MBA 753 : Causal Inference Methods for Business Analytics

Dr. Nivedita Bhaktha

09.09.2024

Agenda

- Regression Discontinuity Design
- Instrumental Variable

- threshold for X
  - deterministic, sharp
  - $E(Y_{1i}|X_i)$, $E(Y_{0i}|X_i)$ continuous at c
- LATE : local interval

Regression Discontinuity Design

$D_i = 0$

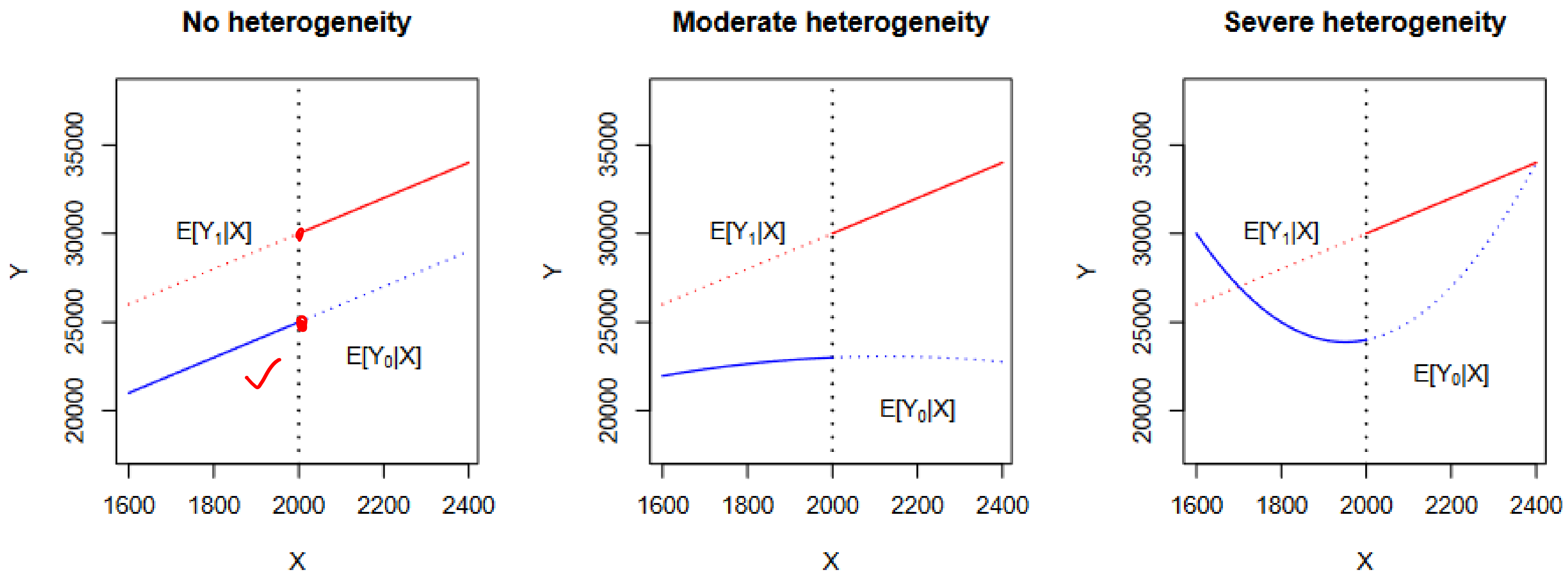
RDD Regression

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 (X_i \overset{0}{-} c) + \beta_3 D_i (X_i - c) + \epsilon_i$$

Z_i

D_i is the treatment variable; X_i is the running variable; c is the threshold

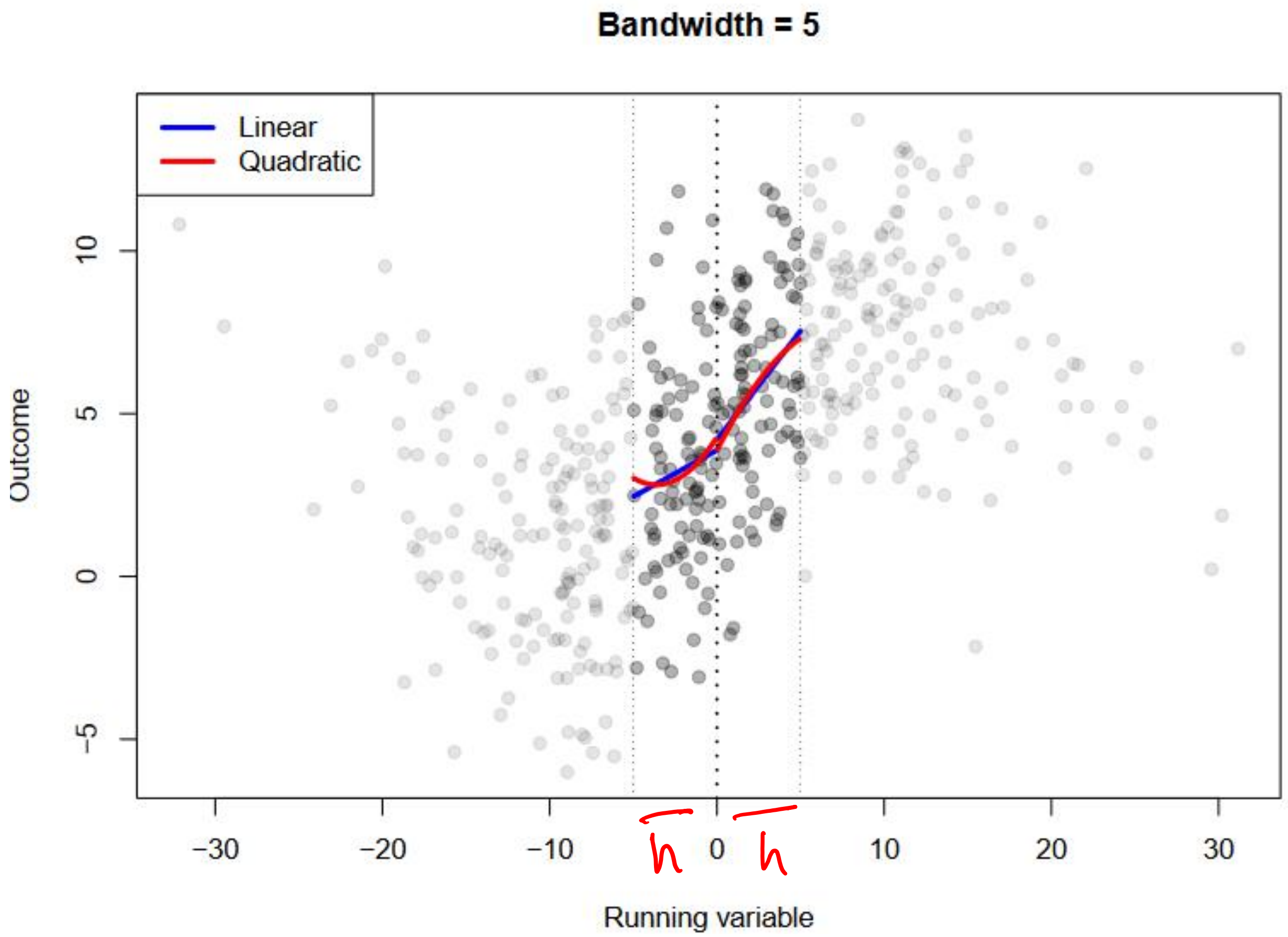
Ideal



Bandwidth Method

LATE

- Idea: Subjects barely to either side of the cutoff are basically the same other than for the cutoff
  - any differences between them are really the fault of treatment
    $$c - h \leq X_i \leq c + h$$
  - h directly affects the properties of the estimation process and empirical findings can be sensitive to the particular value that one chooses for h



Implications of Bandwidth

- Comparing average outcomes in a small neighbourhood to the right and left of the cutoff leads to
  - Estimates of LATE that are less dependent on the functional form specification
  - Decreases the bias that comes from misspecification
  - Leads to a smaller sample size, thus increasing the variance

Bandwidth Approach

- “Optimal” bandwidth selection
  - Use algorithmic bandwidth selection methods
    - Most common → Imbens-Kalyanaraman procedure
  - Choose h to balance bias-variance tradeoff
    - h is chosen to minimise the expected mean-square error of the RD estimator

- Reporting results from multiple bandwidths
  - In practice, it is common to show that how much (if at all) the estimate of $\hat{\tau}_{LATE}$ changes as we vary the bandwidth

X = age   c = 21

Y = mva

Example

- Effect of alcohol consumption on mortality rates - Carpenter and Dobkin (2009)
- Selection in two groups based on their age: young adults who are below the age of 21 are not legally allowed to drink while young adults above the age of 21 are allowed to drink
- Research question: Does alcohol consumption increase mortality rate?

Variable	**Description**
agecell	Age of individual (the study focuses on adults between 19-22 year)
all	Overall mortality rate
alcohol	Mortality rate for alcohol-related causes
homicide	Mortality rate for homicides
suicide	Mortality rate for suicide
mva	Mortality rate for car accidents
drugs	Mortality rate for drug-related causes (alcohol excluded)
externalother	Mortality rate for other external causes

Instrumental Variables

Instrumental Variables

- Three important threats to internal validity are:
  - **omitted variable bias** from a variable that is correlated with X but is unobserved, so cannot be included in the regression;
  - **simultaneous causality bias** (X causes Y, Y causes X);
  - **errors-in-variables bias** (X is measured with error)

- Instrumental variables: provides a solution by introducing a third variable that is correlated with the endogenous variable (X) but not correlated with the error term (ϵ_i)

Instrumental variables scenarios

- Example: X is schooling; Y is wage;
  - "ability" drives both Y and X
- Example: X is number of children; Y is labor force participation;
  - "inclination to remain outside the formal labor force" drives Y down and X up
- Example: X is medical treatment; Y is health;
  - "prior illness" drives Y down and X up

- Problem: biased measure of the causal effect of X on Y
  - Inconsistency of least-squares methods

IV Regression

- One regressor and one instrument

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- IV regression breaks X into two parts:
  - a part that might be correlated with ϵ_i, and
  - a part that is not.
  - By isolating the part that is not correlated with ϵ_i, it is possible to estimate β_1
- Done using instrumental variable Z_i which is uncorrelated with ϵ_i
  - Z_i detects portion of X_i that is uncorrelated with ϵ_i

Terminology

- Endogenous variable: correlated with ϵ_i
  - Determined within the system – jointly determined with Y
  - Simultaneous causality
- Exogenous variable: uncorrelated with ϵ_i

- Instrument relevance: $cor(Z_i, X_i) \neq 0$
- Instrument exogeneity: $cor(Z_i, \epsilon_i) = 0$
  - Instrument relevance and exogeneity are two necessary conditions for a valid instrument

IV Regression - Estimation

- **Two Stage Least Squares (TSLS)**
- First Stage: regress X on Z using OLS

$$X_i = \alpha_0 + \alpha_1 Z_i + \delta_i$$

  - Estimate α_0 & α_1 using OLS
  - $\alpha_0 + \alpha_1 Z_i$ is uncorrelated with ϵ_i because …
  - Compute the predicted values of X_i

$$\hat{X}_i = \hat{\alpha}_0 + \hat{\alpha}_1 Z_i$$

IV Regression - Estimation

- **Two Stage Least Squares (TSLS)**
- Second Stage: regress Y_i on $\hat{X}_i$ using OLS, i.e. replace X_i with $\hat{X}_i$

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \epsilon_i$$

  - Estimate β_0 & β_1 using OLS as assumptions hold
  - $cor\left(\hat{X}_i, \epsilon_i\right) = 0$ because …
  - Requires large sample
  - The resulting β_1 estimator is called the "Two Stage Least Squares" (TSLS) estimator $\widehat{\beta_1}^{TSLS}$

IV Regression Estimation Summary

- Suppose Z_i is a valid Instrument
- Stage 1: Regress X_i on Z_i, obtain predicted values of $\hat{X}_i$
- Stage 2: Regress Y_i on $\hat{X}_i$, the coefficient of $\hat{X}_i$ is the TSLS estimator $\hat{\beta}_1^{TSLS}$
- $\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1

Recap

Summary

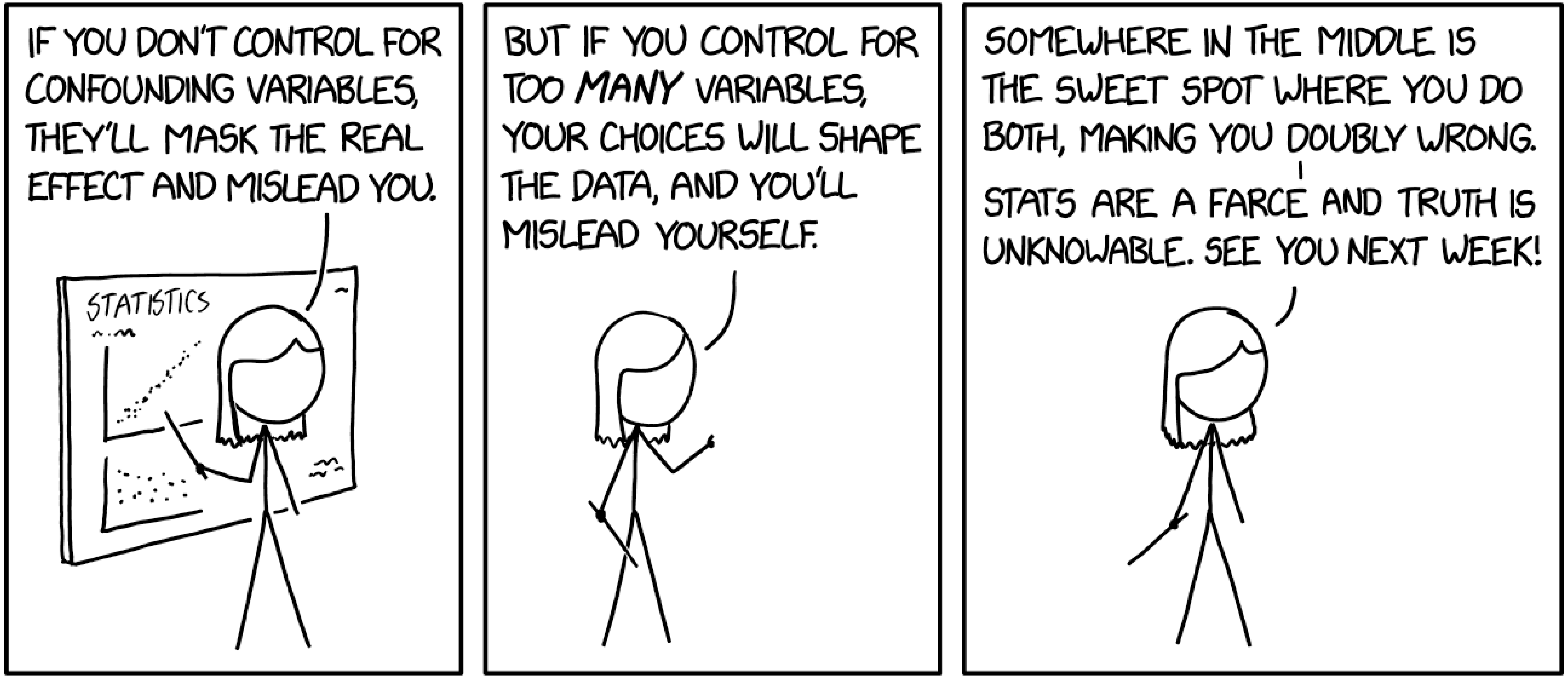
- RDD
  - Setup: continuous running variable, threshold c, and sharp design
  - Can take any functional form
  - LATE can be ATE if the treatment effect is homogenous
- IV
  - Setup: IV Z is correlated with X but uncorrelated with error term
  - Two stage least squares estimation is used

- Objectives achieved:
  - Can interpret the regression coefficients of RDD
  - Can understand the set up of instrumental variables

References

- Scott Cunningham, Causal Inference: The Mix Tape, Yale University Press.
- Cook, T. D., & Campbell, D. T. (2007). *Experimental and quasi-experimental designs for generalized causal inference*.
- Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Thank You ☺


IF YOU DON'T CONTROL FOR CONFOUNDING VARIABLES, THEY'LL MASK THE REAL EFFECT AND MISLEAD YOU.
STATISTICS
BUT IF YOU CONTROL FOR TOO *MANY* VARIABLES, YOUR CHOICES WILL SHAPE THE DATA, AND YOU'LL MISLEAD YOURSELF.
SOMEWHERE IN THE MIDDLE IS THE SWEET SPOT WHERE YOU DO BOTH, MAKING YOU DOUBLY WRONG.
STATS ARE A FARCE AND TRUTH IS UNKNOWABLE. SEE YOU NEXT WEEK!