

Lecture 4 - When Complete Randomization is Infeasible

Matching and Weighting

Dr. Xi Chen

Rotterdam School of Management

2025-06-21



Outline

1 When complete randomization is infeasible

Examples

Identification under “selection on observable”

2 Various estimators under selection on observables

Subclassification

Matching

Weighting with propensity scores

3 Appendix

Complete randomization can be inefficient

Imagine you can only “afford” 20 consumers in your study.
You know beforehand **gender** is an important factor for the study outcomes.

If you were leave it to chance to assign the 10 people:

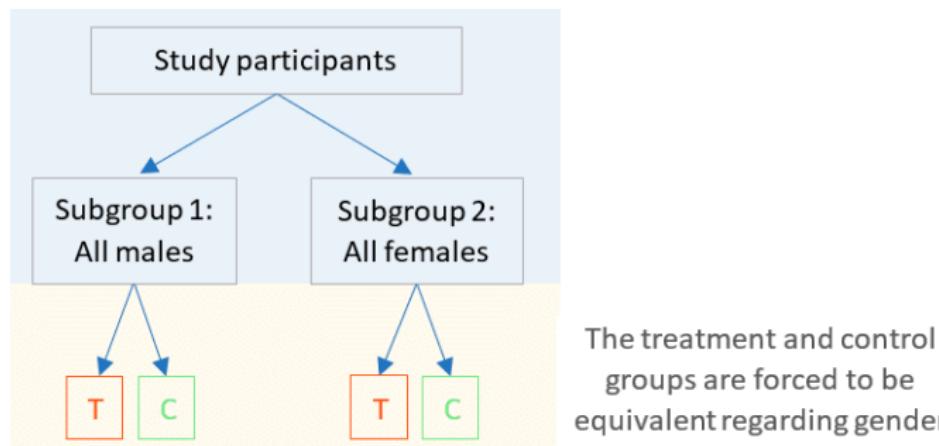
Randomly assign 10 consumers

```
people <- rep(c("female","male"),5)
unlist(rerun(10, sum(sample(people,10,replace = F)=="female")))
[1] 6 4 4 4 5 6 5 7 5 3
```

Blocking to increase efficiency

Instead, we stratify the sample to females and males, and then randomize within each subgroup.

- ① Reduces sampling uncertainty and improves precision;
- ② Ensures that certain subgroups are available for analysis (heterogeneity).



Blocking is efficient: An example

store	block	All Stores		Block A		Block B	
		Y_1	Y_0	Y_1	Y_0	Y_1	Y_0
1	A	4	1	4	1		
2	A	5	3	5	3		
3	A	6	2	6	2		
4	A	4	1	4	1		
5	A	3	2	3	2		
6	A	2	1	2	1		
7	B	10	8			10	8
8	B	13	10			13	10
9	B	11	9			11	9
10	B	10	8			10	8
11	B	12	10			12	10
12	B	11	11			11	11
Mean		7.6	5.5	4.0	1.7	11.2	9.3
Variance		14.2	15.6	1.7	0.6	1.1	1.2

Under complete randomization:

$$\begin{cases} \text{ATE}_{\text{CR}} &= 7.6 - 5.5 = 2.1 \\ \text{SE}_{\text{CR}}^{\text{Neyman}} &= \sqrt{\frac{14.2}{12} + \frac{15.6}{12}} = 1.57 \end{cases}$$

Blocking is efficient: An example

store	block	All Stores		Block A		Block B	
		Y_1	Y_0	Y_1	Y_0	Y_1	Y_0
1	A	4	1	4	1		
2	A	5	3	5	3		
3	A	6	2	6	2		
4	A	4	1	4	1		
5	A	3	2	3	2		
6	A	2	1	2	1		
7	B	10	8			10	8
8	B	13	10			13	10
9	B	11	9			11	9
10	B	10	8			10	8
11	B	12	10			12	10
12	B	11	11			11	11
Mean		7.6	5.5	4.0	1.7	11.2	9.3
Variance		14.2	15.6	1.7	0.6	1.1	1.2

Under block randomization:

$$\begin{cases} \text{ATE}_{\text{BR}} = \sum_{b=1}^B \frac{N_b}{N} \text{ATE}_b \\ \text{SE}_{\text{BR}}^{\text{Neyman}} = \sqrt{\sum_{b=1}^B \left(\frac{N_b}{N} \text{SE}_b^{\text{Neyman}} \right)^2} \end{cases} = \frac{6}{12} \cdot 2.3 + \frac{6}{12} \cdot 1.9 = 2.1 \\ = \mathbf{0.62}$$

① When complete randomization is infeasible

Examples

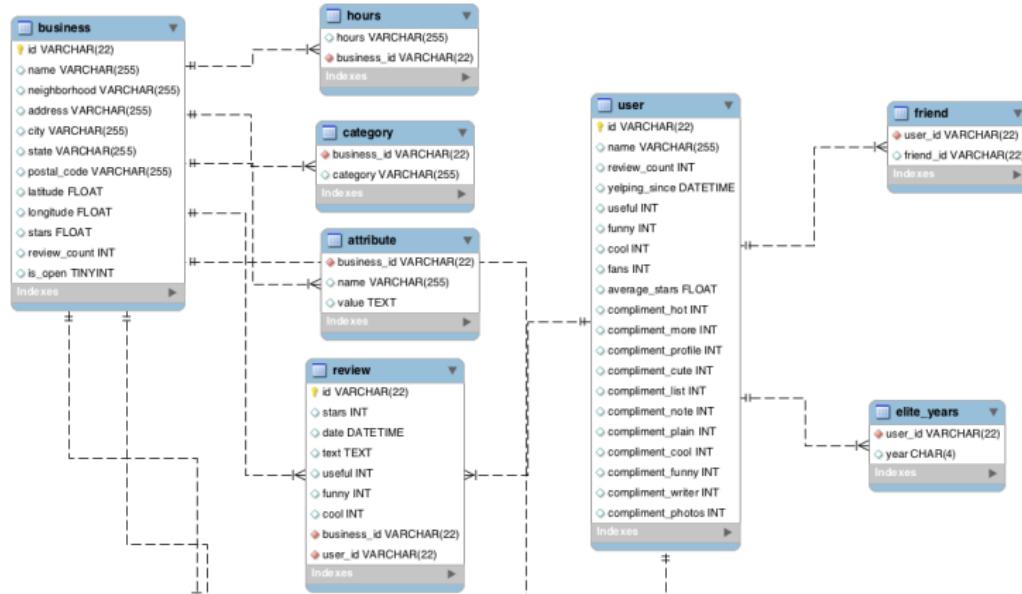
Identification under “selection on observable”

② Various estimators under selection on observables

③ Appendix



Yelp Elite



Almost all information is known, but not how Yelp selects Elites.

① When complete randomization is infeasible

Examples

Identification under “selection on observable”

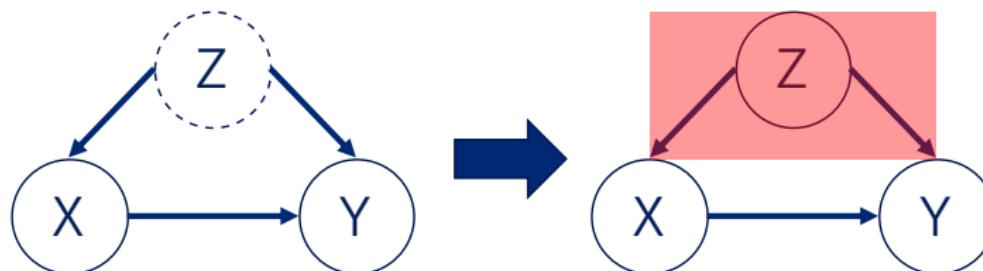
② Various estimators under selection on observables

③ Appendix

Assumptions

Identification Assumptions

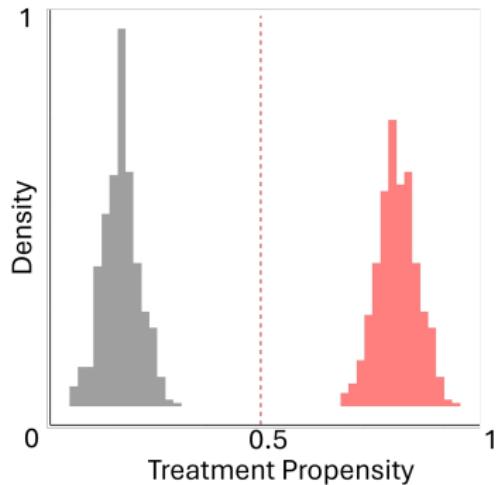
- ① $(Y_i^1, Y_i^0) \perp D_i | X_i$ (conditional unconfoundedness)
- ② $p_i(D_i = 1 | X_i) \in (0, 1)$ (common support)



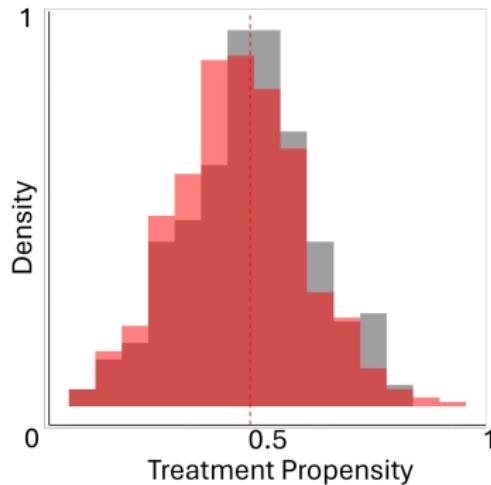
It's essentially a “conditioning” identification strategy

Understanding common support

The treatment propensity of the treated vs. control group must be overlapped.



(a) Non-overlapped



(b) Overlapped

Identification results: ATE

Under the conditional unconfoundedness:

$$\begin{aligned} E[Y_i^1 | X_i, D_i = 1] &= E[Y_i^1 | X_i] \\ E[Y_i^0 | X_i, D_i = 0] &= E[Y_i^0 | X_i] \end{aligned}$$

With common support assumption:

$$\tau_{ATE} = \int \left(E[Y_i^1 | X_i] - E[Y_i^0 | X_i] \right) dP(X_i)$$

Identification results: ATT

Under milder assumptions, in practices, also identify **ATT** (average treatment effects on the treated units).

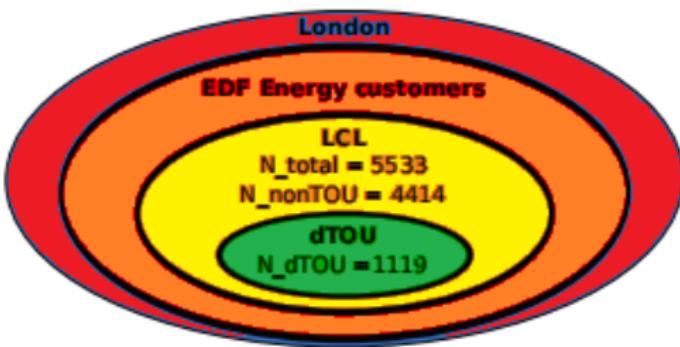
$$\text{ATT} = E \left[Y_i^1 - Y_i^0 \mid D_i = 1 \right]$$

If we assume a milder version of unconfoundedness: $Y_i^0 \perp D_i \mid X_i$ and common support $p_i(D_i = 1 \mid X_i) < 1$, ATT is identified.

$$\begin{aligned} E \left[Y_i^0 \mid X_i, D_i = 1 \right] &= E \left[Y_i^0 \mid X_i \right] \\ E \left[Y_i^1 \mid X_i, D_i = 1 \right], \text{ directly from data} \end{aligned}$$

An example of ATT: “Low Carbon London” trial

Figure 1: Venn diagram illustrating the sample selection in LCL trial



The control group is randomly drawn and therefore ATT is identified.

① When complete randomization is infeasible

② Various estimators under selection on observables

Subclassification

Matching

Weighting with propensity scores

③ Appendix

① When complete randomization is infeasible

② Various estimators under selection on observables

Subclassification

Matching

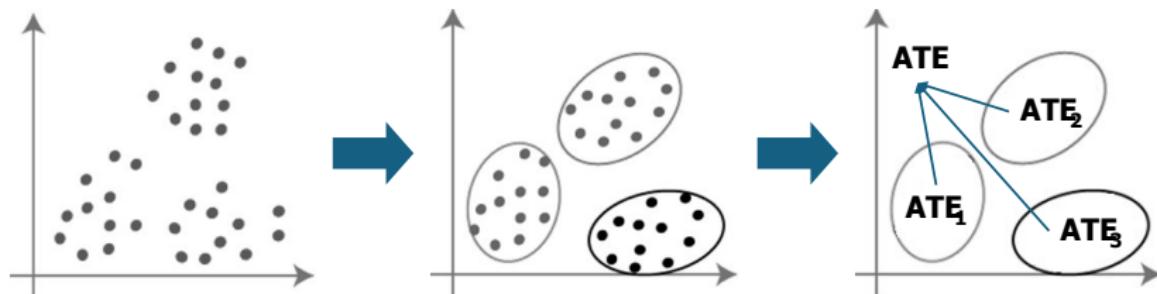
Weighting with propensity scores

③ Appendix

Subclassification: the process

Given $(Y_i^1, Y_i^0) \perp D_i | X_i$, emulate blocking by creating “blocks” based on X_i :

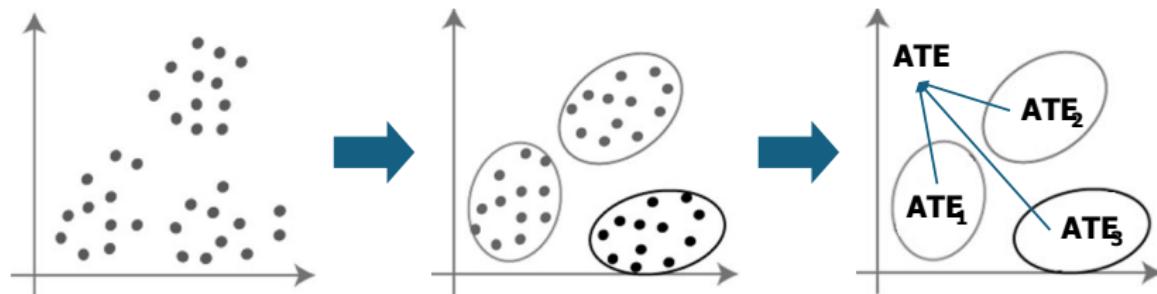
- Discrete variables: to use directly.
- Continuous variables: to first discretize and then use.



Subclassification: the process

By creating K groups, with size N_k for a group k :

$$\begin{cases} \hat{\tau}_{ATE} = \sum_{k=1}^K \frac{N_k}{N} (\bar{Y}_1^k - \bar{Y}_0^k) \\ \hat{\tau}_{ATT} = \sum_{k=1}^K \frac{N_k^1}{N^1} (\bar{Y}_1^k - \bar{Y}_0^k) \end{cases}$$



Subclassification: an example

Usage X_i	Average Purchase		Diff.	# Ad Viewers	# Units
	With Ads	No Ads		N_k^1	N_k
Heavy	€28	€24	€4	3	10
Light	€22	€16	€6	7	10
Total				10	20

$$\hat{\tau}_{ATE} = \sum_{k=1}^K \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \frac{N_k}{N} = 4 \cdot \frac{10}{20} + 6 \cdot \frac{10}{20} = €5$$

$$\hat{\tau}_{ATT} = \sum_{k=1}^K \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \frac{N_k^1}{N^1} = 4 \cdot \frac{3}{10} + 6 \cdot \frac{7}{10} = €5.4$$

Subclassification: potential problems

Usage	Average Purchase		Diff.	# Ad Viewers	# Units
	With Ads	No Ads			
Heavy, Male	€28	22	€4	3	7
Heavy, Female	??	€24	??	0	3
Light, Male	€21	€16	€5	3	4
Light, Female	€23	€17	€6	4	6
Total				10	20

$$\widehat{\tau}_{ATE} = \sum_{k=1}^K \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \frac{N_k}{N} = ???$$

$$\widehat{\tau}_{ATT} = \sum_{k=1}^K \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \frac{N_k^1}{N^1} = ???$$

① When complete randomization is infeasible

② Various estimators under selection on observables

Subclassification

Matching

Weighting with propensity scores

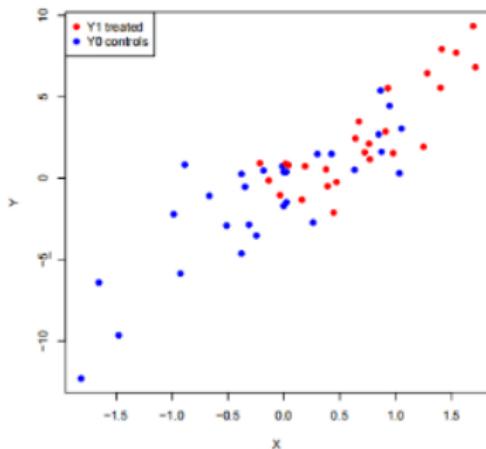
③ Appendix

Matching: the basic idea

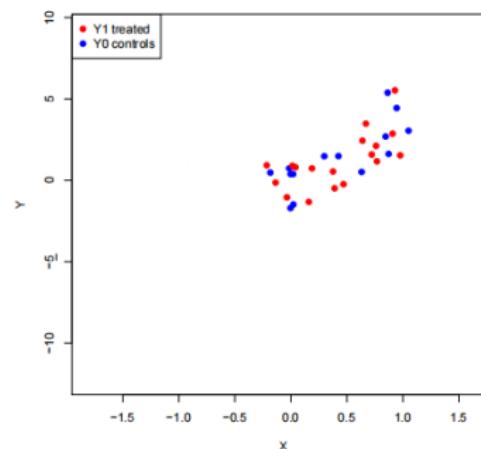
Recall the conditional unconfoundedness:

$$E[Y_i^D | X_i, D_i] = E[Y_i^D | X_i]$$

Implication: **to equalize the distribution of X_i for the treated and control.**



(a) Before Matching



(b) After Matching

Matching: a general procedure



1. SELECT A MATCHING METHOD.



2. OBTAIN THE MATCHED SAMPLES.



3. EVALUATE THE MATCHING PERFORMANCES.



4. CALCULATE THE TREATMENT EFFECTS.

Matching: a simple example

Unit	Potential Outcome		Treatment	Variables
	Under Treatment	Under Control		
i	Y_i^1	Y_i^0	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

What is $\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i:D_i=1} (Y_i^1 - \overline{Y}_j^0)$?
 Match and plug in potential outcomes.

Matching: a simple example

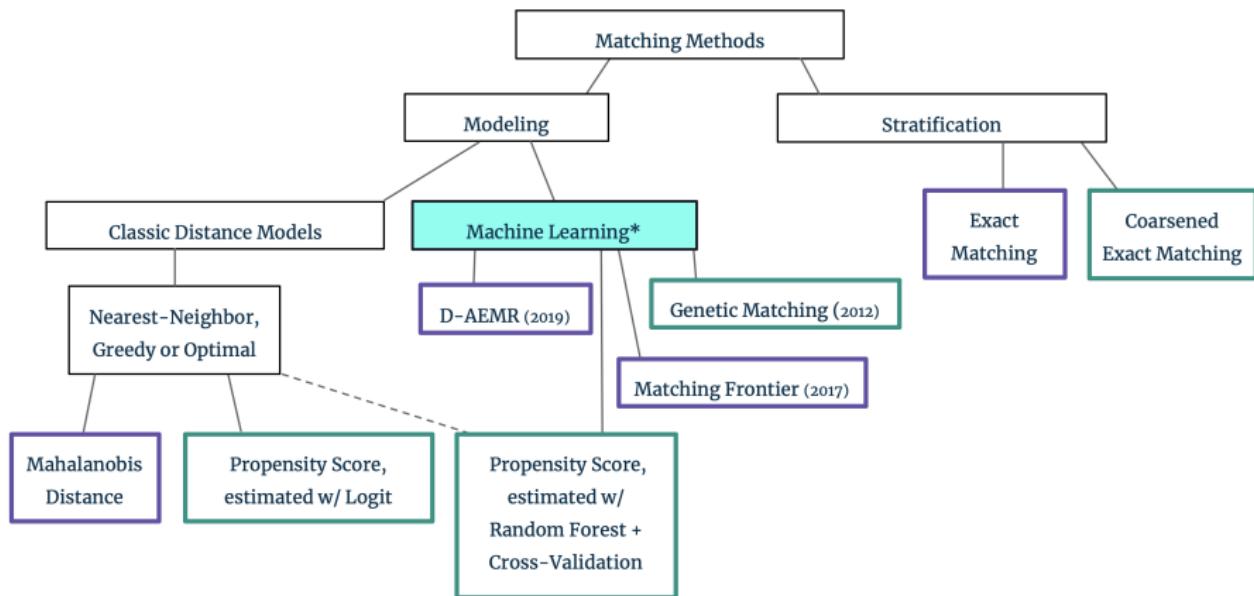
Unit	Potential Outcome		Treatment	Variables
	Under Treatment	Under Control		
i	Y_i^1	Y_i^0	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

What is $\hat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i:D_i=1} (Y_i^1 - \bar{Y}_j^0)^1$?

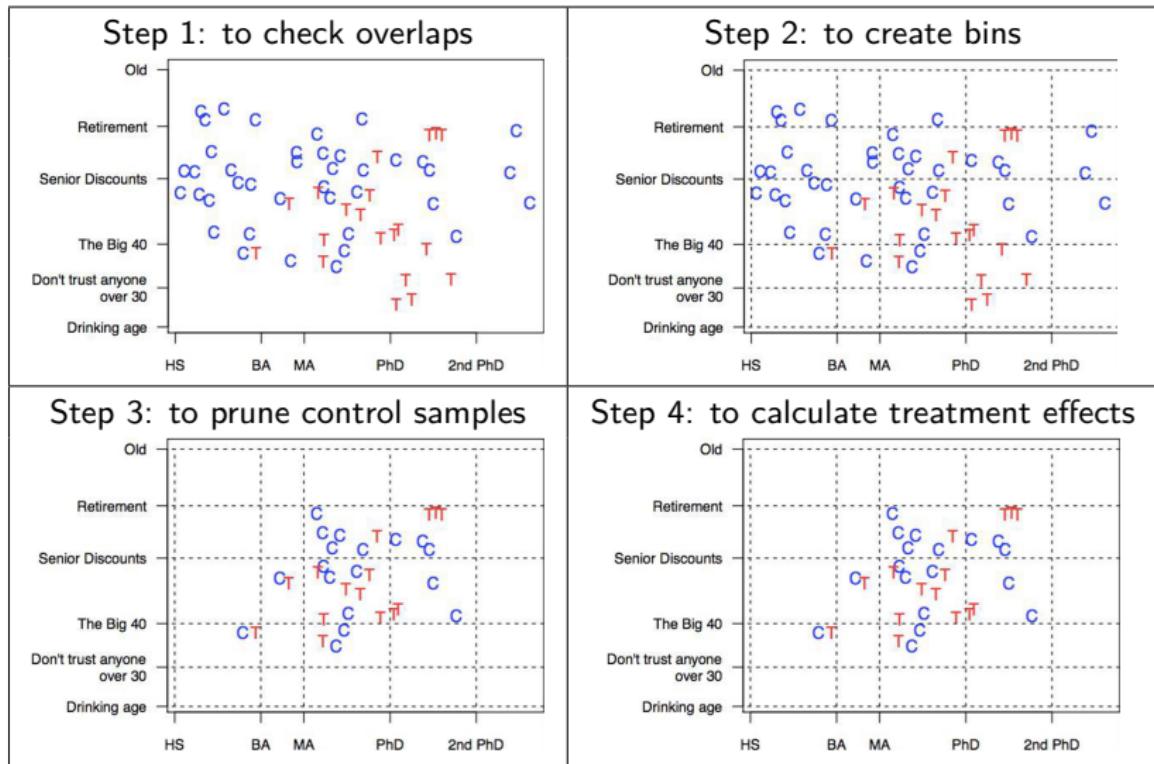
$$\hat{\tau}_{ATT} = \frac{1}{3} ((6 - 9) + (0 - 1) + (9 - 1)) = -3.7$$

¹Note: ATC (ATE on the control units) can be calculated similarly.

A classification of matching methods



Stratification: coarsened exact matching



Modeling: classic distance models

One way is to define “closeness” by a distance metric.

Classic distance metrics

Suppose we have a treated unit i and a control unit j .

A set of continuous characteristics for i and j , X_i and X_j .

Two most frequently used distance metrics:

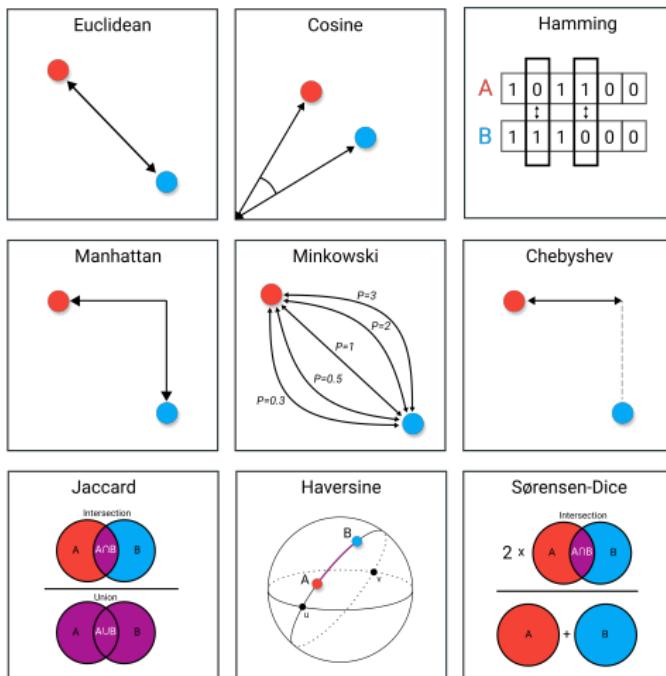
$$\text{Mahalanobis distance: } \sqrt{(X_i - X_j)' \Sigma^{-1} (X_i - X_j)}$$

$$\text{Euclidean distance: } \sqrt{(X_i - X_j)' (X_i - X_j)}$$

For an “exact match”, the distance metrics are zeros.

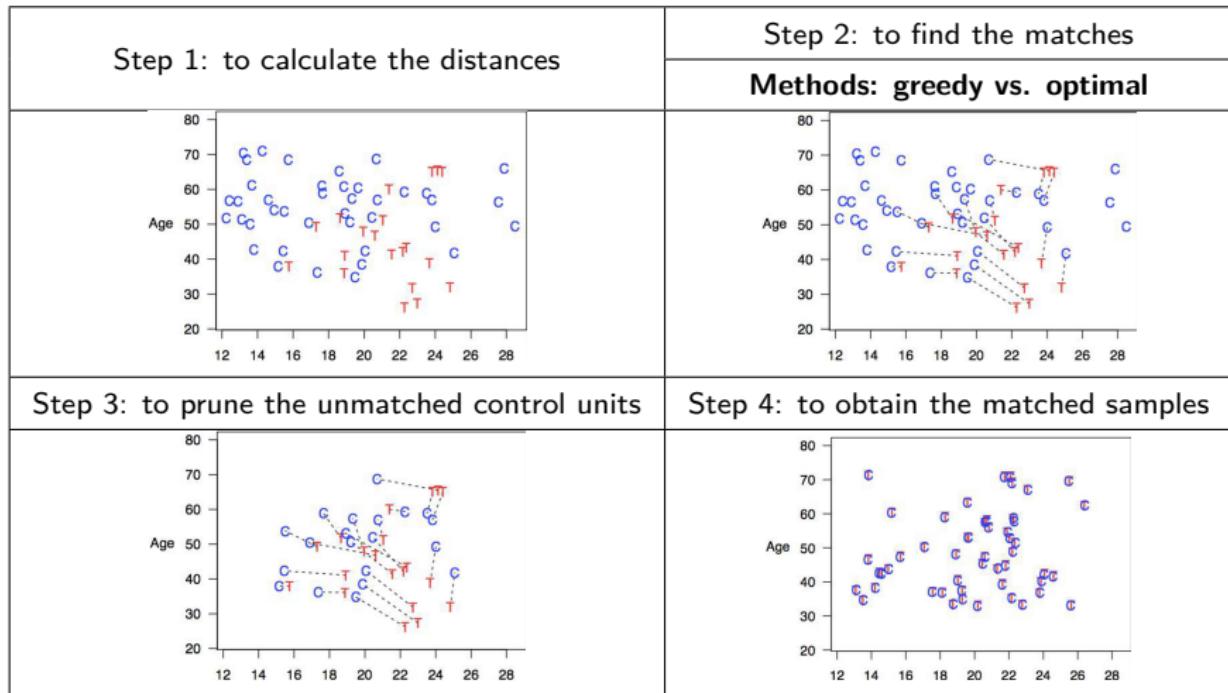
Modeling: classic distance models

A rich selection of distance metrics (probably too many...)

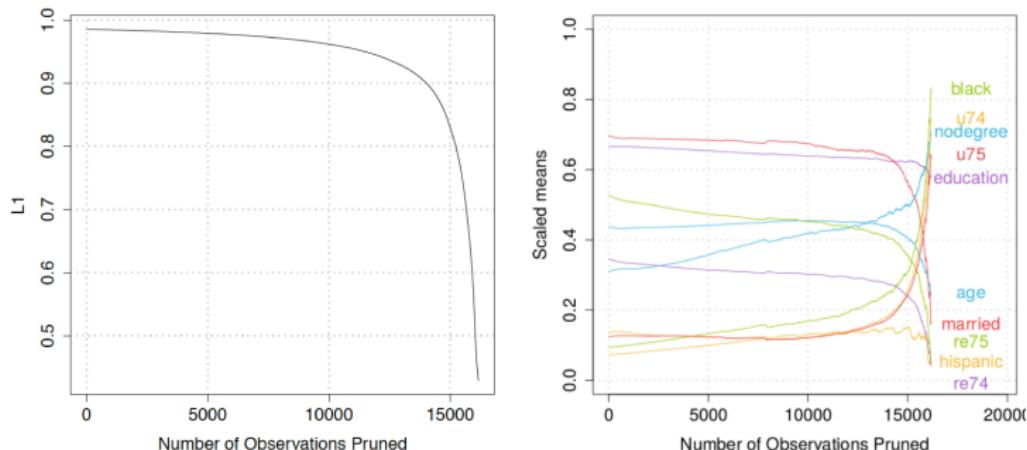


Examples of Distance Metrics

Modeling: classic distance models



Modeling: matching frontier



Source: King et al. (2017)

Formalize a standard bias-variance trade-off in matching:

Poor covariate balance can lead to biased causal effect estimates but pruning too many observations can increase the variance of the estimates.

Evaluate matching: balance tests

The performance of matching depends on **the resulting balance of X_i** between matched samples².

How should one assess the balance of matched data?

²See the appendix for a formal proof.

Evaluate matching: balance tests

The performance of matching depends on **the resulting balance of X_i** between matched samples².

How should one assess the balance of matched data?

- Ideally, compare the joint distribution of all X_i for the matched samples.

²See the appendix for a formal proof.

Evaluate matching: balance tests

The performance of matching depends on **the resulting balance of X_i** between matched samples².

How should one assess the balance of matched data?

- Ideally, compare the joint distribution of all X_i for the matched samples.
- In practice, this is impossible when X_i is high-dimensional and so check various lower-dimensional summaries.

²See the appendix for a formal proof.

Evaluate matching: balance tests

The performance of matching depends on **the resulting balance of X_i** between matched samples².

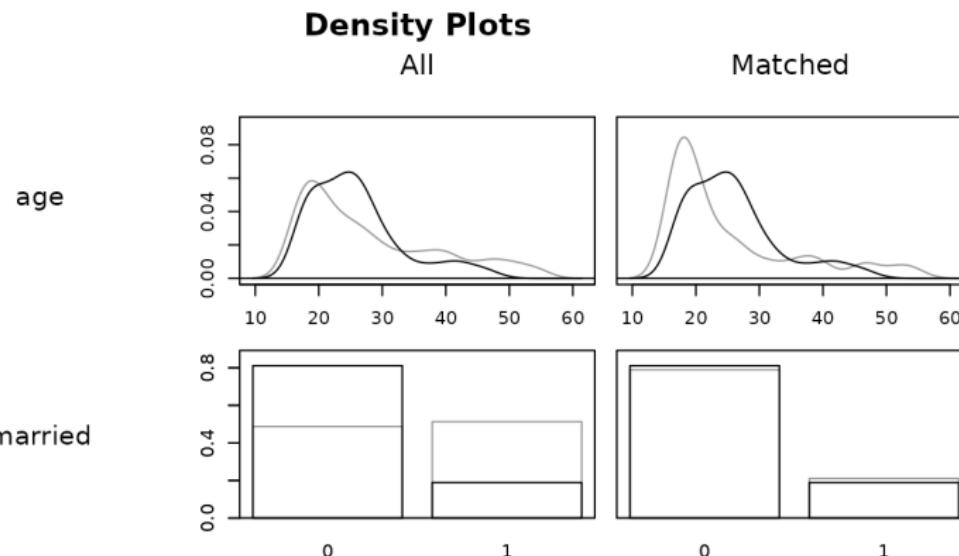
How should one assess the balance of matched data?

- Ideally, compare the joint distribution of all X_i for the matched samples.
- In practice, this is impossible when X_i is high-dimensional and so check various lower-dimensional summaries.
- Statistical tests are often used; plotting is also common in practice.

²See the appendix for a formal proof.

Evaluate matching: balance tests

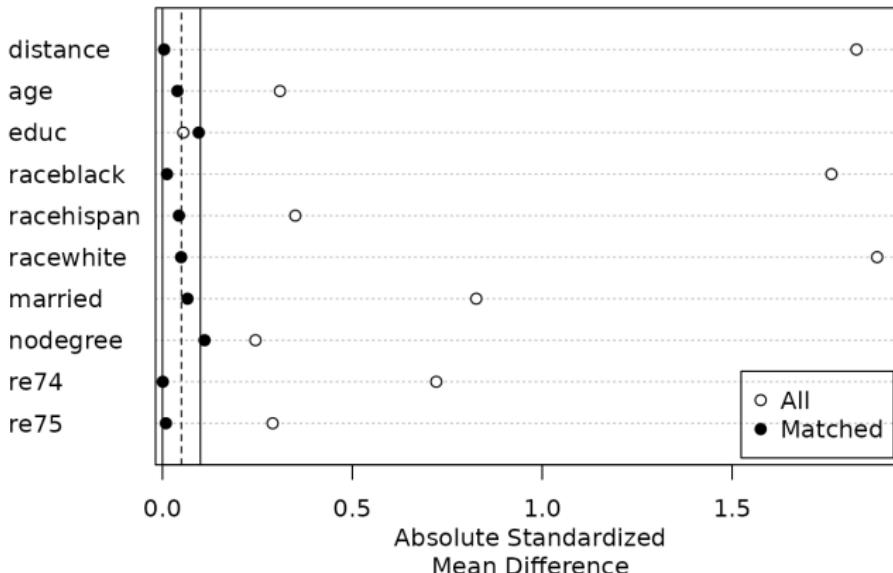
The performance of matching depends on **the resulting balance of X_i between matched samples³.**



³See the appendix for a formal proof.

Evaluate matching: balance tests

The performance of matching depends on **the resulting balance of X_i between matched samples⁴.**



⁴See the appendix for a formal proof.

Balance tests fallacy

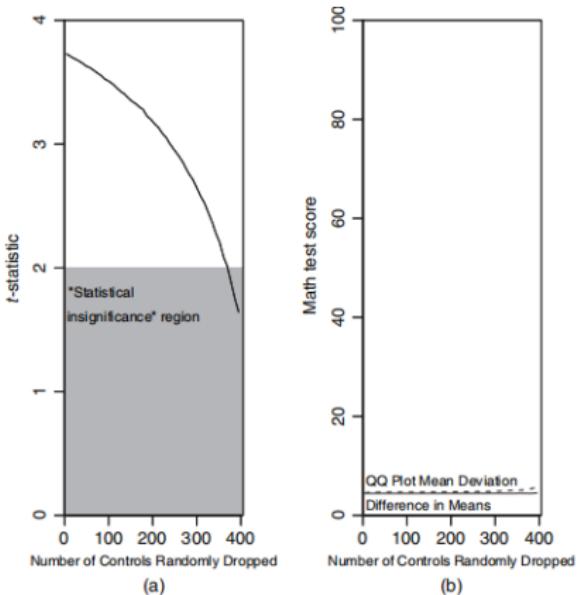


Fig. 1. Dangers in relying on t -statistics as a measure of balance (average value of a measure of balance when a given number of control units are randomly dropped from the data set (out of a total of 434)): with larger numbers of control units dropped (i.e. smaller numbers of control units in the resulting sample), the value of the t -statistic becomes closer to 0, falsely indicating improvements in balance, even though true balance does not vary systematically across the data sets (and efficiency declines); the difference in means and quantile-quantile plot mean deviation, which are given in (b), correctly indicate no change in bias as observations are randomly dropped

Source: Imai et al. (2008)

Matching: potential issues

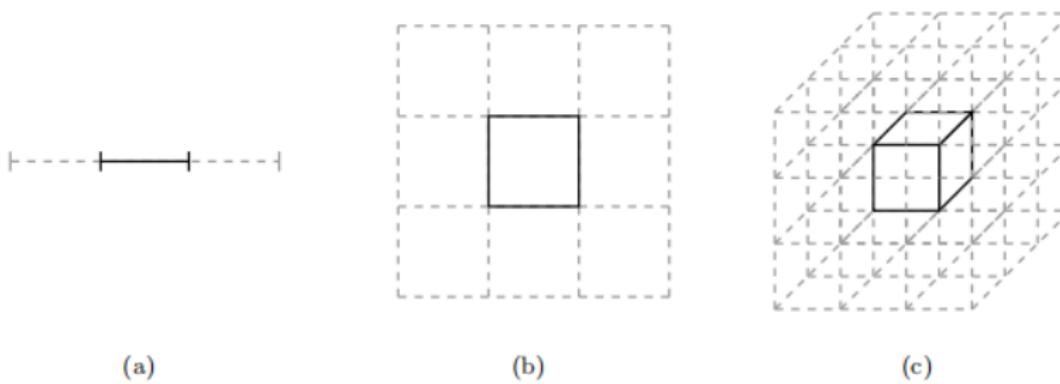
Some difficult decisions in distance matching

- ① With replacement or without replacement?
- ② Which distance metrics to use?
- ③ A single closest control unit or multiple closer control units?

In practice, you almost always need to use multiple matching methods to show the “**robustness**”.

Matching: potential issues

Difficult to find “good matches” with many continuous variables or variables with large variance...



The probability of the “focal” region decreases exponentially:

$$1/3 \rightarrow 1/9 \rightarrow 1/27.$$

Source: Betancourt (2017)

① When complete randomization is infeasible

② Various estimators under selection on observables

Subclassification

Matching

Weighting with propensity scores

③ Appendix

Propensity score to the rescue

Instead of working with X_i , why not a **statistic** of X_i ?

Propensity Score: $e(X_i) = P(D_i = 1 | X_i)$

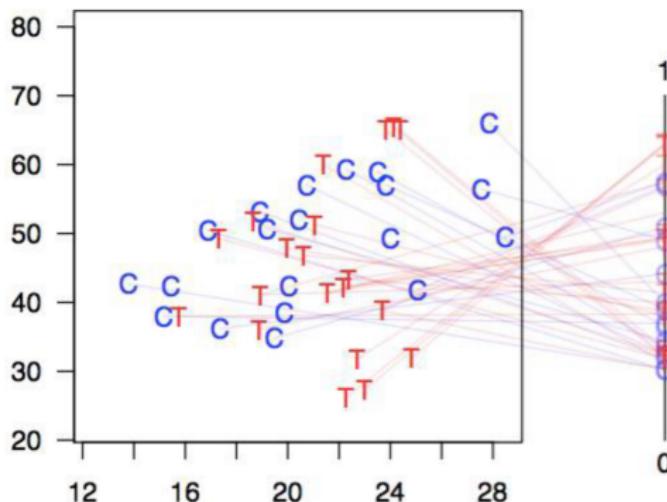
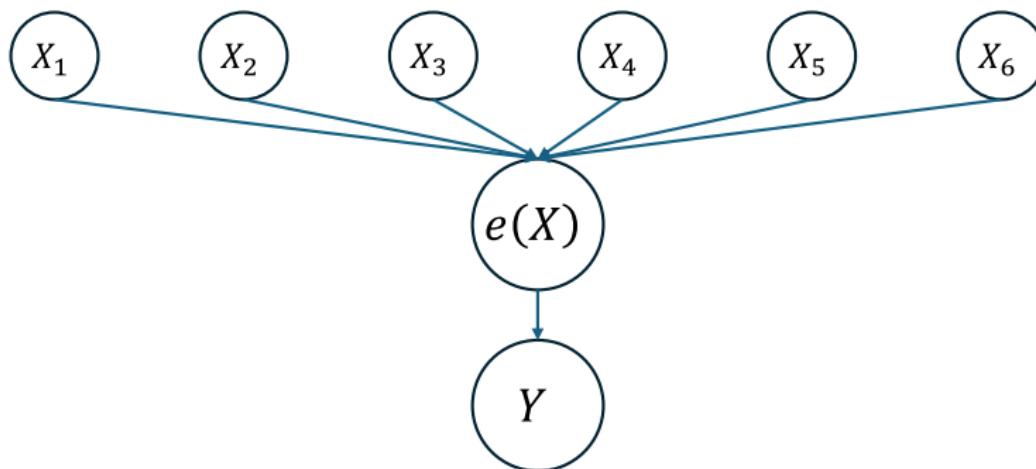


Figure 2: Propensity score “shrinks” X_i into a single index.

Identification results (Proof on slide 63)

Given **conditional unconfoundedness** and **common support**,

- ① The propensity score is a sufficient statistic of X_i , such that $D_i \perp X_i | e(X_i)$.
- ② The treatment is unconfounded given the propensity score: $D_i \perp Y_i^1, Y_i^0 | e(X_i)$.



Implications of the identification results

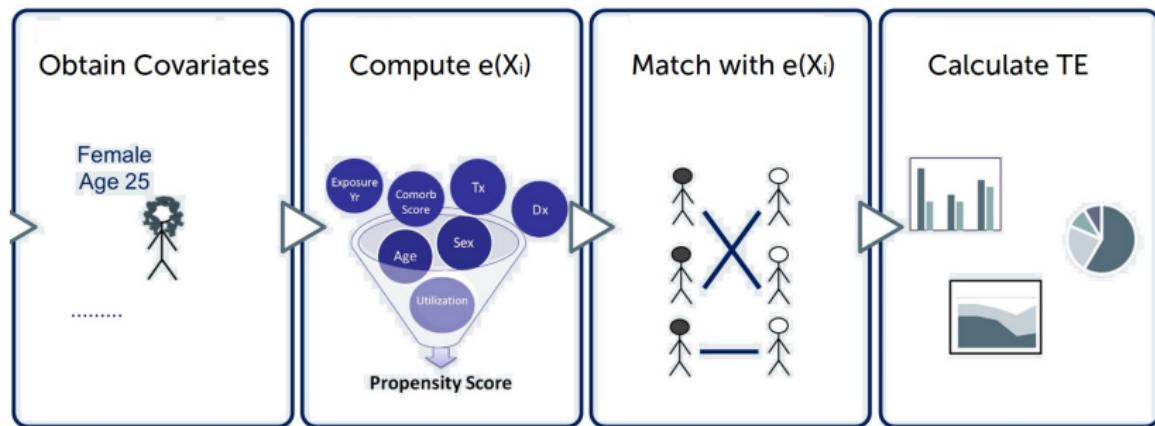
Adjustment of the treatment-control differences in $e(X_i)$ is sufficient to remove all biases from the differences in X_i .

$$\begin{aligned} E \left[Y_i^1 - Y_i^0 \mid D_i, e(X_i) \right] &\stackrel{\text{Unconfounded}}{=} E \left[Y_i^1 - Y_i^0 \mid e(X_i) \right] \\ &\stackrel{\text{Definition}}{=} E \left[Y_i^1 - Y_i^0 \mid P(D_i = 1 \mid X_i) \right] \end{aligned}$$

Suggests a two step procedure to estimate treatment effects:

- ① **Estimate the propensity score $\hat{e}(X_i)$.**
- ② **Estimate the expectations** of potential outcomes conditional on $\hat{e}(X_i)$.

Matching with propensity score



Illustrating propensity score matching

Weighting with propensity scores

Matching throws away samples to gain balance.
The “efficiency” is sacrificed for “precision.”

Question: Can we do better?

Observe two equations given propensity scores⁵:

$$\begin{cases} E \left(\frac{D_i \cdot Y_i^{\text{obs}}}{e(X_i)} \right) &= E [Y_i^1 | X_i] \\ E \left(\frac{(1-D_i) \cdot Y_i^{\text{obs}}}{1-e(X_i)} \right) &= E [Y_i^0 | X_i] \end{cases}$$

⁵The original construction is in Horvitz and Thompson (1952).

Weighting with propensity scores

Technical Proof:

$$\begin{aligned} E\left(\frac{D_i \cdot Y_i^{\text{obs}}}{e(X_i)}\right) &= E\left(\frac{D_i \cdot Y_i^1}{e(X_i)}\right) \\ &= E\left[E\left(\frac{(D_i = 1) \cdot Y_i^1}{e(X_i)} \middle| X_i\right)\right] \\ &= \frac{E_D[D_i = 1 | X_i] E[Y_i^1 | X_i]}{e(X_i)} \\ &= E[Y_i^1 | X_i] \end{aligned}$$

Similar proof holds for the second equality.

The weighting estimator

Based on the equations, we can construct two estimators:

$$E \left[\widehat{Y_i^1} \mid X_i \right] = \frac{1}{N} \sum_{i=1}^N \frac{D_i \cdot Y_i^{\text{obs}}}{e(X_i)}$$

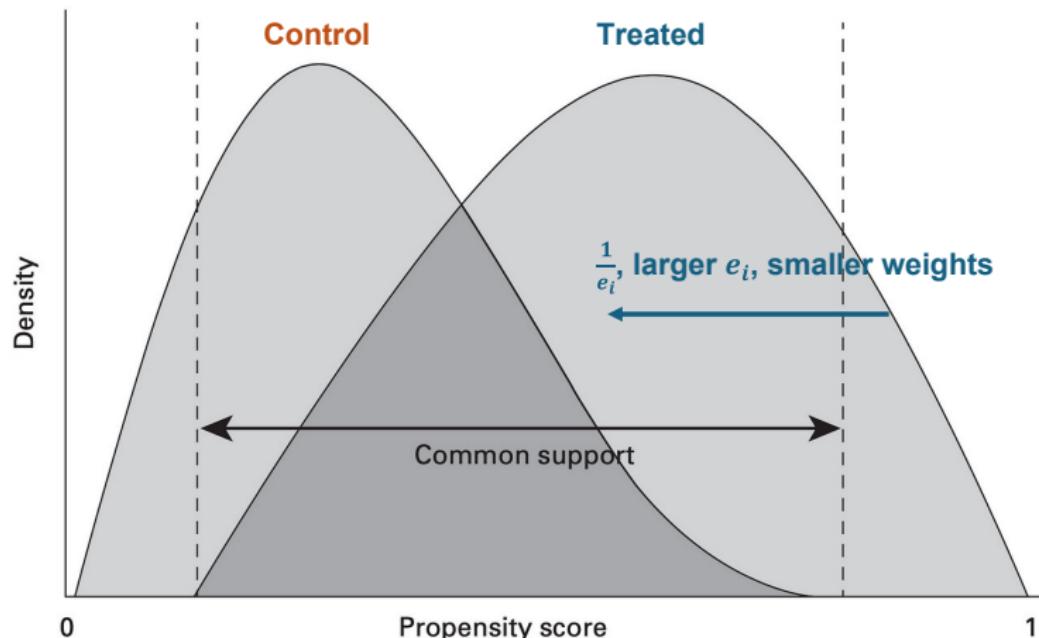
$$E \left[\widehat{Y_i^0} \mid X_i \right] = \frac{1}{N} \sum_{i=1}^N \frac{(1 - D_i) \cdot Y_i^{\text{obs}}}{1 - e(X_i)}$$

The weighting estimator of ATE is defined as,

$$\begin{aligned}\hat{\tau} &= E \left[\widehat{Y_i^1} \mid X_i \right] - E \left[\widehat{Y_i^0} \mid X_i \right] \\ &= \frac{1}{N} \sum_{i=1}^N \left(\frac{D_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{(1 - D_i) \cdot Y_i^{\text{obs}}}{1 - e(X_i)} \right)\end{aligned}$$

Intuitions behind weighting

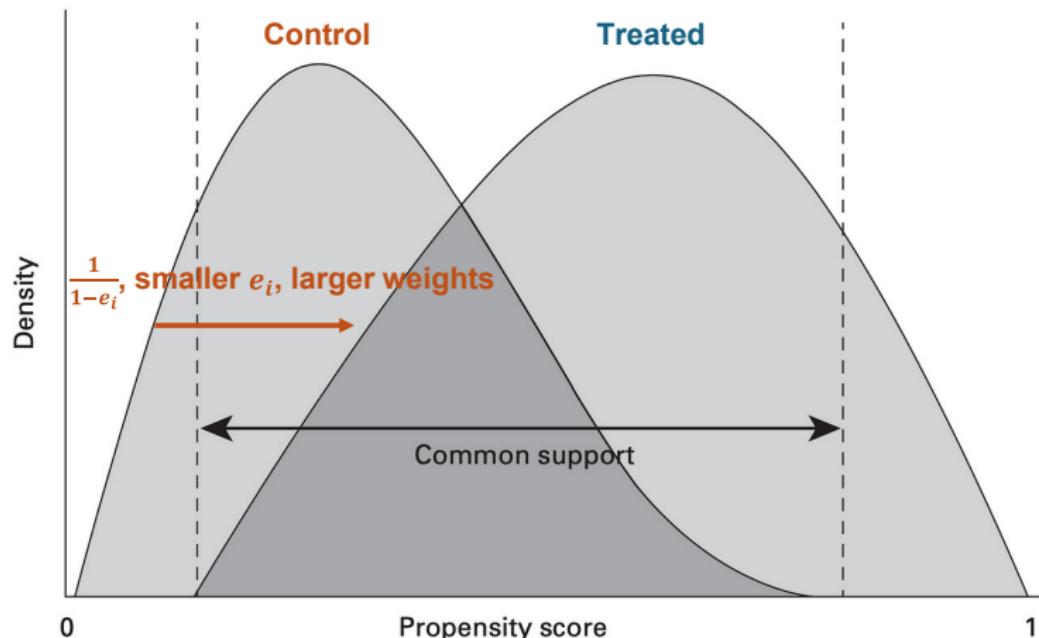
The weights w_i is $\frac{1}{e(X_i)}$ for the treated and $\frac{1}{1-e(X_i)}$ for the control.



“Expanding” the common support

Intuitions behind weighting

The weights w_i is $\frac{1}{e(X_i)}$ for the treated and $\frac{1}{1-e(X_i)}$ for the control.



“Expanding” the common support

The weighted regression estimation

The weighted regression procedure

- ① Estimate the propensity scores with $P(D_i = 1 | X_i)$, possibly with different methods and specifications.
- ② Plug in estimated propensity score $\hat{e}(X_i)$ into a weighted regression.

$$Y_i = \alpha + \tau D_i + X_i \beta + \varepsilon_i \text{ with } w_i = \begin{cases} \frac{1}{\hat{e}(X_i)} & \text{if } D_i = 1 \\ \frac{1}{1 - \hat{e}(X_i)} & \text{if } D_i = 0 \end{cases}$$

The weighted regression for ATT and ATC

Some procedure, but with different weights:

- The weights (odds ratio) for ATT estimation is:

$$w_i = \begin{cases} 1 & \text{if } D_i = 1 \\ \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)} & \text{if } D_i = 0 \end{cases}$$

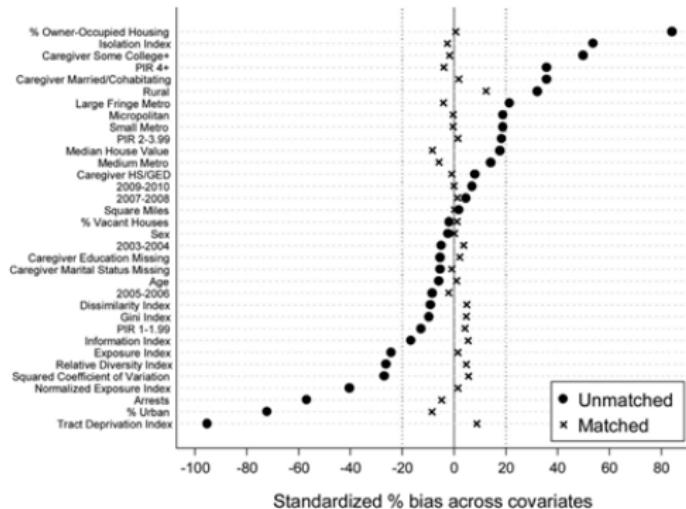
- The weights for ATC estimation is similar:

$$w_i = \begin{cases} \frac{1 - \hat{e}(X_i)}{\hat{e}(X_i)} & \text{if } D_i = 1 \\ 1 & \text{if } D_i = 0 \end{cases}$$

What is a good estimation of propensity score?

Balancing condition: samples after weights should show similar distributions of X_i

$$E \left[\frac{D_i \cdot X_i}{e(X_i)} - \frac{(1 - D_i) \cdot X_i}{1 - e(X_i)} \right] = 0$$



Notes on propensity scores

- The true propensity score is unknown (by definition), and estimated by a model.
- Model misspecification leads to biased treatment effects estimation.
- In practice, *ad hoc* robustness checks across different specs and models.
- In recent years, some machine learning models (non-parametric) prove to be useful.

- ① When complete randomization is infeasible
- ② Various estimators under selection on observables
- ③ Appendix

Other ML methods in matching: see references

Genetic matching:

- Diamond, A., & Sekhon, J. S. (2013)

Dynamic almost exact matching with replacement:

- Dieng et al. (2019)

Propensity score matching with random forest:

- Krief and Diaz-Ordaz (2019), Ferri-Garcia and Rueda (2020)

Proof: matching is about covariate balancing

The conditional ATE or CATE is: $\tau(X_i) = E[Y_i^1 - Y_i^0 | X_i]$.

Under the assumptions of conditionally unconfounded, individualistic and probabilistic assignment, the variance for CATE is⁶,

$$\text{Var}(\tau(X_i)) = E \left[\frac{\sigma_c^2(X_i)}{1 - e(X_i)} + \frac{\sigma_t^2(X_i)}{e(X_i)} + (\tau(X_i) - \tau)^2 \right]$$

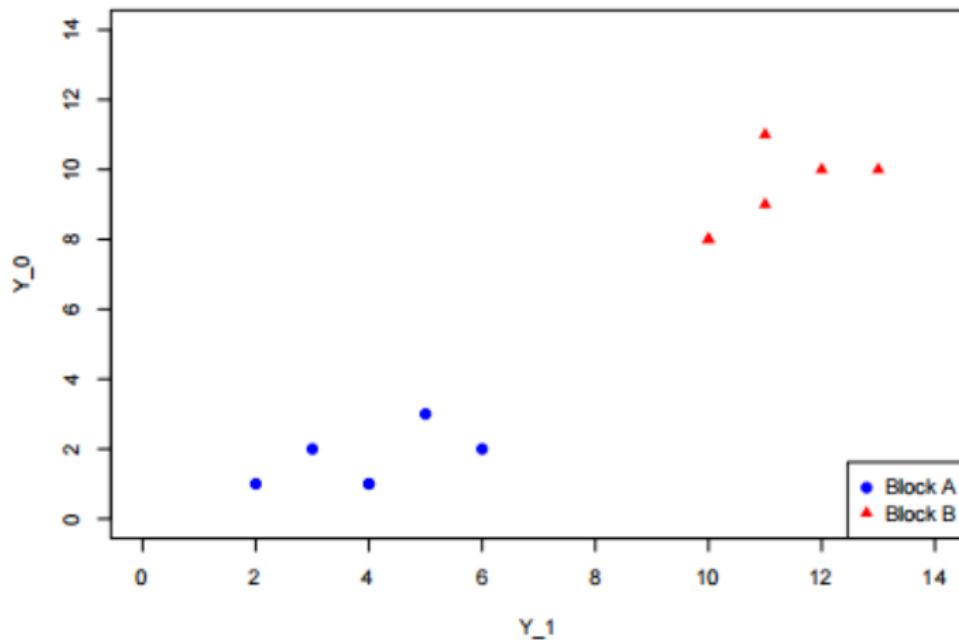
$e(X_i)$ the propensity to be treated, and τ the unconditional ATE.

The third term $(\tau(X_i) - \tau)^2$ goes to zero if the treatment effect is constant for the treated and the control.

This requires X_i has the same distributions in both groups, i.e., X_i is balanced.

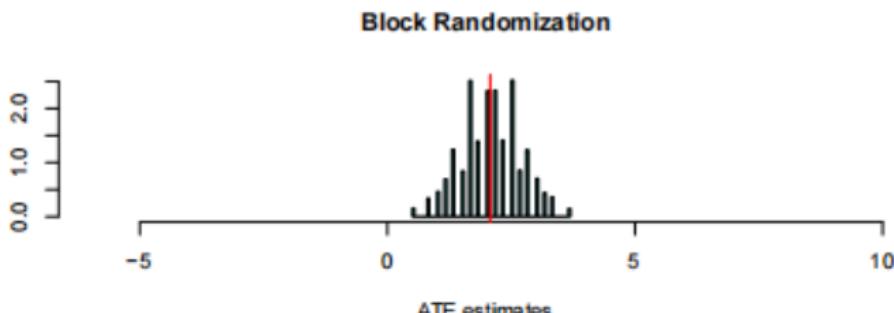
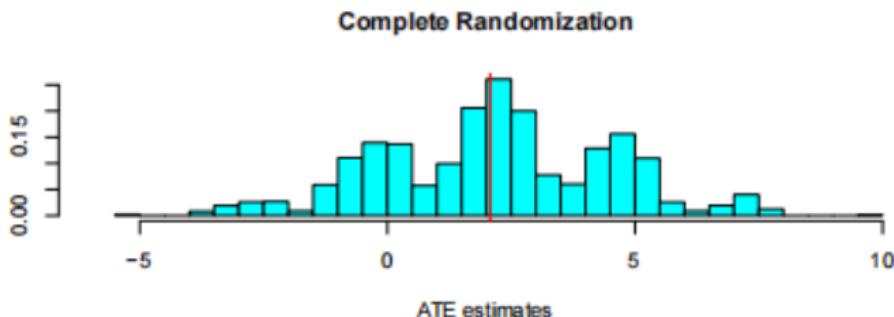
⁶Please see Imbens and Rubin (2015), p. 269 for the derivation.

Blocking is efficient: Intuition



Potential outcomes across blocks

Blocking is efficient: Intuition



Sampling distribution of ATE estimates

Blocking: summary

What to block on?

- **“Block what you can/need, randomize what you can’t or don’t need.”**
- Variables linked to potential outcomes.
- Variables desired for subgroup analysis.

How to block?

- Stratification
- Pairwise assignment...

Proof: the sufficiency of propensity score

Here, we prove that the propensity score $e(X_i)$ is a sufficient statistic of X_i or:

$$D_i \perp X_i \mid e(X_i) \text{ or } P(D_i = 1 \mid X_i, e(X_i)) = P(D_i = 1 \mid e(X_i))$$

First, for the left-hand side:

$$P(D_i = 1 \mid X_i, e(X_i)) = P(D_i = 1 \mid X_i) = e(X_i)$$

Second, for the right-hand side:

$$\begin{aligned} P(D_i = 1 \mid e(X_i)) &= E[D_i \mid e(X_i)] \\ &= E[E[D_i \mid X_i, e(X_i)] \mid e(X_i)] \\ &= E[e(X_i) \mid e(X_i)] = e(X_i) \end{aligned}$$

The second equality is the iterated expectations and the third equality is proved above.

Proof: the conditional unconfoundedness of the propensity score

We want to prove that $D_i \perp Y_i^0, Y_i^1 \mid e(X_i)$, or

$$P(D_i = 1 \mid Y_i^0, Y_i^1, e(X_i)) = P(D_i = 1 \mid e(X_i))$$

For the left-hand side, we have:

$$\begin{aligned} P(D_i = 1 \mid Y_i^0, Y_i^1, e(X_i)) &= E[D_i \mid Y_i^0, Y_i^1, e(X_i)] \\ &= E[E[D_i \mid Y_i^0, Y_i^1, X_i, e(X_i)] \mid Y_i^0, Y_i^1, e(X_i)] \end{aligned}$$

By conditional unconfoundedness, $E[D_i \mid Y_i^0, Y_i^1, X_i, e(X_i)] = E[D_i \mid X_i, e(X_i)]$.

By the sufficiency of the propensity score, $E[D_i \mid X_i, e(X_i)] = E[D_i \mid e(X_i)]$. If we submit it to the equation above, we have:

$$E[E[D_i \mid e(X_i)] \mid Y_i^0, Y_i^1, e(X_i)] = E[D_i \mid e(X_i)] = P(D_i = 1 \mid e(X_i))$$

[Back to Slide 44](#)

References I

-  Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint arXiv:1701.02434.
-  Kosuke Imai, Gary King, and Elizabeth Stuart. 2008. "Misunderstandings Among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A*, 171, part 2, Pp. 481–502.
-  King, G., Lucas, C., & Nielsen, R. A. (2017). The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*, 61(2), 473-489.

References II

- ❖ Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3), 932-945.
- ❖ Dieng, A., Liu, Y., Roy, S., Rudin, C. & Volfovsky, A. (2019). Almost-Exact Matching with Replacement for Causal Inference. Submitted to arXiv.org Statistics / Machine Learning
- ❖ Kreif, N. & DiazOrdaz, K. (2019). Machine Learning in Policy Evaluation: New Tools for Causal Inference. Submitted to arXiv.org Statistics / Machine Learning

References III

-  Ferri-García, R., & Rueda, M. D. M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PloS one*, 15(4), e0231500.
-  Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260), 663-685.