Lecture 4: When Complete Randomization is Infeasible

Matching and Weighting in Experimental Designs

Xi Chen

Rotterdam School of Management
Erasmus University Rotterdam

May 31, 2023

Outline

RSM Erasmus

Outline

RSM Erasmus

Outline

RSM Erasmus

Complete randomization is 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 20 consumers:

Randomly assign 20 consumers

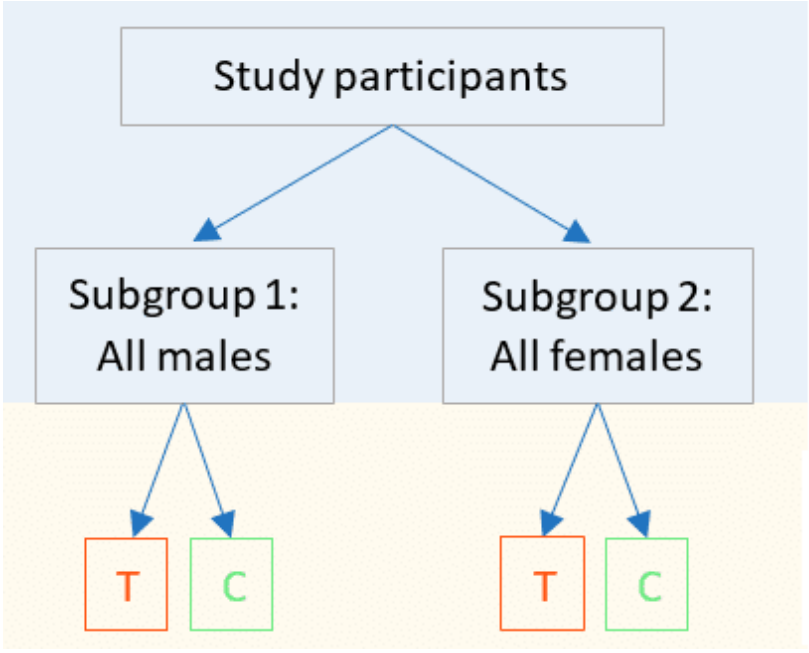
```
people <- rep(c("female","male"),10)
unlist(rerun(10,
sum(sample(people,10,replace = F)=="female"))
[1] 6 4 4 4 5 6 5 7 5 3
```

RSM Erasmus

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).



The treatment and control groups are forced to be equivalent regarding gender

Blocking is efficient: An example

store	block	All Stores Y_1	All Stores Y_0	Block A Y_1	Block A Y_0	Block B Y_1	Block B 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:

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

RSM Erasmus

Blocking is efficient: An example

store	block	All Stores Y_1	All Stores Y_0	Block A Y_1	Block A Y_0	Block B Y_1	Block B 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:

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

RSM Erasmus

Irish energy conservation trial

Tariff	Bi-monthly bill and energy usage statement	Monthly Bill, and energy usage statement	Bi-monthly bill, energy usage statement and electricity Monitor	Bi-monthly bill, energy usage statement plus Overall Load Reduction	Total
Tariff A	342	342	342	342	1,368
Tariff B	127	129	127	128	511
Tariff C	342	342	343	343	1,370
Tariff D	127	129	126	127	509

The treatment conditions in Ireland energy conservation trial

RSM Erasmus

Irish energy conservation trial

Composition of Eigen vectors	Eigen Vectors				
Factors	V1	V2	V3	V4	V5
Peak band	-0.178	0.577	-0.065	0.085	-0.087
Night band	0.103	-0.526	0.101	0.142	0.153
Overall Weekly Usage Variance	-0.039	0.1	0.133	0.164	0.485
Overall Peak Usage Variance	-0.152	0.53	0.01	0.193	0.207
Number in house	-0.395	-0.003	-0.11	-0.288	-0.276
# of bedrooms	-0.312	-0.119	-0.078	-0.247	-0.197
Internet access	-0.414	-0.119	0.006	0.095	-0.078
Income band	-0.463	-0.12	0.043	0.039	0.094
Education classification	-0.361	-0.178	0.071	0.196	0.288
Employment status	-0.401	-0.087	0.036	0.1	0.125
Wet	0.013	0.006	0.028	0.582	-0.304
Electronics	-0.004	-0.092	0.044	0.568	-0.398
Energy reduction engagement band	0.022	-0.047	-0.646	0.08	0.103

Re-randomization with eigenvectors from a factor analysis

RSM Erasmus

Yelp Elite

yelp elite

The 2014 Yelp Providence Elite Squad

Hey there, Phil R.!

Elite badges are so hot right now... and let's be honest, so is your writing! We received a nomination and agree that you embody the spirit of Yelp with your enthusiasm, positivity, constructive honesty and useful funny coolness. So you can consider this your official invite to be on the Providence Elite Squad!

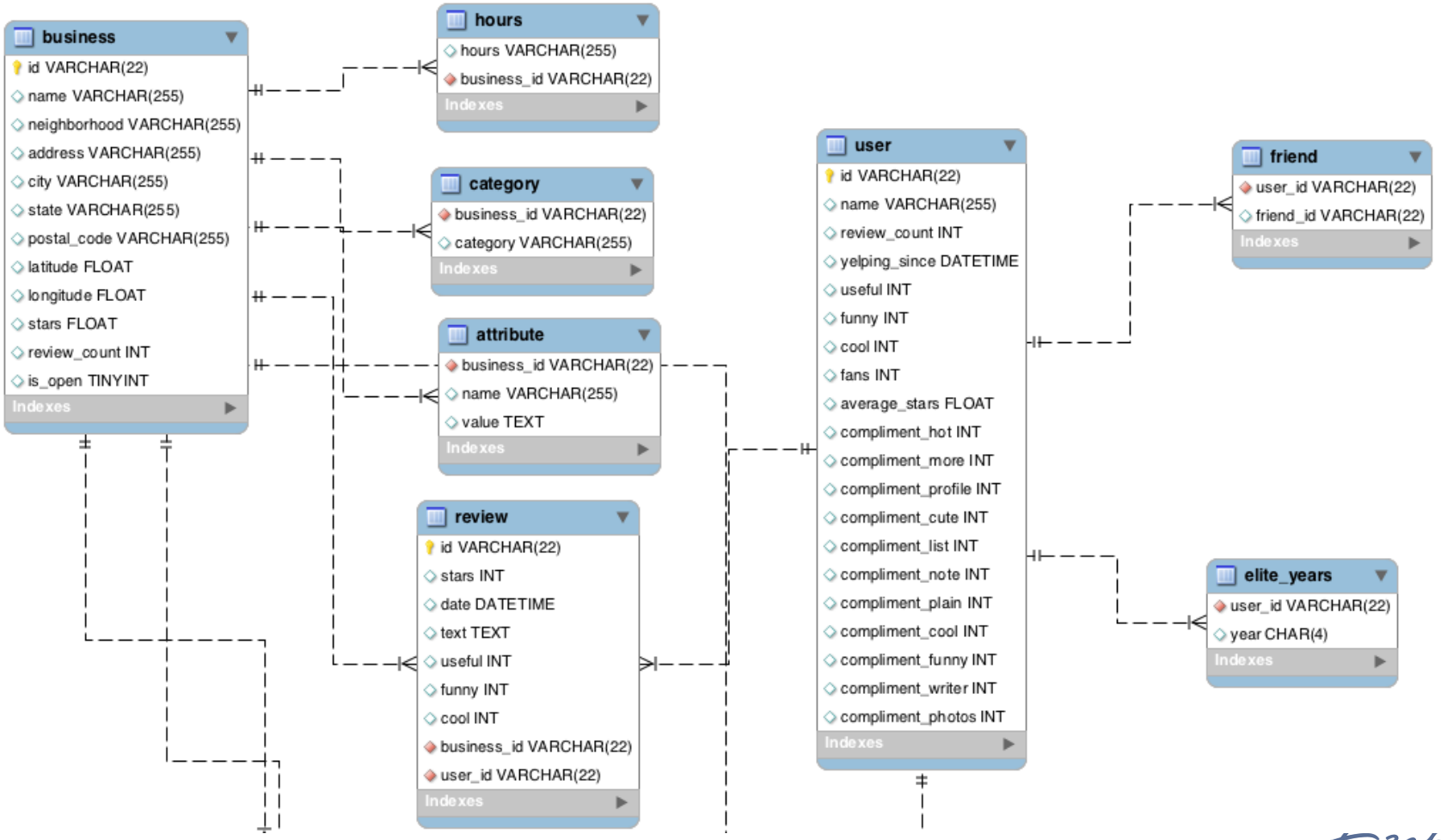
Thunderous applause

As an Elite, you'll be an ambassador both on and offline. We do ask that you keep on being a stellar yelper with rich and consistently contributed reviews. A great elite also compliments other folks, votes on reviews, participates respectfully in Talk and welcomes new members to the site. As an elite, you'll be invited to attend exclusive parties and events that we throw around town and be privy to hot giveaways.

What is Yelp Elite?

RSM Erasmus

Yelp Elite

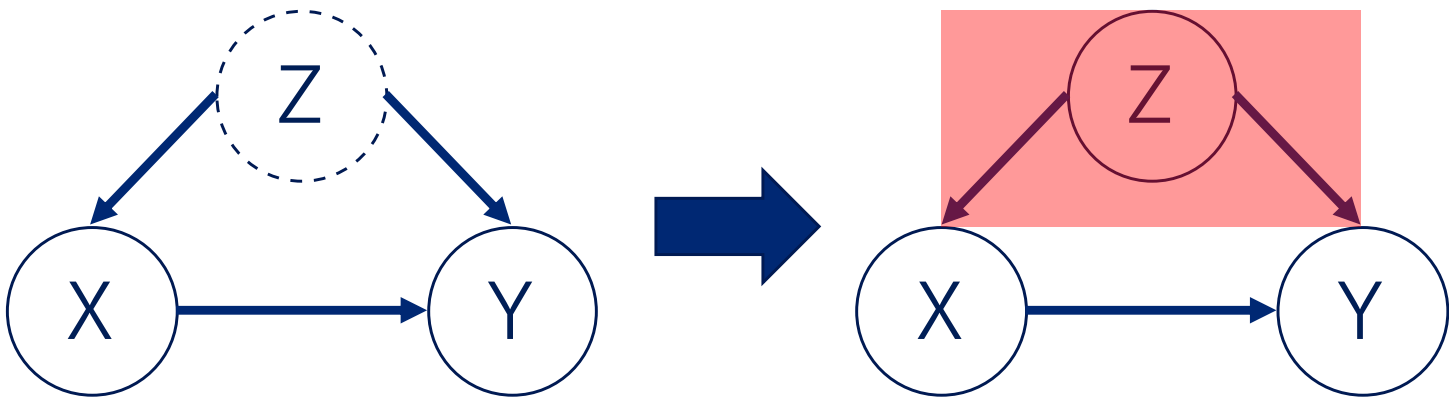


Almost all information is know, but not how Yelp selects Elites.

Identification under "selection on observables"

Identification Assumption

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



It's essentially a "conditioning" identification strategy

RSM Erasmus

Identification results: ATE

Under the conditional unconfoundedness:

$$E\left[Y_i^1 \mid X_i, D_i = 1\right] = E\left[Y_i^1 \mid X_i\right]$$
$$E\left[Y_i^0 \mid X_i, D_i = 0\right] = E\left[Y_i^0 \mid X_i\right]$$

With common support assumption:

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

RSM Erasmus

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 also a weaker version of common support $p_i\left(D_i = 1 \mid X_i\right) < 1$, ATT is identified.

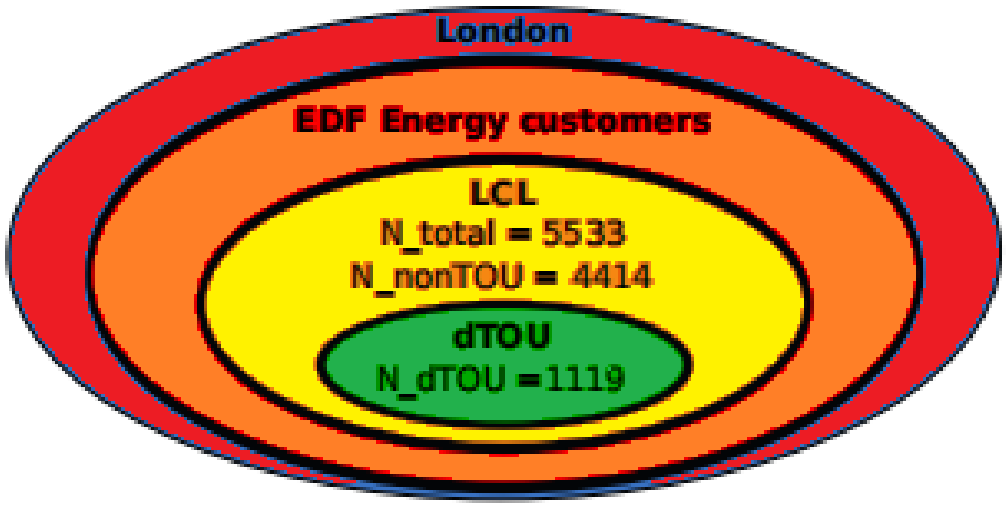
$$E\left[Y_i^0 \mid X_i, D_i = 1\right] = E\left[Y_i^0 \mid X_i\right]$$

$E\left[Y_i^1 \mid X_i, D_i = 1\right]$, directly from data

RSM Erasmus

An example of ATT: "Low Carbon London" trial

Venn diagram illustrating sample selection in LCL trial



The control group is randomly drawn from the super-population, and thus ATT is identified.

RSM Erasmus

Outline

RSM Erasmus

Outline

RSM Erasmus

Subclassification

To apply a technique as in blocking by creating "blocks" based on X_i:

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

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

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

RSM Erasmus

Subclassification: An example

Usage	**Average Purchase**		**Diff.**	**# Ad Viewers**	**# Units**
	With Ads	**No Ads**			
X_i	$\overline{Y}_1^k$	$\overline{Y}_0^k$		N_k^1	N_k
Heavy	€28	€24	€4	3	10
Light	€22	€16	€6	7	10
Total				10	20

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

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

RSM Erasmus

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

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

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

RSM Erasmus

Outline

RSM Erasmus

Matching: the basic idea

With continuous control variables X_i, we can estimate $\widehat{\tau}_{\text{ATT}}$ by "imputing" the missing potential outcomes of each treated unit with the "closest" control units:

$$\widehat{\tau}_{\text{ATT}} = \frac{1}{N_1} \sum_{i:D_i=1} \left(Y_i^1 - \overline{Y_j^0} \right)$$

Where $\overline{Y_j^0}$ is the average outcome of those control units that are "closest" to unit i ($j \in M_i$).

RSM Erasmus

Matching: a general procedure

Step 1 of 2:

1. Selecting a matching method.
2. Obtaining the matched samples.
3. Evaluating the matching performances.

Step 2 of 2:

1. Calculating the treatment effects.

RSM Erasmus

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 $\widehat{\tau}_{\text{ATT}} = \frac{1}{N_1} \sum_{i:D_i=1} \left(Y_i^1 - \overline{Y_j^0} \right)$?
Match and plug in potential outcomes.

RSM Erasmus

Matching: a simple example

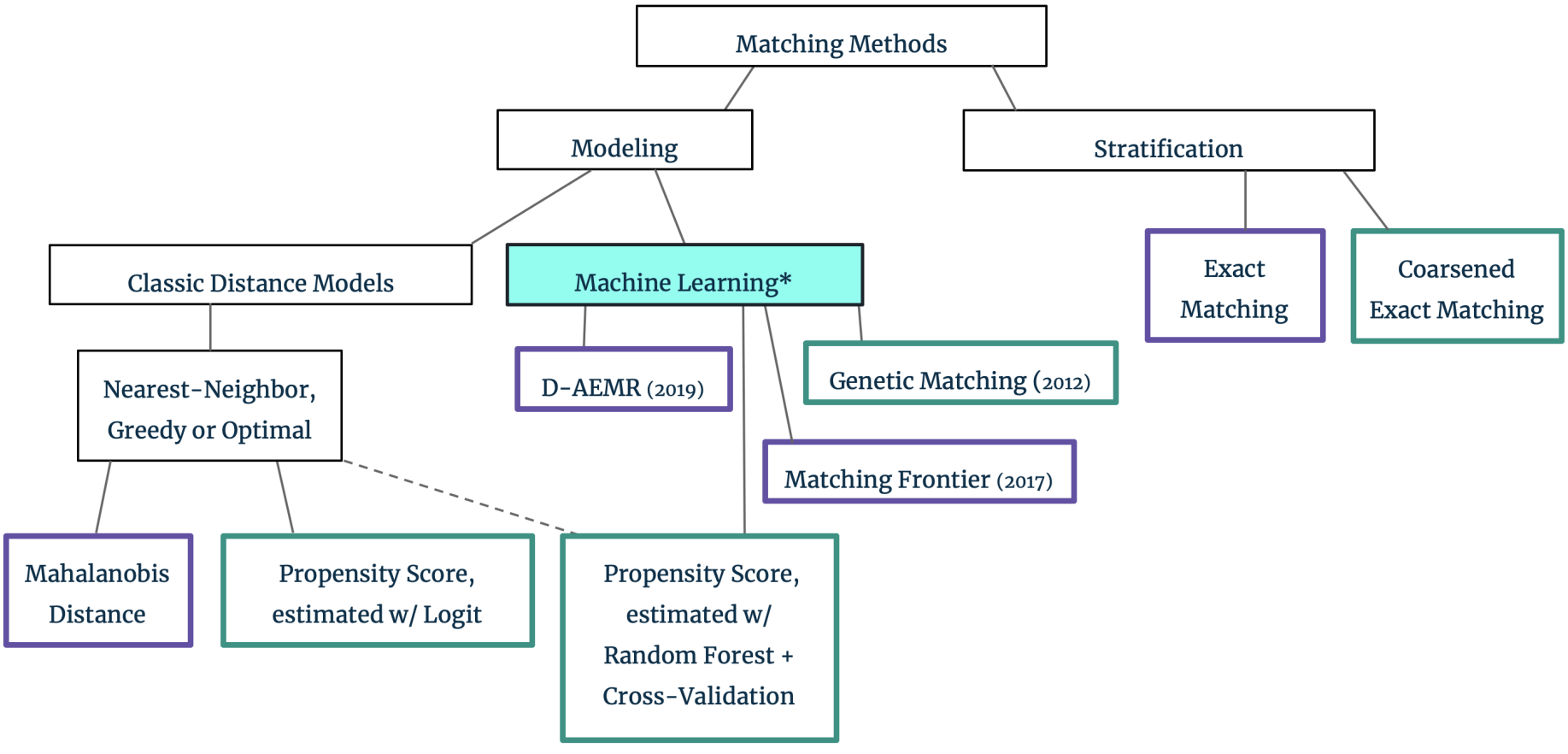
Unit	Potential Outcome		Treatment	Variables
	Under Treatment	Under Control		
i	Y_i^1	Y_i^0	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

What is $\widehat{\tau}_{\text{ATT}} = \frac{1}{N_1} \sum_{i:D_i=1} \left(Y_i^1 - \overline{Y_j^0} \right)$[1]?

$$\widehat{\tau}_{\text{ATT}} = \frac{1}{3} \left((6-9) + (0-1) + (9-1) \right) = -3.7$$

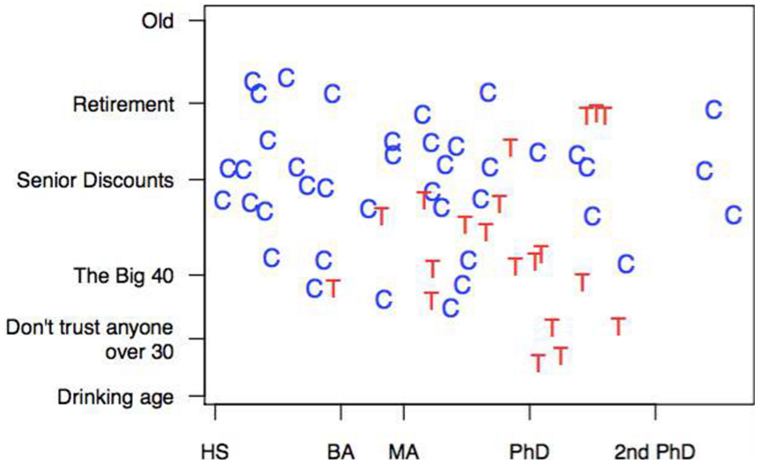
[1] Note: ATC (ATE on the control units) can be calculated similarly.

A classification of matching methods

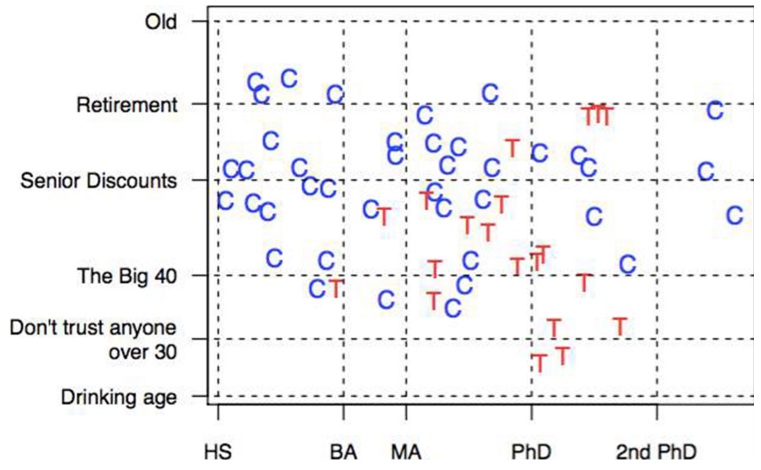


Stratification: coarsened exact matching

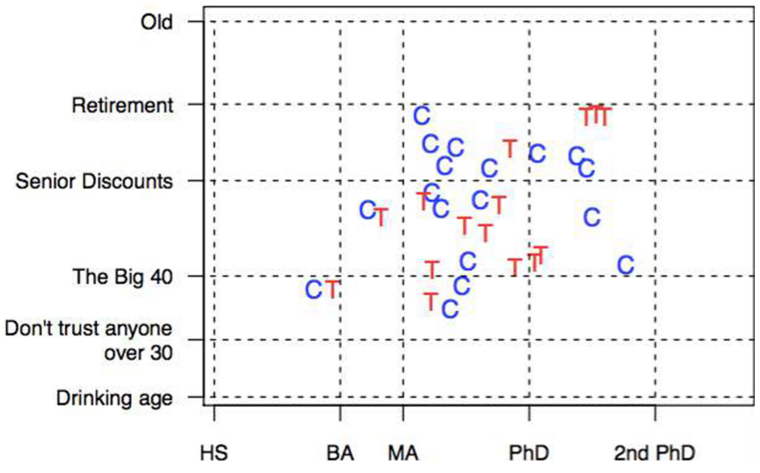
Step 1: to check overlaps



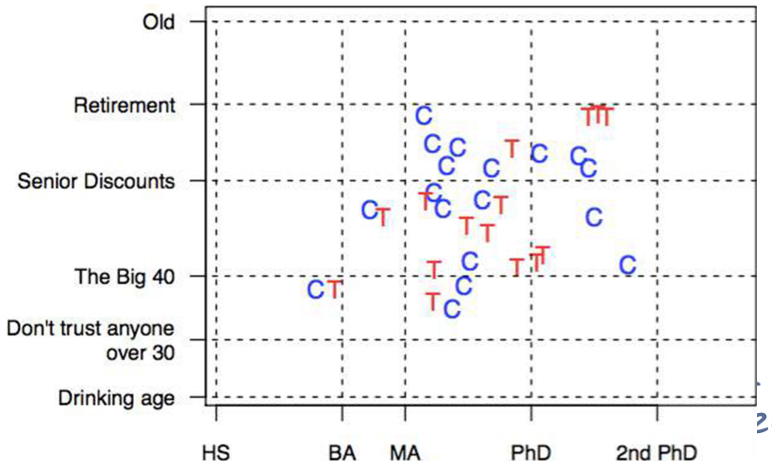
Step 2: to create bins



Step 3: to prune control samples



Step 4: to calculate treatment effects



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:

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

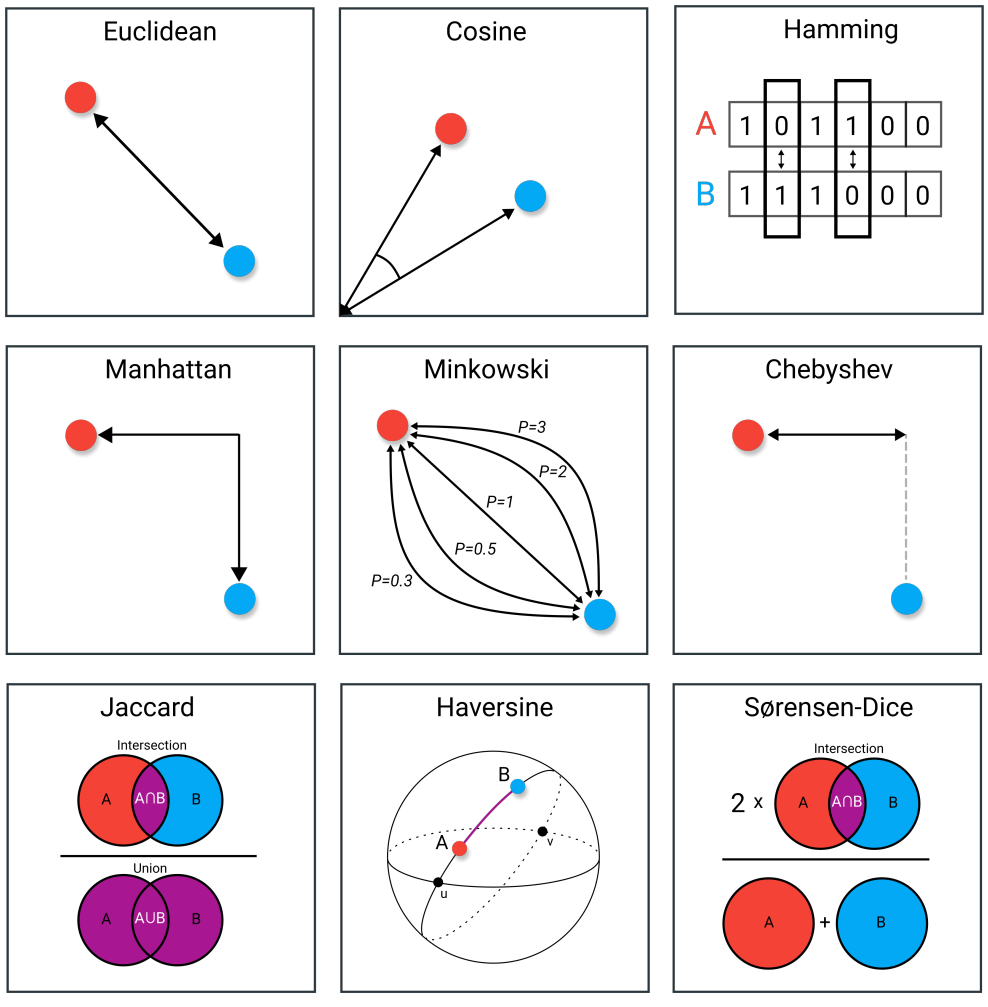
$$\text{Euclidean distance: } \sqrt{(X_i - X_j)' (X_i - X_j)}$$

For an "exact match", the distance metrics are zeros.

RSM Erasmus

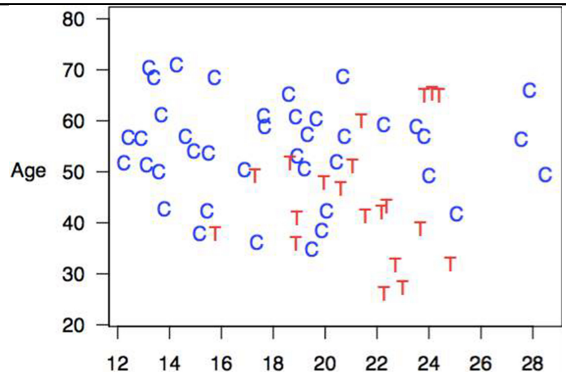
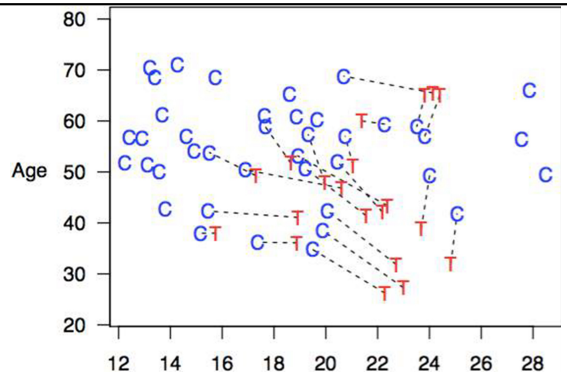
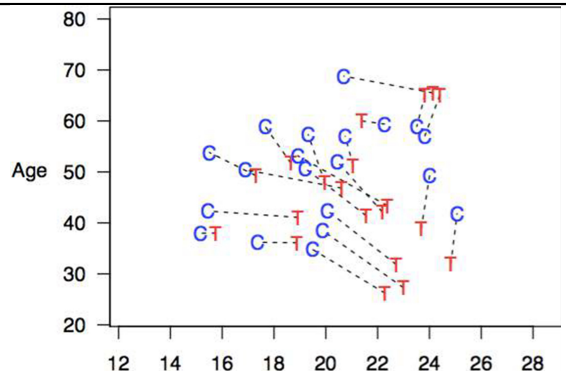
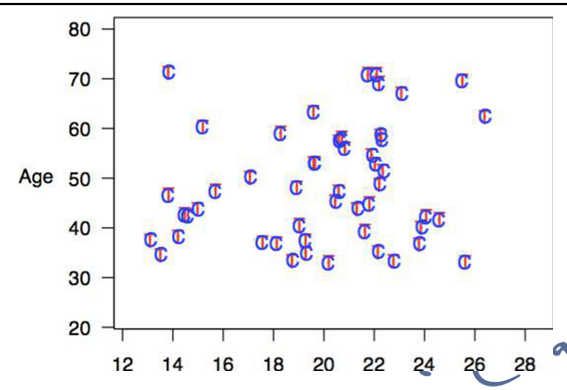
Classic distance models

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

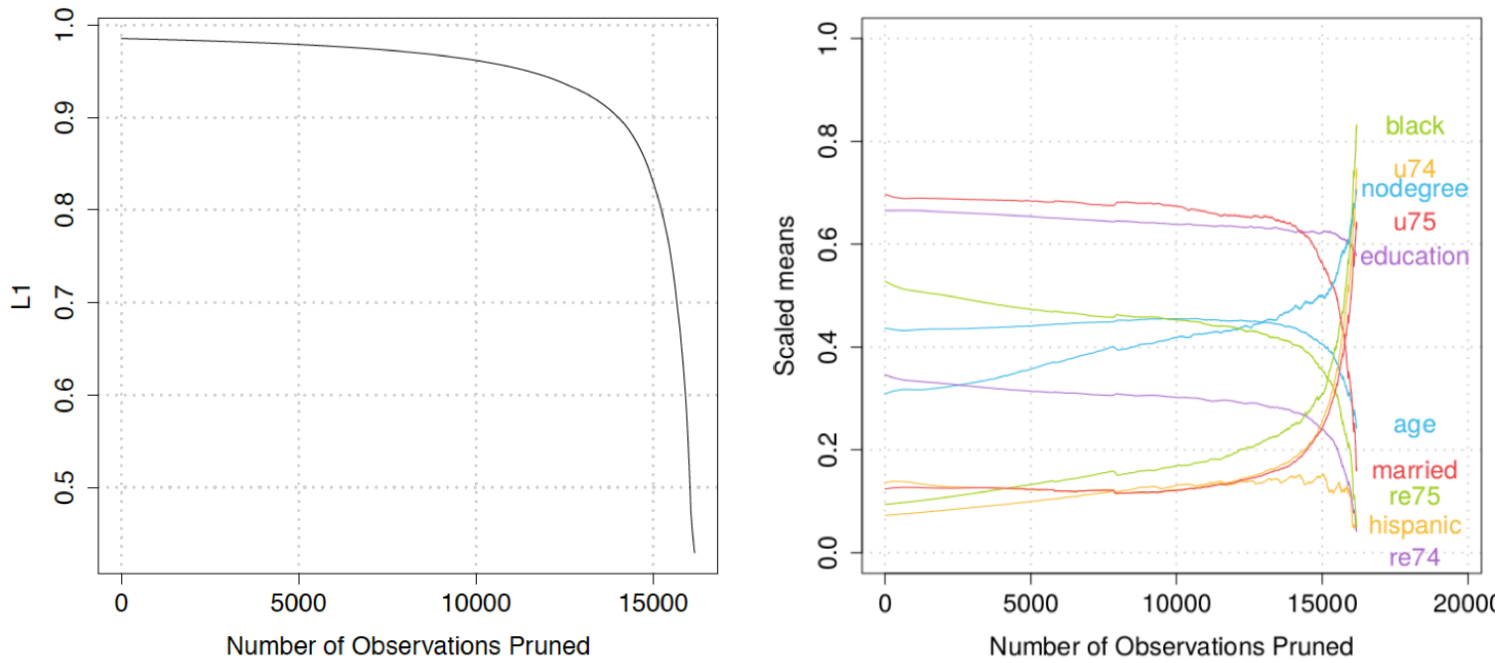


Examples of Distance Metrics

Classic distance models

Step 1: to calculate the distances	Step 2: to find the matches **Methods: greedy vs. optimal**
	
Step 3: to prune the unmatched control units	**Step 4: to obtain the matched samples**
	

Machine learning methods: 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.

Evaluating matching: balance tests

The performance of matching depends on the resulting balance of X_i between matched samples.
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.
- Instead, check various lower-dimensional summaries.

Balance tests

t-test for different in means of X_i
Other test statistics, e.g., F-test, χ^2, K-S test...
Good balance $\rightarrow$ Insignificant tests

RSM Erasmus

Balance tests fallacy

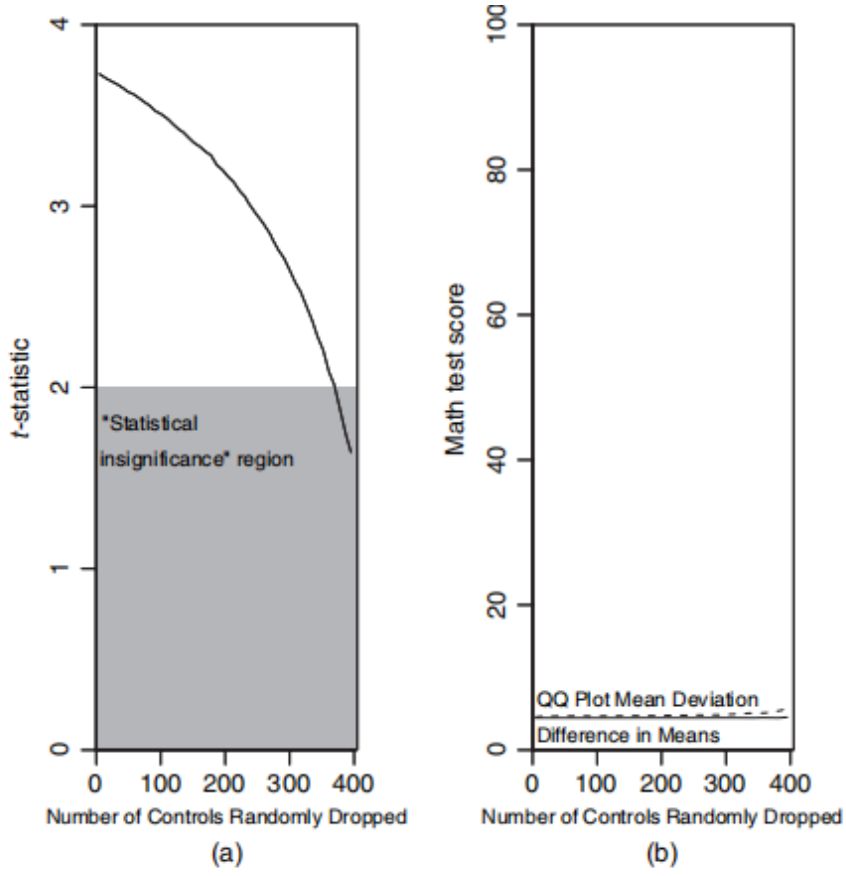


Fig. 1. Dangers in relying on *t*-statistics as a measure of balance (average value of a measure of balance when a given number of control units are randomly dropped from the data set (out of a total of 434)): with larger numbers of control units dropped (i.e. smaller numbers of control units in the 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."

RSM Erasmus

Matching: potential issues

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

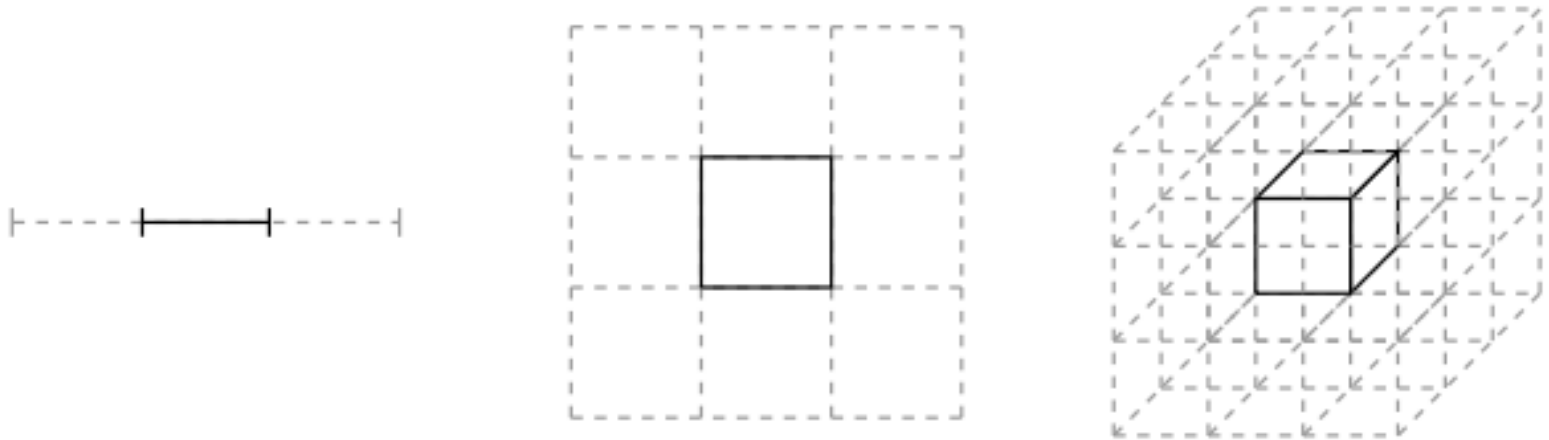


FIG 1. *To understand how the distribution of volume behaves with increasing dimension we can consider a rectangular partitioning centered around a distinguished point, such as the mode. (a) In one dimension the relative weight of the center partition is* $1/3$*, (b) in two dimensions it is* $1/9$*, (c) and in three dimensions it is only* $1/27$*. Very quickly the volume in the center partition becomes negligible compared to the neighboring volume.*

Impossible to find an "optimal" region with dimensionality increases

Source: Betancourt (2017)

Outline

RSM Erasmus

Propensity score to the rescue

Instead of working directly with X_i, why not a sufficient statistic for X_i?

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

Given conditional unconfoundedness and common support, we have:

Identification results

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

RSM Erasmus

Implications the identification results

If we know X_i, adjustment of the treatment-control differences in $e(X_i)$ is sufficient to remove all biases associated with the differences in X_i.

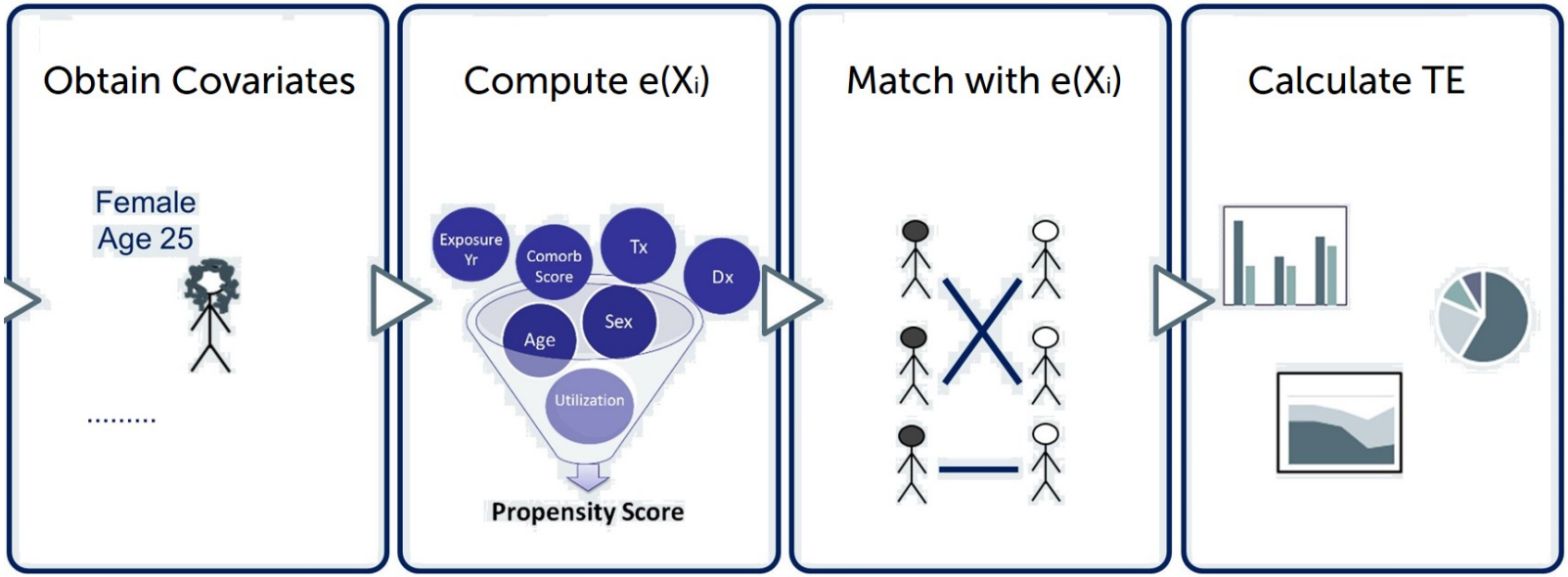
$$E\left[Y_i^1 - Y_i^0 \mid D_i, e(X_i)\right] \overset{\text{Unconfouded}}{=} E\left[Y_i^1 - Y_i^0 \mid e(X_i)\right]$$
$$\overset{\text{Definition}}{=} E\left[Y_i^1 - Y_i^0 \mid P(D_i = 1 \mid X_i)\right]$$

Suggests a two step procedure to estimate treatment effects:

1. Estimate the propensity score $e(X_i)$, e.g., logistic reg, or ML methods etc.
2. Estimate the conditional expectations of potential outcomes, given $\hat{e}(X_i)$.

RSM Erasmus

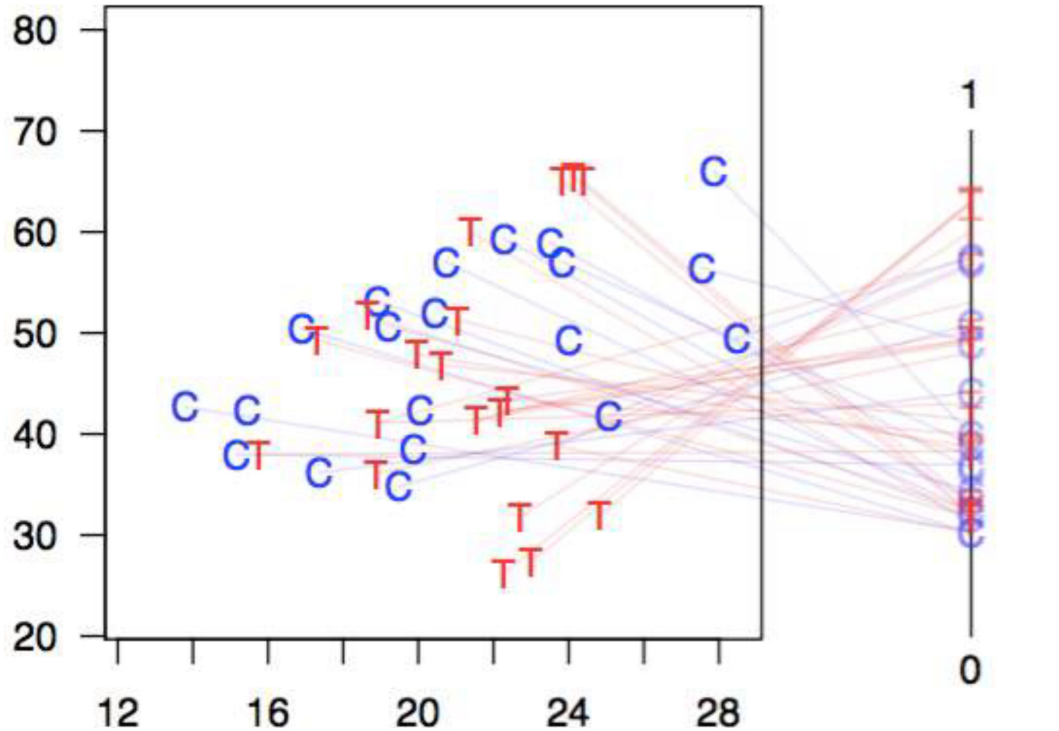
Matching with propensity score



Illustrating propensity score matching

The advantage of propensity score

Scalable with a large set of covariates X_i.
Widely adopted in practices, especially in recent years.



Propensity score "shrinks" the large set of covariates into a single index.

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 scores[2]:

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

[2] The original construction is in Horvitz and Thompson (1952).

Weighting with propensity scores

Technical Proof:

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

Similar proof holds for the second equality.

RSM Erasmus

The weighting estimator

Based on the equations, we can construct two estimators:

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

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

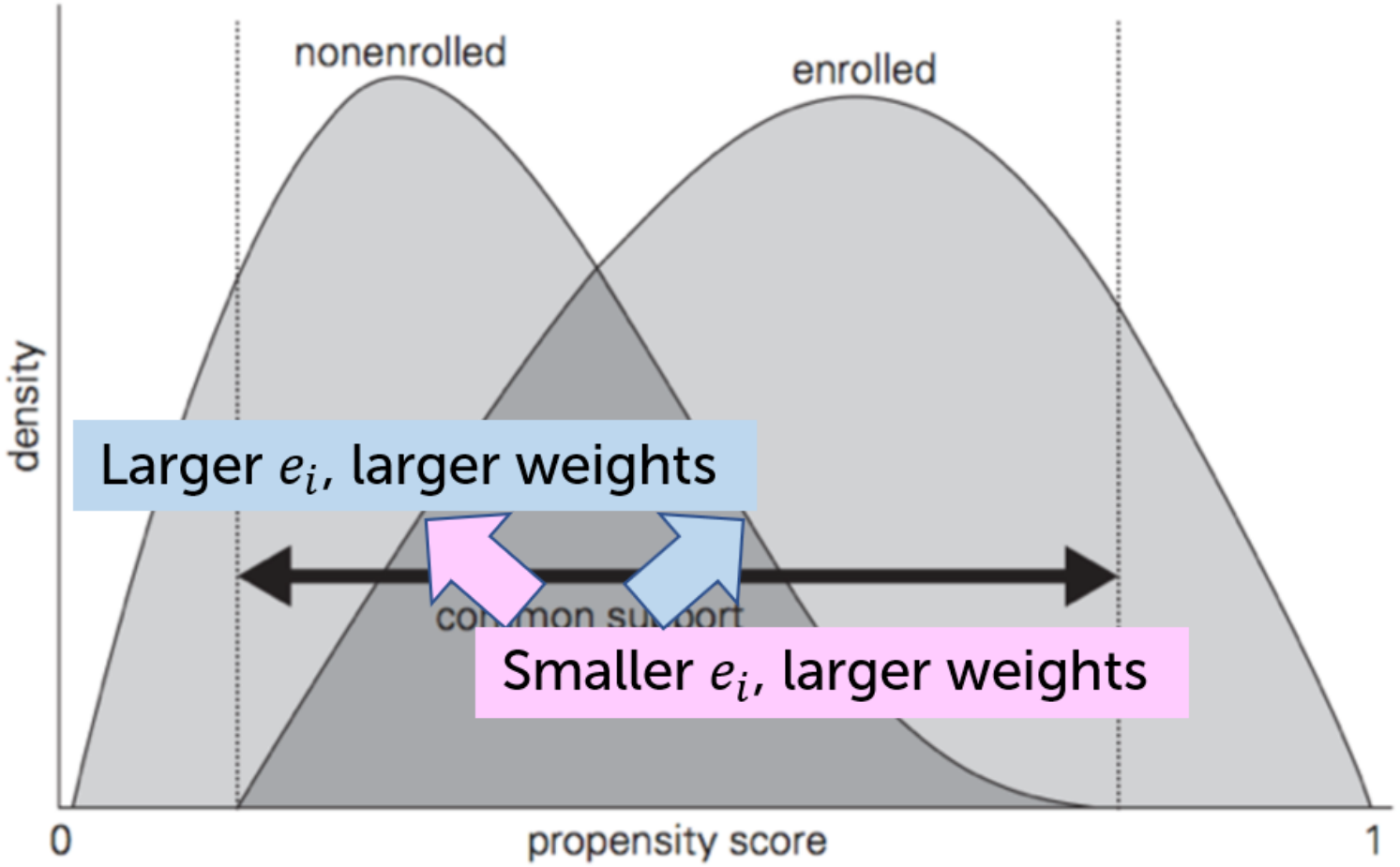
The weighting estimator of ATE is defined as,

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

RSM Erasmus

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

In practice, we do not observe the true propensity scores.
We need to first estimate the propensity scores with $P(D_i = 1 \mid X_i)$, possibly with different methods and specifications.

Plug in estimated propensity score $\hat{e}(X_i)$ into a weighted regression.

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

Normalize weights to improve standard errors, with

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

RSM Erasmus

The weighted regression for ATT and ATC

We can also estimate the average treatment effects on the treated by only weighting the control group.
The weights (odds ratio) for ATT estimation is:

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

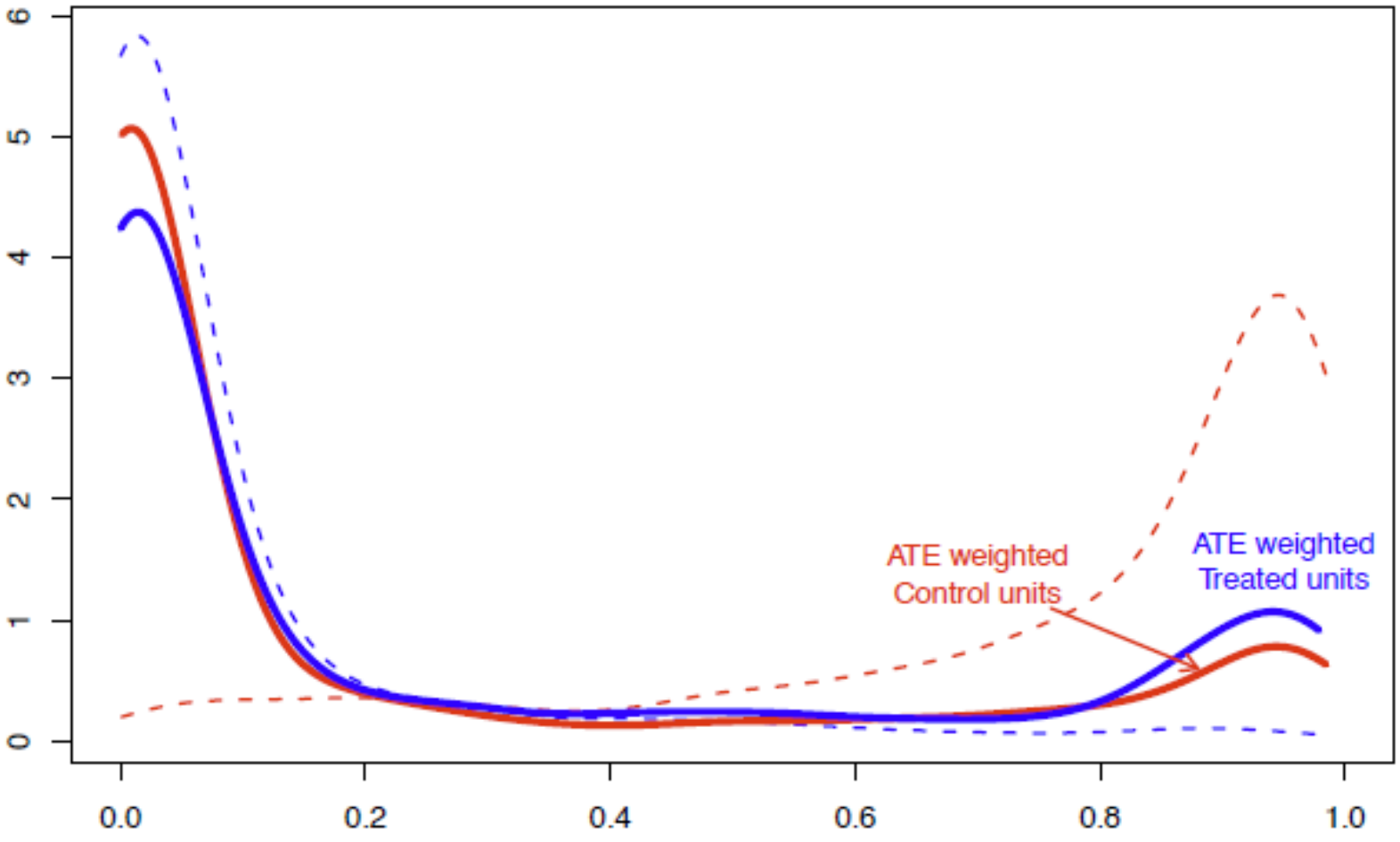
The weights for ATC estimation is similar:

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

RSM Erasmus

What is a good estimation of propensity score?

Balancing condition: samples after weights should show similar distributions of X_i, with $E\left[\frac{D_i \cdot X_i}{e(X_i)} - \frac{(1-D_i)\cdot X_i}{1-e(X_i)}\right] = 0$



The tautology of propensity score

The true propensity score is unknown (by definition), and estimated by a model.

It is always possible to mis-specify the propensity score models.

The treatment effects estimation is sensitive to mis-specification.

In practice, *ad hoc* robustness checks across different specs and models.

In recent years, some machine learning models (non-parametric) prove to be useful.

RSM Erasmus

Outline

RSM Erasmus

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)

RSM Erasmus

Proof: matching is about covariate balancing

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

Under the assumptions of conditionally unconfounded, individualistic and probabilistic assignment, the variance for CATE is[3],

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

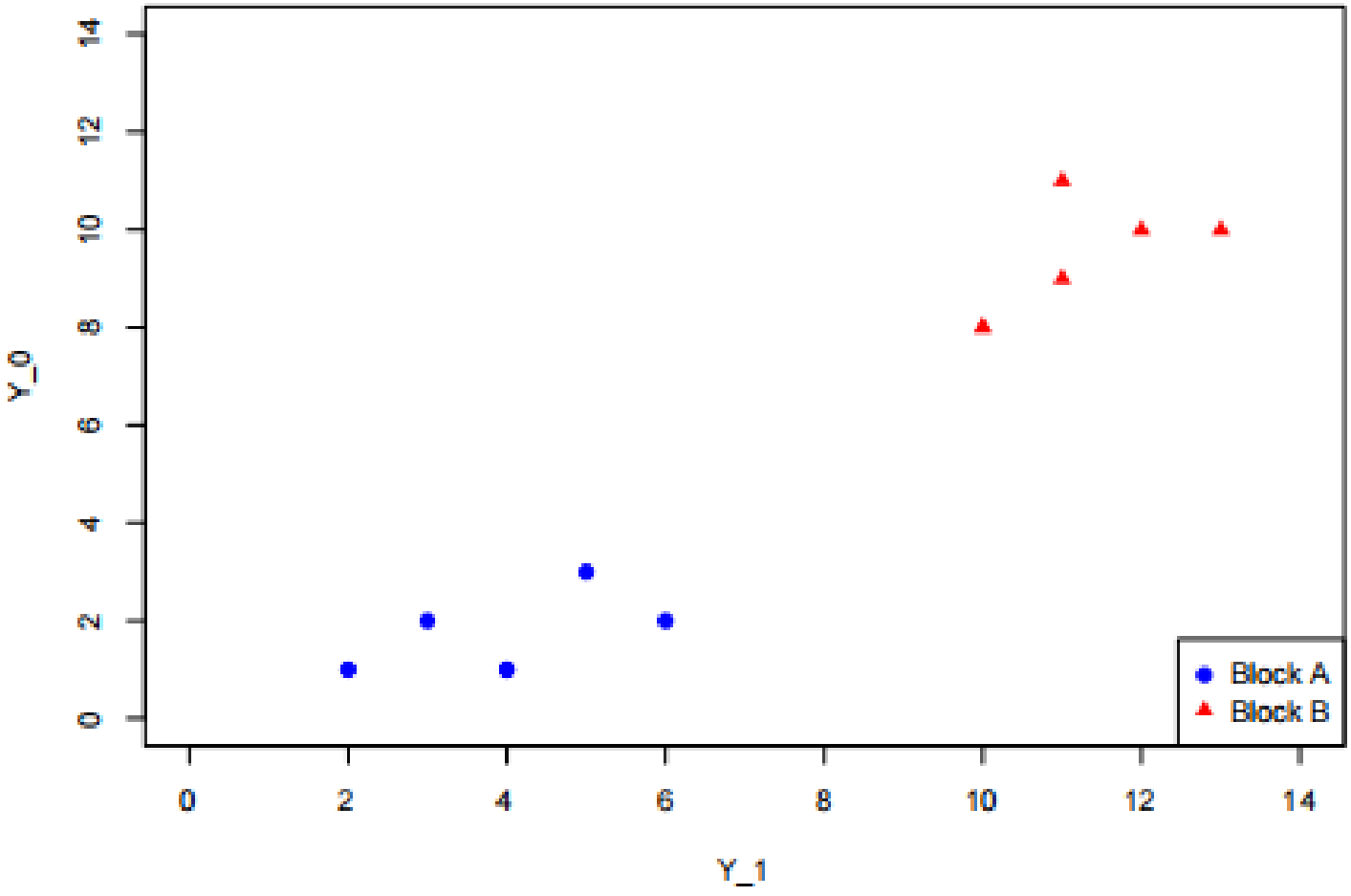
$e(X_i)$ the propensity to be treated, and τ the unconditional ATE.

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.

RSM Erasmus

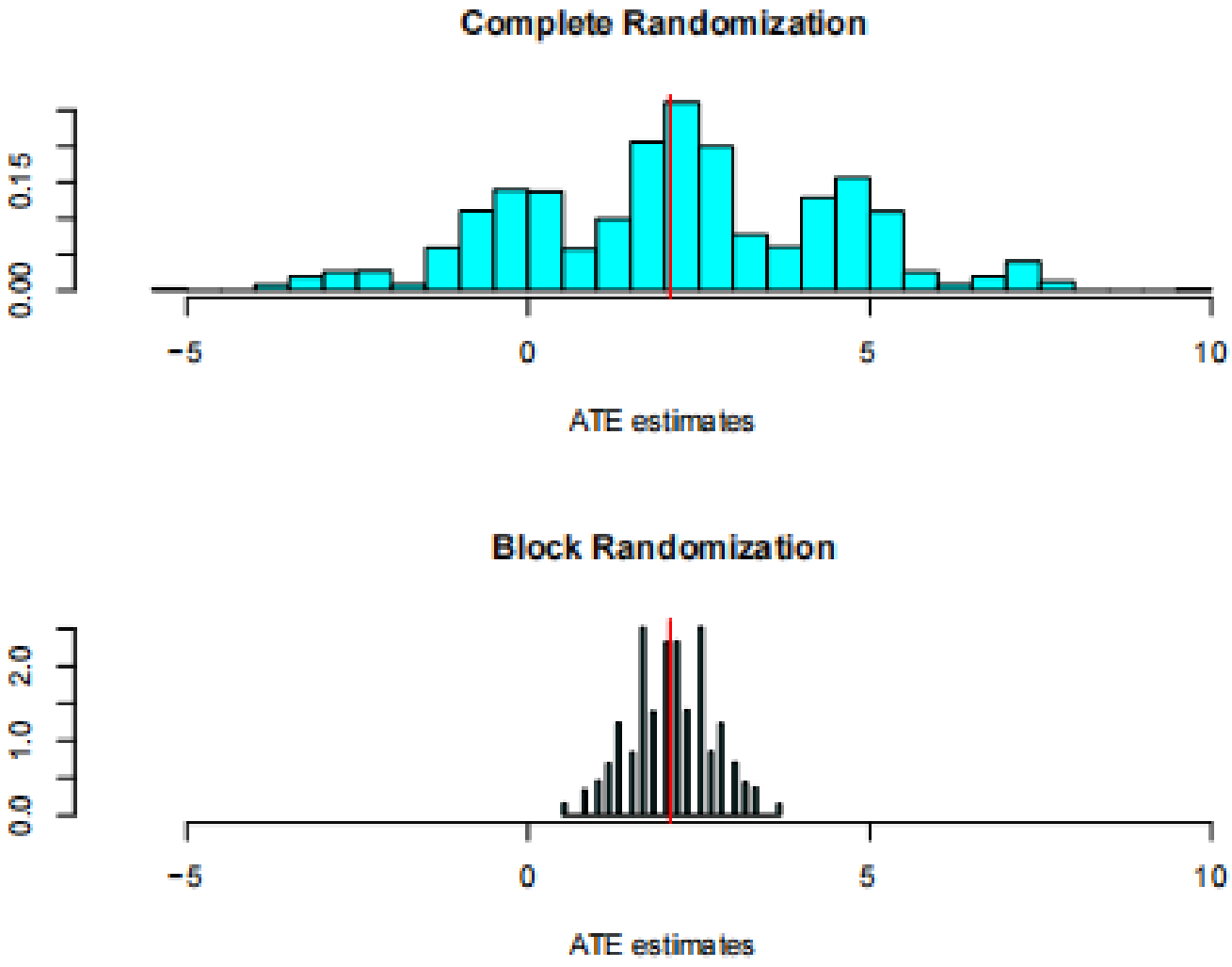
[3] Please see Imbens and Rubin (2015), p. 269 for the derivation.

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...

RSM Erasmus

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 scrore

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

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

For the left-hand side, we have:

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

By conditional unconfoundedness, $E\left[D_i \mid Y_i^0, Y_i^1, X_i, e(X_i)\right] = E\left[D_i \mid X_i, e(X_i)\right]$. By the sufficiency of the propensity score, $E\left[D_i \mid X_i, e(X_i)\right] = E\left[D_i \mid e(X_i)\right]$. If we submit it to the equation above, we have:

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

RSM Erasmus

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.
- 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.

References II

- 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
- 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.

References III