Variation in impacts of letters of recommendation on college admissions decisions

Approximate balancing weights for treatment effect heterogeneity in observational studies

Eli Ben-Michael, Avi Feller, and Jesse Rothstein

UC Berkeley

Polmeth XXXVII

July 2020

Disentangling differential impacts in an observational study

Estimating treatment effects for pre-defined subgroups

- Randomized trial $\rightarrow$ stratified sampling

Disentangling differential impacts in an observational study

Estimating treatment effects for pre-defined subgroups

- Randomized trial $\rightarrow$ stratified sampling
- Observational study $\rightarrow$ different impacts or different covariate balance?

Disentangling differential impacts in an observational study

Estimating treatment effects for pre-defined subgroups

- Randomized trial $\rightarrow$ stratified sampling
- Observational study $\rightarrow$ different impacts or different covariate balance?

Overall effect $\longleftrightarrow$ Global Balance across the study

Disentangling differential impacts in an observational study

Estimating treatment effects for pre-defined subgroups

- Randomized trial $\rightarrow$ stratified sampling
- Observational study $\rightarrow$ different impacts or different covariate balance?

Overall effect	$\longleftrightarrow$	Global Balance across the study
Subgroup effects	$\longleftrightarrow$	Local Balance within subgroup

Controlling for Local Balance to estimate subgroup effects

This paper: Balancing weights for subgroup analysis in observational studies

Find weights that control Local Balance

- Control Global Balance for stability and estimate overall effect

Dual relation to partially pooled propensity score estimation

Augment with outcome model

Do letters of reccommendation affect URM applicants differently?

UC Berkeley pilot study on LORs in 2016

[Rothstein, 2017]

Do letters of reccommendation affect URM applicants differently?

UC Berkeley pilot study on LORs in 2016

[Rothstein, 2017]

Worsen disparity for URM applicants?

Do letters of reccommendation affect URM applicants differently?

UC Berkeley pilot study on LORs in 2016

[Rothstein, 2017]

Take advantage of aspects of design

- Two reader evaluation system
- First reader score $\rightarrow$ invitation for LOR
- Second reader has access to LOR

Worsen disparity for URM applicants?

Do letters of reccommendation affect URM applicants differently?

UC Berkeley pilot study on LORs in 2016

[Rothstein, 2017]

Take advantage of aspects of design

- Two reader evaluation system
- First reader score $\rightarrow$ invitation for LOR
- Second reader has access to LOR

Worsen disparity for URM applicants?

Large imbalance on income/test scores

Accounting for compositional differences across applicants

Are differences due to composition?

Accounting for compositional differences across applicants

Are differences due to composition?

Admissibility Index

- Predicting admission from 2015 data

Accounting for compositional differences across applicants

Are differences due to composition?

Admissibility Index

- Predicting admission from 2015 data

Interaction between URM status and AI

- Second and third order interactions!
- Large sample sizes $\rightarrow$ power

Accounting for compositional differences across applicants

Are differences due to composition?

Admissibility Index

- Predicting admission from 2015 data

Interaction between URM status and AI

- Second and third order interactions!
- Large sample sizes $\rightarrow$ power

AI Range	URM	Number
$<$5%	URM	11,832
	Not URM	6,529
5% - 10%	URM	3,106
	Not URM	2,099
10% - 20%	URM	2,876
	Not URM	2,495
$>$20%	URM	4,645
	Not URM	6,959

Balancing weights to estimate sugroup effects

What we see

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

- Covariates $X_i \in \mathcal{X}$
- Treatment status $W_i \in \{0, 1\}$ and outcome $Y_i \in \mathbb{R}$
- Group indicator $G_i \in \{1, \ldots, K\}$
  - Interaction between URM, AI, reader 1 score, and college

What we see

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

- Covariates $X_i \in \mathcal{X}$
- Treatment status $W_i \in \{0, 1\}$ and outcome $Y_i \in \mathbb{R}$
- Group indicator $G_i \in \{1, \ldots, K\}$
  - Interaction between URM, AI, reader 1 score, and college

Goal: Estimate the ATT

$$\tau = \mathbb{E}[Y(1) - Y(0) \mid W = 1]$$

What we see

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

- Covariates $X_i \in \mathcal{X}$
- Treatment status $W_i \in \{0, 1\}$ and outcome $Y_i \in \mathbb{R}$
- Group indicator $G_i \in \{1, \ldots, K\}$
  - Interaction between URM, AI, reader 1 score, and college

Goal: Estimate the ATT **and subgroup CATT**

$$\tau = \mathbb{E}[Y(1) - Y(0) \mid W = 1]$$

$$\tau_g = \mathbb{E}[Y(1) - Y(0) \mid W = 1, G = g]$$

What we assume

Key Identifying assumption: Strong ignorability

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

What we assume

Key Identifying assumption: Strong ignorability

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

- Flexible ML outcome models
  [Künzel et al., 2019; Nie and Wager, 2019; Hahn et al., 2020]
- Design-based IPW estimators: logistic regression with interactions
  [Li et al., 2013; Lee et al., 2019; Dong et al., 2020; Yang et al., 2020]

The importance of Local Balance for weighting

Estimation error depends on local imbalance in prognostic score $m_0(X_i, g)$

To simplify, assume linearity $m_0(x, g) = \eta_g \cdot \phi(x)$:

$$\text{Error}_g \approx \|\eta_g\|_2 \, \|\text{Local Balance}_g\|_2 + \sigma \|\gamma_g\|_2$$

Can generalize to non-parametric function classes, infinite dimensional bases

[Hirshberg et al., 2019; Hazlett, 2020]

Optimizing for both Global Balance and Local Balance

$$\min_{\gamma} \quad \sum_{g=1}^{K} \|\text{Local Balance}_g\|_2^2$$

Optimizing for both Global Balance and Local Balance

$$\min_{\gamma} \quad \sum_{g=1}^{K} \|\text{Local Balance}_g\|_2^2 + \frac{\lambda_g}{2}\|\gamma_g\|_2^2$$

Optimizing for both Global Balance and Local Balance

$$\begin{aligned} \min_{\gamma} \quad & \sum_{g=1}^{K} \|\text{Local Balance}_g\|_2^2 + \frac{\lambda_g}{2}\|\gamma_g\|_2^2 \\ \text{s.t.} \quad & \sum_{G_i=g} \gamma_i = n_{1g} \quad \gamma_i \geq 0 \end{aligned}$$

Optimizing for both Global Balance and Local Balance

$$\min_{\gamma} \quad \sum_{g=1}^{K} \|\text{Local Balance}_g\|_2^2 + \frac{\lambda_g}{2}\|\gamma_g\|_2^2$$

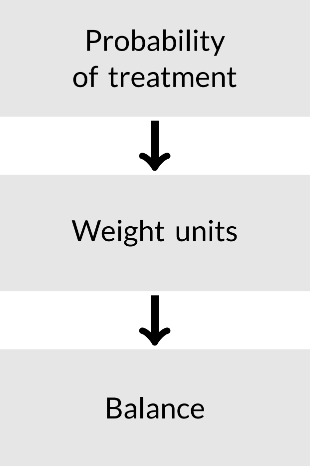
$$\text{s.t.} \quad \sum_{G_i=g} \gamma_i = n_{1g} \quad \gamma_i \geq 0$$

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

Two views of the P-score imply different estimation methods

Conditional probability of treatment

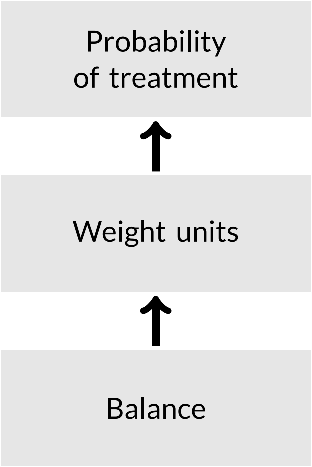
- Estimate $\hat{e}(x, g)$ with MLE, then estimate weights $\hat{\gamma}_i = \frac{\hat{e}(X_i, G_i)}{1-\hat{e}(X_i, G_i)}$
- Indirectly balances covariates
- Poor finite sample performance, especially with many covariates
  [Athey et al., 2018]



Two views of the P-score imply different estimation methods

Balancing score [Rosenbaum and Rubin, 1983]

- Find weights $\hat{\gamma}$ that balance covariates
  [Hainmueller, 2011; Zubizarreta, 2015]
  [Hirshberg et al., 2019; Hazlett, 2020]
  [Imai and Ratkovic, 2013]
- Old history as raking and calibration in survey sampling with non-response
  [Deming and Stephan, 1940; Deville et al., 1993]
- Indirectly estimates the P-score
  [Zhao and Percival, 2016]
  [Wang and Zubizarreta, 2019]
  [Chattopadhyay et al., 2020]



Dual problem: multilevel P-score estimation

Fully interacted (truncated) linear model for **treatment odds**

$$\frac{e(X_i, G_i)}{1 - e(X_i, G_i)} \sim \left[\alpha_g + \beta_g{}' \phi(X_i)\right]_+$$

Dual problem: multilevel P-score estimation

Fully interacted (truncated) linear model for **treatment odds**

$$\frac{e(X_i, G_i)}{1 - e(X_i, G_i)} \sim \left[\alpha_g + \beta_g{}'\phi(X_i)\right]_+$$

Global Balance constraint $\rightarrow$ partial pooling towards global model

$$\frac{\lambda_g}{2} \left\|\beta_g - \mu_\beta\right\|_2^2$$

Dual problem: multilevel P-score estimation

Fully interacted (truncated) linear model for **treatment odds**

$$\frac{e(X_i, G_i)}{1 - e(X_i, G_i)} \sim \left[\alpha_g + \beta_g{}'\phi(X_i)\right]_+$$

Global Balance constraint $\rightarrow$ partial pooling towards global model

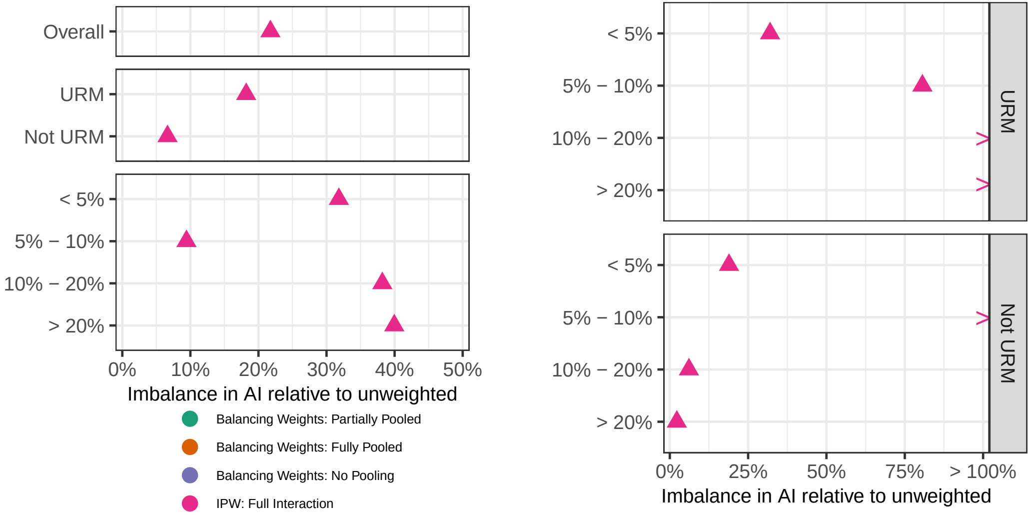
$$\frac{\lambda_g}{2} \left\|\beta_g - \mu_\beta\right\|_2^2$$

Primal weights are estimated treatment odds

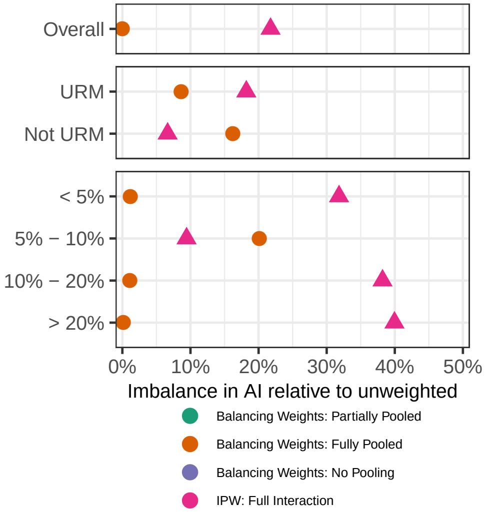
$$\hat{\gamma}_i = \left[\hat{\alpha}_j + \hat{\beta}_j{}'\phi(X_i)\right]_+$$

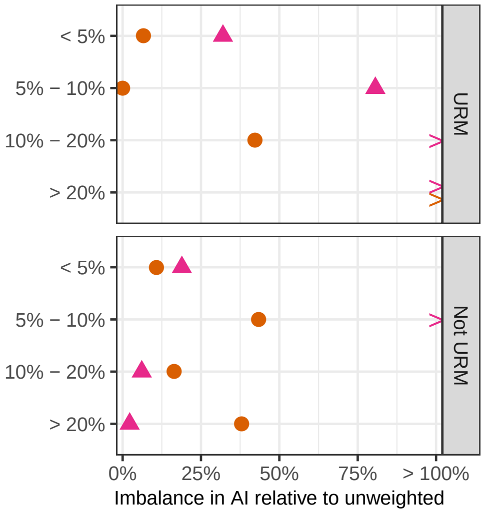
Differential impacts of letters of recommendation

Acheiving good balance by design

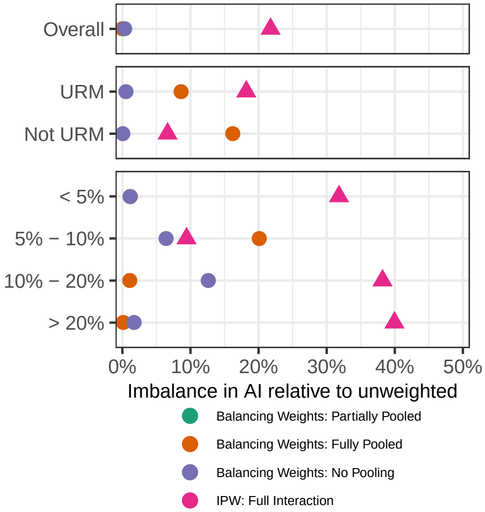


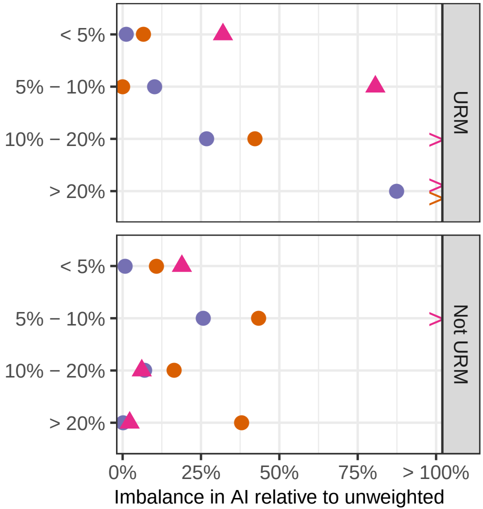
Acheiving good balance by design



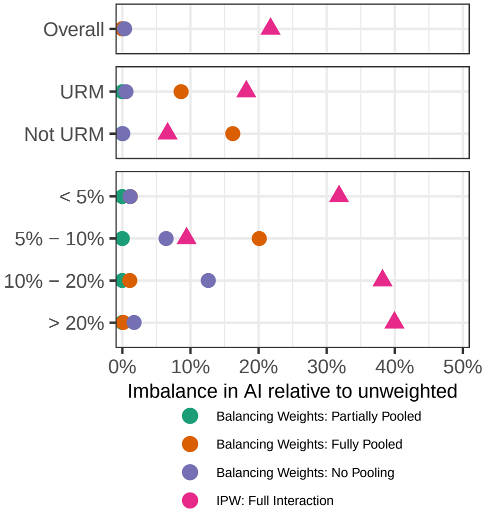


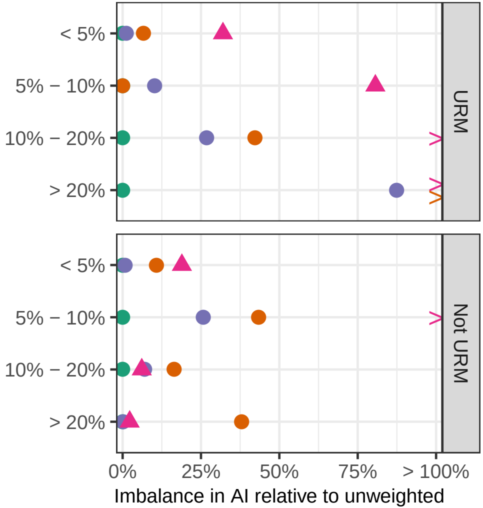
Acheiving good balance by design



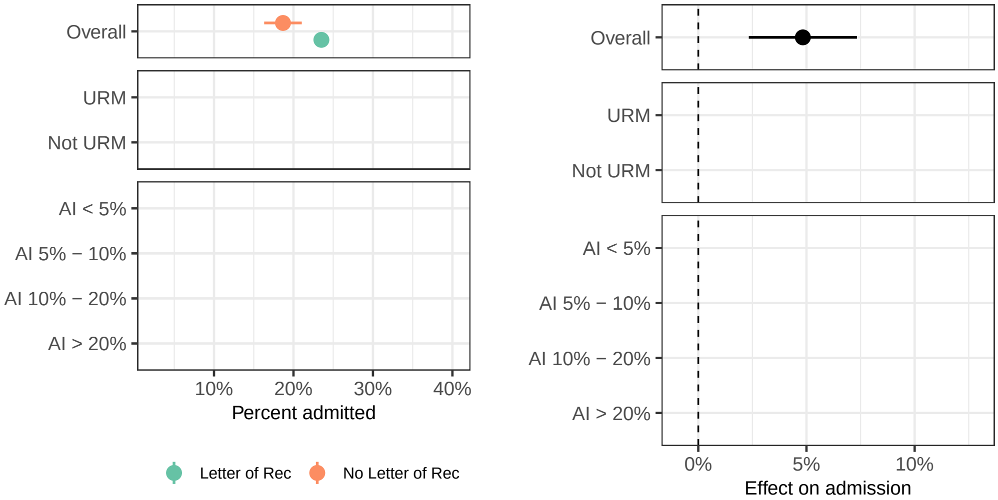


Acheiving good balance by design

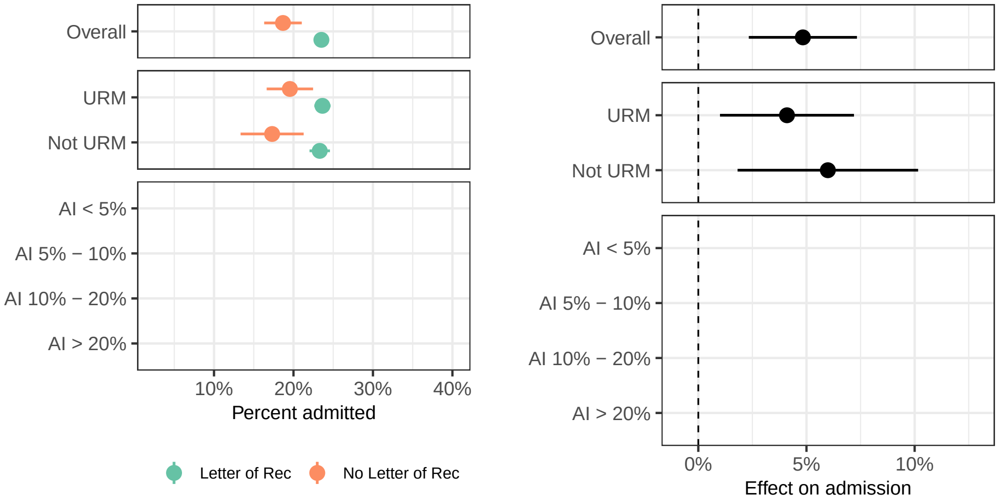




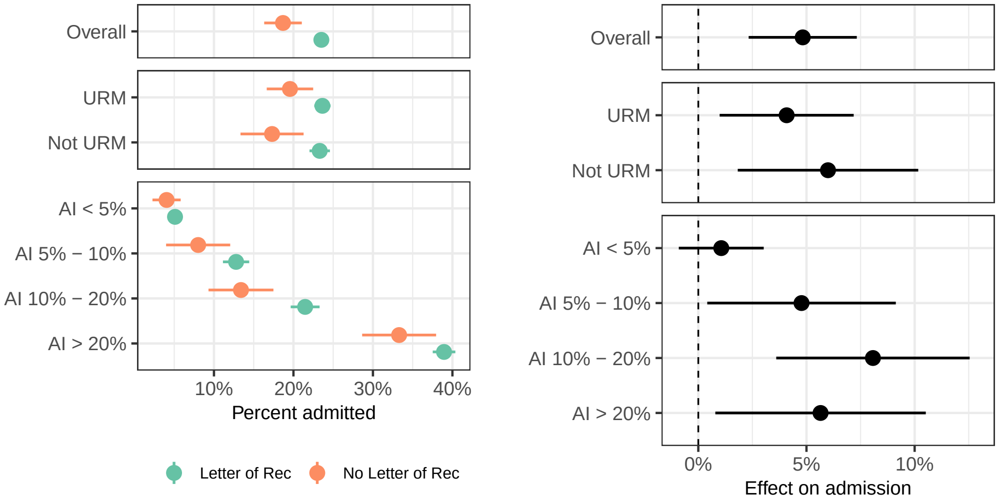
No discernable differences for URMs, some by AI



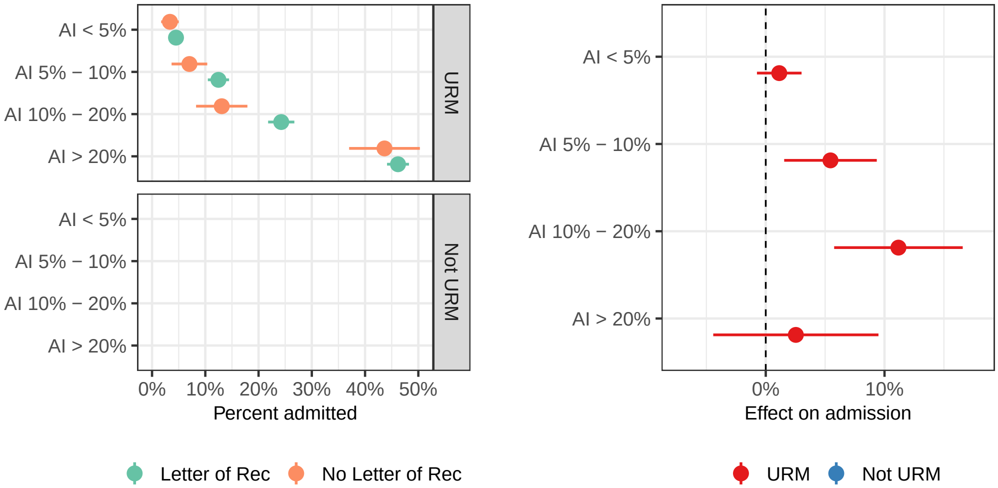
No discernable differences for URMs, some by AI



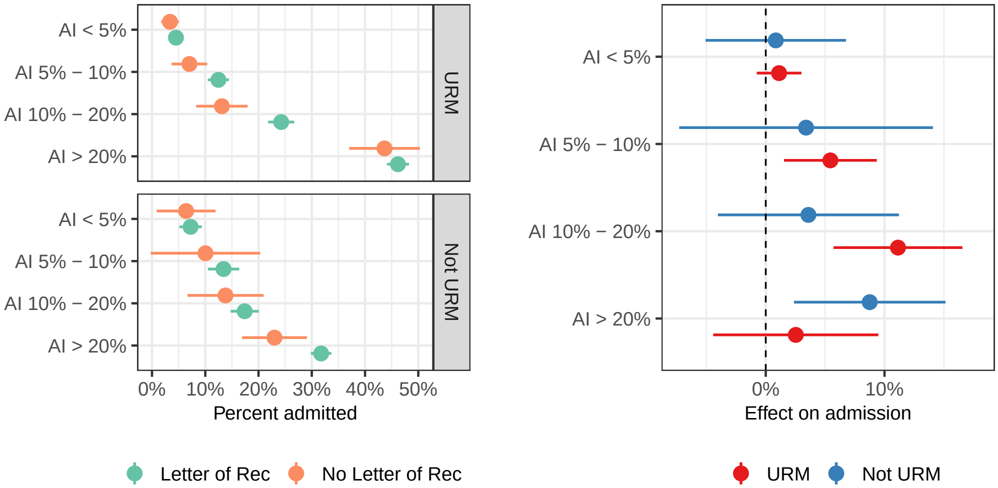
No discernable differences for URMs, some by AI



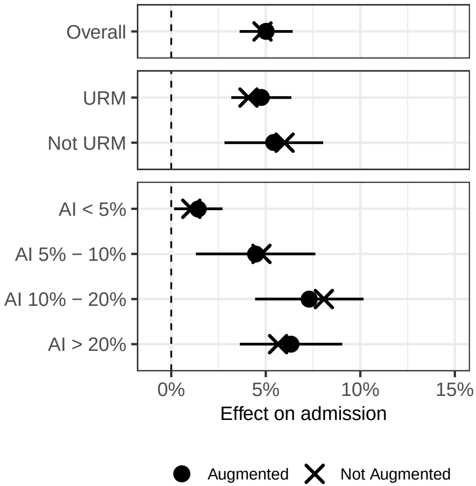
Peak effect for middle tier URM applicants



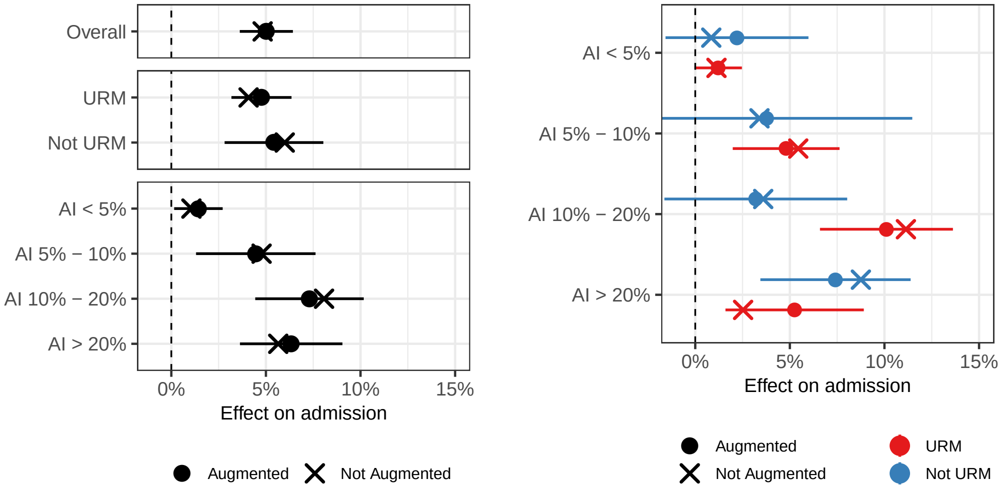
Peak effect for middle tier URM applicants



Augmentation diminishes differences



Augmentation diminishes differences



Recap

Disentangle differential impacts from differential confounding

- Find weights that control Local Balance
- Control Global Balance to estimate the overall effect
- Dual relation to partially pooled IPW
- Augment with outcome model

Recap

Disentangle differential impacts from differential confounding

- Find weights that control Local Balance
- Control Global Balance to estimate the overall effect
- Dual relation to partially pooled IPW
- Augment with outcome model

Next Steps:

- Sensitivity analysis
- Heterogeneity in other observational study settings
- `R` package coming soon!

Recap

Disentangle differential impacts from differential confounding

- Find weights that control Local Balance
- Control Global Balance to estimate the overall effect
- Dual relation to partially pooled IPW
- Augment with outcome model

Next Steps:

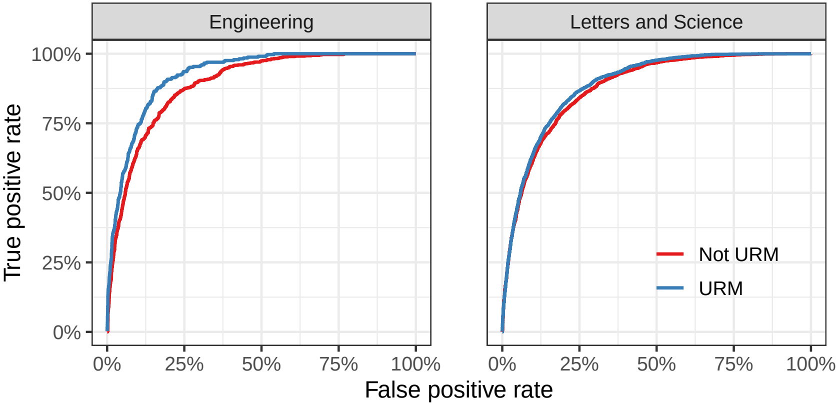
- Sensitivity analysis
- Heterogeneity in other observational study settings
- R package coming soon!

Thank you!

ebenmichael.github.io

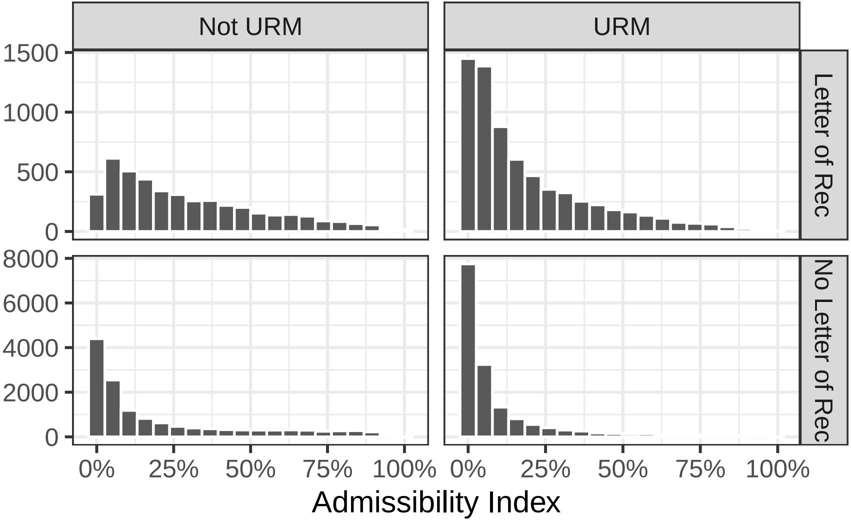
Appendix

Admissibility Index: a strong and simple prognostic

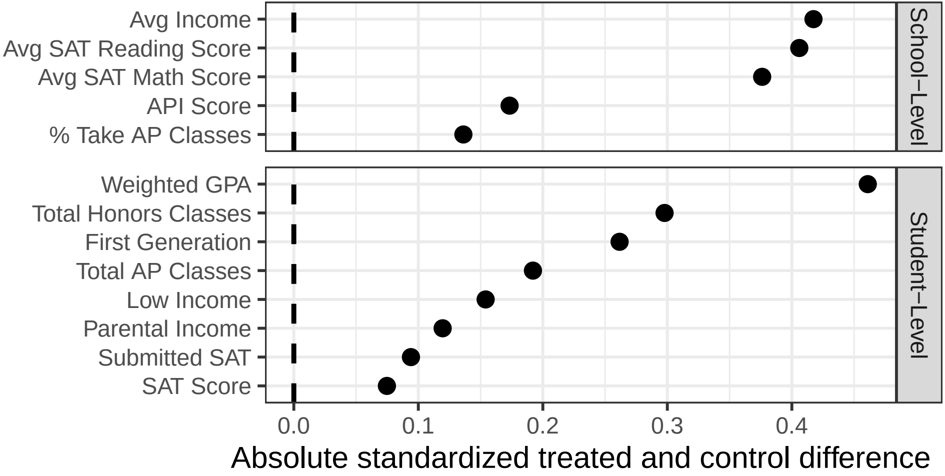


Use logisitic regression on 2015 applicant pool to predict admission for 2016 pool

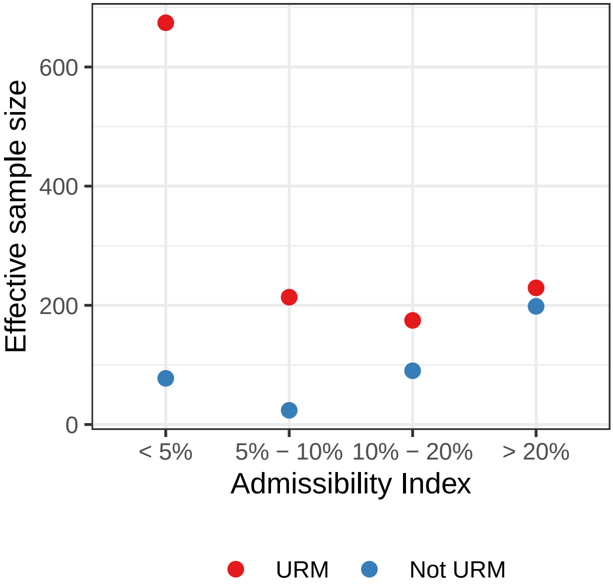
Heterogeneity across admissibility index



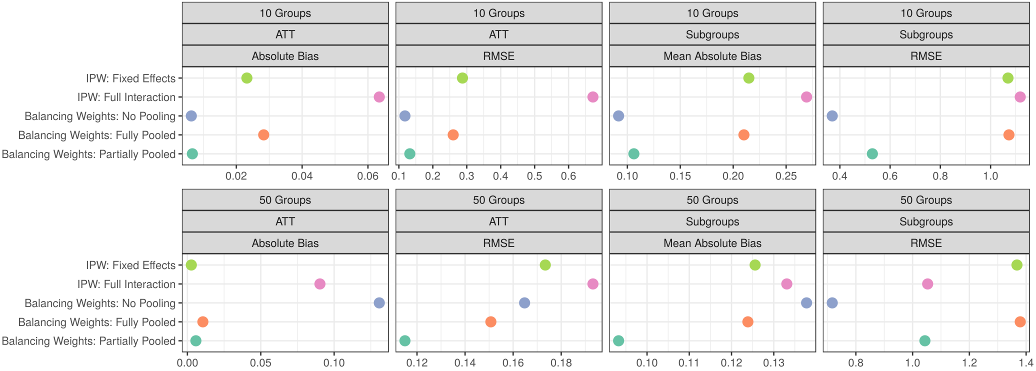
Large imbalance on income, grades, and test scores



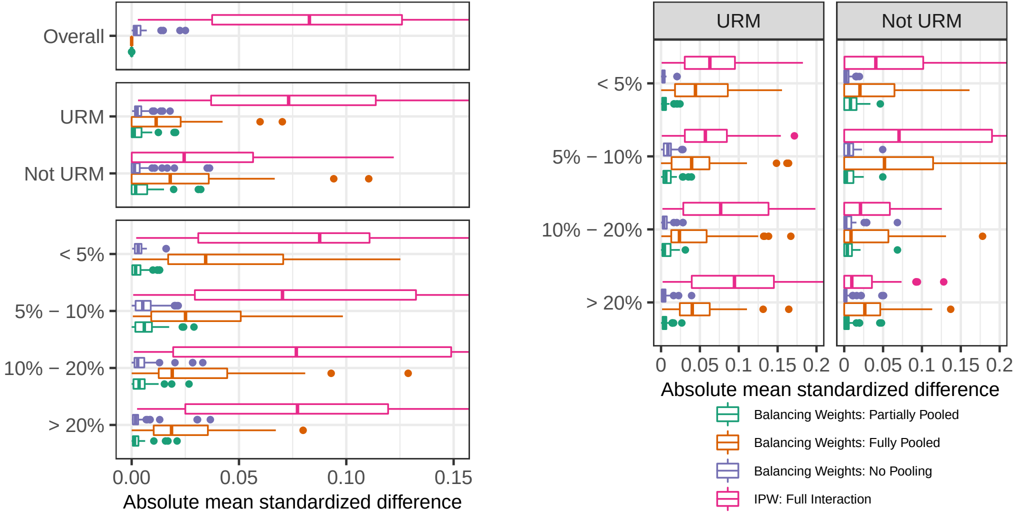
Effective sample sizes



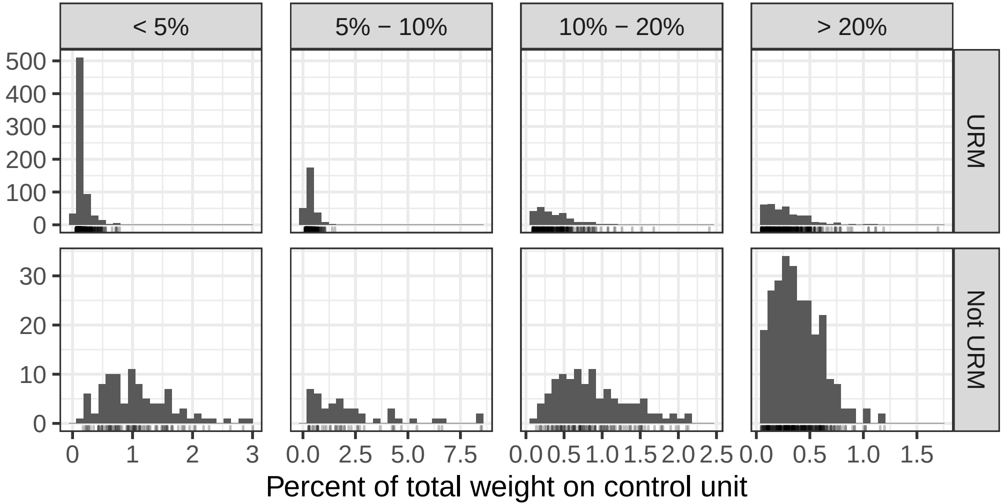
Simulation study



Global and local balance



Good sample overlap for URMs, less so for non-URMs



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.

Chattopadhyay, A., Christopher H. Hase, and Zubizarreta, J. R. (2020). Balancing Versus Modeling Approaches to Weighting in Practice. *Statistics in Medicine*, in press.

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.

Dong, J., Zhang, J. L., Zeng, S., and Li, F. (2020). Subgroup balancing propensity score. *Statistical Methods in Medical Research*, 29(3):659–676.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects. *Bayesian Analysis*, pages 1–33.

References II

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.

Imai, K. and Ratkovic, M. (2013). Covariate balancing propensity score. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(1):243–263.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4156–4165.

Lee, Y., Nguyen, T. Q., and Stuart, E. A. (2019). Partially Pooled Propensity Score Models for Average Treatment Effect Estimation with Multilevel Data.

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19):3373–3387.

References III

Nie, X. and Wager, S. (2019). Quasi-Oracle Estimation of Heterogeneous Treatment Effects.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Rothstein, J. (2017). The impact of letters of recommendation on UC Berkeley admissions in the 2016-17 cycle. Technical report, California Policy Lab.

Wang, Y. and Zubizarreta, J. R. (2019). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*.

Yang, S., Lorenzi, E., Papadogeorgou, G., Wojdyla, D., Li, F., and Thomas, L. (2020). Subgroup covariates balancing via the overlap weights.

Zhao, Q. and Percival, D. (2016). Entropy Balancing is Doubly Robust. *Journal of Causal Inference*.

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.