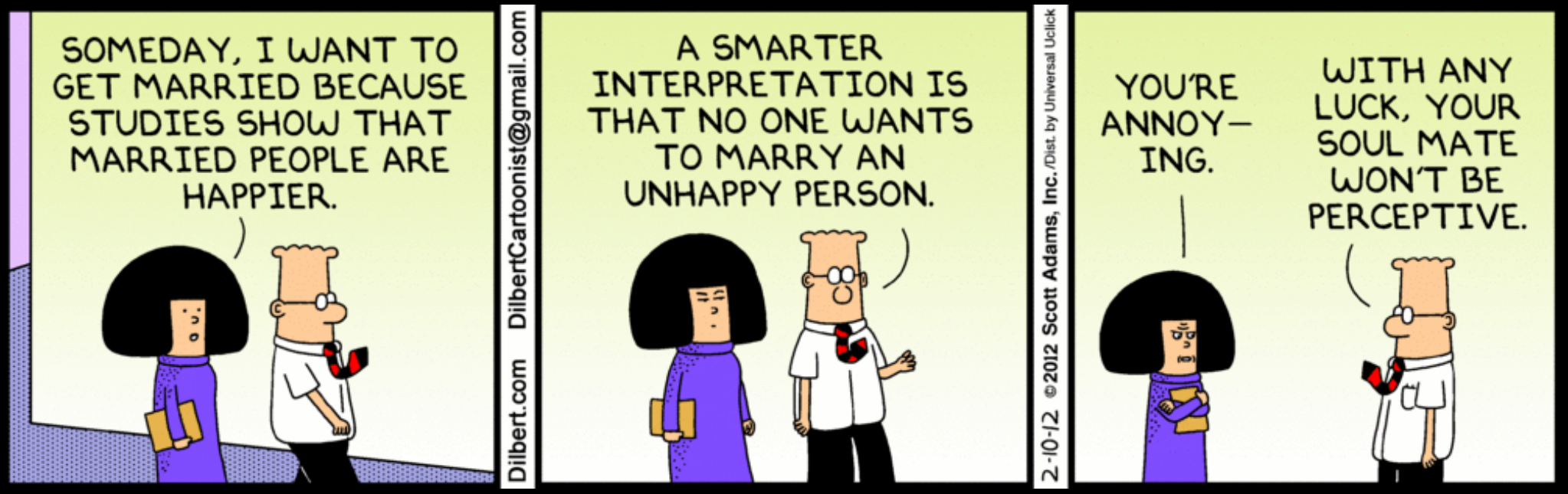


INFO 251: Applied Machine Learning

Nonexperimental Methods for Causal Inference: Instrumental Variables and Regression Discontinuity

Key Concepts (previous lecture)

- Regression for impact evaluation
  - Estimating treat vs. control
  - Estimating Pre vs. Post
  - Estimating difference-in-difference
- Control variables
- Interaction variables
- Heterogeneous treatment effects
- Double difference estimation
- Cross-sectional vs. panel data
- Fixed effects (revisited)

Course Outline

- Causal Inference and Research Design
  - Experimental methods
  - **Non-experimental methods**
- Machine Learning
  - Design of Machine Learning Experiments
  - Linear Models and Gradient Descent
  - Non-linear models
  - Fairness and Bias in ML
  - Neural models
  - Deep Learning
  - Practicalities
  - Unsupervised Learning
- Special topics

Outline

- Instrumental Variables
  - Motivation
  - Intuition
  - Theory
  - Examples
  - Practice
- Regression Discontinuity
- Econometrics summary

Key Concepts (today's lecture)

- Conditional exogeneity
- Instrumental variables
- First Stage
- Second Stage
- Reduced Form
- Exclusion restriction
- Instrument relevance
- Regression discontinuity
- Running variables

Instrumental Variables: Motivation

- We are interested in estimating the (causal) effect of getting a flu vaccine on later sickness

$$GotSick_i = \alpha + \beta Vaccine_i + u_i$$

  - We care about β
  - (Note: Prior slides used ϵ_i to denote idiosyncratic error; in this lecture I'm going to use u_i to be consistent with the readings)

- If we estimate this regression using observational data, can we interpret our estimate $\hat{\beta}$ as causal?
  - Think: what would the identifying assumption need to be?

Refresher: Ordinary Least Squares (OLS)

- Most of these intuitive problems involve the violation of a critical assumption of OLS:
  - $E(u_i|X_i) = 0$ ("Conditional Exogeneity")
    - The conditional distribution of u_i given X_i has mean zero
      - i.e., the "other" factors (like Age, Wealth, Ethnicity) are unrelated to X_i
- Note: Other OLS assumptions important too ☺

Instrumental Variables: Motivation

- How might we obtain causal estimates of the parameter: β?
- Use an experiment!
  - Randomly assign vaccines to people and keep track of who gets sick

$$GotSick_i = \alpha + \beta Vaccine_i + u_i$$

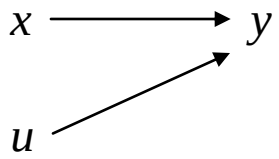
  - If we can truly randomize who gets the vaccine, then the identifying assumption necessary to interpret β causally might be plausible
    - But, this research design is likely not feasible! (why?)

Instrumental Variables: Introduction

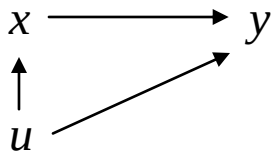
- Often randomization is impossible, and the assumptions required for causal inference using basic techniques are not justified
- "Instrumental variables" (Two Stage Least Squares): An **instrument** variable creates random variation in a "treatment" variable without affecting the outcome (except via the treatment)
- In our example: Something that affects a person's likelihood of getting a vaccine but doesn't directly affect their likelihood of getting sick

Instrumental Variables: Intuition

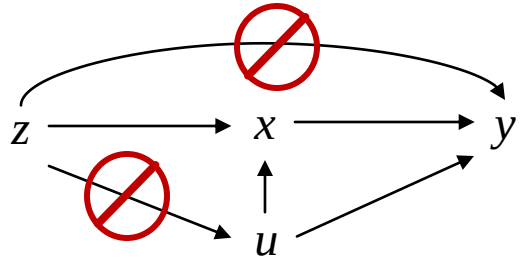
- Normal OLS:



- Often times:



- Enter the Instrument:
  - Note: "exclusion restriction"



- See Chapter 4.8 of Cameron & Trivedi, Microeconometrics (on bCourses) for details

Outline

- Instrumental Variables
  - Motivation
  - Intuition
  - **Theory**
  - Examples
  - Practice
- Regression Discontinuity
- Econometrics summary

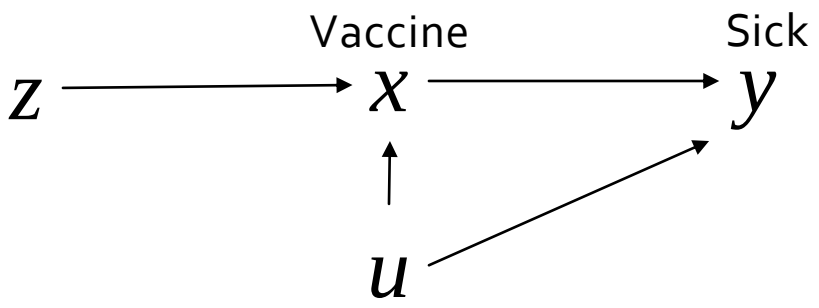
Instrumental Variables: Theory

$$Y_i = \alpha + \beta X_i + u_i$$

- IV regression can eliminate bias when $\mathrm{E}(u|X) \neq 0$
  - IV regression breaks X_i into two parts
    - One part that might be correlated with u_i,
    - One part that is not correlated u_i
  - We use an "instrumental" variable Z_i to isolate the second part
    - The instrument must be uncorrelated with u_i (more on this later)
    - This allows us to isolate (and estimate) the causal part of β
  - Intuitively: the "instrument" induces movements in X_i that are uncorrelated with u_i, and uses this variation to estimate β

Instrumental Variables: Theory

- For an instrumental variable Z to be valid it must satisfy two conditions:
  1. Instrument relevance: $\text{corr}(Z_i, X_i) \neq 0$
  2. Instrument exogeneity: $\text{corr}(Z_i, u_i) = 0$
     - Often called the "Exclusion restriction"



Instrumental Variables: Finding an instrument

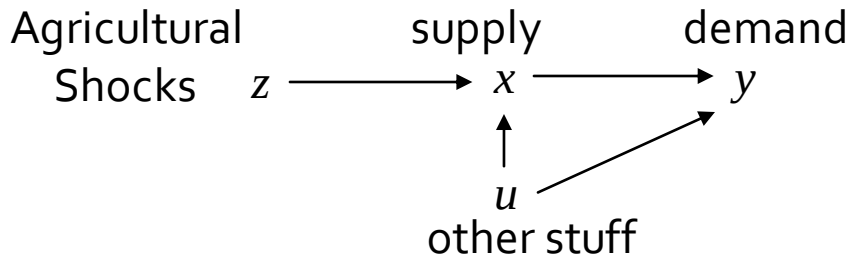
- How do we find a good instrument?
  1. Instrument relevance: $\text{corr}(Z_i, X_i) \neq 0$
  2. Instrument exogeneity: $\text{corr}(Z_i, u_i) = 0$
- Can these be tested empirically?
  1. Sure! We can regress X_i on Z_i and check the corresponding t-statistic. The more significant the better
  2. Sadly, no. We need assumptions (and often theory+experiments) to find a good instrument
     - Identifying assumption: Z only affects Y through X

Outline

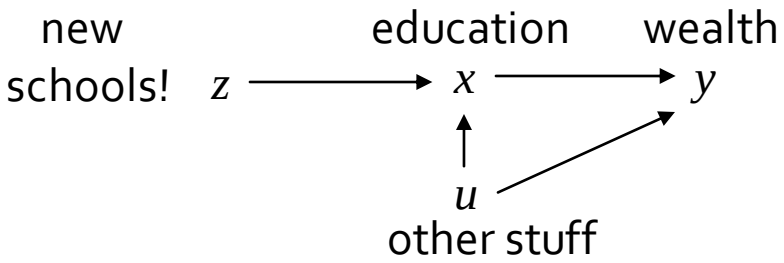
- Instrumental Variables
  - Motivation
  - Intuition
  - Theory
  - **Examples**
  - Practice
- Regression Discontinuity
- Econometrics summary

Instrumental Variables: Canonical examples

- Supply ← → Demand



- Education ← → Wealth



- See also: Angrist, J.; Krueger, A. (2001). "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments". Journal of Economic Perspectives 15(4): 69–85.

Instrumental Variables: Another example

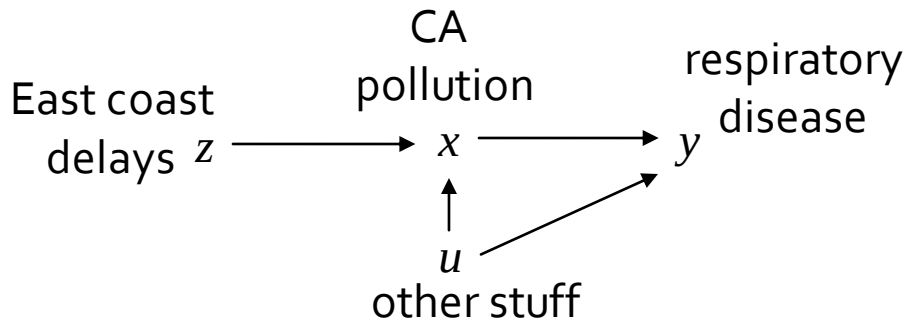
- I want to estimate the causal effect of air pollution on respiratory disease
  - Option A: compare polluted to non-polluted areas
  - Option B: run an experiment
  - Option C: use instrumental variables

AIRPORTS, AIR POLLUTION, AND CONTEMPORANEOUS HEALTH

Wolfram Schlenker
W. Reed Walker

Working Paper 17684
http://www.nber.org/papers/w17684

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2011



Outline

- Instrumental Variables
  - Motivation
  - Intuition
  - Theory
  - Examples
  - **Practice**
- Regression Discontinuity
- Econometrics summary

Instrumental Variables: Practice

- Suppose for now that we have a valid Z_i
  - How can you use Z_i to estimate β (from $Y_i = \alpha + \beta X_i + u_i$)
- Instrumental Variables ("2SLS") is a 2-step procedure

1. Isolate the part of X that is uncorrelated with u by regressing X on Z:

$$X_i = b_0 + b_1 Z_i + \nu_i$$

  - If Z_i is uncorrelated with u_i (assumed earlier), then $b_0 + b_1 Z_i$ is uncorrelated with u_i
  - Compute predicted X_i (i.e. $\widehat{X_i}$) where

$$\widehat{X_i} = \widehat{b_0} + \widehat{b_1} Z_i$$

Instrumental Variables: Practice

- Suppose for now that we have a valid Z_i
  - How can you use Z_i to estimate β (from $Y_i = \alpha + \beta X_i + u_i$)
- Instrumental Variables ("2SLS") is a 2-step procedure

2. Replace X_i with $\widehat{X_i}$ in the regression of interest, i.e.

$$\mathrm{Y_i} = \alpha + \beta \widehat{X_i} + u_i$$

  - Because $\widehat{X_i}$ is uncorrelated with u_i, the first OLS assumption holds
  - Thus, β can be estimated by OLS
  - The resulting estimator is the IV (or 2SLS) estimator

Instrumental Variables: Summary

- You want to estimate: $Y_i = \alpha + \beta X_i + u_i$
- Suppose that you have a valid instrument Z_i
- Stage 1: $\mathrm{X_i} = b_0 + b_1 Z_i + \nu_i$
  - Obtain predicted values $\widehat{X}_i$
- Stage 2: $\mathrm{Y_i} = \alpha + \beta \widehat{X}_i + u_i$
  - $\hat{\beta}$ is a consistent estimator of β

Instrumental Variables: Duflo (2001)

Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment

By ESTHER DUFLO*

Between 1973 and 1978, the Indonesian government engaged in one of the largest school construction programs on record. Combining differences across regions in the number of schools constructed with differences across cohorts induced by the timing of the program suggests that each primary school constructed per 1,000 children led to an average increase of 0.12 to 0.19 years of education, as well as a 1.5 to 2.7 percent increase in wages. This implies estimates of economic returns to education ranging from 6.8 to 10.6 percent. (*JEL* I2, J31, O15, O22)

Instrumental Variables: Duflo (2001)

- "First Stage" (X on Z)
  - Program led to an increase of 0.25 to 0.40 years of education (0.12 to 0.19 years for each new school built per 1,000 children)
- "Reduced Form" (Y on Z)
  - The estimates also suggest that the program led to an increase of 3 to 5.4 percent in wages
- "IV Estimate" (Y on X)
  - Combining the effect of the program on years of schooling and wages generates 2SLS estimates of economic returns to education ranging from 6.8 to 10.6 percent

Instrumental Variables: More examples

Examples of Studies That Use Instrumental Variables to An Natural and Randomized Experiments

Outcome Variable	*Endogenous Variable*	*Source of Instrumental Variable(s)*
	1. Natural Experiments	
Labor supply	Disability insurance replacement rates	Region and time variation in benefit rules
Labor supply	Fertility	Sibling-Sex composition
Education, Labor supply	Out-of-wedlock fertility	Occurrence of twin births
Wages	Unemployment insurance tax rate	State laws
Earnings	Years of schooling	Region and time variation in school construction
Earnings	Years of schooling	Proximity to college
Earnings	Years of schooling	Quarter of birth
Earnings	Veteran status	Cohort dummies
Earnings	Veteran status	Draft lottery number
Achievement test scores	Class size	Discontinuities in class size due to maximum class-size
College enrollment	Financial aid	Discontinuities in financial aid formula
Health	Heart attack surgery	Proximity to cardiac care centers
Crime	Police	Electoral cycles
Employment and Earnings	Length of prison sentence	Randomly assigned federal judges
Birth weight	Maternal smoking	State cigarette taxes
	2. Randomized Experiments	
Earnings	Participation in job training program	Random assignment of admission to training program
Earnings	Participation in Job Corps program	Random assignment of admission to training program
Achievement test scores	Enrollment in private school	Randomly selected offer of school voucher
Achievement test scores	Class size	Random assignment to a small or normal-size class
Achievement test scores	Hours of study	Random mailing of test preparation materials

Outline

- Instrumental Variables
  - Motivation
  - Intuition
  - Theory
  - Examples
  - Practice
- **Regression Discontinuity**
- Econometrics Summary

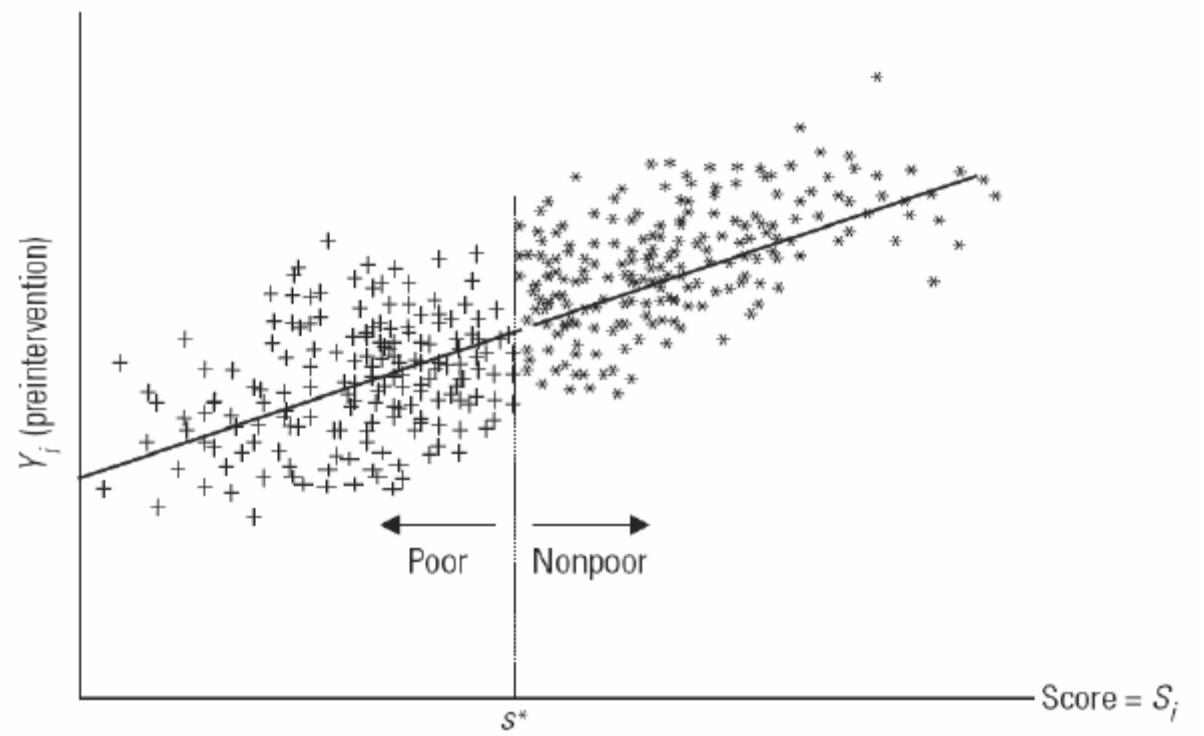
Regression Discontinuity: Intuition

- Many treatments are assigned based on an index or score
  - **Vaccines**: Everyone older than a threshold age is eligible
  - **Promotions**: Targeted to people who generate more than a threshold value of revenue per month
  - **Progresa**: Targeted to households below a wealth threshold
  - **Education**: Scholarships for students who score above a threshold

- Key feature: Treatment (or treatment eligibility) is quasi-random close to the discontinuity

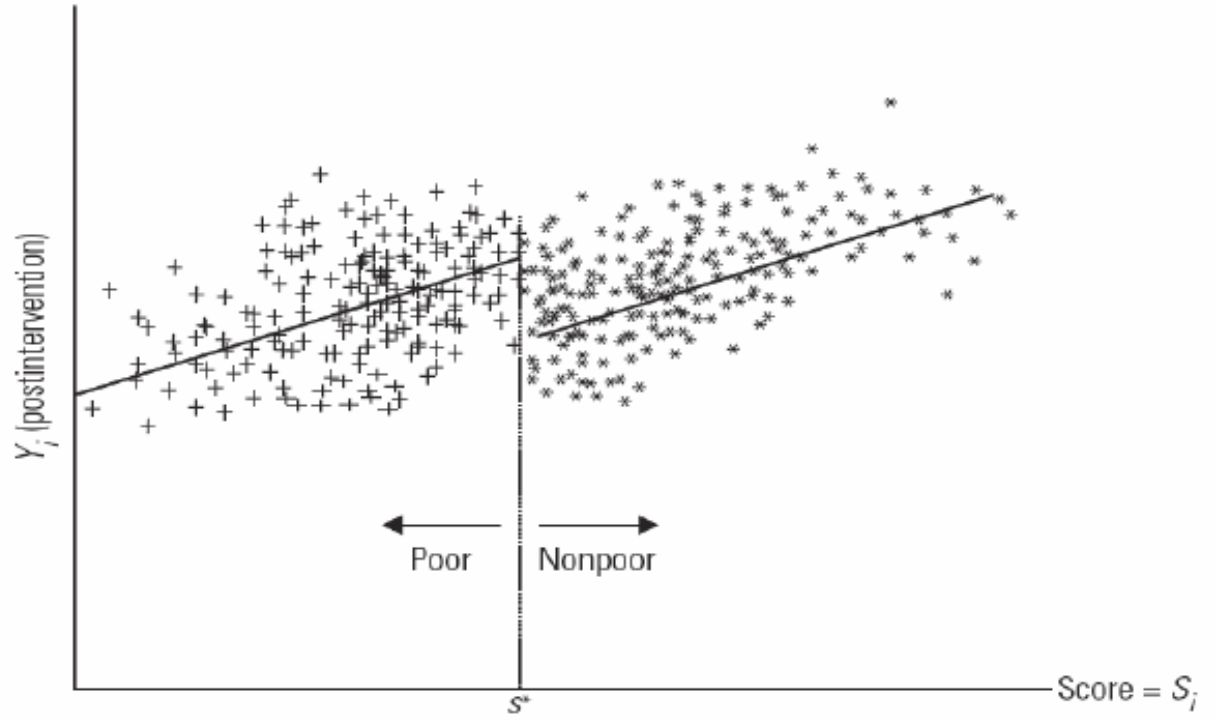
Regression Discontinuity: Intuition

- Outcomes before the program



Regression Discontinuity: Intuition

- Outcomes after the program



Regression Discontinuity: Intuition

- RDD uses individuals just below (or above) the threshold as a counterfactual for the treated individuals just above (or below) the threshold

- This requires that:
  - Other factors that determine the outcome are not discontinuous at the threshold point (smoothness assumption)
  - Running variable has not been "manipulated"

Regression Discontinuity: Example

CREDIT ACCESS AND COLLEGE ENROLLMENT*

Alex Solis†

"Treatment"

"Outcome"

"Running variable"

Abstract

Does limited access to credit explain some of the gap in schooling attainment between children from richer and poorer families? I present new evidence on this important question using data from two loan programs for college students in Chile. Both programs offer loans to students who score above a threshold on the national college admission test, providing the basis for a regression discontinuity evaluation design. I find that students who score just above the cutoff have nearly 20 percentage points higher enrollment than students who score just below the cutoff, which represent a 100% increase in the enrollment rate. More importantly, access to the loan program effectively eliminates the family income gradient in enrollment among students with similar test scores. Moreover, access to loans also leads to 20 percentage points higher enrollment rates in the second and third years of college around the cutoff score, representing relative increases of 213% and 446% respectively, and also eliminating the enrollment gap between the richest and poorest income quintiles. These findings suggest that differential access to credit is an important factor behind the intergenerational transmission of education and income.

Regression Discontinuity: Example

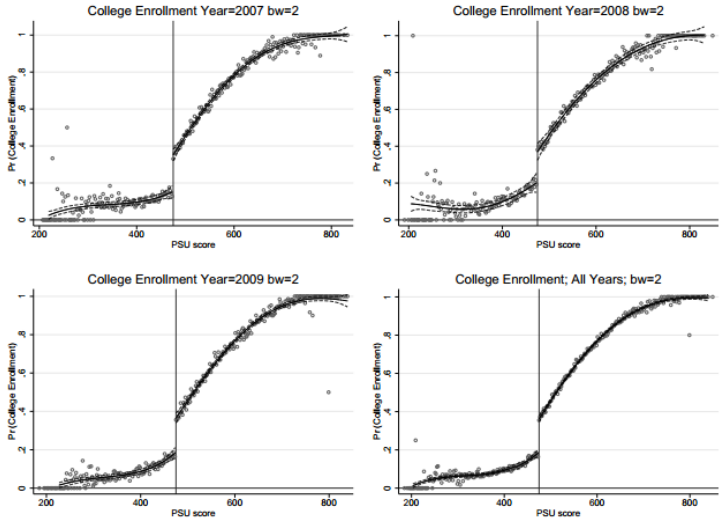
A key strength of RD is that it is often possible to test the presence (or absence) of effects with a few simple figures:

1. Outcomes (Y_i , T_i) vs. Running variable (Z_i)
2. Covariates (X_i) vs. Running variable
3. Density of Running variable

Regression Discontinuity: Example

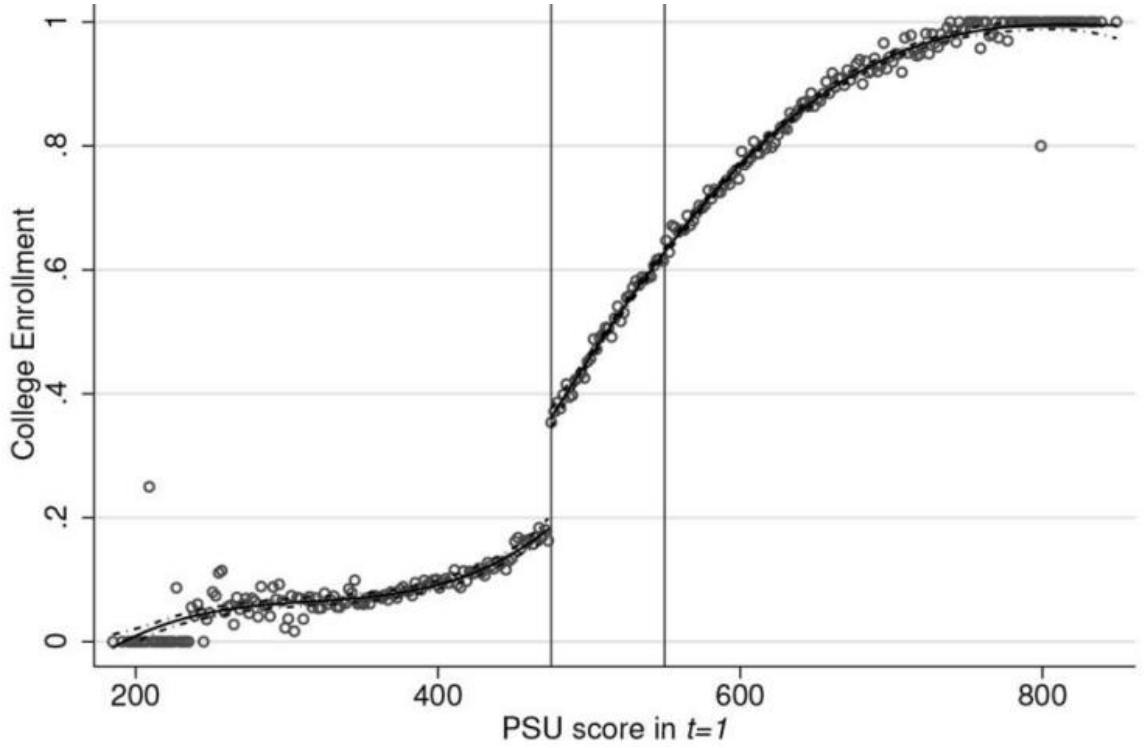
1. **Outcomes (Y_i, T_i) vs. Running variable (Z_i)**
2. Covariates (X_i) vs. Running variable
3. Density of Running variable

Figure 6: RD for College enrollment. Full sample.



Note: Each dot represents average college enrollment in an interval of 2 PSU points.
The dashed lines represent fitted values from a 4th order spline and 95% confidence intervals for each side.
The vertical line indicates the cutoff (475).
These graphs show the full sample of students fulfilling all requirements to be eligible for college loans and taking the PSU immediately after graduating from high school.

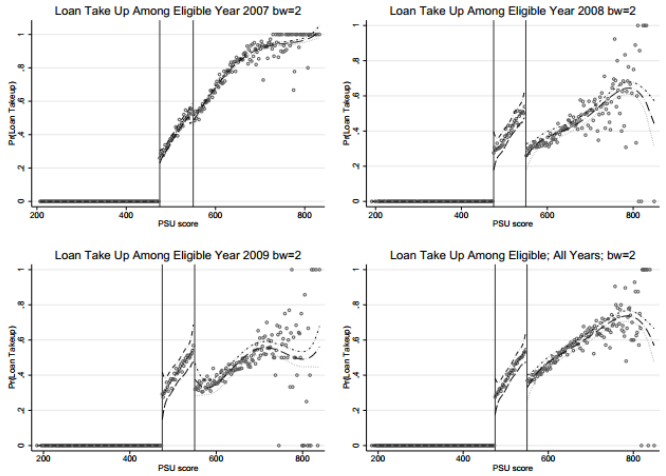
Outcome Y_i: Schooling attainment



Regression Discontinuity: Example

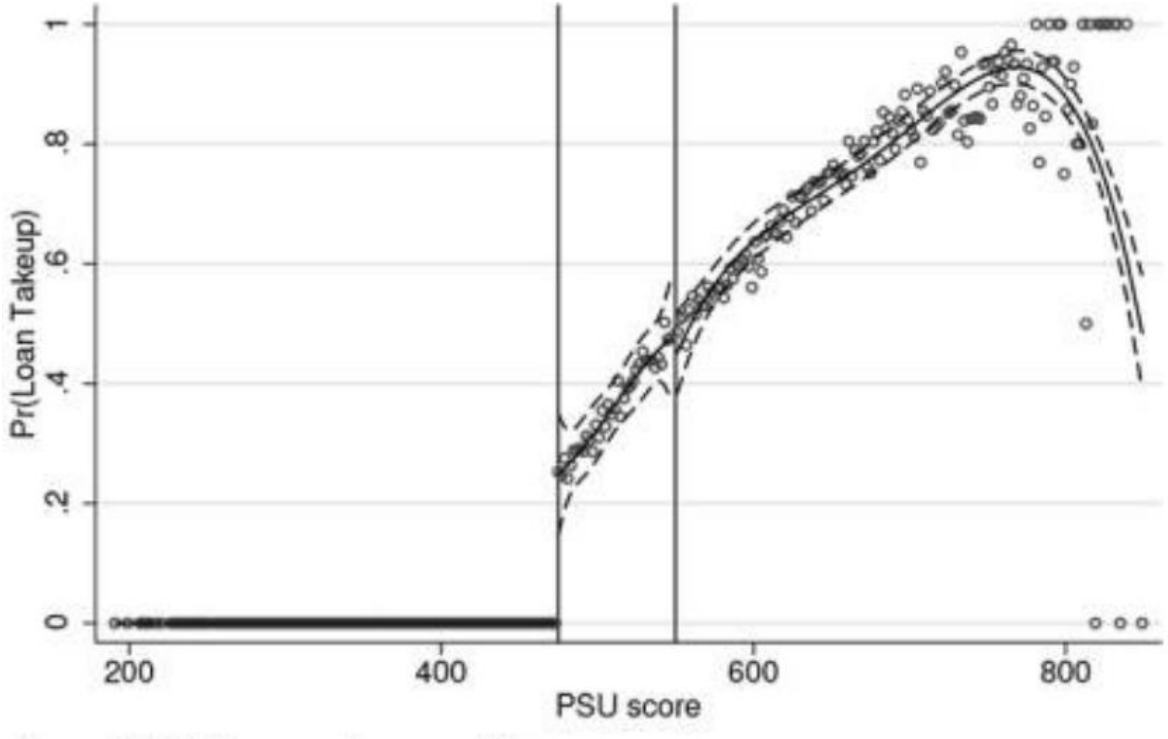
1. **Outcomes (Y_i , T_i) vs. Running variable (Z_i)**
2. Covariates (X_i) vs. Running variable
3. Density of Running variable

Figure 3: Loan take up. Probability of taking up a college tuition loan among preselected eligible students.



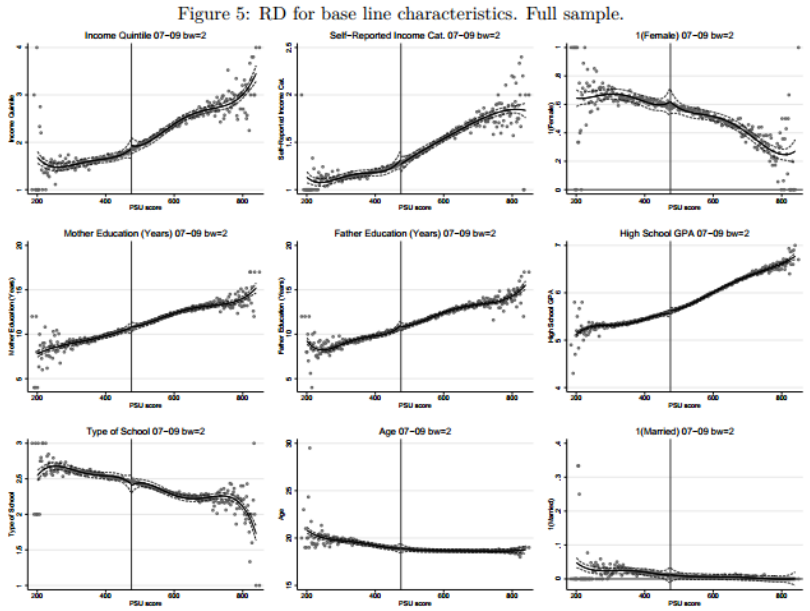
Note: Each dot represents average loan take-up relative to eligible students, in an interval of 2 PSU points. To the right of tye cutoff, each dot contains on avergae rouglhy 441 students receiving the loans. The dashed lines represent fitted values from a 4th order spline and 95% confidence intervals for each side. The vertical line indicates the cutoff (475).

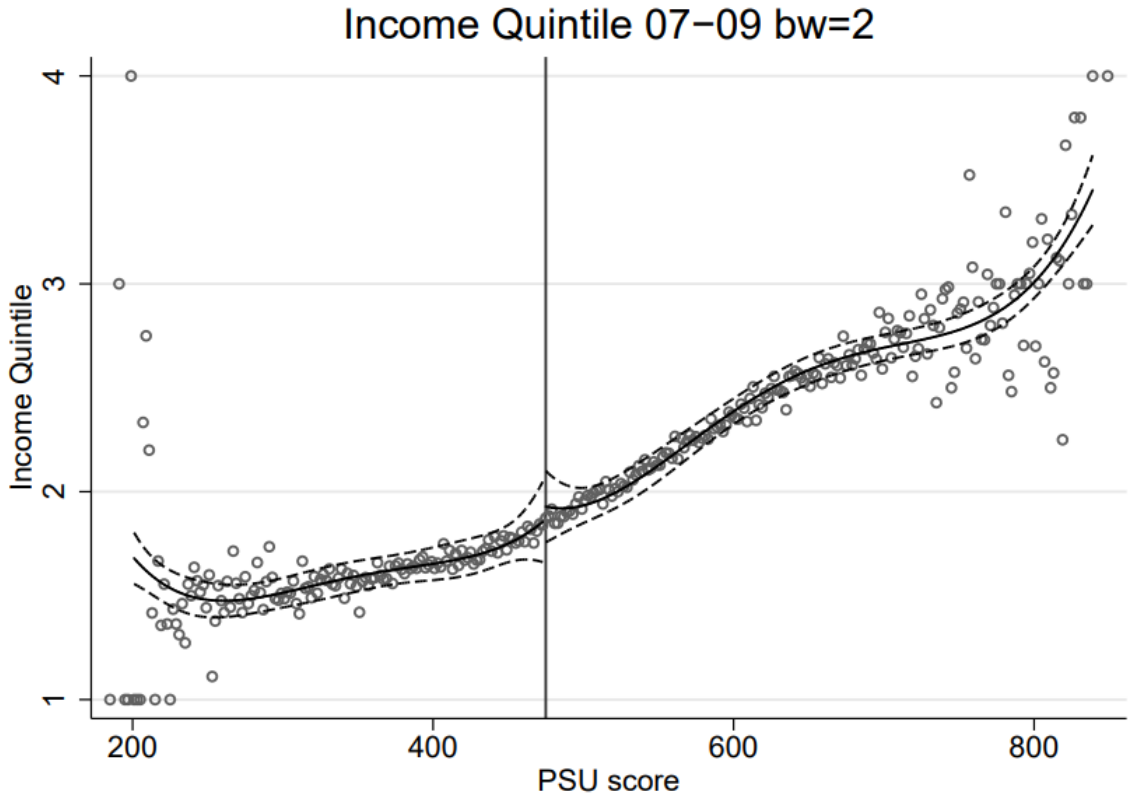
Treatment T_i: Access to Credit



Regression Discontinuity: Example

1. Outcomes (Y_i, T_i) vs. Running variable (Z_i)
2. **Covariates (X_i) vs. Running variable**
3. Density of Running variable

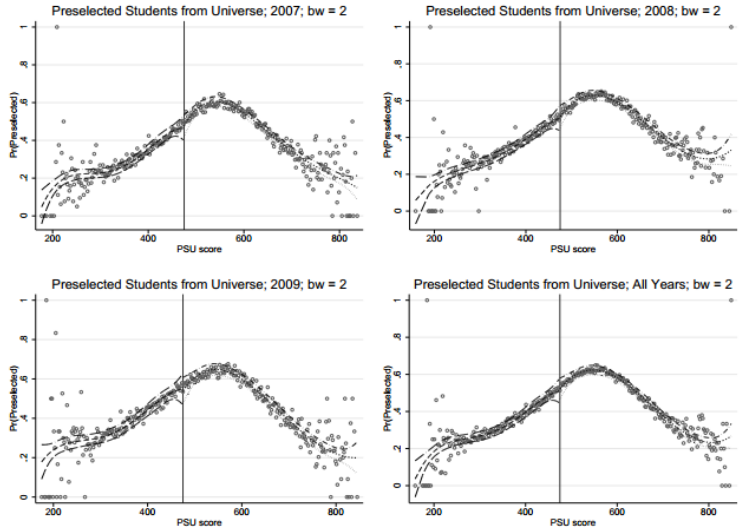




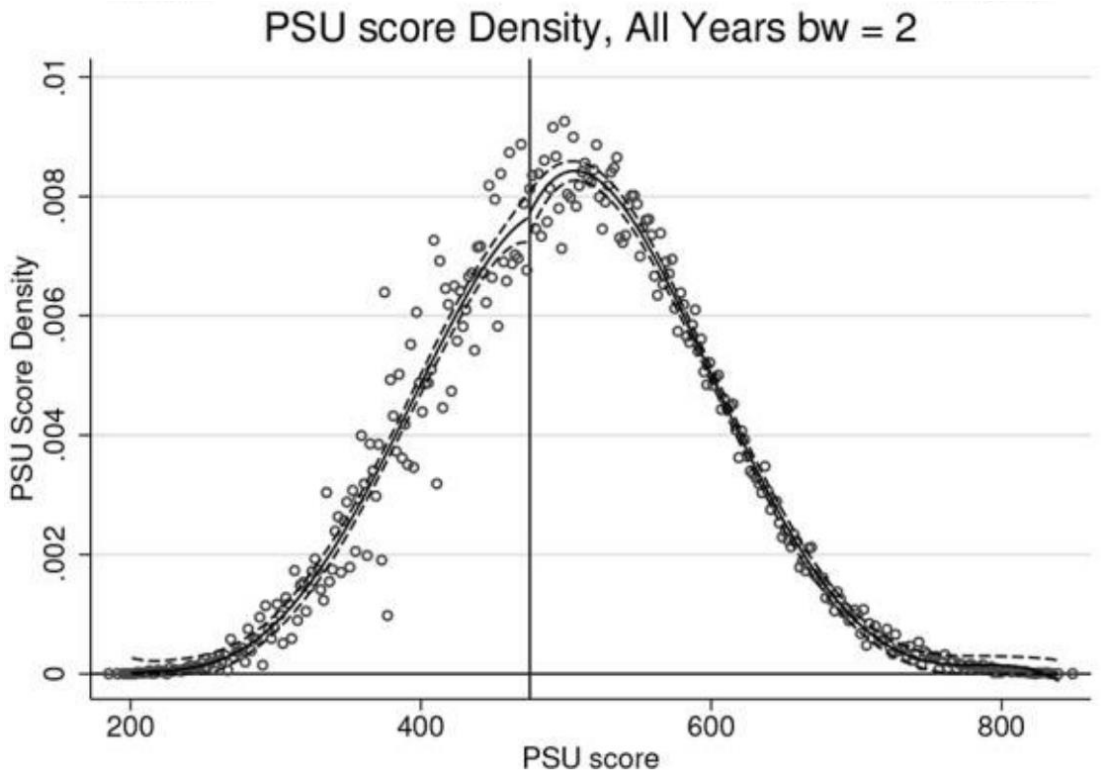
Regression Discontinuity: Example

1. Outcomes (Y_i, T_i) vs. Running variable (Z_i)
2. Covariates (X_i) vs. Running variable
3. **Density of Running variable**

Figure 2: Probability of being *Preselected* for loans.



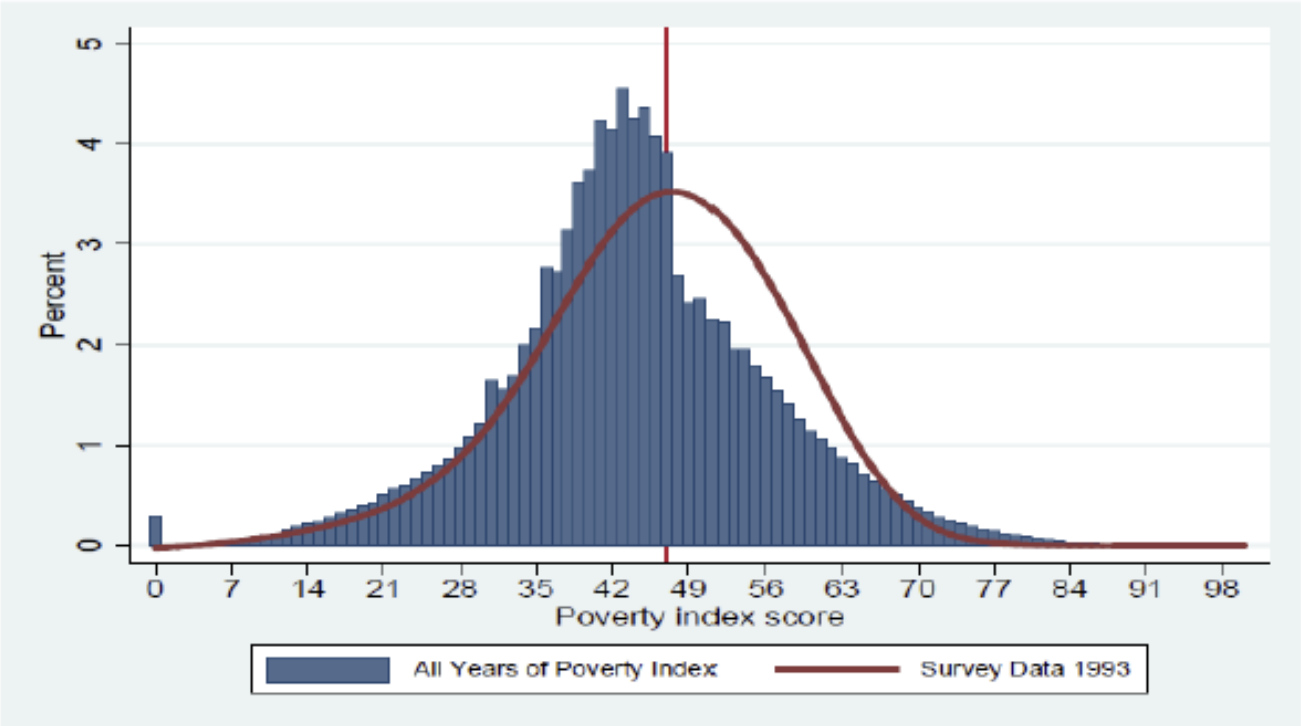
Note: Each dot indicates the preselection rate of students with scores in an interval of 2 PSU points (all students included). On average each dot contains 670 students. The dashed lines represent fitted values from a 4th order polynomial spline and 95% confidence intervals for each side. The vertical line indicates the cutoff (475).



Regression Discontinuity: Manipulation

- Density of Running variable
  - What if the running variable itself is discontinuous at the threshold?

Figure 3: Census of the Poor and 1993 Survey Data Score Distribution



Regression Discontinuity: Example

- When the discontinuity precisely determines treatment, this is equivalent to quasi-random assignment *in a neighborhood*
- For instance:
  - Everyone older than 75 as of Jan 31 2021 is eligible for a Covid vaccine
    - (Let's assume that compliance is perfect)
  - We might compare rates of illness between people born in January 1946 and February 1946
    - Identifying assumption: Rates of illness in 2021 among people aged born in Jan and Feb 1946 *would have been the same* in the absence of the vaccine

Regression Discontinuity: Estimation

- Quantifying the effect of the discontinuity
  - Instead of estimating: $GotSick_i = \alpha + \beta Vaccine_i + u_i$
  - We estimate: $GotSick_i = \alpha + \beta(Over75_i) + \delta(AgeInDays_i) + u_i$
    - $Over75_i$ is a binary "treatment" variable
    - $AgeInDays_i$ is the individual's age, in days
    - δ is a kernel (but just think of it as a constant, for now)
    - Estimated locally, for people with $s_{\min} < AgeInDays_i < s_{max}$
  - Note the similarity to Instrumental Variables!
    - $\beta(Over75_i)$ is an instrument for treatment status

Regression Discontinuity: Summary

- Advantages
  - Yields an unbiased estimate of treatment effect at the discontinuity
  - Takes advantage of a known rule for assigning the benefit that is common in the designs of social policy
  - A group of eligible households or individuals need not be excluded from treatment
  - Can be used in other settings
    - Spatial discontinuities
    - Temporal discontinuities (event studies)

Regression Discontinuity: Summary

- Disadvantages
  - Produces *local average treatment effects (LATE)* that are not always generalizable
  - Effect is estimated at the discontinuity, so generally, fewer observations exist than in a randomized experiment with the same sample size
  - Specification can be sensitive to functional form, including nonlinear relationships and interactions

Outline

- Instrumental Variables
  - Motivation
  - Intuition
  - Theory
  - Examples
  - Practice
- Regression Discontinuity
- **Econometrics Summary**

Econometrics: Summary

- Wikipedia says:
  - **Econometrics** is the application of mathematics, statistical methods, and computer science, to economic data and is described as the branch of economics that aims to give empirical content to economic relations
- For the purposes of this class:
  - **Econometrics** is an enormously useful set of quantitative methods for understanding associations and causal relationships in data

Econometrics: What you've learned

- Experimental methods
  - Design and randomization
  - Simple differences
  - Double differences
  - Regression
  - Fixed effects
- Non-experimental methods
  - All of the above and…
  - Instrumental variables
  - Regression discontinuity

Econometrics: Key lesson

- No single method is "right" or "better"
- Each method requires a different identifying assumption, and implies a different counterfactual
- When deciding which method to use:
  - Determine which methods you *could potentially* use
  - For each candidate, articulate the identifying assumption
  - Brainstorm ways to possibly invalidate that assumption
  - Decide which assumption seems most reasonable, given your context, your data, and your situational knowledge

Additional Resources

Beginner → Advanced