Big Data for Public Policy

Applied Micro Methods II

Sergio Galletta

ETHZ Zurich

23/03/2023

Synthetic control method

- With this method, it is possible to generate a "**synthetic control**" for a specific observation i, enabling the estimation of its causal impact.

- By assigning weights to the Y values of untreated units, a "synthetic control" is produced

$$\hat{Y}_{t,post}(0) = \mu + \sum_{i \in c} \omega_i Y_{i,T}$$

- Typically, it is necessary to estimate the ω_i, which are formed by minimizing the distance between covariates in the pre-period.

$$\{\hat{\omega}\}_i = \arg\min_{\boldsymbol{W}} ||\boldsymbol{X}_{\text{treat}} - \boldsymbol{X}_{\text{control}} \boldsymbol{W}||$$

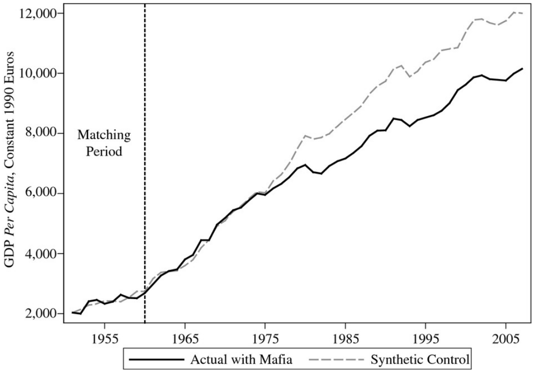
Synthetic control method

- Importantly, $\boldsymbol{X}$ can include both lagged outcomes and covariates.
- Reconnecting with the idea of what is observable and what is not:
  - Unobserved outcomes: $\mathrm{Y}_{t,post}(0)$, $\mathrm{Y}_{c,post}(1)$
  - Observed outcomes: $\mathrm{Y}_{t,post}(1)$, $\mathrm{Y}_{c,post}(0)$
  - Observed covariates / predictors: $\mathrm{Y}_{t,pre}(0)$, $\mathrm{Y}_{c,pre}(0)$, X_t, X_c
- Relevant method if one wants to study just one treated unit

Example

The Economic Costs of Organised Crime - Pinotti 2015

- Is organized crime good or bad for the economy?
- Expansion of Mafia to regions previously unaffected (Apulia and Basilicata)
- From the minimization problem, the control group is created by weights from Abruzzo (0.624) and Molise (0.376).
- GDP growth slows down after mafia activities expanded to new regions



Example

Tax Reform and Foreign Inventors - Akcigit et al 2016

- Do people respond to changes in tax burden?
- In 1992, Denmark reduced taxes on foreign researchers
- From the minimization problem, the control group is created by weights from Switzerland, Canada, and Portugal
- Increased in the share of foreign inventors

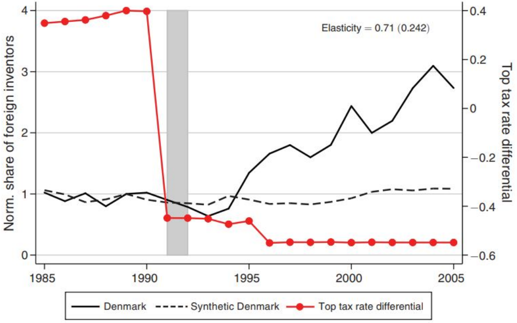


Figure 9. Denmark's 1992 Tax Reform and Foreign Inventors

Instrumental variables

- **Instrumental variables (IV)** is the most popular solution for dealing with endogenous treatments

- The 2021 Nobel prize for economics was assigned to researchers that linked the potential outcome framework with IV and introduced the concept of **local average treatment effect** (LATE), the actual type of effect that IV delivers

Instrumental variables

- **Instrumental variables (IV)** is the most popular solution for dealing with endogenous treatments

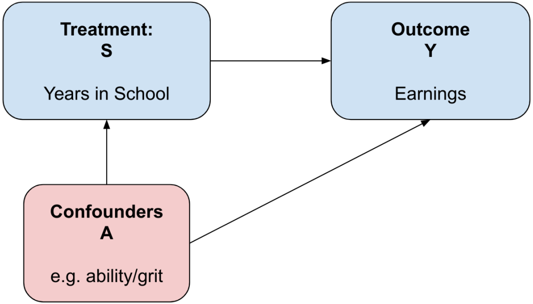
- The 2021 Nobel prize for economics was assigned to researchers that linked the potential outcome framework with IV and introduced the concept of **local average treatment effect** (LATE), the actual type of effect that IV delivers

- Let's go back to the link between education and income

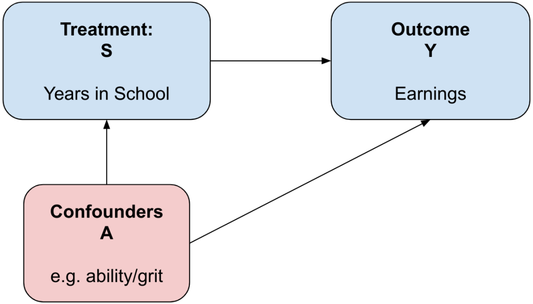
$$
\begin{aligned}
Y_i &= \alpha + \rho S_i + \epsilon_i \\
Y_i &= \alpha + \rho S_i + \phi \underbrace{A_i}_{\text{unobs}} + \eta_i
\end{aligned}
$$

- OLS estimates for $\hat{\rho}$ will be biased.

Instrumental Variables: Main Intuition

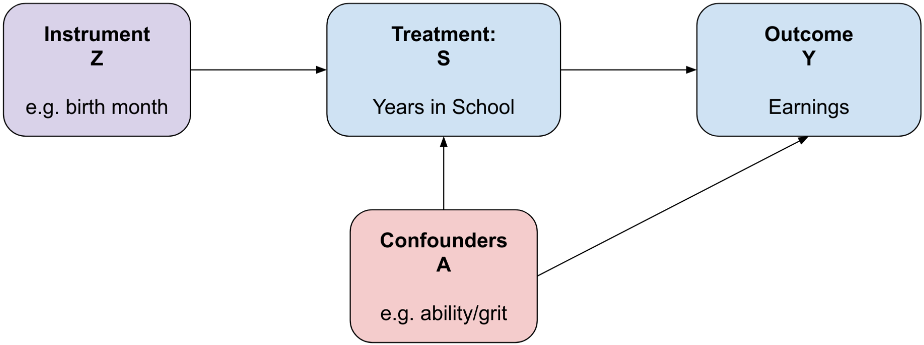


Instrumental Variables: Main Intuition



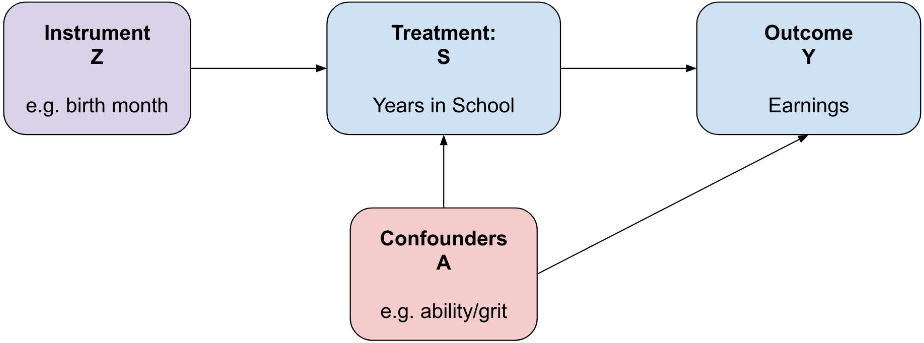
Instrumental Variable (IV): to identify a variable, that is correlated with S_i, but not correlated with anything else affecting Y_i.

Instrumental Variables: Main Intuition



- We identify a source of variation in treatment assignment that is as good as random – orthogonal to any relevant unobserved confounder.
- We compare individuals that, due to the instrument, are shifted between the control group and the treatment group.

What is a valid instrumental variable?



1. Correlated with the causal variable, e.g. S_i:

$$\text{Cov}[Z_i, S_i] \neq 0$$

2. Uncorrelated with any other determinants of outcome Y:

$$\text{Cov}[Z_i, \epsilon_i] = 0$$

What is a valid instrumental variable?

(1) Exogeneity: No unobserved factors affect both the outcome and the instrument:

$$\epsilon_i \not\rightarrow Z_i$$

- **No "Z-confounders"**

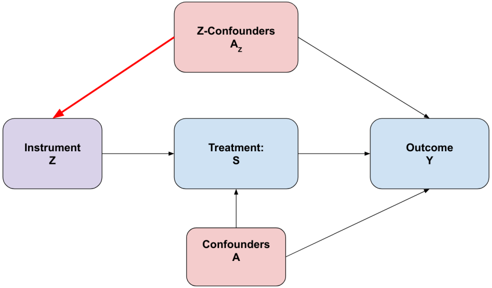
What is a valid instrumental variable?

(1) Exogeneity: No unobserved factors affect both the outcome and the instrument:

$$\epsilon_i \nrightarrow Z_i$$

- **No "Z-confounders"**

Violation of exogeneity:



What is a valid instrumental variable?

(1) **Exogeneity**: No unobserved factors affect both the outcome and the instrument:

$$\epsilon_i \nrightarrow Z_i$$

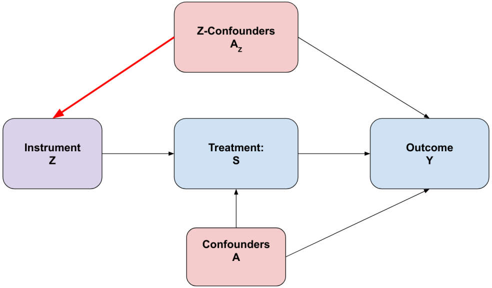
- **No "Z-confounders"**

(2) **Exclusion**: Instrument only affects outcome through treatment variable:

$$Z_i \nrightarrow \epsilon_i$$

- **"single mediator" condition**

Violation of exogeneity:



What is a valid instrumental variable?

(1) Exogeneity: No unobserved factors affect both the outcome and the instrument:

$$\epsilon_i \not\rightarrow Z_i$$

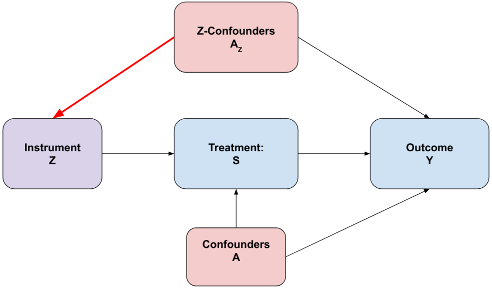
- **No "Z-confounders"**

(2) Exclusion: Instrument only affects outcome through treatment variable:

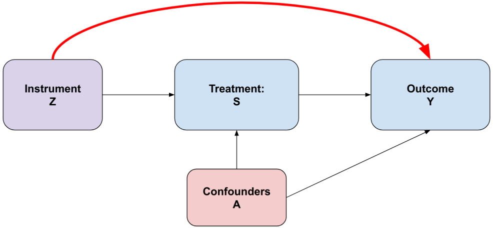
$$Z_i \not\rightarrow \epsilon_i$$

- **"single mediator" condition**

Violation of exogeneity:



Violation of exclusion:



Good instruments are hard to find

- Good instruments come from a combination of three ingredients:
  - Good institutional knowledge
  - Economic theory
  - Last but not least: Originality

Good instruments are hard to find

- Good instruments come from a combination of three ingredients:
  - Good institutional knowledge
  - Economic theory
  - Last but not least: Originality
- Some usual sources of instruments:
  - Nature (e.g., genes, weather)
  - Assignment rules (e.g., random assignment of judges to cases)
  - 'Natural' experiments (e.g. the quarter of birth, conscription lottery, electoral timing...)

Two-Stage Least Squares (2SLS)

IV estimates are equivalent to running two separate OLS regressions:

1. Estimate"first stage", regressing treatment on instrument:

$$S_i = \gamma Z_i + \nu_i$$

Two-Stage Least Squares (2SLS)

IV estimates are equivalent to running two separate OLS regressions:

1. Estimate“first stage”, regressing treatment on instrument:

$$S_i = \gamma Z_i + \nu_i$$

2. Form prediction $\hat{S}_i = \hat{\gamma} Z_i$ and estimate the“second stage”, regressing outcome on first-stage-predicted treatment:

$$Y_i = \rho \hat{S}_i + \epsilon_i$$

Two-Stage Least Squares (2SLS)

IV estimates are equivalent to running two separate OLS regressions:

1. Estimate"first stage", regressing treatment on instrument:

$$S_i = \gamma Z_i + \nu_i$$

2. Form prediction $\hat{S}_i = \hat{\gamma} Z_i$ and estimate the"second stage", regressing outcome on first-stage-predicted treatment:

$$Y_i = \rho \hat{S}_i + \epsilon_i$$

- First stage is driven by "compliers" (responders to instrument).
- Standard 2SLS estimates give a "local average treatment effect" on the complier population.

Can we test validity of IV?

- Is Z_i correlated with causal variable of interest, S_i?
  - YES: check for the significance of the first stage (first-stage F-statistic)
  - The standard is F $>$ 10, but recent studies show you might need more
  - With weak instruments IV bias towards the OLS

Can we test validity of IV?

- Is Z_i correlated with causal variable of interest, S_i?
  - YES: check for the significance of the first stage (first-stage F-statistic)
  - The standard is F $>$ 10, but recent studies show you might need more
  - With weak instruments IV bias towards the OLS
- Is Z_i uncorrelated with any other determinants of Y_i?
  - Untestable, use logic and theory to argue in favor the assumption
  - But often indirect ways to probe exogeneity and exclusion

Can we test validity of IV?

- Is Z_i correlated with causal variable of interest, S_i?
  - YES: check for the significance of the first stage (first-stage F-statistic)
  - The standard is F $>$ 10, but recent studies show you might need more
  - With weak instruments IV bias towards the OLS
- Is Z_i uncorrelated with any other determinants of Y_i?
  - Untestable, use logic and theory to argue in favor the assumption
  - But often indirect ways to probe exogeneity and exclusion
- Additional assumption that is important (and untestable)
  - Monotonicity: the instrument(s) should have a monotonic relationship with the endogenous explanatory variable(s)

Reduced Form

“Reduced Form” (RF) means regressing the outcome directly on the instrument:

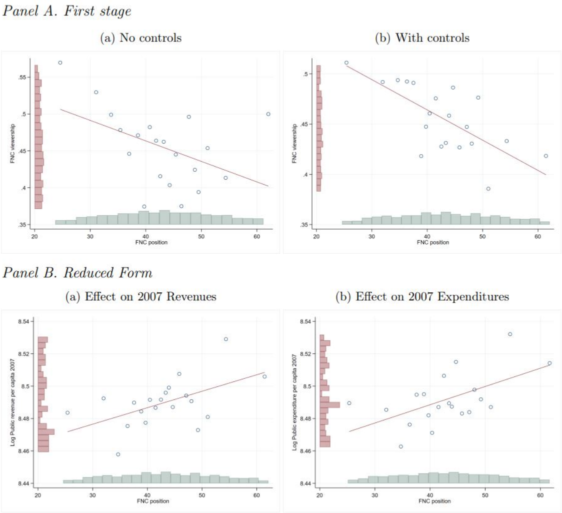
$$Y_i = \alpha + \phi Z_i + \epsilon_i$$

- Papers will normally report this along with 2SLS estimates.
- For causal interpretation, RF requires exogeneity but not exclusion.

Example

Media and local finance - Ash and Galletta 2023

- How national cable news affects local public policy?
- Variation in channel position depending on the area of residence
- Higher channel position lower viewership
- Fox News did decrease the size of local budgets



Panel A. First stage

(a) No controls

(b) With controls

Panel B. Reduced Form

(a) Effect on 2007 Revenues

(b) Effect on 2007 Expenditures

Example

War and Tax Evasion - Galletta and Giommoni 2023

- Does exposure to war violence affect tax evasion?
- Death of a relative in WWI as treatment
- Exogenous allocation of soldiers to more/less risky military units
- Higher tax evasion when a relative died during the war

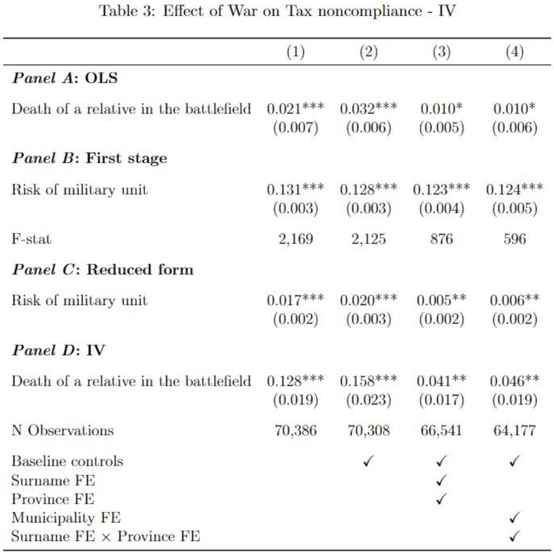


Table 3: Effect of War on Tax noncompliance - IV

	(1)	(2)	(3)	(4)
***Panel A*: OLS**				
Death of a relative in the battlefield	0.021*** (0.007)	0.032*** (0.006)	0.010* (0.005)	0.010* (0.006)
***Panel B*: First stage**				
Risk of military unit	0.131*** (0.003)	0.128*** (0.003)	0.123*** (0.004)	0.124*** (0.005)
F-stat	2,169	2,125	876	596
***Panel C*: Reduced form**				
Risk of military unit	0.017*** (0.002)	0.020*** (0.003)	0.005** (0.002)	0.006** (0.002)
***Panel D*: IV**				
Death of a relative in the battlefield	0.128*** (0.019)	0.158*** (0.023)	0.041** (0.017)	0.046** (0.019)
N Observations	70,386	70,308	66,541	64,177
Baseline controls		✓	✓	✓
Surname FE			✓	
Province FE			✓	
Municipality FE				✓
Surname FE × Province FE				✓

Regression Discontinuity Design (RDD)

- **Regression Discontinuity Design (RDD)** is a quasi-experimental design that has gained popularity among researchers because it can provide more credible causal estimates than other designs.

- It exploits that individuals are assigned to treatment or control groups based on a running variable (e.g., test score, distance, or class size) with a discontinuity at a certain threshold or **cutoff point**.

Sharp vs Fuzzy Regression Discontinuity Design

- In the Sharp RDD, individuals who score above the cutoff receive the treatment, and those who score below the cutoff do not.

- In the Fuzzy RDD, the probability of receiving the treatment changes discontinuously at the cutoff, but not all individuals who score above the cutoff receive the treatment.

- Fuzzy RDD is essentially equivalent to an Instrumental Variables (IV) design, where the running variable serves as the instrument for the treatment.

Notation and Key Assumptions

- Let $Y_i(0)$ and $Y_i(1)$ be the potential outcomes of individual i when they do not receive the treatment and when they do, respectively.

- Let D_i be the treatment indicator such that $D_i = 1$ if individual i receives the treatment, and $D_i = 0$ otherwise.

- Let Z_i be the running variable that assigns individuals to treatment or control.

- We assume that Z_i **has a discontinuity at a threshold** value z_0, where the treatment is assigned to individuals with $Z_i \geq z_0$.

- We assume that $Y_i(0)$ **and** $Y_i(1)$ **are continuous in** Z_i **around the threshold value**, which allows us to estimate the LATE at the threshold.

Regression Discontinuity Design (RDD)

- If the previous assumptions hold

$$\tau_{ATE} = E(Y_i(1) - Y_i(0)|Z_i = 0) = \lim_{z \downarrow 0} E(Y_i|Z_i = z) - \lim_{z \uparrow 0} E(Y_i|Z_i = z)$$

- But, this is a very particular subgroup of individuals right at the cutoff

Regression Discontinuity Design (RDD)

- The basic RDD model can be expressed as:

$$Y_i = \alpha + \beta T_i + \gamma(Z_i - c) + \epsilon_i$$

- where Y_i is the outcome variable
- T_i is a binary treatment
- Z_i is the continuous variable used for assignment
- c is the cutoff point, and ϵ_i is the error term
- Local linear regression method by selecting a small neighborhood around the cutoff point.

RDD Check list

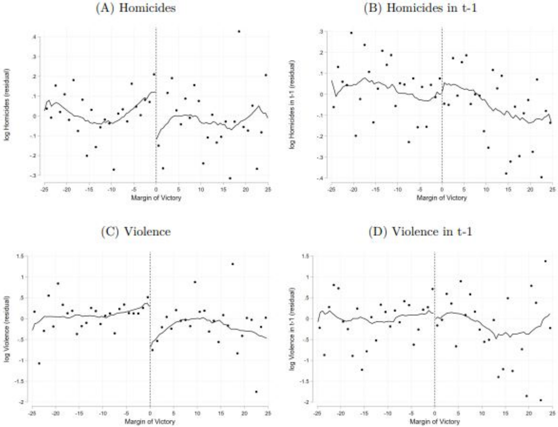
- A graphical representation and test of “balance” and first stage (if fuzzy)
- Permutation test of characteristic at cutoff
- The density of the forcing variable (Mcrary test)
- Placebo checks
- A graphical representation of the outcomes (what we’ve already seen)
- Estimates based on optimal bandwidth choice and robust inference, using local linear analysis
  - These decisions vary depending on running variable. If discrete running variable, need to account for discreteness (Kolesar and Rothe (2018))
  - Should use local linear regression, and not global polynomials (Gelman and Imbens)
- Robustness analysis along bandwidth choice (and other tuning parameters)
  - Present this graphically

Example

Female Mayor and Violence against Women - Bochenkova, Buonanno and Galletta, 2022

- Does female mayor influence violence against women?
- Compare Brazilian municipalities where a female candidate barely won to those where a female candidate barely lost mayoral elections
- Winning is random for those close to 50%
- Yes, having female mayor reduces crime against women

Figure 1: Female Mayor and Violence against Women



(A) Homicides

(B) Homicides in t-1

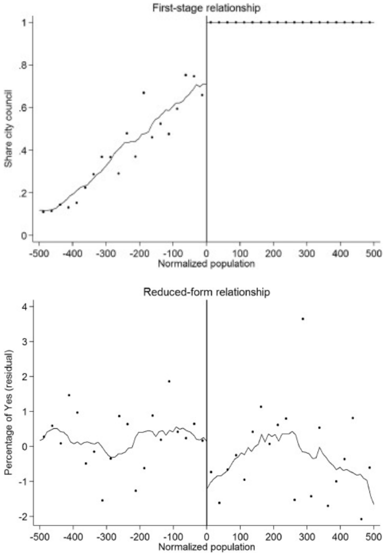
(C) Violence

(D) Violence in t-1

Example

Direct democracy and social preferences - Galletta, 2021

- Does political institutions affect preferences for redistribution?
- Exploits a discrete change in the probability that a municipality has representative democracy based on a population threshold
- $\geq$ 800 inhabitants adopt a city council
- Representative democracy reduces vote shares in favor of spending by around 5 p.p.



To recap

- We care about causality
- Potential outcome framework
- Quasi-experimental methods
  - Diff-in-Diff
  - Synthetic controls
  - IV
  - RDD