Introduction and Methods

Something about me

- I joined Tor Vergata last year after 4 years at the University of St Andrews. I obtained my PhD at Queen Mary University in London
- My research focuses on issues in economics of education and urban economics. This is my website: https://sites.google.com/view/lorenzoneri/home-page
- In a set of projects I study how public housing programs affect student achievement in more deprived neighbourhoods
- In other projects I study how management and governance affects the performance of students attending state-funded schools that are independent
- Lately I have been working on issues in Environmental Economics (the effect of air pollution on children's outcomes) and ethnic discrimination

The admin stuff

- All lectures will be regularly held in-person; lecture slides will be posted on the course page
- IMPORTANT: MAKE SURE YOU ENROL IN THE COURSE: I WILL POST THE MATERIAL **ONLY** ON THE COURSE PAGE
- Final exam: March 31st (first session)
- Course Assignment: Presentation [-1/+3] or a Stata exercise [pass/fail]. Presentations will be held in the last week of the course
- **Office hours**: **Wed 2-3pm**. Please email me if you plan to come to office hous or if you'd like to meet outside of office hours
- **Contact**: lorenzo.neri@uniroma2.it
- Please read the Syllabus on the course webpage carefully and let me know if anything is unclear

Module Outline

- Empirical strategies
- Labor supply
- Labor demand
- Minimum wage
- Immigration
- Compensating wage differentials
- Human capital
- Inequality
- Intergenerational mobility
- Gender/Race differences
- Incentive pay
- Unemployment

- Labor economics is a heavily empirical field, more than others.
- Primarily micro data.
- Labor economics distinguished by the use of cutting edge econometric techniques.
- Many innovations in econometrics (outside of time-series econometrics) motivated by labor economics questions:
  - sample selection
  - censored data
  - survival analysis
  - quantile regression
  - identification issues: instrumental variables, quasi-experimental methods.

Broadly, empirical analysis can be divided into two types:

1. Descriptive analysis:
   - establish facts, yield new insights into economic trends.
   - Sherlock Holmes: "It is a capital offense to theorize before all the facts are in."
2. Causal inference:
   - what is the effect of a particular intervention or policy?
   - estimate features of behavioral relationships suggested by economic theory.

The Structural versus Reduced-Form Debate

Much debate about the best way to do causal inference: structural versus reduced-form ("quasi-experimental" or "experimentalist") approach.

- Structural approach:
  - relies heavily on economic theory to guide empirical work.
  - recover the primitive parameters of interest, then use them and theory to make predictions about behavior.
  - shortcomings:
    1. identification often relies on untestable functional form assumptions;
    2. computationally demanding.

The Structural versus Reduced-Form Debate

- Reduced-form approach:
  - focus is on identifying parameter of interest from specific events or situations.
  - ideal: controlled randomized experiment.
  - shortcomings:
    1. generalizability of findings;
    2. theory is often only in the background, or informal.

The fundamental problem of causal inference

- Causal questions in labor economics:
  - What is the effect of education/unionization/veteran status on earnings?
  - What is the effect of immigration on wages and employment?
  - What is the effect of unemployment benefits on the duration of unemployment?
  - What is the effect of taxes on labor supply?

- In all cases, we are seeking the answer to a "what if" question (a counterfactual):
  - What would the earnings of individual i had been, *if* he had chosen not to go to college instead of going to college?

Potential Outcomes

- Two possible treatments, 0 (the control group), and 1 (the treatment group). Every individual in the sample has two potential outcomes:

$$\begin{aligned} y_{i1} &: \quad \text{outcome if treated} \\ y_{i0} &: \quad \text{outcome if not treated.} \end{aligned}$$

- No matter what the actual treatment one receives, the potential outcome is always well defined.
- Treatment is going to college, and outcome is earnings. Potential outcomes:
  - y_{i1} are earnings of individual i if he goes to college
  - y_{i0} are the earnings of individual i if he does not go to college.
- These potential outcomes are well-defined even *before the actual treatment is received*.

The fundamental problem of causal inference

- The key problem of causal inference: we can only observe one outcome at a time, we can *never* observe the counterfactual...
- How do we learn about counterfactuals?
- Gold standard: controlled, randomized experiments.
  - Subjects randomly assigned to treatment and control groups.
  - Because of randomization, we can be sure that all other factors (observable and unobservable) are uncorrelated with treatment status.
  - Therefore, comparison between treated and control subjects answers the "what if" question.
- Problems
  - True randomized experiments are rare (but growing) in economics.
  - Even with true randomization, many things can go wrong in the experimental design.

Identification strategies

Five basic identification strategies (Angrist and Pischke, *Mastering Metrics*, 2014).

1. Random assignment.
2. Control for confounding variables (i.e., regression).
3. Fixed effects and differences-in-differences.
4. Instrumental variables.
5. Regression discontinuity.

Random Assignment

- Best way to identify causal effects is to conduct randomized experiment.
- Some well-known experiments
  - Job training programs in the 1970s and 1980s
  - Tennesee STAR program: what is the effect of smaller class sizes on student achievement.
  - Oregon Health Insurance Experiment (low-income uninsured randomly assigned to receive Medicaid).
- In good experimental design, observable and unobservable characteristics of treatment and control group should be *balanced*.
- Potential weaknesses:
  - General equilibrium effects
  - Low external validity

Regression: Control for confounding variables

- Example: the relationship between schooling and log earnings (Y).
- People with more schooling have higher earnings: human capital, or wealthier parents and/or higher ability?
- Possible solution: control for confounding factors in a linear regression framework:

$$Y_i = X_i'\beta + \rho S_i + e_i$$

- β and ρ are regression coefficients. Can they be given a causal interpretation?

- (Structural) relationship determining log earnings as a function of schooling for individual i:

$$Y_{S,i} \equiv f_i(S) = \beta_0 + \rho S + \eta_i.$$

- OLS coefficient of ρ in regression of Y_i on S_i has probability limit

$$\begin{aligned} p\lim \hat{\rho} &= \frac{Cov(Y_i, S_i)}{V(S_i)} = \frac{Cov(\beta_0 + \rho S_i + \eta_i, S_i)}{V(S_i)} \\ & \rho + \frac{Cov(S_i, \eta_i)}{V(S_i)} \end{aligned}$$

- Typically, $Cov(S_i, \eta_i)$ is not zero. Omitted variable bias.

- As an example, assume that
$$\eta_i = X_i'\beta + \varepsilon_i$$
- with X_i and ε_i uncorrelated by construction, and X_i is unobserved to the econometrician.
- Key identifying assumption:
$$Cov\left(S_i, \varepsilon_i\right) = 0$$
- In other words, the only reason for why S_i and η_i are correlated is the observable variables, X_i. *Selection on observables*.

- In this framework:

$$\begin{aligned}\frac{Cov\left(Y_i, S_i\right)}{V\left(S_i\right)} &= \frac{Cov\left(\beta_0 + \rho S + X_i' \beta + \varepsilon_i, S_i\right)}{V\left(S_i\right)} \\ &= \rho + \frac{Cov\left(S_i, X_i\right)}{V\left(S_i\right)} = \rho + \Gamma_{SX}' \beta\end{aligned}$$

- This is the omitted variable bias formula: the sign of the omitted variable bias depends on the correlation between S and X (Γ_{SX}') and on β (the effect of X on the dependent variable).
- Schooling example:
  - Omitted variable: ability
  - $\Gamma_{SX} > 0$: high ability people also get more schooling
  - $\beta > 0$: high ability people have higher earnings
- Implication: omitted variable bias is positive, OLS estimate of ρ biased upwards.

Fixed effects

- Example: the effect of union status on earnings:
- Potential outcome framework:

$$\begin{aligned} Y_{0i} &= X_i'\beta + \varepsilon_i \\ Y_{1i} &= Y_{0i} + \delta \end{aligned}$$

- Y_{0i} is outcome if not unionized, Y_{1i} outcome if unionized. Only one is ever observed.
- The observed outcome:

$$Y_i = X_i'\beta + U_i\delta + \varepsilon_i$$

- But union status may be related to potential wages, even after controlling for $X_{i.}$

- With multiple observations on individual i, we can write:

$$
\begin{aligned}
Y_{0it} &= X_{it}'\beta + \lambda\alpha_i + \gamma_{it} \\
Y_{1it} &= X_{it}'\beta + U_{it}\delta + \lambda\alpha_i + \gamma_{it}
\end{aligned}
$$

- Key identifying assumption: individual fixed effect α_i is truly time-invariant, λ also time invariant.
- We can use differencing to eliminate α_i, the part of potential earnings that is correlated with union status.

$$Y_{it} - Y_{it-k} = X_{it}'\beta - X_{it-k}'\beta + U_{it}\delta - U_{it-k}\delta + (\gamma_{it} - \gamma_{it-k}).$$

- Commonly used estimator: deviations from means or "within" estimator

$$Y_{it} - \bar{Y}_i = \beta'(X_{it} - \bar{X}_i) + \delta(U_{it} - \bar{U}_i) + (\gamma_{it} - \bar{\gamma}_i).$$

Weaknesses of fixed effects

- Can only estimate coefficients on time-varying variables.
- Estimator not consistent if there are lagged dependent variables (in many applications reasonable that there are).
- Bias from measurement error often exacerbated.
- Must be careful that fixed effects are truly fixed.

Differences in differences

- Simple panel data method applied to sets of group means in cases where certain groups are exposed to the causing variable of interest, while others are not.
- Example: the effect of immigration on earnings using the Mariel Boatlift (large migration episode from Cuba to Miami in 1980; Card 1990).

	Before	After	After-Before
Miami	$\bar{Y}_{T0}$	$\bar{Y}_{T1}$	$\bar{Y}_{T1} - \bar{Y}_{T0}$
Comparison cities	$\bar{Y}_{C0}$	$\bar{Y}_{C1}$	$\bar{Y}_{C1} - \bar{Y}_{C0}$
Miami - Comparison	$\bar{Y}_{T0} - \bar{Y}_{C0}$	$\bar{Y}_{T1} - \bar{Y}_{C1}$	$(\bar{Y}_{T1} - \bar{Y}_{T0}) - (\bar{Y}_{C1} - \bar{Y}_{C0})$

Table 4
Differences-in-differences estimates of the effect of immigration on unemployment[a]

	Group	Year		
		1979 (1)	1981 (2)	1981–1979 (3)
	Whites			
(1)	Miami	5.1 (1.1)	3.9 (0.9)	−1.2 (1.4)
(2)	Comparison cities	4.4 (0.3)	4.3 (0.3)	−0.1 (0.4)
(3)	Miami-Comparison Difference	0.7 (1.1)	−0.4 (0.95)	−1.1 (1.5)
	Blacks			
(4)	Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
(5)	Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
(6)	Miami-Comparison Difference	−2.0 (1.9)	−3.0 (2.0)	−1.0 (2.8)

[a] Notes: Adapted from Card (1990, Tables 3 and 6). Standard errors are shown in parentheses.

- Convenient econometric framework for DD:

$$\begin{aligned} E\left(Y_{0i}|c,t\right) &= \beta_t + \gamma_c \\ E\left(Y_{1i}|c,t\right) &= Y_{0i} + \delta \end{aligned}$$

- The observed outcome can be written as

$$Y_i = \beta_t + \gamma_c + \delta M_i + \varepsilon_i$$

where $M_i = 1$ if $c = Miami$ and $t = 1981$.

$$\begin{aligned} & \left\{E\left[Y_i|c = Miami, t = 1981\right] - E\left[Y_i|c = Miami, t = 1979\right]\right\} \\ & - \left\{E\left[Y_i|c = Comp, t = 1981\right] - E\left[Y_i|c = Comp, t = 1979\right]\right\} \\ = \; & \delta \end{aligned}$$

- DD effect can be estimated by a regression of outcome on Miami dummy, "after" dummy, and the interaction between the two.
- Easy to incorporate additional regressors:

$$Y_i = X_i'\beta_0 + \beta_t + \gamma_c + \delta M_i + \varepsilon_i$$

Weaknesses of DD

- Key identifying assumption is that the interaction terms are zero at the time of the intervention.
- In other words, no differences in *underlying trends* between Miami and comparison cities at the time of the intervention.
- Many DD studies often show that there is no DD effect in the case of a "placebo" intervention (i.e., no difference in the 1979-1977 DD).

Instrumental variables

- Example: the effect of veteran status on earnings
- Potential outcomes:

$$\begin{aligned} Y_{0i} &= \beta_0 + \eta_i \\ Y_{1i} &= Y_{0i} + \delta \end{aligned}$$

- Observed outcome

$$Y_i = \beta_0 + D_i\delta + \eta_i$$

- D_i not randomly assigned, so OLS is biased. For example, people with low potential civilian earnings more likely to serve in the military.

- But suppose researcher has access to instrument Z that *is* randomly assigned.
- Instrument must be correlated with D, but uncorrelated with potential outcome.
- Example: draft lottery number (Angrist 1990)
  - Between 1970 and 1972, random sequence numbers were associated to each birth date in cohorts of 19-year olds.
  - Those with low lottery numbers were drafted into the military.
  - Draft eligibility and veteran status not perfectly correlated. People with low numbers could still obtain deferments, people with high numbers could still volunteer.

- IV estimator:

$$\hat{\delta}_{IV} = \frac{Cov\left(Y_i, Z_i\right)}{Cov\left(D_i, Z_i\right)}$$

- Why?

$$\begin{aligned} Cov\left(Y_i, Z_i\right) &= Cov\left(\beta_0 + \delta D_i + \eta_i, Z_i\right) \\ &= \delta Cov\left(D_i, Z_i\right). \end{aligned}$$

- Since in this case Z_i is binary, this takes on a particular simple form (the *Wald* estimator)

$$\hat{\delta}_{IV} = \hat{\delta}_{WALD} = \frac{E\left(Y|Z=1\right) - E\left(Y|Z=0\right)}{E\left(D|Z=1\right) - E\left(D|Z=0\right)}.$$

Table 5
IV estimates of the effects of military service on white men[a]

Earnings year	Earnings		Veteran status		Wald estimate of veteran effect
	Mean (1)	Eligibility effect (2)	Mean (3)	Eligibility effect (4)	(5)
A. Men born 1950					
1981	16461	−435.8 (210.5)	0.267	0.159 (0.040)	−2741 (1324)
1970	2758	−233.8 (39.7)			−1470 (250)
1969	2299	−2.0 (34.5)			
B. Men born 1951					
1981	16049	−358.3 (203.6)	0.197	0.136 (0.043)	−2635 (1497)
1971	2947	−298.2 (41.7)			−2193 (307)
1970	2379	−44.8 (36.7)			
C. Men born 1953 (no one drafted)					
1981	14762	34.3 (199.0)	0.130	0.043 (0.037)	No first stage
1972	3989	−56.5 (54.8)			
1971	2803	2.1 (42.9)			

[a] Note: Adapted from Angrist (1990, Tables 2 and 3), and unpublished author tabulations. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 observations with earnings in each cohort.

- If additional control variables, IV becomes 2SLS (two stage least squares estimator):
  1. regress D_i on Z_i, X_i, take fitted values $\hat{D}_i$.
  2. regress Y_i on $\hat{D}_i$ and X_i. Coefficient on $\hat{D}_i$ is $\hat{\delta}_{2SLS}$.
- Usually done in one stage with statistical software programs, gets standard errors right.

Weaknesses of IV framework

- Key identifying assumption is not testable. Must rely on economic theory, institutional knowledge.
- Random assignment alone does not guarantee unbiased estimates: in the draft lottery example, people with low numbers could have chosen to stay in college more, hence the lottery number may have a direct effect on earnings (exclusion restriction).
- Finite sample bias: weak instruments, too many instruments, small samples $\Rightarrow$ all lead to biased inference, IV in finite samples is biased towards OLS.

Regression Discontinuity

- Sometimes assignment to treatment depends on a *forcing variable*, s: agents with $s \geq s^*$ are assigned to treatment, agents with $s < s^*$ are assigned to the control group.
- Agents to the right and the left of the threshold s^* should be identical in all respects, except for their treatment status.
- Basic idea: estimate the effect of the treatment by comparing outcomes of agents just to the right and just to the left of the threshold.

Sharp regression discontinuity design

- assignment to treatment depends deterministically on s.
  - eligibility to receive tax credit based on date of birth.
  - effect of unionization on firm value, using results from union elections (Lee and Mas, 2009).
  - effect of gender on politician's survival in office (Gagliarducci and Paserman, 2012).
- Implementation: estimate (often nonparametrically) $E\left(Y|s\right)$ separately on both sides of the threshold, treatment effect is

$$\lim_{\varepsilon\to 0} E\left(Y_i|s_i = s^* + \varepsilon\right) - E\left(Y_i|s_i = s^* - e\right)$$

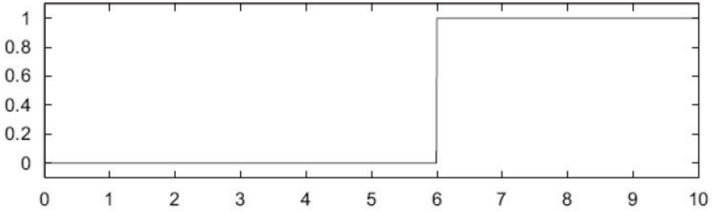

Fig. 1. Assignment probabilities (SRD).

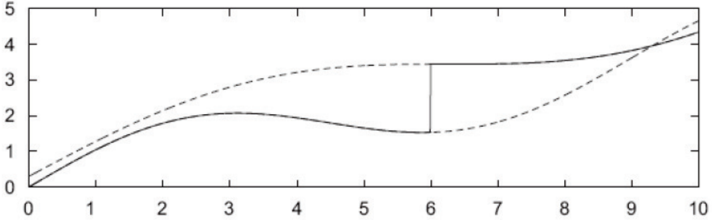

Fig. 2. Potential and observed outcome regression functions.

- **Fuzzy regression discontinuity design**
- Probability of assignment jumps discontinuously at s^* (but not from 0 to 1).
- Example 1: class size rules (Angrist and Lavy, 1997)
  - Many countries have rules that classrooms must be split if enrollment is larger than some threshold (Israel $=$ 40; Chile $=$ 45; California $=$ 20).
  - But some schools will split classrooms even before reaching the threshold, while others will violate the rule.

- Example 2: The effect of financial aid on college attendance (Van der Klaauw, IER 2002).
  - Based on objective admission credentials (SAT scores and high school GPA), an aggregate score s is calculated, and prospective students are put into separate categories (e.g., high, medium, low).
  - The probability of being offered financial aid varies discontinuously at the threshold values of s.
  - But not from 0 to 1: other factors enter decision of the admissions officer (extracurricular activities, essay, personal circumstances, etc.)
  - Also, prospective students with high s more likely to have attractive outside offers.

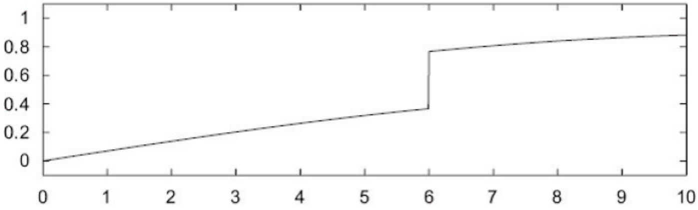

Fig. 3. Assignment probabilities (FRD).

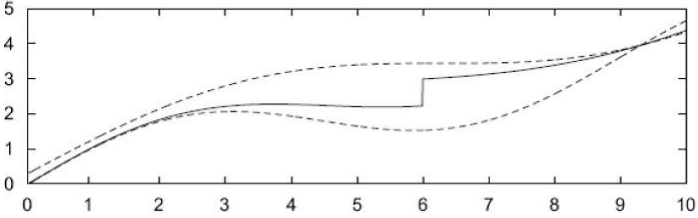

Fig. 4. Potential and observed outcome regression (FRD).

Regression discontinuity estimand is

$$\lim_{\varepsilon \to 0} \frac{E\left(Y_i | s_i = s^* + \varepsilon\right) - E\left(Y_i | s_i = s^* - \varepsilon\right)}{P\left(D_i = 1 | s_i = s^* + \varepsilon\right) - P\left(D_i = 1 | s_i = s^* - \varepsilon\right)}.$$

Very similar to IV.

Weaknesses of RD design

- Can estimate treatment effect only at the discontinuity point. May be non-informative about the effect at other points of the distribution.
- Must have large samples to estimate precisely the effect, otherwise relies on functional form assumptions and data not close to the threshold.
- Must be careful that other variables are not discontinuous at the threshold, and that the density is not discontinuous at the threshold (i.e., no manipulation)
- Fuzzy design is essentially IV, see weaknesses of IV design.