Varying impacts of letters of recommendation on college admissions

Approximate balancing weights for subgroup effects in observational studies

Eli Ben-Michael

Harvard University

(joint work with Avi Feller and Jesse Rothstein)

JSM, August 2021

- *2016:* UC Berkeley pilot program requests **letters of recommendation** (LORs) for undergrad admission
    - No LOR requirement at other UCs/CSUs

- *Goal:* Identify students from non-traditional backgrounds who might be overlooked
    - “Holistic review” of applicants

- *Concern:* Adverse impact on disadvantaged applicants, especially under-represented minority (URM) students

Chair of the UC Berkeley Academic Senate

*"**LORs conflict with UC principles of access and fairness**, because students attending under-resourced schools or from disadvantaged backgrounds will find it more difficult to obtain high-quality letters, and could be disadvantaged by a LOR requirement"*

Chair of the UC Berkeley Academic Senate

*"**LORs conflict with UC principles of access and fairness**, because students attending under-resourced schools or from disadvantaged backgrounds will find it more difficult to obtain high-quality letters, and could be disadvantaged by a LOR requirement"*

University Committee on Affirmative Action, Diversity, and Equity

*"The burden of proof rests on those who want to implement the new letters of recommendation policy, and should include a test of statistical significance demonstrating **measurable impact on increasing diversity** in undergraduate admissions"*

Chair of the UC Berkeley Academic Senate

*"**LORs conflict with UC principles of access and fairness**, because students attending under-resourced schools or from disadvantaged backgrounds will find it more difficult to obtain high-quality letters, and could be disadvantaged by a LOR requirement"*

University Committee on Affirmative Action, Diversity, and Equity

*"The burden of proof rests on those who want to implement the new letters of recommendation policy, and should include a test of statistical significance demonstrating **measurable impact on increasing diversity** in undergraduate admissions"*

UC Berkeley administration requested independent review of LOR impact [Rothstein, 2017]

Chair of the UC Berkeley Academic Senate

*"**LORs conflict with UC principles of access and fairness**, because students attending under-resourced schools or from disadvantaged backgrounds will find it more difficult to obtain high-quality letters, and could be disadvantaged by a LOR requirement"*

University Committee on Affirmative Action, Diversity, and Equity

*"The burden of proof rests on those who want to implement the new letters of recommendation policy, and should include a test of statistical significance demonstrating **measurable impact on increasing diversity** in undergraduate admissions"*

UC Berkeley administration requested independent review of LOR impact [Rothstein, 2017]

⇝ **LORs discontinued** before study results released

LOR Pilot Program: Subgroup Effects

Our question: Impact of submitting LORs on admissions

- Variation across pre-defined subgroups, especially URM status

→ Design an observational study; one of many potential cuts at this problem

LOR Pilot Program: Subgroup Effects

Our question: Impact of submitting LORs on admissions

- Variation across pre-defined subgroups, especially URM status

$\rightarrow$ Design an observational study; one of many potential cuts at this problem

Challenge: Design the study for good overall and subgroup estimates

- Only optimizing for global balance $\rightarrow$ poor subgroup estimates
  - Ignores subgroup structure, assumes away heterogeneity
- Only optimizing for local balance $\rightarrow$ poor overall estimates
  - Potentially small errors compound across subgroups

LOR Pilot Program: Subgroup Effects

Our question: Impact of submitting LORs on admissions

- Variation across pre-defined subgroups, especially URM status

$\rightarrow$ Design an observational study; one of many potential cuts at this problem

Challenge: Design the study for good overall and subgroup estimates

- Only optimizing for global balance $\rightarrow$ poor subgroup estimates
  - Ignores subgroup structure, assumes away heterogeneity
- Only optimizing for local balance $\rightarrow$ poor overall estimates
  - Potentially small errors compound across subgroups

Our paper: Balancing weights for subgroup analysis

- Partially pooled balancing weights, control both local balance and global balance
- *Dual:* IPW with hierarchical propensity score

LOR Pilot Program: Subgroup Effects

Our question: Impact of submitting LORs on admissions

- Variation across pre-defined subgroups, especially URM status

$\rightarrow$ Design an observational study; one of many potential cuts at this problem

Challenge: Design the study for good overall and subgroup estimates

- Only optimizing for global balance $\rightarrow$ poor subgroup estimates
  - Ignores subgroup structure, assumes away heterogeneity
- Only optimizing for local balance $\rightarrow$ poor overall estimates
  - Potentially small errors compound across subgroups

Our paper: Balancing weights for subgroup analysis

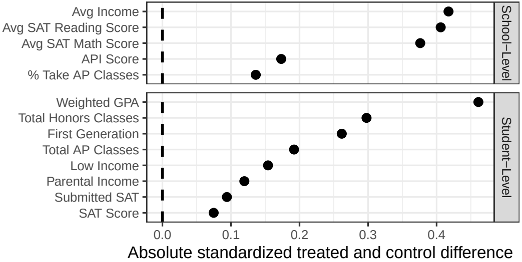
- Partially pooled balancing weights, control both local balance and global balance
- *Dual:* IPW with hierarchical propensity score

No evidence of differential impacts on URM applicants

Letters of Recommendation: Pilot Study

LOR Pilot Study: Overview

- Total $N = 40,541$ applicants in 2016
  [exclude athletes, other groups]
  - 14,596 invited to submit LORs
  - 11,143 submitted LoRs
- Two admissions readers
  - Scores of {No, Possible, Yes}
  - Admitted with 1-2 Yes votes
- Invitation to submit LORs:
  [+ funkiness due to timing]
  - First reader score of "possible"
  - Predicted possible score of >50%
    [nearly all URM]

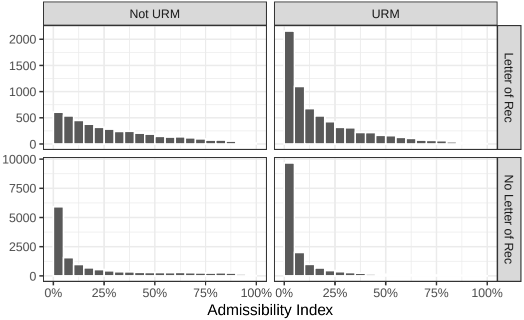


LoR Pilot Study: Subgroups

URM: Under-Represented Minority

- Low-income or first-gen college
- Underrepresented racial/ethnic group
- Low-performing high school

  [$\sim$ 55% of all applicants]



AI: Admissibility Index

- Predicted prob. of admissions using 2015 data

Define subgroups by URM $\times$ AI bin

+ First reader score; college applied to

Setup and Background

Setup

For applicant $i = 1, \ldots, N$ observe

- Outcome $Y_i \in \mathbb{R}$ (admission)
- Treatment status $W_i \in \{0, 1\}$ (submit LoRs)
- Covariates $X_i \in \mathcal{X}$
- Group indicator $G_i \in \{1, \ldots, K\}$

Setup

For applicant $i = 1, \ldots, N$ observe

- Outcome $Y_i \in \mathbb{R}$ (admission)
- Treatment status $W_i \in \{0, 1\}$ (submit LoRs)
- Covariates $X_i \in \mathcal{X}$
- Group indicator $G_i \in \{1, \ldots, K\}$

Estimands: Overall ATT and subgroup CATTs

$$\tau = \mathbb{E}[Y(1) - \underbrace{Y(0)}_{\widehat{\mu}_0 = \sum \hat{\gamma}_i Y_i} \mid W = 1] \quad \text{and} \quad \tau_g = \mathbb{E}[Y(1) - \underbrace{Y(0)}_{\widehat{\mu}_{0g} = \sum_{G=g} \hat{\gamma}_i Y_i} \mid W = 1, G = g]$$

Strong ignorability (sensitivity analysis in paper)

$$Y(1), Y(0) \perp\!\!\!\perp W \mid X, G \qquad \text{and} \qquad e(X, G) \equiv P(W = 1 \mid X, G) < 1$$

Setup

Strong ignorability (sensitivity analysis in paper)

$$Y(1), Y(0) \perp\!\!\!\perp W \mid X, G \qquad \text{and} \qquad e(X, G) \equiv P(W = 1 \mid X, G) < 1$$

Many methods for subgroup effects under ignorability

- **Outcome model** and **design-based** approaches
- *Review:* 2018 ACIC data challenge [Carvalho et al., 2020]

Inverse Propensity Score Weighting identities, with known $e(x, g)$

$$\mu_0 \;=\; \mathbb{E}[\text{missing } Y(0) \mid \text{treated}] = \mathbb{E}\Big[\underbrace{\frac{e(x,g)}{1-e(x,g)}}_{\text{weights}} \; Y^{\text{obs}} \mid \text{control}\Big]$$

Inverse Propensity Score Weighting identities, with known $e(x, g)$

$$\mu_0 \;=\; \mathbb{E}[\text{missing } Y(0) \mid \text{treated}] = \mathbb{E}\Big[\underbrace{\frac{e(x,g)}{1-e(x,g)}}_{\text{weights}} \; Y^{\text{obs}} \mid \text{control}\Big]$$

$$\mathbb{E}[\text{covariates} \mid \text{treated}] = \mathbb{E}\Big[\underbrace{\frac{e(x,g)}{1-e(x,g)}}_{\text{weights}} \; \text{covariates} \mid \text{control}\Big]$$

Inverse Propensity Score Weighting identities, with known $e(x, g)$

$$\mu_0 \;=\; \mathbb{E}[\text{missing } Y(0) \mid \text{treated}] = \mathbb{E}\Big[\underbrace{\frac{e(x,g)}{1-e(x,g)}}_{\text{weights}}\; Y^{\text{obs}} \mid \text{control}\Big]$$

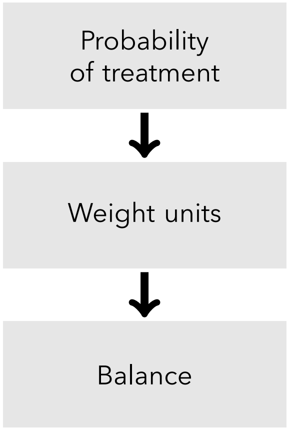
$$\mathbb{E}[\text{covariates} \mid \text{treated}] = \mathbb{E}\Big[\underbrace{\frac{e(x,g)}{1-e(x,g)}}_{\text{weights}}\; \text{covariates} \mid \text{control}\Big]$$

$\leadsto$ How to estimate weights?

Background: Traditional IPW workflow

Goal: $\hat{e}(x, g)$ close to $e(x, g)$

1. **Directly** estimate $\hat{e}(x, g)$, via MLE, ML, etc.
2. Calculate weights $\hat{\gamma} = \frac{\hat{e}(x,g)}{1-\hat{e}(x,g)}$
3. **Indirectly** balance covariates

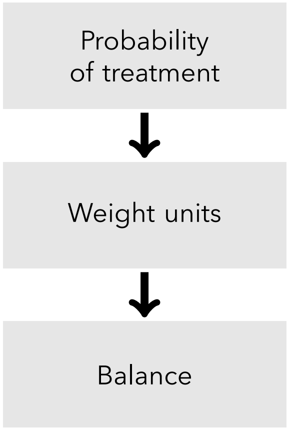


Background: Traditional IPW workflow

Goal: $\hat{e}(x, g)$ close to $e(x, g)$

1. **Directly** estimate $\hat{e}(x, g)$, via MLE, ML, etc.
2. Calculate weights $\hat{\gamma} = \frac{\hat{e}(x,g)}{1-\hat{e}(x,g)}$
3. **Indirectly** balance covariates

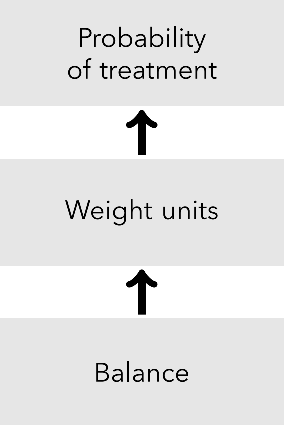
- Poor finite sample performance, esp with many covariates



Background: Balancing weights workflow

Goal: $\hat{\gamma}$ close to $\frac{e(x,g)}{1-e(x,g)}$

1. **Directly** estimate $\hat{\gamma}$ to balance covariates
2. **Indirectly** estimate $\hat{e}(x,g) = \frac{\hat{\gamma}}{1+\hat{\gamma}}$



Background: Balancing weights workflow

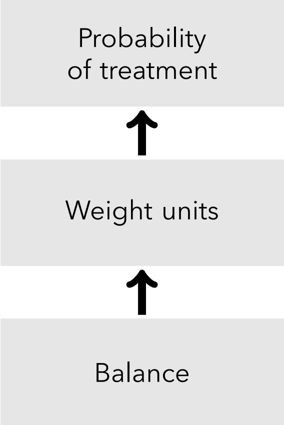
Goal: $\hat{\gamma}$ close to $\frac{e(x,g)}{1-e(x,g)}$

1. **Directly** estimate $\hat{\gamma}$ to balance covariates

2. **Indirectly** estimate $\hat{e}(x,g) = \frac{\hat{\gamma}}{1+\hat{\gamma}}$

- Old history as raking and calibration in survey sampling with non-response
  [Deming and Stephan, 1940; Deville et al., 1993]

- New causal inference literature
  [Hainmueller, 2011; Zubizarreta, 2015; Athey et al., 2018; Chernozhukov et al., 2018]



Balancing weights to estimate subgroup effects

Balancing weights for local balance only

Error for effect in subgroup g

For outcome model $m_0 = \eta_g \cdot \phi(x)$; weighting estimator $\hat{\mu}_{0g} = \sum \gamma Y_i$

$$\text{error}_g \quad \leq \quad \|\eta_g\|_2 \left\|\text{Local Balance}_g\right\|_2 \quad + \quad \|\gamma\|_2$$

Can generalize to flexible outcome models [Hirshberg et al., 2019; Hazlett, 2020]

Balancing local balance only

Balancing weights for subgroup g

$$\min_{\gamma} \quad \|\text{Local Balance}_g\|_2^2 \quad + \quad \frac{\lambda_g}{2}\|\gamma\|_2^2$$

$$\text{s.t.} \quad \sum \gamma_i = 1, \quad \gamma_i \geq 0$$

Balancing local balance only

Balancing weights for subgroup g

$$\min_{\gamma} \quad \|\text{Local Balance}_g\|_2^2 \quad + \quad \frac{\lambda_g}{2}\|\gamma\|_2^2$$

$$\text{s.t.} \quad \sum \gamma_i = 1, \quad \gamma_i \geq 0$$

Challenge:

- Small subgroups can be hard to balance well
- Balancing subgroups separately $\rightarrow$ poor global balance

Balancing global balance and local balance

Partially Pooled Balancing Weights

$$\min_{\gamma} \quad \sum_{g} \|\text{Local Balance}_g\|_2^2 \quad + \quad \frac{\lambda_g}{2}\|\gamma\|_2^2$$

$$\text{s.t.} \quad \sum_{G_i = g} \gamma_i = n_{1g}, \qquad \gamma_i \geq 0$$

$$\text{Global Balance} = 0$$

Balancing global balance and local balance

Partially Pooled Balancing Weights

$$\min_{\gamma} \quad \sum_g \|\text{Local Balance}_g\|_2^2 \quad + \quad \frac{\lambda_g}{2}\|\gamma\|_2^2$$

$$\text{s.t.} \quad \sum_{G_i=g} \gamma_i = n_{1g}, \qquad \gamma_i \geq 0$$

$$\text{Global Balance} = 0$$

- Overall errors depends on both global balance and local balance
- Further expand to control *differences* in local balance

⇝ *Tuning parameter:* Global parameter $\lambda \quad \Rightarrow \quad \lambda_g = \lambda / n_g$

Dual perspective: M estimation of treatment odds

Dual when optimizing for for local balance only

Population:
$$\underbrace{\frac{e(x,g)}{1-e(x,g)}}_{\substack{\text{inverse prop.}\\ \text{score weights}}} \sim \underbrace{\alpha_g + \beta_g'\phi(x)}_{\text{balancing weights}}$$

Dual perspective: M estimation of treatment odds

Dual when optimizing for for local balance only

Population: $$\underbrace{\frac{e(x,g)}{1-e(x,g)}}_{\substack{\text{inverse prop.}\\ \text{score weights}}} \sim \underbrace{\alpha_g + \beta_g' \phi(x)}_{\text{balancing weights}}$$

Sample: $$\min_{\alpha_g, \beta_g} \text{ regression loss } + \underbrace{\frac{\lambda}{2} \|\beta_g\|_2^2}_{\text{ridge penalty}}$$

Global balance constraint $\longleftrightarrow$ partial pooling in the dual problem

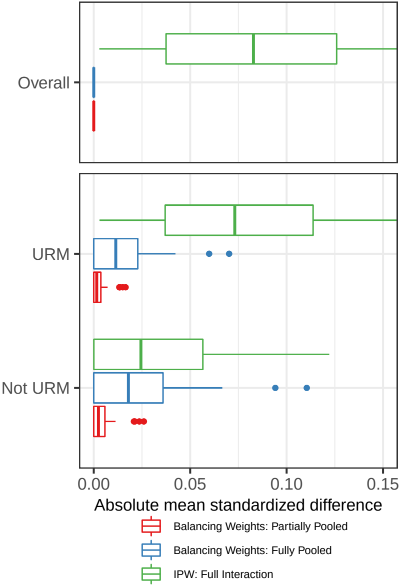
Dual for Partially Pooled Balancing Weights

$$\min_{\alpha_g, \beta_g, \mu_\beta} \quad \text{regression loss} \; + \; \underbrace{\frac{\lambda_g}{2}\|\beta_g - \mu_\beta\|_2^2}_{\text{local} \to \text{global}}$$

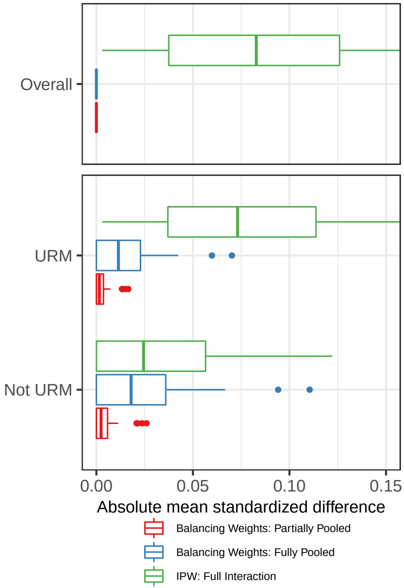
Partially pool local $\to$ global model: regularization *directly* related to imbalance

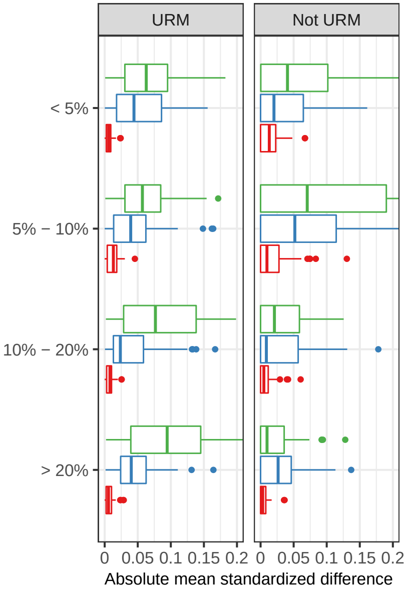
Differential impacts of letters of recommendation

Partially pooled balancing weights → improved balance



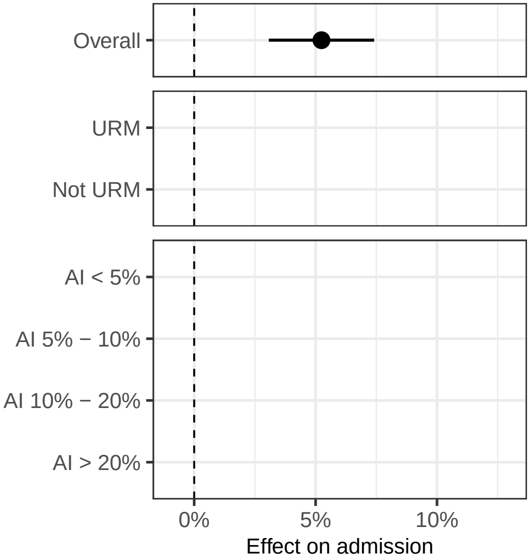
Partially pooled balancing weights → improved balance



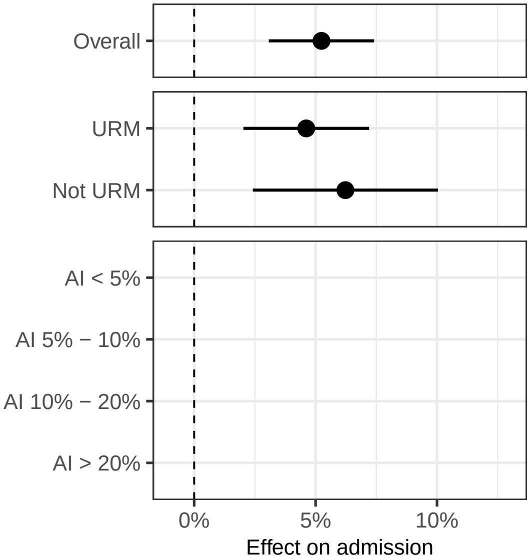


Large overall effect

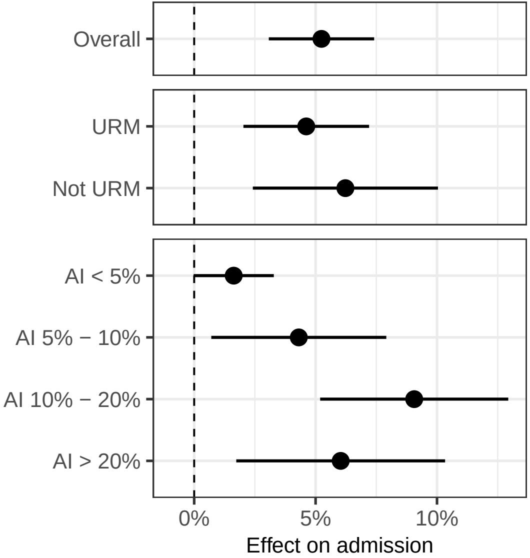
Baseline admissions rate ∼20%



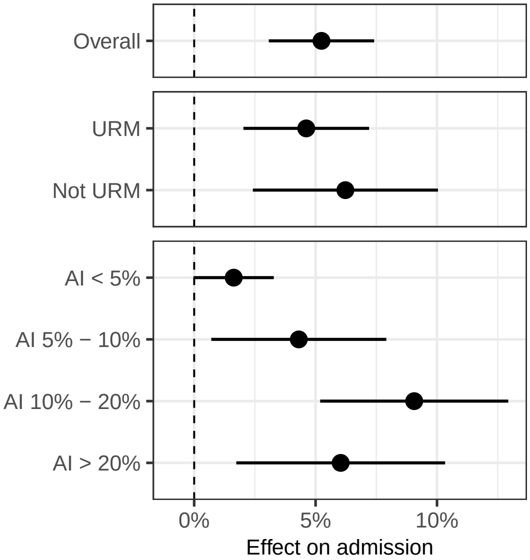
No differences by URM status

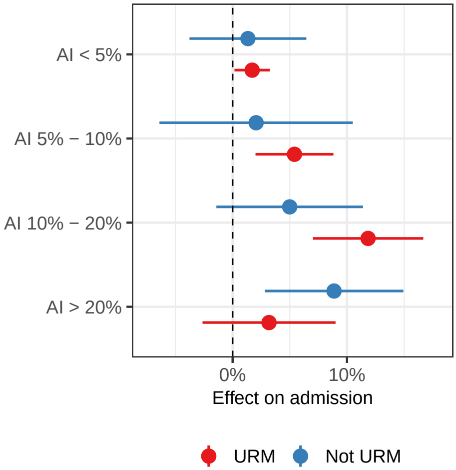


Large differences by Admissibility Index



Relative effect sizes flip





Recap: Varying Impacts of Letters of Recommendation

Partially pooled balancing weights

- Find weights that control both Local Balance and Global Balance
- Dual relation to partially pooled IPW
- R package `balancer`

→ No evidence of different impacts by URM status

Recap: Varying Impacts of Letters of Recommendation

Partially pooled balancing weights

- Find weights that control both Local Balance and Global Balance
- Dual relation to partially pooled IPW
- R package `balancer`

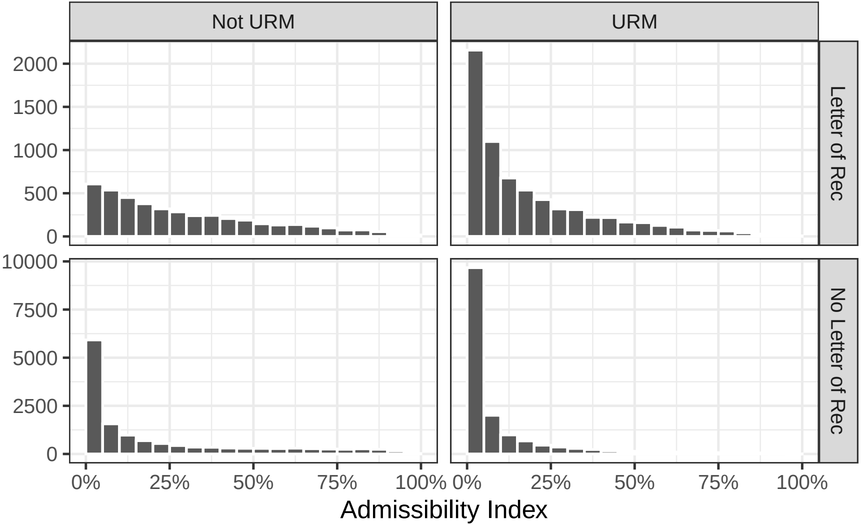
→ No evidence of different impacts by URM status

Thank you!

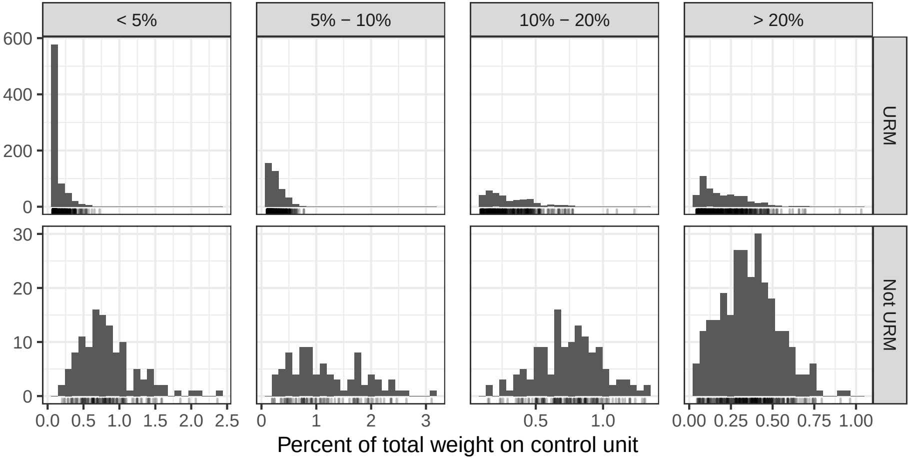
ebenmichael.github.io

Appendix

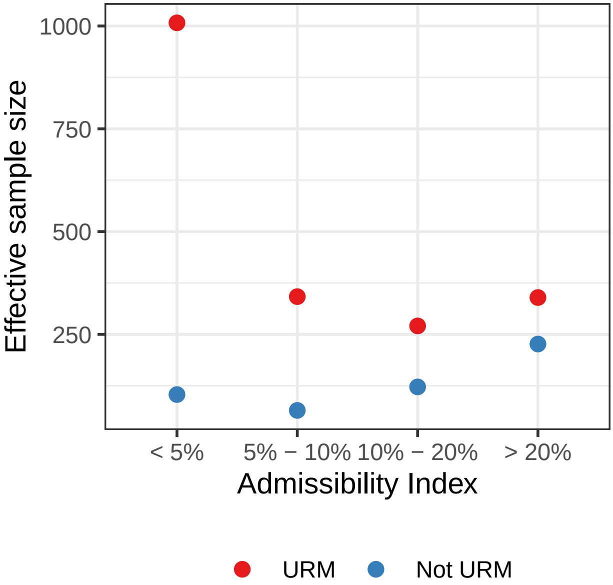
Heterogeneity across admissibility index



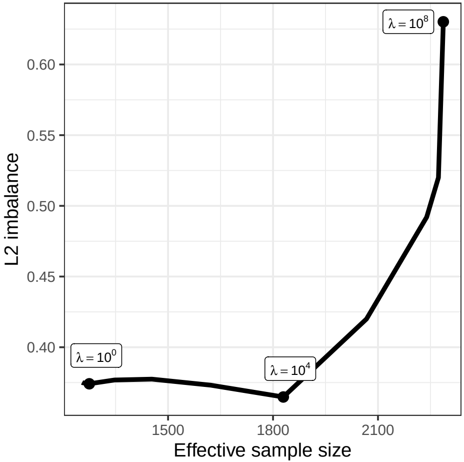
Distribution of the estimated weights



Effective sample sizes

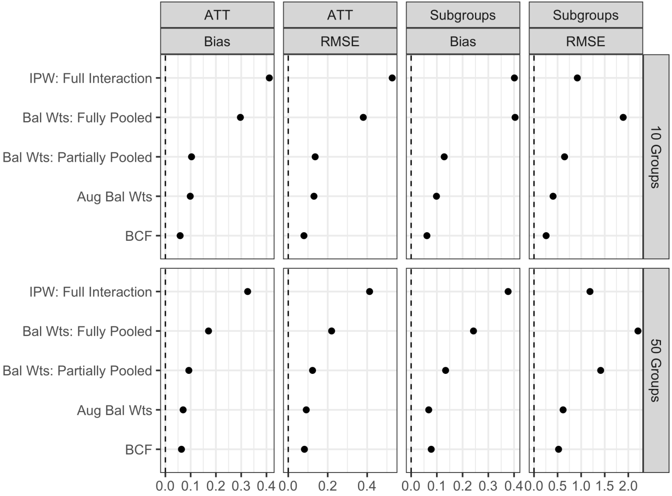


Hyperparameter tuning



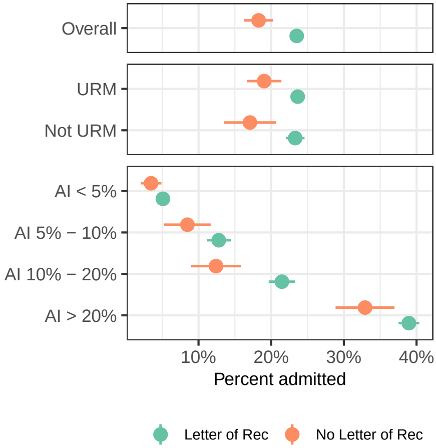
- Evaluate across a range of λ
- Gains in precision, comparable imbalance

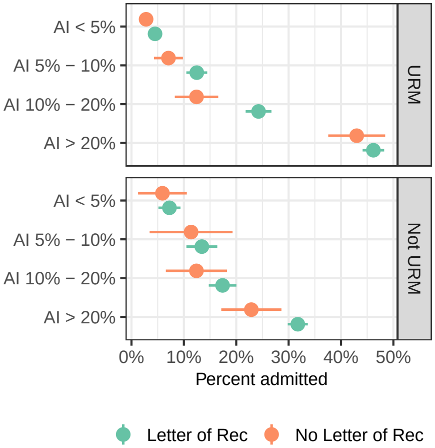
Simulation Study



- Major gains relative to traditional IPW
- Comparable performance to ML methods; retain design-based advantages

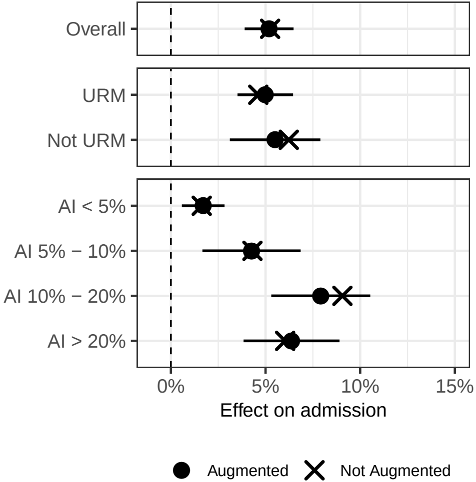
Estimated group means

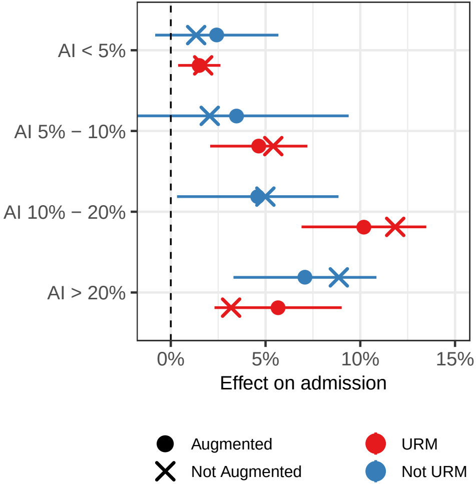




Augmentation diminishes differences

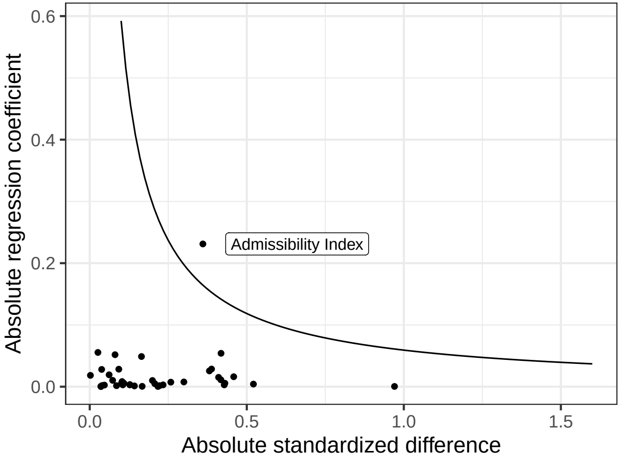
[Random forest outcome model]





Sensitivity to unmeasured confounding

- Adapt Soriano et al. [2021]
- Overall LOR effect still positive with $\Lambda = 1.1$
- Consistent with wide range of subgroup estimates



References I

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Technical report.

Chernozhukov, V., Newey, W., Robins, J., and Singh, R. (2018). Double/de-biased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*.

Deming, W. E. and Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, 11(4):427–444.

Deville, J. C., Särndal, C. E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423):1013–1020.

Hainmueller, J. (2011). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20:25–46.

Hazlett, C. (2020). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statistica Sinica*.

Hirshberg, D. A., Maleki, A., and Zubizarreta, J. (2019). Minimax Linear Estimation of the Retargeted Mean.

Zubizarreta, J. R. (2015). Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. *Journal of the American Statistical Association*, 110(511):910–922.