The 35th NESS Speed Poster Presentation

Overlap, Inverse Probability, and Matching Weights: What Are We Weighting For?

Yi Liu

(Incoming) PhD student
Department of Statistics
North Carolina State University

May 23, 2022

Introduction

- Propensity score (PS) weighting methods usually play a central role in darwing causal conclusions from observational data
- Some PS weights: overlap weights (OW), matching weights (MW), entropy weights (EW) and standard IPW weights. What should we expect when using them?
  - their estimands, target population, weights...
  - sensitivity to model misspecifications
- What is the role of the proportion of treated participants, i.e. $p = P(Z = 1)$, in estimating these estimands?

Why do we care about the impact of p?

- When exposure is rare (small p) and the treated population is of interest, ATT is encouraged, but ATE reaches extreme values and larger biases[1] to what we want
  - Rare exposure could be frequently encountered in pharmacoepidemiologic observational studies, study design does not require a high prevalence of exposure (EHR data, evaluation of newly approved drugs, etc.), ...
  - Then who is closer to ATT, which can also be useful in this case?
  - Similar issues may exist for ATC when p is large...

[1] D. Hajage, F. Tubach, P. G. Steg, *et al.*, "On the use of propensity scores in case of rare exposure," *BMC medical research methodology*, vol. 16, no. 1, p. 38, 2016.

Notation

- Treatment: $Z = 0, 1$
- Covariates: $\boldsymbol{X} = (X_1, ..., X_p)'$
- Propensity score: $e(\boldsymbol{x}) = P(Z = 1 | \boldsymbol{X} = \boldsymbol{x})$
- Potential outcome notation: $Y(z), z = 0, 1$ – associated with the treatment assignment
- A generalized class of ATE–*weighted* ATE (WATE)[2]:

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

  - $\tau(\boldsymbol{x}) = E[Y(1) - Y(0) | \boldsymbol{X} = \boldsymbol{x}]$; $g(\boldsymbol{x})$: *selection function* that re-distributes the covariates
  - Different g defines different causal estimand and target population

[2] F. Li, K. L. Morgan, and A. M. Zaslavsky, "Balancing covariates via propensity score weighting," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 390–400, 2018.

Choice of g

Table: Choices of g, corresponding target population and causal estimands[3]

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

IPW: inverse probability weights;

OW (resp. MW, EW): overlap (resp. matching, entropy) weights;

We choose $\alpha = 0.05, 0.1$ and 0.15.

So, all g we considered is a function of e–the propensity score.

[3] 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.

Hájek-type Estimation

Data $\{(\boldsymbol{X}_i, Y_i, Z_i), i = 1, \ldots, N\}$. $\tau_g = \dfrac{E[g(\boldsymbol{X})\tau(\boldsymbol{X})]}{E[g(\boldsymbol{X})]}$ can be estimated by the *Hájek-type estimator*

$$\widehat{\tau}_g^{\mathrm{H}} = \frac{\sum_{i=1}^N Z_i \widehat{w}_1(\boldsymbol{x}_i) Y_i}{\sum_{i=1}^N Z_i \widehat{w}_1(\boldsymbol{x}_i)} - \frac{\sum_{i=1}^N (1 - Z_i) \widehat{w}_0(\boldsymbol{x}_i) Y_i}{\sum_{i=1}^N (1 - Z_i) \widehat{w}_0(\boldsymbol{x}_i)}$$

$\widehat{w}_z(\boldsymbol{x}),\ z = 0, 1$ is calculated by a propensity score (PS) model (usually logstic regression).

Note: The *balancing weights*[4]: $(w_0(\boldsymbol{x}), w_1(\boldsymbol{x})) \propto \left(\dfrac{g(\boldsymbol{x})}{1 - e(\boldsymbol{x})}, \dfrac{g(\boldsymbol{x})}{e(\boldsymbol{x})} \right)$

[4] F. Li, K. L. Morgan, and A. M. Zaslavsky, "Balancing covariates via propensity score weighting," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 390–400, 2018.

Augmented Estimation

Data $\{(\boldsymbol{X}_i, Y_i, Z_i), i = 1, \ldots, N\}$. The *augmented estimator* of τ_g is given by

$$\widehat{\tau}_g^{\text{aug}} = \frac{\sum_{i=1}^N g(\boldsymbol{x}_i)\{\widehat{m}_1(\boldsymbol{x}_i) - \widehat{m}_0(\boldsymbol{x}_i)\}}{\sum_{i=1}^N g(\boldsymbol{x}_i)} + \frac{\sum_{i=1}^N Z_i \widehat{w}_1(\boldsymbol{x}_i)\{Y_i - \widehat{m}_1(\boldsymbol{x}_i)\}}{\sum_{i=1}^N Z_i \widehat{w}_1(\boldsymbol{x}_i)} - \frac{\sum_{i=1}^N (1 - Z_i)\widehat{w}_0(\boldsymbol{x}_i)\{Y_i - \widehat{m}_0(\boldsymbol{x}_i)\}}{\sum_{i=1}^N (1 - Z_i)\widehat{w}_0(\boldsymbol{x}_i)}$$

where $m_z(\boldsymbol{X}) = E(Y(z)|\boldsymbol{X}),\ z = 0, 1$ is an outcome regression (OR) model.

- Augmentation = Hájek-type + modeling the outcome
- For ATE, ATT and ATC, we can actually use their *doubly robust (DR) estimator*, which have better large-sample properties. For overlap estimands, augmented estimators are not DR.

Sandwich Variance Estimation

- Both $\widehat{\tau}_g^{\mathrm{H}}$ and $\widehat{\tau}_g^{\mathrm{aug}}$ are consistent if the PS model is correctly specified[5];
- Asymptotic normality under regularity conditions[6]:
  $\sqrt{N}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, A(\theta)B(\theta)\{A(\theta)'\}^{-1})$;
- Score equation:

$$\sum_{i=1}^{N} \Psi_\theta(\boldsymbol{X}_i, Z_i, Y_i) = \sum_{i=1}^{N} \begin{bmatrix} \psi_\beta(\boldsymbol{X}_i, Z_i) \\ Z_i \psi_{\alpha_1}(\boldsymbol{X}_i, Y_i) \\ (1 - Z_i)\psi_{\alpha_0}(\boldsymbol{X}_i, Y_i) \\ g(\boldsymbol{X}_i)\{m_1(\boldsymbol{X}_i) - \tau_{1g}^m\} \\ g(\boldsymbol{X}_i)\{m_0(\boldsymbol{X}_i) - \tau_{0g}^m\} \\ Z_i w_1(\boldsymbol{X}_i)(Y_i - m_1(\boldsymbol{X}_i) - \mu_{1g}) \\ (1 - Z_i) w_0(\boldsymbol{X}_i)(Y_i - m_0(\boldsymbol{X}_i) - \mu_{0g}) \end{bmatrix} = 0$$

- $\theta = (\beta', \alpha_1', \alpha_0', \tau_{1h}^m, \tau_{0g}^m, \mu_{1g}, \mu_{0g})'$

---

[5] H. Mao, L. Li, and T. Greene, "Propensity score weighting analysis and treatment effect discovery," *Statistical methods in medical research*, vol. 28, no. 8, pp. 2439–2454, 2019.

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

Simulation - DGP

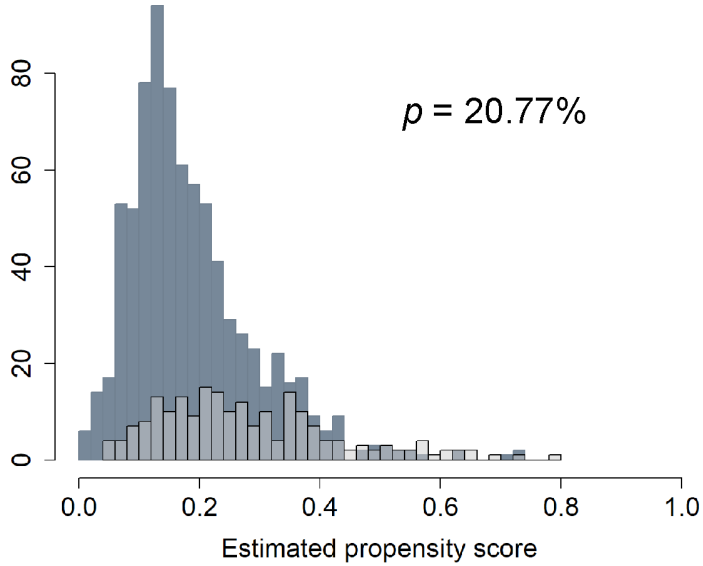
- 7 covariates $\boldsymbol{X} = (X_1, X_2, ..., X_7)'$ (following Li and Li[7])
- Treatment: $Z \sim \text{Bern}(\text{expit}(\boldsymbol{X}\beta))$ (logistic regression)
- Outcome model:
  $Y(0) = 0.5 + X_1 + 0.6X_2 + 2.2X_3 - 1.2X_4 + (X_1 + X_2)^2 + \varepsilon$ and
  $Y(1) = Y(0) + \delta(\boldsymbol{X})$, for $\varepsilon \sim N(0, 4)$ (linear regression)
  - True treatment effect: $\delta(\boldsymbol{X}) = 4 + 3(X_1 + X_2)^2 + X_1 X_3$
- Sample size $N = 1000$ with 2000 replications
- We **only** consider model misspecifications for augmented estimators. For Hájek-type estimator, past research has shown that overlap estimators are more robust than the IPW estimator[8].

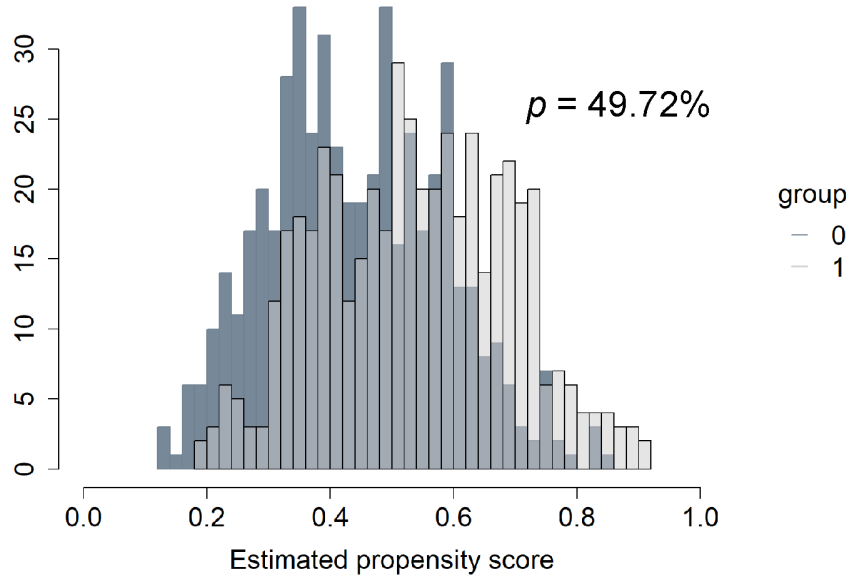
[7] Y. Li and L. Li, "Propensity score analysis methods with balancing constraints: A monte carlo study," *Statistical Methods in Medical Research*, vol. 30, no. 4, pp. 1119–1142, 2021.

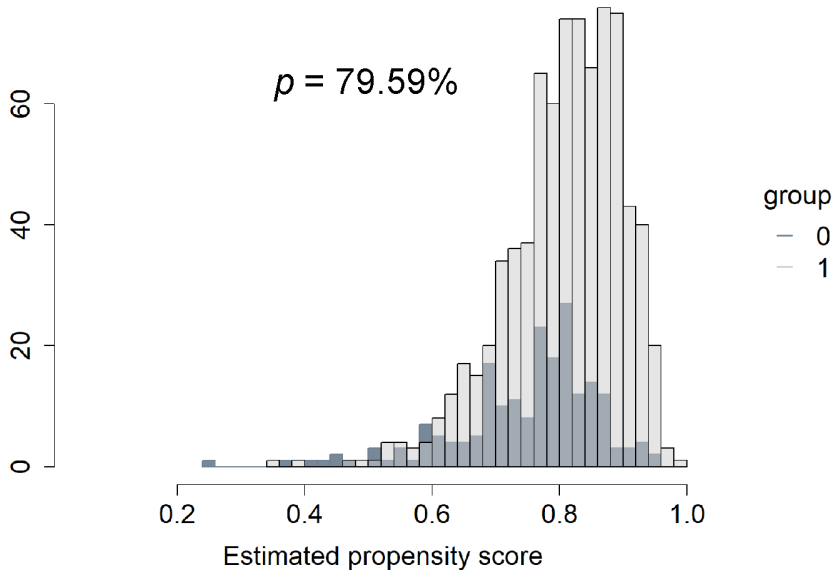
[8] Y. Zhou, R. A. Matsouaka, and L. Thomas, "Propensity score weighting under limited overlap and model misspecification," *Statistical Methods in Medical Research*, vol. 29, no. 12, pp. 3721–3756, 2020.

Simulation - Propensity Score

- Generate 3 PS models, with small, middle and large p





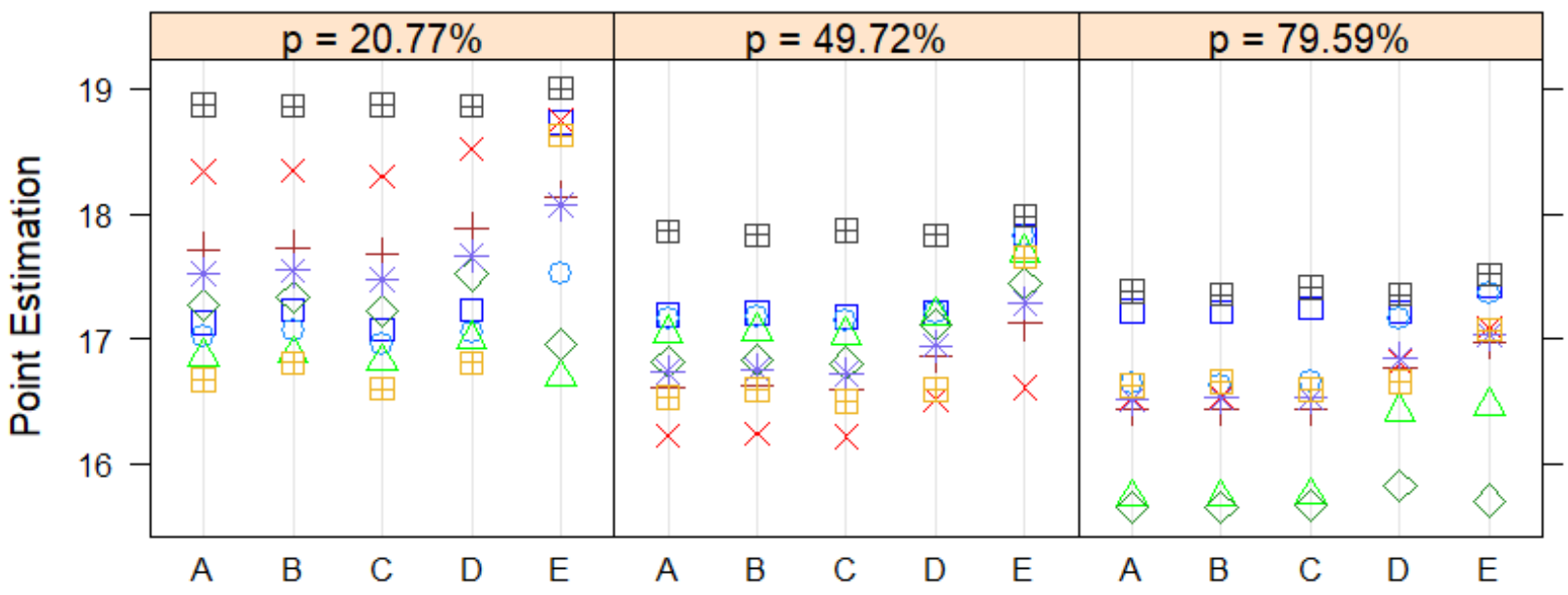


Simulation - Point Estimation

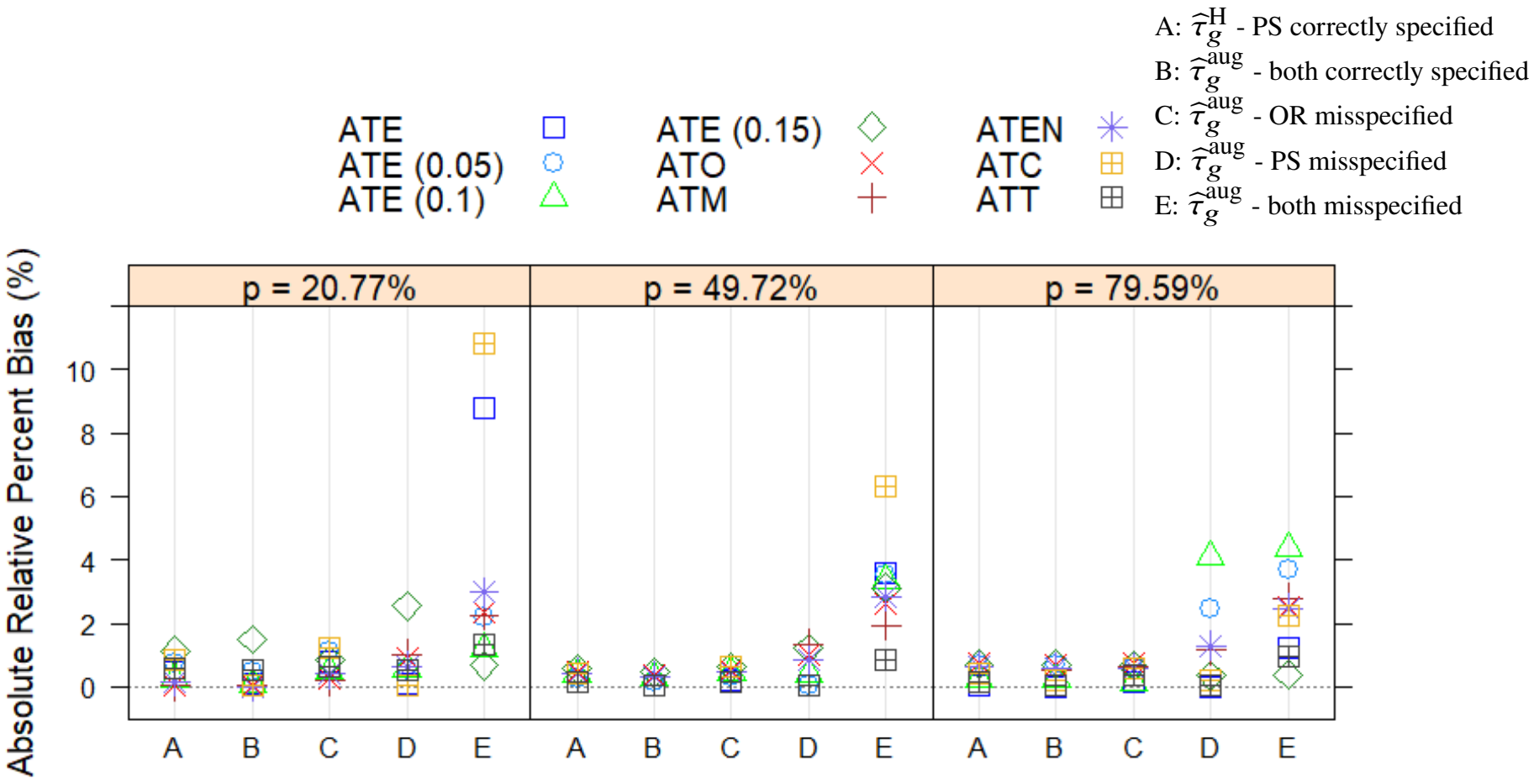
True treatment effect:

p	r	ATE	ATE (0.05)	ATE (0.1)	ATE (0.15)	ATO	ATM	ATEN	ATC	ATT
20.77%	1.80	17.22	17.15	16.90	17.08	17.72	18.33	17.55	16.81	18.76
49.72%	1.13	17.22	17.20	17.12	16.91	16.69	16.30	16.81	16.61	17.83
79.59%	0.75	17.22	16.75	15.78	15.77	16.56	16.63	16.63	16.70	17.35

A: $\widehat{\tau}_g^{\mathrm{H}}$ - PS correctly specified
B: $\widehat{\tau}_g^{\mathrm{aug}}$ - both correctly specified
C: $\widehat{\tau}_g^{\mathrm{aug}}$ - OR misspecified
D: $\widehat{\tau}_g^{\mathrm{aug}}$ - PS misspecified
E: $\widehat{\tau}_g^{\mathrm{aug}}$ - both misspecified

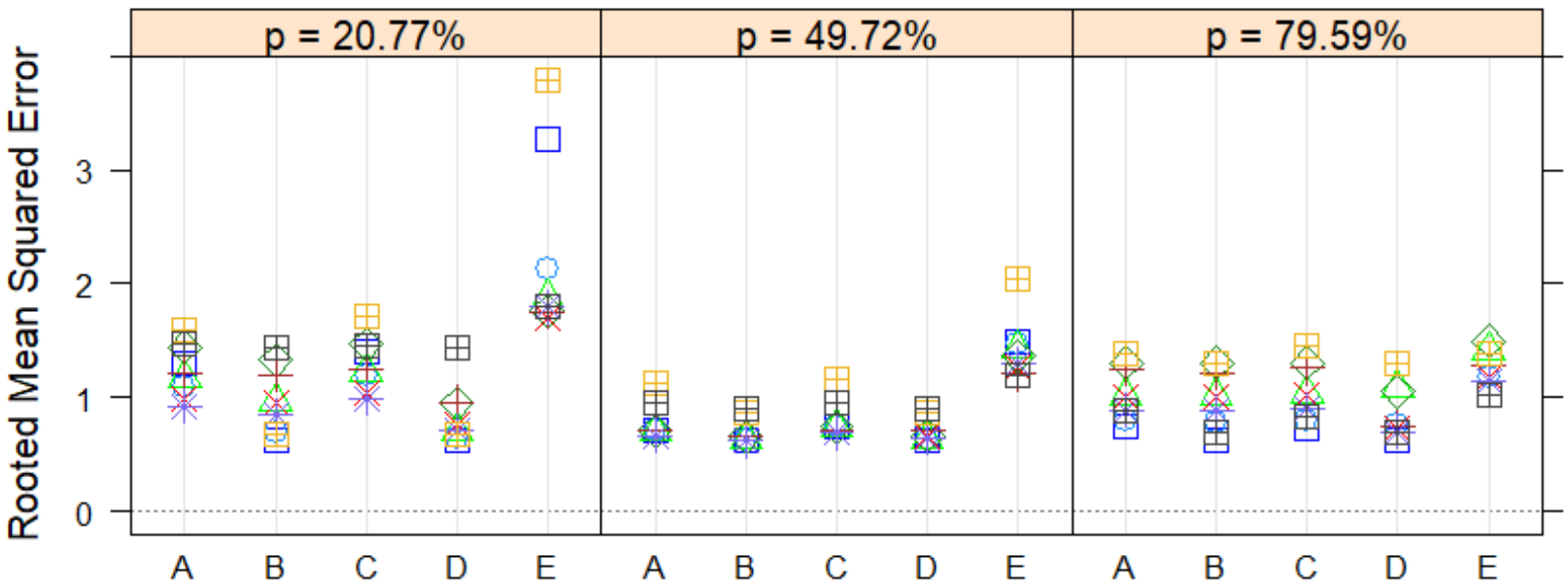


Simulation - ARBias (%)

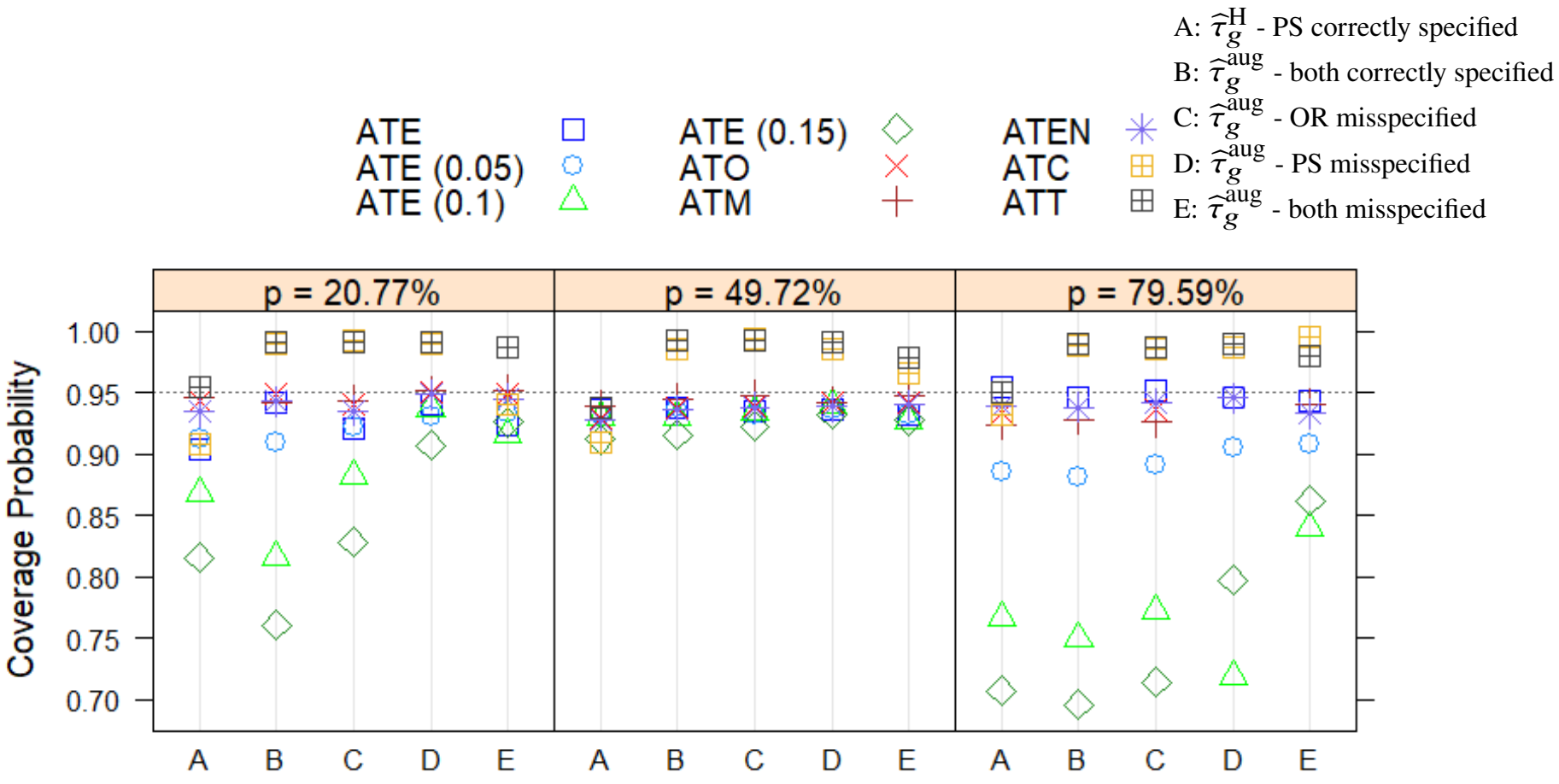


Simulation - RMSE

A: $\widehat{\tau}_g^{\mathrm{H}}$ - PS correctly specified
B: $\widehat{\tau}_g^{\mathrm{aug}}$ - both correctly specified
C: $\widehat{\tau}_g^{\mathrm{aug}}$ - OR misspecified
D: $\widehat{\tau}_g^{\mathrm{aug}}$ - PS misspecified
E: $\widehat{\tau}_g^{\mathrm{aug}}$ - both misspecified



Simulation - Coverage Probability



Simulation - Summary I: Sensitivity Assessment

- The IPW estimates (w/ or w/o trimming) for ATE sometimes have apprently larger biases and errors than overlap estimates for their estimands–more robust!
- The coverage probabilities suggest that the close-form sandwich variance formula works better for overlap estimands

Simulation - Summary II: Impact of p

- When p is high, overlap estimators (ATO, ATM and ATEN) weight toward ATC (not exactly very close), and vice versa
- When $p \approx 0.5$ and no extreme weights exist, IPW and overlap estimations are similar
- The variances of propensity scores of both treated and control groups also play a role–see my poster and a future publication!

The impact of p–is this really a surprise?

- First, clearly ATE = pATT+$(1-p)$ATC
- Second, the overlap weights (OW) $\propto (e(\boldsymbol{x}), 1-e(\boldsymbol{x}))$, so[9]
  - when $e(\boldsymbol{x}) \approx 0.5$, $(e(\boldsymbol{x}), 1-e(\boldsymbol{x})) \approx \left(\frac{0.25}{1-e(\boldsymbol{x})}, \frac{0.25}{e(\boldsymbol{x})}\right)$ (ATE weights)
  - when $e(\boldsymbol{x})$ is small, $(e(\boldsymbol{x}), 1-e(\boldsymbol{x})) \approx \left(\frac{e(\boldsymbol{x})}{1-e(\boldsymbol{x})}, 1\right)$ (ATT weights)
  - when $e(\boldsymbol{x})$ is large, $(e(\boldsymbol{x}), 1-e(\boldsymbol{x})) \approx \left(1, \frac{1-e(\boldsymbol{x})}{e(\boldsymbol{x})}\right)$ (ATC weights)
- Note $p = P(Z=1) = E[e(\boldsymbol{X})]$, so we conjecture that under some conditions, p–the first moment of the propensity score, might reflect how overlap estimators weight ATT and ATC
- MW and EW are similar to OW[10]!

[9] F. Li, K. L. Morgan, and A. M. Zaslavsky, "Balancing covariates via propensity score weighting," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 390–400, 2018.

[10] 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.

Data Example - treatment effects of RHC

- 5735 hospitalized patients enrolled a right heart catheterization (RHC) study
- **2184 (38%)** patients received the RHC treatment (small p)
- Outcome: log-length of stay during the first 24 hours in ICU

Table: Treatment effects of the RHC procedure on patients

		Hájek-type Estimate			**Augmented Estimate**		
Estimand	Prop.	Est.	SE	p-value	Est.	SE	p-value
ATE	100%	**0.130**	0.032	<0.001	**0.129**	0.033	<0.001
ATE (0.05)	93%	0.099	0.030	<0.001	0.102	0.030	<0.001
ATE (0.1)	82%	0.102	0.029	<0.001	0.100	0.029	<0.001
ATE (0.15)	73%	0.078	0.029	0.008	0.079	0.029	0.007
ATO	100%	**0.095**	0.028	<0.001	**0.098**	0.028	<0.001
ATM	100%	**0.094**	0.028	<0.001	**0.095**	0.028	<0.001
ATEN	100%	**0.100**	0.028	<0.001	**0.102**	0.028	<0.001
ATC	100%	**0.156**	0.035	<0.001	**0.148**	0.037	<0.001
ATT	100%	**0.090**	0.046	0.049	**0.099**	0.043	0.021

Prop.: proportion of sample used; Est.: point estimation; SE: standard error

Concluding Remarks

- This work showed why sometimes ATE fails to identify the logical treatment effect, and provided some facts to researchers when they make choices according to what they want to accomplish
- Overlap estimators are overall more robust to model misspecifications, and have better coverage probabilities using the close-form sandwich variance estimations

Upcoming Research

Variance estimation for equipoise treatment effect estimators (ATO, ATM, ATEN, BET)

- Sandwich
- Standard bootstrap
- Wild bootstrap (works very well for ATT and ATC!)
- ...

References

[1] D. Hajage, F. Tubach, P. G. Steg, D. L. Bhatt, and Y. De Rycke, "On the use of propensity scores in case of rare exposure," *BMC medical research methodology*, vol. 16, no. 1, p. 38, 2016.

[2] F. Li, K. L. Morgan, and A. M. Zaslavsky, "Balancing covariates via propensity score weighting," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 390–400, 2018.

[3] 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.

[4] H. Mao, L. Li, and T. Greene, "Propensity score weighting analysis and treatment effect discovery," *Statistical methods in medical research*, vol. 28, no. 8, pp. 2439–2454, 2019.

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

[6] Y. Li and L. Li, "Propensity score analysis methods with balancing constraints: A monte carlo study," *Statistical Methods in Medical Research*, vol. 30, no. 4, pp. 1119–1142, 2021.

[7] Y. Zhou, R. A. Matsouaka, and L. Thomas, "Propensity score weighting under limited overlap and model misspecification," *Statistical Methods in Medical Research*, vol. 29, no. 12, pp. 3721–3756, 2020.

Related Working Paper

Variance estimation for the average treatment effects on the treated and on the controls.

Journal Title
XX(X):1–25

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Roland A. Matsouaka[1,2], Yi Liu[1], and Yunji Zhou[1,3]

Revision (April 17th): Statistical Methods in Medical Research

ARTICLE TYPE

Overlap weights: what are we weighting for?

Roland A. Matsouaka[1,2] | Yi Liu[1] | Yunji Zhou[1,3]

Close to submit: Statistics in Medicine

Acknowledgements

- Roland Matsouaka (Duke Biostatistics & Bioinformatics)
- Yunji Zhou (Duke Global Health Institute, Duke Biostatistics & Bioinformatics)
- Fan Li (Duke Statistical Science)
- Hwanhee Hong (Duke Biostatistics & Bioinformatics)

Thank you!

Yi Liu

yliu297@ncsu.edu