Propensity Score Matching
Hebrew University of Jerusalem
Afternoon Session

Alberto Stefanelli and Prof. Sharon Gilad

2019-12-11

Structure (1/2)

Afternoon (tentative)

1. 14.00 - 14.15 Recap
2. 14.15 - 15.45 Propensity Matching Score
   - Matching, PS, Weights: Making sense of the terms
   - Why we should use matching
3. 14.45 - 15.30 How it works
   - Steps
   - What is happening under the hood
4. 15.30 - 16.00
   - Selection on the observables
   - Covariate Balancing
5. 16.00 - 16.15 Break
6. 16.15 - 16.45 Lalonde dataset
   - Description
   - Assessing covariate Balancing
   - Estimating the ATE
7. 16.45 - 17.00 Setting up R Studio
8. 17.00 - 17.15 Look at some code
9. 17.15 - 18.15 Exercise: Casual inference with Observational data

Section 1

Afternoon Session

Afternoon Session

$$ \end{aligned} ATE&= (roductivity Y_i1 - Productivity Y_i0 &=230 - 155 &=75/6 &=12.5 \end{aligned}

$$

$$
\begin{aligned}
ATT &= (40 + 50)/2 - (35 + 20)/2 \\
&= (90/2) - 55/2 \\
&= 45 - 27.5 \\
&= 17.5
\end{aligned}
$$

$$
\begin{aligned}
ATC &= (40 + 25 + 40 + 35)/4 - (20 + 10 + 40 + 30)/4 \\
&= (140/4) - (100/4) \\
&= 35 - 25 \\
&= 10
\end{aligned}
$$

Section 2

Identification and Assumptions

Identification

1. Usually we cannot fill in the counterfactuals by ourself (unfortunately!!) since are unobservable
2. We must rely on assumptions to identify the treatment effect (fill em in)
   1. IA
   2. SUTVA

Assumption: IA (1/2)

1. IA = Treatment status is independent of potential outcomes
2. The "assignment mechanism" is unconfounded
3. The assignment status (control and treatment) are unrelated to potential outcomes.
    1. Potential outcomes (pain/no pain) do not affect whether person takes an aspirin
    2. People in the treatment group (Academic Degree) would do as bad as the control group (Professional Degree) if they were not treated (they were in a Professional Degree)
4. **People from the treatment group and control group are – on average – the same**
5. In RCT we achieve this with random assignment

Assumption: IA (2/2)

Units	D_i (Collage: Yes/No)	Y_i (Level of Productivity)	Productivity $Y_i1 \mid D_i = 1(Yes)$	Productivity $Y_i0 \mid D_i = 0(No)$
Chris	Yes	40	40	?
Julia	Yes	50	50	?
Paul	No	20	?	20
Trump	No	10	?	10
Fred	No	40	?	40
Diego	No	30	?	30

1. The independence assumption means that
   1. the expected value of the green observations can be equated with the observed values of brown observation (always on average)
   2. the expected value of the blue observations can be equated with the observed values of the red colours (always on average)
2. Formally: $(Y_i1, Y_i0 \perp\!\!\!\perp D_i)$
3. $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$

Assumption 2: SUTVA

1. SUTVA: Stable Unit Treatment Values Assumption
2. **No interference:** The potential outcomes for any unit do not vary with the treatments assigned to other units
   - A subject's potential outcome is not affected by other subjects' exposure to the treatment
   - No interaction between subjects assigned to the experimental condition
3. **No hidden variations of treatment**
   - There are no different forms or versions of the treatment treatment
   - Consistency over time
4. Summarized in: **Potential outcomes are a function of only our individual actions**
5. Q: Imagine that the the local in your home town implement a new training program to reduce unemployment. Why the **independence assumption** (despite random assignment) and/or the **SUTVA** assumptions may be violated?

Section 3

Recap

Why the two assumption are important?

4. In the ideal scenario of a randomised control trial (RCT) ATE equals ATT because we assume that:
   - $E[y_0|D=1] = E[y_0|D=0]$: the baseline of the treatment group equals the baseline of the control group (layman terms: people in the treatment group would do as bad as the control group if they were not treated)
   - $E[y_1|D=1] = E[y_1|D=0]$: the treatment effect on the treated group equals the treatment effect on the control group (layman terms: people in the control group would do as good as the treatment group if they were treated)
   - If $ATE \neq ATT$: the treatment assignment mechanism was probably not random
5. Layman term: The average productivity of Academic collage students would be a good estimator for mean of productivity outcome of all students in the population

Section 4

Matching and Standardization

PS: Some general notions

1. What is it?
2. "broadly [. . .] any method that aims to equate (or"balance") the distribution of covariates in the treated and control groups" (Stuart 2010, 2)
3. find one (or more) non-treated unit(s) with similar observable characteristics of the treated unites to assess the effect of the treatment

Goal:

1. "Recreate" what would happen in a RCT
   1. Simulate that the treatment assignment mechanism is at random
2. IA and SUTVA assumptions still need to hold.
   1. However in their more general form, usually **they do not hold**
   2. IA: treatment status is **NOT independent** of potential outcomes. Most of the cases people self select in the treatment.
   3. SUTVA: People interact with each others
   4. Q: Are the IA and the SUTVA satisfied if we study the impact of public/private collages on success in life?
3. We need Treatment and control group as similar as possible except for the treatment status (**Balance!**)
4. Approaches: so many!!!

PS: So many approaches

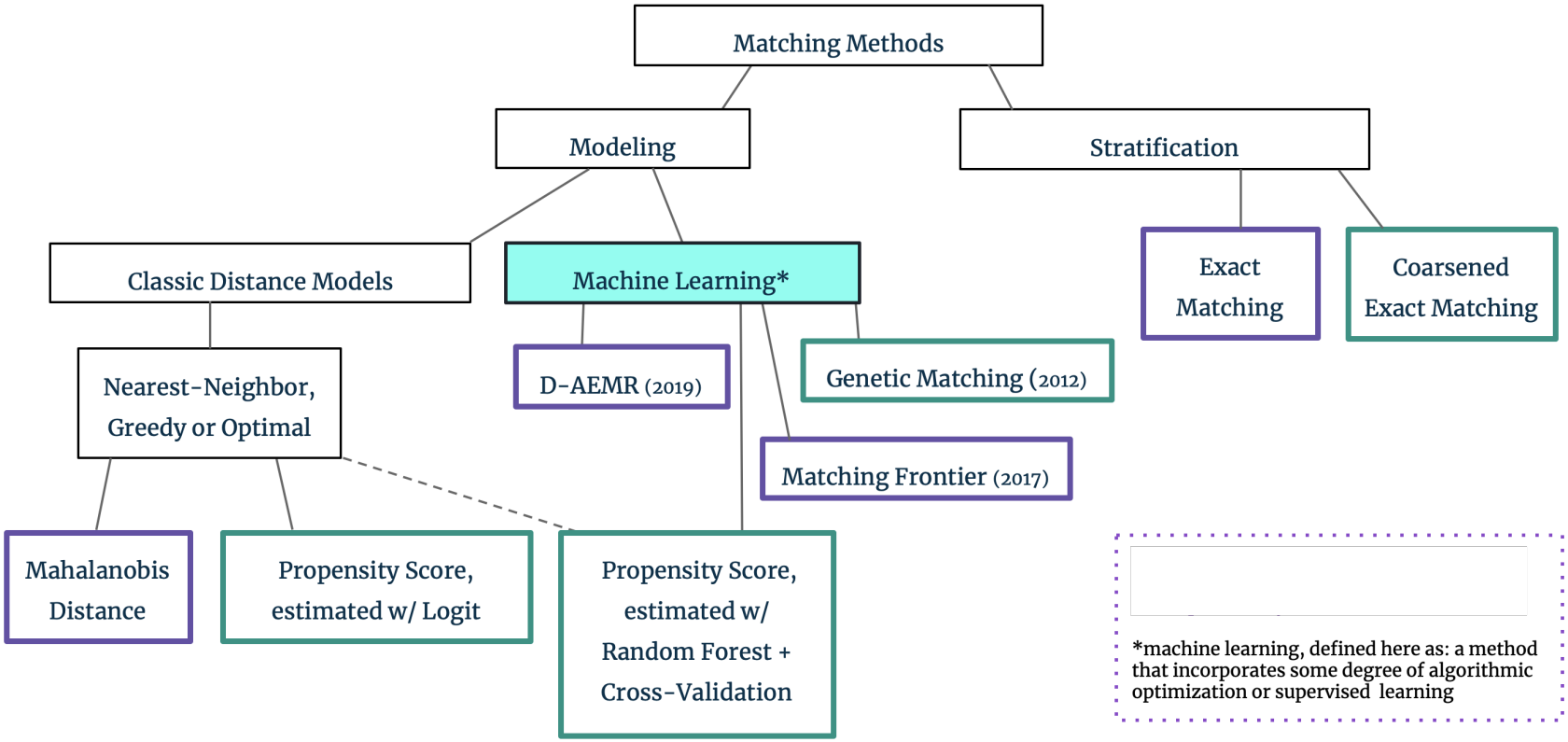

Figure 1: Graph POF

General Logic: Exclude, prune or down-weight observations without comparable units in both treatment and control groups

Why PS is relevant (1/2)

1. Failed Randomization
2. Unethical to run a RCT
   - Smoking
3. Unfeasable to manipulate certain characteristics
   - Gender
   - Genes
   - Discrimination
4. External validity usually higher with observational Data
5. Costs: Secondary data are usually cheaper

Why PS is relevant (1/2)

1. Pure regression approach is increasingly questioned (e.g. Aronow and Samii 2015)
2. Matching methods presents several advantages (Stuart 2010, 2)
   - Guard against extrapolation: Highlight areas of covariate distribution without sufficient overlap/common support between treatment/control
   - Estimation of treatment effects without parametric assumptions
   - Straightforward diagnostics to assess performance
   - Makes you think theoretically about selection of the observables (we will get there in a sec)
   - Complementary to regression adjustment (doubly robust estimands)

Steps: (1/2)

1. Assess quantitatively how different the treatment and control groups are
   - Plotting
   - T-test and means differences
2. Assess theoretically how different your treatment and control group are
   - Are the two populations really comparable?
   - Yes: Why?
   - No: Why?
   - Example: BMI and Asian population
3. **Selection on the observables**
   1. Select those covariates that allow us to assume that the distribution is as it is at random.

Steps: (2/2)

4. Select a series of Matching methods. You need to make a choice based on:
   1. Trade-off between internal and external validity
   2. Model complexity
   3. Covariance balance (see next point)
   4. For a review of matching methods see the Vignette of the R package **MatchIt**
5. **Assessing balance** after matching
   1. Compare multivariate covariate distribution
   2. One-dimensional measures (std. difference in means, variance ratio etc)
   3. Repeat prior steps until balance is good
6. Based on the previous points, your theoretical argument, and your RQ, select the right estimands
   - Conditional means (OLS with controls)
   - ATE
   - ATT
   - ATC

What is happening under the hood (1/3)

Generally, it's a 3-step procedure

1. Estimation of the propensity score in many many different ways. The goal is to estimate a single score that represent the probability of being assigned to the treatment, conditional on pre-treatment (or baseline) characteristics.
   1. logistic regression
   2. porbit regression
   3. GAM and GAM smoothers (generalized additive models)
   4. Decision Trees
   5. ...

What is happening under the hood (2/3)

2. Matching or Weighting
   1. **Matching:** Exclusion of a set of units based on a distance criterion that usually involve some form of geometric mean difference
      - You end up with a dataset containing control and treatment units that have been matched.
   2. **Up-Weighting and Down-Weighting:** Associate a weight with each observation similar to what we do in weighted OLS.
      - End up with a dataset containing all the control and treatment units and a PS weight.
   3. What is the trade-of of pruning units (think representativeness)?

What is happening under the hood (3/3)

3. Parametric or Non-parametric techniques
   1. Estimation of the conditional effects. This is the coefficient of your normal OLS
   2. Estimation of the marginal effects (ATE,ATT, ATC. . .)

A graphical representation of matching (1/2)

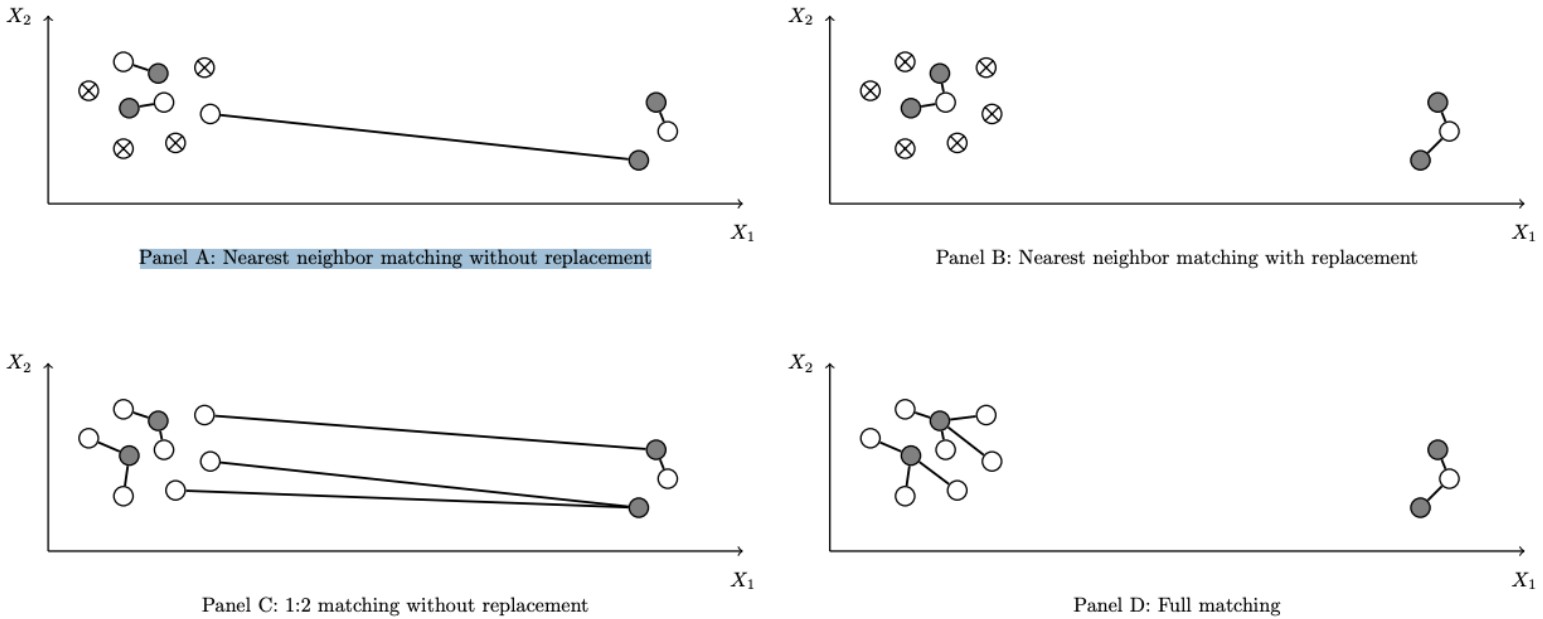

Figure 1: Illustration of different matching methods. The sample consists of 12 units belonging to either of two treatment conditions. We observe two covariates, X_1 and X_2, for each unit. The units are presented as circles in the covariate plane. The color of the circles indicates the units' treatment assignment. Edges indicate matched groups, and crossed out circles indicate discarded units. Well-performing matchings avoid discarded units and long edges (as they correspond to matches between dissimilar units).

Figure 2: Graph POF

A graphical representation of matching (2/2)

- Q: Discuss the previous graph with your neighbour and explain what you see to the others!
  - How many dimensions/variables are shown in the graph?
  - What are the white/gray/strikethrough circles?
  - **Hint:** read the Figure Caption

Selection on the observables: General Logic

1. In RCT the missing potential outcome is missing completely at random so we ensure that we can assume that the missing information can be replaced with the data at hand.
2. In observational studies this is usually not the case
3. **Assumption:** there is some set of pre-treatment covariates such that treatment assignment is random conditional on these covariates (Barnow, Cain, and Goldberger 1980)
4. Also called conditional ignorability, no omitted variables or conditional on observables
5. Meaning (1): If we control for the right set of pre-treatment covariates, we achieve conditional randomization
6. Meaning (2): Instead of assuming that the treatment is randomly assigned, we assume that the treatment is as good as randomly assigned after conditioning on covariates
7. How we pick up the variable to condition upon
   - Algorithmically. Include variables to achieve best fit such as LL or AIC/BIC... (highly debated)
   - Theoretically

Selection on the observables: Brain Cancer Example

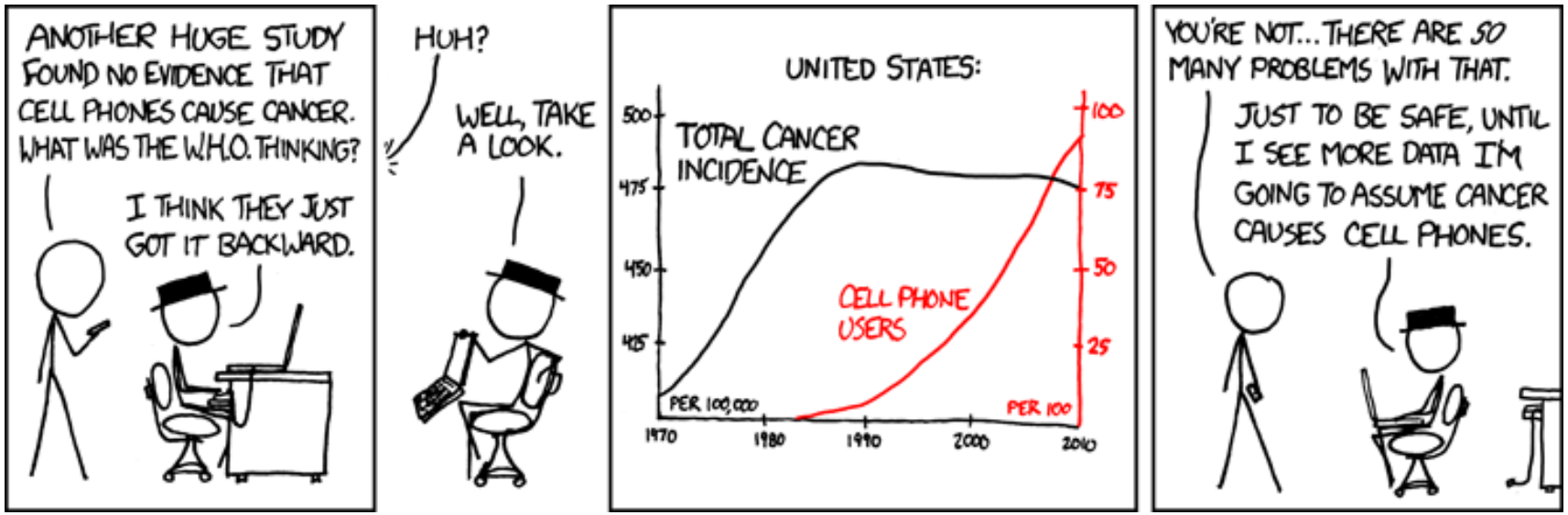

Figure 3: Graph POF

- **NB: Be sure to include all the confounding variables!!**

Selection on the observables: Pre-treatment

1. We need to use pre-treatment variables to find comparable units
2. Pre-treatment variables means a set of variables that is unaffected by treatment assignment.
3. Background variables (gender, age, income ....) also counts as pre-treatment variables
4. Q: We are studying the impact of Party ID on Vote Choice
    1. Treatment/Control= Democrats/Republicans
    2. Vote Choice= Trump/Clinton
    3. Control: Issue Proximity
    4. Q: Is Issue Proximity a Pre-Treatment variable ? (Think about survey design and questionnaire)

Selection on the observables: Guidelines

1. Match on all observed variables that may affect both treatment D and outcome Y
2. Conceptual: Find subset in your data in which X –> Y (Tam Cho et al. 2013)
3. Avoid post-treatment and endogenous selection bias (Elwert and Winship 2014b)
4. Match on quadratic/polynomial terms where it makes sense theoretically (e.g. age X, education D and income Y)

Selection on the observables: bad research in the social sciences?

All empirical papers, top 3 top political science journals, 2010-2015. (Acharya, Blackwell, and Sen 2015, 1)

1. 40% explicitly conditioned on a post treatment variable
2. 27% conditioned on a variable that could plausibly be post treatment
3. 1/3 of the above make causal claims condition on post treatment variables
4. **Fortunately:** 33% no post-treatment variables included in their analyses

Covariates balancing

1. **Assumption**: The two populations needs to overlap to ensures that each individual could get any treatment level
2. Even if there is overlap, the comparison needs to makes sense
3. Formally: Each individual needs to have a probability above zero to receive the treatment
4. If the probability is 0 or close to zero
   - Severely Down-weight
   - Purging
5. The way you purge of down-up weight is the **CENTRAL point** of matching methods
6. **BE CAREFUL** If many propensity matching score close to 0 or 1
7. Q: Why is this the case?

Section 5

Recap

Recap main differences between PS and RCT

Probability of Treatment:	Known	Unknown, may be 0 or 1
Covariate Balance:	Relationship between covariates and treatment assignment are known from study design. Usually the study is designed so that there is no relationship	Relationship between covariates and treatment assignment is unknown. There may be covariate imbalance.

The Lalonde dataset: Dataset description I

- Analyzed in Lalonde (1986) and Dehejia and Wahba (1999).
- The "treatment" refers to participation in a employment program that took place between 1976 and 1977 to help disadvantage to move into the labour market
- Employment program (Treatment) -> Earnings
- DV:
  - re78: real earnings in 1978.
- Control/Treatment:
  - treat: treatment indicator
- Background and Pretreatment variables
  - age: age in years.
  - educ: years of schooling.
  - nodegr: indicator variable for high school diploma.
  - black: indicator variable for blacks.
  - hisp: indicator variable for Hispanics.
  - married: indicator variable for martial status.
  - re74: real earnings in 1974.
  - re75: real earnings in 1975.

The Lalonde dataset: Dataset description II

- Sample of 614
- Treatment of 185
- Control of 429

Let's take a look at it: Age Distribution

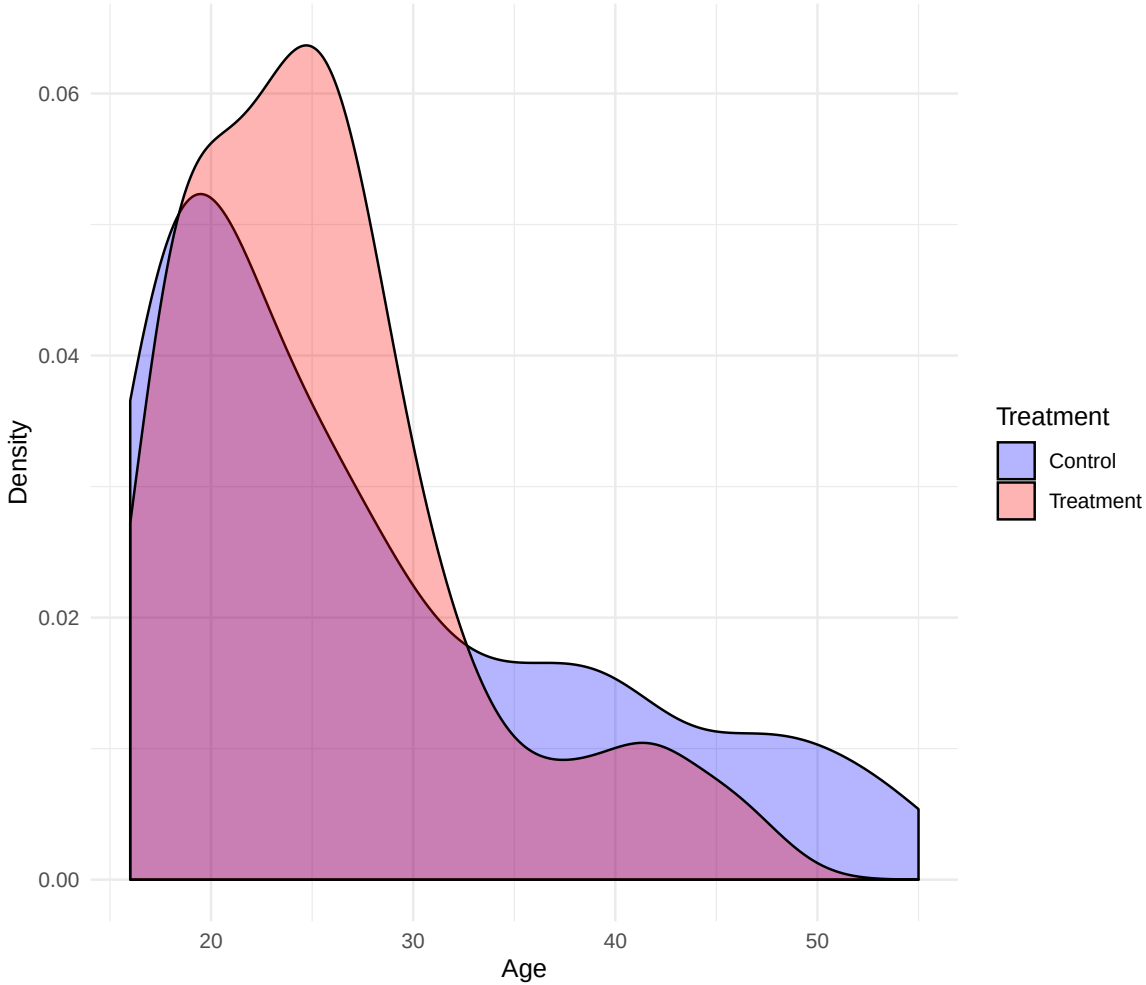

Figure 4: Graph POF

Let's take a look at it: Age t-test

```
## The Welch Two Sample t-test suggests that the difference
## between treatment$age and control$age (mean of x = 25.82,
## mean of y = 28.03, difference = -2.21) is significant
## (t(510.57) = -2.99, 95% CI [-3.67, -0.76], p = 0.003).
```

Let's take a look at it: Education Distribution

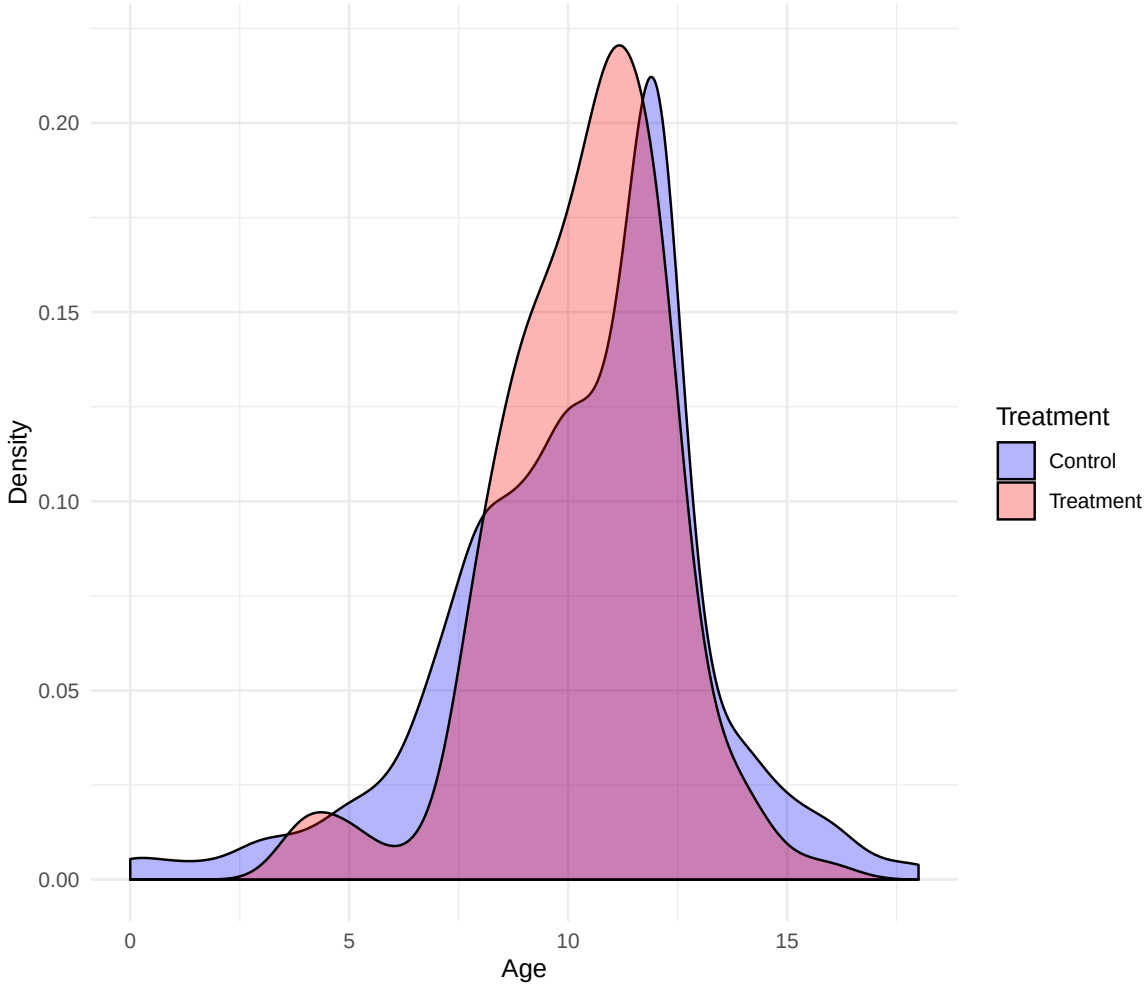

Figure 5: Graph POF

Let's take a look at it: Education t-test

```
## The Welch Two Sample t-test suggests that the difference
## between treatment$educ and control$educ (mean of x = 10.35,
## mean of y = 10.24, difference = 0.11) is not significant
## (t(485.37) = 0.55, 95% CI [-0.29, 0.51], p = 0.585).
```

Probability of receiving the treatment based on pre-treatment variables

1. Predict treatment assignment using all pre-treatment variables

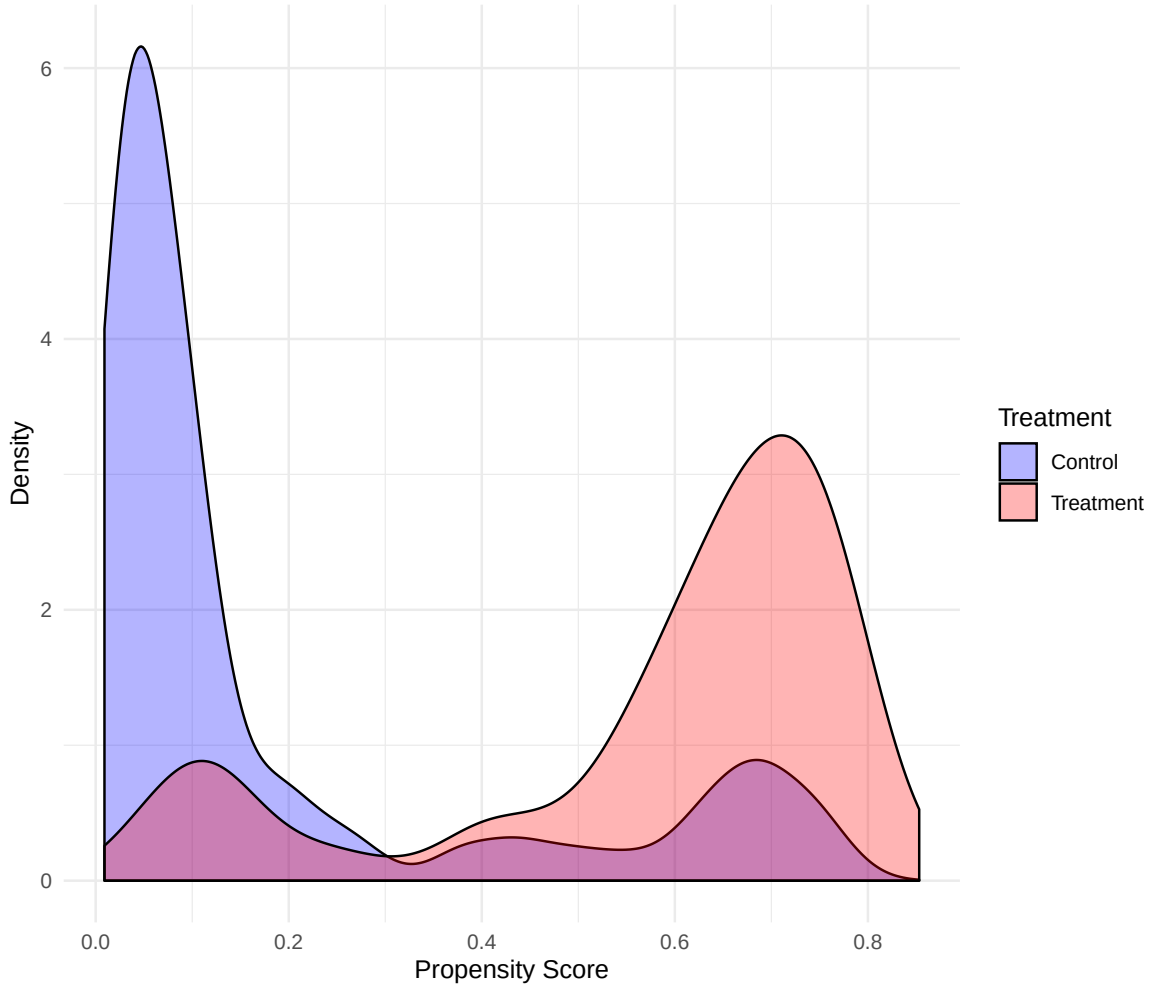

Figure 6: Graph Propensity

Let's look at the ATE and ATT

- ATE: Estimate the the treatment effect with the whole population, both treated and controlled, as target
- ATT: Estimates the treatment effect with the treated population as the target population.
- ATO: Estimating the treatment effect among those likely to have received either treatment or control.

The propensity score for participant i is defined here as e_i and the treatment assignment is Z_i, where Z=1 indicates the participant received the treatment and Z=0 indicates they received the control.

- $w_{ATE} = (1 - e_i)Z_i + e_i(1 - Z_i)$
- $w_{ATT} = \frac{e_i Z_i}{e_i} + \frac{e_i(1-Z_i)}{1-e_i}$
- $w_{ATO} = (1 - e_i)Z_i + e_i(1 - Z_i)$

Another way to look at the PS distribution

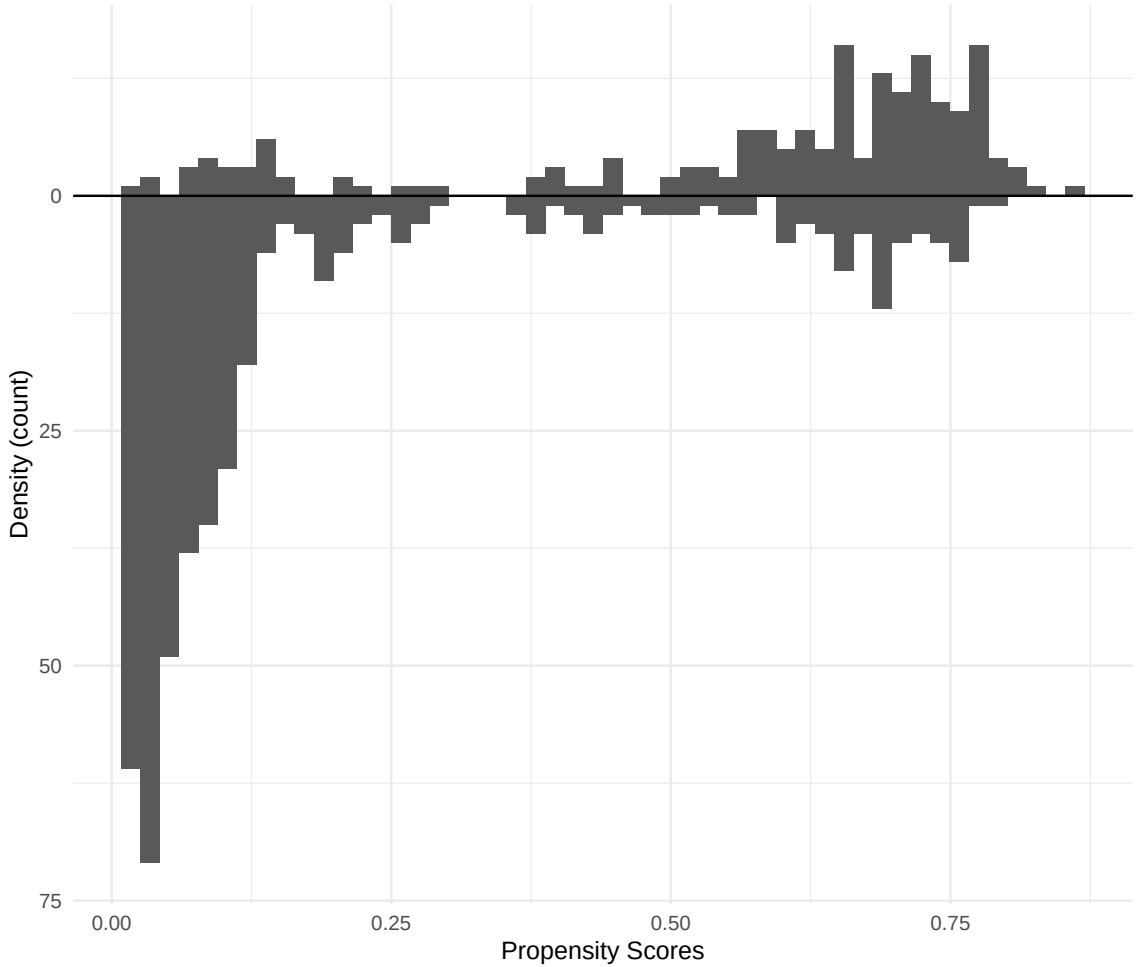

Figure 7: Another Propensity distribution Graph

Overlapping observation and the ATE weights

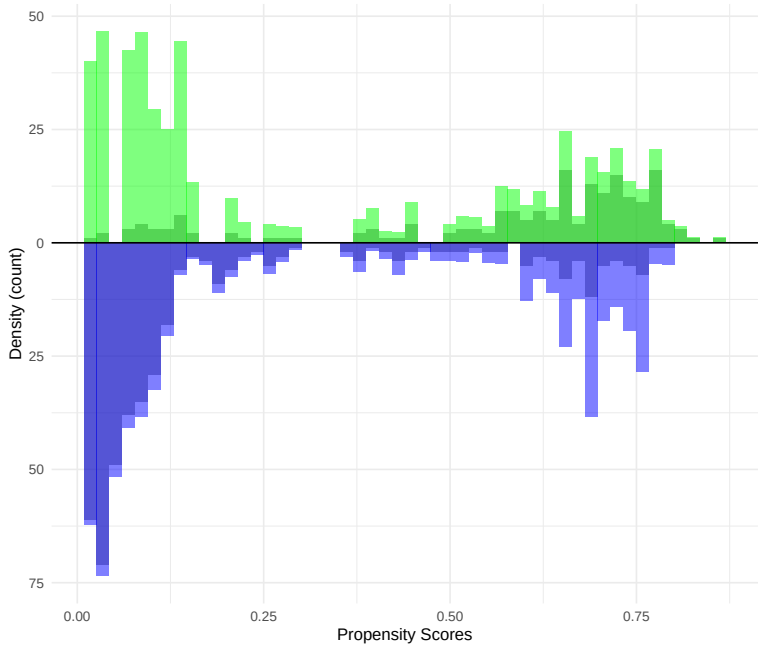

Figure 8: ATE Propensity distribution Graph

- Green: "pseudo-population" for the treated group
- Blue: "pseudo-population" for the control group
- Darker Green/Blue: Weights

Overlapping observation and the ATT weights

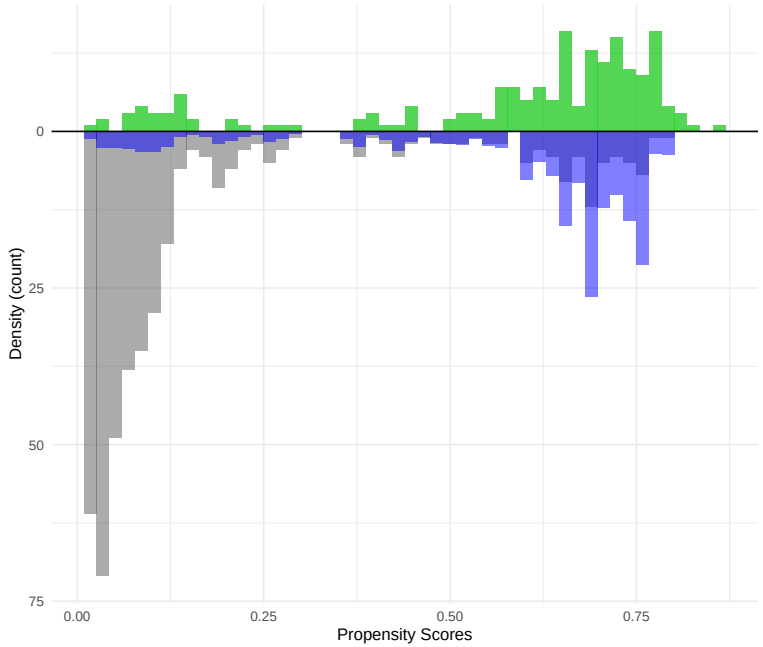

Figure 9: ATT Propensity distribution Graph

- Q: What is happening here?

Overlapping observation and the ATT weights

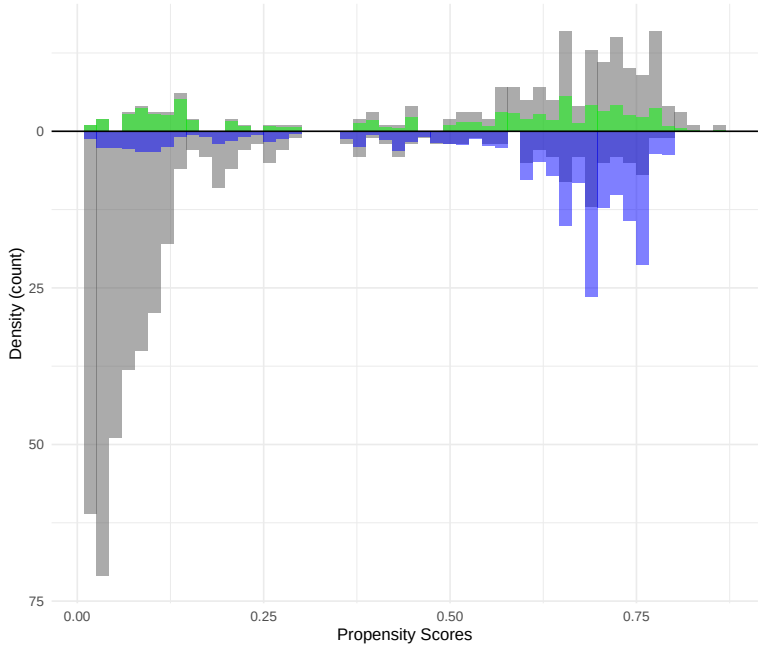

Figure 10: ATT Propensity distribution Graph

- The "pseudo-population" for the treatment group is exactly the same as the actual population.
- We leave the treatment population as it
- Weight only the control population to match the treatment distribution.

Overlapping observation and the ATO weights

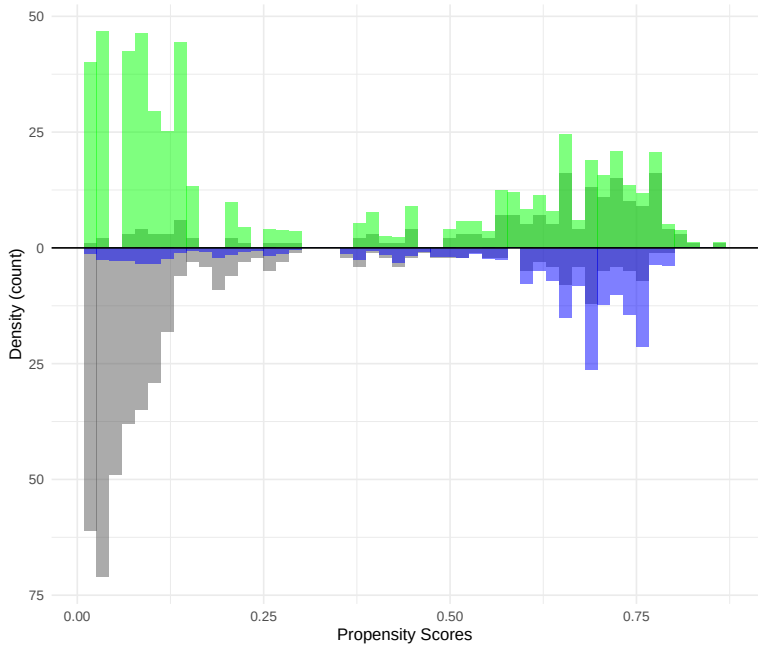

Figure 11: ATO Propensity distribution Graph

- This "pseudo-population" looks like a 1:1 matched population.
- In the regions where the treatment group is in the minority, on the left side of the graph, the controls are weighted to match the distribution of the treated.
- In the regions where the control group is in the minority, on the right side of the graph, the treated are weighted to match the distribution of the controls.

Covariates balancing: Marijuana Example

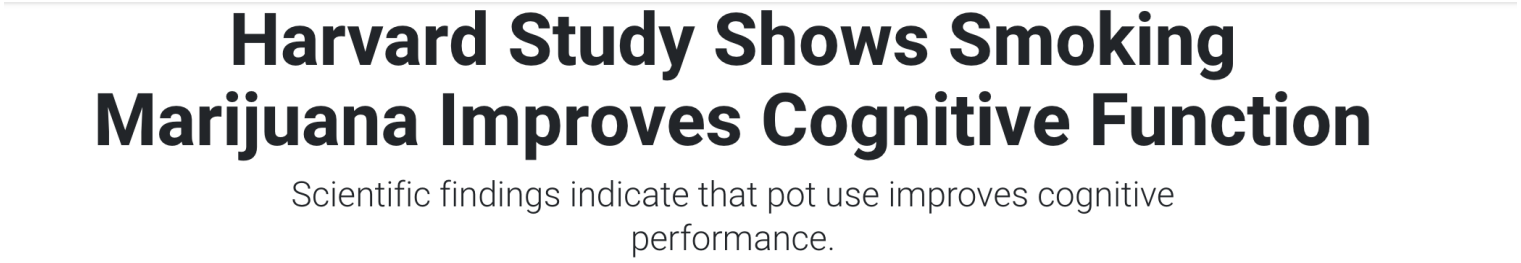

Figure 12: Graph POF

- RTC
- Stratified sample design
- Prefect covariate balance
- Q: Does this sound right to you?