TA Session 6: Treatment Effects

Microeconometrics with Hanna Wang
IDEA, Fall 2022

TA: Conghan Zheng

Overview

1. Introduction

2. Exogenous treatment effects
   - Regression adjustment
   - Matching

3. Endogenous treatment effects
   - Regression Discountinuity
   - Differences-in-Differences

4. Appendix

Causality

- Causal relationships in terms of the potential outcomes notation: what would happen to a given individual in a hypothetical scenario (potential outcomes in the parallell worlds).

$$\begin{array}{lll} & y_1(T+C) & y_0(T+C) \\ T: & y_1(T), & \hat{y}_0(T) \\ C: & \hat{y}_1(C), & y_0(C) \end{array}$$

  - $y_1(T+C)$: the outcome of an individual had he received the treatment, irrespective of whether he actually received.
    - $y_1(T)$: observed outcome for the treated individuals (group T)
    - $\hat{y}_1(C)$: the outcome that would happen if the non-treated individuals are treated
  - $y_0(T+C)$: the outcome of an individual had he not received the treatment, irrespective of whether he actually received.
    - $y_0(C)$: observed outcome for the non-treated individuals (group C)
    - $\hat{y}_0(T)$: the outcome that would happen if the treated individuals are not treated
  - *Average causal effect of the treatment on those who were treated*:

$$y_1(T) - \hat{y}_0(T)$$

    Differences in average potential outcomes for a fixed reference population.

Independence

$$\underbrace{\mathbb{E}(y_i|D_i=1)-\mathbb{E}(y_i|D_i=0)}_{\text{Observed: } y_1(T)-y_0(C)} = \underbrace{\mathbb{E}(y_{1i}|D_i=1)-\mathbb{E}(y_{0i}|D_i=1)}_{\text{ATT: } y_1(T)-\hat{y}_0(T)} + \underbrace{\mathbb{E}(y_{0i}|D_i=1)-\mathbb{E}(y_{0i}|D_i=0)}_{\text{Selection bias: } \hat{y}_0(T)-y_0(C)}$$

- Selection bias: the sick (low y) are more likely than the healthy to seek treatment ($D=1$).
- Random treatment assignment $(y_0, y_1) \perp D$ solves the selection problem.
  - $(y_0, y_1) \perp D$ implies the *mean independence*: $\mathbb{E}(y_0|D)=y_0$, $\mathbb{E}(y_1|D)=y_1$, which further implies

$$\underbrace{\mathbb{E}(y_{1i}|D_i=1)-\mathbb{E}(y_{0i}|D_i=1)}_{ATT} = \underbrace{\mathbb{E}(y_{1i})-\mathbb{E}(y_{0i})}_{ATE}$$

The effect of randomly assigned treatment on the treated is the same as the effect of the treatment on a randomly chosen individual.

Conditional independence

- The conditional independence assumption: $(y_{1i}, y_{0i}) \perp D_i | X_i$.
- D and (y_0, y_1) are allowed to be correlated, $(y_{1i}, y_{0i}) \perp D_i | X_i$ implies the *conditional mean independence*:

$$\mathbb{E}(y_0|D_i, X_i) = \mathbb{E}(y_0|X_i), \;\; \mathbb{E}(y_1|D_i, X_i) = \mathbb{E}(y_1|X_i)$$

which further implies

$$\underbrace{\mathbb{E}(y_{1i}|D_i = 1, X_i) - \mathbb{E}(y_{0i}|D_i = 1, X_i)}_{ATT} = \underbrace{\mathbb{E}(y_{1i}|X_i) - \mathbb{E}(y_{0i}|X_i)}_{ATE}$$

- Notice that the expectation is taken over the distribution of X, estmating ATE will require being able to observe both control and treated units for every outcome on X. $\rightarrow$ The overlap assumption in matching.
- Control for covariates can increase the likelihood that regression estimates have a causal interpretation.
- The conditional independence assumption is fundamentally untestable because we only observe (y, D, X)

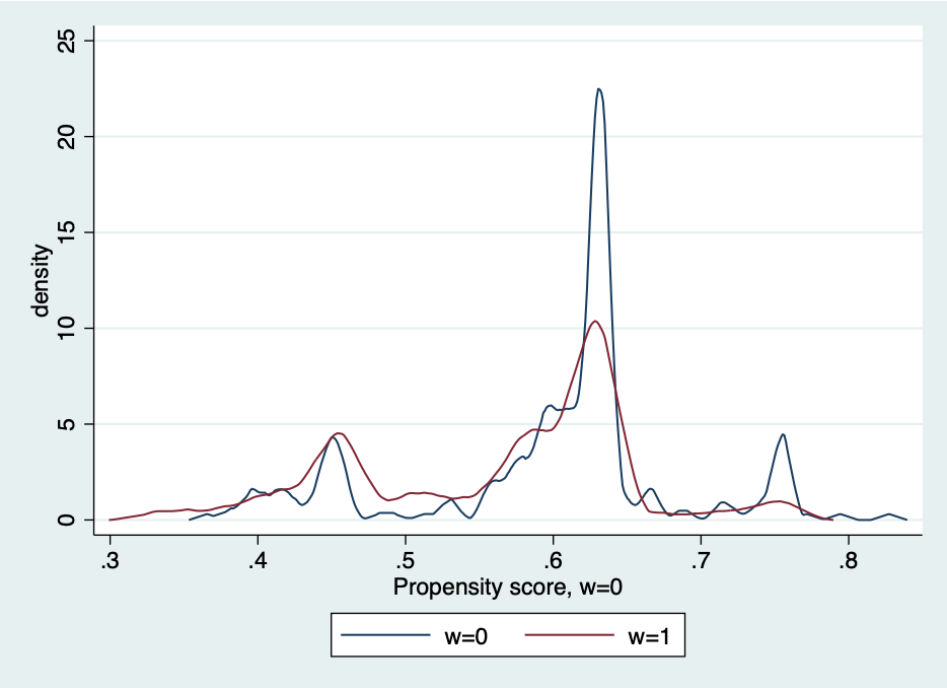
Introduction

- **Exogenous treatment**: the selection into treatment is assumed to be exogenous. The causal interpretation in this case is based on the conditional independece assumption: conditional on enough controls, any selection into treatment is uncorrelated with the potential outcomes.
- **Endogenous treatment**: the selection into treatment is endogenous. The selection is not only on observables.

Data

`TA5_1.dta`: a subset of Lalonde (1986) data.

- Source: the National Supported Work (NSW)
- Goal: effects of on-the-job training on labor market outcomes
- Treatment: on-the-job training ($w = 1$) lasting between 9 months and one year (1976-1977)
- Sample size: treatment - 185, controls - 260



Regression adjustment

Regression Adjustment

- Assumption: conditional independence $(y_{1i}, y_{0i}) \perp D_i | X_i$
- *Regression adjustment* (RA): two separate OLS of y on X

$$\mathbb{E}(y_{1i}) = X_i \beta_1 \text{ if } D_i = 1$$
$$\mathbb{E}(y_{0i}) = X_i \beta_0 \text{ if } D_i = 0$$

- Predictions: potential outcome means $\hat{y}_{1i}, \hat{y}_{0i}$

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_{1i} - \frac{1}{N} \sum_{i=1}^{N} \hat{y}_{0i}$$
$$\widehat{\text{ATT}} = \frac{1}{N_1} \sum_{i:D_i=1} \hat{y}_{1i} - \frac{1}{N_1} \sum_{i:D_i=1} \hat{y}_{0i}$$

Regression Adjustment

- Homogenous response to treatment: α, $\forall i$
- Heterogenous response to treatment: α_i for i

	(1) Hom:w/o X	(2) Hom:w X	(3) Het:ATE	(4) Het:ATT
main				
d	1.794**	1.625*		
	(0.633)	(0.640)		
r1vs0.d			1.545*	1.764**
			(0.662)	(0.672)
N	445	445	445	445

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Effect size

- Effect size as a percentage of the untreated potential outcome mean: the training program (d) increases earnings (y) by 33.8%.

y	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	.3382046	.1589424	2.13	0.033	.0266832	.649726

- Shortcoming of *Regression adjustment*: Covariates are used to ensure the treatment assignment is random, the choice of the set of controls and the functional form is crucial. RA estimator is not robust in most observational data settings.

Matching

Matching

- Regressions and matching are both control strategies, the core assmption (conditional independece) underlying causal inference is the same for the two strategies.
- Difference: whether the counterfactual is identified or not.
  - Matching amounts to covariate-specific treatment-control comparisons, **weighted** together to produce a single overall average treatment effect.
  - Rregression adjustment is not robust to substantially different control and treatment groups (e.g., C: millionaires, T: workers, D: on-the-job training).
  - Matching is by design robust to outliers, whose influence is down-weighted in matching.
- Matching formula:

$$\hat{\alpha}^{M}_{ATE} = \frac{1}{N}\left\{\sum_{i=1}^{N} D_i[y_i - \hat{y}_0(X_i)] + (1 - D_i)[\hat{y}_1(X_i) - y_i]\right\}$$

- $\hat{y}_0(X_i)|D_i = 1$: matched counterparts in the control group for the treated individual
- $\hat{y}_1(X_i)|D_i = 0$: matched counterparts in the treatment group for the non-treated individual

Identifying Assumptions

1. **Overlap**:

$$0 < \mathbb{P}(D = 1|X) \equiv \pi(X) < 1$$

For each value of X there are both treated ($\mathbb{P} > 0$) and nontreated ($\mathbb{P} < 1$) cases.

2. **Conditional independence**:

$$(y_0, y_1) \perp D|X \Leftrightarrow (y_0, y_1) \perp D|\pi(X)$$

Participation in the treatment program does not depend on outcomes, after controlling for the variation in outcomes induced by differences in X.

3. **Regressor balance** (a testable hypothesis):

$$D \perp X|\text{Matching}$$

A well-specified matching model should balance the covariates, that two groups should look identical in terms of their X vector

Propensity score

- Propensity score: $\pi(X) \equiv \mathbb{E}(D|X) = \mathbb{P}(D = 1|X)$

The Propensity Score Theorem (Rosenbaum and Rubin, 1983)

Suppose the conditional independence assumption holds such that $(y_0, y_1) \perp D|X$. Then $(y_0, y_1) \perp D|\pi(X)$.

- For proof, it's enough to show that $P[D_i = 1|y_{ji}, \pi(X_i)] = \pi(X_i)$ does not depend on $y_{ji}, j = 0, 1$.
- $\pi(X)$ is a sufficient statistic for the distribution of D by construction.
- The propensity score theorem says that you need only control for covariates that affect the probability of treatment. The only covariate you really need to control for is the probability of treatment itself.
- *Propensity score matching*: first, $\pi(X_i)$ is estimated using some kind of parametric model, say, logit or probit. Then estimates of the effect of treatment are computed by matching based on some algorithm.

Measures and Algorithms

- Matching measures:
  1. regressors;
  2. propensity score.
- Various matching algorithms can be used to find potential matches based on different measures:
  1. **Exact matching**: practicable when the vector of covariates is discrete and the sample contains many observations at each distinct value of X_i.
  2. **Nearest-neighbor matching method**: taking each treated unit and searching for the control unit with the closest propensity score. In the Nearest-Neighbor method, all treated units find a match, but could be a fairly poor match, Kernel matching provides a solution to this problem.
  3. **Kernel matching method**: all treated are matched with a weighted average of all controls (with weights that are inversely proportional to the distance between the propensity scores of treated and controls).

k Nearest-neighbor matching

- Each observation is matched with k observations (k closest individuals to i in propensity score) from the other treatment level. The k matches for individual i are denoted by set $\mathcal{J}_i$.
- The matching estimator imputes the potential outcomes as

$$\hat{y}_i(0) = \begin{cases} y_i, & \text{if } D_i = 0 \\ \frac{1}{k}\sum_{j\in\mathcal{J}_j} y_j, & \text{if } D_i = 1 \end{cases}$$

and

$$\hat{y}_i(1) = \begin{cases} \frac{1}{k}\sum_{j\in\mathcal{J}_j} y_j, & \text{if } D_i = 0 \\ y_i, & \text{if } D_i = 1 \end{cases}$$

where $k \leq N_0$, $k \leq N_1$, N_0 and N_1 are the number of subjects in the control group and the treatment group, respectively.

- This leads to the matching estimator for the ATE and the ATT:

$$ATE_M = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i(1) - \hat{y}_i(0)\right), \quad ATT_M = \frac{1}{N_1}\sum_{i:D_i=1}\left(y_i - \hat{y}_i(0)\right)$$

Nearest-neighbor matching

- For any given subject i, there could be no counterfactual available for estimating the treatment effects. This could happen if there is insufficient overlap of the distribution of propensity between the treatment groups. In the interests of obtaining better balance, or sufficient overlap, such an observation may be dropped.
- After matching, the resulting trimmed (smaller) sample is expected to yield a less biased estimate of the treatment effects. Smaller bias is traded off against a wider confidence interval resulting from higher variance due to shrinkage in sample size.

Nearest-neighbor matching

- Matching is more likely to be poor if there are more than one continuous variable and cause a bias in the treatment effects estimators.
- *Bias adjustment*: to balance the remaining imbalance in X (or part of the regressors) after matching

```
. teffects nnmatch (y $x) (d), biasadj(age re74)
```

```
Treatment-effects estimation                    Number of obs      =        445
Estimator      : nearest-neighbor matching      Matches: requested =          1
Outcome model  : matching                                   min =          1
Distance metric: Mahalanobis                                max =         16
```

	y	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATE							
	d						
	(1 vs 0)	**1.517626**	**.6661188**	**2.28**	**0.023**	**.212057**	**2.823195**

Kernel matching

- Kernel matching uses weighted averages of all individuals in the control group to construct the counterfactual outcome. The weight of each control observation j in the counterfactual for observation i is

$$w(i,j) = \frac{k\left(\frac{\pi(X_j)-\pi(X_i)}{h}\right)}{\sum_{j\neq k:D_j=0} k\left(\frac{\pi(X_k)-\pi(X_i)}{h}\right)}$$

  where k is a kernel function which downweights distant observations and h is a bandwidth parameter. Increasing the bandwidth h will decrease the variance (a smoother density) but increase bias (underlying features could be also smoothed away).
- The matching estimator imputes the missing potential outcomes as

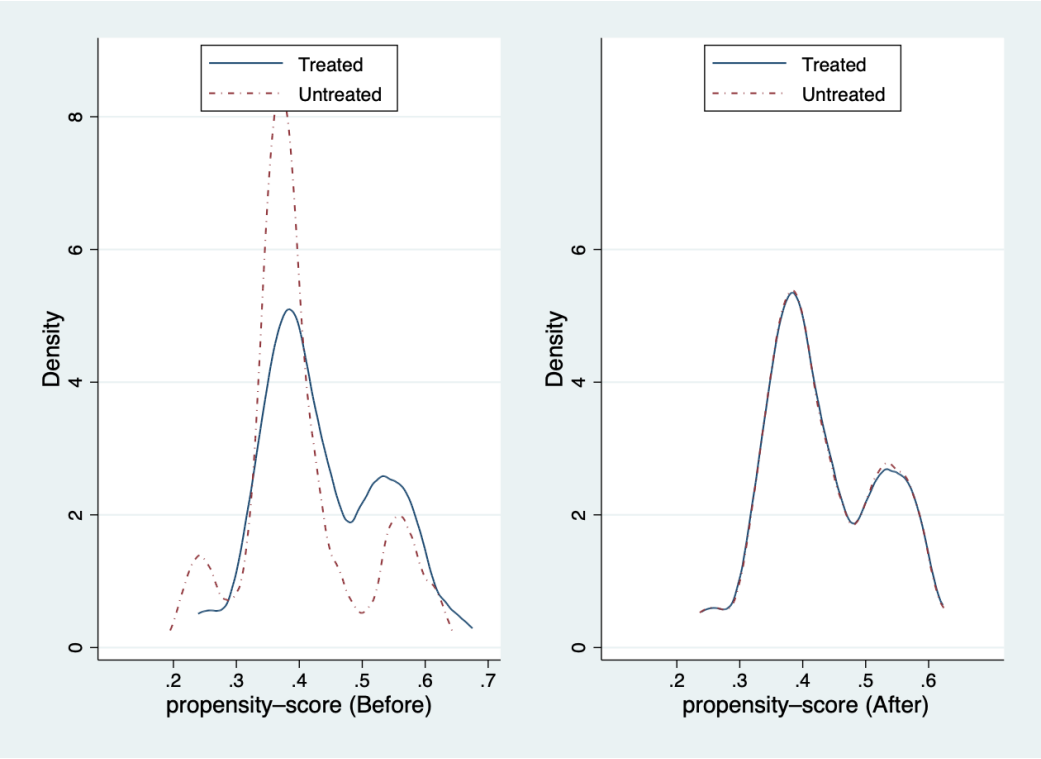
$$\hat{y}_i(0) = \begin{cases} y_i, & \text{if } D_i = 0 \\ \sum_{j:D_j=0} w(i,j)y_j, & \text{if } D_i = 1 \end{cases}$$

$$\hat{y}_i(1) = \begin{cases} \sum_{j:D_j=1} w(i,j)y_j, & \text{if } D_i = 0 \\ y_i, & \text{if } D_i = 1 \end{cases}$$

  Based on the potential outcomes, we can easily calculate the ATT and ATE.

Propensity score matching

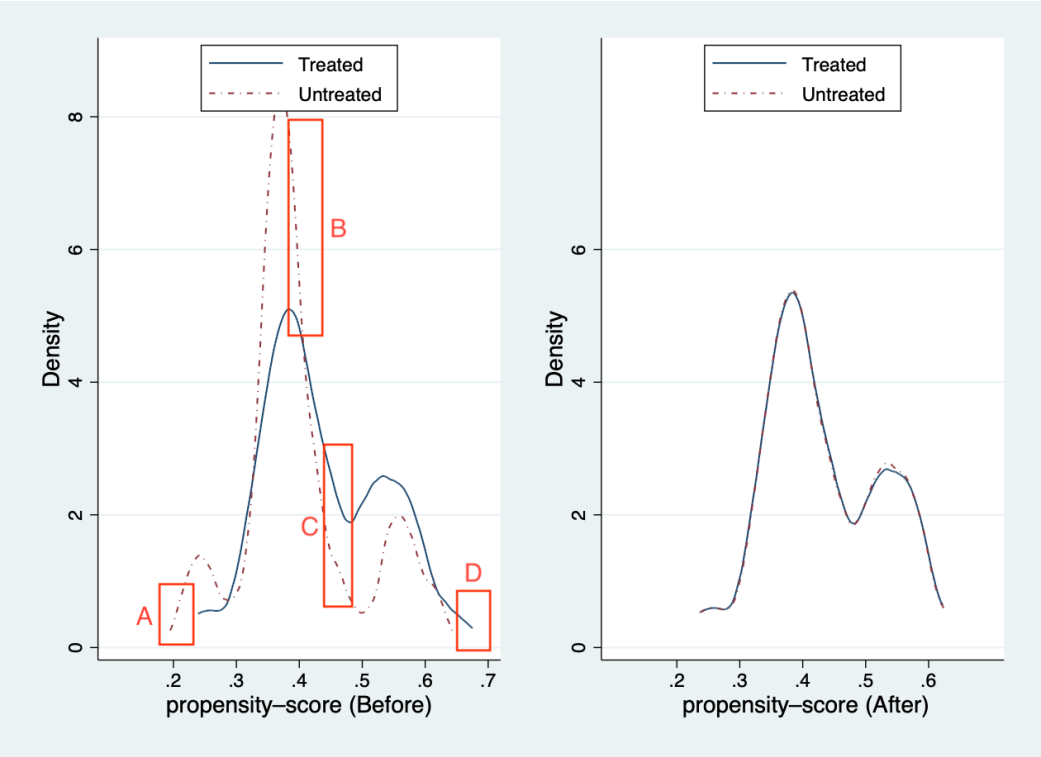
k Nearest neighbor matching



- Message from the left panel (before matching): the propensity score overlap is essentially in the range of 0.25 to 0.65. Without matching, analysis (or at least in a robustness test) might best be restricted to this range.
- Message from the right panel (after matching): a clear overlapping of the distributions is achieved by matching (notice the x-axises).

Propensity score matching

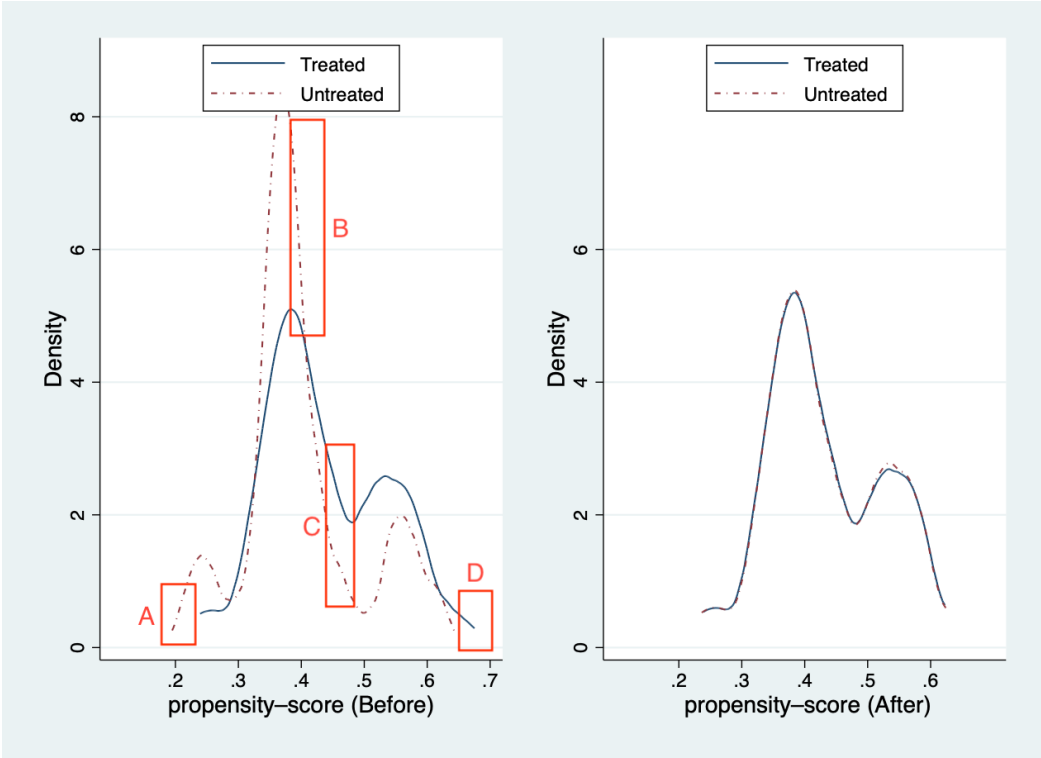
k Nearest neighbor matching



- Region A: Non-treated individuals in this region find no matches in the treated set, so they are missing from the distributions on the right panel.

Propensity score matching

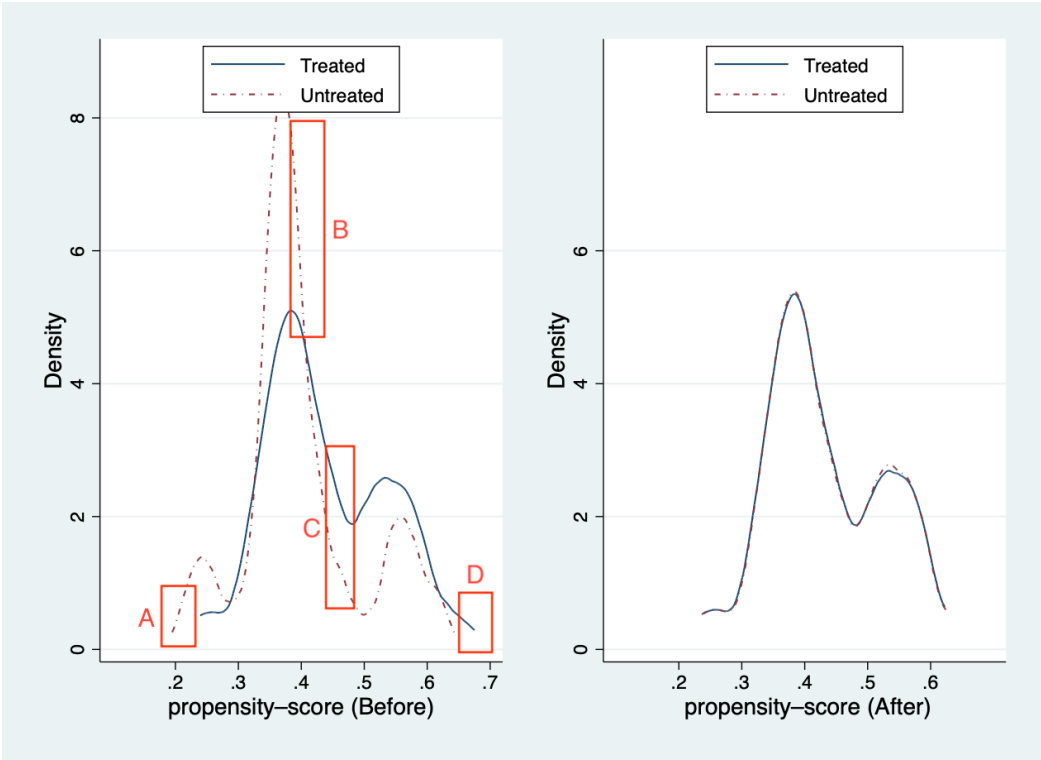
k Nearest neighbor matching



- Region B: There are more non-treated individuals than treated individuals in this region. If $k = 1$ and more than one counterparts are found in the control group, either one of them is picked, or the arithmetic mean is taken across the multiple counterparts (equivalent). The distribution for the control group of this region is then trimmed.

Propensity score matching

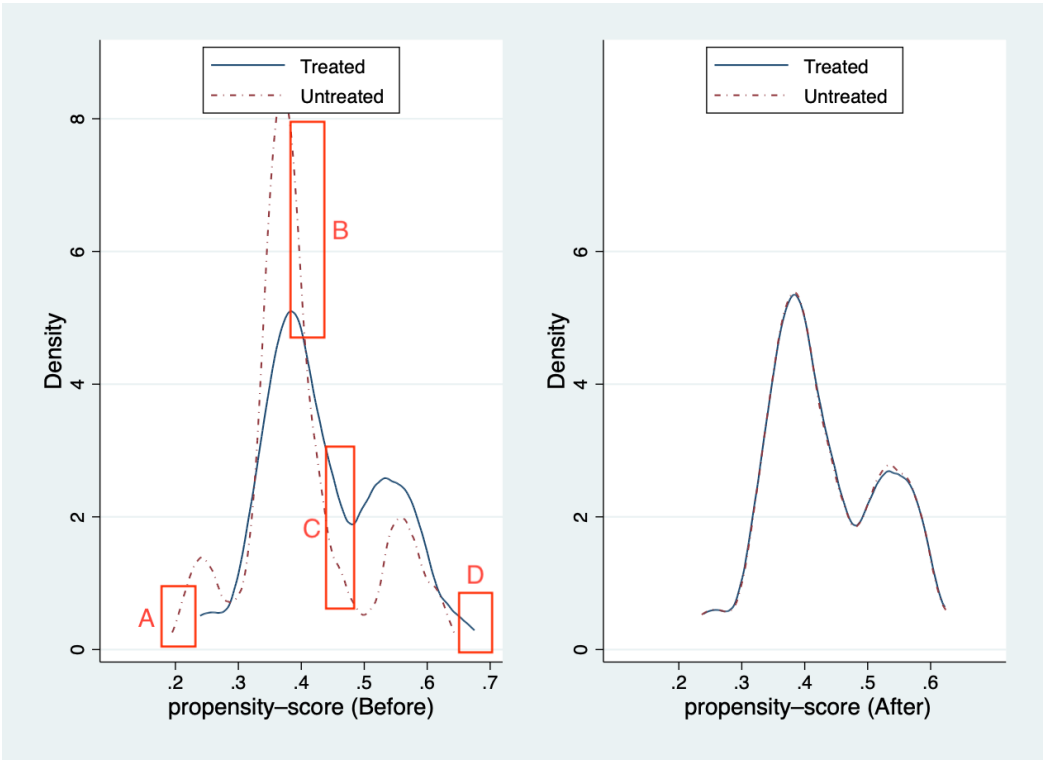
k Nearest neighbor matching



- Region C: There are more individuals in the treated group than in the control group. Therefore each control individual will be matched with more than one treated individuals, the control distribution is lifted up from the left panel to the right panel for this region.

Propensity score matching

k Nearest neighbor matching



- Region D: The treated individuals in this region find no counterpart in the control group, and are dropped in the left panel. There is no matching estimate for the treatment effects on this region.

Propensity score matching

Matching procedure	#Treated	#Non-treated	ATT	Std. Err.
Nearest neighbor	185	152	2.93	0.75
Kernel	185	248	1.92	0.65
Hetero RA	185	260	1.76	1.59

- Lower variance is achieved by kernel matching because more information is used (but there can be bad matches): NN matching is local matching, Kernel matching does global matching.

Regression Discountinuity

Regression Discountinuity

- Regression continuity (RD) research designs exploit precise knowledge of the rules determining treatment (some rules are arbitrary and therefore provide good experiments).
- RD comes in two styles: fuzzy and sharp.
  - The sharp design can be seen as a selection-on-observables story.
  - The fuzzy design leads to an instrumental variables (IV) type of setup.

Sharp RD

- Sharp RD is used when treatment status is a deterministic and discontinuous function of a covariate, x_i. Suppose, for example, that

$$D_i = \begin{cases} 1, & \text{if } x_i \geq x_0 \\ 0, & \text{if } x_i < x_0 \end{cases}$$

  where x_0 is a known threshold or cutoff.
- *Deterministic*: once we know x_i, we know D_i.
- *Discontinuous*: no matter how close x_i gets to x_0 (from the left), threatment is unchanged until $x_i = x_0$.
- Example: a standardized test for college entrance, sharp RD compares the post college performance of students with scores just above and just below the threshold.

Sharp RD

- Consider the following regression that formalizes the RD idea,

$$y_i = f(x_i) + \rho D_i(x_i) + \eta_i$$

  where ρ is the causal effect of interest, and $D_i(x_i) = \mathbb{1}(x_i \geq x_0)$.
- As long as $f(x_i)$ is continuous in a neighborhood of x_0, it should be possible to estimate this model.

Sharp RD

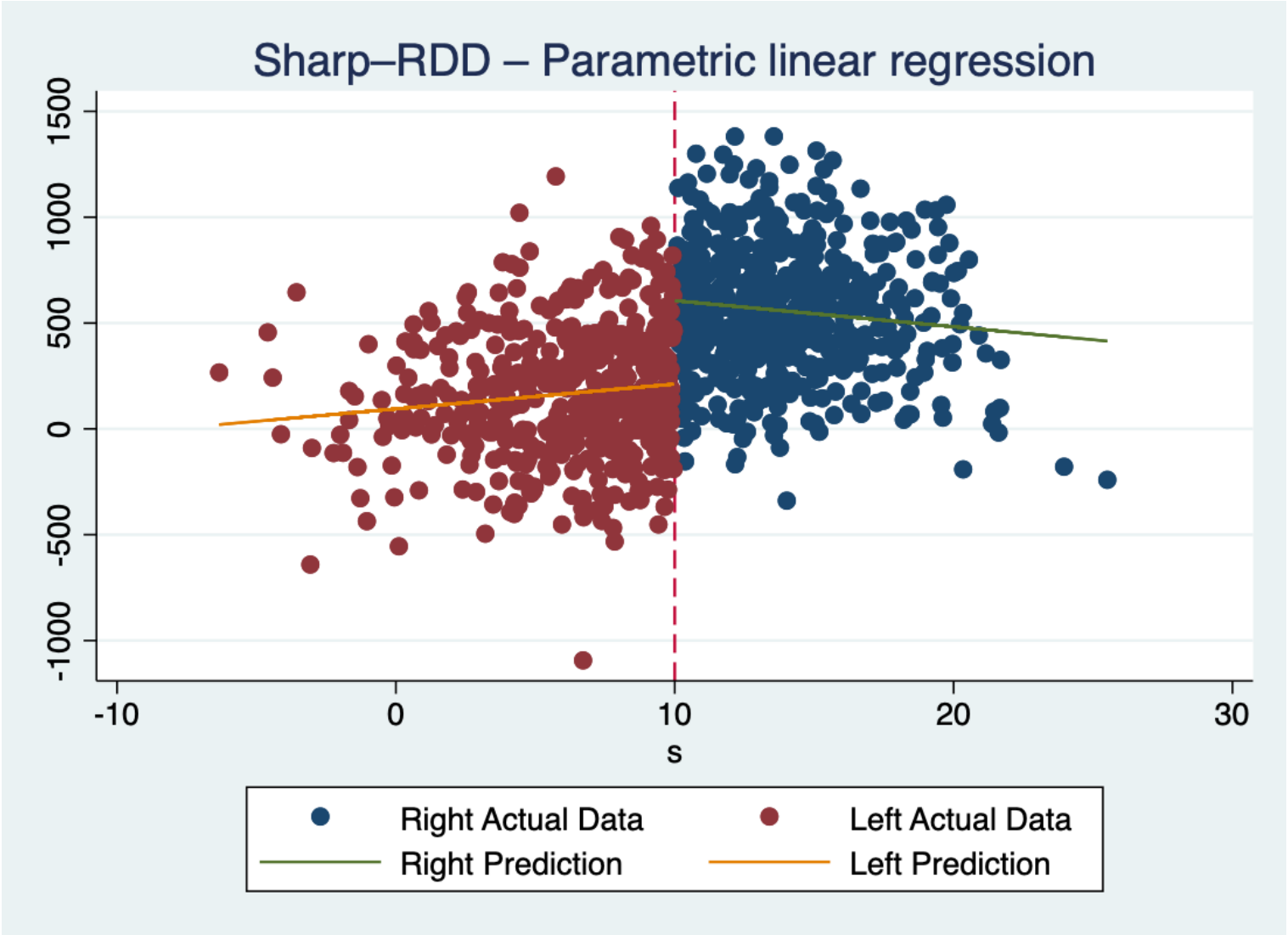
- For example, consider modeling $f(x_i)$ with a pth-order polynomial,

$$y_i = \underbrace{\alpha + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p}_{=f(x_i)} + \rho D_i + \eta_i$$

- A generalization of RD model allows different trend functions $f_0(x_i)$ for $\mathbb{E}(y_{0i}|x_i)$ and $f_1(x_i)$ for $\mathbb{E}(y_{1i}|x_i)$.
- Calculate the treatment effect:

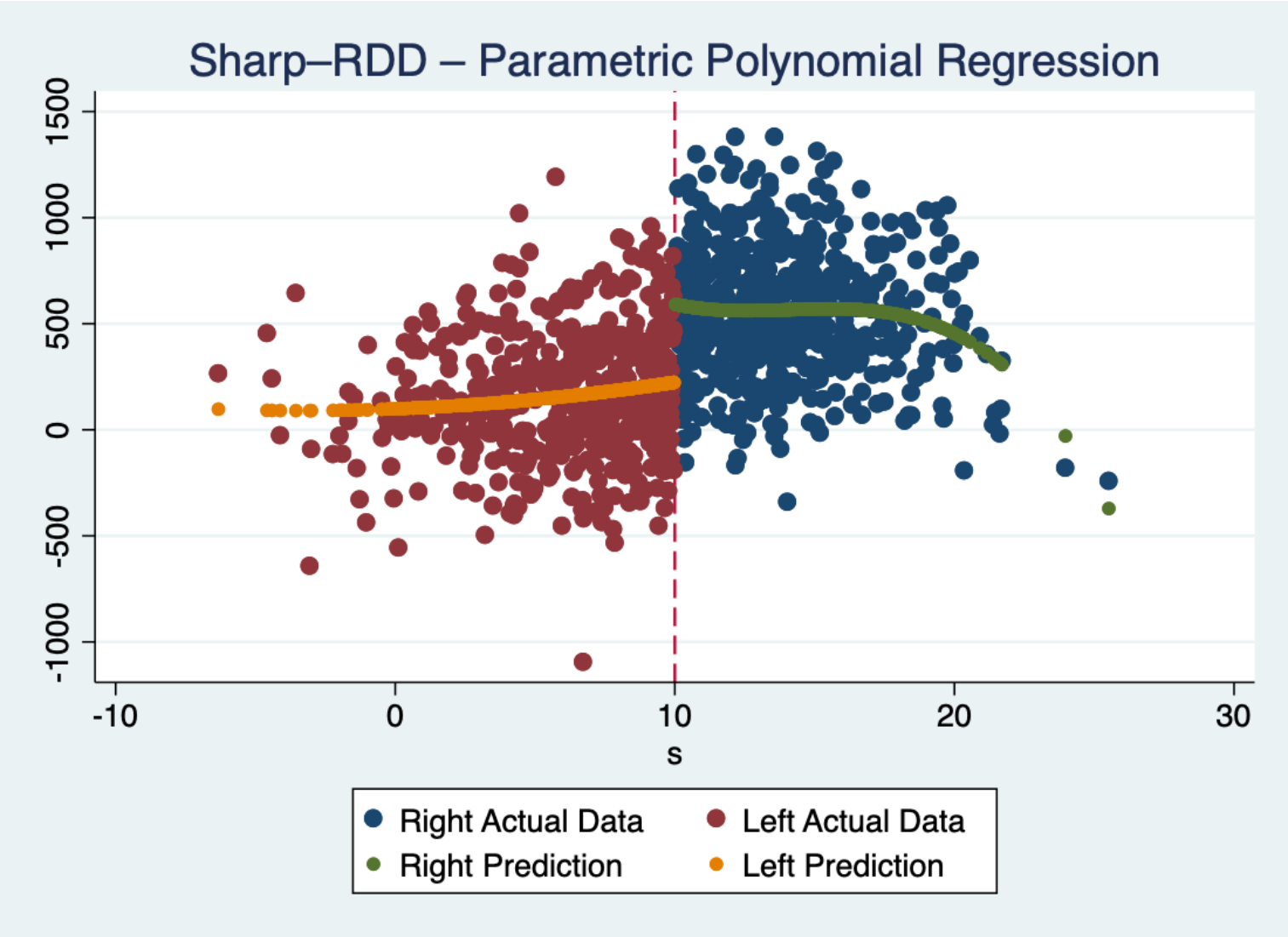
$$\alpha_{ATE,RD} = \underbrace{\lim_{x \to x_0^+} \mathbb{E}[y_i|x_i = x]}_{\simeq \mathbb{E}[y_{1i}|x_i]} - \underbrace{\lim_{x \to x_0^-} \mathbb{E}[y_i|x_i = x]}_{\simeq \mathbb{E}[y_{0i}|x_i]}$$

Sharp RD



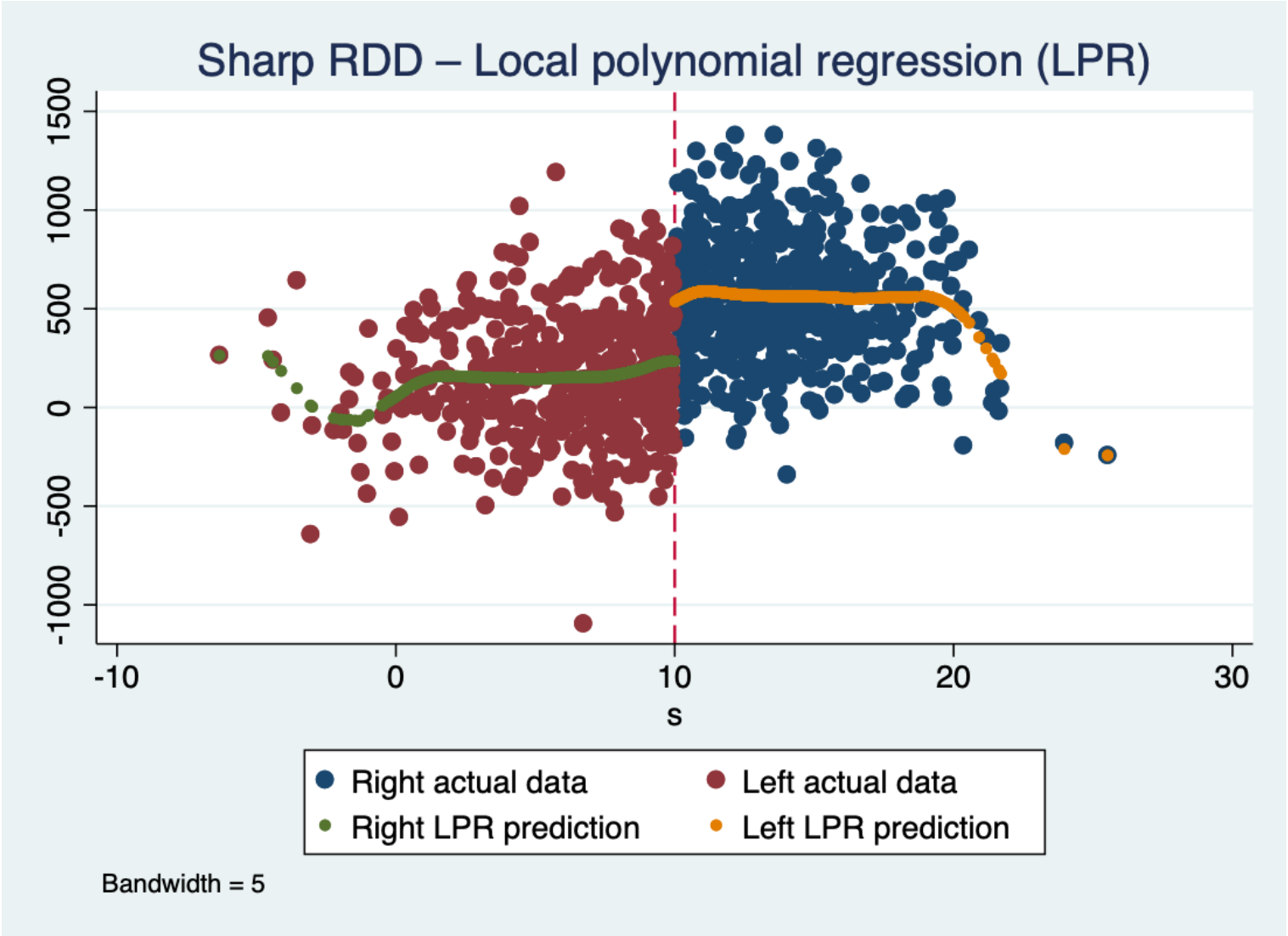
Linear $\mathbb{E}(y_{0i}|X_i)$

Sharp RD



Nonlinear $\mathbb{E}(y_{0i}|X_i)$: third-degree polynomials

Sharp RD

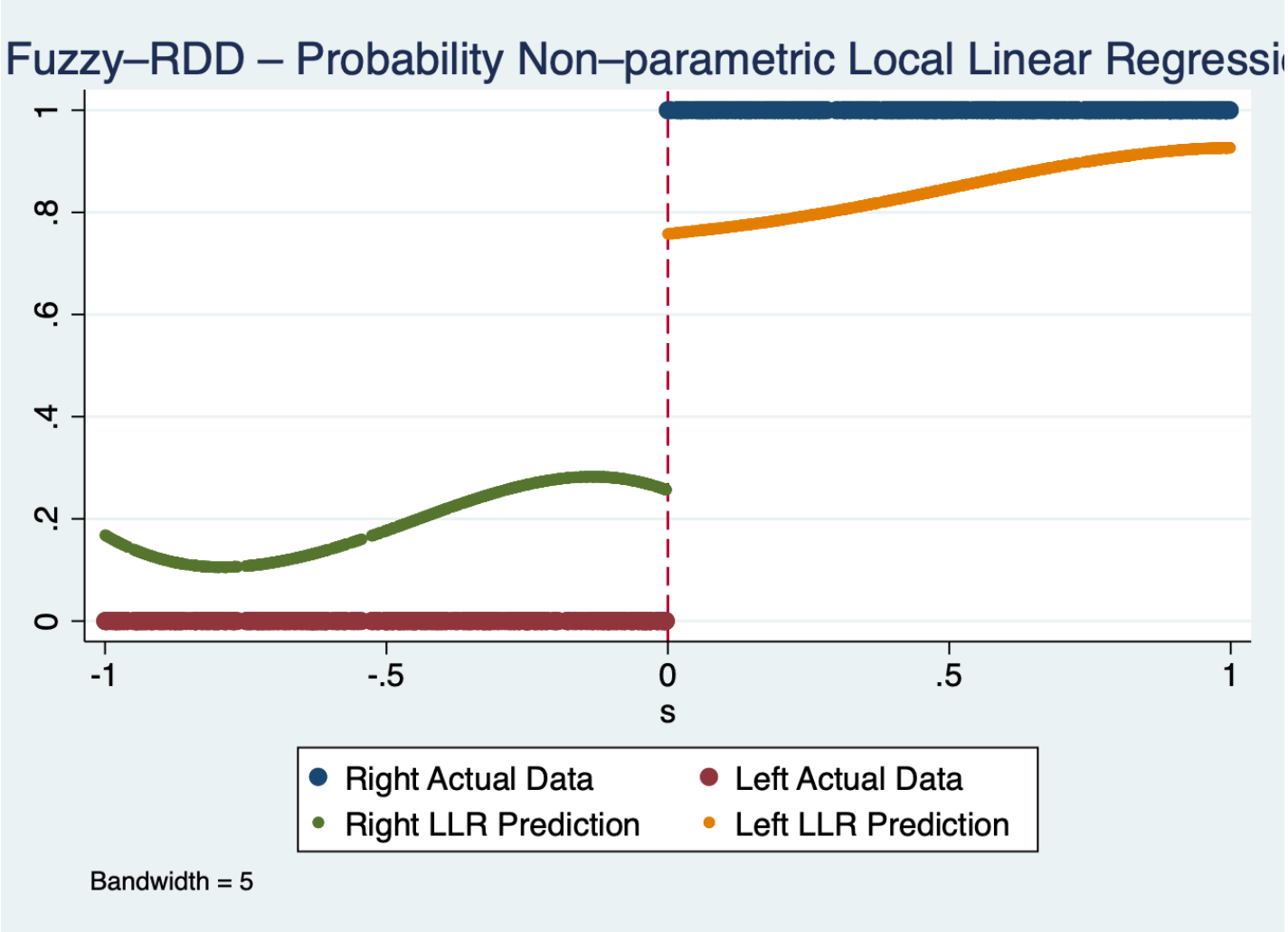


Local polynomials for f_0 and f_1

Identification

- The regression discontinuity design identifies the conditional ATE at the treatment cut-off.
- What we need for identification is the continuity of $f_1(X)$ and $f_0(X)$: which means that the conditional expectation of the untreated and treated outcome are continuously affected by the running variable.

Sharp RD vs. Fuzzy RD



Discontinuity in D

Fuzzy RD

- Now, there is a jump in the probability in the probability of treatment at x_0, such that

$$\mathbb{P}(D_i = 1|x_i) = \begin{cases} g_1(x_i), & \text{if } x_i \geq x_0 \\ g_0(x_i), & \text{if } x_i < x_0 \end{cases}$$

  where $g_1(x_0) \neq g_0(x_0)$. We assume $g_1(x_0) > g_0(x_0)$ so that $x_i \geq x_0$ makes treatment more likely.
- Fuzzy RD is IV: the discontinuity becomes an instrumental variable for treatment status instead of deterministically switching treatment on or off.
- The ATE at the cutoff:

$$\alpha_{ATE,RD} = \frac{\mathbb{E}(y|x_0^+) - \mathbb{E}(y|x_0^+)}{\pi(x_0^+) - \pi(x_0^-)}$$

  In sharp RD, $\pi(x_0^+) - \pi(x_0^-) = 1 - 0 = 1$.

Fuzzy RD

The Oreopoulos (2006) Example

- The treatment variable D_i has conditional density:

$$f(D_i|x_i) = \begin{cases} g_1(x_i), & x_i \geq 47 \\ g_0(x_i), & x_i < 47 \end{cases}, \quad g_1(47) > g_0(47)$$

where $x_i \geq 47$ makes the treatment more likely. Instrument variable:

$$Z_i = \begin{cases} 1, & x_i \geq 47 \\ 0, & x_i < 47 \end{cases}$$

- It follows that

$$\begin{aligned} \mathbb{E}(D_i|x_i) &= \int D_i f(D_i|x_i) dD_i \\ &= \int D_i \Big[g_0(x_i) + \Big(g_1(x_i) - g_0(x_i) \Big) \cdot Z_i \Big] dD_i \\ &= \mathbb{E}(D_i) \Big[g_0(x_i) + \Big(g_1(x_i) - g_0(x_i) \Big) \cdot Z_i \Big] \end{aligned}$$

Fuzzy RD is IV

The Oreopoulos (2006) Example

- We repeat the last line here

$$\mathbb{E}(D_i|x_i) = \mathbb{E}(D_i)\Big[g_0(x_i) + \Big(g_1(x_i) - g_0(x_i)\Big) \cdot Z_i\Big]$$

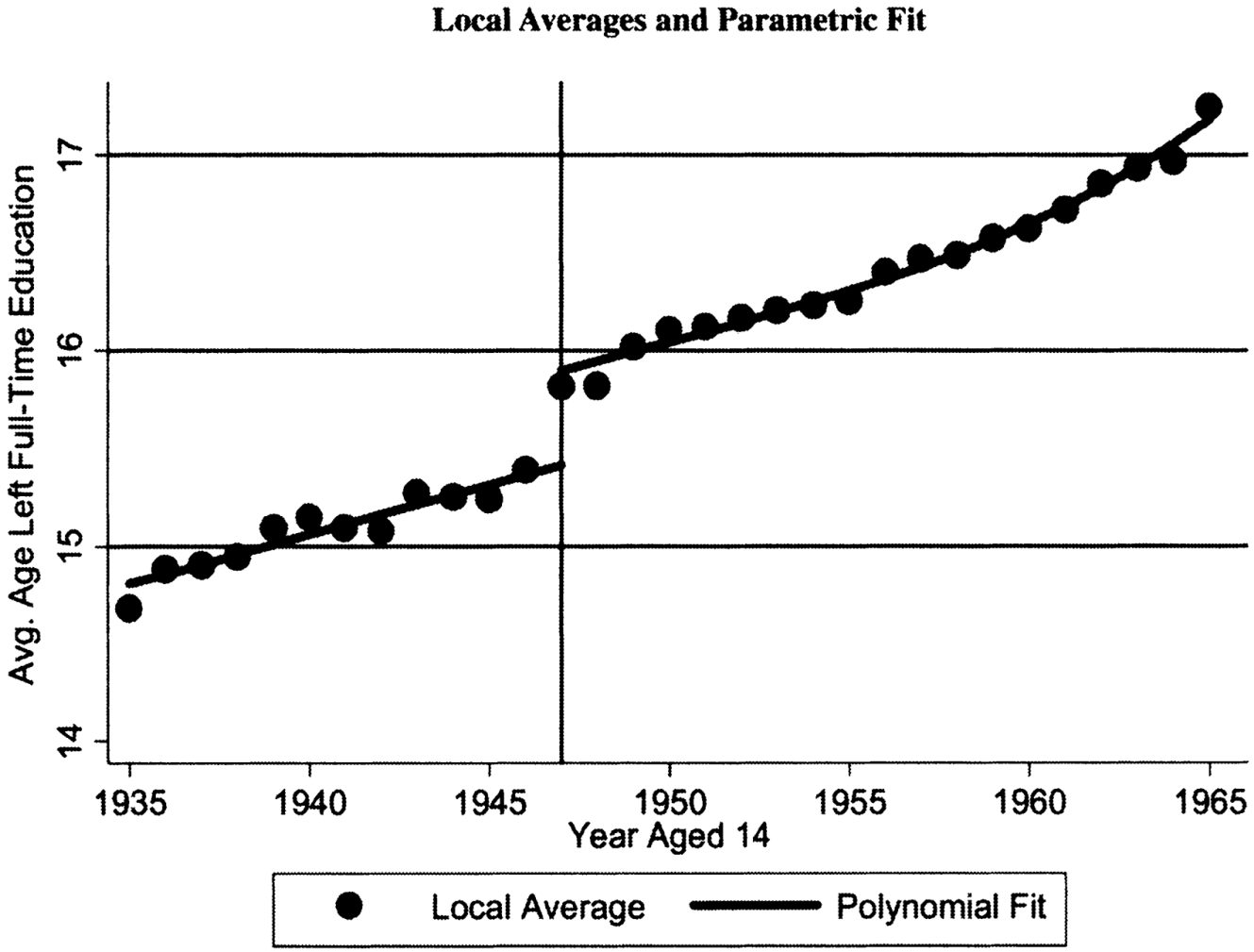
  The dummy variable Z_i indicates the point of discontinuity in $E(D_i|x_i)$.
- To capture the non-linearity of the trend, we assume $g_1(x_i)$ and $g_0(x_i)$ each be some resonably smooth function, for example, a p-th order polynomial:

$$E(D_i|x_i) = \mathbb{E}(D_i)\Big[\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \Big(\beta_0^* + \beta_1^* x_i + \beta_2^* x_i^2 + \cdots + \beta_p^* x_i^p\Big) \cdot Z_i\Big]$$

- From this (the relevance condition) we see that Z_i as well as the interaction terms $\{x_i Z_i, x_i^2 Z_i, \ldots, x_i^p Z_i\}$ can be used as instruments for D_i.

Fuzzy RD

The Oreopoulos (2006) Example



Discontinuity in D

Fuzzy RD is IV

The Oreopoulos (2006) Example

1. The first stage of the 2SLS: discontinuity in D

$$D_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \tilde{\beta}_2 x_i^2 + \cdots + \tilde{\beta}_p x_i^p + \gamma Z_i + u_i \tag{1}$$

2. The second stage of the 2SLS: discontinuity in y

   Assume $E(Y_{0i}|x_i) = h(x_i)$, where $h(x_i)$ is also a p-th order polynomial of x_i.

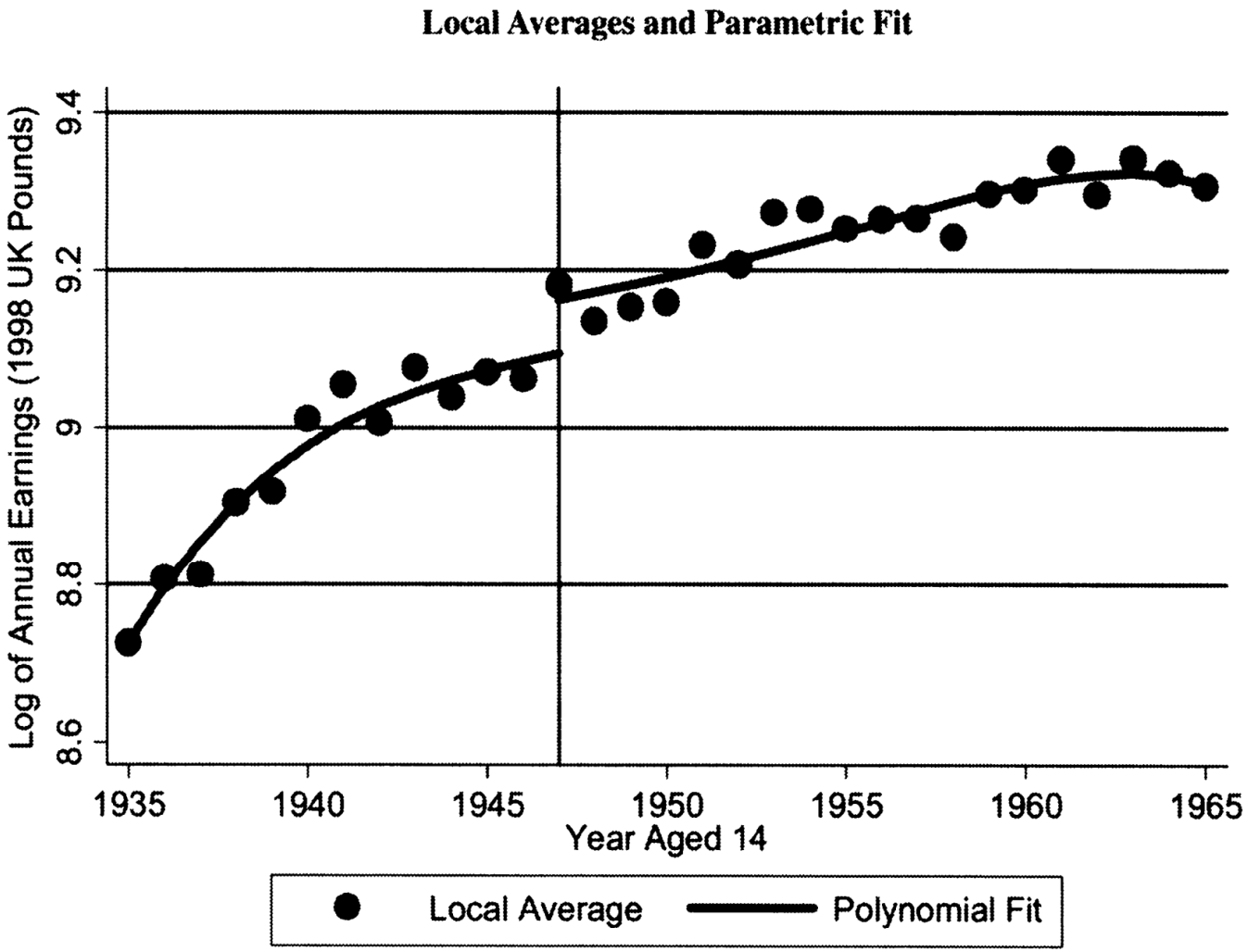
$$\begin{aligned} Y_i &= \alpha_i D_i + h(x_i) + \epsilon_i \\ &= \alpha_i D_i + \rho_0 + \rho_1 x_i + \rho_2 x_i^2 + \cdots + \rho_p x_i^p + \epsilon_i \end{aligned} \tag{2}$$

   The fuzzy RD reduced form is obtained by substituting (1) into (2):

$$\begin{aligned} Y_i =& \alpha_i \left[\tilde{\beta}_0 + \tilde{\beta}_1 x_i + \tilde{\beta}_2 x_i^2 + \cdots + \tilde{\beta}_p x_i^p + \gamma Z_i + u_i \right] \\ &+ \rho_0 + \rho_1 x_i + \rho_2 x_i^2 + \cdots + \rho_p x_i^p + \epsilon_i \\ =& \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_p x_i^p + \tilde{\epsilon} \end{aligned}$$

Fuzzy RD is IV

The Oreopoulos (2006) Example



Discontinuity in y

Exercise

Exam question fall 2021

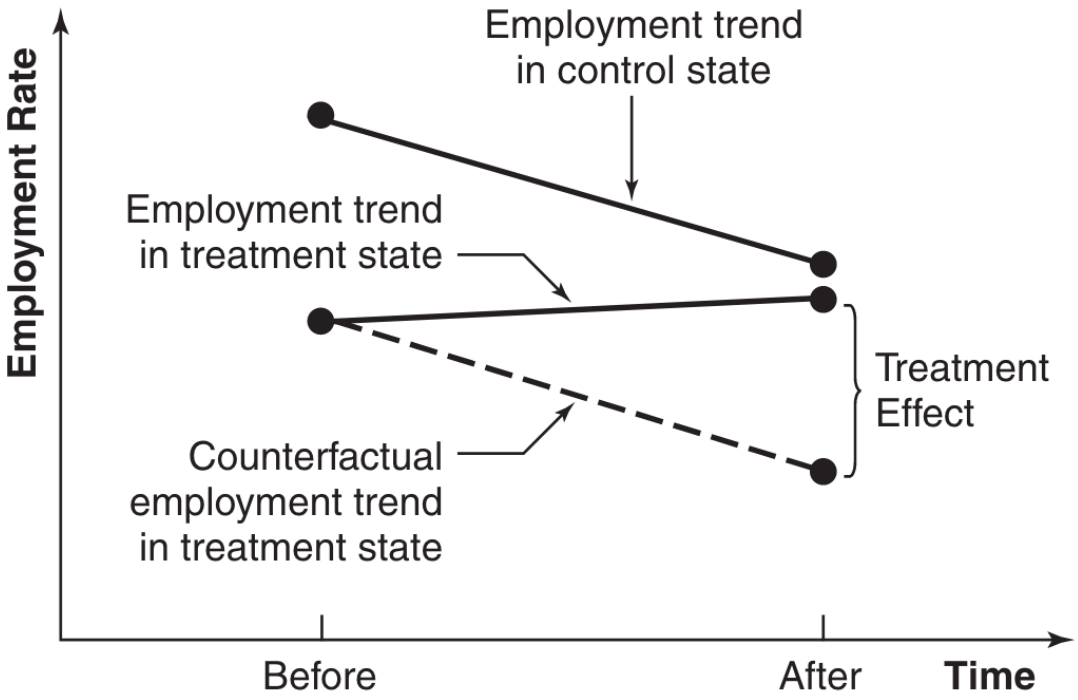
2. (20 points) You are interested in studying the effect of an artists death on the price of their artwork. You have data on the auction sales of a number of renowned artists through their life-time and after their death. Each observation is an artwork sold (e.g. a painting sold at \$10 000 when the artist was 35 years old, another sold at \$20 000 two years after the same artist's death etc.).

   (a) (15 points) Describe how you would estimate the effect using regression discontinuity. What is your Z variable? Illustrate how your data would look like if the death of an artist causes prices to increase, in particular plot Y and D against Z. Is this a sharp or fuzzy design?

   (b) (5 points) You also have some variables X on the characteristics of the painting (e.g. size, medium, motif etc.). Does it makes sense to include these in the estimation? How do you expect X to behave around the Z cut-off if the RD design is valid?

Differences-in-Differences

Differences-in-Differences



- The basic DID model is a two-way fixed effects model:

$$y_{it} = \alpha D_{it} + X_{it}'\beta + \nu_i + \gamma_t + \varepsilon_{it}$$

Trend specification

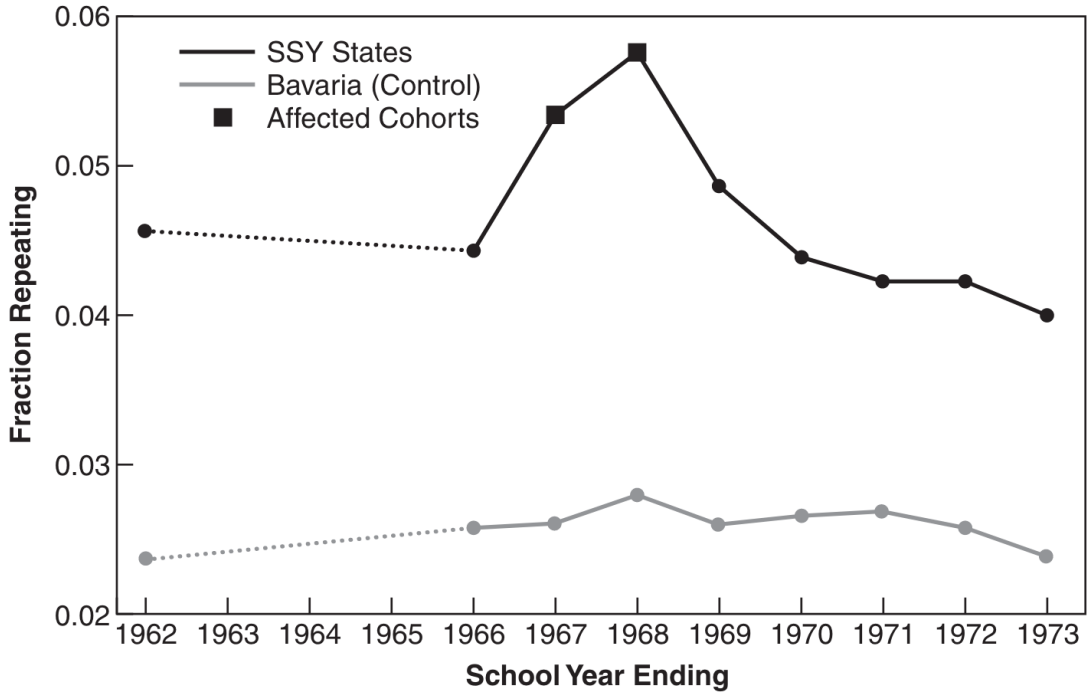


Figure 5.2.3 Average grade repetition rates in second grade for treatment and control schools in Germany (from Pischke, 2007). The data span a period before and after a change in term length for students outside Bavaria (SSY states).

- Message from the graph:
  1. strong visual evidence of treatment and control states with a common underlying trend, and
  2. a treatment effect that induces a sharp but transitory deviation from this trend.

References

- Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: methods and applications. Cambridge university press. Chapter 25.
- Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. econometrica, 74(1), 235-267.
- Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press. Chapter 21.
- Imbens, G. W. (2015). Matching methods in practice: Three examples. Journal of Human Resources, 50(2), 373-419.
- Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. Econometrica, 84(2), 781-807.
- Cameron, A. C., & Trivedi, P. K. (2022). Microeconometrics using stata (Second Edition). Stata press. Chapters 24, 25.
- Hansen, B. E. (2022). Econometrics. Chapters 18, 21.