the period after the policy change with two periods before it).

What happens at the individual level may also happen at the level of governments

With regression to the mean, those who experienced negative shock and therefore joined the treatment - will recover and, therefore, experience a larger earnings growth than those who had a positive shock (and, therefore, did not join the treatment). At least part of the positive effect of the treatment may be just due to this phenomenon.

Endogeneity of policy changes?

”Obviously, DID estimation also has its limitations. It is appropriate when the interventions are as good as random, conditional on timeand group fixed effects. Therefore, much of the debate around the validity of a DID estimate typically revolves around the possible endogeneity of the interventions themselves.” (Bertrand et al. [2004], p. 250).

- Are the policy changes truly exogenous?
- We need to study the political economy of policy changes

For instance, the reason why a minimum wage policy was introduced in one state rather than in another may also explain part of the future differences in the outcomes: e.g. were the states which raised the minimum wage experiencing a particularly positive/negative business cycle?

In general: ”(...) time varying state level policies can be studied as either left or right hand side variables.(...) If state policy making is purposeful action, responsive to economic and political conditions within the state, then it may be necessary to identify and control for the forces that lead policies to change if one wishes to obtain unbiased estimates of a policy’s incidence.” (Besley and Case [2000], p. F672)

At least three good reasons for investigating the determinants of policies (Besley and Case [2000]):

1) it is an important prerequisite to understanding when and whether one can legitimately put policy on the right hand side.
2) it gives us a basis for selecting ‘control groups’: good control groups will be those whose fortunes have evolved similarly to the those of the group experiencing the policy change and who respond similarly to changes in the variables that drive policies to change.
3) it also allows for the identification of instruments for the policy change.

Extensions of DID

- Using repeated cross-section.
- Multiple pre-intervention and/or post.intervention periods: test for Granger causality (Granger [1969]) and perform placebo DID treatments. See Angrist and Pischke [2010] (chapter 5.2.1).
- Multiple treatment and control groups.
- Synthetic control group. (Abadie and Gardeazabal [2003])
- Inconsistent standard errors with DID (Bertrand et al. [2004])
- Non-linear DID

Reading list

Compulsory readings

- Angrist and Pischke [2015]: chapter 5

Additional readings:

- Card and Krueger [1994]
- Krueger and Card [2000]

LECT 5: DID - EXAMPLES

Table of contents

Introduction

We now look at some examples of how a DiD approach can be used to estimate causal parameters of interest

The basic "ingredients" for a DiD are:

- a natural experiment (e.g. event, policy change, etc.)
- two comparable regions / groups of individuals: one never treated and one treated after the event/policy change
- data before and after the natural experiment

The setting, however, can be more complex than the basic case of "2 periods and 2 regions"

We can have multiple periods, several groups, groups treated with different intensity (not just treated or untreated), etc.

Estimating returns to schooling

We have already discussed (see lecture 1) the empirical challenges in estimating the returns to schooling

Individuals who select into more/better education are different with respect to those who do not

The schooling decision has:

- a demand side (a choice of individuals and their parents) which is endogenous (i.e. is determined by unobservable characteristics of the individual)
- a supply side (existence of schools, proximity to them, costs of education, etc.) which is arguably exogenous

Hence, we can try and find some exogenous variation in the supply–side to identify the causal effect of schooling on some outcome of interest

E.g. changes in fees, in number of schools/universities, in number of places available in existing schools

School construction in Indonesia

Does investment in school infrastructure improve educational and labor market attainments of individuals in developing countries?

Duflo [2001] exploits a natural experiment in Indonesia: oil revenues from the 1973 oil boom used to finance public programs INPRES program of school construction:

- 62 thousand new schools constructed between 1973 and 1979
- i.e. more than one school per 500 children in 1971
- Stock of schools doubled and number of teachers grew by 43 percent
- Fastest primary school construction program ever undertaken in the world
- Allocation rule: number of schools constructed in each district proportional to the number of children in primary school age NOT enrolled in school in 1972

Identification strategy:

- Construction program started in 1974, with different intensity in different districts
- More schools were constructed in districts that had less in 1974
- Having more schools in one area increases the average proximity to school for residents, reducing the costs of sending kids to school
- However, if all district were treated what can we use as control group?
- We can compare districts where many schools were constructed with district where few schools were constructed

In the 70s, Indonesian children attended primary school between the ages of 7 and 12

Therefore, students who were already 12 years old in 1974 probably did not benefit at all from the policy: by the time the schools were constructed, they had already finished their education and entered the labour market

The effect of the program, instead, should be positive for younger cohorts, who may have benefitted from having schools closer to their homes

Date of birth and region of birth jointly determine (exogenously) an individual's exposure to the program: people cannot adjust their (or their kids') date and place of birth in response to the policy

Two sources of variation in the policy:

- variation across districts in policy intensity
- **variation within district** across cohorts in exposure to the policy

Empirical approach:

- Outcomes: education and labor market outcomes of adults (measured in 1995 by the intercensal survey of Indonesia, SUPAS)
- Each individual is matched with the number of schools built between 1974 and 1979 in her district of birth

Duflo [2001] implements a DID approach:

Divide regions in "high program intensity" (H) and "low program intensity" (L) (depending on the number of schools constructed)

Divide individuals in three cohorts:

- aged 2 to 6 in 1974 (2-6; treated);
- aged 12 to 17 in 1974 (12-17; untreated);
- aged 17 to 24 in 1974 (17-24; untreated)

Test: difference in average outcome of treated cohort (i) in H and L program regions, minus the difference of untreated cohort (ii) in H and L program regions:

$$DID = E(Y_{2-6,H} - Y_{2-6,L}) - E(Y_{12-17,H} - Y_{12-17,L})$$

Placebo test: DID of the two untreated cohorts

$$DID = E(Y_{12-17,H} - Y_{12-17,L}) - E(Y_{17-24,H} - Y_{17-24,L})$$

"The difference in these differences can be interpreted as the causal effect of the program, under the assumption that in the absence of the program, the increase in educational attainment would not have

been systematically different in low and high program regions" (Duflo [2001], p. 798)

Findings: significant increase in years of education and log wages; no effect in the placebo test

FIGURE: DID ESTIMATES ON EDUCATION AND WAGES; DUFLO [2001]

TABLE 3—MEANS OF EDUCATION AND LOG(WAGE) BY COHORT AND LEVEL OF PROGRAM CELLS

	Years of education			Log(wages)		
	Level of program in region of birth			Level of program in region of birth		
	High (1)	Low (2)	Difference (3)	High (4)	Low (5)	Difference (6)
Panel A: Experiment of Interest						
Aged 2 to 6 in 1974	8.49	9.76	−1.27	6.61	6.73	−0.12
	(0.043)	(0.037)	(0.057)	(0.0078)	(0.0064)	(0.010)
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Difference	0.47	0.36	0.12	−0.26	−0.29	0.026
	(0.070)	(0.038)	(0.089)	(0.011)	(0.0096)	(0.015)
Panel B: Control Experiment						
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Aged 18 to 24 in 1974	7.70	9.12	−1.42	6.92	7.08	−0.16
	(0.059)	(0.044)	(0.072)	(0.0097)	(0.0076)	(0.012)
Difference	0.32	0.28	0.034	0.056	0.063	0.0070
	(0.080)	(0.061)	(0.098)	(0.013)	(0.010)	(0.016)

The table shows means of education (columns 1-3) and wages (columns 4-6) for different cohorts (2-6, 12-17 and 18-24 in 1974) and program levels (high vs low)

Panel A reports the "Experiment of interest": comparing the educational attainment and the wages of individuals who had little or no exposure to the program (they were 12 to 17 in 1974) to those of

individuals who were exposed the entire time they were in primary school (they were 2 to 6 in 1974), in both types of regions.

Note that in both cohorts, the average educational attainment and wages in regions that received fewer schools are higher than in regions that received more schools

This reflects the program provision that more schools were to be built in regions where enrollment rates were low.

In both types of regions, average educational attainment increased over time.

However, it increased more in regions that received more schools.

The difference in these differences can be interpreted as the causal effect of the program, under the assumption that in the absence of the program, the increase in educational attainment would not have been systematically different in low and high program regions.

Individuals young enough, born in a high program region, received on average 0.12 more years of education, and the logarithm of their wage in 1995 was 0.026 higher.

Both these differences in differences are not significantly different from 0.

Panel B reports the “control experiment”, comparing the cohort aged 18 to 24 in 1974 and a cohort aged 12 to 17 in 1974

The first cohort should not have been affected by the policy, while the second may have been only partially affected.

If anything, we should find a smaller - or zero - DID estimate here

Indeed, both figures are substantially smaller than the corresponding values in Panel A

After comparing means to estimate a basic DiD, Duflo [2001] develops her econometric analysis

Econometric analysis: effect on years of schooling

- Duflo [2001] estimates:

$$S_{ijk} = c_1 + \alpha_{1j} + \beta_{1k} + (P_j T_i)\gamma_1 + (C_j T_i)\delta_1 + \varepsilon_{ijk}$$

where:

- S_{ijk}: years of schooling of individual i born in region j in year k
- c_1: a constant
- α_{1j}: district of birth fixed effect
- β_{1k}: cohort of birth fixed effect
- P_j: program intensity in district j
- T_i: equal one if individual i belongs to the cohort aged 2 to 6 in 1974
- C_j: vector of region-specific controls

The coefficient of interest is γ1 : the effect of the program on years of schooling of individuals aged 2 to 6 in 1974

Two potential threats to identification (both would imply an upward bias):

- The allocation of schools to each region was an explicit function of the enrollment rate in the region in 1972. Therefore, the estimate could potentially confound the effect of the program with mean reversion that would have taken place even in its absence.
- The allocation of other governmental programs initiated as a result of the oil boom (and potentially affecting education) may be correlated with the allocation of INPRES schools.

Solution: use specifications that control for the interactions between cohort dummies and...

- ...the enrollment rate in the population in 1971
- ...the allocation of the water and sanitation program, the second largest INPRES program centrally administered at the time.

TABLE 4—EFFECT OF THE PROGRAM ON EDUCATION AND WAGES: COEFFICIENTS OF THE INTERACTIONS BETWEEN COHORT DUMMIES AND THE NUMBER OF SCHOOLS CONSTRUCTED PER 1,000 CHILDREN IN THE REGION OF BIRTH

		Dependent variable					
		Years of education			Log(hourly wage)		
	Observations	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Experiment of Interest: Individuals Aged 2 to 6 or 12 to 17 in 1974 (Youngest cohort: Individuals ages 2 to 6 in 1974)							
Whole sample	78,470	0.124	0.15	0.188			
		(0.0250)	(0.0260)	(0.0289)			
Sample of wage earners	31,061	0.196	0.199	0.259	0.0147	0.0172	0.0270
		(0.0424)	(0.0429)	(0.0499)	(0.00729)	(0.00737)	(0.00850)
Panel B: Control Experiment: Individuals Aged 12 to 24 in 1974 (Youngest cohort: Individuals ages 12 to 17 in 1974)							
Whole sample	78,488	0.0093	0.0176	0.0075			
		(0.0260)	(0.0271)	(0.0297)			
Sample of wage earners	30,225	0.012	0.024	0.079	0.0031	0.00399	0.0144
		(0.0474)	(0.0481)	(0.0555)	(0.00798)	(0.00809)	(0.00915)
Control variables:							
Year of birth*enrollment rate in 1971		No	Yes	Yes	No	Yes	Yes
Year of birth*water and sanitation program		No	No	Yes	No	No	Yes

The table reports estimates of the parameter γ1 for the experiment of interest (2-6 vs 12-17 in 1974; panel A) and for the control experiment (12-17 vs 18-24 in 1974; panel B)

Estimates for education (columns 1-3) and for log wages (columns4-6)

Panel A shows that the effect on education is sizeable and significant: one school built per 1,000 children increased the education of children aged 2-6 in 1974 by 0.12 years for the whole sample (col 1, row 1)

and by 0.2 years in the sample of wage earners (col 1, row 2)

The results are robust to the inclusion of the interaction terms with enrollment rate in 1971 (col 2) and with the allocation of the water and sanitation program (col 3)

There is also a positive a significant effect on wages (columns 4-6): the treated cohort has wages that are 1.5-2.7 percent higher.

In panel B (the control experiment), instead, the impact of the program is very small and never significant.

Duflo [2001] uses a more flexible specification, allowing the policy to have produced different impact on different cohorts (interaction terms analysis)

$$S_{ijk} = c_1 + \alpha_{1j} + \beta_{1k} + \Sigma_{l=2}^{23}(P_j \times d_{il})\gamma_{1l} + \Sigma_{l=2}^{23}(C_j \times d_{il})\delta_{1l} + \varepsilon_{ijk}$$

where dil is a dummy that indicates whether individual i is aged l in 1974.

22 year-of-birth dummies: individuals aged 24 in 1974 form the control group, and this dummy is omitted from the regression

We expect the policy effect to be zero for all the kids who had already left school before the policy introduction (i.e. aged 12 or more in 1974)

The effect should instead become positive for individuals younger than 12 in 1974

Among this latter cohort, we expect the affect to be stronger for younger individuals (who had longer exposure to the increased number of schools)

Estimated coefficients reported in the next figure are in line with our predictions

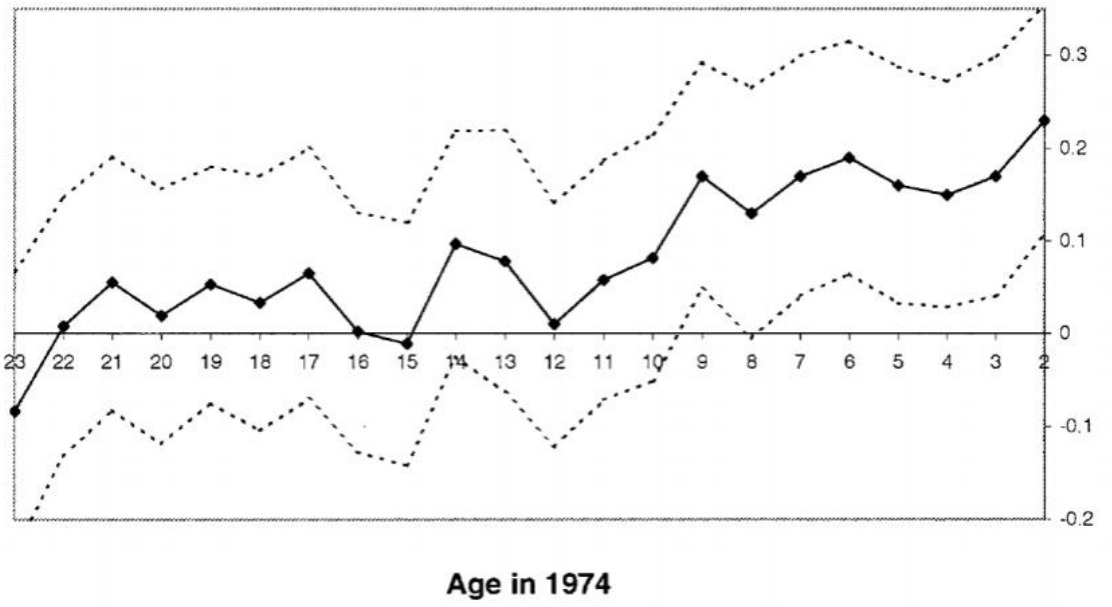


FIGURE 1. COEFFICIENTS OF THE INTERACTIONS AGE IN 1974* PROGRAM INTENSITY IN THE REGION OF BIRTH IN THE EDUCATION EQUATION

Immigration and wages

Do immigrants lower natives' wages and employment?

Although many politicians think that immigrants cannot but have negative effects on the labour market outcomes of natives, the existing empirical evidence generally fail to uncover this negative impact

Obtaining causal estimates on this issue is complex

This is one case where correlations may bear very little information regarding underlying causal effects

By definition, immigrants are a highly mobile population

After having paid a very high cost to reach the host country, the immigrants generally face low costs when it comes to moving across regions within the host country

Immigrants, therefore, are more responsive than natives to shocks in labor demand: they will move to regions where labor demand is stronger

With cross-sectional data, we will probably see that larger populations of immigrants will be residing in regions with higher employment/lower unemployment.

If we have a panel (longitudinal) dataset of regions, recording the number of immigrants and various economic outcomes, we will probably see that regions that experience stronger economic growth (or stronger reduction in unemployment rate) will also receive larger inflows of immigrants

If immigrants are systematically associated to better economic outcomes in the regions where they live, can we conclude that they are beneficial to the local economy?

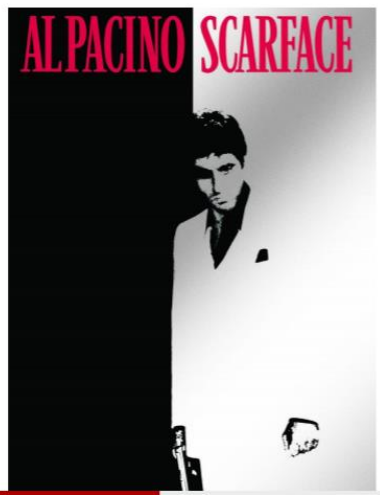
No, unfortunately these correlations are not informative about the causal relationship: immigrants tend to "select into" better regions, the correlation we find may be just due to this selection

The ideal experiment in this area of research would be to have immigrants randomly allocated to different regions within a country and then analyze their impact on local economies

Although we can hardly have this random allocation, there are case where either particular events or government policies can be used to achieve identification

Card [1990] is the first paper which exploits a natural experiment in this field

The 1980 Mariel Boatlift



The 1980 Mariel Boatlift

In 1980, the Cuban president Fidel Castro allowed all Cubans who wished to emigrate to the US to do so from the port of Mariel.

As a result, some 125,000 Cuban immigrants arrived in Miami between May and September 1980, increasing by 20% the number of Cuban immigrants in Miami and Miami's labour force by 7%

The impact on the population of low skilled workers was even stronger given that the majority of the Cuban immigrants were low skilled

This inflow of low-skilled immigrants was exogenous to the local labour market conditions in Miami: a push factor, rather than a Miami-specific pull factor, was determining the inflow

One could compare labor market outcomes in Miami before and after the inflow (Before-After estimator): but the difference observed may be partially/entirely due to the economic cycle which affected Florida

and/or the US in that period (independently from the Mariel boatlift inflow)

We need a comparison group which - under the common trend assumption - would allow us to remove the economic trend

Card [1990] chooses four other major cities (Atlanta, Houston, Los Angeles and Tampa-St. Petersburg) with:

- relatively similar economic growth to Miami in the late '70s and '80s;
- relatively large population of blacks and Hispanics (as in Miami)

Card [1990] uses a DID framework to compare mean wages, employment and unemployment in the pre-migration situation with those occurring after the Mariel boatlift, in both the treatment (Miami) and the control group (the other 4 cities)

He distinguishes between labor market effects on whites, blacks, Cubans and Hispanics.

Table 3. Logarithms of Real Hourly Earnings of Workers Age 16–61 in Miami and Four Comparison Cities, 1979–85.

Group	*1979*	*1980*	*1981*	*1982*	*1983*	*1984*	*1985*
Miami:							
Whites	1.85	1.83	1.85	1.82	1.82	1.82	1.82
	(.03)	(.03)	(.03)	(.03)	(.03)	(.03)	(.05)
Blacks	1.59	1.55	1.61	1.48	1.48	1.57	1.60
	(.03)	(.02)	(.03)	(.03)	(.03)	(.03)	(.04)
Cubans	1.58	1.54	1.51	1.49	1.49	1.53	1.49
	(.02)	(.02)	(.02)	(.02)	(.02)	(.03)	(.04)
Hispanics	1.52	1.54	1.54	1.53	1.48	1.59	1.54
	(.04)	(.04)	(.05)	(.05)	(.04)	(.04)	(.06)
Comparison Cities:							
Whites	1.93	1.90	1.91	1.91	1.90	1.91	1.92
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
Blacks	1.74	1.70	1.72	1.71	1.69	1.67	1.65
	(.01)	(.02)	(.02)	(.01)	(.02)	(.02)	(.03)
Hispanics	1.65	1.63	1.61	1.61	1.58	1.60	1.58
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.02)

Table 3:

- Whites: real earning levels are fairly constant between 1979 and 1985 in both Miami and the comparison cities; in contrast with general decline in real wages in the U.S. economy over this period (evidence of close correspondence between Miami and comparison cities)
- Blacks: wages in Miami were constant from 1979 to 1981, drop in 1982 and 1983 (but, according to Card, this drop should be attributed to the 1982-83 recession: effect is stronger for more educated black workers), and rose to their previous level in 1984; in the comparison cities, instead, showed a steady downward trend
- non-Cuban Hispanics: fairly stable (a slight dip in 1983) in Miami; 6 percent fall in comparison cities
- Cubans: real wages fell by 6-7 percentage points between 1979-1981; decline consistent with the inflow of 45 thousand Mariel workers which were less skilled than the existing pool of Cubans in Miami

Consistent picture when looking at *unemployment rates* (Table 4)

Table 4. Unemployment Rates of Individuals Age 16–61 in Miami and Four Comparison Cities, 1979–85. (Standard Errors in Parentheses)

Group	*1979*	*1980*	*1981*	*1982*	*1983*	*1984*	*1985*
Miami:							
Whites	5.1	2.5	3.9	5.2	6.7	3.6	4.9
	(1.1)	(0.8)	(0.9)	(1.1)	(1.1)	(0.9)	(1.4)
Blacks	8.3	5.6	9.6	16.0	18.4	14.2	7.8
	(1.7)	(1.3)	(1.8)	(2.3)	(2.5)	(2.3)	(2.3)
Cubans	5.3	7.2	10.1	10.8	13.1	7.7	5.5
	(1.2)	(1.3)	(1.5)	(1.5)	(1.6)	(1.4)	(1.7)
Hispanics	6.5	7.7	11.8	9.1	7.5	12.1	3.7
	(2.3)	(2.2)	(3.0)	(2.5)	(2.1)	(2.4)	(1.9)
Comparison Cities:							
Whites	4.4	4.4	4.3	6.8	6.9	5.4	4.9
	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)	(0.4)
Blacks	10.3	12.6	12.6	12.7	18.4	12.1	13.3
	(0.8)	(0.9)	(0.9)	(0.9)	(1.1)	(0.9)	(1.3)
Hispanics	6.3	8.7	8.3	12.1	11.8	9.8	9.3
	(0.6)	(0.6)	(0.6)	(0.7)	(0.7)	(0.6)	(0.8)

We do not see an effect on average earnings, but the Mariel inflow may have affected the earnings distribution

Possibly, a negative effect on low skilled natives (through substitution) and a positive effect on high skilled natives (through complementarities in the labor market)

Card [1990] looks at wage distribution of non-Cuban workers in Miami (Table 5)

Wage distribution remarkably stable over the period 79-85: no evidence of distributional effects

No effects on labour market outcomes of Blacks in Miami (Table 6)

Table 5. Means of Log Wages of Non-Cubans in Miami by Quartile of Predicted Wages, 1979–85. (Standard Errors in Parentheses)

	Mean of Log Wage by Quartile of Predicted Wage				
Year	*1st Quart.*	*2nd Quart.*	*3rd Quart.*	*4th Quart.*	*Difference of Means: 4th − 1st*
1979	1.31	1.61	1.71	2.15	.84
	(.03)	(.03)	(.03)	(.04)	(.05)
1980	1.31	1.52	1.74	2.09	.77
	(.03)	(.03)	(.03)	(.04)	(.05)
1981	1.40	1.57	1.79	2.06	.66
	(.03)	(.03)	(.03)	(.04)	(.05)
1982	1.24	1.57	1.77	2.04	.80
	(.03)	(.03)	(.03)	(.04)	(.05)
1983	1.27	1.53	1.76	2.11	.84
	(.03)	(.04)	(.03)	(.05)	(.06)
1984	1.33	1.59	1.80	2.12	.79
	(.03)	(.04)	(.04)	(.04)	(.05)
1985	1.27	1.57	1.81	2.14	.87
	(.04)	(.04)	(.04)	(.05)	(.06)

Table 6. Comparison of Wages, Unemployment Rates, and Employment Rates for Blacks in Miami and Comparison Cities.
(Standard Errors in Parentheses)

	All Backs				*Low-Education Blacks*			
	Difference in Log Wages, Miami − Comparison		*Difference in Emp./Unemp., Miami − Comparison*		*Difference in Log Wages, Miami − Comparison*		*Difference in Emp./Unemp., Miami − Comparison*	
Year	*Actual*	*Adjusted*	*Emp. − Pop. Rate*	*Unemp. Rate*	*Actual*	*Adjusted*	*Emp. − Pop. Rate*	*Unemp. Rate*
1979	−.15	−.12	.00	−2.0	−.13	−.15	.03	−.8
	(.03)	(.03)	(.03)	(1.9)	(.05)	(.05)	(.04)	(3.8)
1980	−.16	−.12	.05	−7.1	−.07	−.07	.03	−8.2
	(.03)	(.03)	(.03)	(1.6)	(.05)	(.05)	(.04)	(3.5)
1981	−.11	−.10	.02	−3.0	−.05	−.11	.04	−7.7
	(.03)	(.03)	(.03)	(2.0)	(.05)	(.05)	(.04)	(4.2)
1982	−.24	−.20	−.06	3.3	−.17	−.20	−.04	.6
	(.03)	(.03)	(.03)	(2.4)	(.05)	(.05)	(.04)	(4.7)
1983	−.21	−.15	−.02	.1	−.13	−.11	.04	−3.3
	(.03)	(.03)	(.03)	(2.7)	(.06)	(.05)	(.04)	(4.7)
1984	−.10	−.05	−.04	2.1	−.04	−.03	.05	.1
	(.03)	(.03)	(.03)	(2.4)	(.06)	(.05)	(.04)	(4.7)
1985	−.05	−.01	−.06	−5.5	.18	.09	.00	−4.7
	(.04)	(.04)	(.04)	(2.6)	(.07)	(.07)	(.06)	(5.6)

Overall findings:

- the Mariel boatlift inflow had no effect on wages and employment outcomes of non-Cuban workers (blacks included)
- ...and it had no strong effect on the other Cuban workers: negative effects mainly due to “the “dilution” of the Cuban labor force with less-skilled Mariel workers” (Card [1990], p. 255) - i.e. a change in the characteristics of the workers rather than in the returns to their skills

But, how was the Miami labor market able to absorb an inflow of 7 percent of the workforce without adverse effects?

Card [1990] suggests that:

- The Mariel Boatlift seems to have displaced other potential migrants:
  domestic native and earlier immigrant migration into Miami slowed down significantly after the Mariel inflow in comparison with the rest of Florida
- Miami’s industry structure, with a high concentration of industries intensive in low skilled labour, was particularly well-suited to incorporate Cuban immigrants
- The high existing concentration of Hispanics in Miami could have facilitated integration.

Police and crime

Crime imposes high costs on society

Governments can intervene to reduce crime in different ways:

- **Prevention**: education, poverty reduction, etc.
- **Detection** and deterrence: police, judicial system, etc.
- **Incapacitation**: prison (and death penalty)

All these types of interventions are expensive

Let's focus on the impact of policing on crime: **do more policemen lead to lower crime?**

Clearly, if we could increase the enforcement to infinite, such that the probability of being caught when committing crime is equal to one (and the sanctions are sufficient severe) we would drive crime to zero

Only some violent - irrational - crime would still be committed

However, increasing enforcement is costly and we want to know what is the return (i.e. reduction in crime) of investing more in prosecuting criminals

Together with estimates of the marginal benefit of crime (see the costs of crime literature), we could then choose the optimal level of enforcement (i.e. MC = MB)

The optimal level of enforcement is clearly below infinite

Major identification issue: reverse causality

Cross-sectional evidence: more policemen are deployed and more resources are invested in areas with higher levels of crime.

Therefore, areas with high crime rates, may tend to have large police forces, even if police reduce crime.

Longitudinal evidence: within a particular city, if more police are hired when crime is increasing, a positive correlation between police and crime can emerge, even if police reduces crime.

Empirically, we tend to find positive correlations between measures of enforcement and crime

See Freeman [1999] for a survey of early work in this area

This positive correlation is hardly informative about the underlying causal effect that goes from police to crime

Having more policemen can have no effect on crime (for instance, because they are poorly managed) while it is difficult to think that they will increase crime (unless they are all very corrupt)

The recent literature has tried to break the simultaneity between enforcement and crime with IV strategies or natural experiments

Panic on the streets of London

Some recent studies have used terrorism-related events to look at the crime-police relationship, since terror attacks can induce an increased police presence in particular locations (Tella and Schargrodsky [2004], Klick and Tabarrok [2005] and Draca et al. [2011])

Draca et al. [2011] looks at the increased security presence following the terrorist bombs that hit central London in July 2005

“Panic on the streets of London / Panic on the streets of Birmingham / I wonder to myself / Could life ever be sane again ?” (The Smiths)

London terrorist attacks:

- 7 July 2005: series of coordinated suicide attacks (detonation of four bombs) on public transport; 52 people killed and 700 injured
- 21 July 2005: four unsuccessful attempts at detonating bombs on trains

Police activity in central London increased by over 30 percent in the six weeks following the July 7 bombings as part of a police deployment policy titled Operation Theseus by the authorities.

This dramatic “shock” could be used as natural experiment to look at the impact of having more police officers deployed on crime rate

FIGURE: OPERATION THESEUS: TREATMENT GROUP

FIGURE 1. MAP OF LONDON BOROUGHS

Do we have a good setting to address this question?

A crucial part of identifying a causal impact in this type of setting is establishing the exclusion restriction that terrorist attacks affect crime through the post-attack increase in police deployment, and not via other observable and unobservable factors correlated with the attack or shock.

The challenge here is to convincingly isolate the effect on crime of the higher number of policemen from other consequences the terrorist attack may have produced in London (on victims, criminals, interactions between individuals, etc.

They exploit two discontinuities:

1) the increase in policing in the six weeks after the attack;

2) the fall in policing after six weeks

The immediate period surrounding the introduction of the policy was also characterized by a series of potentially correlated observable and unobservable shocks related to the attack

Therefore, the deployment of police and the shock of the attack occurred exactly at the same time (they are both triggered by the attack)

Are criminals offending less in the aftermath of the crisis because there is more police around? Or because they are shocked, because everyone tend to stay home most of the time, because there are less tourists to rob, because everyone is more vigilant to suspicious behavior, etc.?

The second discontinuity is crucial for the identification

Indeed, police deployment was discretely switched off after a six-week period:

→ the Metropolitan Police never made an official public announcement that the police deployment was being significantly reduced.

In this case, the observable and unobservable shocks associated with the attack were still in effect and dissipating gradually.

They observe an increase in crime that is timed exactly with this change: it is difficult to attribute such a clear change in crime rates to observable and unobservable shocks arising from the terrorist attacks.

If these types of shocks significantly affected crime rates, we would expect this to continue even as the police deployment was withdrawn.

FIGURE: LONDON BOMBINGS AND TUBE JOURNEYS

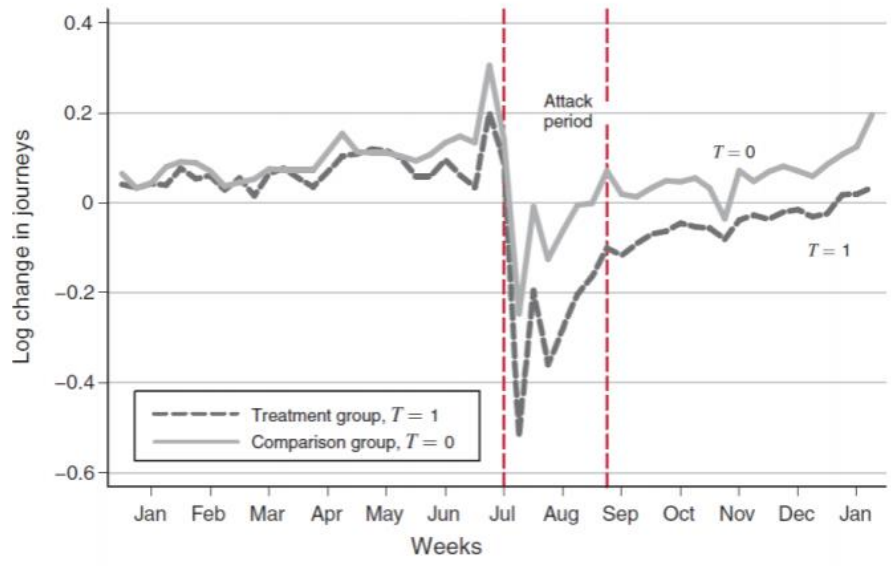


FIGURE 4. YEAR-ON-YEAR CHANGES IN NUMBER OF TUBE JOURNEYS, JANUARY 2004–JANUARY 2006

- Draca et al. [2011] compare 5 areas in Inner London (treated) with 27 areas in Outer London (untreated) in a DID setup
- Data: weekly panel on crime and police covering 32 London boroughs over two years, giving 3,328 observations.
- Basic DID with seasonally adjusted differences (crime is strongly seasonal): before (July 8 - August 19, 2004) and after (July 7 August 18, 2005) - Table 1
- No controls

DID findings (table 1):

- Treatment boroughs experienced a very large relative change in police deployment (34.6 percent increase)
- The relative change was driven by an increase in the treatment group (of 72.8 hours per capita) with little change in hours worked for the comparison group (only 2.2 hours per capita more). This was feasible because of the large number of overtime shifts worked.
- In practice, this means that, while there was a diversion of police resources from the comparison boroughs to the treatment boroughs, the former areas were able to keep their levels of police hours constant.
- Crime rates fell by 11.1 percent: again, this change is driven by a fall in treatment group crime rates and a steady crime rate in the comparison group.

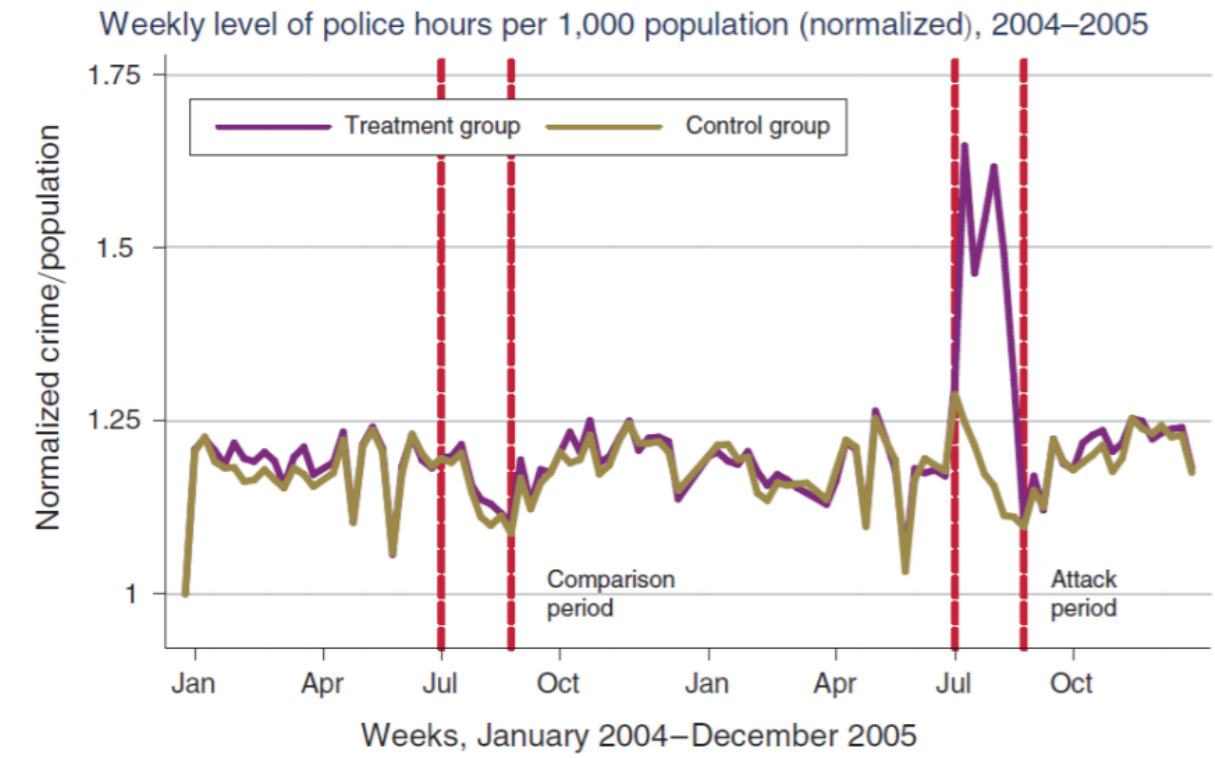
TABLE 1—POLICE DEPLOYMENT AND MAJOR CRIMES, DIFFERENCE-IN-DIFFERENCES, 2004–2005

	Panel A. Police deployment (*Hours worked per 1,000 population*)			Panel B. Crime rate (*Crimes per 1,000 population*)		
	Pre-period (1)	Post-period (2)	Difference (post–pre) (3)	Pre-period (4)	Post-period (5)	Difference (post–pre) (6)
$T = 1$	169.46	242.29	72.83	4.03	3.59	−0.44
$T = 0$	82.77	84.95	2.18	1.99	1.97	−0.02
Difference-in-differences (levels)			70.65*** (7.52)			−0.43** (0.16)
Difference-in-differences (logs)			0.35*** (0.04)			−0.11*** (0.04)

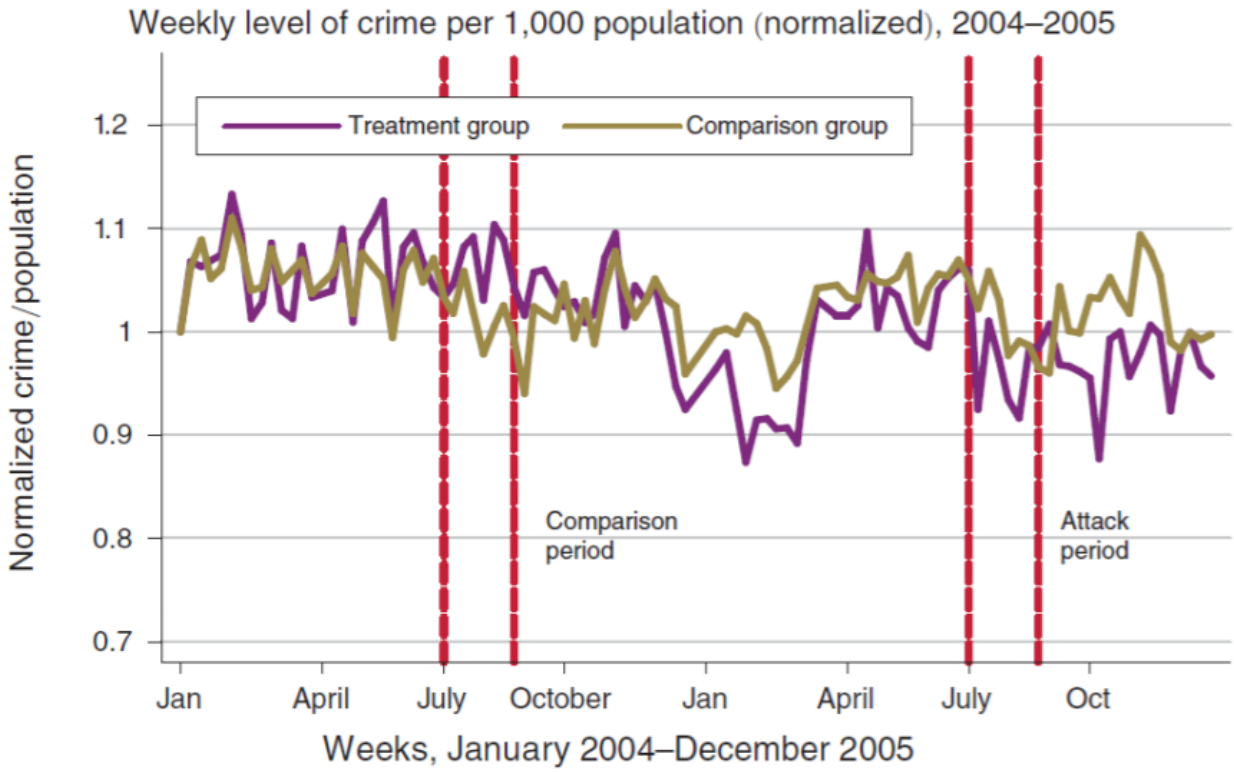
We can look at this empirical exercise with graphs (figure 2)

- Both police hours and crime rate have been normalized to 1 in January 2004 (beginning of the period)
- The visual evidence shows a substantial and clear jump in police hours (panel A) while for crime rate (panel B) the picture is is less decisive:
    - weekly crime rates are clearly more volatile than the police hours data.
- This volatility does raise the possibility that the fall in crime rates seen in the Table 1 DiD estimates may simply be due to naturally occurring, short-run time series volatility rather than the result of a policy intervention - more on this later

Panel A. Police hours (per 1,000 population)



Panel B. Total crimes (per 1,000 population)



In order to test whether the findings of the basic DID are reliable, they run DID regressions

They report regressions of police hours on policy change (Panel A) and regressions of crime on policy change (Panel B)

Define with Tb a dummy which identifies treatment boroughs and with POSTt a dummy variable equal to one in the post-attack period, they estimate:

$$\Delta Y_{bt} = \alpha_1 + \beta_1 POST_t + \delta_1 (T_b * POST_t) + \lambda_1 \Delta x_{bt} + \Delta \varepsilon_{bt}$$

- where Y_{bt} is either police hours (panel A) or crime rate (Panel B)

In order to exploit the two discontinuities (extra-deployment and withdrawal of the extra-deployment), they split the post-attack period into two periods

Define with POSTt1 a dummy identifying the 6 weeks after the attack and with POSTt2 a dummy identifying the time period subsequent to the deployment until the end of the year (that is, from August 19, 2005, until December 31, 2005).

The estimating equation becomes:

$$\begin{aligned} \Delta Y_{bt} &= \alpha_1 + \beta_1 POST_t + \delta_{11} (T_b * POST_t^1) + \delta_{12} (T_b * POST_t^2) \\ &+ \lambda_1 \Delta x_{bt} + \Delta \varepsilon_{bt} \end{aligned}$$

TABLE 2—DIFFERENCE-IN-DIFFERENCES REGRESSION ESTIMATES, POLICE DEPLOYMENT AND TOTAL CRIMES, 2004–2005.

	Full (1)	Split (2)	+Controls (3)	+Trends (4)
Panel A. Police deployment (Hours worked per 1,000 population)				
T × *Post-Attack*	0.081***			
	(0.010)			
T × *Post-Attack1*		0.341***	0.342***	0.356***
		(0.028)	(0.029)	(0.027)
T × *Post-Attack2*		−0.001	0.001	0.014
		(0.011)	(0.010)	(0.016)
Controls	No	No	Yes	Yes
Trends	No	No	No	Yes
Number of boroughs	32	32	32	32
Observations	1,664	1,664	1,664	1,664

	Full (1)	Split (2)	+Controls (3)	+Trends (4)
Panel B. Total crimes (Crimes per 1,000 population)				
T × *Post-Attack*	−0.052**			
	(0.021)			
T × *Post-Attack1*		−0.111***	−0.109***	−0.056*
		(0.027)	(0.027)	(0.030)
T × *Post-Attack2*		−0.033	−0.031	0.024
		(0.027)	(0.028)	(0.054)
Controls	No	No	Yes	Yes
Trends	No	No	No	Yes
Number of boroughs	32	32	32	32
Observations	1,664	1,664	1,664	1,664

Findings:

- Panel A: increase in police hours only in the six weeks after the attack
- Panel B: reduction in crime only in the six weeks after the attack

However, remember that the visual evidence shows a substantial and clear jump in police hours while for crime rate the picture is less decisive: are we finding a significant effect on crime only by chance?

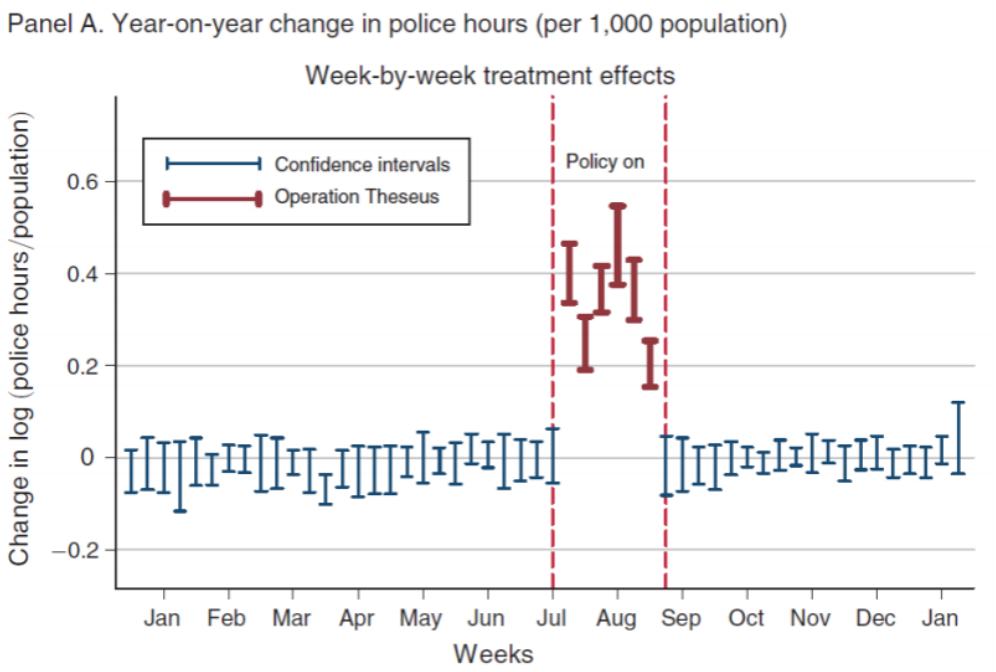
Placebo tests: testing every week for hypothetical or "placebo" policy effects.

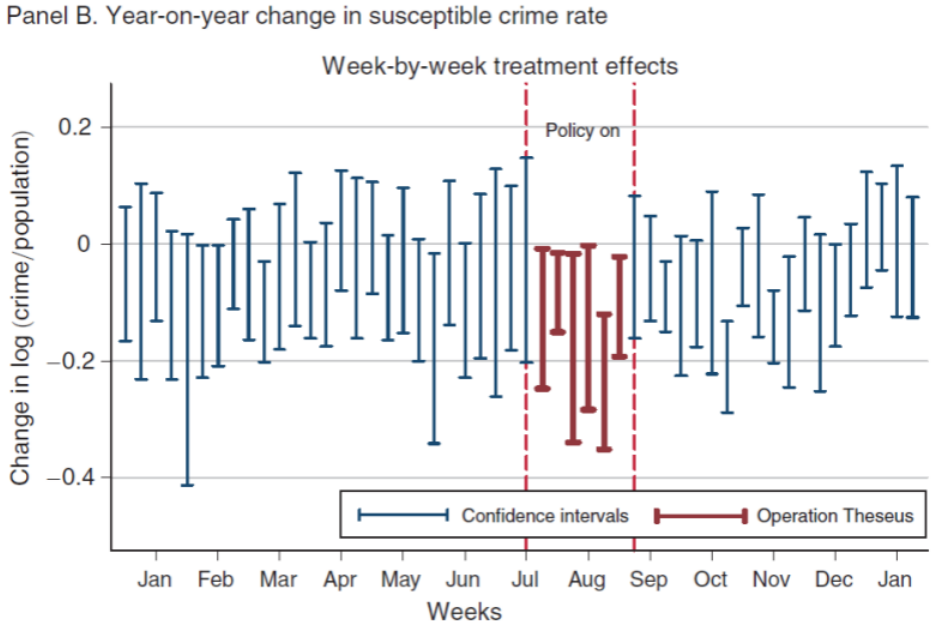
They estimate the same regressions, defining a single week-treatment group interaction term for each of the 52 weeks in our data

They run 52 DID regressions, each featuring a different week × Tb interaction, and plot the estimated coefficients and confidence intervals

The effect should be significant only during the period the Theseus operation was in place

Findings: the pattern of six consecutive weeks of significant, negative treatment effects in the crime rate is not repeated in any other period of the data except Operation Theseus.





IV estimates

In addition to reporting regressions of police hours on policy change (first stage) and regressions of crime on policy change (reduced form), they also report structural regressions In the structural model, the two equation are combined:

$$\Delta crime_{bt} = \alpha_3 + \beta_3 POST_t + \delta_3 PoliceHours_{bt} + \lambda_3 \Delta x_{bt} + \Delta \varepsilon_{bt}$$

This equation is estimated with OLS and with IV. In the latter, the interaction ($T_b * POST_t$) is used to instrument the change in police deployment

IV estimates: substantially larger in size than OLS (as expected); a 10 percent increase in police activity reduces crime by around 3.2 percent.

	OLS Estimates		*IV Estimates*		
	Levels (1)	Differences (2)	Full (3)	Split (4)	+Trends (5)
Panel C. Structural form					
ln(police hours)	0.785*** (0.053)				
Δln(police hours)		−0.031 (0.051)	−0.641** (0.301)	−0.318*** (0.093)	−0.183*** (0.066)
Controls	Yes	Yes	Yes	Yes	Yes
Trends	No	No	No	No	Yes
Number of boroughs	32	32	32	32	32
Observations	3,328	1,664	1,664	1,664	1,664

Crime displacement

Another serious threat to identification is crime displacement: spatial displacement (to other areas) or temporal displacement (crime is postponed)

With DID, spatial displacement would downward bias the estimates (negative spillover on untreated areas), exaggerating the real effect of police on crime

Temporal displacement could impart an upward bias on our estimate. Criminals operating in the treatment group could delay their actions, thus contributing to a larger fall in crime during the policy-on period, but subsequently there will be a compensating increase in crime in the wake of the policy.

See section F in the paper

Reading list

Compulsory readings:

- my lecture slides
- Duflo [2001]
- Draca et al. [2011]

LECT 6: PANEL DATA

Table of contents I

What are Panel Data?

A time series of cross sections where the same individual units are followed over a number of time periods - that is, a collection of N time series.

Individual units can be individuals, firms, regions, countries, etc.

Two sample dimensions: cross-sectional (N, indexed by i = 1, ...N) and time-series (T, indexed by t = 1, ...T).

Two types of longitudinal samples:

- **Unbalanced panel**: individual units observed until they ”disappear” and some may also ”appear” in the sample: individual units are observed for different number of periods.
- **Balanced panel**: Individual units observed for a finite number of time periods and then dropped; all individuals are observed for the same number of periods.

Unbalanced panels

Unbalanced panels usually result from attrition, whereby individual units can disappear at some point:

- If the attrition process is independent of the dependent variable, then attrition is exogenous and balanced and unbalanced panels share the same properties
- Otherwise, attrition is endogenous (for example, the dependent variable is firm profit and firm bankruptcy results from negative profits). As time passes the sample of individuals becomes less and less representative of the population. This selection process must be modelled.

In this course, we assume data are balanced

Balanced panels

Balanced panel with K explanatory variables:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{2T} \\ \vdots \\ y_{NT} \end{bmatrix} \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} x_{111} & x_{112} & \cdots & x_{11K} \\ \vdots & \vdots & & \vdots \\ x_{1T1} & x_{1T2} & \cdots & x_{1TK} \\ x_{211} & x_{212} & \cdots & x_{21K} \\ \vdots & \vdots & & \vdots \\ x_{2T1} & x_{2T2} & \cdots & x_{2TK} \\ \vdots & \vdots & & \vdots \\ x_{NT1} & x_{NT2} & \cdots & x_{NTK} \end{bmatrix}$$

where we use small letters for variables, small bold letters for vectors and capital letters for matrices. Above, y is a column vector of dimension NT , y_i is a column vector of dimension T for each i = 1, . . . , N, X is a matrix NT × K and X_i is a matrix T × K for i = 1, . . . , N.

Different types of panel data:

1) Household panels.
2) Individual level panels.
3) Firm level panels.
4) Countries followed over time.
5) Industries followed over time.

The Model

The basic model we consider is:

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\beta + e_{it} = \\ &= \mathbf{x}_{it}'\beta + f_i + u_{it} \end{aligned}$$

for $i = 1, ..., N$ and $t = 1, ..., T$.
Where

- $e_{it} = f_i + u_{it}$ is the unobservable component: f_i is the unobserved time invariant effect and u_{it} is the idiosyncratic time-varying shock.
- The absence of a t subscript from f_i implies that it does not vary over time.
- The regressors $\mathbf{x}_{it}$ may or may not vary over time.
- $\mathbf{x}_{it}$ is $K \times 1$ and β is $K \times 1$.

Basic assumptions

The two unobservable components are mean-independent:

$$E(u_{it}|f_i) = E(u_{it}) = 0$$

The idiosyncratic temporary shock is not serially correlated:

$$E(u_{it}, u_{is}) = \delta_{ts}\sigma^2$$

where $\delta_{ts} = 1$ if $t = s$ and 0 otherwise.

No correlation across individuals due to unobservable idiosyncratic shocks:

$$\forall i \neq j: \quad E(u_{it}, u_{js}) = 0 \quad \forall t, s$$

Strict versus weak exogeneity

How is *u* related to **x**?

- Strict exogeneity

$$\begin{aligned} E(u_{it}|\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}, f_i) &= E(u_{it}|X_i, f_i) \\ &= E(u_{it}) = 0 \end{aligned}$$

Notice that $E(\mathbf{x}_{it}u_{is}) = 0$ is implied by this assumption.

- **Weak exogeneity** or predetermined regressors

$$E\left(u_{it}\left|\{\mathbf{x}_{is}\}_{s\leqslant t}, f_i\right.\right) = 0$$

In this course, we will assume strict exogeneity.

Random effects versus Fixed effects

How is f_i related to x?

- Random effects

$$E(f_i|\mathbf{x}_{i1}, ...\mathbf{x}_{iT}) = E(f_i) = 0$$

Mean independence implies that: $\text{Cov}(\mathbf{x}_{it}, f_i) = 0$

- Fixed effects

$$E(f_i|\mathbf{x}_{i1}, ...\mathbf{x}_{iT}) = E(f_i|X_i) = g(X_i)$$

where g is a non-constant function of X_i.
In general, this implies that: $\text{Cov}(\mathbf{x}_{it}, f_i) \neq 0$

What does it mean?

Let's consider individual data

Think of fi as the set of persistent traits of someone's personality Would we expect personality to be correlated with observable characteristics of individuals (i.e. with xit)?

For sure whenever the xit contain variables chosen by the individuals (education, labor market status, marital status, number of children, etc.)

Less obvious if the xit are exogenous and predetermined characterizes (gender, age, ethnicity)

With regional or firm data, fi captures persistent unobservable regional (e.g. culture, history, climate, etc.) or firm (e.g. managerial culture, brand, etc.) characteristics

Still likely to be correlated with at least some xit

The Fixed Effects Model

Hence, unless there are strong reasons to argue otherwise (rarely the case), we should generally assume that Cov(x_{it} , f_i) 6= 0 and estimate **Fixed Effects Models**

In other words, we should assume that some regressors are endogenous (i.e. correlated with the error term), but the endogeneity can be modelled as a dependence between the regressors and an unobserved component that is fixed over time

Consider the model:

$$y_{it} = \mathbf{x}'_{it}\beta + \mathbf{z}'_i\gamma + f_i + u_{it} \qquad i = 1, ..., N \text{ and } t = 1, ..., T$$

where we have distinguished between those explanatory variables that vary with time (x_{it}) and those that do not vary over time (z_i).

Assumptions:

1. *Fixed individual effects:* $E(f_i|\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}) \neq 0.$

2. *Strictly exogenous regressors:*
   $E(u_{it}|\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}, f_i) = E(u_{it}) = 0.$

Fixed effects models can be estimated with:

- **Within Groups** (WG) estimator
- First Differences (FD) estimator

Both estimators eliminate the individual fixed effect fi from the equation, eliminating the source of endogeneity

Within Groups (WG) estimator

The idea of the within group estimator is that of removing the fixed effect f_i by de-meaning the data
Define the individual-specific means,

$$\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \qquad \overline{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}, \qquad \overline{\mathbf{z}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_i = \mathbf{z}_i$$

Clearly, the average of z_i (the time-invariant controls) and of f_i are z_i and f_i , respectively

The Within Groups (WG) estimator uses centered (de-meaned)
observations,

$$\widetilde{y}_{it} = y_{it} - \overline{y}_i, \qquad \widetilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - \overline{\mathbf{x}}_i, \qquad \widetilde{\mathbf{z}}_i = \mathbf{z}_i - \overline{\mathbf{z}}_i = \mathbf{0}$$

The average model is:

$$\overline{y}_i = \overline{\mathbf{x}}_i'\beta + \mathbf{z}_i'\gamma + f_i + \overline{u}_i$$

and the centered model is:

$$\widetilde{\mathbf{y}}_{it} = \widetilde{\mathbf{x}}_{it}\beta + \widetilde{\mathbf{u}}_{it}$$

Note that both z_i and f_i have now disappeared from the equation

After centering (de-meaning) the data, we no longer have an endogeneity issue:

$$Cov(\widetilde{\mathbf{x}}_{it}, \widetilde{\mathbf{u}}_{it}) = 0$$

We can use OLS on the centered model

The WG estimator of β is the OLS estimator applied to the centered model:

$$\widehat{\beta}^{WG} = \left(\sum_{i=1}^{N} \widetilde{X}_i'\widetilde{X}_i\right)^{-1}\left(\sum_{i=1}^{N} \widetilde{X}_i'\widetilde{\mathbf{y}}_i\right)$$

Note that all covariates that are time-independent disappear from the centered model, so we will not be able to identify their impact

e.g. given that my ethnicity does not vary over time, I cannot estimate the effect of my ethnicity on some of my outcomes

The Least-Square Dummy-Variable estimator

Instead of de-meaning the variables, one can include individual-specific dummies in the regression, allowing each individual to have a different intercept

This is a different estimator with respect to the WG, but it is conceptually equivalent: the individual-specific dummy captures the individual-specific mean

In some cases, one may be interested in measuring these individual-specific intercepts

e.g. in a panel of Italian regions, we may want to estimate the region fixed effect

We can view the f_i in model:

$$y_{it} = \mathbf{x}'_{it}\beta + \mathbf{z}'_i\gamma + f_i + u_{it} \qquad i = 1, ..., N \text{ and } t = 1, ..., T$$

as parameters to be estimated along with β: N intercepts corresponding to each cross-sectional unit

When xit does not contain a constant term, estimation of a_i , i = 1, ..., N can be achieved defining N dummy variables over the NT observations:

- When $\mathbf{x}_{it}$ does not contain a constant term, estimation of a_i, $i = 1, ..., N$ can be achieved defining N dummy variables over the NT observations:

$$D_{1i} = \begin{cases} 1 & if\ i = 1 \\ 0 & otherwise \end{cases} \quad , ..., \quad D_{Ni} = \begin{cases} 1 & if\ i = N \\ 0 & otherwise \end{cases}$$

- and running the pooled OLS regression

$$y_{it} = a_1 D_{1i} + a_2 D_{2i} + ... + a_N D_{Ni} + \mathbf{x}_{it}\beta + u_{it}$$

with $i = 1, ..., N$ and $T = 1, ..., T$

If x_{it} contains a constant term, we can estimate N − 1 individual specific intercepts and a common intercept

The procedure is simple, though not practical when N is large

Notice that the estimated individual-specific intercepts capture both the individual fixed effect f_i and all the time invariant variables z_i

The resulting estimator, denoted β^{DV} , coincides with β^{WG} and they share the same properties

The First Difference Estimator

This is an alternative to a WG (or DV) estimator

Consider again the model,

$$y_{it} = \mathbf{x}'_{it}\beta + \mathbf{z}'_i\gamma + f_i + u_{it} \qquad i = 1, ..., N \text{ and } t = 1, ..., T$$

where we have distinguished between those explanatory variables that vary with time ($\mathbf{x}_{it}$) and those that do not vary over time ($\mathbf{z}_i$).

We can take first differences (first differences = difference between t and t − 1) to obtain:

$$\Delta y_{it} = \Delta \mathbf{x}'_{it}\beta + \Delta u_{it}$$

$$\text{where } \Delta y_{it} = y_{it} - y_{it-1},\ \Delta \mathbf{x}'_{it} = \mathbf{x}'_{it} - \mathbf{x}'_{it-1} \text{ and } \Delta u_{it} = u_{it} - u_{it-1}.$$

Notice that by differencing the regression equation, we got rid of the fixed effect.

Note again that all covariates $\mathbf{z}_i$ that are time-independent disappear from the differenced model, so we will not be able to identify their impact.

Now we can run an OLS regression on the first-differenced equation

The first difference estimator $\widehat{\beta}^{FD}$ is given by

$$\widehat{\beta}^{FD} = \left(\sum_{i=1}^{N} \Delta X_i' \Delta X_i\right)^{-1} \left(\sum_{i=1}^{N} \Delta X_i' \Delta \mathbf{y}_i\right),$$

Some remarks:

- Using FD we always lose 1 time period, so that we now have T – 1 time periods for each i.
- When we have only two time periods, fixed effects estimation and first differencing produce identical estimates and inference.
- When T > 2, the choice between FE and FD depends on the assumptions about the idiosyncratic errors uit and about strict/weak exogeneity.
- Good practice: always check your results by using both FE and FD estimators

Examples

Example: smoking and income

A common finding in the data is that higher income is associated with healthier behaviours

Is that a causal relationship?

Suppose we want to understand whether increasing individual income would lead to more or less smoking and we have panel data on a sample of individuals

We can estimate the following equation:

$$lcigs_{it} = \beta lrhi_{it} + \gamma age_{it} + \delta age_{it}^2 + \lambda_t + \eta_i + \nu_{it}$$

where $lcigs_{it}$ is the log of average number of daily cigarettes smoked by smokers and $lhri_{it}$ is the log of income

Frank Windmeijer and coauthors have estimated this equation using data from the BHPS (British Household Panel Survey) from 1991 to 2001

- They have 5,300 smokers and a total of 24,972 observations
- They use pooled OLS, WG estimator and FD estimator
- They find a negative β coefficient with the first estimator (random effects) and positive ones with the other two estimator (fixed effects)

lcigs	OLS		Fixed Effects		1st Differences	
	coeff	std err	coeff	std err	coeff	std err
lrhi	-.0837	.0130	.0116	.0092	.0051	.0081
age	.5658	.0327				
age^2	-.0597	.0040	-.0396	.0057	-.0687	.0082
*d*92	-.0326	.0128	-.0056	.0115	.0392	.0115
*d*93	-.0320	.0141	.0130	.0154	.0602	.0122
*d*94	-.0363	.0150	.0390	.0196	.0591	.0128
*d*95	-.0433	.0156	.0931	.0236	.0930	.0121
*d*96	-.0409	.0156	.1346	.0276	.0876	.0119
*d*97	-.0464	.0157	.1733	.0323	.0747	.0120
*d*98	-.0679	.0162	.1958	.0367	.0520	.0121
*d*99	-.0632	.0163	.2199	.0415	.0499	.0120
*d*00	-.0516	.0172	.2657	.0464	.0811	.0124
*d*01	-.0781	.0177	.2964	.0510	.0657	.0128
const	1.568	.0620				

The OLS estimator predominately captures cross-sectional variation in smoking individuals: wealthier individuals smoke less than poorer individuals

The fixed effects estimators (WG and FD), instead, exclusively consider within-individual variation in income and smoking behaviour, showing that increases in income are associated to increases in cigarettes consumption

Example: Manager fixed effects

In some cases, we are interested in estimating the individual fixed effects rather than in eliminating them by de-meaning (WG) or first-differencing the data (FD)

Bertrand and Schoar [2003] estimate CEO fixed effects in order to:

1) show that CEOs matter for firm performance;
2) study how different CEO fixed effects are correlated with performance

How much do individual managers matter for firm behavior and economic performance?

Providing an empirical answer is not obvious: “good firms” will hire “good managers”, finding a positive association between firm performance and manager’s qualities is not very informative Bertrand and Schoar [2003] want to quantify how much of the observed variation in firm policies can be attributed to manager fixed effects

They construct a manager-firm matched panel data set which enables us to track the top managers across different firms over time (using data from Forbes and Execucomp).

They need to observe the same individual managing different firms in order to be able to estimate the CEO fixed effects (sample restriction) separately from the firm fixed effects

Imagine a case in which none of the CEOs change firm: the fixed effect would capture the characteristics of both the firm and the CEO

They estimate how much of the unexplained variation in firm practices can be attributed to manager fixed effects, after controlling for firm fixed effects and time-varying firm characteristics

They analyze a large set of corporate variables: investment policy, financial policy, organizational strategy and performance.

They consider three major executive categories: CEOs, CFOs and “other” (subdivision CEOs, Executive Vice-presidents, COOs, etc.)

They identify approximately 500 managers that either move from one firm to another maintaining the same position (e.g. from CEO in firm A to CEO in firm B; 117 individuals) or move from one position in one firm to a different position in another firm (e.g. from CFO in firm A to CEO in firm B; 7 individuals)

TABLE II

EXECUTIVE TRANSITIONS BETWEEN POSITIONS AND INDUSTRIES

to:	CEO	CFO	Other
from:			
CEO	117	4	52
	63%	75%	69%
CFO	7	58	30
	71%	71%	57%
Other	106	*0*	145
	60%		42%

a. This table summarizes executives’ transitions across positions and industries in the manager-firm matched panel data set (as described in subsection III.A and Table I). All transitions are across firms. The first entry in each cell reports the number of transitions from the row position to the column position. The second line in each cell reports the fraction of the transitions in that cell that are between different two-digit industries.

b. “Other” refers to any job title other than CEO or CFO.

- They estimate the following regression

$$y_{it} = \alpha_t + \gamma_i + \beta X_{it} + \lambda_{CEO} + \lambda_{CFO} + \lambda_{Other} + \epsilon_{it}$$

  where: y_{it} is one of the corporate policy outcomes of firm i in time t; α_t are year fixed effects; γ_i are firm fixed effects; X_{it} is a vector of time-varying firm level controls; ϵ_{it} is an error term.
- The λs are executive fixed effects: λ_{CEO} are fixed effects for the group of managers who are CEOs in the last firm where they can be observed, λ_{CFO} are fixed effects for the group ...

It is evident from the equation above that the estimation of the manager fixed effects is not possible for managers who never leave a given company during our sample period.

Consider, for example, managers who never switch companies and advance only through internal promotions, maybe moving from a CFO to a CEO position in their firm: the effect of these managers on corporate practices cannot be estimated separately from their firm fixed effect.

They report F-tests and adjusted R^2 from the estimation of their main equation for the different sets of corporate policy variables.

For each variable, they report in the first row the fit of a benchmark specification that includes only firm fixed effects, year fixed effects, and time-varying firm controls.

The next two rows, respectively, report the change in adjusted R^2 when we consecutively add the CEO fixed effects and the fixed effects for all three groups of executives (CEOs, CFOs, and other top positions).

Main finding: the inclusion of manager FE increases R^2 for most outcomes.

Similarly, they find that the F-tests are large and allow us to reject in most cases the null hypothesis that all the manager fixed effects are zero.

Hence, managers matter!

TABLE III

EXECUTIVE EFFECTS ON INVESTMENT AND FINANCIAL POLICIES

	Panel A: Investment policy *F-tests on fixed effects for*				
	CEOs	*CFOs*	*Other executives*	*N*	*Adjusted R^2*
Investment				6631	.91
Investment	16.74 (<.0001, 198)			6631	.94
Investment	19.39 (<.0001, 192)	53.48 (<.0001, 55)	8.45 (<.0001, 200)	6631	.96
Inv to Q sensitivity				6631	.95
Inv to Q sensitivity	17.87 (<.0001, 223)			6631	.97
Inv to Q sensitivity	5.33 (<.0001, 221)	9.40 (<.0001, 58)	20.29 (<.0001, 208)	6631	.98
Inv to CF sensitivity				6631	.97
Inv to CF sensitivity	2.00 (<.0001, 205)			6631	.98
Inv to CF sensitivity	0.94 (.7276, 194)	1.29 (.0760, 55)	1.28 (.0058, 199)	6631	.98
N of acquisitions				6593	.25
N of acquisitions	2.01 (<.0001, 204)			6593	.28
N of acquisitions	1.68 (<.0001, 199)	1.74 (.0006, 55)	4.08 (<.0001, 203)	6593	.36

Note that, for each executive in the sample, they have estimated several fixed effects, one for each outcome considered

That is, for executive A, they have estimated a fixed effect on investment, one on performance, etc.

Are these fixed effects correlated? is the correlation positive or negative?

They are interested in understanding whether there are particular managerial “styles”

They analyze the correlation structure between the manager specific fixed effects which they retrieve from the set of regressions they ran.

They form a data set that, for each manager, contains the estimated fixed effects for the various corporate variables.

They estimate regressions as follows:

$$F.E.(y)_j = \alpha + \beta F.E.(z)_j + \epsilon_j$$

where j indexes managers, and y and z are any two corporate policy variables.

They find, for instance, that mangers seem to differ in their approach toward external versus internal growth (see Table VII)

There is a strong negative correlation between capital expenditures, which can be interpreted as internal investments, and external growth through acquisitions and diversification (last 2 rows of column 1)

In a similar vein, managers who follow expansion strategies through external acquisitions and diversification engage in less RD expenditures (last 2 rows of column 1)

(Note that coefficients that are significant at the 10 percent level are highlighted in bold)

TABLE VII

RELATIONSHIP BETWEEN THE MANAGER FIXED EFFECTS

	Investment	Inv to *Q*	Inv to CF	Cash holdings	Leverage	R&D	Return on assets
Investment							0.00
							(0.00)
Inv to *Q* sensitivity	**6.8**						**0.03**
	(0.92)						**(0.01)**
Inv to CF sensitivity	0.02	**−0.23**					−0.01
	(0.6)	**(0.11)**					(0.01)
Cash holdings	−1.10	−0.79	−0.46				**−0.12**
	(1.62)	(1.71)	(1.72)				**(0.05)**
Leverage	−0.39	−0.28	−0.63	**−0.40**			−0.02
	(0.55)	(0.59)	(0.60)	**(0.17)**			(0.02)
R&D	**0.07**	**0.08**	**−0.03**	**−0.23**	**−0.02**		0.11
	(0.00)	**(0.02)**	**(0.01)**	**(0.04)**	**(0.01)**		(0.11)
Advertising	0.01	**0.02**	−0.01	−0.01	0.00	0.25	**0.31**
	(0.01)	**(0.01)**	(0.01)	(0.04)	(0.01)	(0.15)	(0.15)
N of acquisitions	**−0.27**	0.08	**0.23**	**0.01**	**0.02**	**−0.01**	**−0.01**
	(0.11)	(0.10)	**(0.10)**	**(0.00)**	**(0.01)**	**(0.00)**	**(0.00)**
N of divers. acquis.	**−0.30**	−0.14	0.14	0.01	0.01	**−0.01**	**−0.01**
	(0.13)	(0.15)	(0.14)	(0.01)	(0.02)	**(0.00)**	**(0.00)**
SG&A	**−0.22**	**−0.30**	**0.10**	0.54	0.06	**−4.32**	**−3.36**
	(0.01)	**(0.04)**	**(0.03)**	(0.56)	(0.21)	**(0.90)**	**(0.62)**

LECT 7: INSTRUMENTAL VARIABLES

Table of contents I

Introduction

We have an **endogeneity problem** when the error term is correlated with a regressor:

$$y_i = \alpha + \beta x_i + u_i \qquad E(x_i u_i) \neq 0$$

Main sources of endogeneity:

- omitted variable
- measurement error in regressors
- simultaneity (or reverse causality)
- lagged dependent variable in the presence of autocorrelation

In our policy evaluation problem, we were worried about individuals selecting into treatment D according to some unobservable characteristics. This is an omitted variable problem.

With endogeneity, OLS estimator is not consistent

The use of Instrumental Variables (IV) strategy can solve endogeneity problems

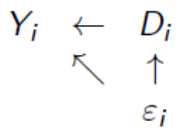
An IV is a variable which influences (= has a causal effect on) the endogenous variable (e.g. treatment), but has no direct effect on the outcome

Examples:

- Randomized experiments with imperfect compliance: the randomization Z influences the participation decision D but has no (or, should not have) direct effect on the outcome Y
- Fuzzy RD: being above or below the threshold... (in a neighbourhood of the threshold)
- College enrolment decision: distance from the university, a change in university fees, etc. will influence the decision, but should not have a direct impact on future earnings

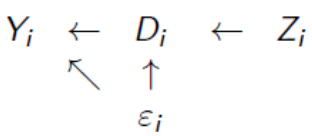
We can represent:

- the endogenous regressor D_i as:



some unobservable ε_i influences both Y_i and D_i

- the instrumental variable Z_i as:



Z_i influences Y_i only through D_i, and it is not correlated with the error term ε_i

What is an IV?

Consider the linear model:

$$y_i = \mathbf{x}_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i \qquad (1)$$

where $\mathbf{x}_{1i}$ includes a constant and is $1 \times K$, β_1 is $K \times 1$, x_{2i} and β_2 are scalars, $E(\varepsilon_i) = 0$ and

$$cov(\mathbf{x}'_{1i}\varepsilon_i) = \mathbf{0} \qquad (2)$$
$$cov(x_{2i}\varepsilon_i) \neq 0 \qquad (3)$$

that is:

- x_{1i} are **exogenous regressors** (i.e. not correlated with ε_i)
- x_{2i} is an **endogenous regressor** (i.e. correlated with ε_i)

The fact that x_{2i} is endogenous implies that our estimate of the coefficient β_2 would be biased

Unfortunately, it can be shown that the presence of one endogenous regressor will also bias the estimated coefficients on all other regressors x_{1i}

If we can find an **instrumental variable** z_{2i} for the endogenous regressor x_{2i}, we can "instrument" it and solve this endogeneity problem (= remove the bias)

What characteristics must a variable have in order to qualify for being a "good" instrumental variable?

Exclusion restriction and Rank condition

We need to find a variable that:

- is observable.
- is not already included among the regressors x1i in the equation
- and satisfies two conditions:
  - Exclusion restriction
  - Rank condition

1) Exclusion restriction

z2i must be exogenous in the main equation (1):

- z_{2i} must be **exogenous** in the main equation (1):

$$cov(z_{2i}\varepsilon_i) = 0 \qquad (4)$$

That is, the instrument should not be correlated with the unobservables (for example ability) contained in the error term ε_i.

This condition also implies that z_{2i} does not determine directly y_i (otherwise it would be contained in the error term ε_i or among the other regressors x_{1i})

We say that z_{2i} determines y_i only via its effect on x_{2i}.

An instrument satisfying this condition is said to be **valid**.

2) Rank condition

z_{2i} must **be partially correlated with** x_{2i} once the other exogenous variables x_{1i} have been netted out.

This means that if we regress the endogenous variable x_{2i} on all the other exogenous variables and on the instrument:

$$x_{2i} = \pi_0 + \mathbf{x}_{1i}\pi_1 + z_{2i}\pi_2 + \nu_i \qquad (5)$$

where by definition $E(\nu_i) = 0,\ E(\mathbf{x}_{1i}\nu_i) = 0,\ E(z_{2i}\nu_i) = 0$

π_2 is the partial correlation between the instrument z_{2i} and the endogenous regressor x_{2i} after removing the effect of all other exogenous regressors $\mathbf{x_{1i}}$.

We require that:

$$\pi_2 \neq 0 \qquad (6)$$

That is, the instrument must predict the endogenous variable, conditional on all other exogenous regressors $\mathbf{x_{1i}}$.

An instrument satisfying this condition is said to be **relevant**:

conditioning on the other exogenous regressors, z_{2i} is informative on the variation of x_{2i}.

Hence, a “good instrument” must be both **valid and relevant.**

What is the crucial difference between the two conditions?

1) Exclusion restriction:

This is an **identifying assumption** which can not be tested empirically.

The exclusion needs to come from economic theory, or from the knowledge of some specific setup which creates quasi random variation in the endogenous variable.

The credibility of the instrumental variable strategy crucially depends on the credibility of this exclusion restriction.

2) Rank condition:

This is a **testable condition** (called First Stage regression) and not an assumption

One does not need to assume that the instrument is partially correlated with the endogenous regressor

One needs to show that this is the case

We have to compute a t-test for the null hypothesis $H_0 : \pi_2 = 0$, after OLS estimation of the equation (called First Stage regression) where the endogenous variable is regressed on the other exogenous regressors and on the instrument:

$$x_{2i} = \pi_0 + \mathbf{x}_{1i}\pi_1 + z_{2i}\pi_2 + \nu_i$$

Some examples of “famous” IV strategies

- Returns to schooling: Quarters of birth and compulsory schooling (Angrist and Krueger [1991]; proximity to college (Card [1993]); school construction in Indonesia (Duflo [2001])
- Effect of Vietnam-era military service on veterans’ earnings: Draft lottery (based on birthdays) (Angrist [1990])
- The effect of family size on mothers employment: Siblings’ sex mix (and twins births)(Angrist and Evans [1998])

Identification of IV estimator

We want to estimate the following univariate model (i.e. an equation with just one regressor):

- We want to estimate the following univariate model (i.e. an equation with just one regressor):

$$y = \beta x + \varepsilon$$

but the regressor x is endogenous: $cov(x, \varepsilon) \neq 0$. Using OLS, we would obtain a biased estimate of the β coefficient.

- Now, suppose we take the covariance of each term in the previous equation with the instrument z: :

$$cov(z, y) = \beta cov(z, x) + cov(z, \varepsilon)$$

We can write the β coefficient as:

$$\beta = \frac{cov(z, y)}{cov(z, x)} - \frac{cov(z, \varepsilon)}{cov(z, x)}$$

Now, let’s see how the exclusion restriction and the rank condition allows us to identify β

- Exclusion restriction implies: cov (z, ε) = 0
- Rank condition implies that cov (z, x) is different from zero and, therefore, the ratio cov (z, y)/cov (z, x) is well defined

Hence, we have:

$$\beta = \frac{cov(z, y)}{cov(z, x)}$$

If we have a valid and relevant instrument z, we can identify the β coefficient as the ratio between the covariance of z and y and the covariance between z and x

In matrix notation, the IV instrument is:

$$\widehat{\beta}_{IV} = \left(\sum_{i=1}^{N} \mathbf{z}_i' \mathbf{x}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i' y_i = (Z'X)^{-1}(Z'y)$$

IV and 2SLS

The IV estimator is often referred to as Two-Stages Least Squares (2SLS)

- The IV estimation can be performed in two steps:
  - **OLS regression of the endogenous variable on ALL the instruments** (i.e. the instrument and all the exogenous regressors); predict the endogenous variable using the estimated coefficients
  - OLS regression of the outcome on the predicted endogenous variable and all the exogenous regressors

NB: separately performing the two steps would lead to wrong standard errors (standard errors in the second stage need to be adjusted for the fact that a predicted regressor is used instead of the endogenous one)

Some definitions:

- Structural equation:

$$y_i = \mathbf{x}_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$$

this equation defines the causal relationship of interest

- **First-Stage** regression

$$x_{2i} = \pi_0 + \mathbf{x}_{1i}\pi_1 + z_{2i}\pi_2 + \nu_i$$

this regression tests whether the instrument is relevant (rank condition)

- **Second-Stage** regression:

$$y_i = \mathbf{x}_{1i}\beta_1 + \widehat{x_{2i}}\beta_2 + \varepsilon_i$$

$$\text{where: } \widehat{x_{2i}} = \widehat{\pi_0} + \mathbf{x}_{1i}\widehat{\pi_1} + z_{2i}\widehat{\pi_2}$$

this equation is identical to the structural equation, but we have replaced the endogenous variable x_{2i} with its predicted value $\widehat{x_{2i}}$ obtained from estimating the first stage regression.

- **Reduced form** regression:

is obtained by substituting the first-stage into the structural equation:

$$\begin{aligned} y_i &= \mathbf{x}_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i = \\ &= \mathbf{x}_{1i}\beta_1 + (\pi_0 + \mathbf{x}_{1i}\pi_1 + z_{2i}\pi_2 + \nu_i)\beta_2 + \varepsilon_i = \\ &= \pi_0\beta_2 + \mathbf{x}_{1i}(\beta_1 + \pi_1\beta_2) + z_{2i}(\pi_2\beta_2) + (\varepsilon_i + \nu_i\beta_2) = \\ &= a + \mathbf{x}_{1i}b + z_{2i}c + u_i = \end{aligned}$$

$$\text{where } b = (\beta_1 + \pi_1\beta_2) \text{ and } c = (\pi_2\beta_2)$$

We can think of the IV estimator as a technique to decompose an endogenous regressor into two components:

- one component correlated with the error term
- one component uncorrelated with the error term

And the last component is used to estimate the parameters.

Conditional on the other exogenous covariates, the IV estimator retains only the variation in the endogenous variable that is generated by quasi-experimental variation, that is, by the instrumental variable

E.g. consider the choice of getting college education: part of this choice is driven by unobservable individual characteristics (ability, motivation, etc.) which are endogenous, but another part is driven by external determinants (e.g. a change in college fees) which are exogenous (i.e. not correlated with ability or motivation)

The mechanics of 2SLS

Consider the simplest framework of a univariate regression:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

We now estimate the first and second stage regressions

1) First stage

$$x_i = \pi_0 + z_i\pi_1 + \nu_i$$

- exogeneity of the instrument, $Corr(z_i\varepsilon_i) = 0$, implies $Corr(\pi_0 + z_i\pi_2, \varepsilon_i) = 0$
- as we do not know $\pi_0 + z_i\pi_1$, we estimate it consistently through OLS:

$$\widehat{x_i} = \widehat{\pi}_0 + z_i\widehat{\pi}_1$$

- and in $\widehat{x_i}$ we have "isolated" the part of x which is not correlated with ε

2) Second stage: use $\widehat{x_i}$ in place of x_i in our model of interest, and run the OLS regression:

$$y_i = \alpha + \beta\widehat{x_i} + error$$

The OLS estimator of the coefficient of $\widehat{x_i}$ in the second stage regression is the two stage least squares estimator $\widehat{\beta}_{2SLS}$.
That is:

$$\widehat{\beta}_{2SLS} = \frac{cov(\widehat{x_i}, y_i)}{var(\widehat{x_i})}$$

Note that an alternative way to write $\hat{\beta}_{IV}$ is:

$$\hat{\beta}_{IV} = \frac{cov(y,z)}{cov(z,x)} = \frac{\frac{cov(y,z)}{var(z)}}{\frac{cov(x,z)}{var(z)}}$$

The IV coefficient is the ratio of the coefficients from two regressions:

- regression of y on z (the reduced form equation)
- regression of x on z (the first stage equation)

Multiple IVs and multiple endogenous variables

The 2SLS easily applies to the cases of multiple IVs and of multiple endogenous variables

With multiple IVs:

- the only difference is that the First Stage will now include all the instruments (and all the exogenous regressors)
- E.g. suppose we have two instruments: z'_{2i} and z''_{2i}
- The first-stage regression is now:

$$x_{2i} = \pi_0 + \mathbf{x}_{1i}\pi_1 + z'_{2i}\pi_2 + z''_{2i}\pi_3 + \nu_i$$

With multiple endogenous variables:

- (at least) one IV needed for each endogenous variable
- one First Stage for each endogenous variable (BUT, each FS must contain all instruments and all exogenous variables)

Properties of $\widehat{\beta}_{IV}$

- IV is **biased**
- $\widehat{\beta}_{IV}$ is a **consistent** and asymptotically normal estimator for β under the following assumptions:
  - (IV1) the population model is $y_i = \mathbf{x}_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$
  - (IV2) we have a random sample from the population on $(y, \mathbf{x}_1, x_2, z_2)$
  - (IV3) z_{2i} is partially correlated with x_{2i} and there are no perfect linear relationships among the IV variables $\mathbf{z}_i = (\mathbf{x}_{1i}, z_{2i})$
  - (IV4) $E(\mathbf{z}_i\varepsilon_i) = \mathbf{0}$ ($\Rightarrow E(\mathbf{x}'_{1i}\varepsilon_i) = \mathbf{0}$, $E(z_{2i}\varepsilon_i) = 0$)
- $\widehat{\beta}_{IV}$ is asymptotically efficient in the class of IV estimators that use instruments linear in $\mathbf{z}$ under the additional homoskedasticity assumption: (IV5) $E(\varepsilon_i|\mathbf{z}_i) = \sigma^2$
- IV is less efficient than OLS (it uses only a fraction of the variation in the endogenous variable)

Weak instruments

The IV estimator is consistent but not unbiased: if we have a good instrument, we should worry about our estimates only if the sample size is small

In the early 1990s a number of papers have highlighted that IV can be severely biased in particular if:

- instruments are weak (i.e. the first stage relationship is weak)
- many instruments are used to instrument for one endogenous variable (i.e. there are many overidentifying restrictions).

When an instrumental variable exhibits only weak partial correlation with the endogenous regressors the **instrument** is said to be **weak** (or poor)

The consequences of a weak instrument on the properties of $\widehat{\beta}_{IV}$ are:

- if the exclusion restriction holds (Cov (z, ε) = 0):
    - it makes the asymptotic variance of $\widehat{\beta}_{IV}$ large, and much larger than OLS asymptotic variance
    - in small samples, the IV is biased towards the OLS estimate
- if the exclusion restriction does not hold perfectly (Cov (z, ε) ≠ 0):
    - The weakness of the instrument exacerbates the inconsistency of the IV estimate: even a mild violation of the exclusion restriction may lead to a large inconsistency
    - the inconsistency of the IV estimator may be even larger than that of the OLS estimator

Let's first consider finite sample properties (unbiasedness)

Suppose we estimate the following model:

$$y = \beta x + \varepsilon$$

and the first stage equation is:

$$x = \pi z + \nu$$

If ε and v are correlated (e.g. $\sigma_{\varepsilon,v}$ 6 ≠ 0 because they both contain the same unobservable characteristics that determine both the outcome y and the endogenous regressor x) estimating the main equation with OLS would deliver biased results

The OLS bias is:

$$E[\beta_{OLS} - \beta] = \frac{Cov[\varepsilon, x]}{Var[x]} = \frac{\sigma_{\varepsilon,\nu}}{\sigma_x^2}$$

- It can be shown that the bias of 2SLS is approximately:

$$E[\beta_{2SLS} - \beta] \approx \frac{Cov[\varepsilon, \nu]}{Var[\nu]} \frac{1}{F+1}$$

  where F is the F-statistic for the joint significance of the excluded instruments in the first stage regression
- If the first stage is very strong ($F \to \infty$), the bias goes to zero
- If the first stage is weak (i.e. $F \to 0$), the bias of 2SLS approaches $\frac{\sigma_{\varepsilon,\nu}}{\sigma_\nu^2}$
- Note that this is the same bias we would get from using OLS if $\pi = 0$ in the first stage (i.e. if there is no first stage relationship): in that case, $x = \nu$ and the OLS bias $\frac{\sigma_{\varepsilon,\nu}}{\sigma_x^2} = \frac{\sigma_{\varepsilon,\nu}}{\sigma_\nu^2}$

"Unity makes strength"?

Not really, in this case. Using many weak instruments rather than one weak instrument likely makes matters worse.

Adding further instruments without any predictive power will further reduce the F-statistic, increasing the 2SLS bias

It can be shown that if the model is just identified (i.e. number of instruments = number of endogenous variables) and the F-statistics is not zero, the IV estimator is approximately unbiased.

Let's now consider consistency:

- It can be shown that the probability limit of $\widehat{\beta}_{IV}$ in terms of population moments is:

$$p\lim \widehat{\beta}_{1,IV} = \beta + \frac{Cov(z,\varepsilon)}{Cov(z,x)} = \beta + \frac{Corr(z,\varepsilon)}{Corr(z,x)} \frac{\sigma_\varepsilon}{\sigma_x}$$

Hence, if the exclusion restriction holds perfectly (Cov (z, ε) = 0), the IV estimates are consistent even with weak instrument (i.e. but we should still worry if we are using a small sample).

However, even a minor violation of the exclusion restriction (i.e. a small Corr (z, ε) ≠ 0) can lead to large inconsistency when Corr (z, x) is small

Minor violations of the exclusion restriction (i.e. a small Corr (z, ε) ≠ 0), instead, would not be too worrying as long as Corr (z, x) is sufficiently large.

- It can also be shown that the probability limit of $\widehat{\beta}_{OLS}$ is

$$p\lim \widehat{\beta}_{OLS} = \beta + \frac{Cov(x,\varepsilon)}{Var(x)} = \beta + Corr(x,\varepsilon)\frac{\sigma_\varepsilon}{\sigma_x}$$

- The IV inconsistency can therefore be larger than the OLS inconsistency if $Corr(x,\varepsilon) < \frac{Corr(z,\varepsilon)}{Corr(z,x)}$

This implies that using a weak instrument may be worse than using OLS: "Archimedes said, "Give me the place to stand, and a lever long enough, and I will move the Earth" (...) "But, like Archimedes' lever, instrumental variable estimation requires both a valid instrument on which to stand and an instrument that isn't too short (or "too weak")" (Murray [2006]; p. 111).

To discuss the strength of the instruments:

- Run the First Stage regression
- Perform an F-test under the H_0 that all instruments are not relevant (i.e. all coefficients are equal to zero)
- Rule of thumb: an F-stat above 10 is considered as evidence of a sufficiently strong instrument (Stock et al. [2002]).
- Adding more weak instruments (i.e. instruments without predictive power) to instrument the same endogenous regressor would reduce the F-stat, increasing the 2SLS bias
- With multiple weak instruments, it is generally preferable to use only one IV, choosing the strongest
- You can use a limited information maximum likelihood estimator (LIML). It provides the same asymptotic distribution as 2SLS (under constant effects) but provides a finite-sample bias reduction.

Finding "good" instruments?

It is not easy!

We need to find a variable that:

1) is observable
2) is not already (and should not be) included among the regressors x_{1i} in the main equation
3) has a clear effect on the endogenous variable (i.e. satisfies the rank condition)
4) has no effect on the outcome y_i other than through the first stage: the effect of z_{2i} on x_{2i}

Exogenous variability is often generated by institutional constraints and policies: natural experiment or quasi-experiments

And instruments can also be constructed (e.g. shift-share instruments)

The IV-Wald estimator

Basic setting:

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

- D is a dummy for a treatment status and is endogenous
- Z is an IV for D and it is also a dummy

The IV estimator is:

$$\beta_{IV} = \frac{cov(Y, Z)}{cov(D, Z)}$$

Consider the **reduced form** regression:

$$Y_i = a + bZ_i + u_i$$

Note that

- $b = E(Y|Z = 1) - E(Y|Z = 0)$
- $b = cov(Y, Z)/var(Z)$

Hence:

$$cov(Y, Z) = [E(Y|Z = 1) - E(Y|Z = 0)] \cdot var(Z)$$

Analogously, consider the **first stage** regression:

$$D_i = \gamma + \delta Z_i + \nu_i$$

Note that:

- $\delta = E(D|Z = 1) - E(D|Z = 0)$
- $\delta = cov(D, Z)/var(Z)$

Hence:

$$cov(D, Z) = [E(D|Z = 1) - E(D|Z = 0)] \cdot var(Z)$$

The IV estimator can be written as:

$$\beta_{IV} = \frac{cov(Y, Z)}{cov(D, Z)} = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)} = \beta_{WALD}$$

This is called the **Wald estimator**.

When both the endogenous variable and the instrument are dummy variables, the Wald estimator and the IV estimator are equivalent.

Note that:

$$E(D|Z=1) = 1*P(D=1|Z=1)+0*P(D=0|Z=1) = P(D=1|Z=1)$$

$$E(D|Z=0) = 1*P(D=1|Z=0)+0*P(D=0|Z=0) = P(D=1|Z=0)$$

Hence, we can rewrite the Wald estimator as:

$$\beta_{WALD} = \frac{E(Y|Z=1) - E(Y|Z=0)}{P(D=1|Z=1) - P(D=1|Z=0)}$$

The Wald estimator is the difference in expected outcomes between those with Z = 1 and those with Z = 0 divided by the difference in the probability of receiving the treatment between those with Z = 1 and those with Z = 0

We have already met two standard examples of Wald-IV estimator:

1) Randomized experiments with imperfect compliance
    - D is the treatment status
    - Z is the initial random assignment
2) Fuzzy RD design
    - D is the treatment status
    - Z is a dummy equal to one if the subject is above/below the threshold

Randomized experiments with imperfect compliance

As already discussed (see lecture 3), with imperfect compliance the initial random assignment (Z) and the actual participation into treatment (D) do not fully overlap

Some individuals randomly assigned to treatment will decide not to take it and, possibly, some of the individuals "randomized out" will manage to get treated

Using the initial assignment (which is random), we can identify the **Intention To Treat (ITT)** parameter:

$$ITT = E(Y|Z=1) - E(Y|Z=0)$$

ITT is the difference in expected outcomes between those randomized in and those randomized out and it measures the impact of being offered the treatment

Note that the WALD estimator can be written as:

$$\beta_{WALD} = \frac{ITT}{P(D=1|Z=1) - P(D=1|Z=0)}$$

It is the ratio between the ITT and the difference in the probability to be treated between those randomized in and those randomized out

The denominator is equal to 1 with perfect compliance: in that case Z = D and ITT = ATE

With imperfect compliance, the denominator will be smaller than 1

Suppose we estimate that the ITT is a 4 percent increase in the outcome (e.g. earnings)

This is the effect of being offered the treatment (e.g. some training), but we know that not all those who were offered the treatment actually took it, and not all those who were not offered the treatment did not take it

Hence, when comparing the outcomes of those offered and not offered the treatment, we have treated and untreated individuals in both groups

This will make the estimated effect of the treatment look smaller than it actually is

Suppose the share of treated individuals is 60 percent among those randomized in and 20 percent among those randomized out

The Wald estimator is:

$$\beta_{WALD} = \frac{ITT}{P(D=1|Z=1) - P(D=1|Z=0)} = \frac{4}{0.6 - 0.2} = 10$$

Hence, if we use the initial random assignment Z as an instrument for treatment status D, we find that the causal effect of taking the training on expected earnings is plus 10 percent

The effect of taking the treatment is obviously larger than the effect of being offered it

See lecture 3 for which parameters are identified by the Wald-IV estimator (ATT with one-sided compliance and LATE with two-sided non-compliance)

Note that if the share of treated individuals is identical among individuals randomized in and those randomized out, the denominator is equal to zero

The Wald-IV estimator is not defined.... because the instrument is not relevant (i.e. rank condition is not satisfied)

If the initial assignment to treatment increases the chances of being treated, it must be that:

$$P(D=1|Z=1) > P(D=1|Z=0)$$

And, is the exclusion restriction credible?

Yes, assignment to treatment is randomized and there is no reason to expect that the initial assignment per se will have a direct effect on the outcome

Fuzzy RDD

As we already discussed (see lecture 4), with fuzzy RD design the probability of being treated discontinuously increases at the threshold but it does not jump from zero to one (which is the case of a sharp RDD)

This is the case of admission to flagship universities: not everyone who is admitted then decides to enroll

If D is the treatment status, we can define Z as a dummy equal to one if the subject is above/below the threshold

We can then use Z as instrument for D

The exclusion restriction is credible: being admitted to a flagship university does not affect future earnings, but it affects the probability of enrolling in a flagship university (which will possibly affect future earnings)

Heterogenous effects and LATE

With homogenous treatment effects, the IV estimator identifies the ATE (Average Treatment Effect)

As already discussed (see lecture 2), we can generally expect to have heterogenous treatment effects (i.e. different individuals benefit with different intensity from the same treatment)

In this case, the IV estimator identifies a **Local Average Treatment Effect (LATE)**

In particular, it is the ATE on a specific group of individuals: the "**compliers**"

The fact that the effect only refers to a group makes it "local"

Consider the case of a binary treatment D and a binary instrument Z :

For instance, D is going to college and Z is being offered a scholarship

The causal chain is: Z → D → Y

We can divide the population in four groups:

	$Z_i = 0$ $D_i(Z=0) = 0$	$Z_i = 0$ $D_i(Z=0) = 1$
$Z_i = 1 \quad D_i(Z=1) = 0$	**Never-takers**	**Defiers**
$Z_i = 1 \quad D_i(Z=1) = 1$	**Compliers**	**Always-taker**

- **Never-takers** are those who would not go to college in any case, irrespectively of being offered a scholarship (i.e. even if they were offered a scholarship)
- **Always-takers** are those who would go to college in any case, irrespectively of being offered a scholarship (i.e. even if they were not offered a scholarship)
- **Compliers** are those who decide to go to college because they were offered a scholarship, but would have not gone if they had not received the scholarship (e.g. credit constrained individuals)
- **Defiers** are those who decide not to go to college because they were offered a scholarship, but would have gone if they had not received the scholarship (this may sound silly... these individuals may not exist)
- We define **Non-Compliers**: Never-takers + Defiers + Always-takers

It can be shown (Imbens and Angrist [1994]) that under thec assumption that there are no defiers (monotonicity condition) the IV estimator identifies the ATE on the population of "**compliers**"

In other words, on the individuals whose decision was "shifted" from non-participation into participation by the instrument.

The “compliers” are those individuals for which the probability to take the treatment shifts from zero to one if the instrument is equal to one (e.g. if they are randomized in):

$$Pr[D_i(Z_i = 1) - D_i(Z_i = 0) = 1]$$

We can write the LATE parameter as:

$$\begin{aligned} LATE &= E[(Y_{1i} - Y_{0i})|D_i(Z_i = 1) - D_i(Z_i = 0) = 1] = \\ &= \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{Pr[D_i(Z_i = 1) - D_i(Z_i = 0) = 1]} \end{aligned}$$

That is, the LATE effect is the difference in average outcomes between those with Z = 1 and Z = 0 (e.g. randomized in and randomized out) divided by the share of compliers

But, can you identify the compliers in our data?

Not really!

We only observe the choice actually made by individuals, not the choice they would have made if the instrument had taken a different value (e.g. if they had been randomized out rather than in)

Note that if Z = 1 and D = 1, compliers and always-takers are not distinguishable; and if Z = 0 and D = 1, defiers and always-takers are not distinguishable; etc.

		$Z_i = 0$ $D_i(Z = 0) = 0$	$Z_i = 0$ $D_i(Z = 0) = 1$
$Z_i = 1$	$D_i(Z = 1) = 0$	**Never-takers**	**Defiers**
$Z_i = 1$	$D_i(Z = 1) = 1$	**Compliers**	**Always-takers**

The LATE is not the average treatment effect for either the entire population or for a subpopulation identifiable from observed values

We need to speculate about who the compliers could be

In general, the estimates are specific to the instrument used and are not generalizable to other contexts because changing instrument will change the group of compliers

This is clearly a serious issue when we think about external validity of the results

Beyond binary instruments

The LATE interpretation can be easily extended to cases where the instrument is not a binary variable (Angrist and Imbens, 1995)

Intuitively, the LATE will be a weighted average of the effect for all values of Z

Suppose Z is a continuous variable: we consider two values Z* and Z**

e.g.: two levels of college fees; two distances from the nearest college

- LATE assumptions:
  - Existence of an instrument Z
  - Monotonicity: assume Z* < Z**

$$D_i(Z_i = Z^*) \geq D_i(Z_i = Z^{**}) \quad \forall i = (1, ..., N)$$

or, alternatively:

$$D_i(Z_i = Z^*) \leq D_i(Z_i = Z^{**}) \quad \forall i = (1, ..., N)$$

Suppose Z is college fees: a reduction in Z should induce more people to enrol

In this case, the monotonicity assumption would imply, for any Z* < Z** :

$$D_i(Z_i = Z^*) \geq D_i(Z_i = Z^{**}) \quad \forall i = (1, ..., N)$$

Therefore:

$$E[(Y_{1i} - Y_{0i})|D_i(Z^*) - D_i(Z^{**}) = 1] = \frac{E(Y_i|Z_i = Z^*) - E(Y_i|Z_i = Z^{**})}{E[D_i|Z_i = Z^*] - E[D_i|Z_i = Z^{**}]}$$

Using college fees as an instrument for college education in a wage regression, would identify the causal average treatment effect of college education on future earnings for those whose decision to

college was shifted by the college fees reduction (compliers)

in this case: compliers are those who chose to go to college because of the reduction in fees, but would not have gone to college in the absence of this reduction

Learning from LATE

What do we learn from LATE?

- IV can be meaningless when effects are heterogeneous (without monotonicity assumption)
- If the monotonicity assumption can be justified, IV estimates identify the average treatment effect for a particular subset of the population (the compliers)
- In general, the estimates are specific to that instrument and are not generalizable to other contexts: changing instrument will change the group of compliers

An example

Consider two alternative policies that can increase enrollment into College:

- Free tuition is randomly allocated to young people to attend College
  (Z = 1 means that the subsidy is available)
- The possibility of entering a competition for a fees waiver based on merit is randomly allocated (Z = 1 means that the individual is allowed to compete for the scholarship)

Suppose the aim is to use these two policies to estimate the returns to College education: the outcome is log earnings, the treatment is going to College and the instrument is one of the two randomly allocated programmes

First, we need to assume that no one who intended to go to College will be discouraged from doing so as a result of the policies (monotonicity assumption)

This could fail as a result of a General Equilibrium response of the policy; for example if it is perceived that the returns to College decline as a result of the increased supply, those with better outside opportunities may drop out.

Now compare the two instruments.

The subsidy is likely to draw poorer liquidity constrained students into College but not necessarily those with the highest returns.

The scholarship is likely to draw in the best students, who will probably have higher returns.

It is not a priori possible to believe that the two policies will identify the same parameter, or that one experiment will allow us to learn about the returns for a broader/different group of individuals

Reading list

Compulsory readings:

- my lecture slides
- chapter 3 - Angrist and Pischke [2015

THE EFFECT OF IMMIGRATION ALONG THE DISTRIBUTION OF WAGES

Christian Dustmann, Tommaso Frattini, Ian Preston

This paper

- Re examines the effect of immigration on wages for Britain over period 1997 2005
- Research question: “What is the impact of immigration on native wages?”
- Empirical observation: considerable downgrading of recent immigrants.
- Empirical Approach:
  - Proposes theory based empirical framework
  - Estimation of wage effects of immigration across the wage distribution
  - No pre allocation of immigrants to skill groups

Motivation and previous literature

Large literature on the effects of immigration on native wages and employment, mostly for the US (e.g. Altonji and Card 1991, Card 2001, 2007, Borjas 2003, Ottaviano and Peri 2008)

Less studies for Europe. (e.g. D’Amuri et al. 2008 Manacorda et al. 2006, Pischke and Velling 1997)

Italy: Gavosto et al. (1999), Venturini and Villosio (2006)

Most papers find modest effects on wages (e.g. Card 2001) (some find positive effects (e.g. Card 2007, Friedberg 2001, Ottaviano and Peri 2008)), with some exceptions (e.g. Borjas 2003)

Latest literature looks at alternative adjustment mechanisms (e.g. Lewis 2005, Peri and Sparber 2009

Structure of Talk

Immigration to Britain: Data Sources and Some Facts

Impact of Immigration on Wages: Empirical Estimation

Results

Conclusion

Data Sources: LFS

Labour Force Survey:

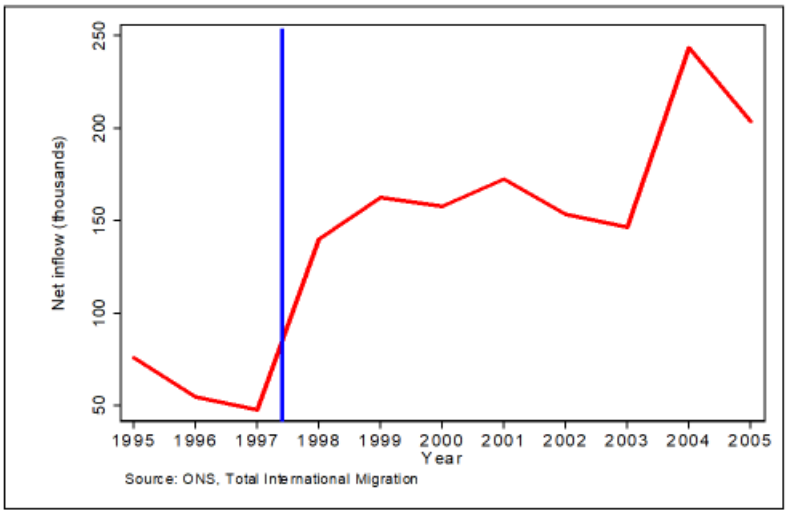
- Available since 1975. In its current quarterly format since 1992 (December 1994 for Northern Ireland)
- About 60000 households interviewed every quarter; 1 st and 5 th quarter have earnings information since 1997
- Small sample size may be a problem when dealing with immigrants

Data Sources: Census

Census

- Available every ten years. Most recent is 2001
- Very good coverage, but not frequent.
- Microdata not available. Tables produced by the ONS.
- A 2% sample of individual records (SARS) available for the 1991 and 2001 Censuses

Immigrant inflows



Foreign born working age population, 1993-2005

	Percentage of total working age population
1993	8.3
1995	8.3
1997	8.7
1999	9.1
2001	9.7
2003	10.4
2005	11.5

Source: LFS, various years

Education, 1997 and 2005

	Natives		*Foreign Born*			
			Earlier		*Recent*	
	1997	*2005*	*1997*	*2005*	*1997*	*2005*
High	11.64	16.02	25.49	33.81	49.25	45.40
Intermediate	23.38	26.41	33.31	34.22	39.62	40.73
Low	64.98	57.57	41.20	31.97	11.13	13.87

Source: LFS, various years

- Low: left FT education at 16 or earlier
- Intermediate: left between 17-20
- High: left>21

Education, 1997 and 2005

	Natives		*Foreign Born*			
			Earlier		*Recent*	
	1997	*2005*	*1997*	*2005*	*1997*	*2005*
High	11.64	16.02	25.49	33.81	49.25	45.40
Intermediate	23.38	26.41	33.31	34.22	39.62	40.73
Low	64.98	57.57	41.20	31.97	11.13	13.87

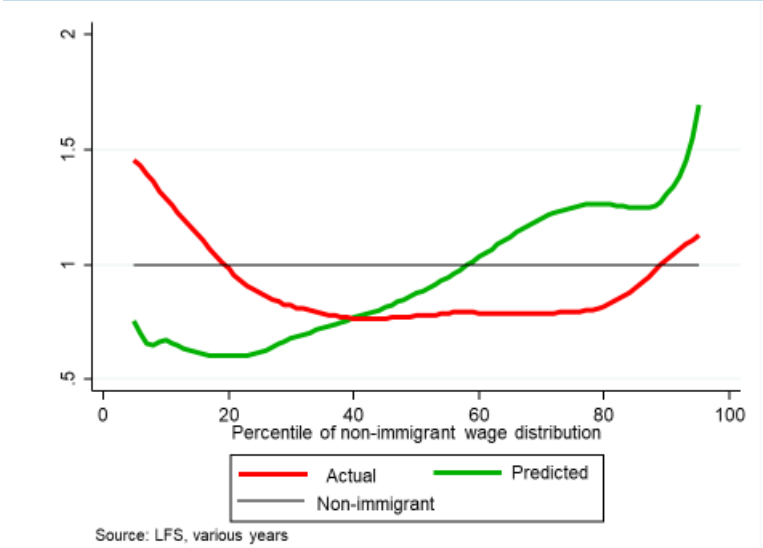
Source: LFS, various years

- Low: left FT education at 16 or earlier
- Intermediate: left between 17-20
- High: left>21

Occupational distribution by level of education

	High education			*Intermediate education*			*Low education*		
	Natives	*Foreign Born*		*Natives*	*Foreign Born*		*Natives*	*Foreign Born*	
		Earlier	*Recent*		*Earlier*	*Recent*		*Earlier*	*Recent*
Higher managerial and professional	36.9	39.7	29.0	15.4	12.9	4.5	7.1	5.8	3.6
Lower managerial and professional	47.3	36.4	29.2	37.8	33.9	13.6	22.7	20.0	6.3
Intermediate occupations	8.2	8.7	9.5	18.5	14.7	9.8	13.7	10.1	1.8
Lower supervisory and technical	2.7	4.3	5.2	9.1	9.7	8.3	17.5	16.8	8.5
Semi-routine occupations	3.8	7.4	15.0	13.3	18.4	29.3	21.6	25.5	29.6
	5.0	**10.9**	**27.2**	**19.2**	**28.9**	**63.7**	**39.0**	**47.3**	**79.8**
Routine occupations	1.2	3.5	12.2	5.9	10.5	34.4	17.4	21.8	50.2

Relative density of recent immigrants along native wage distribution



Structure of Talk

- Immigration to Britain: Data Sources and Some Facts
- **Impact of Immigration on Wages: Empirical Estimation**
- Results
- Conclusion

Empirical strategy

- Empirical research concerned with estimating the *causal* effect of immigration on wages of residents
- Observed: Wages of residents before and after Immigration
- Not Observed: Wages of residents after Immigration *if Immigration had not taken place*
- Empirical challenge: reconstruction of the *missing counterfactual*

We exploit heterogeneity in immigrant inflows across different regions within the UK (see e.g. Altonji and Card 1991)

Correlate (changes in) wages with (changes in) immigration in different regions

Identification: Variation within spatial units over time

Empirical Model

Let w_{prt} denote the p^{th} percentile of the native wage distribution in region r at time t .

We adopt a model

$$\ln w_{prt} = a_{pr} + b_{pt} + c_{prt} + \gamma_p m_{rt} + \varepsilon_{prt}$$

We estimate this model in differences for different quantiles of the wage distribution, using spatial variation over time,

$$\Delta \ln w_{prt} = b_{pt} + \Delta c_{prt} + \gamma_p \Delta m_{rt} + \Delta \varepsilon_{prt}$$

Empirical strategy

- Problems:
  - Endogeneity: immigrants may tend to go to economically successful areas (direction of causation unclear)
  - Measurement Error: due to small sample size. Emphasized by difference estimation.
- Solution (to both): IV, based on previous settlement of immigrants (Bartel 1989, Munshi 2003)
  - Lags
  - Historical settlement from the Census
  - Predicted inflows

Structure of Talk

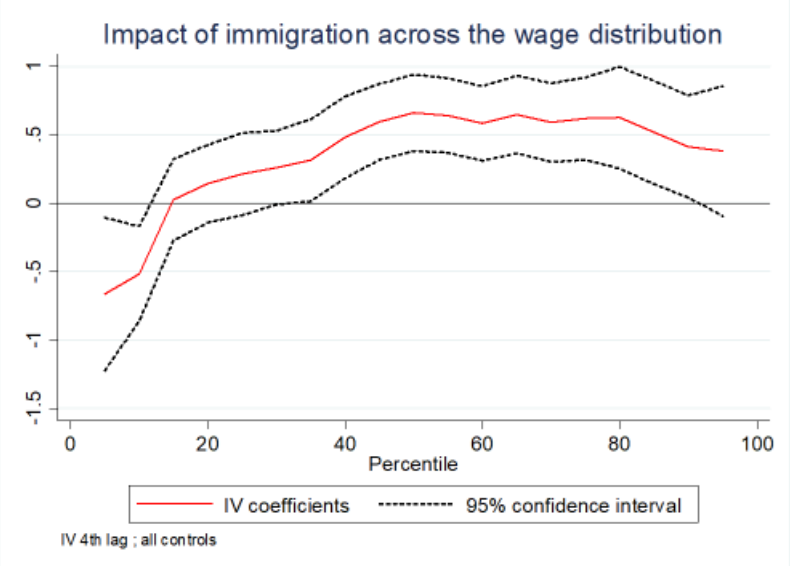
- Immigration to Britain: Data Sources and Some Facts
- Impact of Immigration on Wages: Theory
- Impact of Immigration on Wages: Empirical Estimation
- **Results**
- Conclusion

The Effect of Immigration on Wage Distribution

Dependent Variable	OLS		IV [1991 Immigration Share]		IV [4 period lag]	
	(1)	(2)	(3)	(4)	(5)	(6)
5th Percentile	-0.165	-0.221	-0.353	-0.340	-0.750	-0.665
	(0.383)	(0.383)	(0.181)	(0.186)	(0.286)	(0.282)
10th Percentile	-0.079	-0.094	-0.217	-0.219	-0.536	-0.516
	(0.231)	(0.237)	(0.109)	(0.115)	(0.173)	(0.175)
25th Percentile	0.175	0.124	0.237	0.305	0.119	0.212
	(0.210)	(0.207)	(0.099)	(0.101)	(0.156)	(0.152)
50th Percentile	0.264	0.234	0.409	0.444	0.615	0.660
	(0.192)	(0.190)	(0.091)	(0.093)	(0.144)	(0.141)
75th Percentile	0.407	0.375	0.441	0.500	0.561	0.617
	(0.210)	(0.207)	(0.099)	(0.101)	(0.156)	(0.152)
90th Percentile	0.341	0.314	0.299	0.340	0.379	0.414
	(0.262)	(0.257)	(0.124)	(0.125)	(0.194)	(0.188)
95th Percentile	0.251	0.230	0.301	0.286	0.387	0.381
	(0.325)	(0.327)	(0.153)	(0.159)	(0.241)	(0.239)
F- stat for significance of excluded instruments			172.06	115.53	156.03	163.71
Partial R^2 for first stage regression			0.454	0.463	0.322	0.333
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Other Controls	No	Yes	No	Yes	No	Yes
Observations	136	136	136	136	136	136

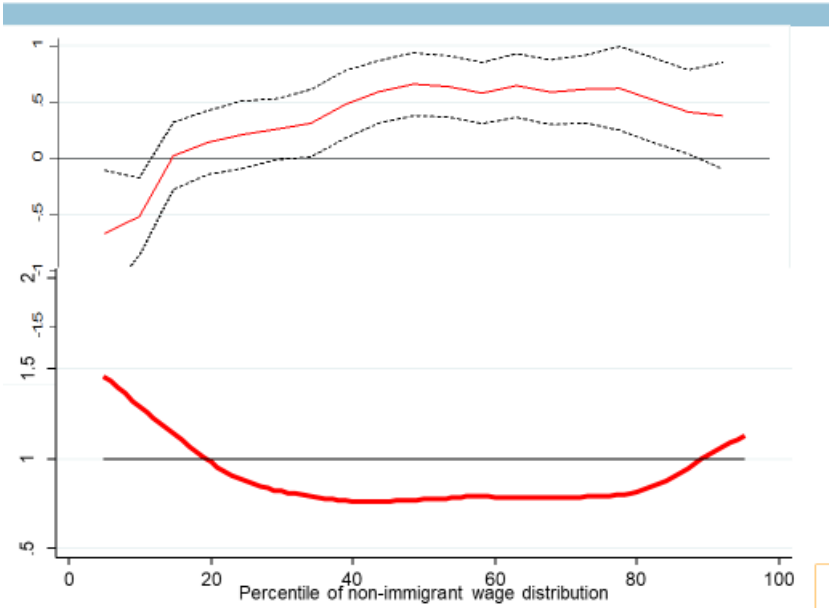
19

The Effect of Immigration on Wage Distribution



2

Wage effects and wage location compared



The Effect of Immigration on Average Wages

Dependent Variable	OLS		IV [1991 Immigration Share]		IV [4 period lag]	
	(1)	(2)	(3)	(4)	(5)	(6)
Average	0.410	0.389	0.213	0.256	0.428	0.465
	(0.187)	(0.181)	(0.088)	(0.088)	(0.138)	(0.133)
Robust average	0.296	0.272	0.268	0.302	0.356	0.396
	(0.156)	(0.153)	(0.074)	(0.074)	(0.116)	(0.112)
Wage index	0.322	0.311	0.100	0.132	0.306	0.338
	(0.168)	(0.169)	(0.079)	(0.083)	(0.124)	(0.124)
Robust index	0.228	0.215	0.168	0.192	0.270	0.301
	(0.137)	(0.139)	(0.064)	(0.068)	(0.101)	(0.102)
F-stat for significance of excluded instruments			172.06	115.53	156.03	163.71
Partial R^2 for first stage regression			0.454	0.463	0.322	0.333
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Other Controls	No	Yes	No	Yes	No	Yes
Observations	136	136	136	136	136	136

Robustness checks: different instruments

Instrumental variable	*Average wage* (1)	*Robust Average* (2)
4th lag of immigrant-native ratio	0.428	0.356
	(0.138)	(0.116)
14th lag of immigrant-native ratio	0.369	0.326
	(0.136)	(0.114)
1991 immigrant-native ratio (Census 1991)	0.213	0.268
	(0.088)	(0.074)
1981 immigrant-native ratio (Census 1981)	0.193	0.258
	(0.093)	(0.077)
change 91-81	0.284	0.300
	(0.082)	(0.068)
Predicted inflow by ethnic group (LFS 91)	0.411	0.320
	(0.168)	(0.140)
Predicted inflow by ethnic group (LFS 85)	0.326	0.266
	(0.186)	(0.155)
Predicted inflow by ethnic group (LFS 81)	0.332	0.291
	(0.172)	(0.144)

Conclusion

- Recent immigrants are well-educated relative to the native population
- However they work in relatively poorly paid jobs, particularly in early years after arrival
- Estimated effects on native wages line up closely with location in the wage distribution
- Growth of wages at low end held back by immigration while wages benefit around the middle
- Average wage growth modestly encouraged by immigration