Overlap, Matching, or Entropy Weights: *What are we weighting for?*

Yi Liu

Department of Statistics, North Carolina State University

yliu297@ncsu.edu

Acknowledgement. I am grateful to the supervision by Prof. Roland Matsouaka (B&B, Duke) and many helpful discussions with Yunji Zhou (PhD student, University of Washington) for this project.

Scan QR code for arXiv paper

Introduction

There has been a recent surge in statistical methods for handling the lack of adequate positivity when using inverse probability weighted (IPW) estimator. However, these nascent developments have raised a number of questions.

Thus, we demonstrate the ability of equipoise estimators (overlap, matching, and entropy weights) to handle the lack of positivity.

- To infer causality, what are they really estimating and what are their target populations?
- We specifically look into the impact imbalances in treatment allocation can have on the positivity and, ultimately, on the estimates of the treatment effect.

Setup

- Treatment: $Z \in \{0,1\}$; covariates vector: X; potential outcome: $Y(z),\ z=0,1$; observed outcome: $Y = ZY(1) + (1-Z)Y(0)$; PS: $e(x) = P(Z=1|X=x)$
- Common assumptions in causal inference literature are made: SUTVA, consistency, positivity (overlap), unconfoundeness. The positivity assumption is the heart of research here, which stated that $0 < e(X) < 1$ w.p.1.
- *Weighted average treatment effect (WATE)* estimand:

$$\tau_g = \frac{\mathbb{E}[g(X)\tau(X)]}{\mathbb{E}[g(X)]}$$

where $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X=x]$, and $g(x)$ is a *tilting function* which specifies a target population. We considered the following choices of $g(x)$, which are all functions of $e(x)$.

Target	$g(x)$	Estimand	Weights
overall	1	ATE	IPW
treated	$e(x)$	ATT	IPWT
control	$1-e(x)$	ATC	IPWC
restricted	$\mathbf{1}\{\alpha \le e(x) \le 1-\alpha\}$	ATE	IPW trimming
overlap	$e(x)(1-e(x))$	ATO (overlap)	OW
overlap	$\min\{e(x), 1-e(x)\}$	ATM (matching)	MW
overlap	$u(e(x)) + u(1-e(x))$	ATEN (entropy)	EW

$u(t) = -t\log t;\ 0 < \alpha < 0.5$, e.g., $0.05, 0.1$

Table 1: Examples of tilting functions, causal estimands and propensity score weights

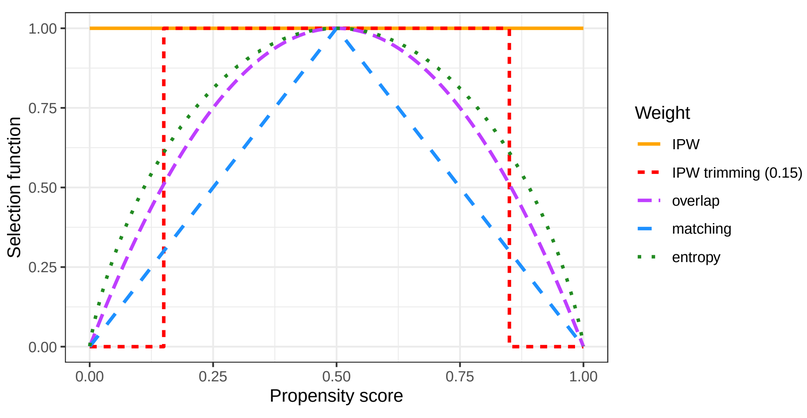


Figure 1: Visualizations of $g(x)$ vs. $e(x)$

OW, EW and MW weigh most on those with PS of 0.5 (the "clinical equipoise"), and water-down evenly and smoothly at both sides of 0.5 to 0 weight. They avoid deciding some ad hoc parameters in their estimands, e.g., the threshold for IPW trimming.

- Estimators: We considered two commonly used estimators for WATE in practice: Hájek-type (i.e., normalized PS weighting) and augmented estimators. The latter requires modelling the potential outcomes.

What are we weighting for?

We assessed the impact of proportion of the treated participants $p = P(Z=1) = \mathbb{E}\{e(X)\}$ to the relationship of equipoise estimators, ATE, ATT and ATC estimators.

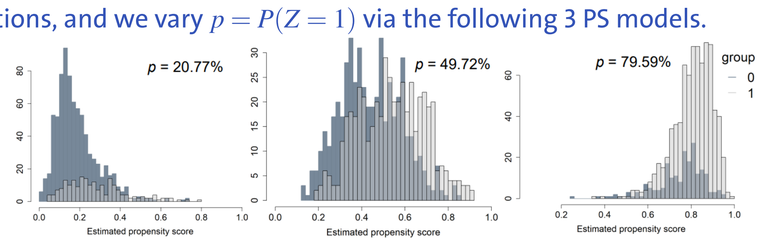
First, clearly ATE = pATT + $(1-p)$ATC. Second,

- when $e(x) \approx 0.5$, $(e(x), 1-e(x)) \approx \left(\frac{0.25}{1-e(x)}, \frac{0.25}{e(x)}\right)$ (ATE weights)
- when $e(x)$ is small, $(e(x), 1-e(x)) \approx \left(\frac{e(x)}{1-e(x)}, 1\right)$ (ATT weights)
- when $e(x)$ is large, $(e(x), 1-e(x)) \approx \left(1, \frac{1-e(x)}{e(x)}\right)$ (ATC weights)

Our hunch is that under some conditions, p might be sufficient to reflect how ATO weights ATT and ATC.

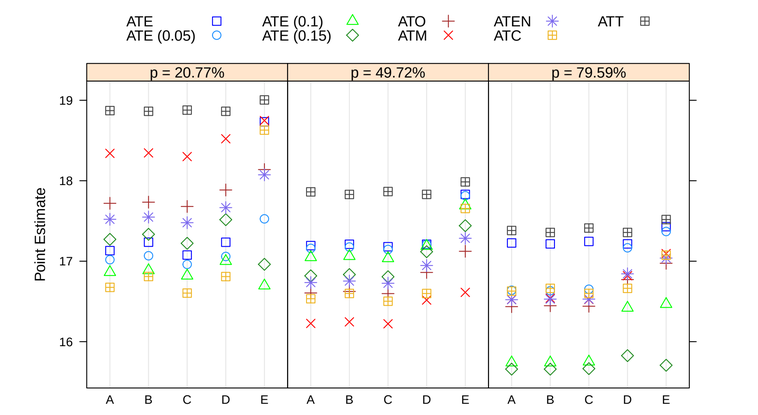
Simulation findings

We generated some observational data under common assumptions, and we vary $p = P(Z=1)$ via the following 3 PS models.

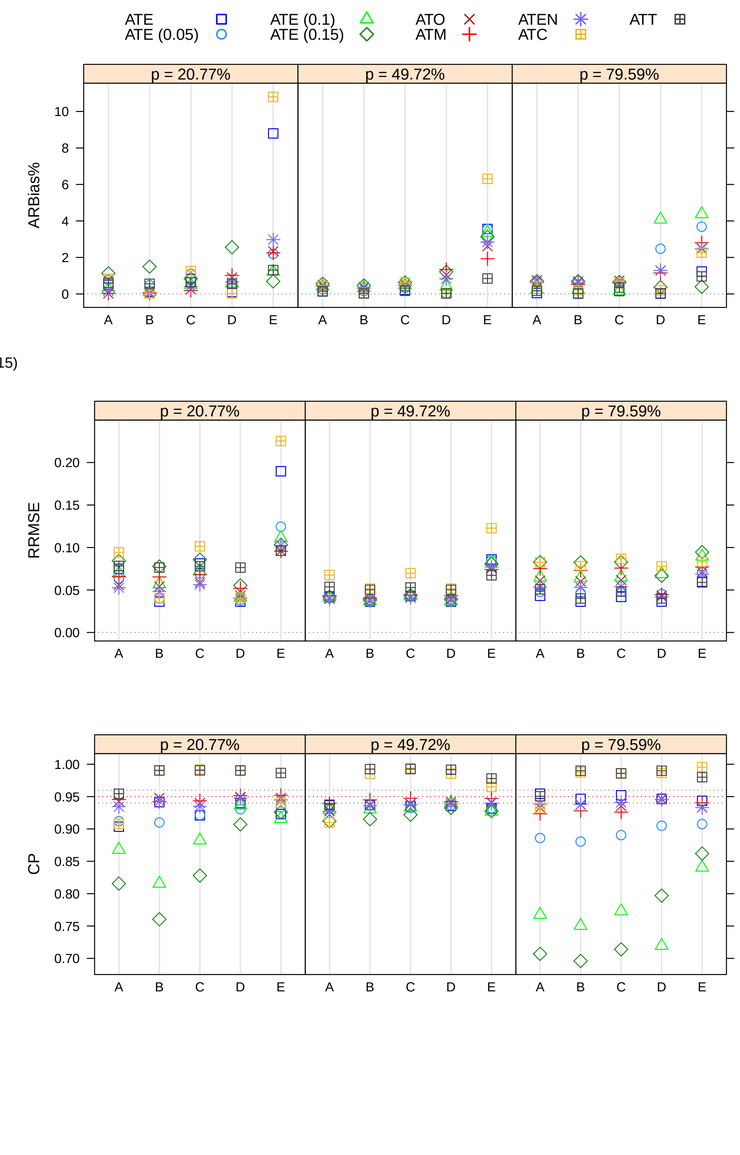


We have the following main simulation findings.

In following 2 figures, A: Hájek-type (weighted) estimator; B (resp. C, D, and E): augmented estimator, with both the PS and OR models correctly specified (resp. only the PS model correctly specified, only the OR model correctly specified, both the PS and OR models misspecified).



When p is small, estimators of equipoise estimands (resp. ATE) move toward ATT (resp. ATC), and vice versa. When $p \approx 0.5$ and no extreme weights exist, they have similar values to ATE.



- Augmented estimators for ATO, ATM and ATEN are more robust to model misspecifications than that of ATE and ATE trimming. In addition, the Hájek-type estimator of equipoise effects has been proved more robust than that of ATE [2], so we did not further investigate it here.
- Coverage probabilities (CP) of equipoise estimators are closer to the nominal 95% level. While trimming shows to have a good bias-variance trade-off, their variance estimations overestimate the efficiency from their poor CPs.

Analysis of racial disparities in health care expenditure

We evaluate racial disparities in the health care expenditure using data from the Medical Expenditure Panel Survey (MEPS). We focus on three specific 2-by-2 comparisons: White vs. Hispanic, White vs. Black, and White vs. Asian, with White as the reference group ($Z=1$) and the minority racial or ethnic group as control ($Z=0$).

Figure 2: Racial disparities in the health care expenditure

The proportion p of White participants is 65.06% in White vs. Hispanic, 70.97% in White vs. Black, and 87.18% in White vs. Asian. Our data analysis further confirms our findings from the simulation study and our hunch.

Take-away messages

- We provided a coherent assessment of the different estimands of the propensity score weighting methods to dispel confusion we may have around their use via a series of Monte Carlo simulations.
- We also demonstrate why and how ATE estimators can fail to identify logical treatment effect estimands and why using IPW trimming is not always a good idea. We must choose our estimand and corresponding weights wisely to recover the estimated or specific causal effects (or parameters) of interest that align with our scientific question(s).
- Beware of what you ultimately get when using a specific weighting method. ATE may not lead you where you expected; ATO, ATM, and ATEN take you in the overlap/equipoise land and provide the estimate of the treatment effect on the subgroup of participants for whom there is clinical equipoise [1]. Thus, the answer to this very simple question: when you are using a weighting method, what are you weighting for?

References

[1] Roland A Matsouaka and Yunji Zhou. A framework for causal inference in the presence of extreme inverse probability weights: the role of overlap weights. *arXiv preprint arXiv:2011.01388*, 2020.

[2] Yunji Zhou, Roland A Matsouaka, and Laine Thomas. Propensity score weighting under limited overlap and model misspecification. *Statistical methods in medical research*, 29(12):3721–3756, 2020.