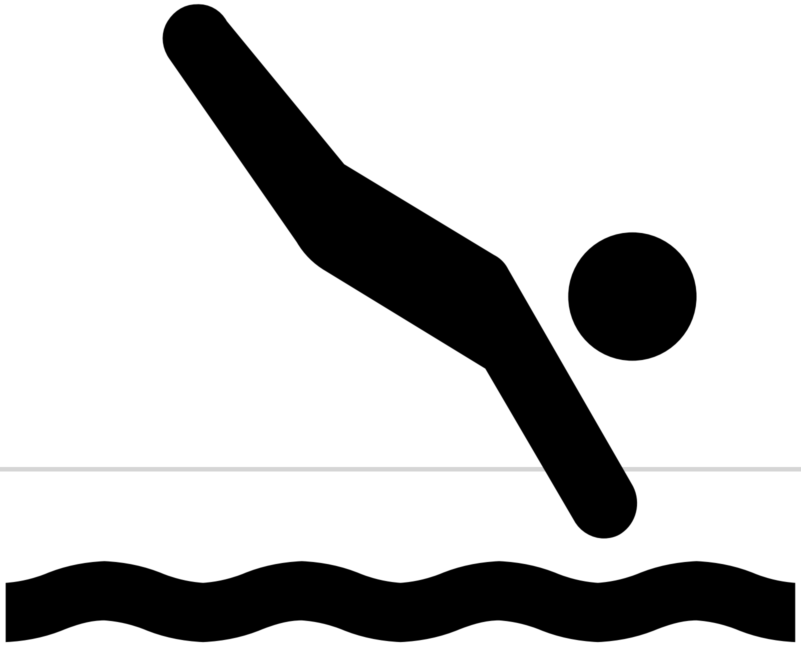
Lecture 8 – Quasi-experimental: Regression discontinuity design

Xi Chen

6/13/2023

Table of Contents



Getting a little “jumpy”

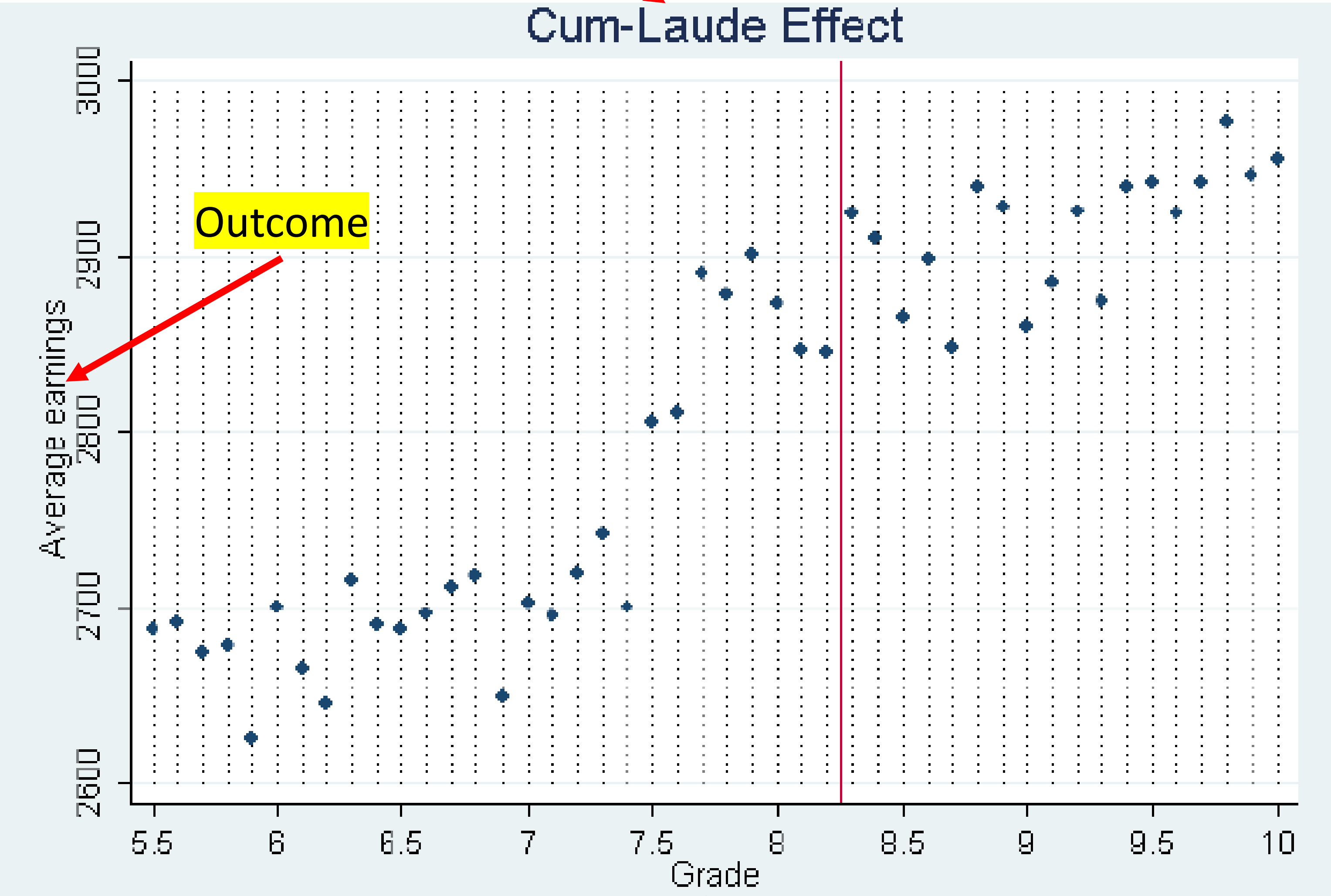
The basic idea of
Regression **D**iscontinuity **D**esign **(RDD)**

The cum laude effects on earnings at RSM

- We cannot run an experiment!

- How the cum laude is determined?
- Eligibility
- $\begin{cases} D_i = 1, if\ X_i \geq 8.25 \\ D_i = 0, if\ X_i < 8.25 \end{cases}$

Treatment



Outcome

Running (or forcing) variables

Exploiting variations from “rules”

Policies, rules or regulations create an assignment of treatment based on eligibility…

Some examples of "eligibility design"

Naturalization (years) on political integration of immigrants

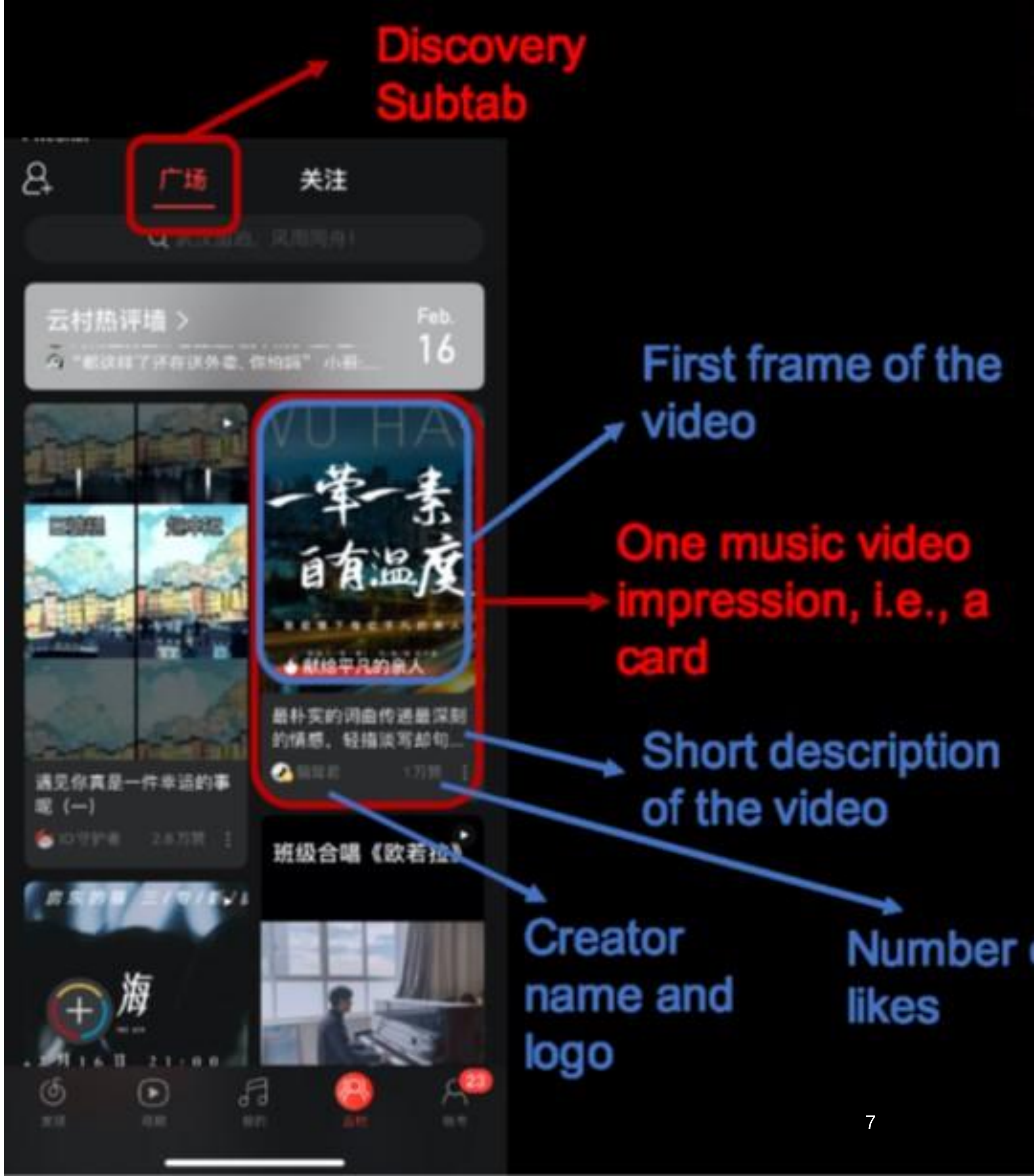
Loyalty program (consumer scores) on consumer purchases

College admission (GPA scores) on future earnings

Alcohol consumption (legal drinking age) on health outcomes

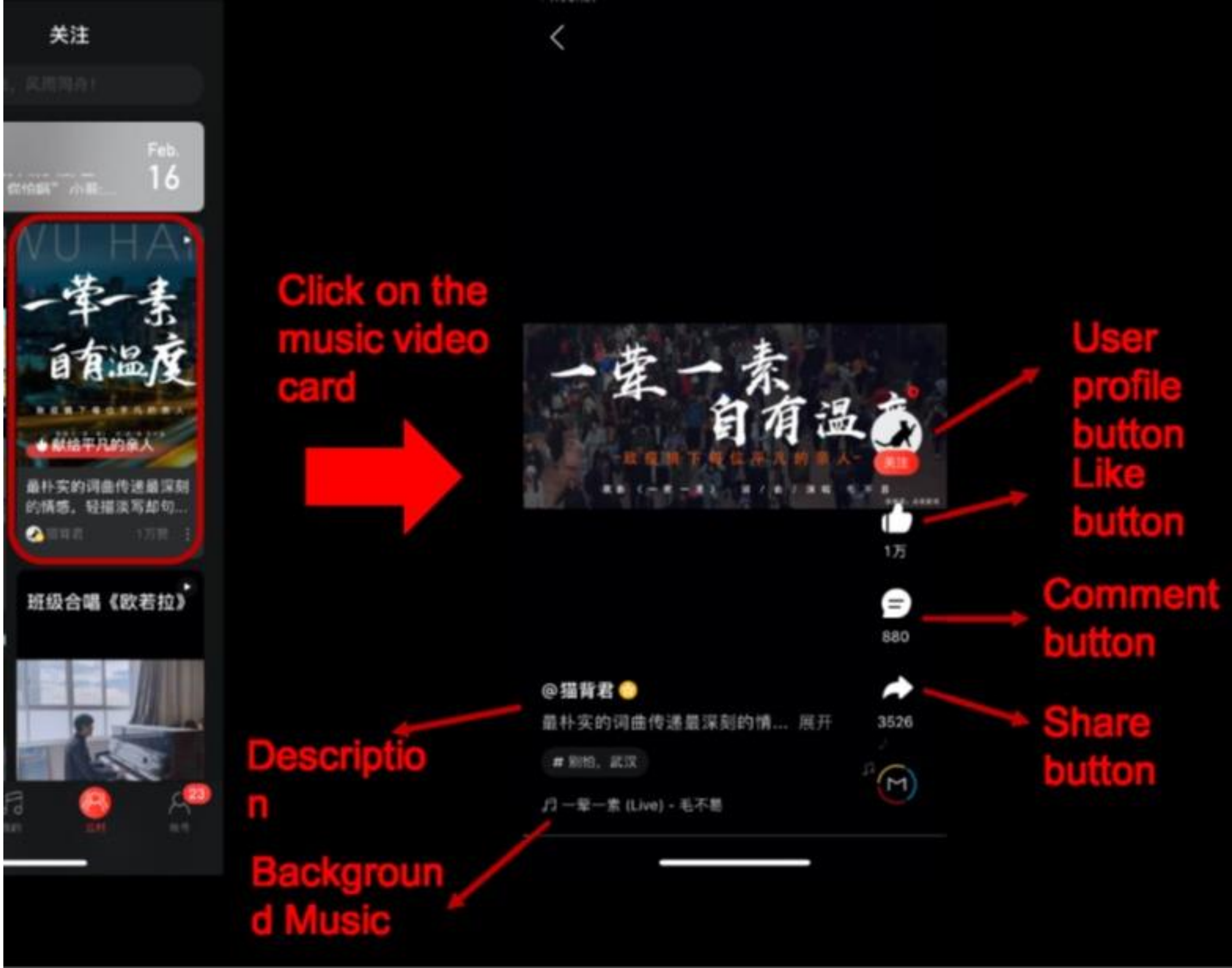
Some examples of “eligibility design”

- Evaluating the effects of the recommender system at NetEase.
- NetEase is a Chinese music streaming service (e.g., Spotify).
- Recommender system called “discovery” in their app.



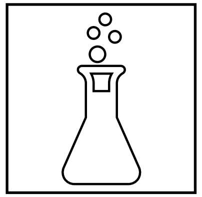
Some examples of “eligibility design”

- Clicking through the recommended cards



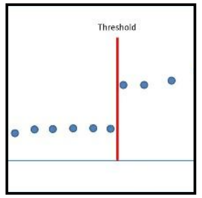
Some examples of “eligibility design”

- NetEase wanted to evaluate the performance of their recommender system.



Experiments

A random recommendation to the treatment group



RDD as a pre-experiment

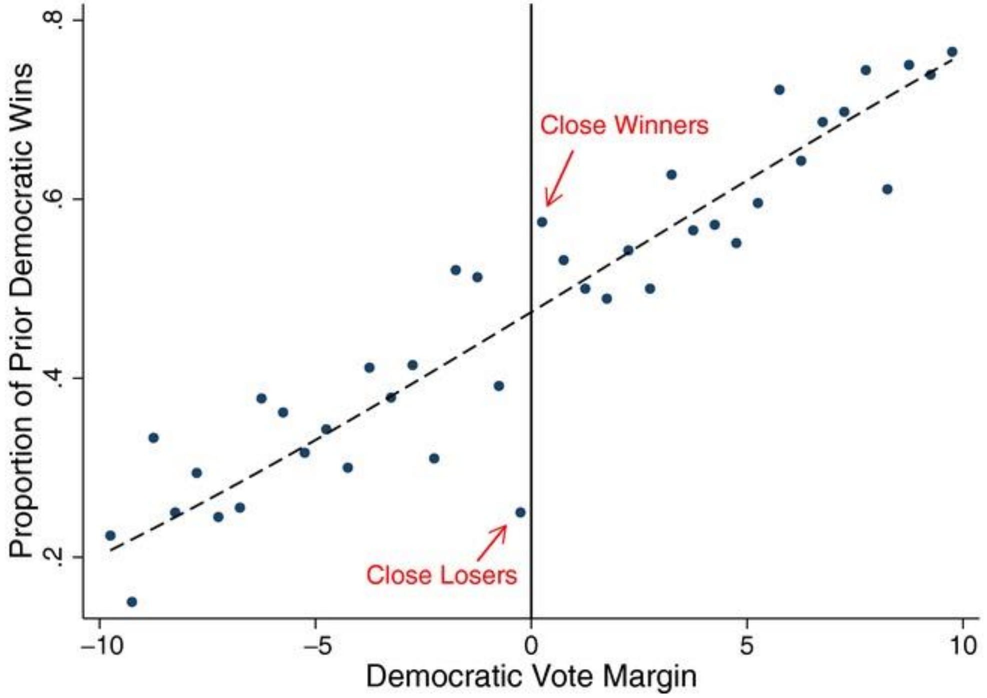
Running variable: matching scores calculated from the recommender system

Treatment: Recommended on the discovery tab

Outcome: the clicks

Some examples of "eligibility design"

- In political science, using "close elections" to examine the incumbency effects – the incumbent candidates / parties can partially control the election results.



Source: Eggers, A. C., Fowler, A., Hainmueller, J., Hall, A. B., & Snyder Jr, J. M. (2015). On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. *American Journal of Political Science*, *59*(1), 259-274.

Some examples of “eligibility design”

- In accounting research, Bird and Karolyi (2017, retracted) exploited the discontinuous relationship between **market capitalization** and **assignment to either the Russell 1000 or the Russell 2000 index** to test for the effect of institutional ownership on tax planning.

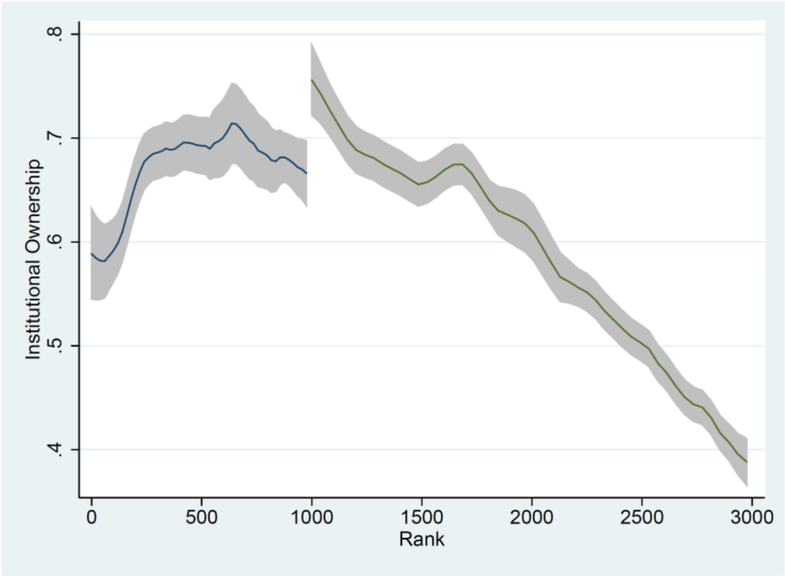
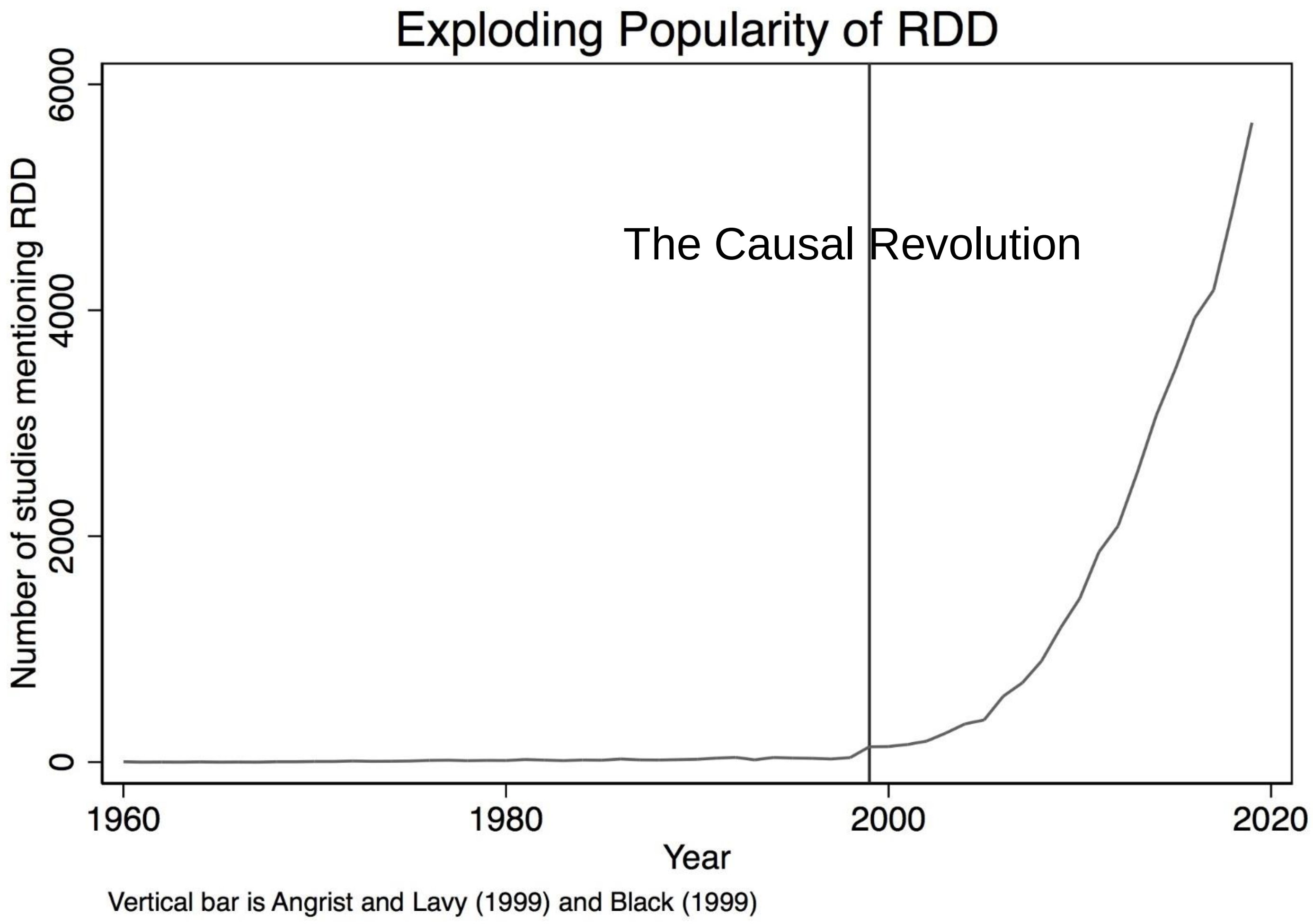


Figure 1. Institutional Ownership over Russell Index Ranks

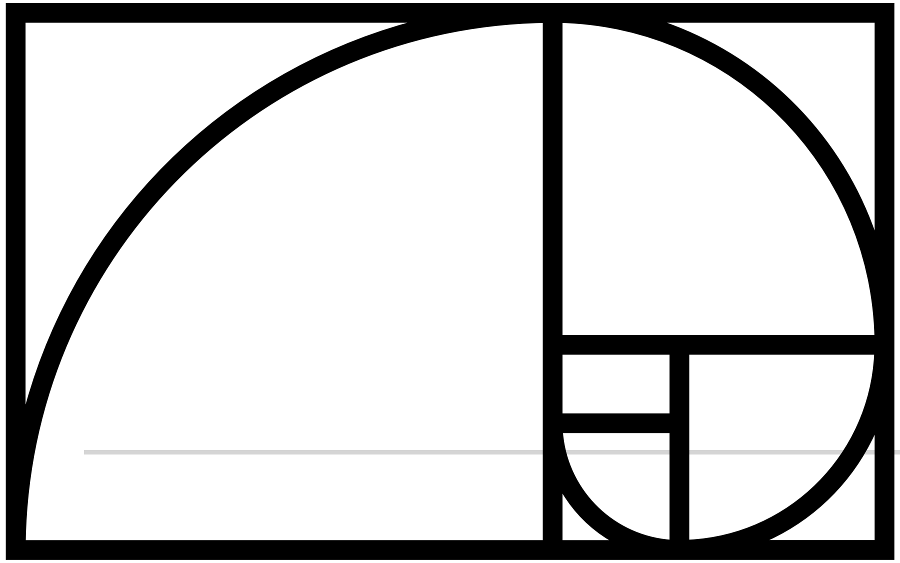
Source: Bird, A., & Karolyi, S. A. (2017). Governance and taxes: evidence from regression discontinuity (retracted). *The Accounting Review*, *92*(1), 29-50.

The origin of RDD and its explosion in recent years



Source: Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.

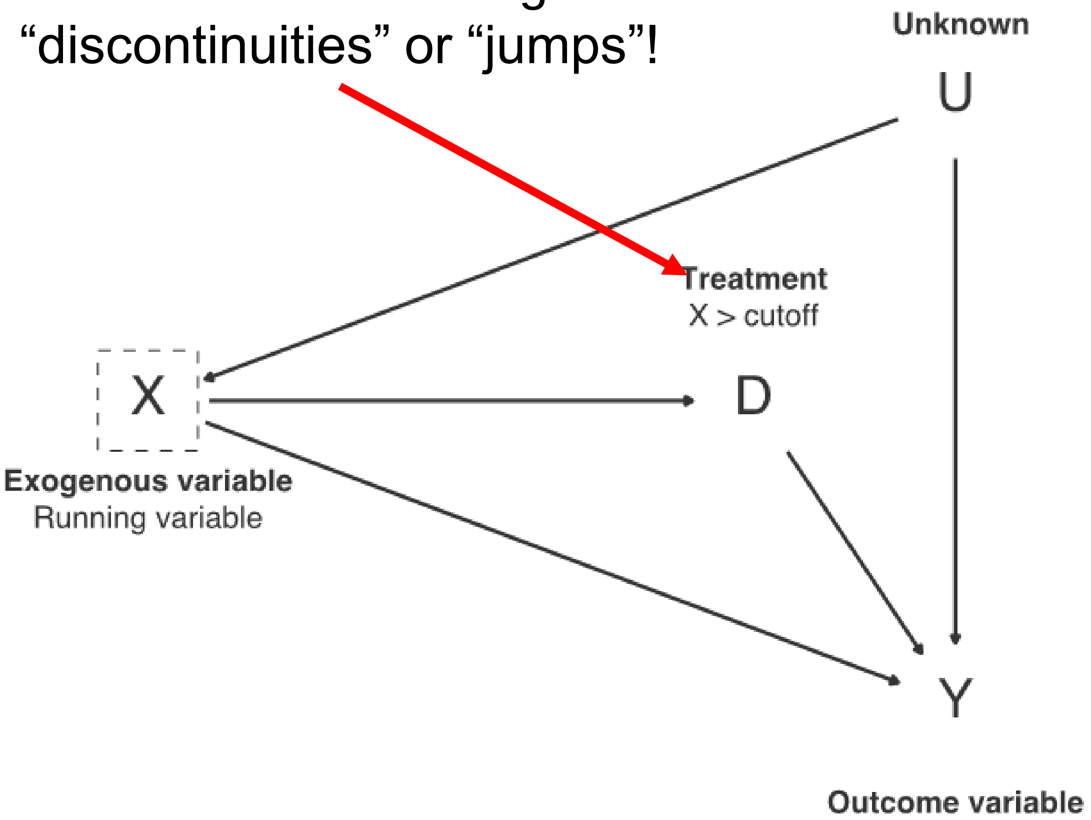
- The first research that used RDD is Thistlehwaite and Campbell (1960).
- They studied the effect of merit awards on future academic outcomes with scores as the running variable.



RDD: A formal examination

The logic of RDD

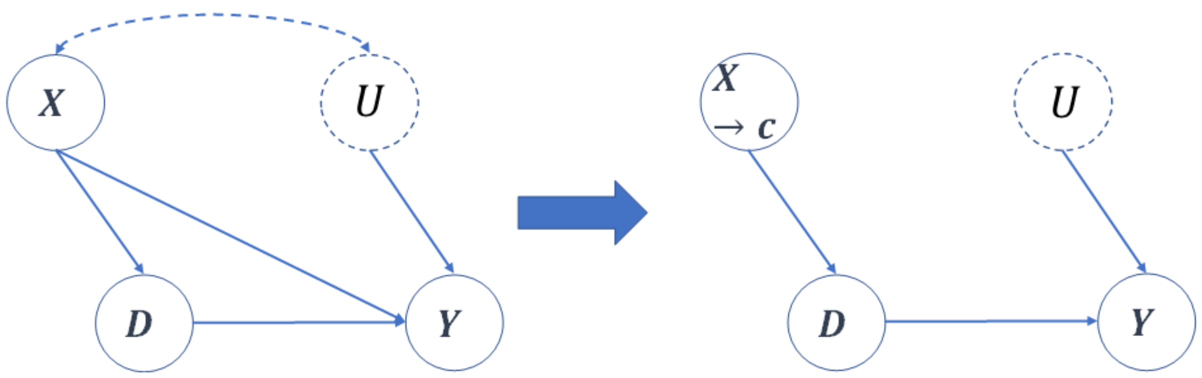
We are searching
“discontinuities” or “jumps”!



- Effects of the treatment on the outcome
- **Confounders exist!**

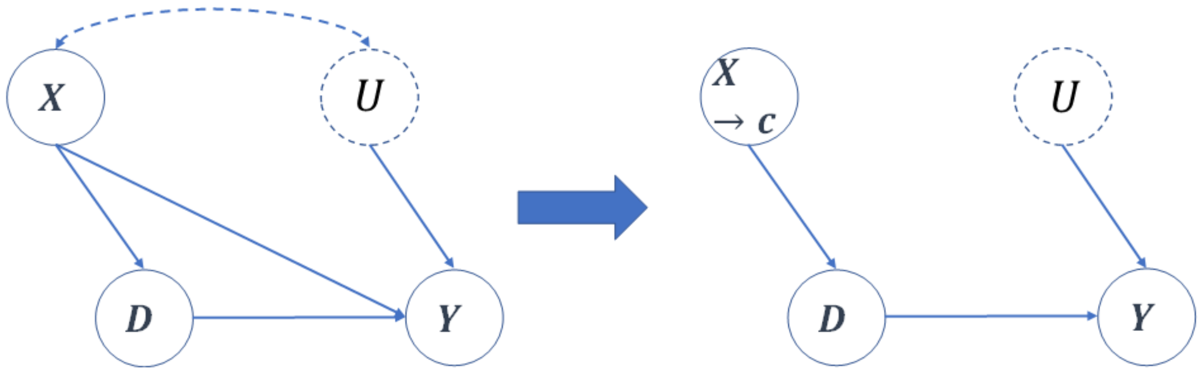
- Knowledge about the assignment:
  - Running (or forcing) variables
  - Cutoff

The DAG representation of RDD



- The treatment D is assigned based on the running variable X ($X \to D$).
- X is likely to be associated with the confounders U ($X \leftrightarrow Z$).
- For the DAG on the left, $P(Y \mid D)$ cannot be identified due to unblocked paths.

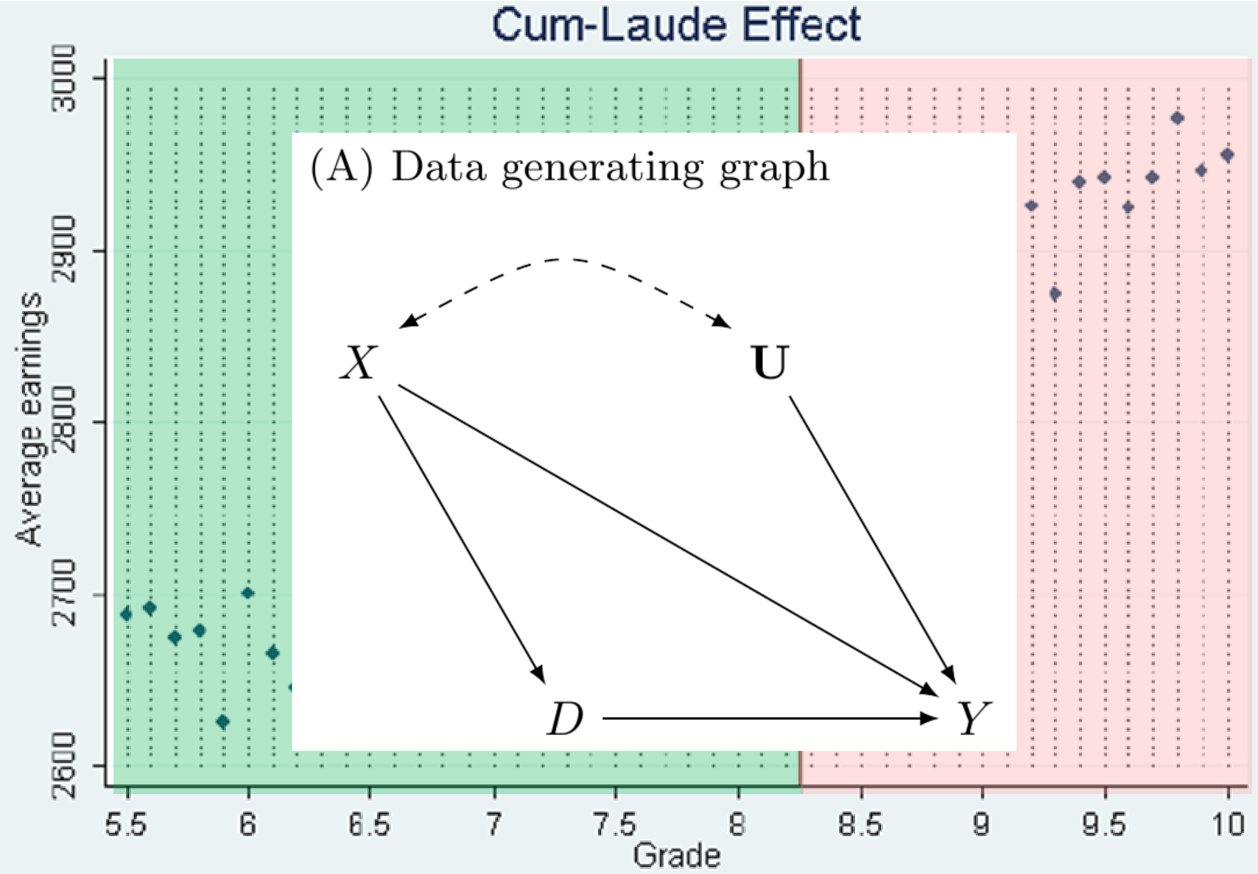
The DAG representation of RDD

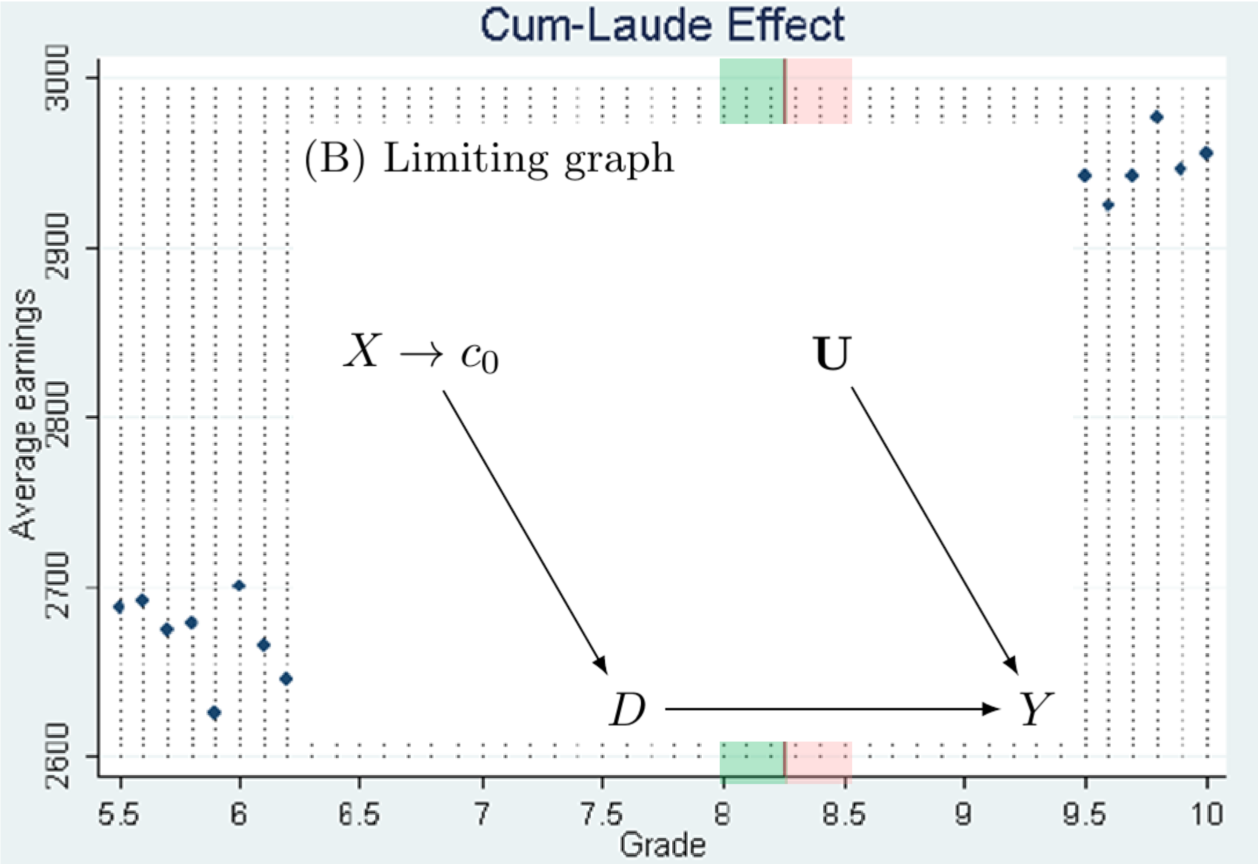


- The magic happens **at the limit or the cutoff $\boldsymbol{c}$**.
- At the cutoff c, the treatment D only depends on c.
- Therefore, X is excluded or the path $X \rightarrow Y$ is blocked by the cutoff c.
- For the DAG on the right, $P(Y \mid D)$ is identified as the back-door paths between D and Y are blocked.

DAG on data

Focusing on a close neighborhood around the cutoff.



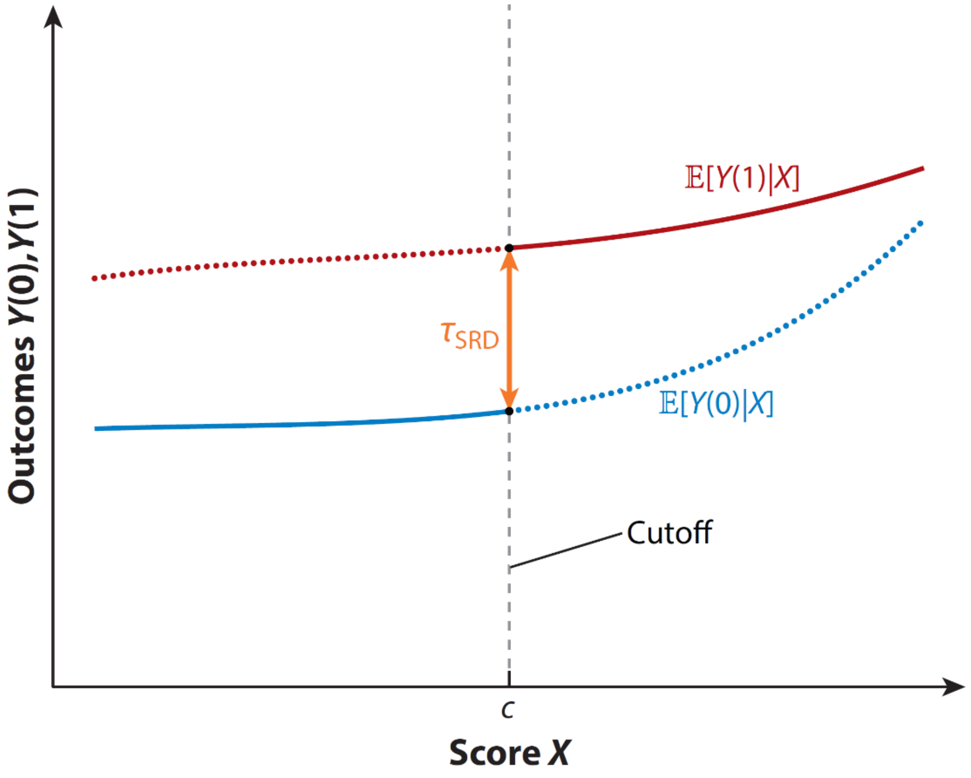


A formal representation of RDD

- Suppose we have:
  - A running variable X_i with the cutoff c.
  - The treatment of a unit D_i.
  - The outcome of a unit Y_i.

- Given the setup of RDD, the treatment assignment is:
  - $P(D_i = 1 \mid X_i \geq c) = 1$
  - $P(D_i = 1 \mid X_i < c) = 0$

- The potential outcome and the observed outcome are:
  - $Y_i = Y_i^1 D_i + Y_i^0 (1 - D_i)$



A formal representation of RDD

- The treatment effect is identified at the cutoff $X_i = c$:

$$\tau_{RDD} = E\left(Y_i^1 - Y_i^0 \mid X_i = c\right)$$

- To define τ_{RDD} in terms of limits:

$$\tau_{RDD} = \lim_{X_i \downarrow c}\left(E\left(Y_i^1 \mid X_i = x\right)\right) - \lim_{X_i \uparrow c}\left(E\left(Y_i^0 \mid X_i = x\right)\right)$$

Assumptions of RDD: A call-back to matching

- **Conditional unconfoundedness**: Treatment assignment D_i is unconfounded conditional on X_i.

$$Y_i^1, Y_i^0 \perp D_i \mid X_i$$

- ~~**Overlap assumption**: for all values of the covariates there are both treated and control units.~~

$$\cancel{0 < P(D_i \mid X_i) < 1}$$

- **But, we have**

$$P(\, D_i = 1 \mid X_i < c\,) = 0$$
$$P(D_i = 1 \mid X_i \geq c) = 1$$

6/13/2023

Assumptions of RDD:
The implication of violating the overlap assumption

$$\tau_{RDD} = \lim_{X_i \downarrow c} \left(E\left(Y_i^1 \mid X_i = x \right) \right) - \lim_{X_i \uparrow c} \left(E\left(Y_i^0 \mid X_i = x \right) \right)$$

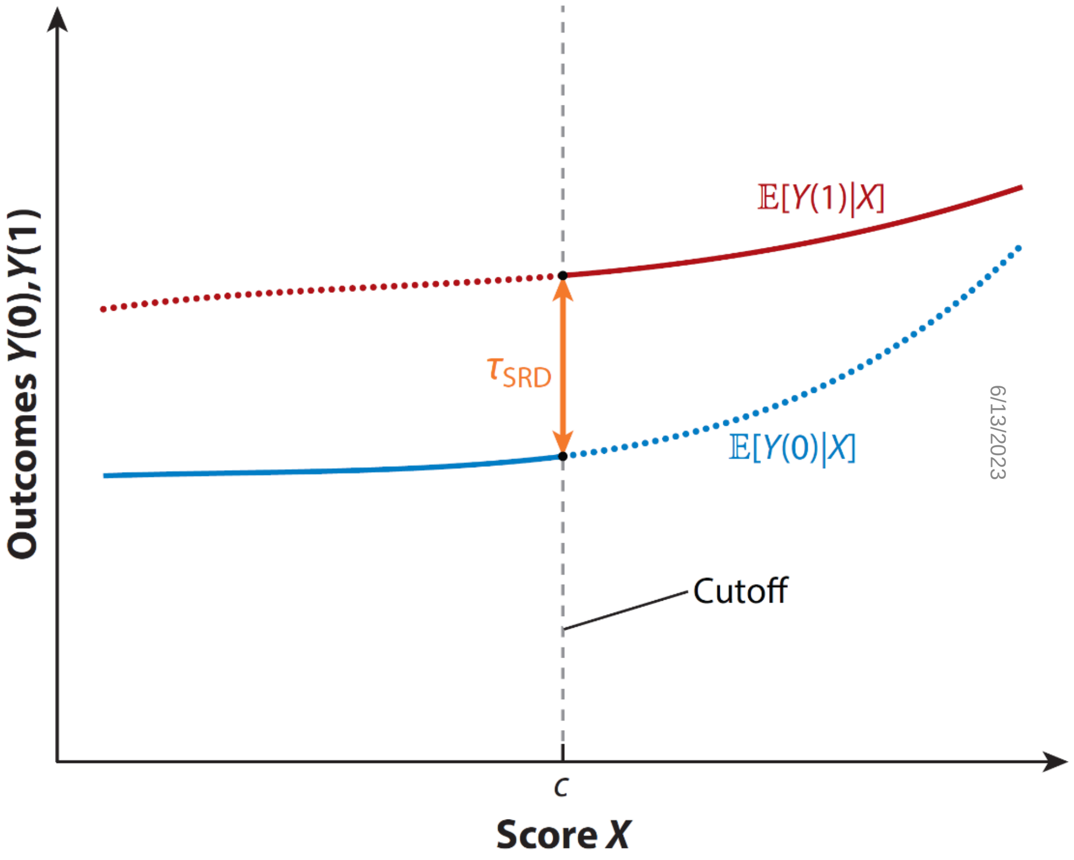
- We need both $\lim_{X_i \downarrow c} \left(E\left(Y_i^1 \mid X_i = x \right) \right)$ and $\lim_{X_i \uparrow c} \left(E\left(Y_i^0 \mid X_i = x \right) \right)$.
- **However, we at best observe $\boldsymbol{\lim_{X_i \downarrow c} \left(E\left(Y_i^1 \mid X_i = x \right) \right)}$ because $\boldsymbol{P(D_i = 1 \mid X_i \geq c) = 1}$.**
- Extra assumptions are needed!

Assumptions of RDD:
What extra assumptions?

- Because we do not observe $\lim_{X_i \uparrow c}\left(E\left(Y_i^0 \mid X_i = c\right)\right)$, we need to somehow “predict” it.

- Under unconfoundedness, we can use the observation of the outcome of untreated group Y^0 and their forcing variable X^0 ($X^0 < c$) to learn the function:
$$\boldsymbol{\hat{E}\left(Y_i^0 \mid X_i < c\right)}$$

- **With this function, we can “extrapolate” $\boldsymbol{\hat{E}\left(Y_i^0 \mid X_i = c\right)}$.**



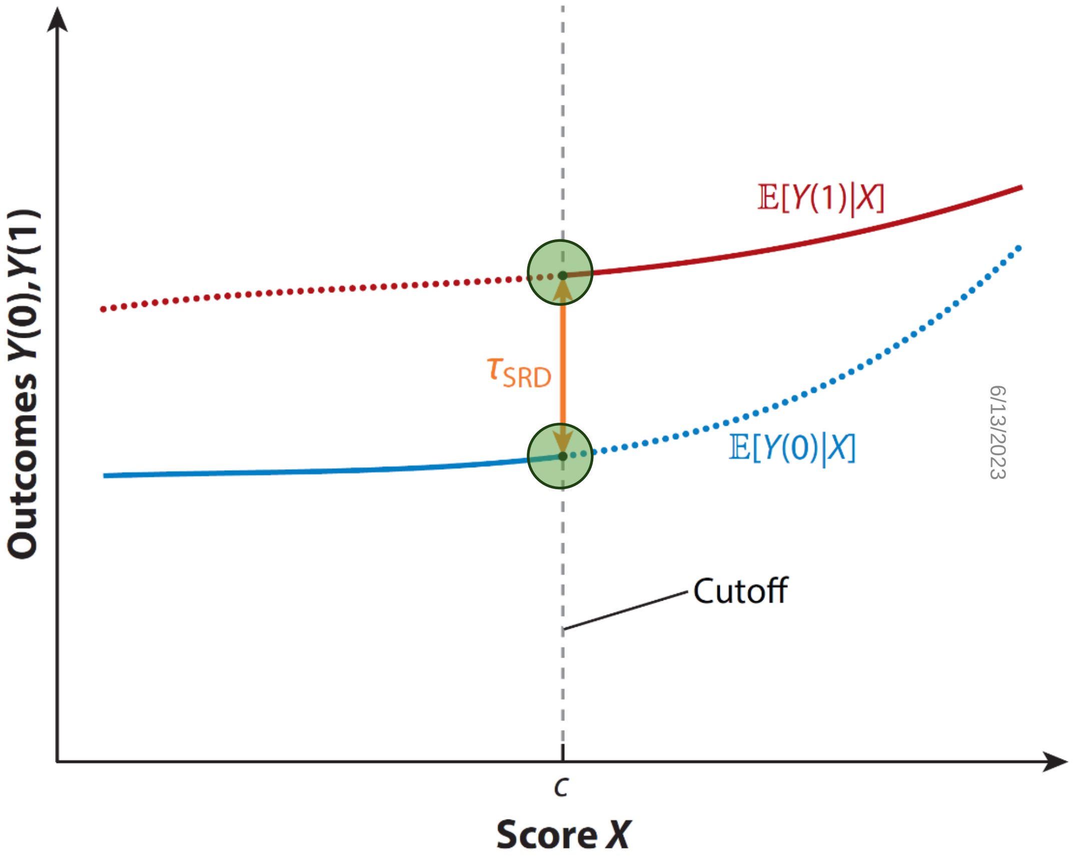
6/13/2023

Assumptions of RDD:
What extra assumptions?

- With this function, we can “extrapolate” $\hat{E}\left(Y_i^0 \mid X_i = c\right)$.

- **Question: for a valid extrapolation, what assumption is needed?**

- **Continuity assumption:** The expectation function of potential outcomes conditional on the running variable is continuous **at the cutoff point**.

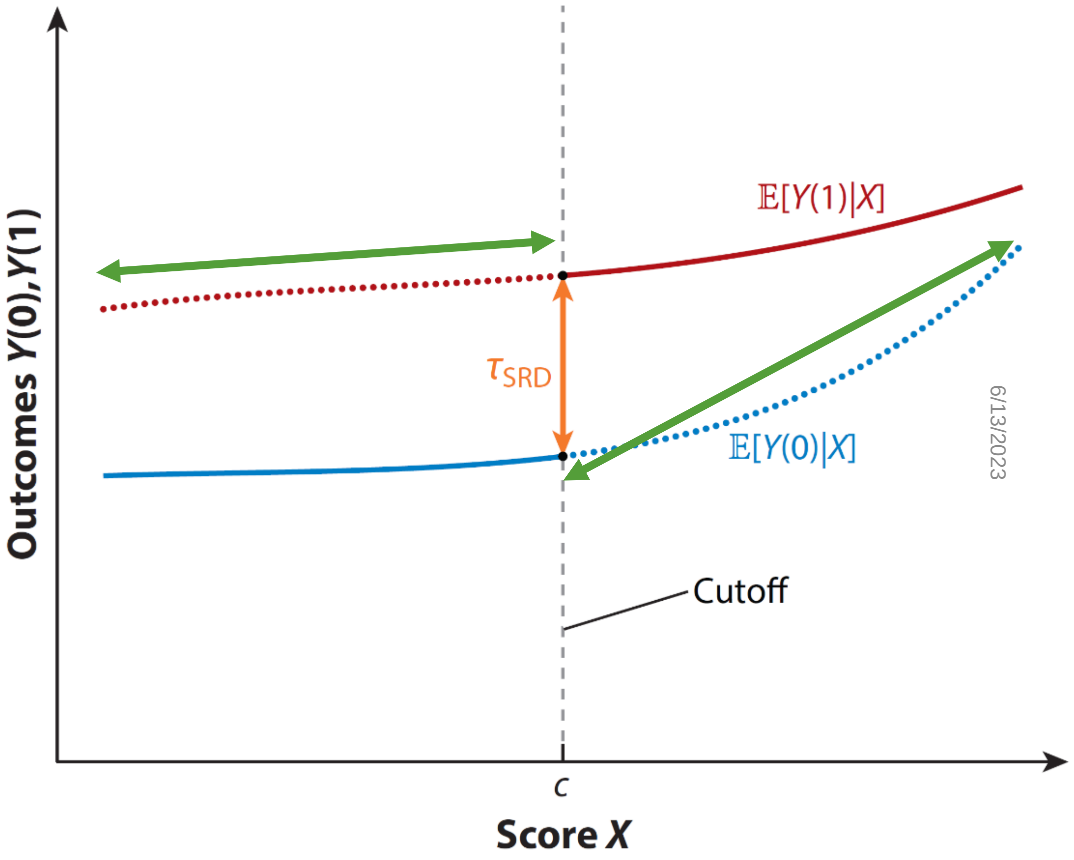


6/13/2023

Assumptions of RDD:
What extra assumptions?

- **Strong continuity assumption:** The expectation function of potential outcomes conditional on the running variable is continuous **everywhere**.

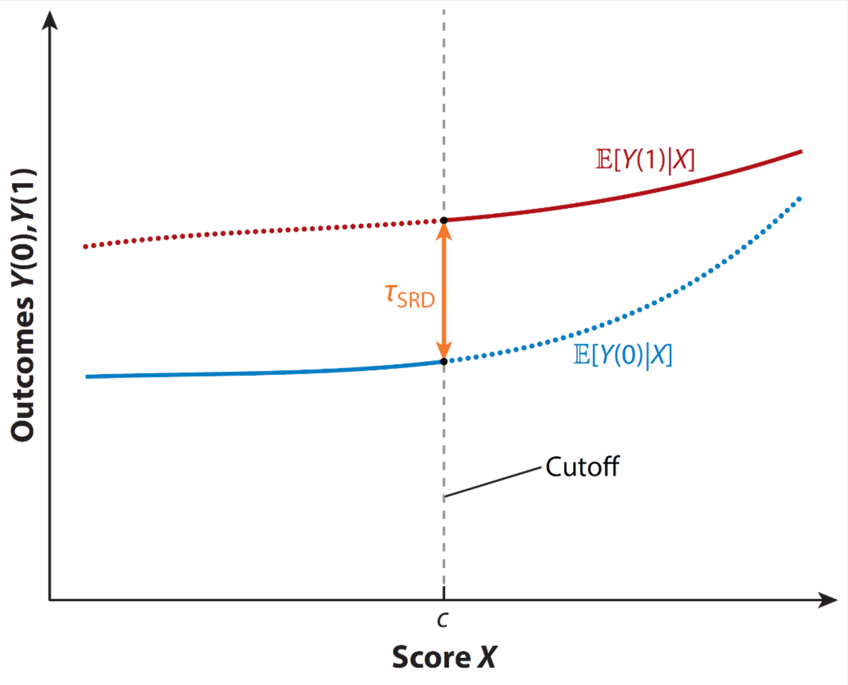
- **This assumption is sufficient but not necessary.**



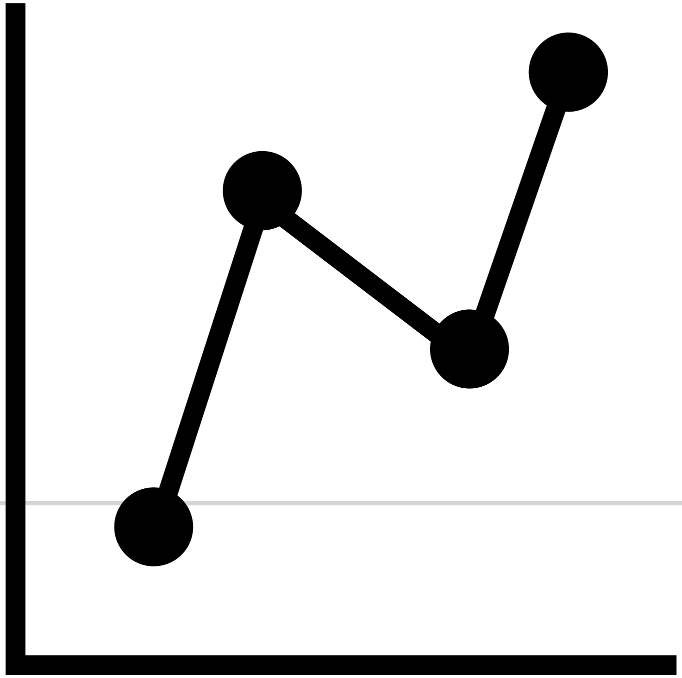
6/13/2023

Is any of the two assumptions testable?

Assumptions of RDD: A summary



- **Conditional unconfoundedness**: Treatment assignment D_i is unconfounded conditional on X_i.
- **Strong continuity assumption:** The expectation function of potential outcomes conditional on the running variable is continuous **everywhere**.

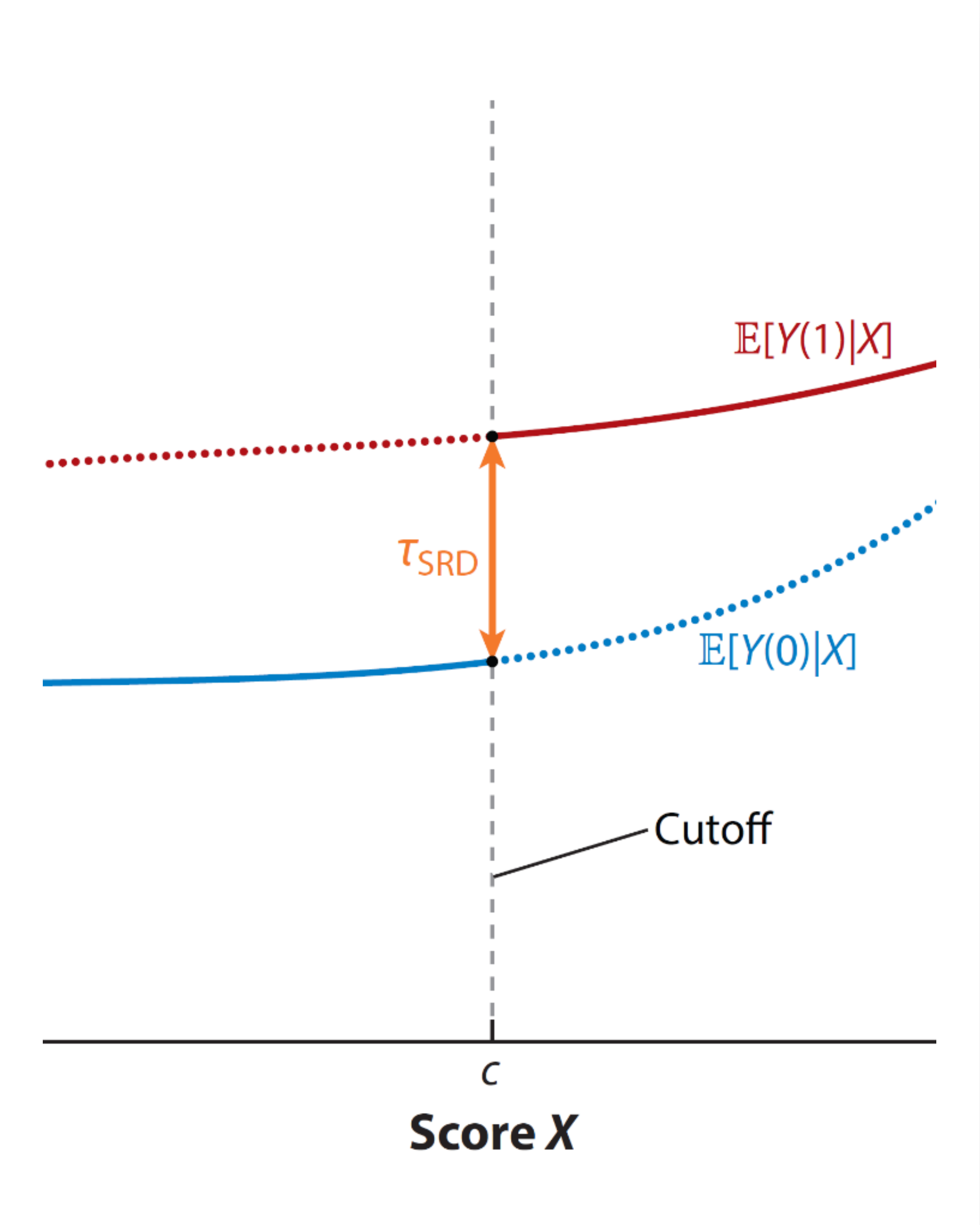


RDD: Estimation issues

RDD estimation:
how to predict the conditional expectations?

- Under the unconfoundedness assumption, we can estimate the conditional expectation function.

- Using data above the cutoff $X > c$ to estimate $E(Y^1 \mid X)$.
- Using data below the cutoff $X < c$ to estimate $E(Y^0 \mid X)$.



RDD estimation:

how to predict the conditional expectations?

- **Parametric models**

- $Y_i = \alpha + \beta D_i + \gamma X_i + e_i$
  - The base model.

- $Y_i = \alpha + \beta D_i + \gamma X_i + \theta D_i X_i + e_i$
  - A flexible spec which considers the change in slopes of the running variable.

- $Y_i = \alpha + \beta D_i + \gamma_1 X_i + \gamma_2 X_i^2 + \theta_1 D_i X_i + \theta_2 D_i X_i^2 + e_i$
  - Adding square terms or higher order terms to capture the non-linearity.

RDD estimation:

how to predict the conditional expectations?

- A side note about the polynomial model.

- Gelman and Imbens (2019) pointed our several problems and recommend using only polynomials up to the second degree (quadratic).

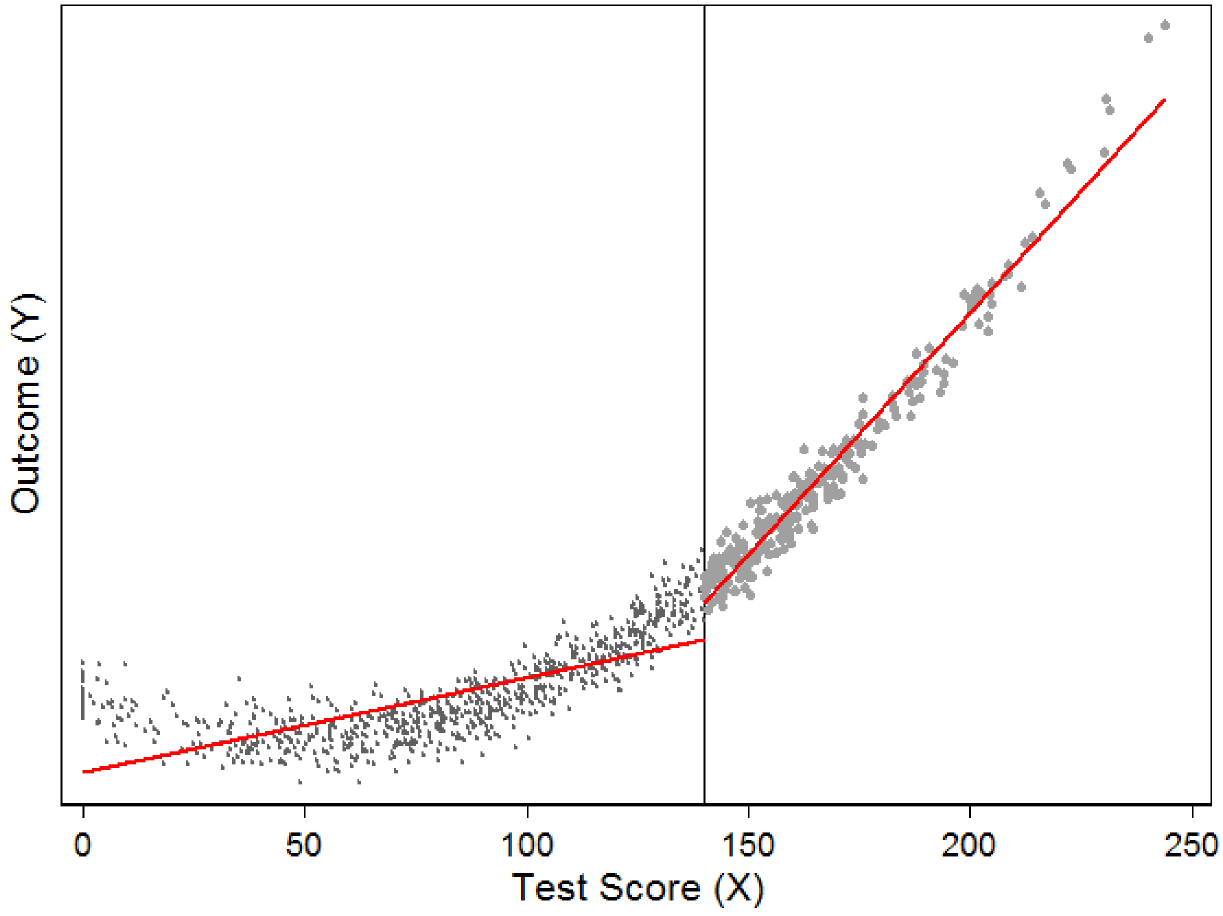
- They justify the approach in three ways:
  - Polynomials impose weights that can be noisy with polynomials of higher order (the average treatment effect is a weighted function of X).
  - Estimates can be sensitive to the degree of the polynomial
  - Confidence intervals don't have good coverage with higher order polynomials

Understanding specification errors.

In the graph, there is not jump, but a linear specification leads to “pseudo-jumps.”

RDD estimation:

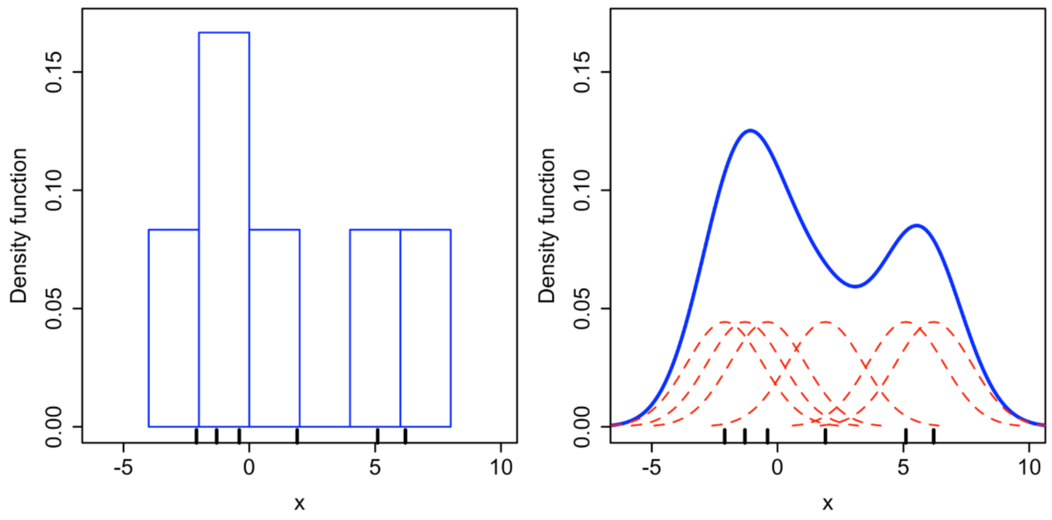
how to predict the conditional expectations?



RDD estimation:
how to predict the conditional expectations?

- Non-parametric approach to “**hopefully**” reduce specification errors.

- Every parametric model makes assumptions of the shape of the function.
- Instead, **why not make no functional assumptions and let the data inform us?**
- For this, we adopt the non-parametric approach.



RDD estimation:
how to predict the conditional expectations?

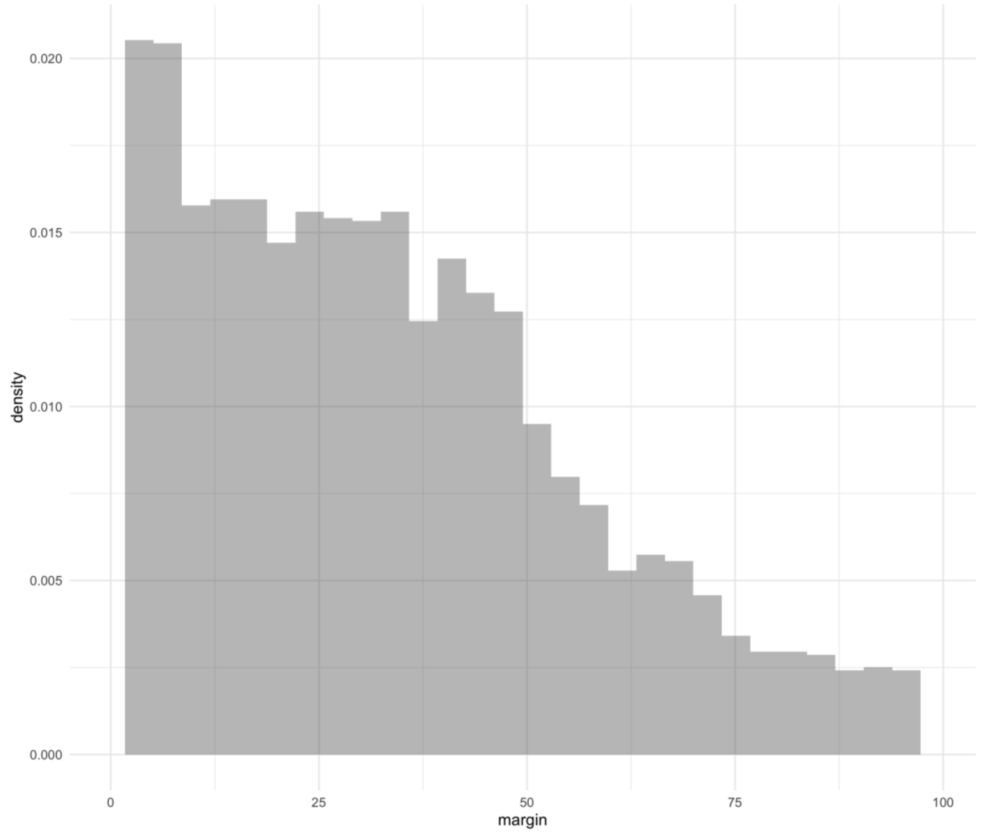
- Non-parametric estimation

- In simple cases, it's straight forward, e.g., estimate the mean $\hat{E}(X) = N^{-1}\sum X_i$ (no parameters).
- For discrete variables, use we can non-parametrically estimate the probability density function, e.g., $p_1(X) = N_1^{-1}\sum(X_i == 1)$.

- How about continuous variables?

RDD estimation:

how to predict the conditional expectations?



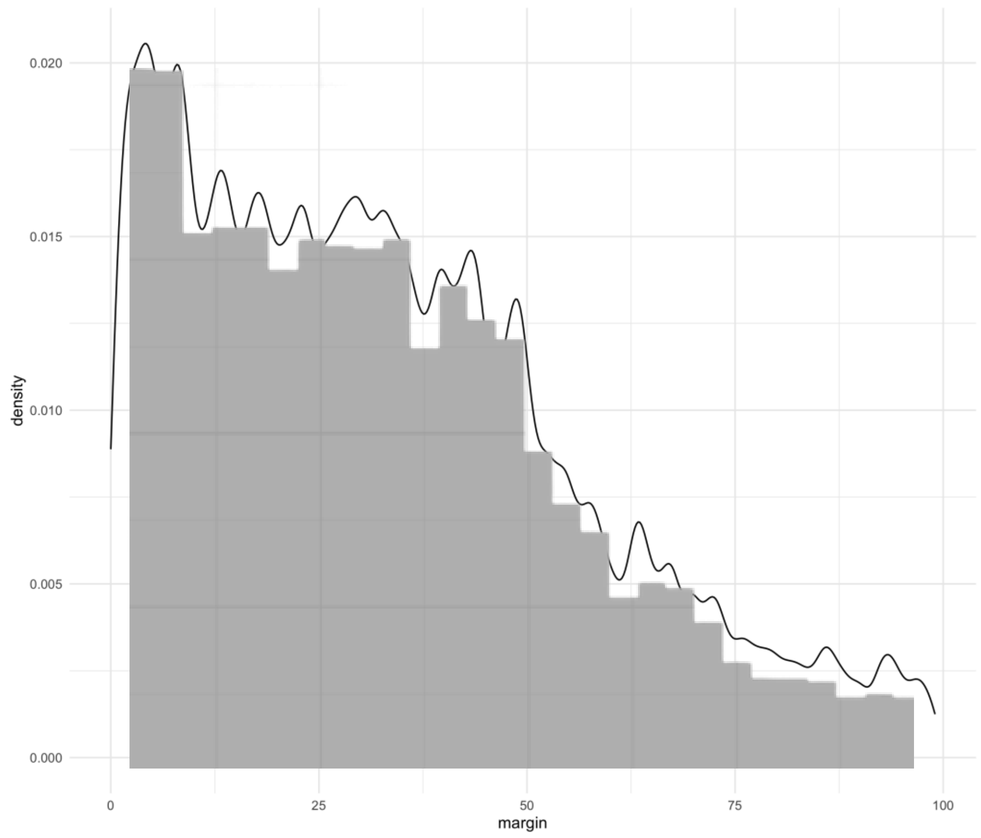
- For continuous variables, one approach to non-parametric estimation is to use histogram.
- Put data into small bins and count the frequency of bins.

RDD estimation:

how to predict the conditional expectations?

- For continuous variables, another approach is to use kernel estimation.

- To use small kernels (a smooth function) to smooth out the histogram.



RDD estimation:

how to predict the conditional expectations?

- The true functional form is unknown to us → No ground truth.
- It’s not guaranteed a non-parametric model is better than a parametric one. Or a complex model is better than a simple one.
- In practice, we often try different model specs to show the robustness.
- However, if the “jump” is big, the functional forms probably do not matter.

RDD estimation: what data to use?

- This it a problem of the variance-bias tradeoff.

To use more data away from the cutoff

→ a better recovery of the conditional expectation function

More data we use away from the cutoff

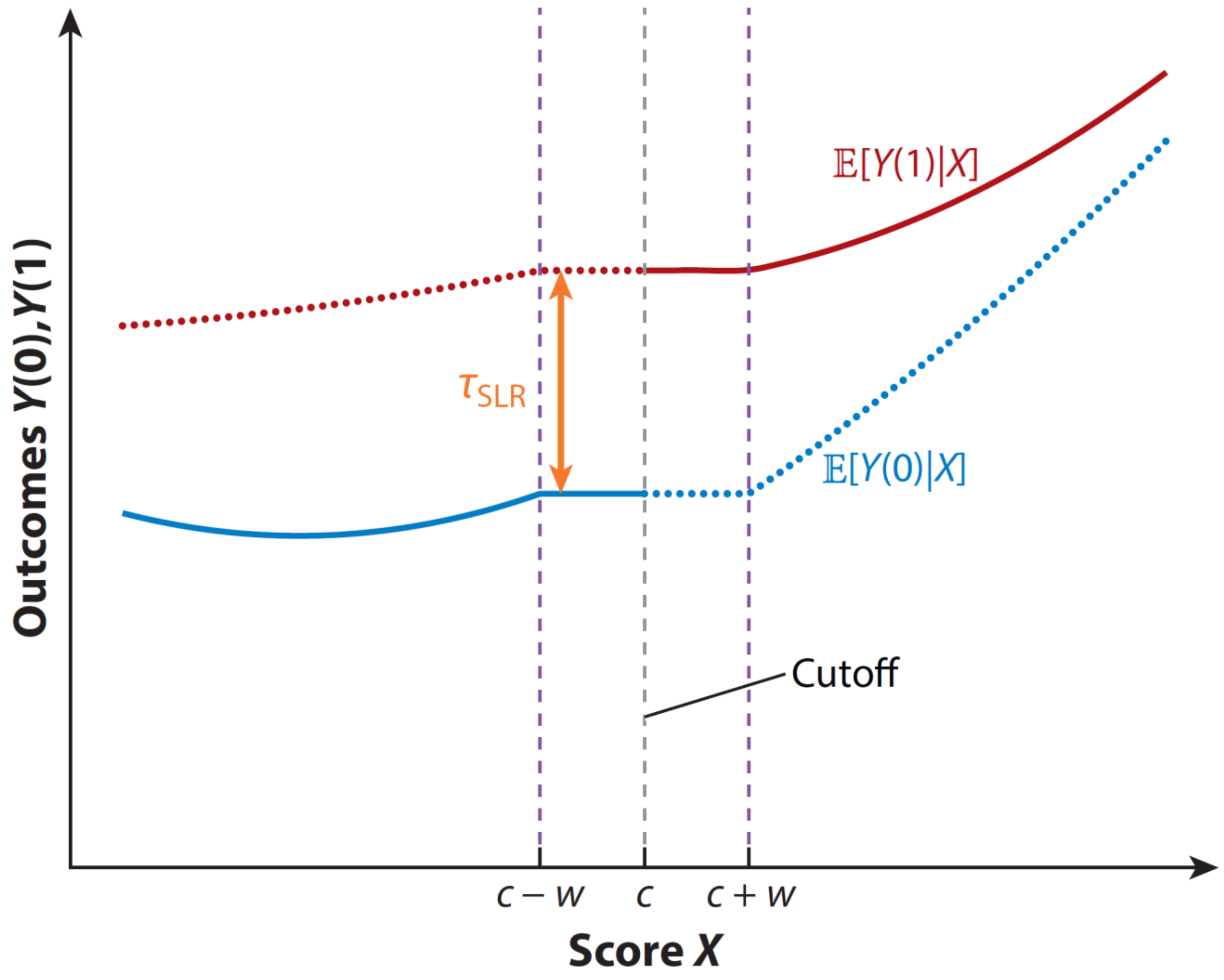
→ More rely on extrapolation

→ More bias in the treatment effects

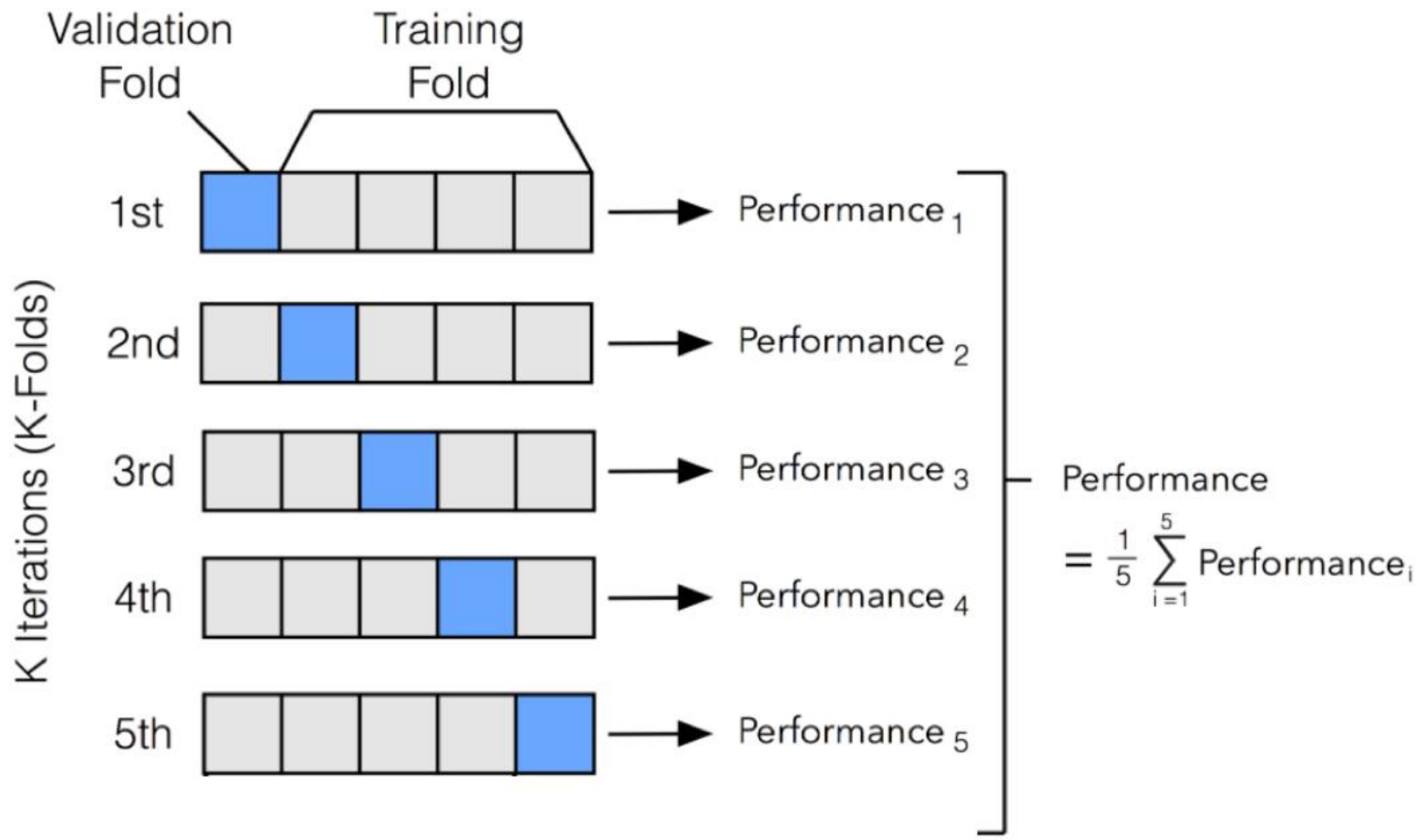
The key of RDD estimation: Choosing the bandwidth

- In practice, we need to choose the neighborhood around the cutoff (“w” in the plot).

- This is known as the “bandwidth selection.”



The key of RDD estimation: Choosing the bandwidth



A popular method of choosing the bandwidth is **cross-validation**.

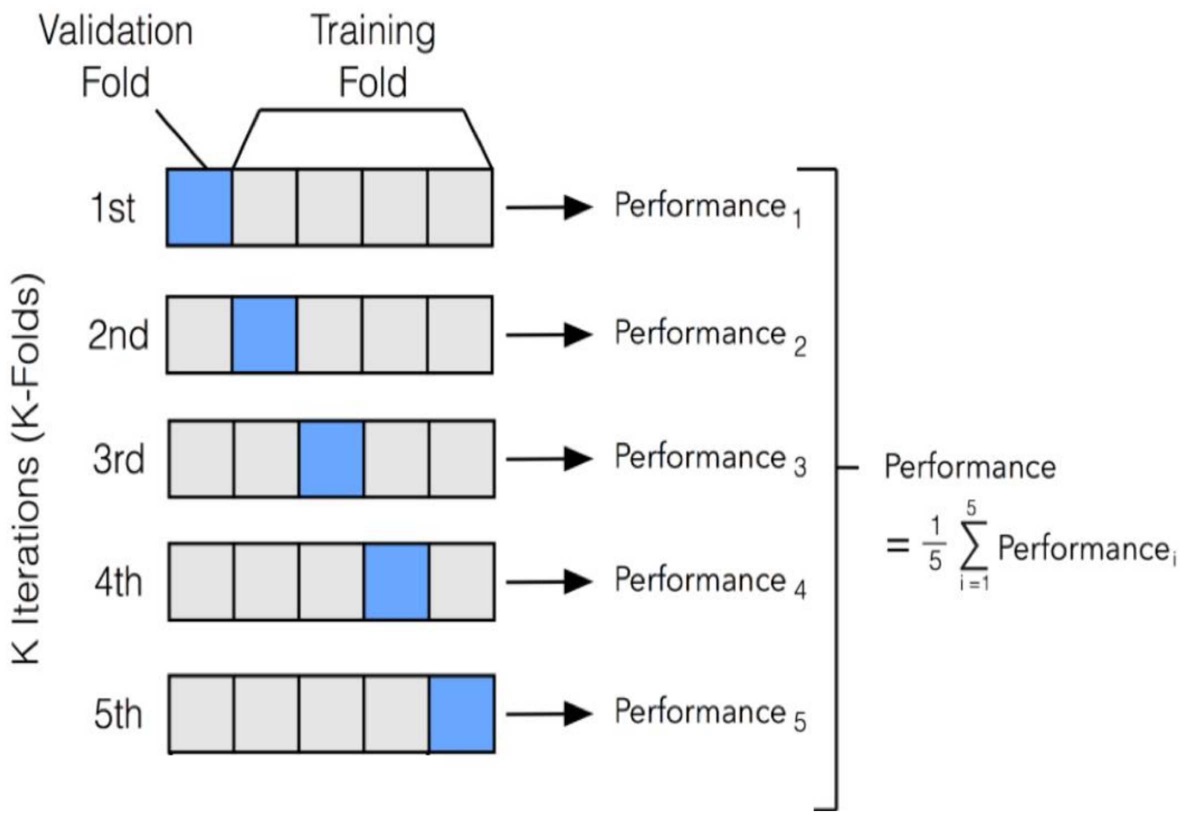
In-sample vs. out-sample fitting.

The key of RDD estimation: Choosing the bandwidth

For RDD, we split data randomly to K subsamples.

h is set as proportion to the standard deviation of the running variable with, $h = \rho SD(X_i)$.

1. We start with many bandwidths for a grid search.
2. To use h to trim the test set and train set.
3. Fit the model of conditional expectation to the train set and calculate the performance with the test set.
4. Select the optimal bandwidth.



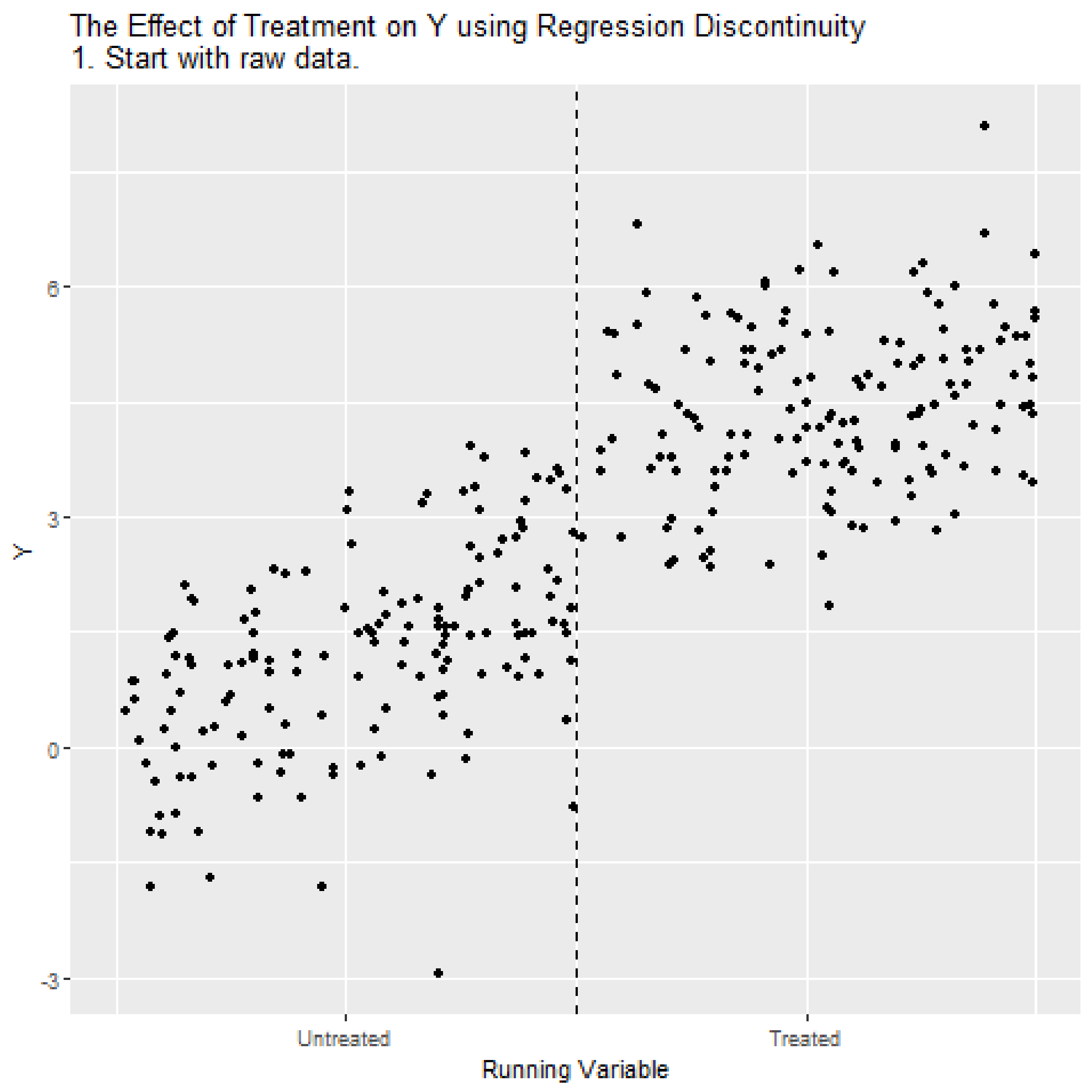
The general procedure of estimating RDD

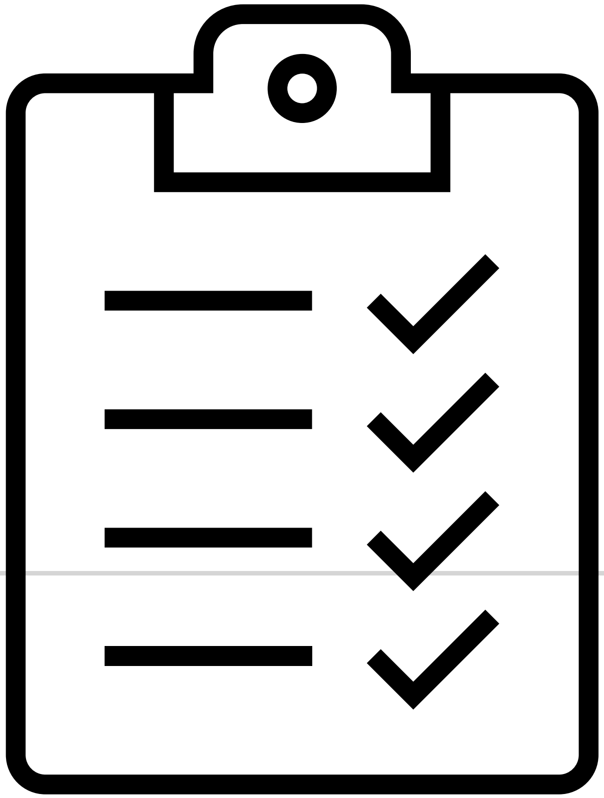
Choosing a function for conditional expectation

Choosing bandwidth

Cut out data outside the bandwidth

Get the treatment effects





RDD Checklist

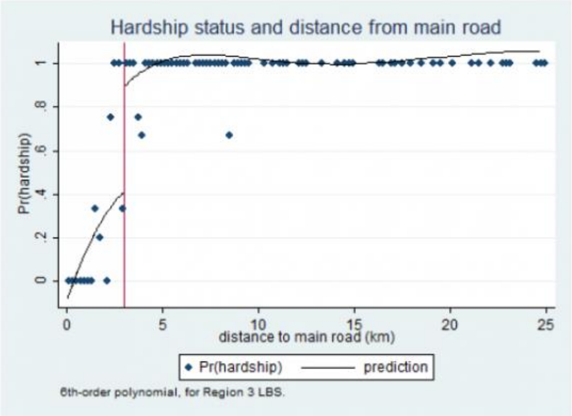
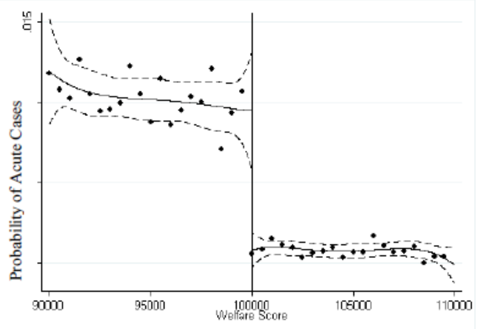
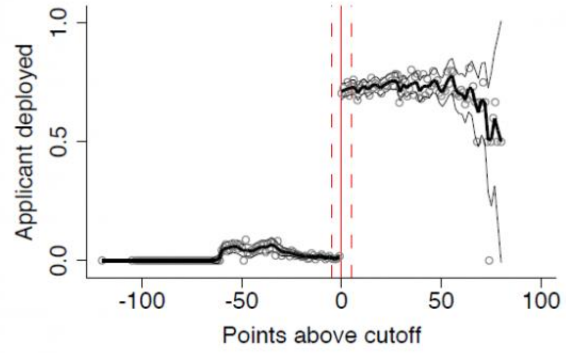


Figure 3: The Effect of MAP on Utilization of Acute Surgeries/In-Patient Services





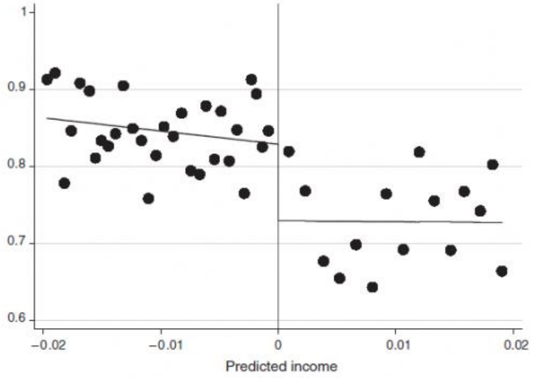
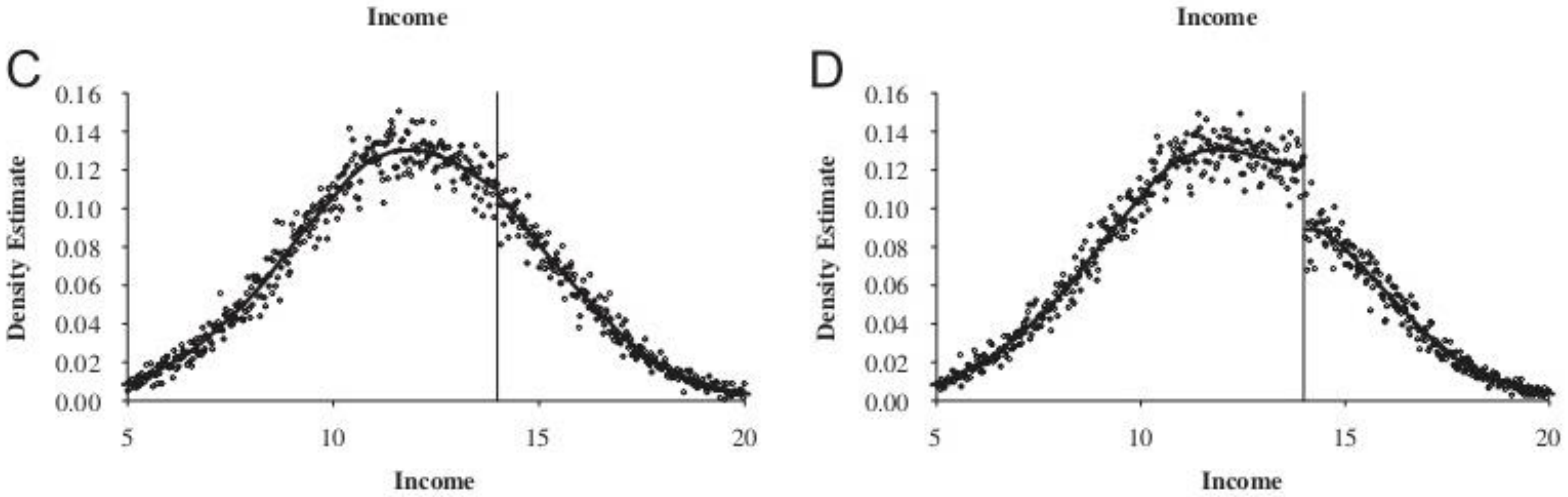


Figure 4. PANES Program Eligibility and Political Support for the Government, 2008 Follow-up Survey Round

A picture is worth thousand words

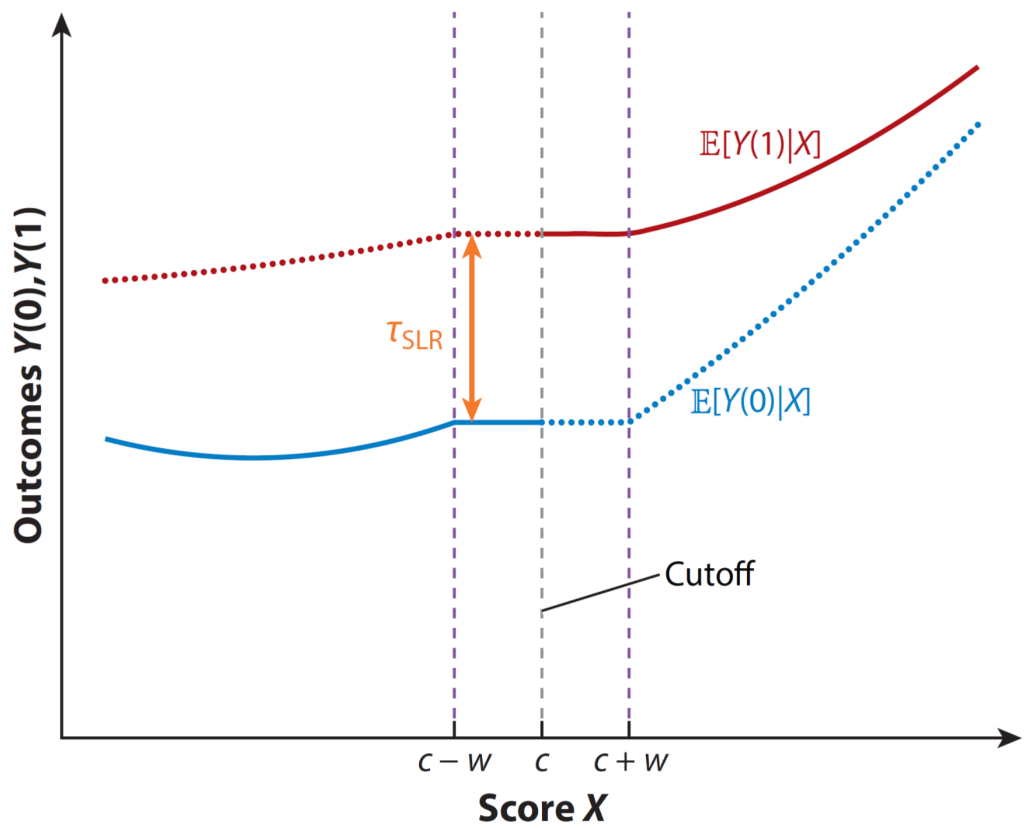


Source: McCray (2008)

Examining the unconfoundedness assumption

- Another condition for unconfoundedness is “no manipulation” – people should not be able to game the system.
- **To check this, we use a density test to see if there is a break in the distribution of the running variable.**

Examining the unconfoundedness assumption



- **Conditional unconfoundedness**: Treatment assignment D_i is unconfounded conditional on X_i.

$$Y_i^1, Y_i^0 \perp D_i \mid X_i$$

- **Implication**: in the close-neighborhood of the cutoff c, the assignment of treatment should be "as-if" random.

- The idea of "local experiments."

Examining the unconfoundedness assumption

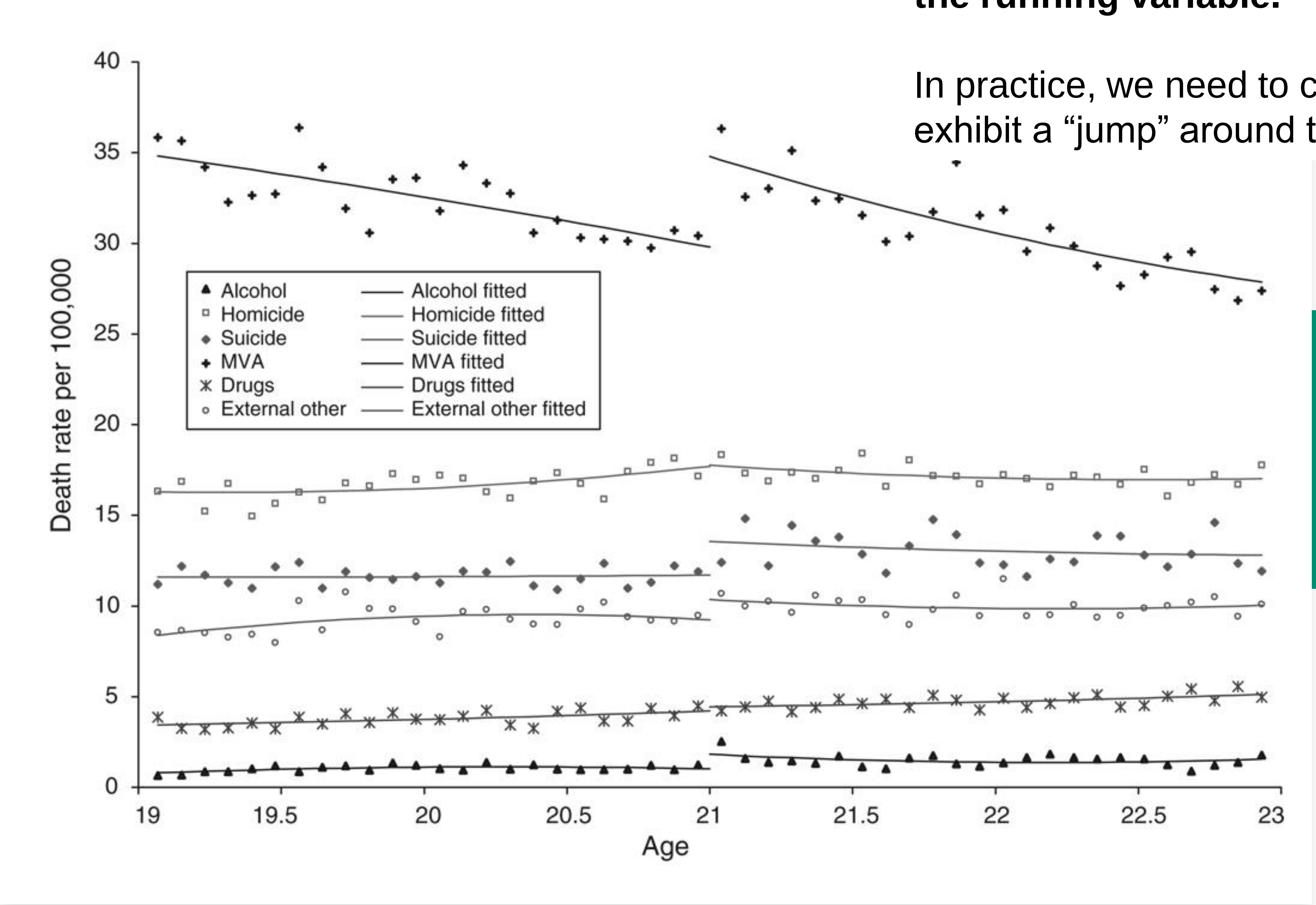
- If we have a “”local completely random experiment,” then characteristics of people should be balanced above and below the cutoff.
- **Testing for covariate balance within the cutoff.**

Variable	Std. Mean Difference	P-value
Age		
25 - 44	0.018	0.849
45 - 64	-0.018	0.849
65 ≥	-0.087	0.360
Education Level		
Primary	0.078	0.409
Secondary	0.058	0.542
Post-secondary	-0.151	0.111
Household Wealth Index		
Q1	-0.056	0.538
Q2	0.00	1.000
Q4	0.142	0.135
Q5	-0.154	0.104
Region		
Northern	-0.155	0.103
Middle	0.158	0.102
Other		
Household size	-0.007	0.942
Household size squared	-0.051	0.599
Overweight(25≤BMI≤29.9)	0.00	1.000
Renewed insurance at least once	0.111	0.241
Female	0.036	0.701
Married	0.051	0.593
Urban	0.032	0.737
At least one child in the household	0.00	1.000
N		450

** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

We assume that the treatment status **ONLY** depends on the running variable.

In practice, we need to check whether other variables also exhibit a "jump" around the cutoff of the running variable.



Examining the unconfoudedness assumption

Carpenter, C., & Dobkin, C. (2009). The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics*, *1*(1), 164-182.

Examining the continuity assumption

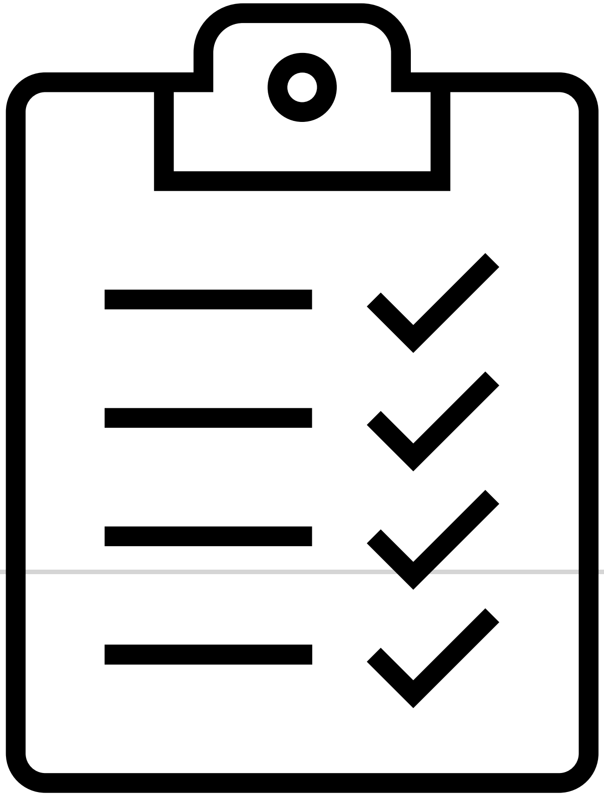
- **Continuity assumption:** The expectation function of potential outcomes conditional on the running variable is continuous **at the cutoff point**.
- We test if there's any discontinuity of the running variable above or below the cutoff.
- Run the RDD with the same model spec and bandwidth with pseudo-cutoffs.
- **If the continuity assumption holds, no effects from the pseudo-cutoffs.**

Interpreting the estimated effects from RDD

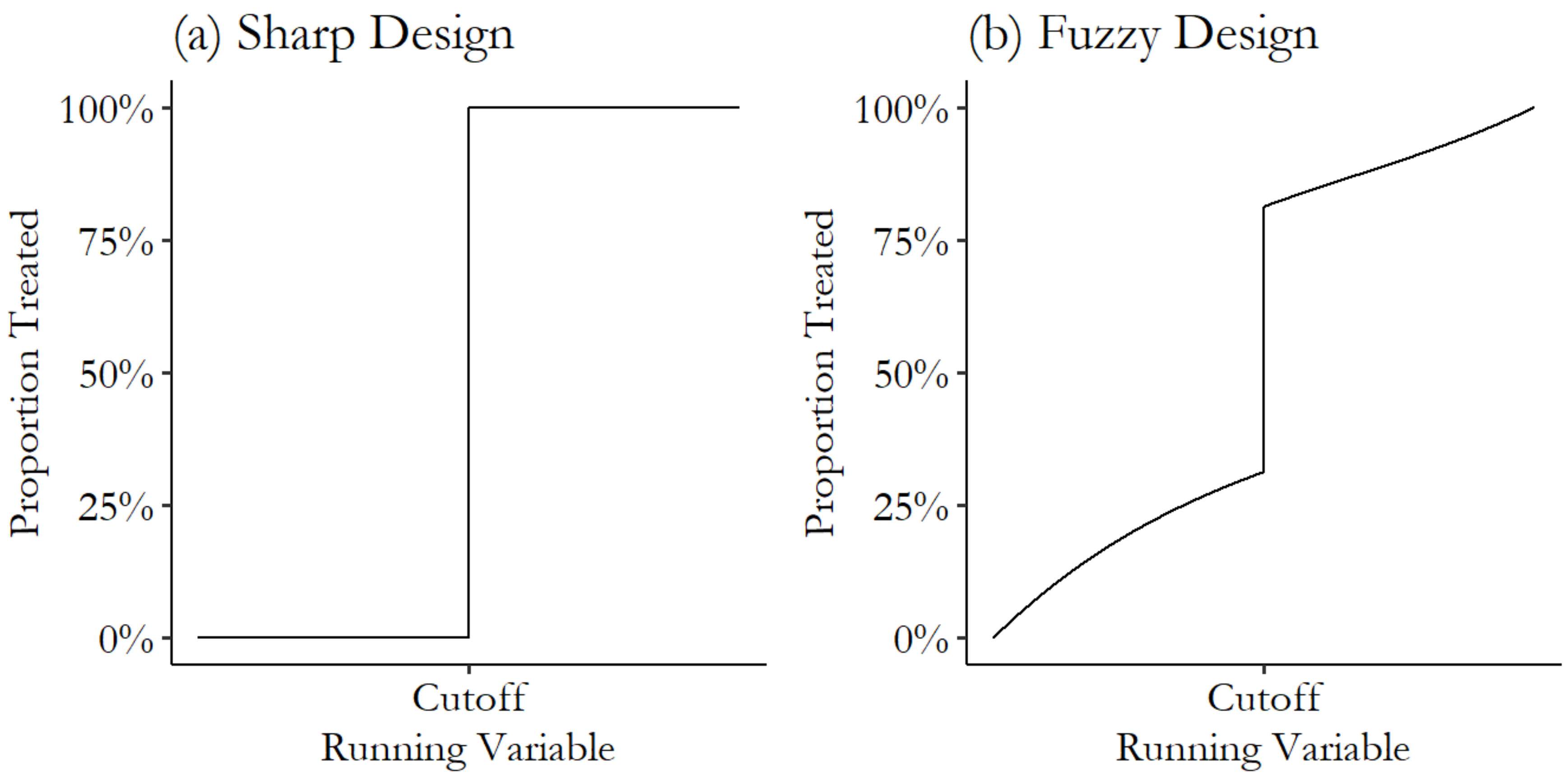
In RDD, treatment effects are local average treatment effects or LATE.

We don't estimate the effect of getting the treatment, but rather the effect of getting the treatment for units that were close to the cutoff, not everybody in the sample.

In a sense, this is the price we pay for being able to estimate treatment effects. However, in some applications, we might be interested in this group and not others



Other variants of RDD

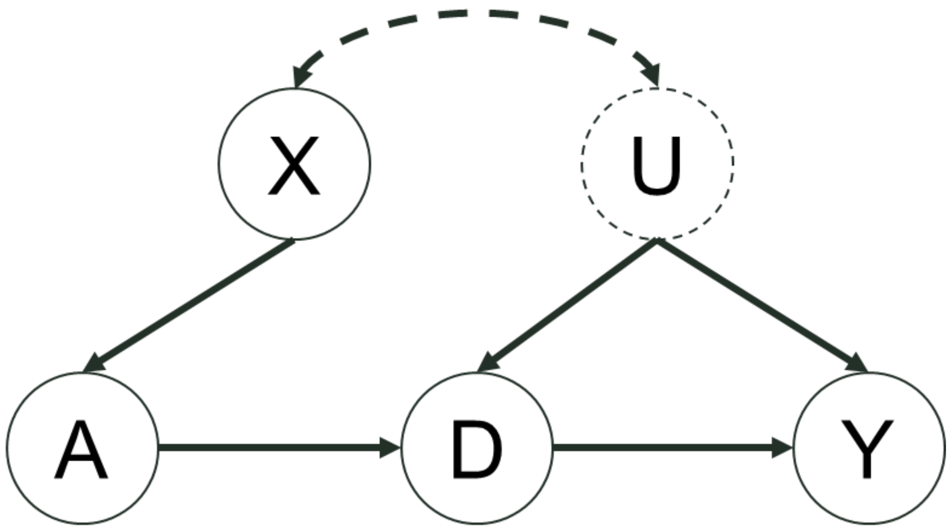


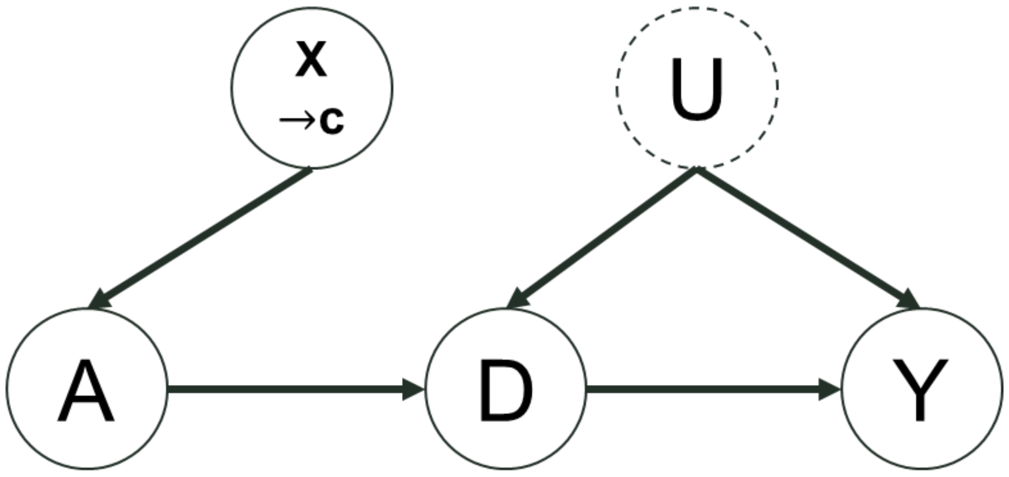
Fuzzy RDD

Comparing sharp RDD and fuzzy RDD

Fuzzy RDD

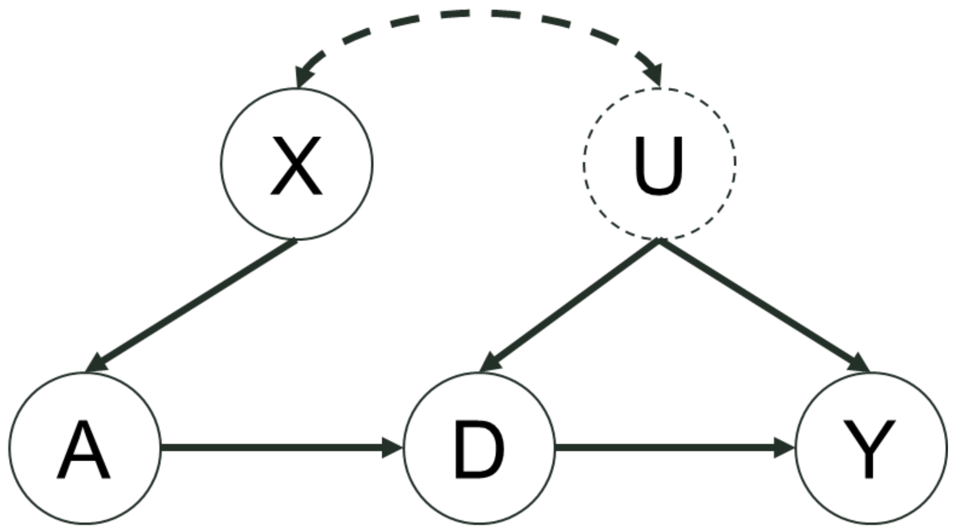
- It is essentially the sharp RDD with a non-compliance issue.
- The assignment A depends on the cutoff of the running variable.
- The actual treatment D depends on the assignment.

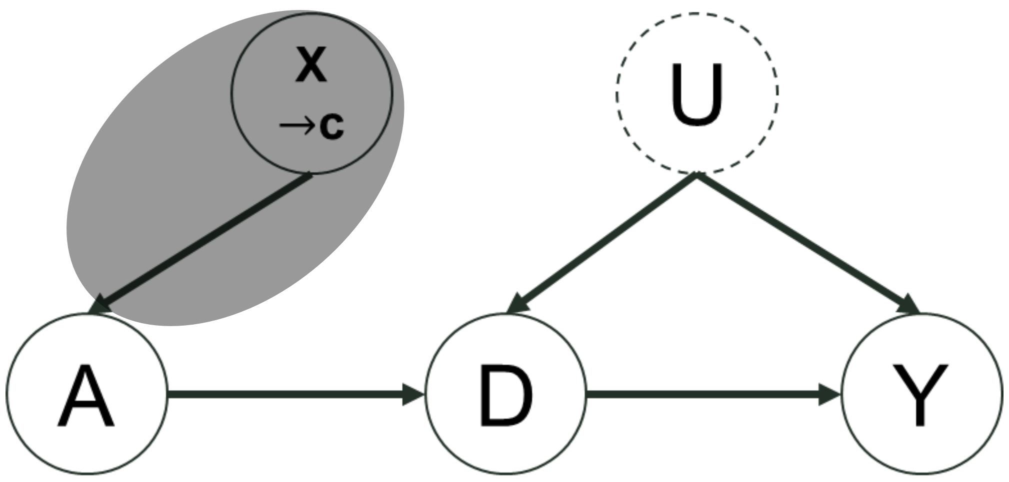




Fuzzy RDD

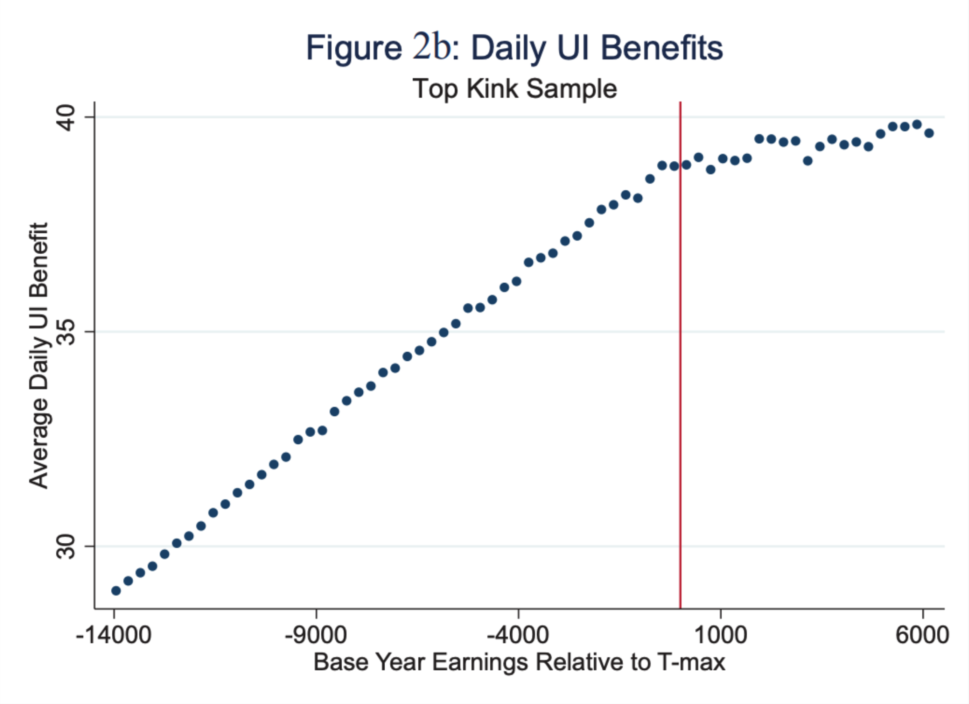
- It is essentially the sharp RDD with a non-compliance issue.
- In the limit $X \rightarrow c$, become just like in the non-compliance setting.
- **Assignment A becomes an instrument for the treatment D.**





Regression kink design (RKD)

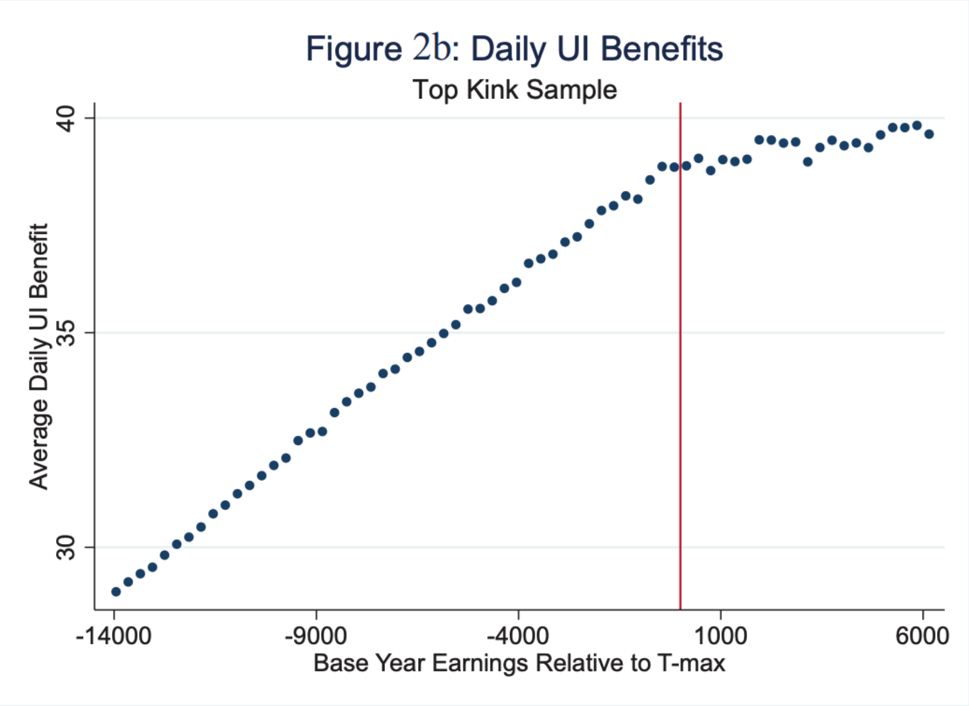
- Much of our discussion centered around a discontinuous jump in the outcome (and treatment variable).
- Many policy tools have shifts in the treatment intensity based on the running variables, rather than jumps.
- Example: income taxes, usage-based electricity prices etc.



Card, D., Lee, D. S., Pei, Z., & Weber, A. (2015). Inference on causal effects in a generalized regression kink design. Econometrica, 83(6), 2453-2483.

Regression kink design (RKD)

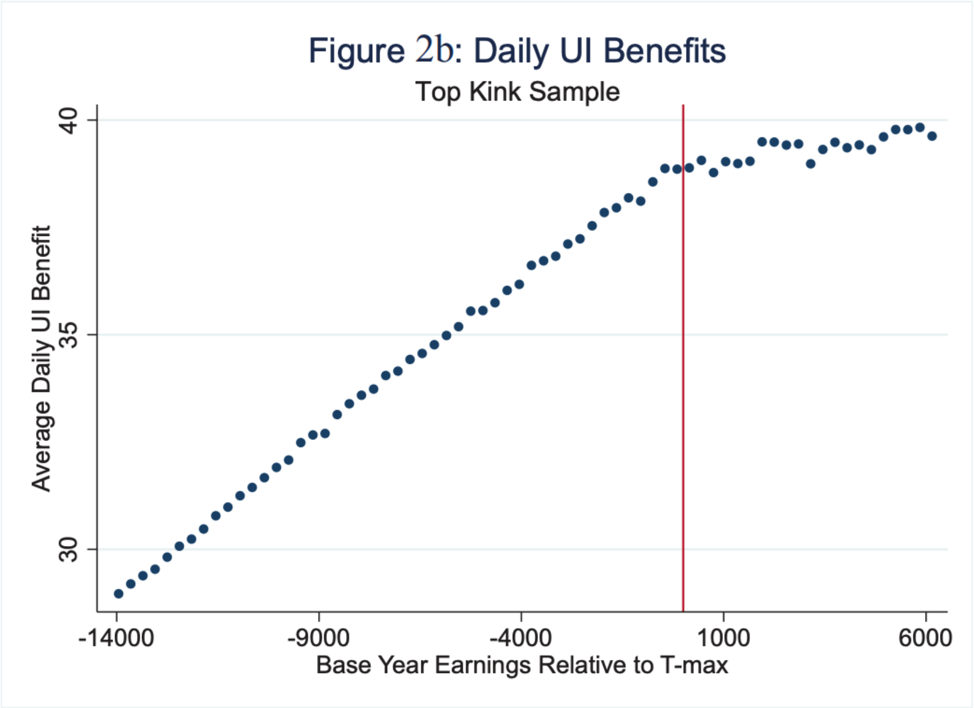
- RKD works with a continuous cause variable and outcome variable.
- The treatment effect is defined as the first derivative.
- We can use the kinks to identify the first derivative.



Card, D., Lee, D. S., Pei, Z., & Weber, A. (2015). Inference on causal effects in a generalized regression kink design. Econometrica, 83(6), 2453-2483.

Regression kink design (RKD)

- The effect of tax payment (cause) on consumption (outcome).
- Tax payment has kinks that are created by the income tax policy (i.e., different tax rates at different income levels).
- At the critical income level, we can identify the change in tax payment on the change on consumption.



Card, D., Lee, D. S., Pei, Z., & Weber, A. (2015). Inference on causal effects in a generalized regression kink design. Econometrica, 83(6), 2453-2483.

Slides for the R notebook

How to select bandwidth? A cross-validation procedure

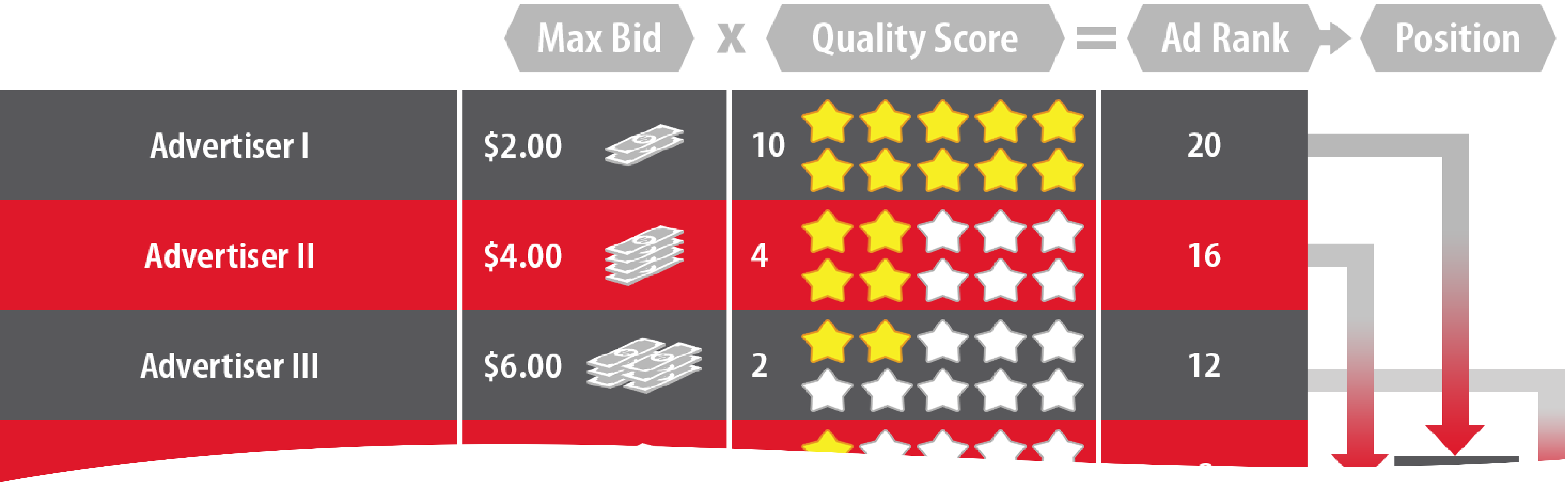
Pseudo code

(Note: assume we do a simple 2-fold cross-validation; with more folds, just loop over all K subsamples and take the average performance across these subsamples.)

Given a bandwidth h, the training set $Train$ and the test set $Test$

1. Trim the training set by retaining observations with $|X_i - x^*| \leq h$
2. Run a regression on the trimmed training set with $Y_i = \alpha + \beta D_i + \gamma X_i + \theta D_i X_i + e_i$.
3. Trim the test set by retaining observations with $|X_i - x^*| \leq h$
4. Predict the outcome $\hat{Y}_i^{Test}$ for the trimmed test set with the estimated regression in Step 2.
5. Calculate the mean squared errors for the trimmed test set with $MSE(h) = mean\left[\left(Y_i^{Test} - \hat{Y}_i^{Test}\right)^2\right]$

Do it for all the candidate bandwidth and choose the one with the smallest MSE.



How to apply the RDD design to keywords auction?

- Objective: to measure the value of positions.
- Observations:
  - The position is determined by Ad Rank.
  - For two positions j and $j+1$, we must have $AdRank_j - AdRank_{j+1} > 0$

How to apply the RDD design to keywords auction?

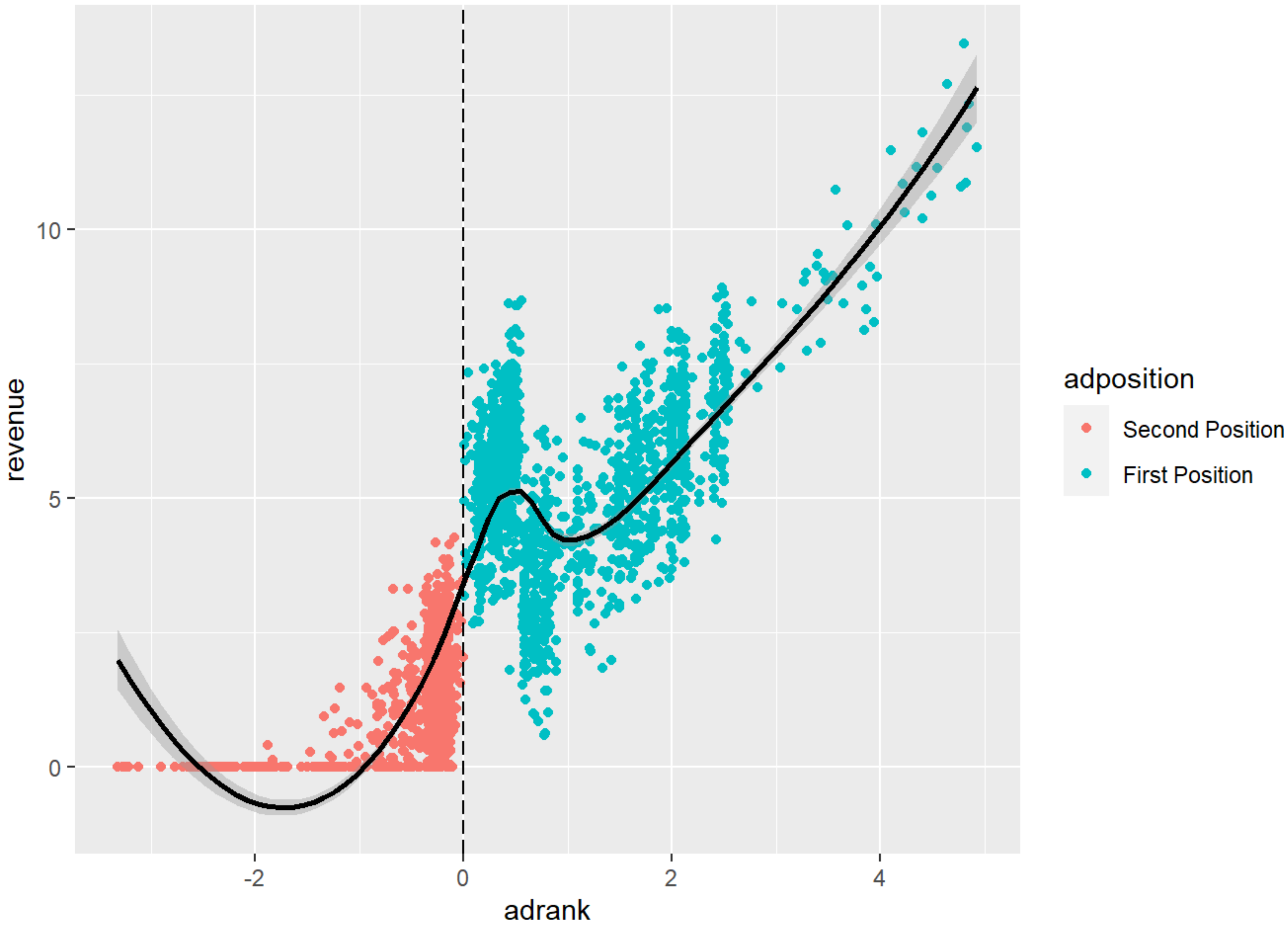
Objective:
to measure the value of positions.

- The running variable:
  - If an advertiser is in position j, $AdRank_i - AdRank_{j+1} > 0$
  - If an advertiser is in position $j+1$, $AdRank_i - AdRank_j < 0$
  - So, we use $\Delta AdRank$ as the running variable with the cutoff point as 0.
- The treatment:
  - The ad of an advertiser is moved up from position $j+1$ to j.
- The outcome:
  - E.g., revenue generated or other conversions

Narayanan, S., & Kalyanam, K. (2014). *Position effects in search advertising: A regression discontinuity approach*. Technical Report (Stanford University, Stanford, CA).

Data visualization

- Running variable:
- AdRank
- Treatment variable:
- Frist vs. Second Position
- Outcome variable:
- Revenue



Pseudo code

Pseudo code
Note:

- We will do a simple 2-fold cross-validation by splitting to data in one test set and one train set with equal sizes;
- With more folds, just loop over all K subsamples and take the average performance across these subsamples.

Step 1: Calculate the standard deviation of AdRank as SD_{AdRank} and set 4 candidate bandwidths with $\{0.25, 0.05, 1.00, 2.00\}$ times SD_{AdRank} .

Step 2: Randomly split the data into a train set and a test set for the cross validation. You may use $sample(\ldots, \ldots, replace = F)$

Step 3: Given a bandwidth h, the training set $Train$ and the test set $Test$, do the following:

(1) Trim the training set by retaining observations with $|AdRank_i| \leq h$
(2) Run a regression on the trimmed training set with $Revenue_i = \alpha + \beta FirstPosition_i + \gamma AdRank_i + \theta FirstPosition_i AdRank_i + e_i$.
(3) Trim the test set by retaining observations with $|AdRank_i| \leq h$
(4) Predict the outcome $\hat{Y}_i^{Test}$ for the trimmed test set with the estimated regression in (2). You may use $predict(\ldots, newdata = \cdots)$
(5) Calculate the mean squared errors (MSE) for the trimmed test set with $MSE(h) = mean\left[\left(Y_i^{Test} - \hat{Y}_i^{Test}\right)^2\right]$

Step 4: Repeat Step 3 for all 4 candidate bandwidths $\{0.25, 0.05, 1.00, 2.00\} \times SD_{AdRank}$ and select the bandwidth with the smallest MSE as $h^{optimal}$.

Step 5: With the bandwidth selected in Step 4, trim the full data (all observations) by retaining only those observations with $|AdRank_i| \leq h^{optimal}$.

Step 6: After trimming the full data, rerun the main regression on the data to obtain the effects of the ad in the first position over the second position, with $Revenue_i = \alpha + \beta FirstPosition_i + \gamma AdRank_i + \theta FirstPosition_i AdRank_i + e_i$.

References

- Bird, A., & Karolyi, S. A. (2017). Governance and taxes: evidence from regression discontinuity (retracted). *The Accounting Review*, *92*(1), 29-50.
- Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, *51*(6), 309.
- Eggers, A. C., Fowler, A., Hainmueller, J., Hall, A. B., & Snyder Jr, J. M. (2015). On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. American Journal of Political Science, 59(1), 259-274.
- Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*(3), 447-456.
- McCrary, Justin. (2008). Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. *Journal of Econometrics* 142 (2), 698–714.