MBA 753 : Causal Inference Methods for Business Analytics

Dr. Nivedita Bhaktha

07.09.2024

Agenda

- Housekeeping
- DID - Issues
- Regression Discontinuity Design

Announcement

- Final Exam: 15th September, Sunday, 13:00-15:00
- Project Presentation: 14th September, Saturday, 10:30-12:30 – C5
  - 6 teams
- Pop quiz in every class

DID

DID Issues

- Standard Errors in DID
  - Practical applications of DID strategies use data from many years: not just 1 pre and 1 post period
  - The variables of interest only vary at a group level and outcome variables are often serially correlated

- Solution:
  - Clustering standard errors at the group level and time-level
  - Bootstrapped SE

DID Issues

- Threats to validity
  - Non-parallel trends
    - Ashenfelter dip: earnings often fall just prior to entering a training program, which complicates measurement of treatment effect
  - Other simultaneous treatment/intervention
  - Functional form dependence
    - Assumption: DID regression equation is linear
    - Matching DID
  - Multiple treatment times
    - Treatment occurs at different times for treated

- Robustness check for DID design: Falsification test using placebo outcome

Regression Discontinuity Design

Regression Discontinuity Design

- A precise rule, based on a continuous characteristic determines participation in intervention/program
- Examples:
  - Academic test scores: scholarships or prizes, higher education admission, certificates of merit
  - Poverty scores: (proxy-)means-tested anti-poverty programs
  - Date: age cutoffs for pensions; dates of birth for starting school with different cohorts

RDD Treatment Assignment

- A continuous variable X that determines who gets treatment - running var., assignment var., or forcing var.
- We should know the cut-off or threshold c for X
- Assumption: there is no manipulation at the threshold
- Treatment variable: D_i

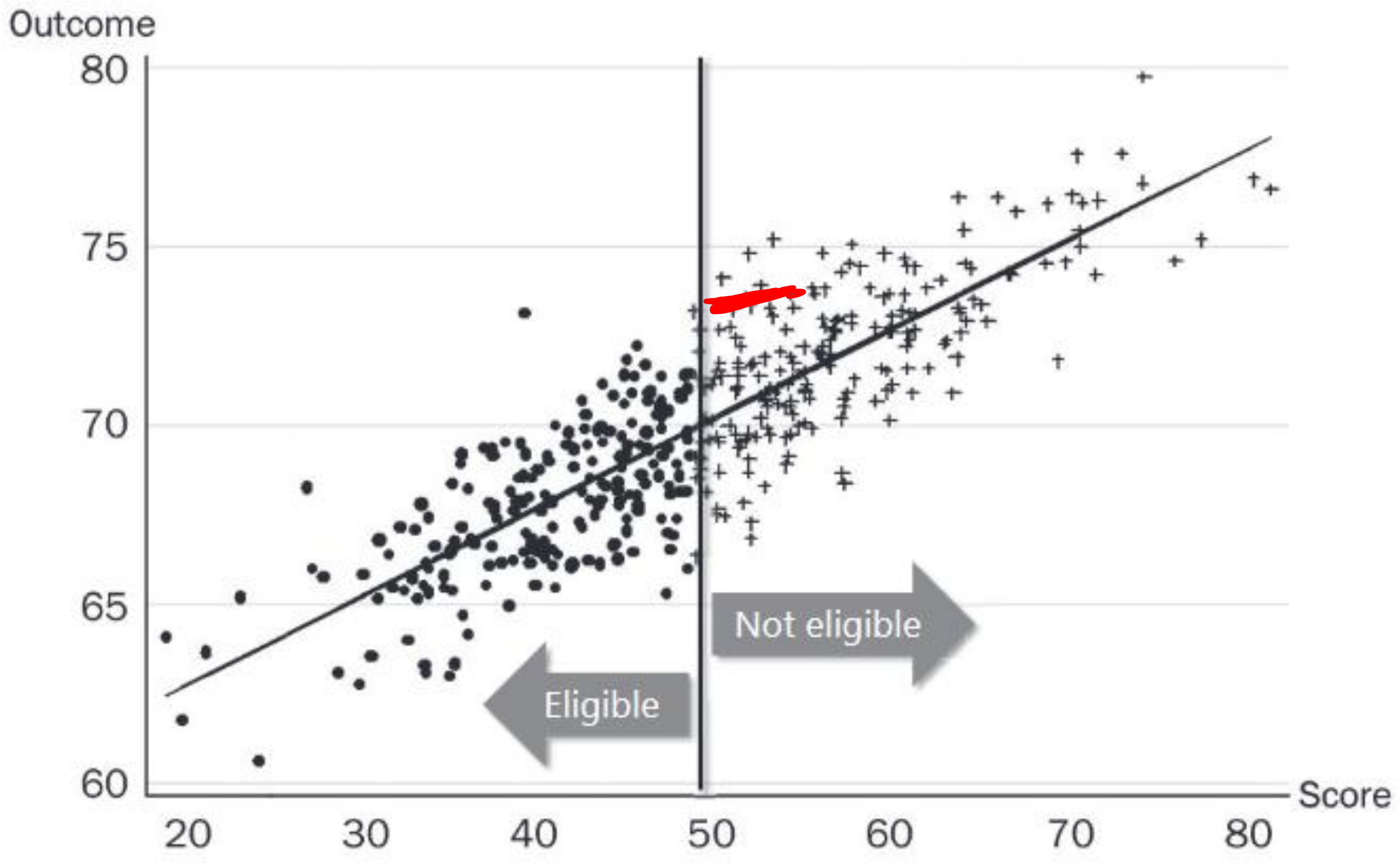
$$D_i = \begin{cases} 1 \; if \; X_i \geq c \\ 0 \; if \; X_i < c \end{cases}$$

RDD Treatment Assignment Type

- In general, depending on enforcement of treatment assignment, RDD can be categorized into two types:
  - **Sharp RDD**: nobody below the cutoff gets the “treatment”, everybody above the cutoff gets it (or vice versa)
    - Everyone follows treatment assignment rule (all are compliers).
    - Local randomized experiment with perfect compliance around cutoff.
  - **Fuzzy RDD**: the probability of getting the treatment jumps discontinuously at the cutoff (NOT jump from 0 to 1)
    - Not everyone follows treatment assignment rule.
    - Local randomized experiment with partial compliance around cutoff.
    - Using initial assignment as an instrument for actual treatment

$X_i - C = Z_i$ $\quad Z_i = 0$

RDD – Treatment Assignment Example



$\beta_2 = 0.3$

$X \to Y$

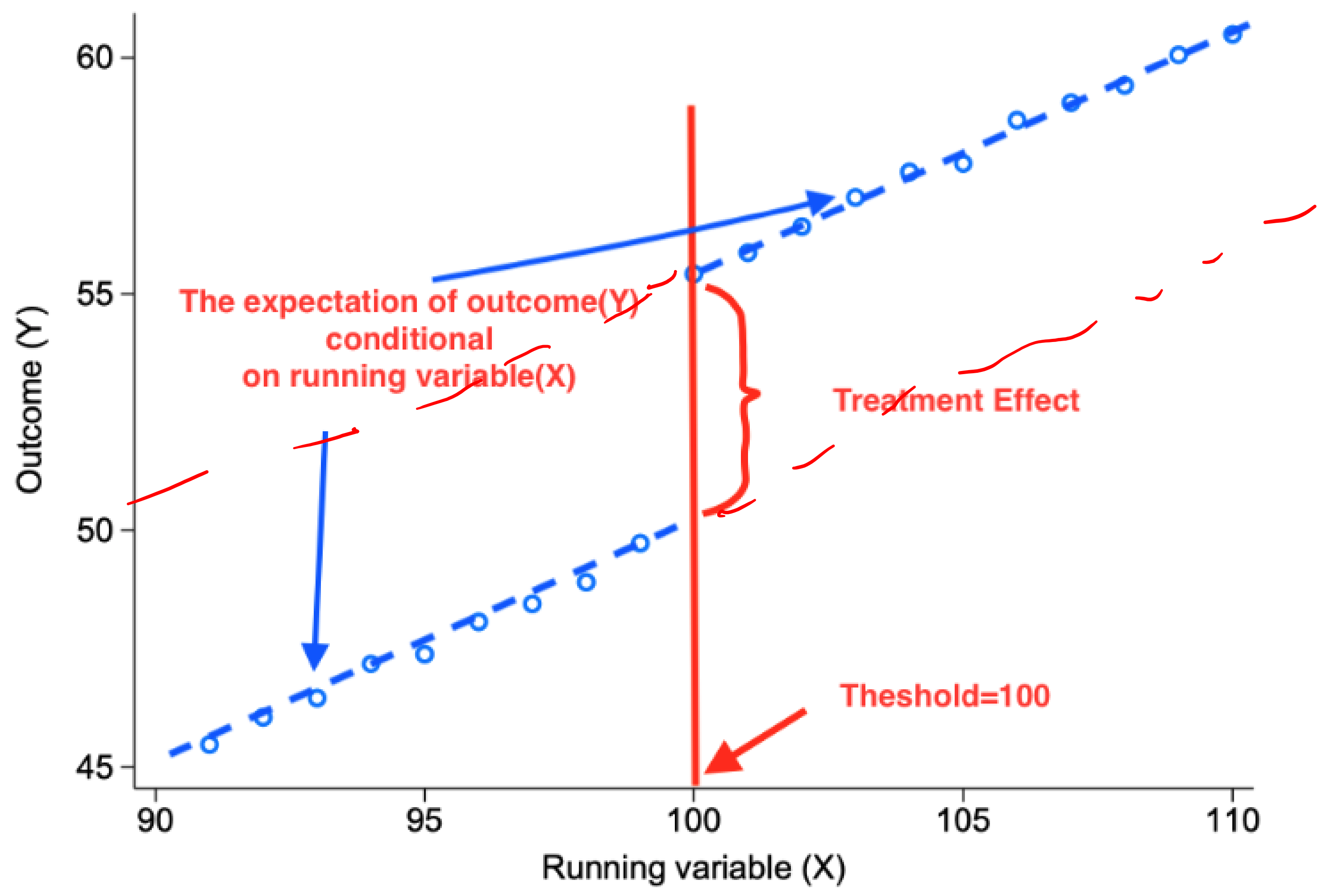
$Y = \beta_0 + 0.3X$

$Y = \beta_0 + 0.3(X-C)$

$Y = \beta_0 + 0.3Z_i$

Source: Gertler, P. J.; Martinez, S., Premand, P., Rawlings, L. B. and Christel M. J. Vermeersch, 2010, Impact Evaluation in Practice: Ancillary Material, The World Bank, Washington DC (www.worldbank.org/ieinpractice)

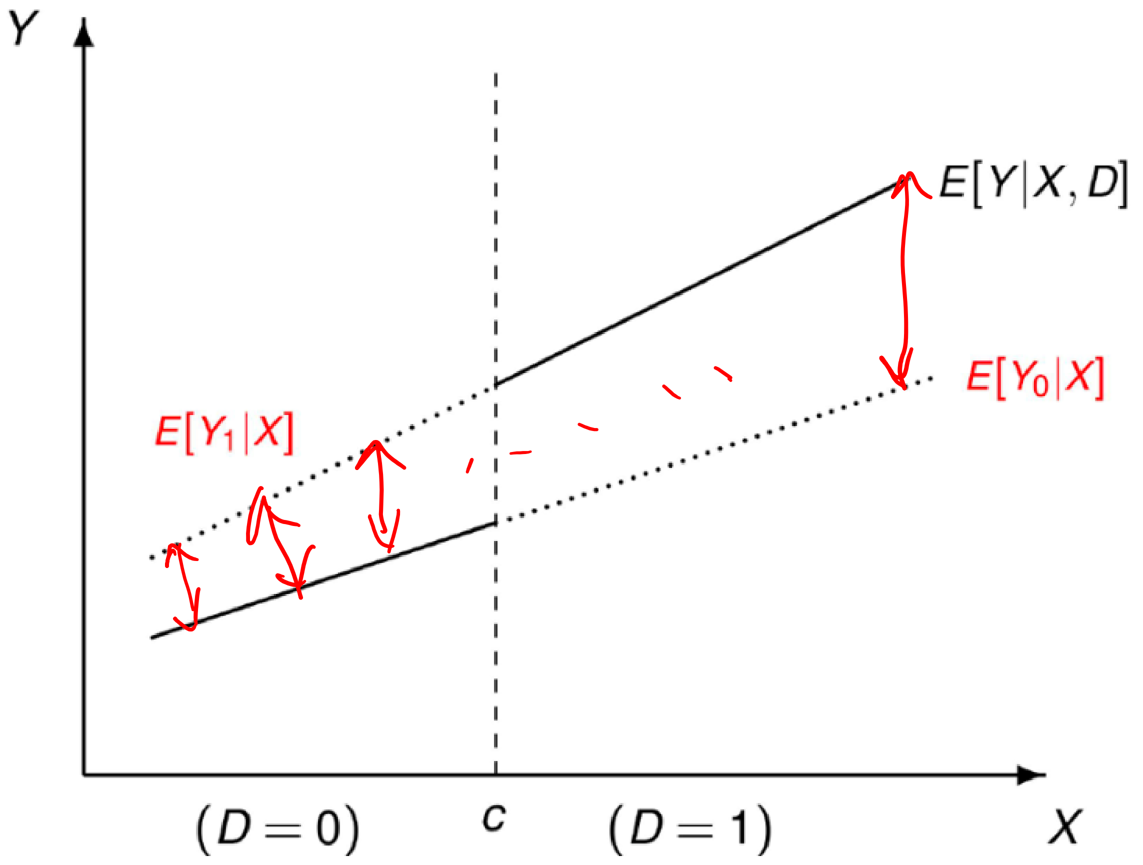
RDD – Basic Idea

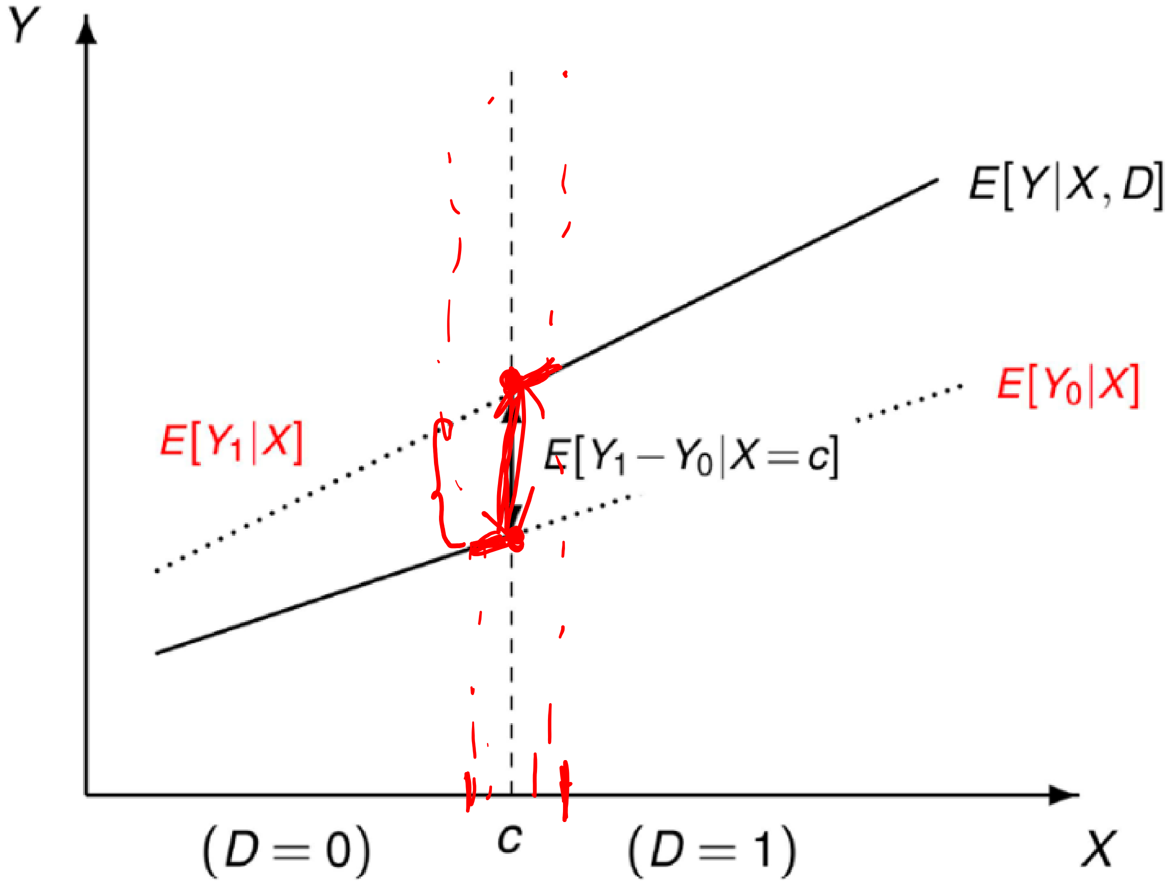


RDD Assumptions

- Deterministic assumption: Treatment assignment is a deterministic function of the assignment variable X and the threshold c
- Continuity assumption: $E(Y_{1i}|X_i)$ and $E(Y_{0i}|X_i)$ are continuous at X_i = c
  - Potential outcomes do not change at cutoff
  - All other unobserved determinants of Yi are continuous at cutoff c
  - No other confounding factor affects outcomes at cutoff c
  - Any observed discontinuity in the outcome can be attributed to treatment assignment

Assumptions





LATE – Local ATE

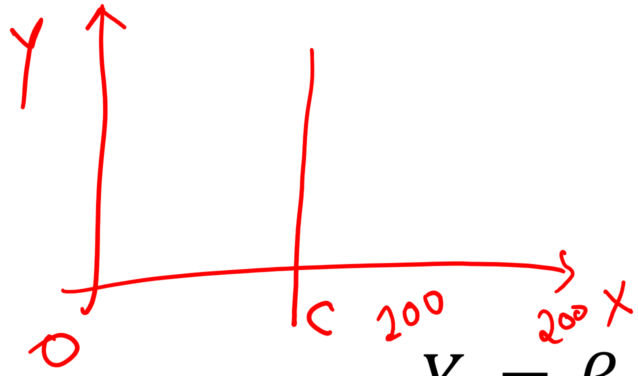
Identification Result

The treatment effect at the threshold c is identified by:

$$\begin{aligned} \tau_{\text{LATE}} &= E[Y_{1i} - Y_{0i} | X = c] \\ &= E[Y_{1i} | X = c] - E[Y_{0i} | X = c] \end{aligned}$$

But we can't observe both of these! However, **if the potential outcomes are continuous around** c, then we can estimate these values through the limits of the observed outcomes from above and below c:

$$\hat{\tau}_{\text{LATE}} = \lim_{X \downarrow c} E[Y_i | X = c, D_i = 1] - \lim_{X \uparrow c} E[Y_i | X = c, D_i = 0]$$



$D_i = 1$ if $X_i > C$
$\quad\;\; 0$ if $X_i \le C$

D X Y X−100
−ve
−ve
0
+ve
+ve

RDD Regressions

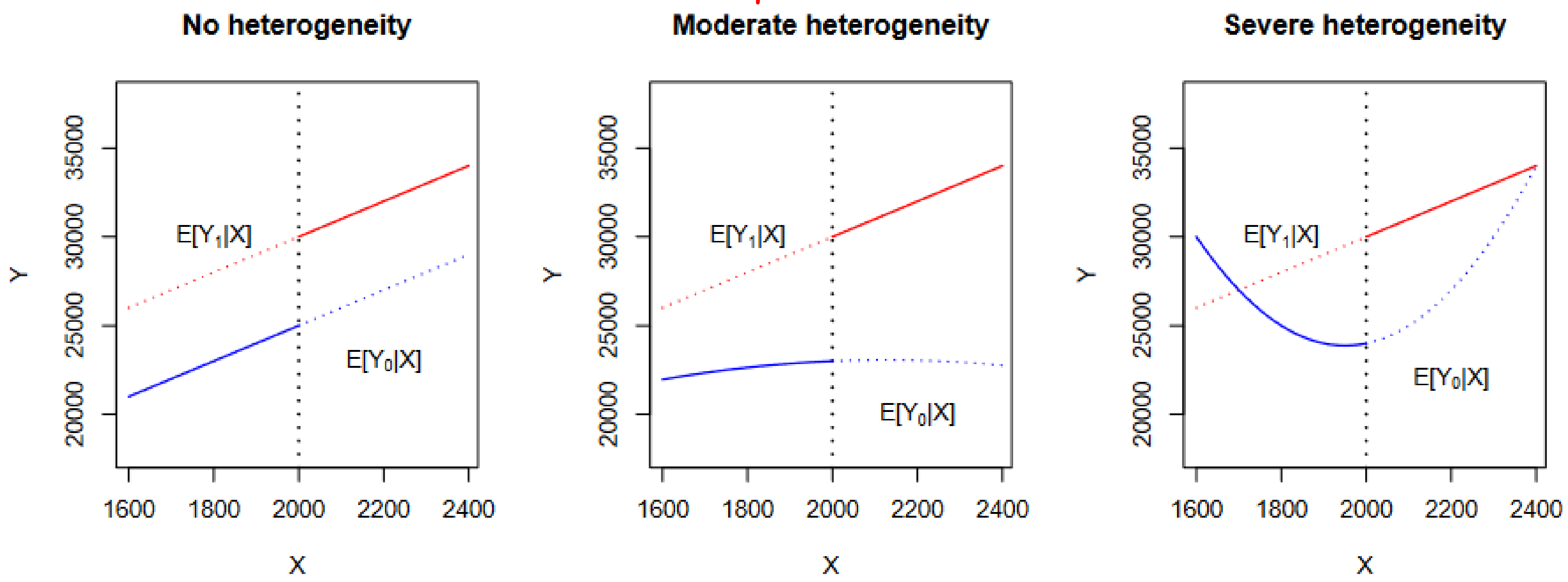
X_i Threshold

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

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

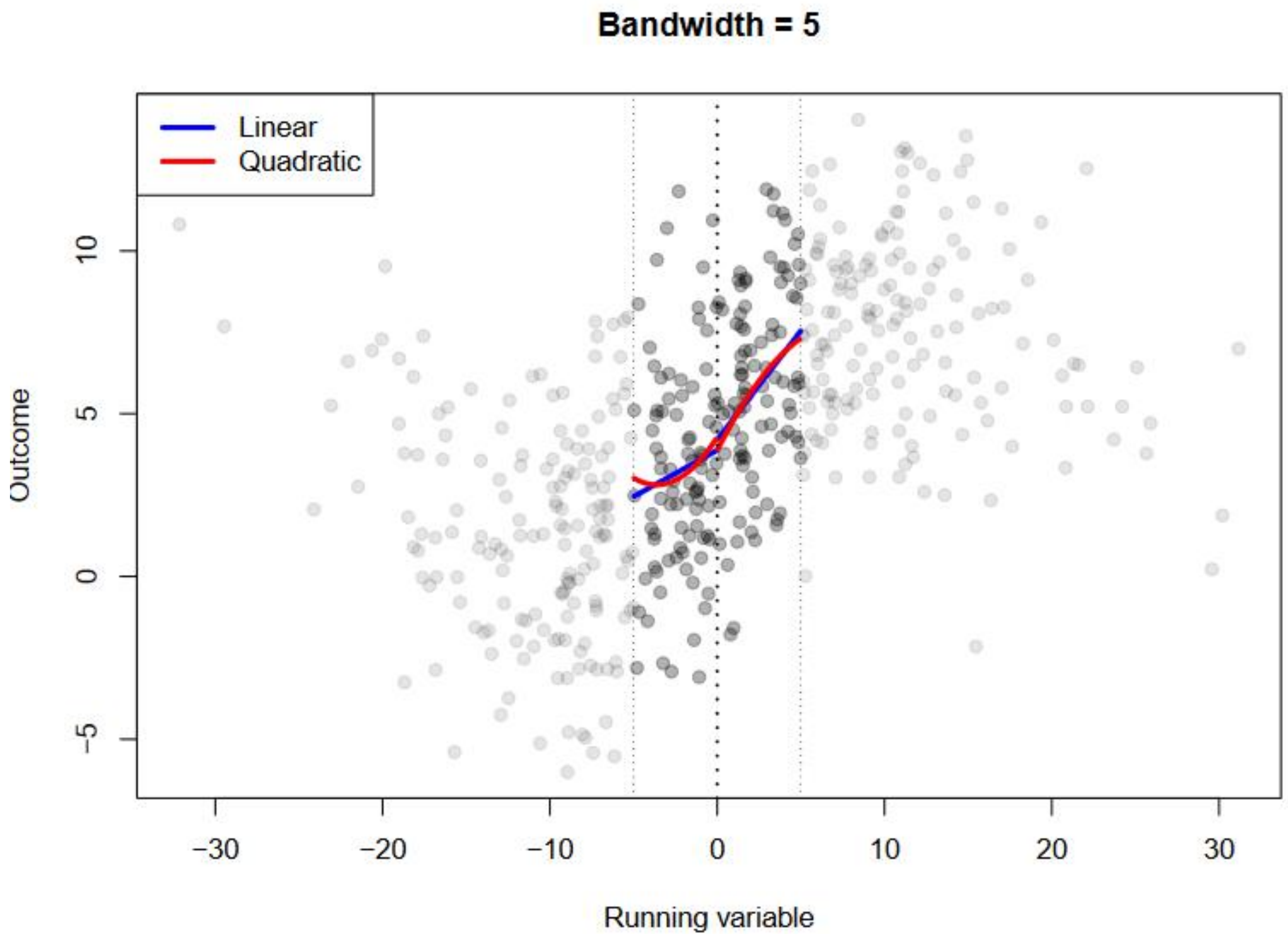
Predictors

$\beta_2 = 0.3$



Bandwidth Method

- 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

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

Recap

Summary

- Use robustness checks for DID
- 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

- Objectives achieved:
  - Can understand the advantages and design of Regression Discontinuity
  - Can interpret the regression coefficients of RDD

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 ☺

Source: Matheus Facure Alves