DISS 2023 Presentation

Causal inference on the treated and on the control under lack of positivity

Yi Liu, PhD student

Department of Statistics, North Carolina State University

NC STATE UNIVERSITY

Introduction

The use of propensity score method has become ubiquitous in causal inference from observational studies as it provides a proper framework to assess treatment effects while properly adjusting for confounding or selection bias. At the heart of PS methods is the **positivity** assumption.[1]

Notations

Data $\mathcal{O} = \{(\boldsymbol{X}_i, Y_i, Z_i), i = 1, \ldots, N\}$

- $\boldsymbol{X}_i = (X_0, X_{i1}, \ldots, X_{ip})$: design matrix by (baseline) covariates, where $X_0 = (1, 1, \ldots, 1)'$
- Y_i: observed outcome
- $Z_i \in \{0, 1\}$: the binary treatment status, $Z_i = 0$ for control, $Z_i = 1$ for treated

Potential outcome: $Y_i(z), i = 1, \ldots, N$ for $z = 0, 1$ is the potential outcome associated with treatment status z assumed for each participant, probably contrary to the fact

Goals and Assumptions

The goal is to estimate the average treatment effect on the treated (ATT) and on the control (ATC), defined, respectively, by:

- $\tau_{att} = \mathbb{E}\{Y(1) - Y(0)|Z = 1\}$
- $\tau_{atc} = \mathbb{E}\{Y(1) - Y(0)|Z = 0\}$

Key assumptions for identifying ATT and ATC from observational data:

- Consistency: $Y = ZY(Z) + (1 - Z)Y(1 - Z)$
- Unconfoundness: $Y(z) \perp Z|X$ for $z = 0, 1$
- **Positivity**: $\text{pr}(Z = 1|X) \leq \bar{e} < 1$ almost everywhere for **control** participants when estimating **ATT**; $0 < \underline{e} \leq \text{pr}(Z = 1|X)$ almost everywhere for **treated** participants when estimating **ATC**

$\text{pr}(Z = 1|X) := e(X)$ is the propensity score.

Why **positivity** important? For estimating ATT (resp. ATC), $1 - e(X_i)$ for control (resp. $e(X_i)$ for treated) too close to 0 incurs extreme inverse probability weight, leading inefficient estimates under finite sample![2]

Methods: WATT and WATC

It can be shown that

$$\tau_{att} = \mathbb{E}\{Y|Z=1\} - \mathbb{E}\left\{(1-Z)Y\frac{e(X)}{1-e(X)}\right\}\frac{1}{\text{pr}(Z=1)},$$
$$\tau_{atc} = \mathbb{E}\left\{\frac{1-e(X)}{e(X)}ZY\right\}\frac{1}{\text{pr}(Z=0)} - \mathbb{E}\{Y|Z=0\},$$

where $\text{pr}(Z=1)$ can also be replaced by $\mathbb{E}\{e(X)\}$.

We propose the following generalized **estimand**, called **weighted** ATT and **weighted** ATC (WATT and WATC)

$$\tau_{watt} = \mathbb{E}\{Y|Z=1\} - \mathbb{E}\left\{(1-Z)Y\frac{e(X)h(X)}{1-e(X)}\right\}\frac{1}{\mathbb{E}\{e(X)h(X)\}},$$
$$\tau_{watc} = \mathbb{E}\left\{\frac{(1-e(X))h(X)}{e(X)}ZY\right\}\frac{1}{\mathbb{E}\{(1-e(X))h(X)\}} - \mathbb{E}\{Y|Z=0\}.$$

$h(X)$: tilting function which "tilts" the distribution of treated/control population.

When choosing $h(X) = e(X)(1-e(X))$, the overlap function[3], we call the corresponding estimands equipoise ATT and equipoise ATC (EATT and EATC).

The overlap function emphasizes those on the "clinical equipoise"—having substantial probabilities to receive both treatment or control. The tilting function can also be other choices: beta family function, entropy function[4], matching function[5], etc.

Asymptotic Properties

For being concise, we only discuss WATT (and EATT) in this section. Similar arguments apply to WATC (and EATC) by symmetry.

- Simple weighting estimator (needs models for $e(X)$)

$$\widehat{\tau}_{watt} = \frac{\sum_{i=1}^N Z_i Y_i}{\sum_{i=1}^N Z_i} - \frac{\sum_{i=1}^N \widehat{h}(X_i)\widehat{e}(X_i)\{1-\widehat{e}(X_i)\}^{-1}(1-Z_i)Y_i}{\sum_{i=1}^N \widehat{h}(X_i)\widehat{e}(X_i)},$$

- Type-I augmented estimator (needs models for $e(X)$ and $\mathbb{E}\{Y(0)|X\}$)

$$\widehat{\tau}_{watt}^{\text{aug-I}} = \frac{\sum_{i=1}^N Z_i Y_i}{\sum_{i=1}^N Z_i} - \frac{\sum_{i=1}^N \frac{\widehat{h}(X_i)\widehat{e}(X_i)}{1-\widehat{e}(X_i)}(Z_i - \widehat{e}(X_i))\widehat{m}_0(X_i)}{\sum_{i=1}^N \widehat{h}(X_i)\widehat{e}(X_i)} - \frac{\sum_{i=1}^N \frac{\widehat{h}(X_i)\widehat{e}(X_i)}{1-\widehat{e}(X_i)}(1-Z_i)Y_i}{\sum_{i=1}^N \widehat{h}(X_i)\widehat{e}(X_i)},$$

- Type-II augmented estimator (needs models for $e(X)$, $\mathbb{E}\{Y(0)|X\}$, and $\mathbb{E}\{Y(1)|X\}$)

$$\begin{aligned}\widehat{\tau}_{watt}^{\text{aug-II}} =& \frac{\sum_{i=1}^N Z_i\{Y_i - \widehat{m}_1(X_i)\}}{\sum_{i=1}^N Z_i} - \frac{\sum_{i=1}^N \frac{\widehat{h}(X_i)\widehat{e}(X_i)}{1-\widehat{e}(X_i)}(1-Z_i)\{Y_i - \widehat{m}_0(X_i)\}}{\sum_{i=1}^N \widehat{h}(X_i)\widehat{e}(X_i)} \\ &+ \frac{\sum_{i=1}^N \widehat{e}(X_i)\widehat{m}_1(X_i)}{\sum_{i=1}^N \widehat{e}(X_i)} - \frac{\sum_{i=1}^N \widehat{h}(X_i)\widehat{e}(X_i)\widehat{m}_0(X_i)}{\sum_{i=1}^N \widehat{h}(X_i)\widehat{e}(X_i)},\end{aligned}$$

  where $\widehat{m}_z(X) = \widehat{\mathbb{E}}\{Y(z)|X\}$ for $z = 0, 1$.

Remark All these estimators are consistent when needed models are correctly specified. Augmented estimators are not in general double-robust, but they have good finite-sample performances. All these estimators are asymptotic linear, which enables the use of bootstrap for variance estimation. Type-I augmented estimator does not require estimating $m_1(X)$ compared to type-II, but the latter one is **locally efficient** in the sense that if needed models are correctly specified, it achieves the semiparametric variance bound (next page).

Asymptotic Properties

Theorem

Under regularity conditions[6]–[8] and correctly specified propensity score and outcomes models, assume $|e(X)h(X)|$ is bounded with $\mathbb{E}\{e(X)h(X)\} > 0$, then $\sqrt{N}(\widehat{\tau}^{aug\text{-}II}_{watt} - \tau_{watt}) \xrightarrow{d} \mathcal{N}(0, \mathsf{V})$, where

$$\begin{aligned}\mathsf{V} = &\mathbb{E}\left\{\frac{e(X)}{\mu_e^2}\mathbb{V}\{Y(1)|X\}\right\} + \mathbb{E}\left\{\frac{e(X)^2h(X)^2}{\mu_{eh}^2(1-e(X))}\mathbb{V}\{Y(0)|X\}\right\} \\ &+ \mathbb{E}\left\{\frac{e(X)^2}{\mu_e^2}\{\mathbb{E}\{Y(1)|X\} - \tau^{(1)}_{watt}\}^2\right\} + \mathbb{E}\left\{\frac{e(X)^2h(X)^2}{\mu_{eh}^2}\{\mathbb{E}\{Y(0)|X\} - \tau^{(0)}_{watt}\}^2\right\} \\ &+ 2\mathbb{E}\left\{\frac{e(X)^2h(X)}{\mu_e\mu_{eh}}\{\mathbb{E}\{Y(1)|X\} - \tau^{(1)}_{watt}\}\{\mathbb{E}\{Y(0)|X\} - \tau^{(0)}_{watt}\}\right\},\end{aligned}$$

where (1) V is also the ***semiparametric variance bound*** *for all WATT estimators; (2) $\tau^{(1)}_{watt} = \mathbb{E}\{Y|Z=1\}$, $\tau^{(0)}_{watt} = \mathbb{E}\left\{(1-Z)Y\frac{e(X)h(X)}{1-e(X)}\right\}\frac{1}{\mathbb{E}\{e(X)h(X)\}}$; (3) $\mu_b = \int b(x)f(x)\mu(dx)$ for a function $b(\cdot)$, and $f(\cdot)$ is the density function for X under some base measure $\mu(\cdot)$.*

In short, $\widehat{\tau}^{aug\text{-}II}_{watt}$ is consistent and (locally) efficient under correctly specified models.

Similar arguments and results apply to WATC.

Inference

- Using the typical empirical estimator of **V** for variance estimation is not realistic. How to get the empirical version of $\mathbb{V}\{Y(z)|X\}$ for $z = 0, 1$?
- We have also shown that all our point estimators are asymptotic linear everywhere, if the models of propensity score and outcomes are correctly specified by some generalized linear models, which enables the use of **bootstrap variance estimation**.
- We also developed the close-form **sandwich variance estimators** for EATT and EATC estimators, which involve uncertainty in estimating propensity score, potential outcomes, and the estimands themselves.
  The unbiased estimating equation in most general form:

$$\sum_{i=1}^{N} \Psi(X_i, Y_i, Z_i; \theta) = \sum_{i=1}^{N} \begin{bmatrix} \psi_\beta(X_i, Z_i) \\ Z_i \psi_{\alpha_1}(X_i, Y_i) \\ (1 - Z_i)\psi_{\alpha_0}(X_i, Y_i) \\ e(X_i)\{m_1(X_i) - \eta_1\} \\ e(X_i)h(X_i)\{m_0(X_i) - \eta_0\} \\ Z_i\{Y_i - m_1(X_i) - \mu_1\} \\ (1 - Z_i)\omega_0(X_i)\{Y_i - m_0(X_i)\} - e(X_i)h(X_i)\mu_0 \end{bmatrix} = 0$$

  $\theta = (\beta', \alpha_1', \alpha_0', \eta_1, \eta_0, \mu_1, \mu_0)'$, so the estimand of interest is $c'\theta$ with $c = (\tilde{\mathbf{0}}, 1, -1, 1, -1)'$, and $\tilde{\mathbf{0}}$ has the same demesion of $(\beta', \alpha_1', \alpha_0')'$.
  Under certain regularity conditions, the variance can be consistently estimated by $\frac{1}{N} c' A_N(\widehat{\theta})^{-1} B_N(\widehat{\theta}) \{A_N(\widehat{\theta})^{-1}\}' c$, where

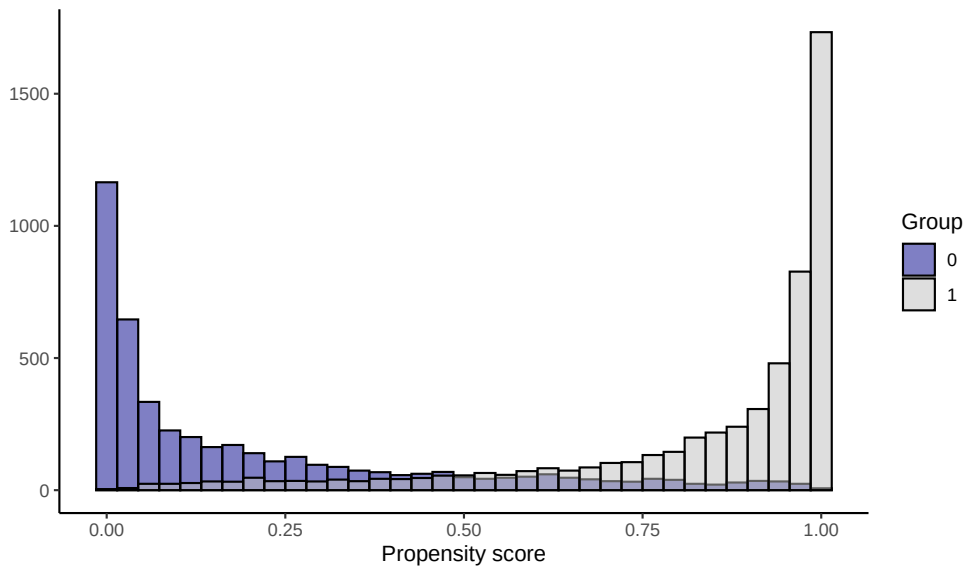
$$A_N(\widehat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \left\{ -\frac{\partial \Psi(X_i, Y_i, Z_i; \widehat{\theta})}{\partial \theta} \right\}, \qquad B_N(\widehat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \Psi(X_i, Y_i, Z_i; \widehat{\theta}) \Psi(X_i, Y_i, Z_i; \widehat{\theta})'$$

Simulation

Data generating process

- $X_1 \sim \mathcal{N}(0, 1)$
- $X_2 = \sqrt{0.3}X_1 + \sqrt{0.7}A$, $A \sim \mathcal{N}(0, 4)$
- $Z \sim Bern(p)$ where $p = \{1 + \exp\{-0.5 + X_1 + 2X_2\}\}^{-1}$
- $Y(0) = 0.25(X_1^2 + X_2) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 1)$
- $Y(1) = Y(0) + 3$

Propensity score distribution (using a random sample of $N = 10000$): extreme IPW weights exist (some control participants have PS close to 1)



Simulation

Monte Carlo arguments

- Sample size $N = 1000$ in each independent replication
- Number of MC replications $M = 500$

Results under correctly specified models

Method	ARBias%	MSE	Std. errors	Relative efficiency	Coverage
Simple weighting estimator					
ATT	0.56	0.61	0.42	0.55	1.00
ATT trimming ($\alpha = 0.1$)	0.06	0.00	0.15	0.20	1.00
EATT	0.05	0.01	0.17	0.46	0.99
Type-I augmented estimator					
ATT	0.37	0.32	0.27	1.28	0.94
ATT trimming ($\alpha = 0.1$)	4.06	0.03	0.12	0.97	0.85
EATT	0.03	0.01	0.11	0.94	0.95
Type-II augmented estimator					
ATT	0.10	0.09	0.20	2.00	0.85
ATT trimming ($\alpha = 0.1$)	4.02	0.04	0.15	1.03	0.89
EATT	0.03	0.01	0.11	0.93	0.95

Concluding Remarks

Simulation findings

- EATT performs well, results in smallest relative bias and is the most precise method
- Augmented estimators improved efficiency (using sandwich variance estimator)
- When extreme weights exist, trimming can still result in bad performance

Summary

- When extreme IPW weights exist for identifying ATT from observational data, classical ATT estimators can be unreliable and unstable in quantifying causal relationship. Trimming is ad hoc and there is no general rule for a trimming threshold, and we have seen it may "overkill". We propose an alternative causal estimand EATT and its inferential framework, which is a smoothed and non-ad hoc strategy with more internal validity for targeting equipoise subpopulation on the control participants.

Future work

- Extend the simulation setting
- Apply methods to real data
- Wild bootstrap variance estimation[9]
- Extensions to multi-level treatments[10]
- ...

References

[1] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[2] R. Crump, V. J. Hotz, G. Imbens, and O. Mitnik, *Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand*, 2006.

[3] F. Li, L. E. Thomas, and F. Li, "Addressing extreme propensity scores via the overlap weights," *American journal of epidemiology*, vol. 188, no. 1, pp. 250–257, 2018.

[4] R. A. Matsouaka and Y. 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.

[5] L. Li and T. Greene, "A weighting analogue to pair matching in propensity score analysis," *The international journal of biostatistics*, vol. 9, no. 2, pp. 215–234, 2013.

[6] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[7] K. Hirano and G. W. Imbens, "Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization," *Health Services and Outcomes research methodology*, vol. 2, pp. 259–278, 2001.

[8] K. Hirano, G. W. Imbens, and G. Ridder, "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, vol. 71, no. 4, pp. 1161–1189, 2003.

[9] R. A. Matsouaka, Y. Liu, and Y. Zhou, "Variance estimation for the average treatment effects on the treated and on the controls," *Statistical Methods in Medical Research*, vol. 32, no. 2, pp. 389–403, 2023.

[10] F. Li and F. L, "Propensity score weighting for causal inference with multiple treatments," *The Annals of Applied Statistics*, vol. 13, no. 4, pp. 2389–2415, 2019.

Acknowledgement Roland Matsouaka, PhD (Biostat & Bioinfo, Duke University)

Thank you!