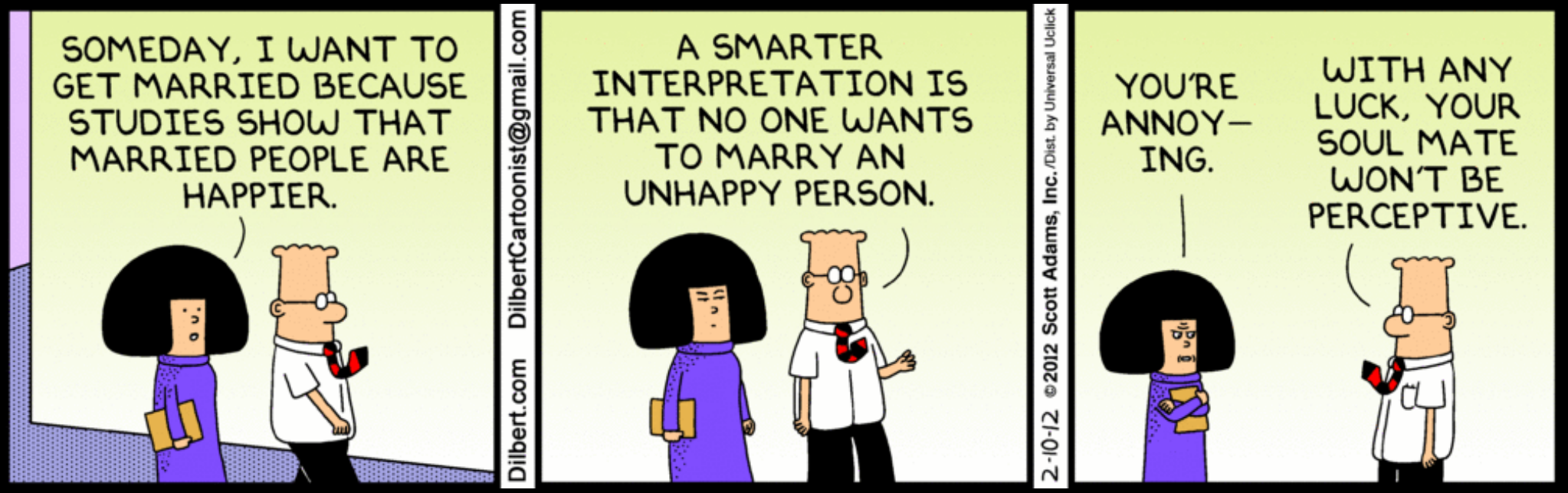


INFO 251: Applied Machine Learning

Nonexperimental Methods: Instrumental Variables

Announcements

- In person instruction – *to resume Feb 8*
  - Starting on Feb 8, I plan to lecture from SH202
  - Class will be split between two rooms (SH202 and SH210)
- My office hours next week will also be on Zoom
  - Future office hours TBD
- For future PS submissions: please also submit a PDF of your responses (in addition to the ipynb notebook)

Course Outline

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

Outline

- Instrumental Variables
  - Motivation
  - Intuition
  - Theory
  - Practice
  - Examples

Random breakout session #4

- You will be randomly divided into groups of 2-4
  - This breakout room will only last a few minutes
  - Introduce yourselves
    - Name, program, favorite class you've taken thus far at UC Berkeley (aside from INFO251, of course!)
    - Are you potentially looking for study partners?
    - Consider exchanging contact information!

Key Concepts (previous lecture)

- Progresa
- Interpreting regression coefficients
- Dummy variables, "one-hot" vectors
- Heterogeneous treatment effects
- Regression and impact evaluation
  - Estimating treat vs. control
  - Interaction variables
  - Estimating difference-in-difference
- Cross-sectional vs. panel data
- Between vs. within variation
- Difference regressions
- Normalization
- Fixed effects

Key Concepts (today's lecture)

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

Instrumental Variables: Motivation

- We are interested in estimating the effect of getting a Covid-19 vaccine later sickness

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

  - We care about β

- If we estimate this regression using observational data, can we interpret our estimate $\hat{\beta}$ as causal?

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
    - For a given value of X_i, the mean of the distribution of these other factors is zero
- 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

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

  - What is our identifying assumption now?
  - Why might this be problematic?

Instrumental Variables: Motivation

- How else might we obtain causal estimates of β?
  - Let's assume we can't experiment, but we have panel data (i.e., we observe the same person's behavior over several years)
- A "panel fixed effects" regression:

$$GotSick_{it} = \alpha + \beta Vaccine_{it} + \mu_i + u_{it}$$

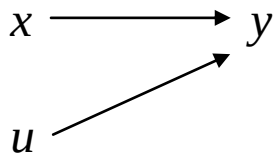
- What is the identifying assumption?
  - Is this credible?
  - When might it be violated?

Instrumental Variables: Introduction

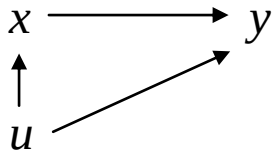
- Often randomization is impossible, and the assumptions required for causal inference using basic techniques are not justified
- "Instrumental variables" (Two Stage Least Squares): An **instrument** variable creates random variation in a "treatment" variable without affecting the outcome (except via 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

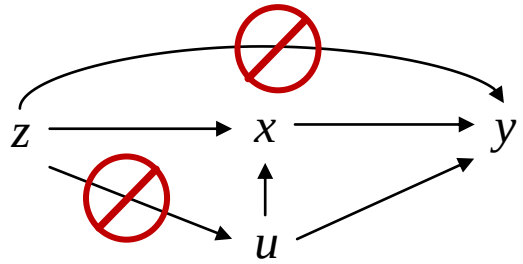
- Normal OLS:



- 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**
  - Practice
  - Examples

Instrumental Variables: Theory

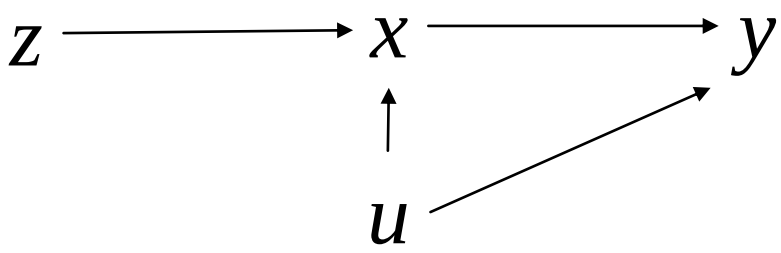
- Instrumental variables (IV) regression can eliminate bias when $\mathrm{E}(u|X) \neq 0$ by using an instrumental variable, Z

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

- IV regression breaks X_i into two parts: a part that might be correlated with u_i and a part that is not
  - By isolating the part that is not correlated with u, we estimate β
  - This is done by using an "instrumental" variable Z_i
    - The instrument must be uncorrelated with u_i (more on this later)
  - 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
  - **Practice**
  - Examples

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

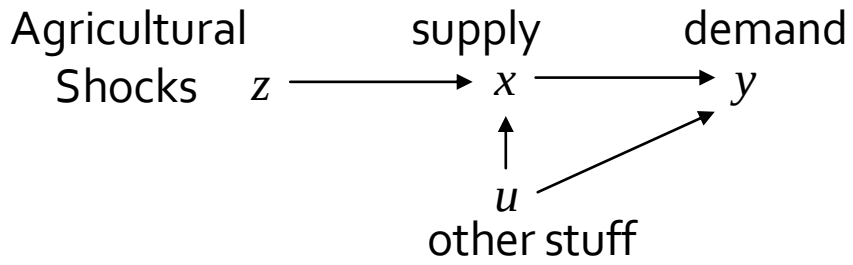
- You want to estimate: $Y_i = \alpha + \beta X_i + u_i$
- Suppose that you have a valid instrument Z_i
- Stage 1: $\mathrm{X_i} = b_0 + b_1 Z_i + \nu_i$
  - Obtain predicted values $\widehat{X}_i$
- Stage 2: $\mathrm{Y_i} = \alpha + \beta \widehat{X}_i + u_i$
  - $\hat{\beta}$ is a consistent estimator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

Outline

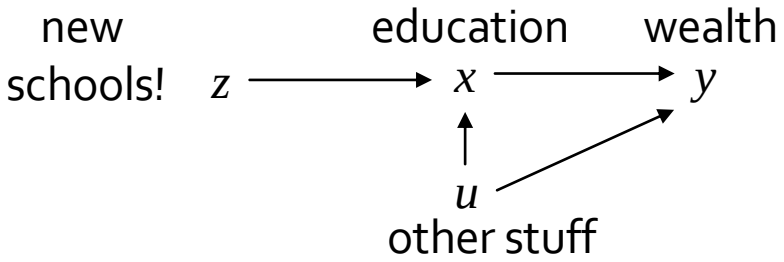
- Instrumental Variables
  - Motivation
  - Intuition
  - Theory
  - Practice
  - **Examples**

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: 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: Another example

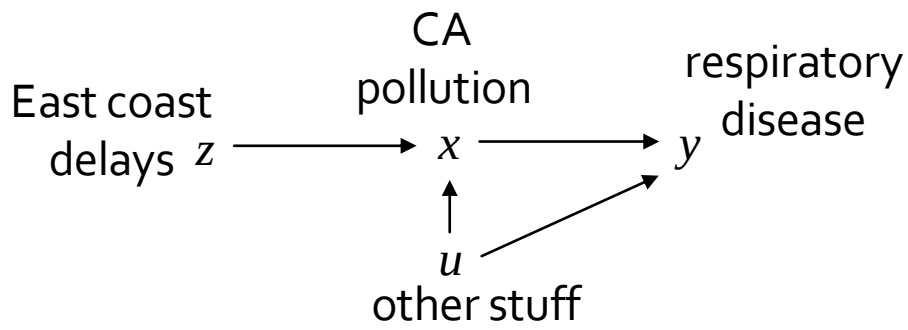
- I want to estimate the causal effect of air pollution on respiratory disease
  - Option A: compare polluted to non-polluted areas
  - Option B: run an experiment
  - Option C: use instrumental variables

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



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

Additional Resources

Beginner → Advanced