TA Session 1
Treatment Effects I: RCT, Matching, and IV

Microeconometrics II with Joan Llull
IDEA, Fall 2024

TA: Conghan Zheng

November 08, 2024

Overview

Introduction

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)

Exogenous treatment

- **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.

RCT

Difference in Means

- When the strong treatment assignment assumptions for causal inference in RCTs are met ...
- Treatment effects in RCT:

$$\alpha_{ATE} = \frac{1}{N}\sum_{i=1}^{N} y_{1i} - \frac{1}{N}\sum_{i=1}^{N} y_{0i}$$

$$\alpha_{ATT} = \frac{1}{N_1}\sum_{i:D_i=1} y_{1i} - \frac{1}{N_1}\sum_{i:D_i=1} y_{0i}$$

- **t-test** of the null ($\alpha = 0$) is a test of differences in means, $H_0 : \mu_1 - \mu_0 = 0$.

Matching

Data

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

- Source: the National Supported Work (NSW)
- Causal relationship of interest: 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

Matching

- Regressions (including the binary treatment indicator as a regressor) 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.
  - Regression (including the binary treatment indicator as a regressor) 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

Measures

- Matching measures:
  1. regressors;
  2. propensity score.

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.

Algorithms

- 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**: 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**: 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

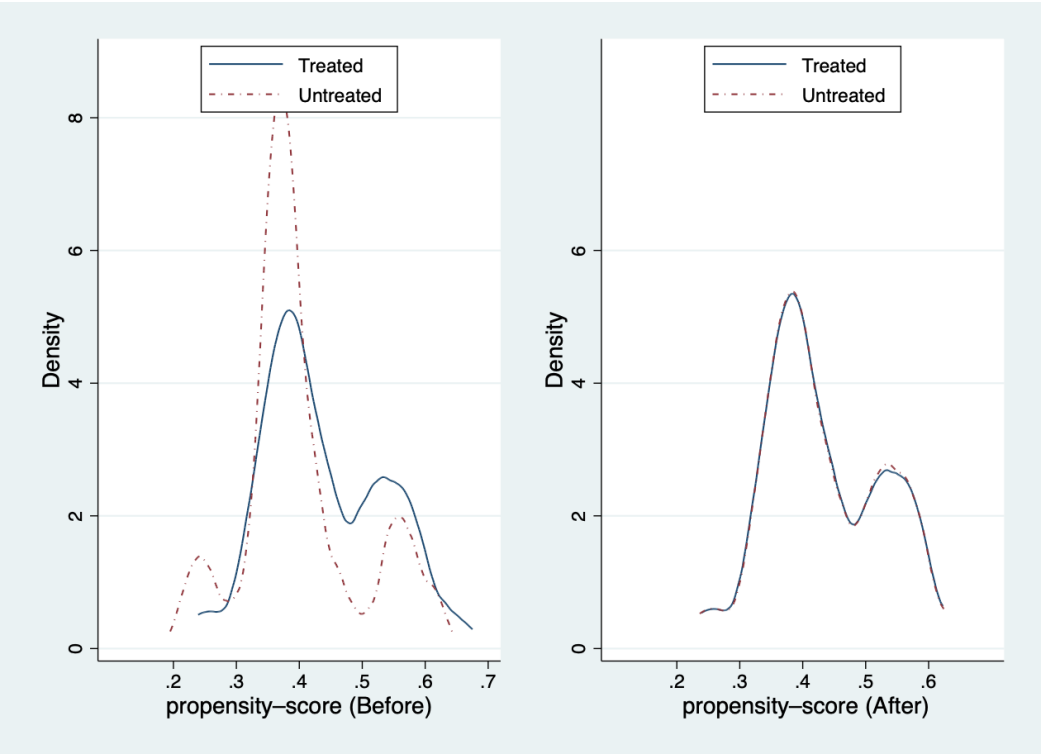
$$\hat{y}_i(0) = \begin{cases} y_i, & \text{if } D_i = 0 \\ \sum_{j:D_j=0} w(i,j)y_j, & \text{if } D_i = 1 \end{cases}$$

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

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

Propensity score matching

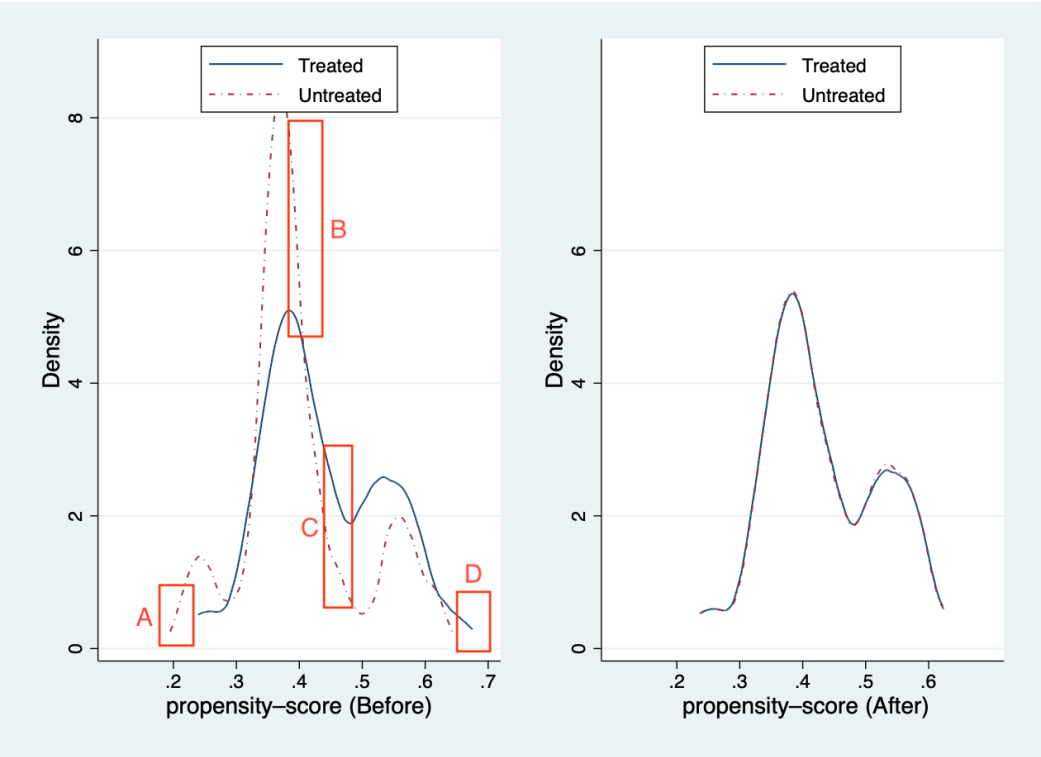
k Nearest neighbor matching



- 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

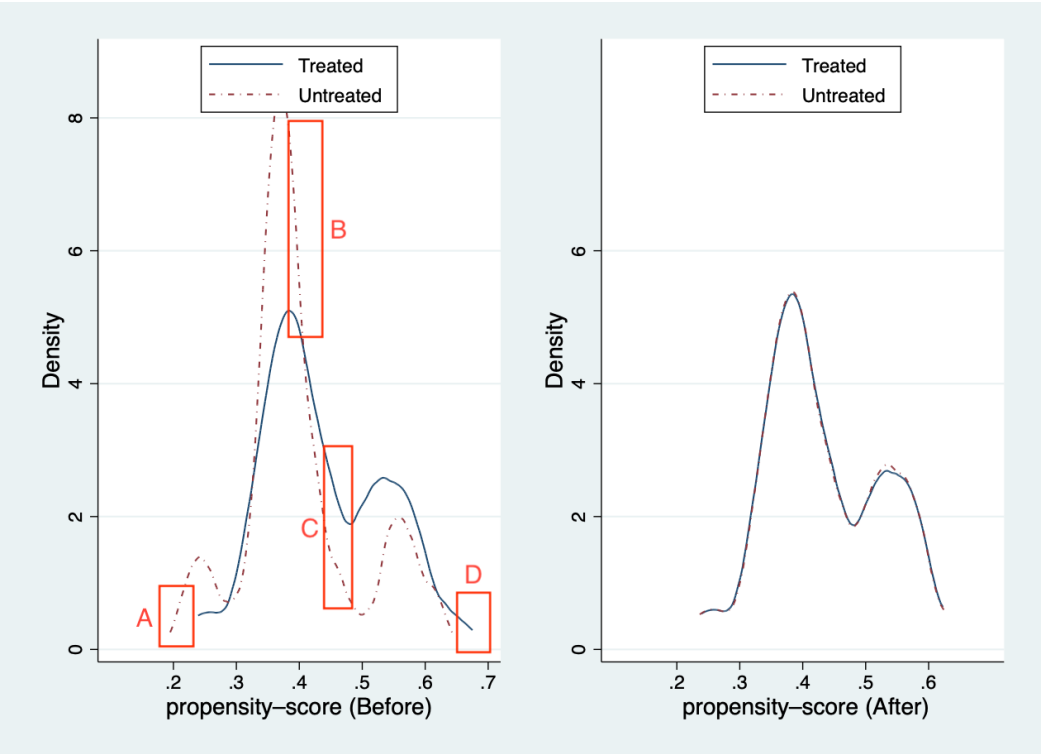
k Nearest neighbor matching



- 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

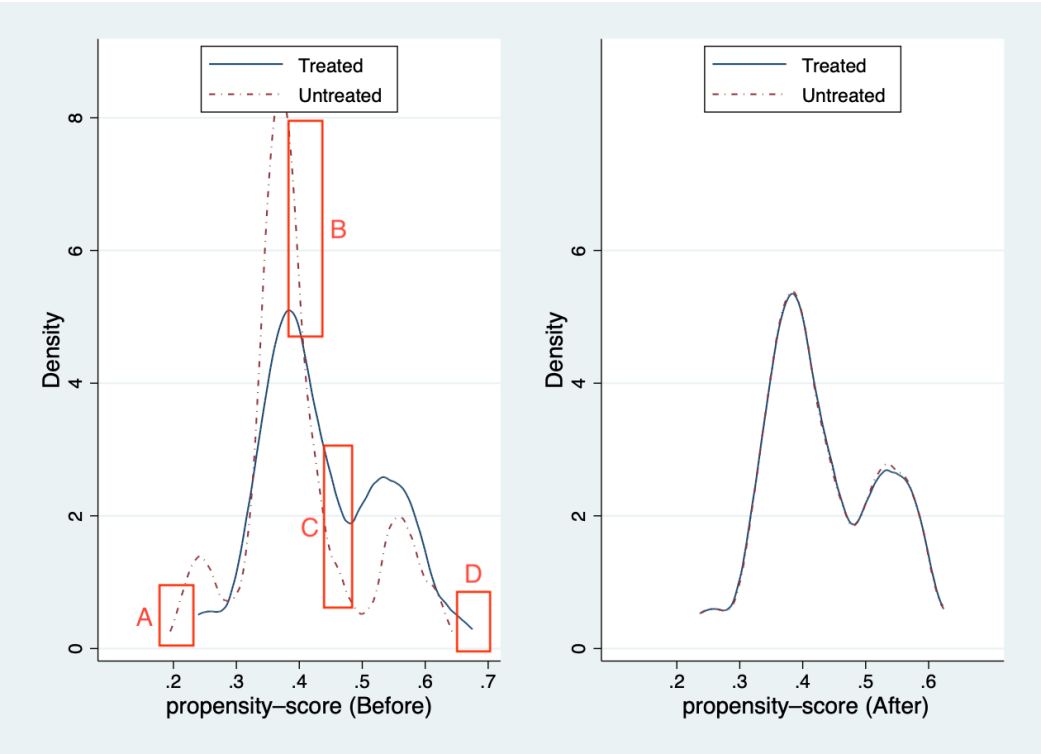
k Nearest neighbor matching



- 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

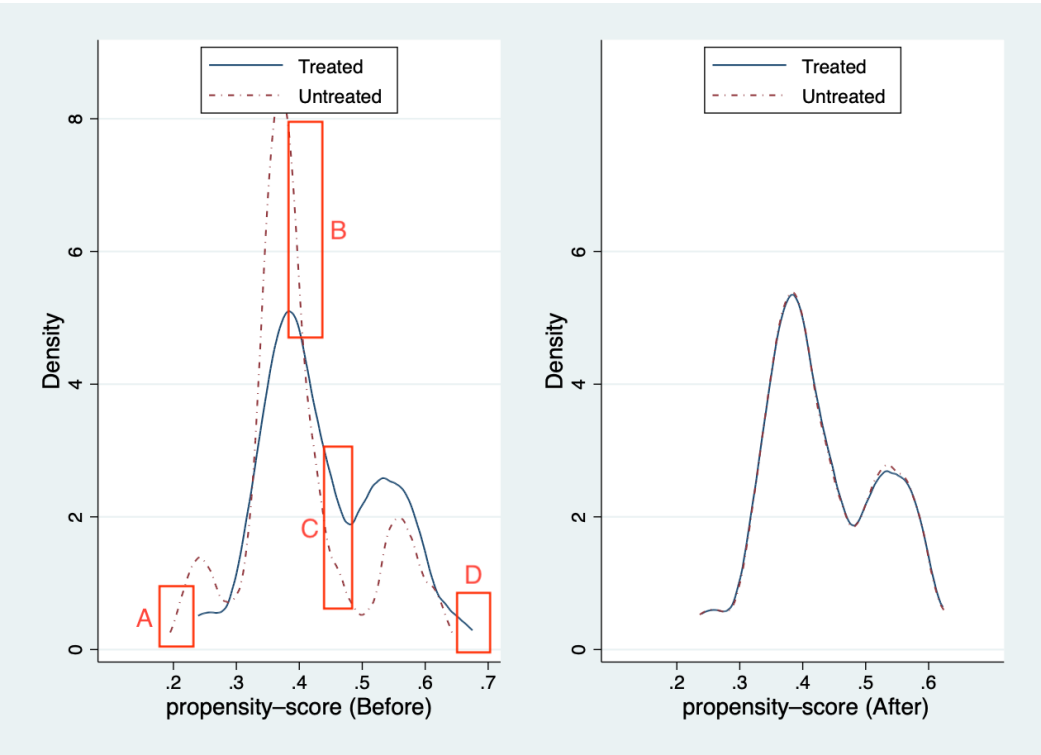
k Nearest neighbor matching



- 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

- 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.

Instrumental variables

Identification of Causal Effects using IV

- Conditional independence is not satisfied: $(y_{0i}, y_{1i}) \not\perp D_i|X_i$
- Exogenous source of variation Z_i such that $(y_{0i}, y_{1i}) \perp Z_i|X_i$ and $Z_i \not\perp D_i|X$
- IV estimate: $\alpha^{IV}_{LATE} = \frac{Cov(Z_i,y_i)}{Cov(Z_i,D_i)}$
- L for Local in LATE: LATE measures the treatment effect on the "compliers" (at the margin of participating) that are induced to participate in the treatment as a result of the change in Z.
- Compliers (comply with the treatment assignment); always-takers (received treatment regardless of eligibility); never-takers (refuse treatment regardless of eligibility); and assume there are no defiers.

Estimation

- Probit 2SLS:
  1. Participation equation using Probit - predicted probability $\hat{p}$
  2. Predict participation indicator $\hat{D}$ using $\hat{p}$
  3. Outcome equation using OLS of y on observables and $\hat{D}$
- Stata Commands:
  - IV Probit

    ```
    . ivprobit depvar (endog=exog) $x
    ```

    where `depvar` is the dependent variable, `endog` is the endogenous regressor, `exog` is the instrument, and `x` is the vector of covariates.
  - Visualize marginal effects

    ```
    . margins, at(endog=(min(step)max)) predict(pr)
    . marginsplot
    ```

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.