Double Regression with Post-stratification (DRP)

for high-dimensional survey data

Eli Ben-Michael, Avi Feller, and Erin Hartman

Harvard & UC Berkeley & UCLA

AsianPolmeth VIII & ASQPS IX

January 2021

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

Raking

More desirable

Post-stratification

Even in moderate dimensions, survey adjustment is hard

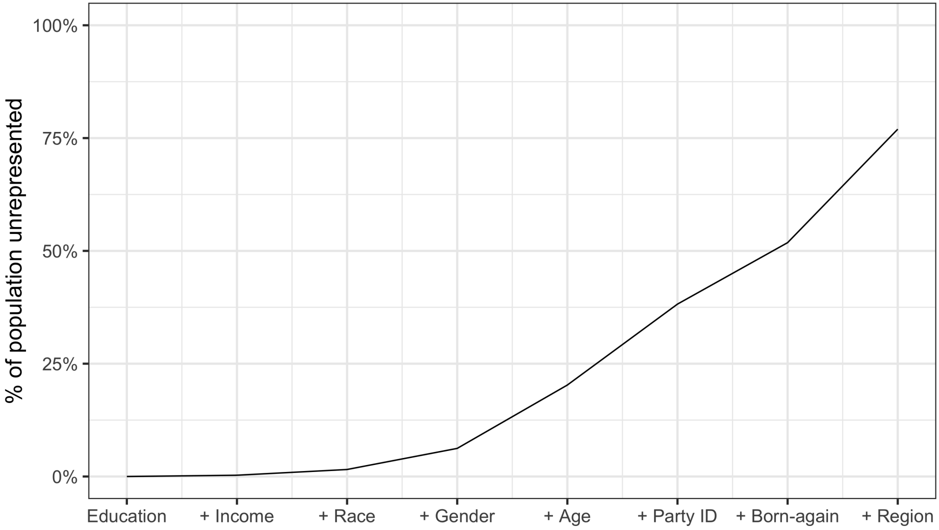
Raking

More desirable

Less feasible

Post-stratification

Quickly run into empty cells



How can we account for interactions in a principled way?

Adjusting for interactions in non-response is important

- Example: 2016 presidential election polling [Kennedy et al., 2018]

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

- Example: 2016 presidential election polling [Kennedy et al., 2018]

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 $\rightarrow$ 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

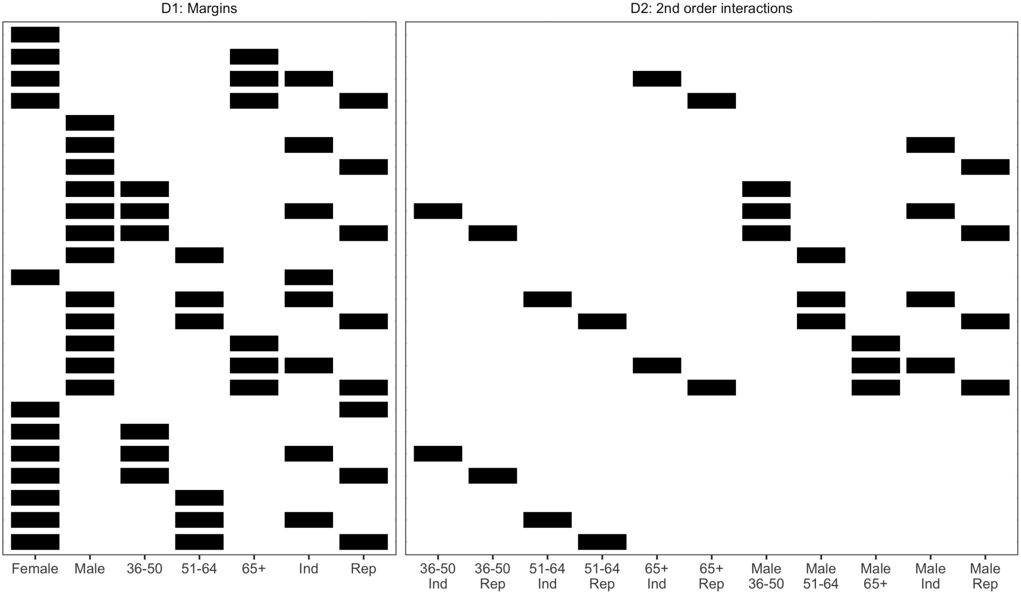
$i = 1, \ldots, N$ individuals

- 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
Male
36-50
51-64
65+
Ind
Rep
36-50 Ind
36-50 Rep
51-64 Ind
51-64 Rep
65+ Ind
65+ Rep
Male 36-50
Male 51-64
Male 65+
Male Ind
Male 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:

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

- Impossible to compute with empty cells
- Unbiased when feasible

What do we want from calibration weights?

Is it worth it to try to post-stratify?

What do we want from calibration weights?

Is it worth it to try to post-stratify?

Population regression

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

What do we want from calibration weights?

Is it worth it to try to post-stratify?

Population regression

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

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

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

Conditional MSE

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

What do we want from calibration weights?

Is it worth it to try to post-stratify?

Population regression

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

Conditional MSE

$$\mathbb{E}\left[(\hat{\mu} - \mu)^2 \mid n^{\mathcal{R}}\right] \leq \frac{1}{N^2}\left(\sum_{k=1}^{K} \|\beta_k\|_2 \left\| \sum_s D_s^k n_s^{\mathcal{R}} \gamma(s) - D_s^k N_s^{\mathcal{P}} \right\|_2\right)^2 + \frac{\sigma^2}{N^2} \sum_s \left(n_s^{\mathcal{R}}\right)^2 \gamma(s)^2$$

What do we want from calibration weights?

Is it worth it to try to post-stratify?

Population regression

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

Conditional MSE

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

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

- Bias depends on strength of interaction $\times$ imbalance
- Variance depends on sum of the squared weights $\|\hat{\gamma}\|_2^2$

Approximately post-stratify while at least raking on margins

Find weights via convex optimization:

$$\min_{\gamma} \sum_{k=2}^{d} \frac{1}{\lambda_k} \left\| \sum_{s} D_s^k n_s^{\mathcal{R}} \gamma(s) - D_s^k N_s^{\mathcal{P}} \right\|_2^2 + \sum_{s} n_s^{\mathcal{R}} \gamma(s)^2$$

Approximately post-stratify while at least raking on margins

Find weights via convex optimization:

$$\min_{\gamma} \quad \sum_{k=2}^{d} \frac{1}{\lambda_k} \left\| \sum_s D_s^k n_s^{\mathcal{R}} \gamma(s) - D_s^k N_s^{\mathcal{P}} \right\|_2^2 + \sum_s n_s^{\mathcal{R}} \gamma(s)^2$$

$$\text{subject to} \quad \sum_s D_s^1 n_s^{\mathcal{R}} \gamma(s) = \sum_s D_s^1 N_s, \quad \gamma(s) > 0$$

Approximately post-stratify while at least raking on margins

Find weights via convex optimization:

$$\min_{\gamma} \sum_{k=2}^{d} \frac{1}{\lambda_k} \left\| \sum_s D_s^k n_s^{\mathcal{R}} \gamma(s) - D_s^k N_s^{\mathcal{P}} \right\|_2^2 + \sum_s n_s^{\mathcal{R}} \gamma(s)^2$$

$$\text{subject to} \quad \sum_s D_s^1 n_s^{\mathcal{R}} \gamma(s) = \sum_s D_s^1 N_s, \quad \gamma(s) > 0$$

Move smoothly between two extremes

- With $\lambda_k \to \infty$, recover raking on margins
- With $\lambda_k \to 0$, recover post-stratification
- With $\lambda_k = 1$, cell weights are regularized by cell size

Based on calibration weighting and approximate balancing weights

[Deville and Särndal, 1992; Deville et al., 1993; Zubizarreta, 2015; 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]

$$\frac{1}{\pi_i} \sim \left[\theta_1 \cdot D^1_{S_i} + \sum_{k=2}^{K} \theta_k \cdot D^k_{S_i}\right]_+$$

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]

$$\frac{1}{\pi_i} \sim \left[\theta_1 \cdot D^1_{S_i} + \sum_{k=2}^{K} \theta_k \cdot D^k_{S_i}\right]_+$$

Regularization makes approximate post-stratification feasible

$$0 \times \|\theta_1\|_2^2 + \sum_{k=2}^{K} \lambda_k \|\theta_k\|_2^2$$

- At least raking $\rightarrow$ no regularization for marginal probabilities [Little and Wu, 1991]
- Approximate post-stratification $\rightarrow$ 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

MRP-style approaches

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

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

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

$$\hat{\mu}^{\text{drp}} = \hat{\mu}(\hat{\gamma}) \quad + \frac{1}{N} \sum_{s} \hat{\mu}_s \times \underbrace{\left(N_s^{\mathcal{P}} - n_s^{\mathcal{R}} \hat{\gamma}(s)\right)}_{\text{imbalance in cell } s}$$

DRP: a more surgical approach

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

$$
\begin{aligned}
\hat{\mu}^{\text{drp}} &= \hat{\mu}(\hat{\gamma}) \quad + \frac{1}{N} \sum_s \hat{\mu}_s \times \underbrace{\left(N_s^{\mathcal{P}} - n_s^{\mathcal{R}} \hat{\gamma}(s)\right)}_{\text{imbalance in cell } s} \\
&= \hat{\mu}^{\text{mrp}}
\end{aligned}
$$

DRP: a more surgical approach

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

$$
\begin{aligned}
\hat{\mu}^{\text{drp}} &= \hat{\mu}(\hat{\gamma}) \quad + \frac{1}{N}\sum_s \hat{\mu}_s \times \underbrace{\left(N_s^{\mathcal{P}} - n_s^{\mathcal{R}}\hat{\gamma}(s)\right)}_{\text{imbalance in cell } s} \\
&= \hat{\mu}^{\text{mrp}} \quad + \frac{1}{N}\sum_s n_s^{\mathcal{R}}\hat{\gamma}(s) \times \underbrace{(\bar{Y}_s - \hat{\mu}_s)}_{\text{error in cell } s}
\end{aligned}
$$

DRP: a more surgical approach

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

$$
\begin{aligned}
\hat{\mu}^{\text{drp}} &= \hat{\mu}(\hat{\gamma}) \quad + \frac{1}{N}\sum_s \hat{\mu}_s \times \underbrace{\left(N_s^{\mathcal{P}} - n_s^{\mathcal{R}}\hat{\gamma}(s)\right)}_{\text{imbalance in cell } s} \\
&= \hat{\mu}^{\text{mrp}} \quad + \frac{1}{N}\sum_s n_s^{\mathcal{R}}\hat{\gamma}(s) \times \underbrace{\left(\bar{Y}_s - \hat{\mu}_s\right)}_{\text{error in cell } s}
\end{aligned}
$$

Augmented balancing weights estimator, related to double robust and bias-corrected estimators

[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 $\hat{\mu}^{\text{drp}} = \hat{\mu}^{\text{weight}}$

Bias variance tradeoff

Improvement in MSE depends on how good the weights and the model are together

Bias variance tradeoff

Improvement in MSE depends on how good the weights and the model are together

Bias adjustment:

$$\mathbb{E}\left[\hat{\mu}^{\text{drp}} - \mu \mid n^{\mathcal{R}}\right]$$

Bias variance tradeoff

Improvement in MSE depends on how good the weights and the model are together

Bias adjustment:

$$\mathbb{E}\left[\hat{\mu}^{\text{drp}} - \mu \mid n^{\mathcal{R}}\right] = \frac{1}{N} \sum_{s} \left(n_s^{\mathcal{R}} \hat{\gamma}(s) - N_s^{\mathcal{P}}\right)$$

Bias variance tradeoff

Improvement in MSE depends on how good the weights and the model are together

Bias adjustment:

$$\mathbb{E}\left[\hat{\mu}^{\text{drp}} - \mu \mid n^{\mathcal{R}}\right] = \frac{1}{N}\sum_{s}\left(n_s^{\mathcal{R}}\hat{\gamma}(s) - N_s^{\mathcal{P}}\right) \times \mathbb{E}\left[\hat{\mu}_s - \mu_s \mid n^{\mathcal{R}}\right]$$

Bias variance tradeoff

Improvement in MSE depends on how good the weights and the model are together

Bias adjustment:

$$\mathbb{E}\left[\hat{\mu}^{\text{drp}} - \mu \mid n^{\mathcal{R}}\right] = \frac{1}{N}\sum_{s}\left(n_s^{\mathcal{R}}\hat{\gamma}(s) - N_s^{\mathcal{P}}\right) \times \mathbb{E}\left[\hat{\mu}_s - \mu_s \mid n^{\mathcal{R}}\right]$$

Potentially small increase in variance

$$\text{Var}\left(\hat{\mu}^{\text{drp}} \mid n^{\mathcal{R}}\right)$$

Bias variance tradeoff

Improvement in MSE depends on how good the weights and the model are together

Bias adjustment:

$$\mathbb{E}\left[\hat{\mu}^{\text{drp}} - \mu \mid n^{\mathcal{R}}\right] = \frac{1}{N}\sum_s \left(n_s^{\mathcal{R}}\hat{\gamma}(s) - N_s^{\mathcal{P}}\right) \times \mathbb{E}\left[\hat{\mu}_s - \mu_s \mid n^{\mathcal{R}}\right]$$

Potentially small increase in variance

$$\text{Var}\left(\hat{\mu}^{\text{drp}} \mid n^{\mathcal{R}}\right) = \text{Var}\left(\hat{\mu}(\hat{\gamma}) \mid n^{\mathcal{R}}\right)$$

Bias variance tradeoff

Improvement in MSE depends on how good the weights and the model are together

Bias adjustment:

$$\mathbb{E}\left[\hat{\mu}^{\text{drp}} - \mu \mid n^{\mathcal{R}}\right] = \frac{1}{N}\sum_s \left(n_s^{\mathcal{R}}\hat{\gamma}(s) - N_s^{\mathcal{P}}\right) \times \mathbb{E}\left[\hat{\mu}_s - \mu_s \mid n^{\mathcal{R}}\right]$$

Potentially small increase in variance

$$\text{Var}\left(\hat{\mu}^{\text{drp}} \mid n^{\mathcal{R}}\right) = \text{Var}\left(\hat{\mu}(\hat{\gamma}) \mid n^{\mathcal{R}}\right) + \frac{1}{N^2}\sum_s (N_s^{\mathcal{P}} - n_s^{\mathcal{R}}\hat{\gamma}(s))^2 \text{Var}(\hat{\mu}_s \mid n_s^{\mathcal{R}})$$

Case study

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
2. Look at 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

Calibrated simulation study

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

Calibrated simulation study

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

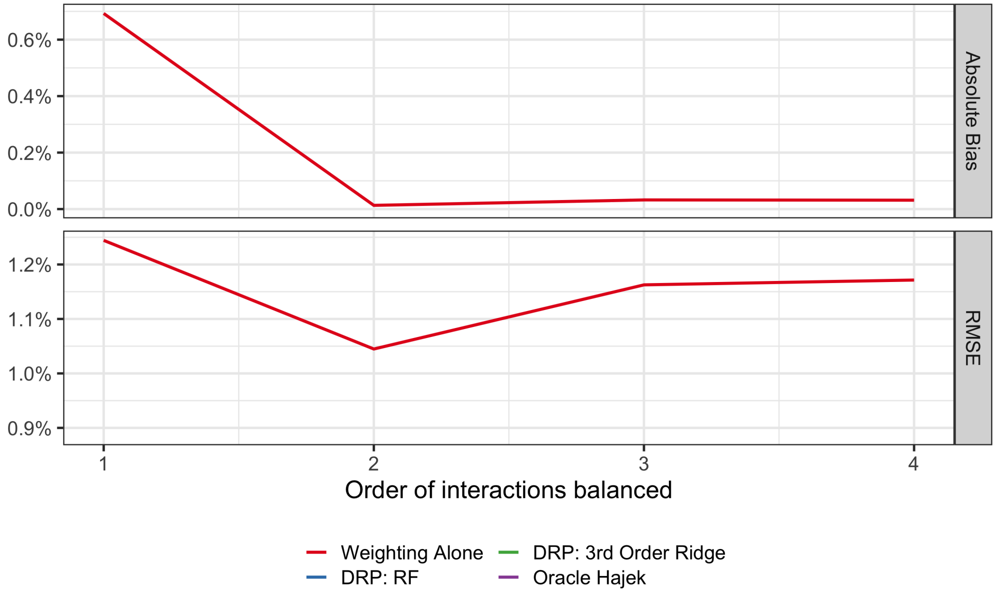
Fit two different response models:

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

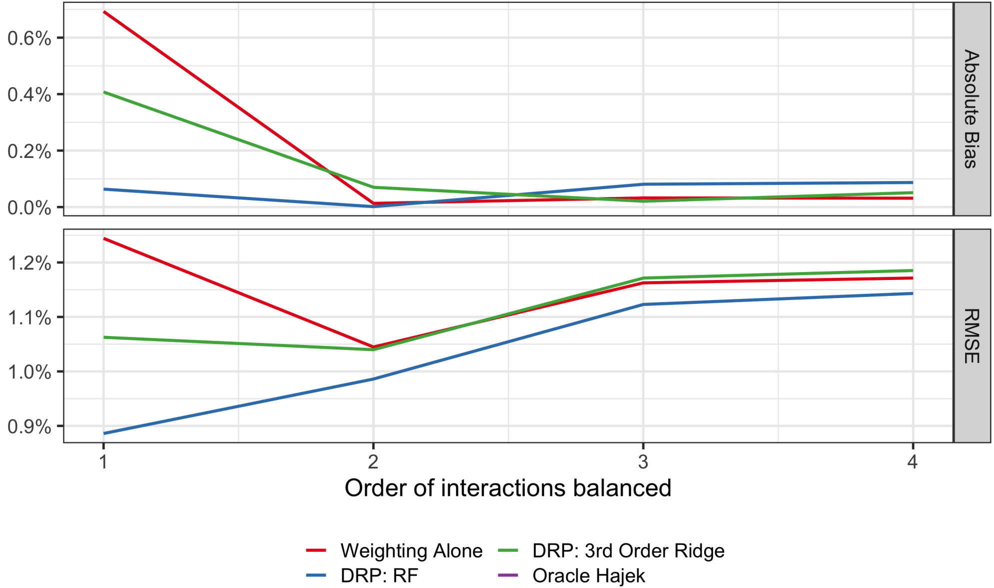
Two different outcomes:

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

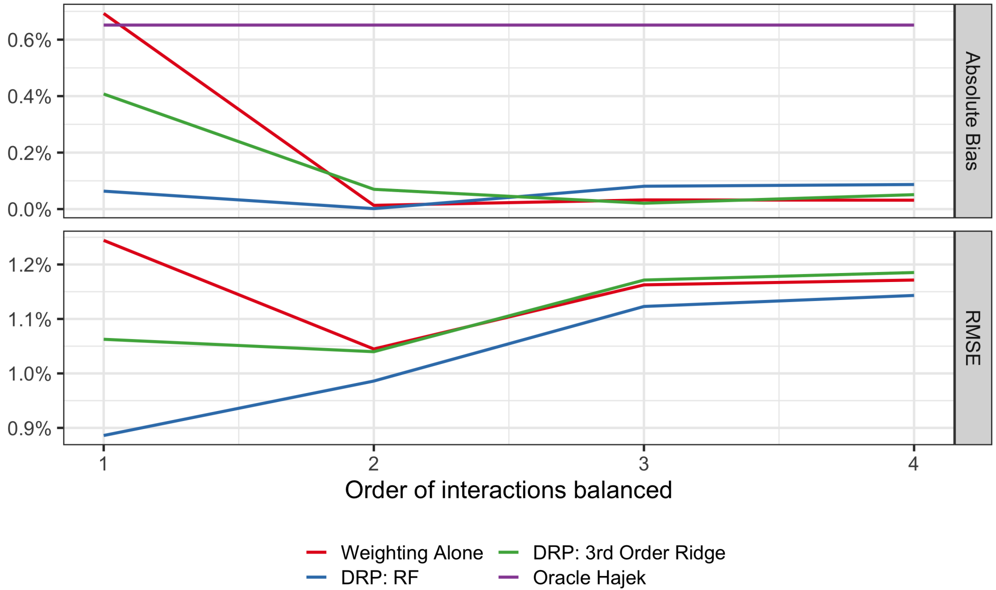
Outcome and Response: 4th order logit



Outcome and Response: 4^{th} order logit

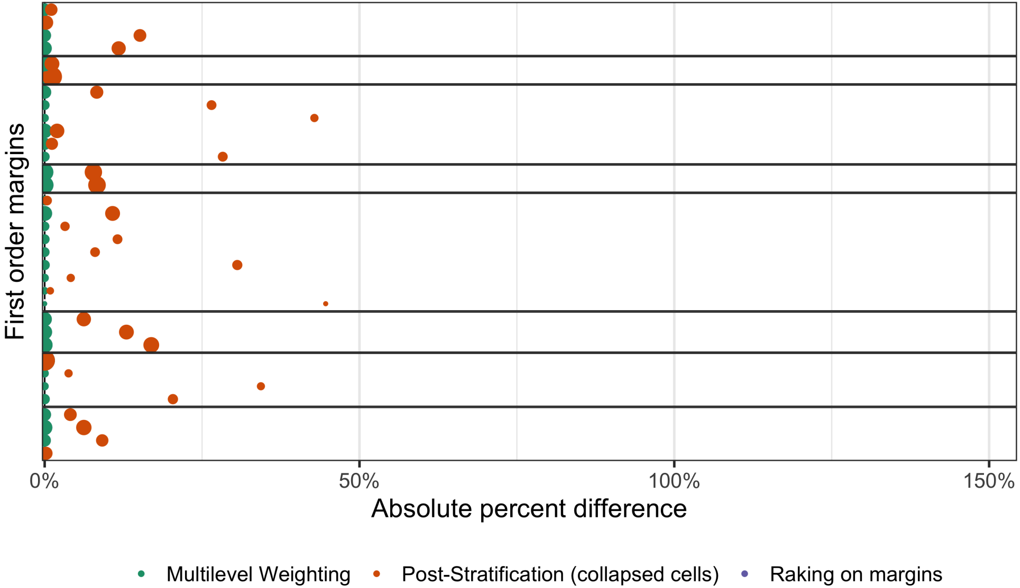


Outcome and Response: 4th order logit

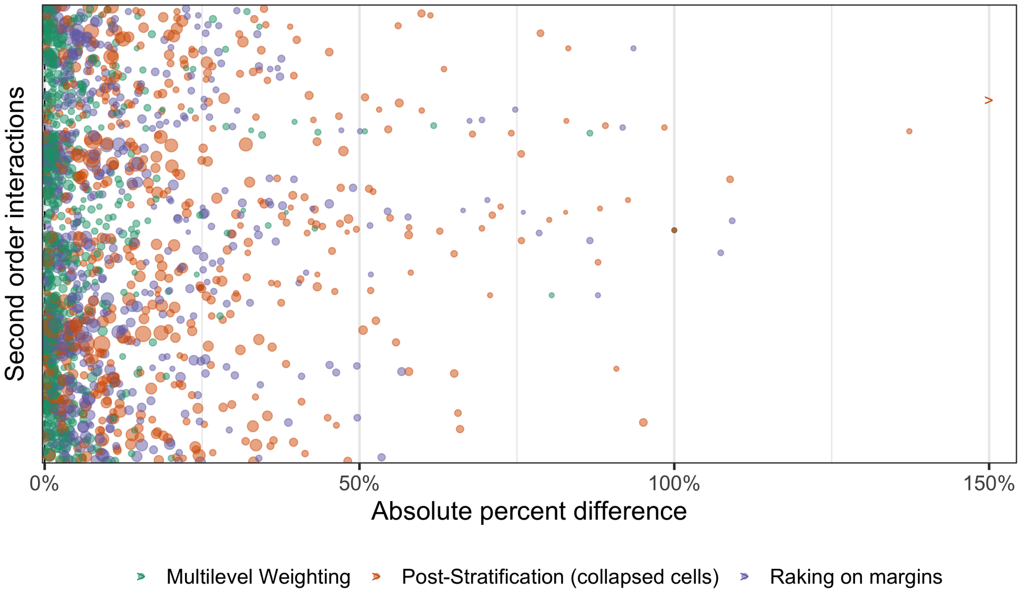


Full CCES

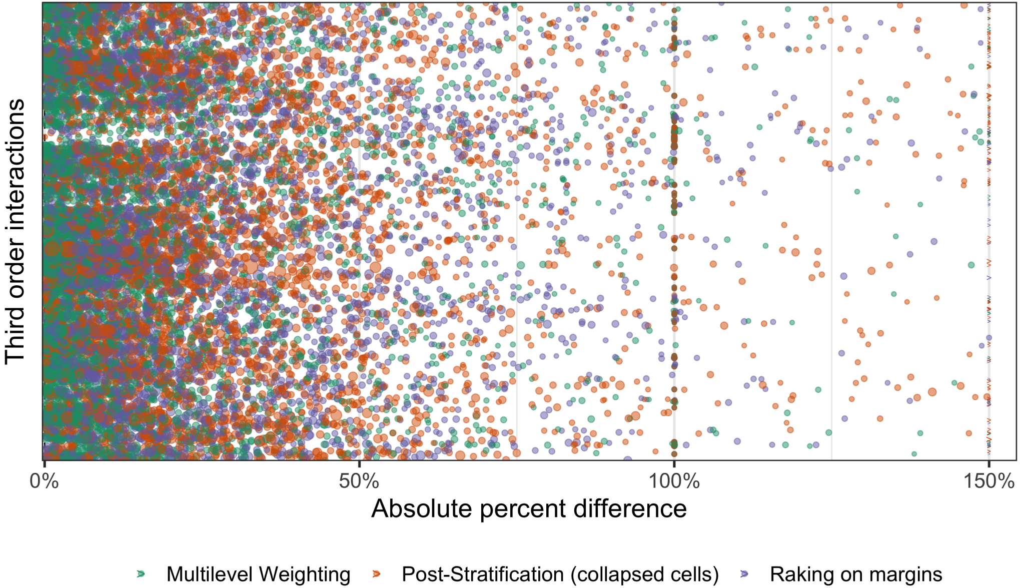
Both multilevel weighting and raking exactly match margins...

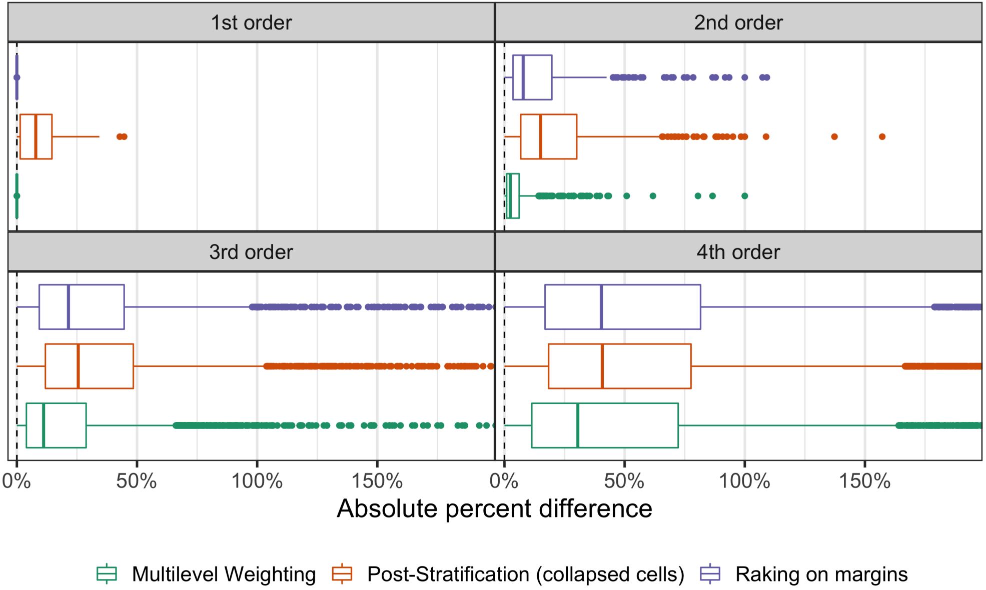


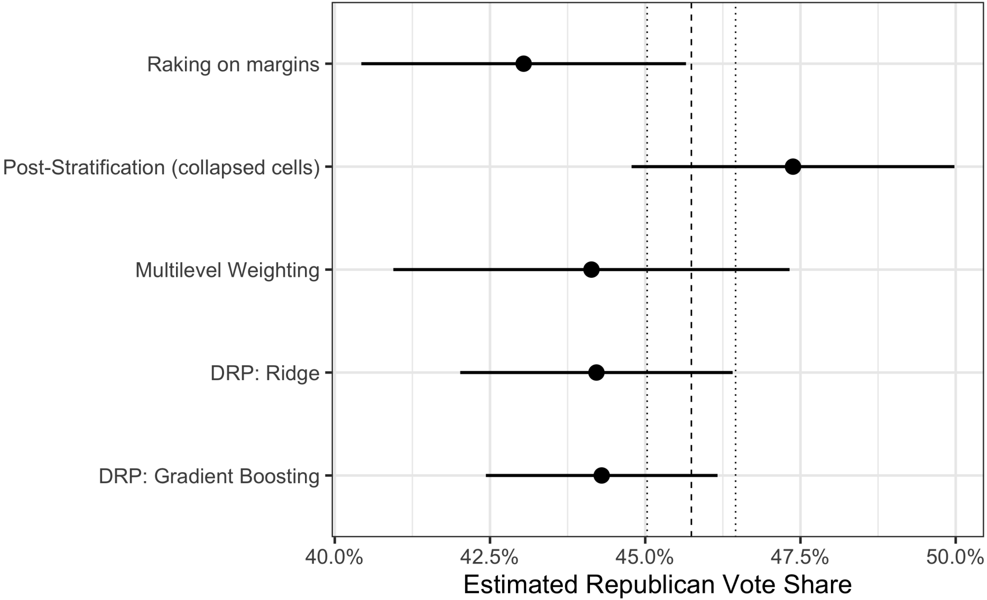
...but raking fails to balance higher order interactions



Approximate balance in 3rd order interactions

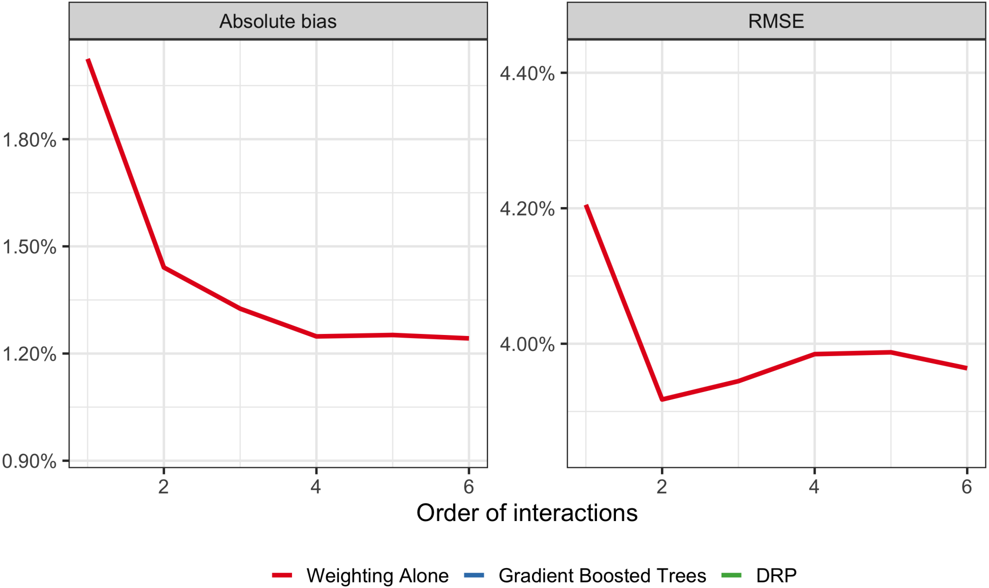



1st order
2nd order
3rd order
4th order
0%
50%
100%
150%
Absolute percent difference
Multilevel Weighting
Post-Stratification (collapsed cells)
Raking on margins

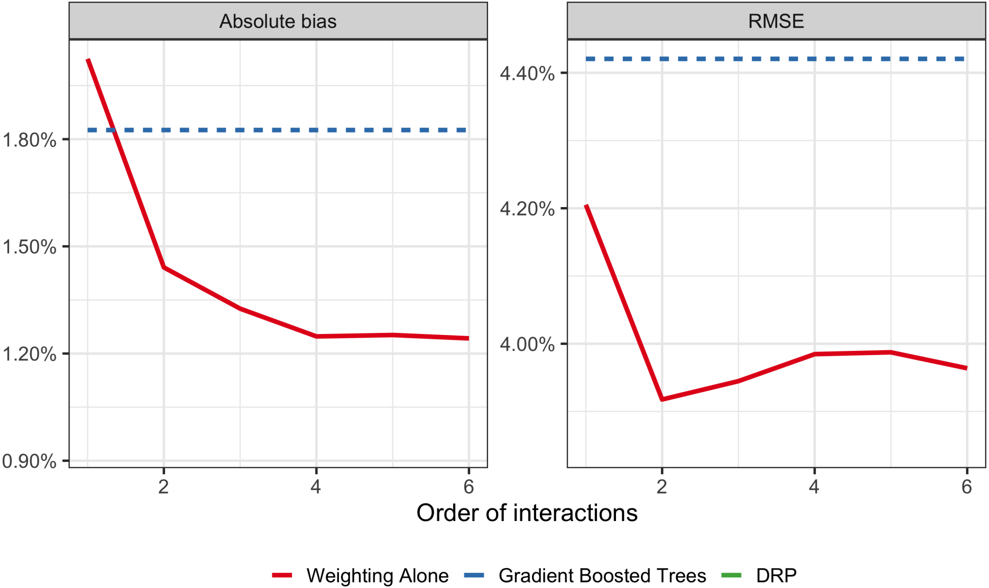

Raking on margins
Post-Stratification (collapsed cells)
Multilevel Weighting
DRP: Ridge
DRP: Gradient Boosting
40.0%
42.5%
45.0%
47.5%
50.0%
Estimated Republican Vote Share

State-level results

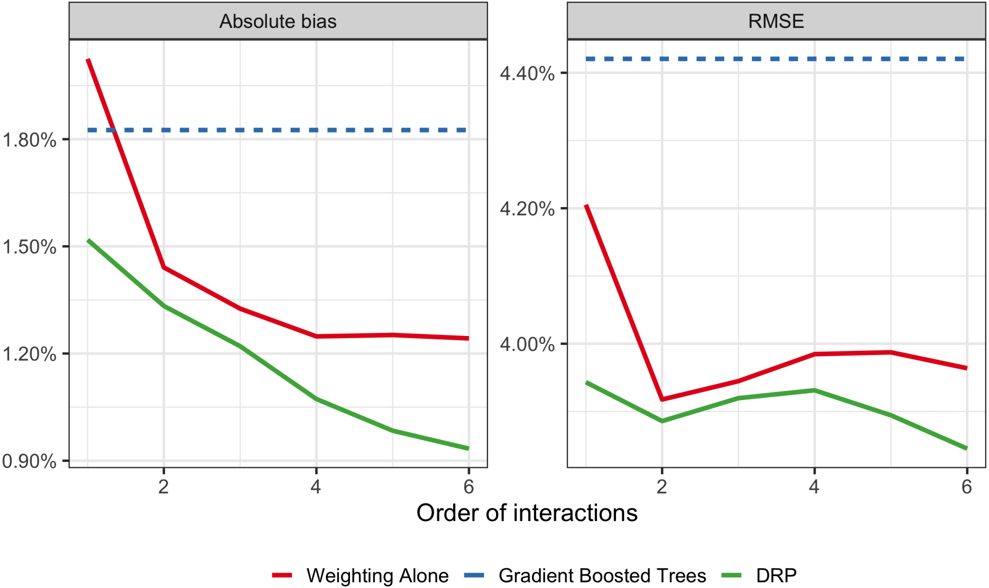
Weighting on higher order interactions reduces bias



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

Recap

Principled, parsimonious way of leveraging the value of interactions

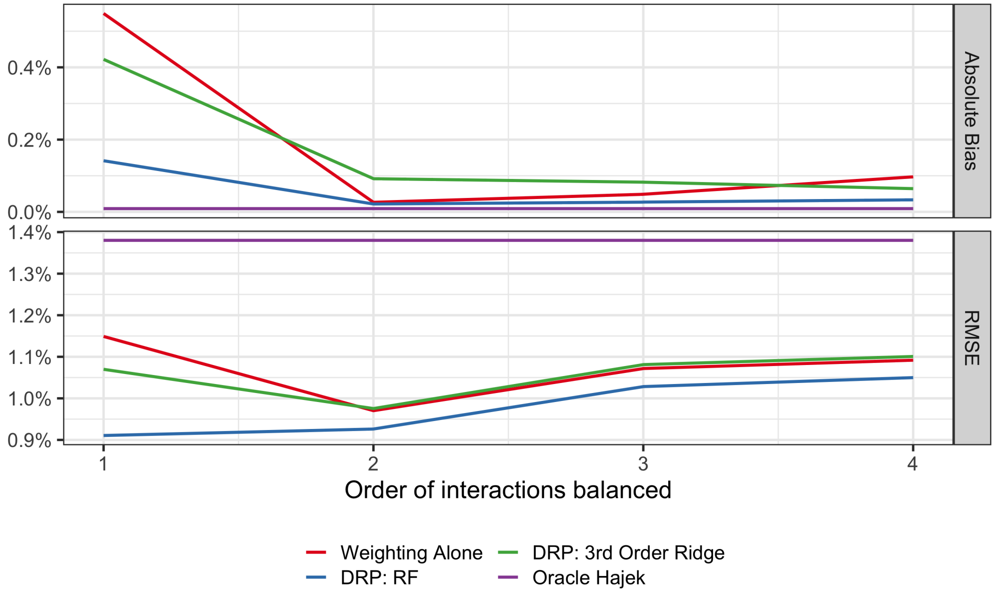
- Multilevel weighting as middle ground between raking and post-stratification
- Dual view as multilevel model for non-response
- DRP adjusts cells where weighting misses the mark

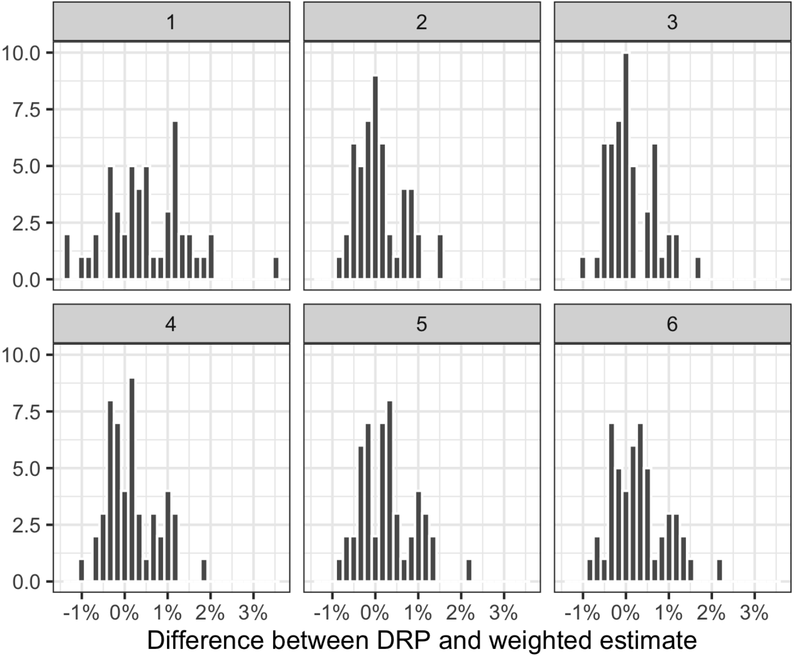
Thank you!

ebenmichael.github.io

Appendix

Actual Republican vote and RF response




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.

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.

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.

References II

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.

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.

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.