Multilevel Calibration Weighting for Survey Data

Eli Ben-Michael, Avi Feller, and Erin Hartman

Harvard & UC Berkeley

Polmeth XXXVIII

July 2021

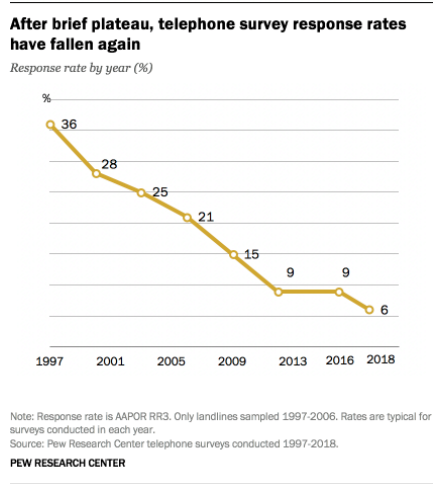
Survey non-response is a pernicious problem

- ↑ reliance on non-probability samples
- ↑ need for non-response adjustment

New and richer sources of data can help

Case study: 2016 presidential election

- Failing to adjust for **interactions** led to substantial bias
  [Kennedy et al., 2018]
- Retrospective study
  - Pre-election Pew poll
  - Post-election CCES as "ground truth"



Even in moderate dimensions, survey adjustment is hard

Even in moderate dimensions, survey adjustment is hard

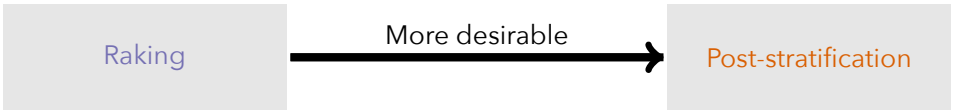
Raking

Even in moderate dimensions, survey adjustment is hard

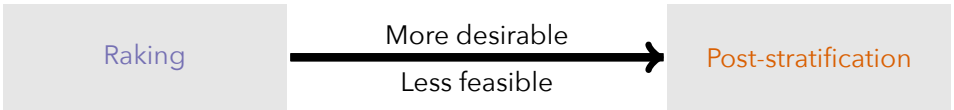
Raking

Post-stratification

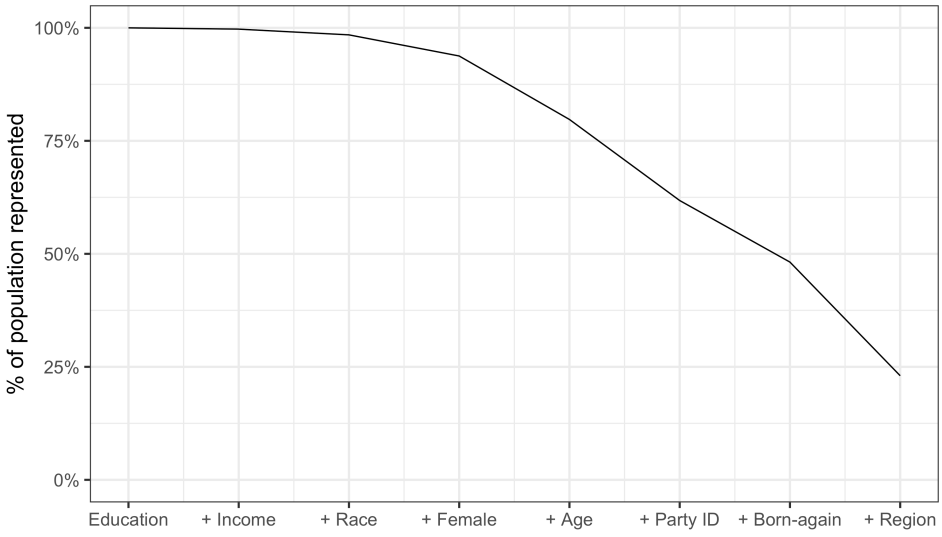
Even in moderate dimensions, survey adjustment is hard



Even in moderate dimensions, survey adjustment is hard



Quickly run into empty cells



How can we account for interactions in a principled way?

Adjusting for interactions in non-response is important

Ideally we'd post-stratify, but we can't

How can we account for interactions in a principled way?

Adjusting for interactions in non-response is important

Ideally we'd post-stratify, but we can't

This paper: Approximately post-stratify while at least raking on margins

- Leverage the value of interactions in a parsimonious way
- Dual representation as multilevel model of non-response

Combine with outcome model → Double Regression with Post-stratification (DRP)

- *Explicitly* adjusting for interactions via weighting and *implicitly* via machine learning

Approximately post-stratifying
while at least raking

Notation and setup

$i = 1, \ldots, N$ individuals

- Outcome Y_i, Response R_i with prob $P(R_i = 1) = \pi_i$
- d categorical covariates w/levels $J_1, \ldots, J_d$

Notation and setup

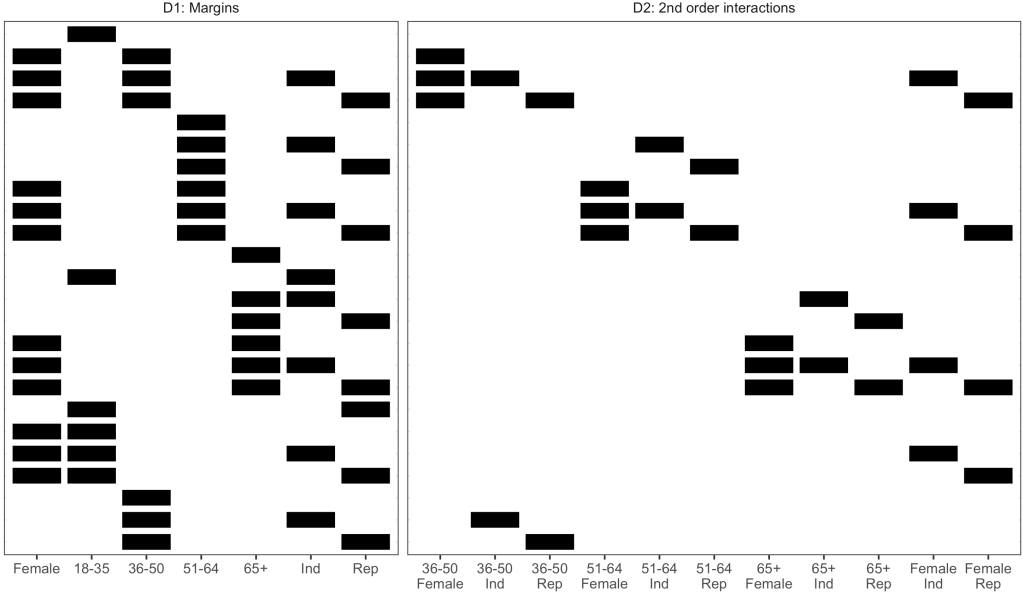
$i = 1, \ldots, N$ individuals

- Outcome Y_i, Response R_i with prob $P(R_i = 1) = \pi_i$
- d categorical covariates w/levels $J_1, \ldots, J_d$

Combine into cells $S_i \in \{1, \ldots, J_1 \times \ldots \times J_d \equiv J\}$

- Overall count vector $N^{\mathcal{P}} \in \mathbb{N}^J$ and response count vector $n^{\mathcal{R}} \in \mathbb{N}^J$
- Probability of responding conditional on cell s: $\pi(s) = \frac{1}{N_s^{\mathcal{R}}} \sum_{S_i = s} \pi_i$

Binary vector of k^{th} order interaction terms for cell s: $D_s^{(k)}$


D1: Margins
D2: 2nd order interactions
Female
18-35
36-50
51-64
65+
Ind
Rep
36-50 Female
36-50 Ind
36-50 Rep
51-64 Female
51-64 Ind
51-64 Rep
65+ Female
65+ Ind
65+ Rep
Female Ind
Female Rep

Goal

Impute the average

$$\mu = \frac{1}{N}\sum_{i=1}^{n} Y_i = \frac{1}{N}\sum_{s} N_s^{\mathcal{P}}\mu_s$$

Goal

Impute the average

$$\mu = \frac{1}{N}\sum_{i=1}^{n} Y_i = \frac{1}{N}\sum_{s} N_s^{\mathcal{P}}\mu_s$$

From the respondents

$$\hat{\mu}(\hat{\gamma}) = \frac{1}{N}\sum_{i=1}^{N} R_i\hat{\gamma}_i Y_i = \frac{1}{N}\sum_{s} n_s^{\mathcal{R}}\hat{\gamma}(s)\bar{Y}_s$$

Goal

Impute the average

$$\mu = \frac{1}{N}\sum_{i=1}^{n} Y_i = \frac{1}{N}\sum_{s} N_s^{\mathcal{P}}\mu_s$$

From the respondents

$$\hat{\mu}(\hat{\gamma}) = \frac{1}{N}\sum_{i=1}^{N} R_i\hat{\gamma}_i Y_i = \frac{1}{N}\sum_{s} n_s^{\mathcal{R}}\hat{\gamma}(s)\bar{Y}_s$$

Assume responses are Missing At Random (MAR) so that within each cell

$$\mathbb{E}\left[\bar{Y}_s\right] = \mu_s$$

And positivity

$$\pi(s) > 0$$

Choosing weights: Raking and Post-stratification

Raking on margins

Exactly match the counts for margins:

$$\sum_s D_s^{(1)} n_s^{\mathcal{R}} \hat{\gamma}(s) = \sum_s D_s^{(1)} N_s^{\mathcal{P}}$$

- Can usually compute if d is moderate
- "Only" accounts for the linear variables

Choosing weights: Raking and Post-stratification

Raking on margins

Exactly match the counts for margins:

$$\sum_s D_s^{(1)} n_s^{\mathcal{R}} \hat{\gamma}(s) = \sum_s D_s^{(1)} N_s^{\mathcal{P}}$$

- Can usually compute if d is moderate
- "Only" accounts for the linear variables

Post-stratification

Exactly match the counts within each cell:

$$\hat{\gamma}(s) = \frac{\text{\# in Population}}{\text{\# in Sample}}$$

- Impossible to compute with empty cells
- Unbiased when feasible

Choosing weights: Raking and Post-stratification

Raking on margins

Exactly match the counts for margins:

$$\sum_s D_s^{(1)} n_s^{\mathcal{R}} \hat{\gamma}(s) = \sum_s D_s^{(1)} N_s^{\mathcal{P}}$$

- Can usually compute if d is moderate
- "Only" accounts for the linear variables

Post-stratification

Exactly match the counts within each cell:

$$\hat{\gamma}(s) = \frac{\text{\# in Population}}{\text{\# in Sample}}$$

- Impossible to compute with empty cells
- Unbiased when feasible

"Soft" or "penalized" calibration loosens the raking constraints

[Huang and Fuller, 1978; Rao and Singh, 1997; Park and Fuller, 2009; Guggemos and Tillé, 2010]

What do we want from calibration weights?

Is it worth it to try to post-stratify?

Population regression

$$Y_i = \sum_{k=1}^{K} \eta_k \cdot D^k_{S_i} + e_i$$

What do we want from calibration weights?

Is it worth it to try to post-stratify?

Population regression

$$Y_i = \sum_{k=1}^{K} \eta_k \cdot D_{S_i}^k + e_i$$

Conditional MSE

$$\mathbb{E}\left[(\hat{\mu} - \mu)^2 \mid n^{\mathcal{R}}\right]$$

What do we want from calibration weights?

Is it worth it to try to post-stratify?

Population regression

$$Y_i = \sum_{k=1}^{K} \eta_k \cdot D^k_{S_i} + e_i$$

Conditional MSE

$$\mathbb{E}\left[(\hat{\mu} - \mu)^2 \mid n^{\mathcal{R}}\right] \leq \frac{1}{N^2} \left(\sum_{k=1}^{K} \|\eta_k\|_2 \underbrace{\left\| \sum_s D^{(k)}_s n^{\mathcal{R}}_s \gamma(s) - D^{(k)}_s N^{\mathcal{P}}_s \right\|_2}_{\text{Imbalance}_k} \right)^2$$

What do we want from calibration weights?

Is it worth it to try to post-stratify?

Population regression

$$Y_i = \sum_{k=1}^{K} \eta_k \cdot D^k_{S_i} + e_i$$

Conditional MSE

$$\mathbb{E}\left[(\hat{\mu} - \mu)^2 \mid n^{\mathcal{R}}\right] \leq \frac{1}{N^2} \left(\sum_{k=1}^{K} \|\eta_k\|_2 \underbrace{\left\| \sum_s D^{(k)}_s n^{\mathcal{R}}_s \gamma(s) - D^{(k)}_s N^{\mathcal{P}}_s \right\|_2}_{\text{Imbalance}_k} \right)^2 + \frac{\sigma^2}{N^2} \sum_s \left(n^{\mathcal{R}}_s\right)^2 \gamma(s)^2$$

What do we want from calibration weights?

Is it worth it to try to post-stratify?

Population regression

$$Y_i = \sum_{k=1}^{K} \eta_k \cdot D^k_{S_i} + e_i$$

Conditional MSE

$$\mathbb{E}\left[(\hat{\mu} - \mu)^2 \mid n^{\mathcal{R}}\right] \leq \frac{1}{N^2} \left(\sum_{k=1}^{K} \|\eta_k\|_2 \underbrace{\left\| \sum_s D^{(k)}_s n^{\mathcal{R}}_s \gamma(s) - D^{(k)}_s N^{\mathcal{P}}_s \right\|_2}_{\text{Imbalance}_k} \right)^2 + \frac{\sigma^2}{N^2} \sum_s \left(n^{\mathcal{R}}_s\right)^2 \gamma(s)^2$$

If interactions are weak, approximate post-stratification may be enough

- Bias depends on strength of interaction $\times$ imbalance
- Variance depends on inverse of effective sample size

Approximately post-stratify while at least raking on margins

Find weights via convex optimization:

$$\min_{\gamma} \sum_{k=2}^{d} \frac{1}{\lambda_k} \|\text{Imbalance}_k\|_2^2 + \sum_s n_s^{\mathcal{R}} \gamma(s)^2$$

Approximately post-stratify while at least raking on margins

Find weights via convex optimization:

$$\min_{\gamma} \sum_{k=2}^{d} \frac{1}{\lambda_k} \|\text{Imbalance}_k\|_2^2 + \sum_s n_s^{\mathcal{R}} \gamma(s)^2$$
$$\text{subject to} \quad \text{Imbalance}_1 = 0, \quad \gamma(s) \geq 0$$

Approximately post-stratify while at least raking on margins

Find weights via convex optimization:

$$\min_{\gamma} \sum_{k=2}^{d} \frac{1}{\lambda_k} \|\text{Imbalance}_k\|_2^2 + \sum_s n_s^{\mathcal{R}} \gamma(s)^2$$
$$\text{subject to} \quad \text{Imbalance}_1 = 0, \quad \gamma(s) \geq 0$$

Move smoothly between two extremes

- With $\lambda_k \to \infty$, recover raking on margins
- With $\lambda_k \to 0$, recover post-stratification

Based on calibration weighting and approximate balancing weights

[Deville and Särndal, 1992; Deville et al., 1993; Zubizarreta, 2015; Wong et al., 2018; Hirshberg et al., 2019]

Dual view: multilevel regression for response

A **regularized** model for the **inverse** probability of response:

[Zhao and Percival, 2016; Wang and Zubizarreta, 2020; Chattopadhyay et al., 2020]

$$\frac{1}{\pi_i} \sim \beta_1 \cdot D^1_{S_i} + \sum_{k=2}^{K} \beta_k \cdot D^k_{S_i}$$

Dual view: multilevel regression for response

A **regularized** model for the **inverse** probability of response:

[Zhao and Percival, 2016; Wang and Zubizarreta, 2020; Chattopadhyay et al., 2020]

$$\frac{1}{\pi_i} \sim \beta_1 \cdot D^1_{S_i} + \sum_{k=2}^{K} \beta_k \cdot D^k_{S_i}$$

Regularization makes approximate post-stratification feasible

$$0 \times \|\beta_1\|_2^2 + \sum_{k=2}^{K} \lambda_k \|\beta_k\|_2^2$$

- At least raking → no regularization for marginal probabilities [Little and Wu, 1991]
- Approximate post-stratification → regularizing interaction terms

Key difference with GLM: regularizing for balance

Double Regression with Post-Stratification (DRP)

MRP-style approaches

Start with an outcome model for the cells $\hat{\mu}_s$

- Multilevel model, high dimensional regression, tree-based methods

  [Gelman and Little, 1997; Ghitza and Gelman, 2013; Gao et al., 2020; Montgomery and Olivella, 2018; Bisbee, 2019]

MRP-style approaches

Start with an outcome model for the cells $\hat{\mu}_s$

- Multilevel model, high dimensional regression, tree-based methods

  [Gelman and Little, 1997; Ghitza and Gelman, 2013; Gao et al., 2020; Montgomery and Olivella, 2018; Bisbee, 2019]

Post-stratify using predictions instead of outcomes

$$\hat{\mu}^{\text{mrp}} = \frac{1}{N} \sum_s \frac{N_s^{\mathcal{P}}}{n_s^{\mathcal{R}}} n_s^{\mathcal{R}} \hat{\mu}_s = \frac{1}{N} \sum_s N_s^{\mathcal{P}} \hat{\mu}_s$$

DRP: a more surgical approach

Instead of relying on the model everywhere, use it to pick up slack from weighting

DRP: a more surgical approach

Instead of relying on the model everywhere, use it to pick up slack from weighting

$$\hat{\mu}^{\text{drp}} = \hat{\mu}(\hat{\gamma})$$

DRP: a more surgical approach

Instead of relying on the model everywhere, use it to pick up slack from weighting

$$\hat{\mu}^{\text{drp}} = \hat{\mu}(\hat{\gamma}) \quad + \frac{1}{N} \sum_{s} \hat{\mu}_s \times \text{imbalance in cell } s$$

DRP: a more surgical approach

Instead of relying on the model everywhere, use it to pick up slack from weighting

$$\hat{\mu}^{\text{drp}} = \hat{\mu}(\hat{\gamma}) \quad + \frac{1}{N} \sum_s \hat{\mu}_s \times \text{imbalance in cell } s$$

$$= \hat{\mu}^{\text{mrp}}$$

DRP: a more surgical approach

Instead of relying on the model everywhere, use it to pick up slack from weighting

$$\hat{\mu}^{\text{drp}} = \hat{\mu}(\hat{\gamma}) \quad + \frac{1}{N} \sum_s \hat{\mu}_s \times \text{imbalance in cell } s$$

$$= \hat{\mu}^{\text{mrp}} \quad + \frac{1}{N} \sum_s n_s^{\mathcal{R}} \hat{\gamma}(s) \times \text{error in cell } s$$

DRP: a more surgical approach

Instead of relying on the model everywhere, use it to pick up slack from weighting

$$\hat{\mu}^{\text{drp}} = \hat{\mu}(\hat{\gamma}) \quad + \frac{1}{N} \sum_s \hat{\mu}_s \times \text{imbalance in cell } s$$

$$= \hat{\mu}^{\text{mrp}} \quad + \frac{1}{N} \sum_s n_s^{\mathcal{R}} \hat{\gamma}(s) \times \text{error in cell } s$$

Augmented balancing weights estimator, related to AIPW and bias-correction

[Cassel et al., 1976; Robins et al., 1994; Abadie and Imbens, 2006; Hirshberg and Wager, 2019]

- Relies on outcome model in cells where weighting doesn't get it right
- Relies on weights to adjust cells where model is off
- If post-stratifying, collapses to weighting estimator $\hat{\mu}^{\text{drp}} = \hat{\mu}(\hat{\gamma})$

The role of the outcome model: bias correction

Bias term depends on how well the model adjusts for remaining imbalance

$$\hat{\mu}^{\text{drp}} - \mu \;=\; \frac{1}{N}\sum_{s} \text{error in cell } s \times \text{imbalance in cell } s \;+\; \text{noise}$$

The role of the outcome model: bias correction

Bias term depends on how well the model adjusts for remaining imbalance

$$\hat{\mu}^{\text{drp}} - \mu \;=\; \frac{1}{N}\sum_{s} \text{error in cell } s \times \text{imbalance in cell } s \;+\; \text{noise}$$

Because we can only approximately post-stratify, bias correction is key!

- Outcome model primarily performs bias correction, prioritize bias over variance
- Include higher order interactions than usual, or deeper trees
- Limits extrapolation

$\hat{\mu}^{\text{mrp}}$ relies entirely rely on outcome model

- Choose a lower variance, higher bias solution
- Fewer interaction terms, shallower trees
- Potentially unchecked extrapolation

Case study: 2016 presidential election

Case study: 2016 presidential election

Pre-election Pew poll of vote intention with $\sim$ 2,000 respondents

- Age, gender, race, region, party ID, education, income, born again Christian

Ground truth: weighted CCES, large sample $N \sim 45{,}000$

Interactions should be important here [Kennedy et al., 2018]

Case study: 2016 presidential election

Pre-election Pew poll of vote intention with $\sim$ 2,000 respondents

- Age, gender, race, region, party ID, education, income, born again Christian

Ground truth: weighted CCES, large sample $N \sim 45{,}000$

Interactions should be important here [Kennedy et al., 2018]

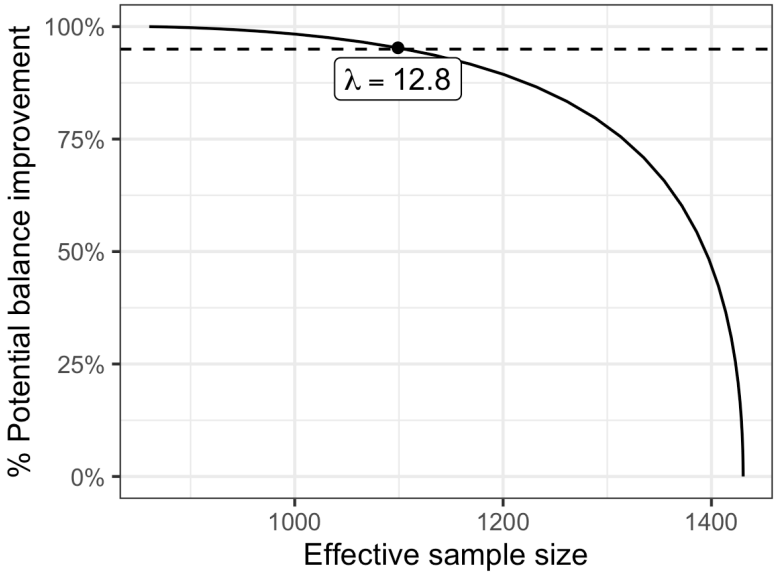
Use this case study in 3 ways

1. Calibrated simulation study [see paper!]
2. Imbalance and estimates with full weighted CCES as target

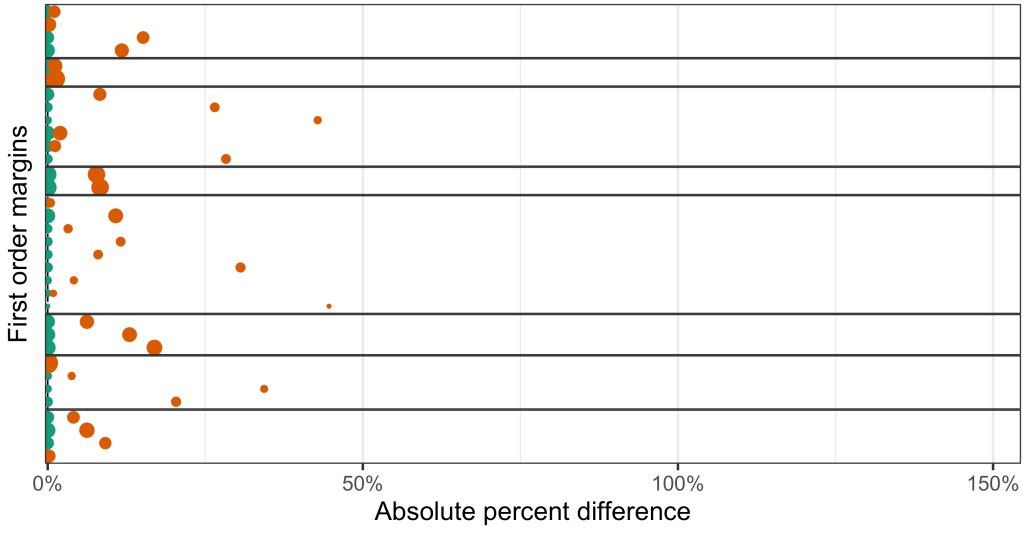
$$\frac{\left|N_s^{\mathcal{P}} - n_s^{\mathcal{R}} \hat{\gamma}(s)\right|}{N_s^{\mathcal{P}}}$$

3. Impute Republican vote share within each state
   - This is challenging! Pew not sampled to represent states well

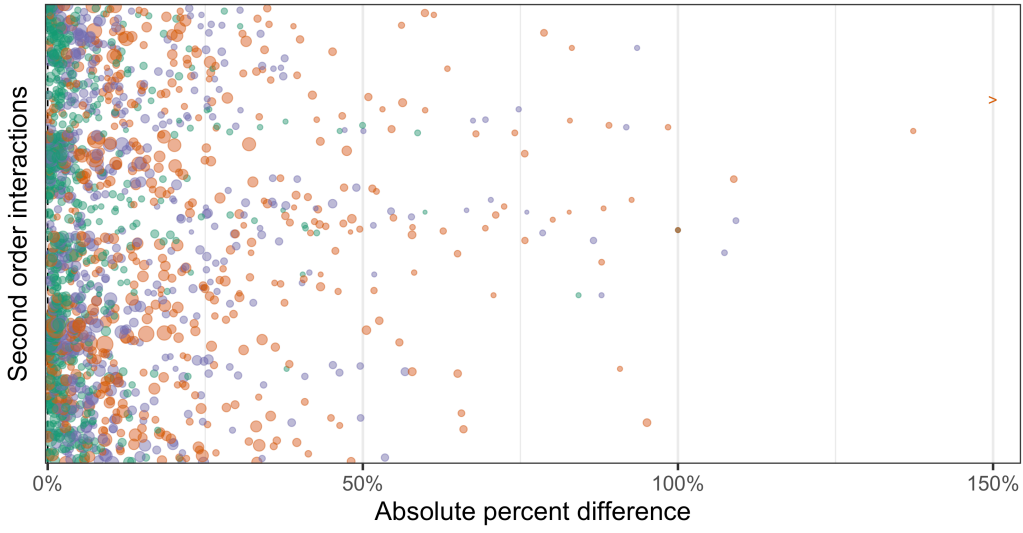
Negotiating balance vs effective sample size



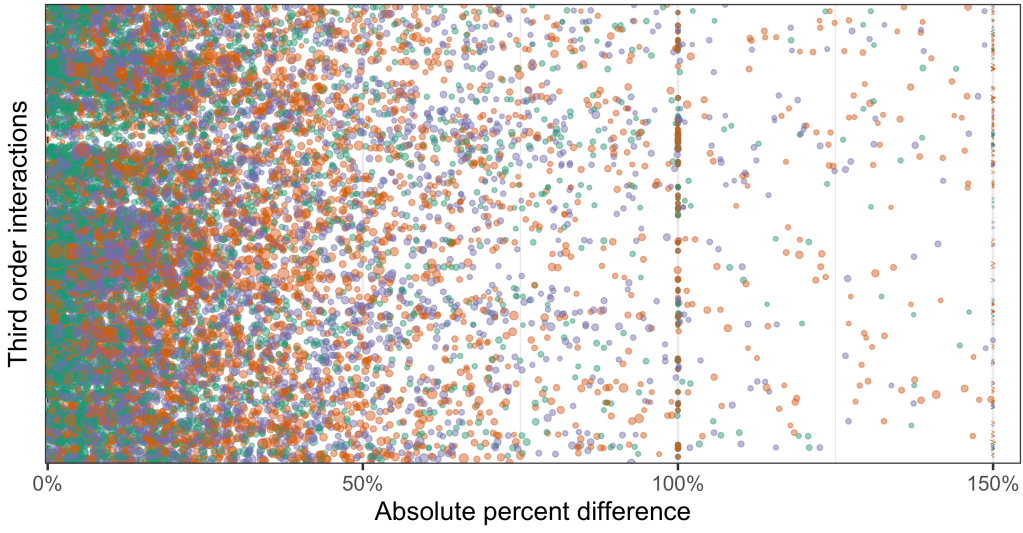
Both multilevel weighting and raking exactly match margins...



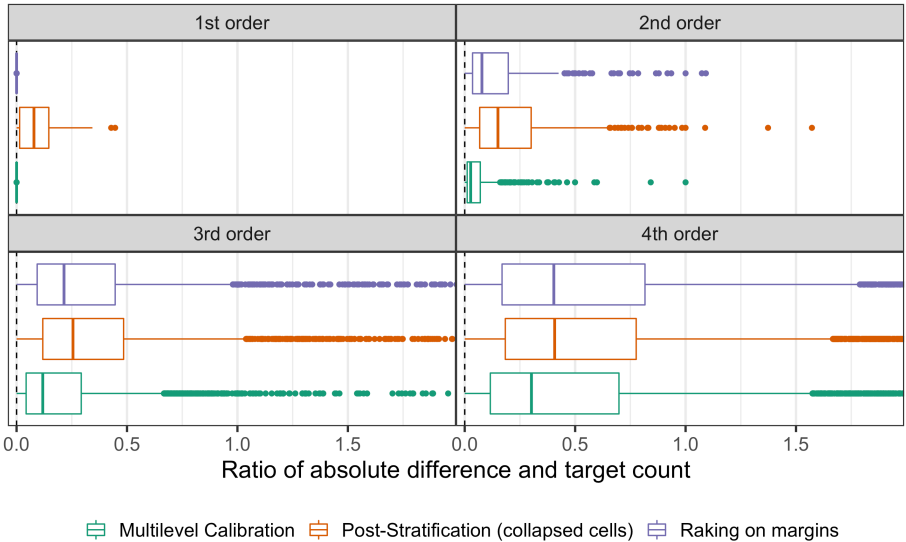
...but raking fails to balance higher order interactions



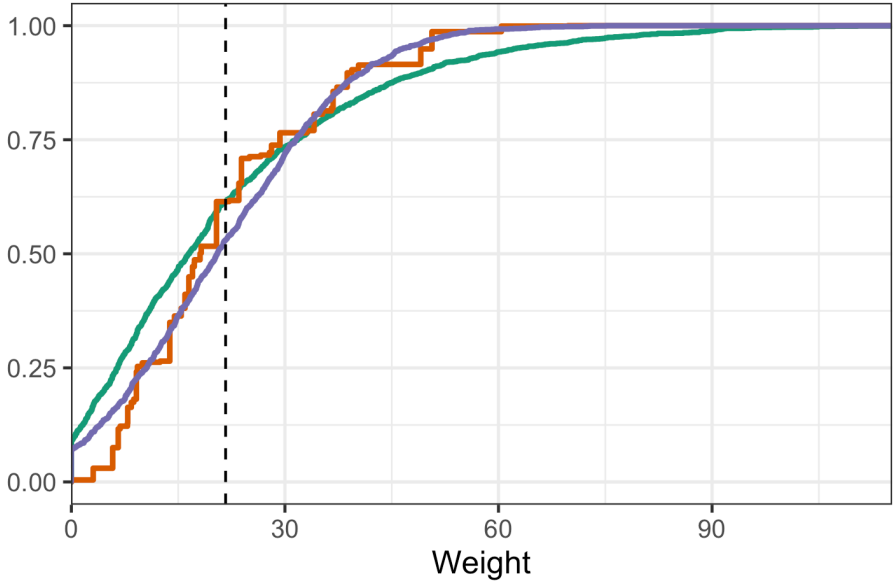
Approximate balance in 3rd order interactions



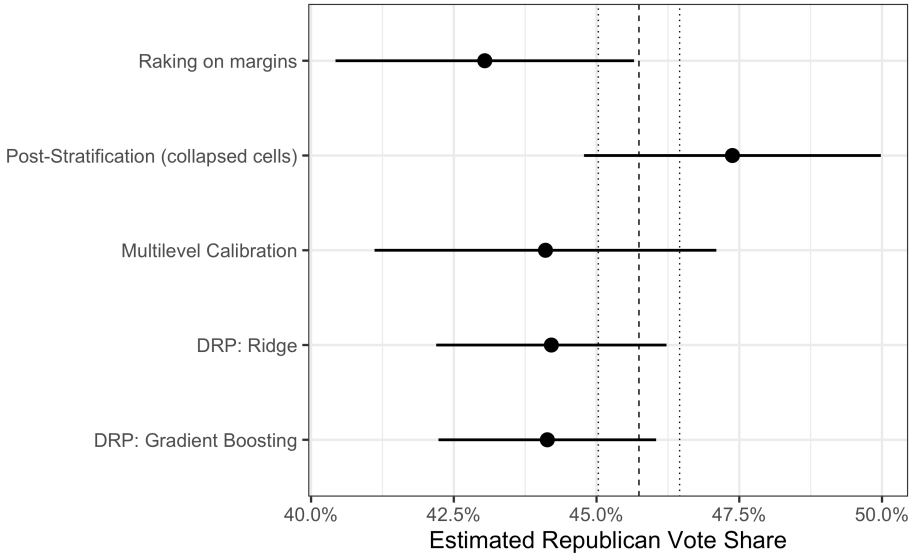
Overall multilevel calibration yields better balance



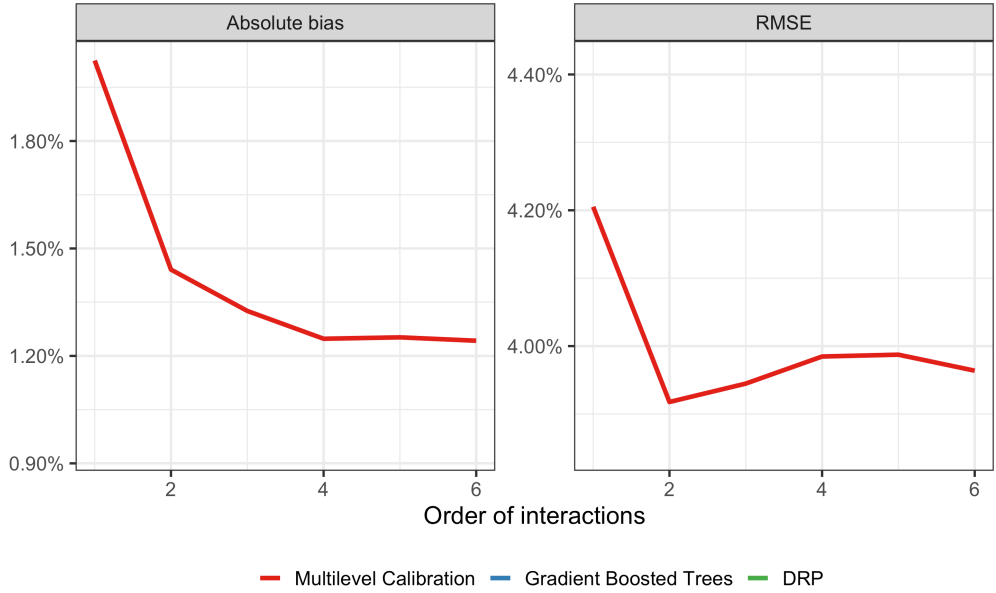
At the cost of lower effective sample sizes



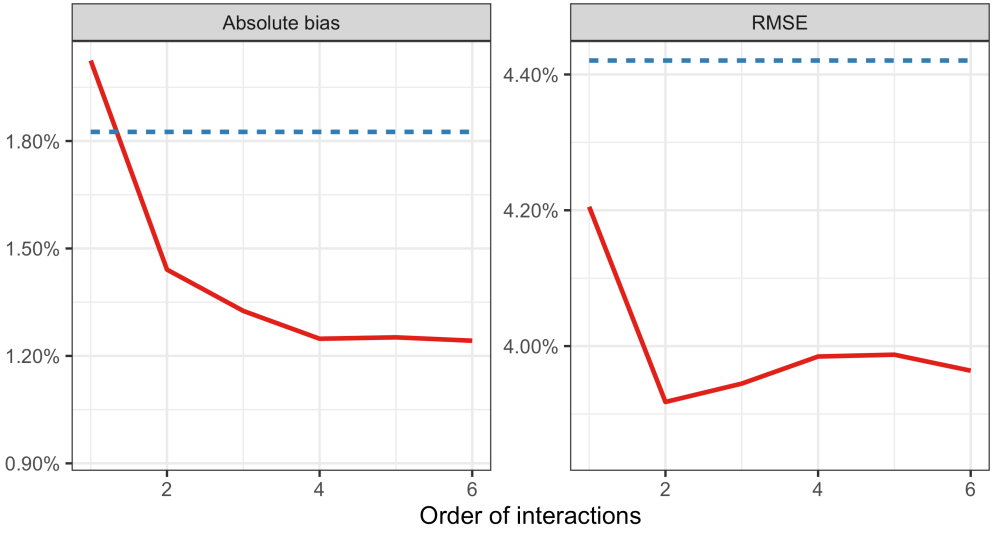
Overall estimate of Republican vote share



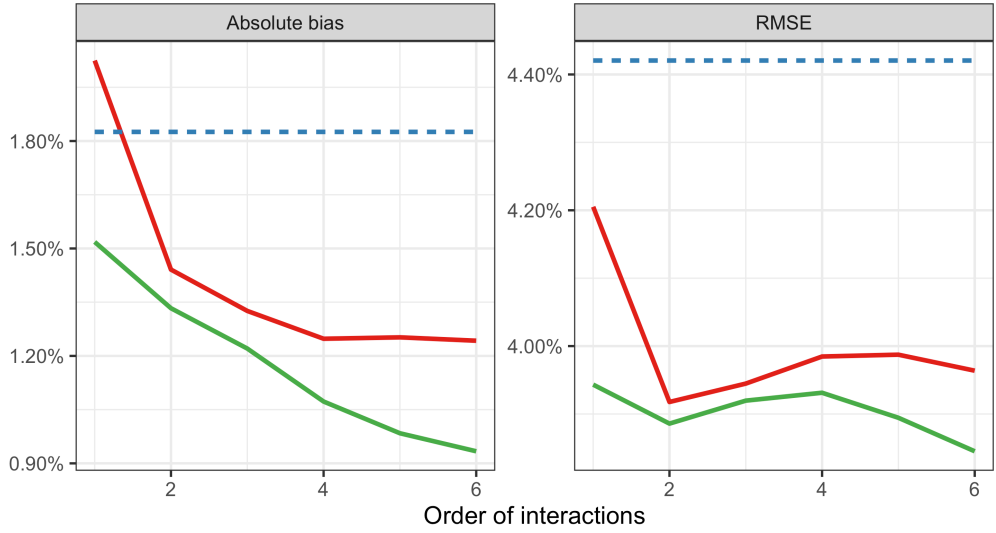
Weighting on higher order interactions reduces bias



Weighting on higher order interactions reduces bias



DRP gives further improvements



Recap

Principled, parsimonious way of leveraging the value of interactions

- Multilevel weighting as middle ground between raking and post-stratification
- Dual view as multilevel model for non-response
- DRP adjusts cells where weighting misses the mark

Implemented in `multical` `R` package

Recap

Principled, parsimonious way of leveraging the value of interactions

- Multilevel weighting as middle ground between raking and post-stratification
- Dual view as multilevel model for non-response
- DRP adjusts cells where weighting misses the mark

Implemented in `multical` `R` package

Thank you!

ebenmichael.github.io
arxiv.org/abs/2102.09052
multical

Appendix

Calibrated simulation study

Create a combined sample from Pew and CCES, with Pew respondents as $R_i = 1$

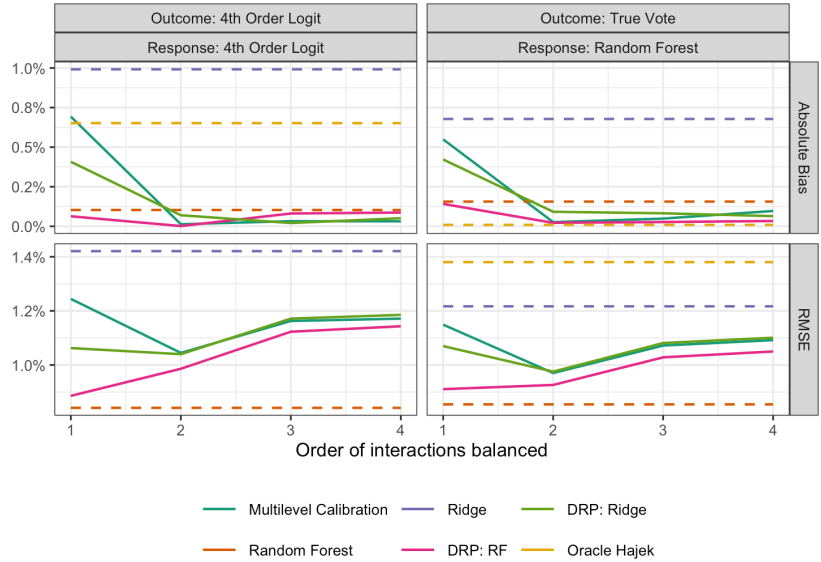
Fit two different response models:

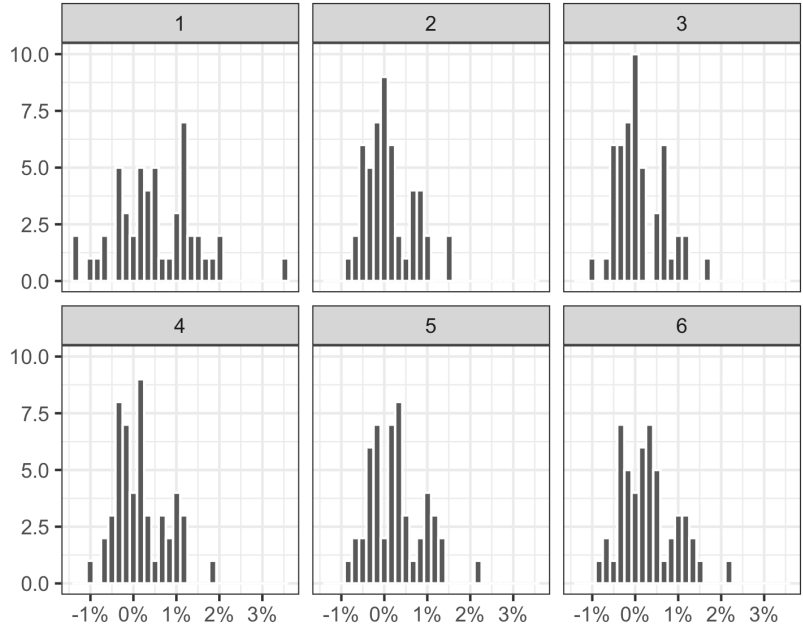
1. Random forest
2. 4^{th} order ridge penalized logistic regression with **low regularization**

Two different outcomes:

1. Actual Republican vote
2. Sampled from 4^{th} order ridge penalized logistic regression with **low regularization**

Balancing higher order interactions helps, so does including an outcome model




1
2
3
4
5
6
10.0
7.5
5.0
2.5
0.0
-1%
0%
1%
2%
3%
Difference between DRP and weighted estimate

References I

Abadie, A. and Imbens, G. W. (2006). Large Sample Properties of Matching Estimators. *Econometrica*, 74(1):235–267.

Bisbee, J. (2019). Barp: Improving mister p using bayesian additive regression trees. *American Political Science Review*, 113(4):1060–1065.

Cassel, C. M., Sarndal, C.-E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

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

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

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.

Gao, Y., Kennedy, L., Simpson, D., Gelman, A., et al. (2020). Improving multilevel regression and poststratification with structured priors. *Bayesian Analysis*.

Gelman, A. and Little, T. C. (1997). Poststratification Into Many Categories Using Hierarchical Logistic Regression. *Survey Methodology*, 23(2):127–135.

References II

Ghitza, Y. and Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3):762–776.

Guggemos, F. and Tillé, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference*, 140(11):3199–3212.

Hirshberg, D. and Wager, S. (2019). Augmented Minimax Linear Estimation.

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

Huang, E. T. and Fuller, W. A. (1978). Nonnegative regression estimation in sample survey data. In *Proceedings of the Section on Survey Research Methods*, pages 300–305.

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., Saad, L., Witt, G. E., and Wlezien, C. (2018). An Evaluation of the 2016 Election Polls in the United States. *Public Opinion Quarterly*, 82(1):1–33.

Little, R. J. and Wu, M. M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association*, 86(413):87–95.

References III

Montgomery, J. M. and Olivella, S. (2018). Tree-Based Models for Political Science Data. *American Journal of Political Science*, 62(3):729–744.

Park, M. and Fuller, W. A. (2009). The mixed model for survey regression estimation. *Journal of Statistical Planning and Inference*, 139(4):1320–1331.

Rao, J. N. K. and Singh, A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. In *SA Proceedings of the Section on Survey Research Methods*, pages 57–85.

Robins, J. M., Rotnitzky, A., Ping Zhao, L., and Ping ZHAO, L. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89427:846–866.

Wang, Y. and Zubizarreta, J. R. (2020). Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105.

Wong, R. K. W., Chuen, K., and Chan, G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1):199–213.

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.