Supplementary Materials: Multilevel calibration weighting for survey data

Eli Ben-Michael, Avi Feller, and Erin Hartman

A Additional figures and results

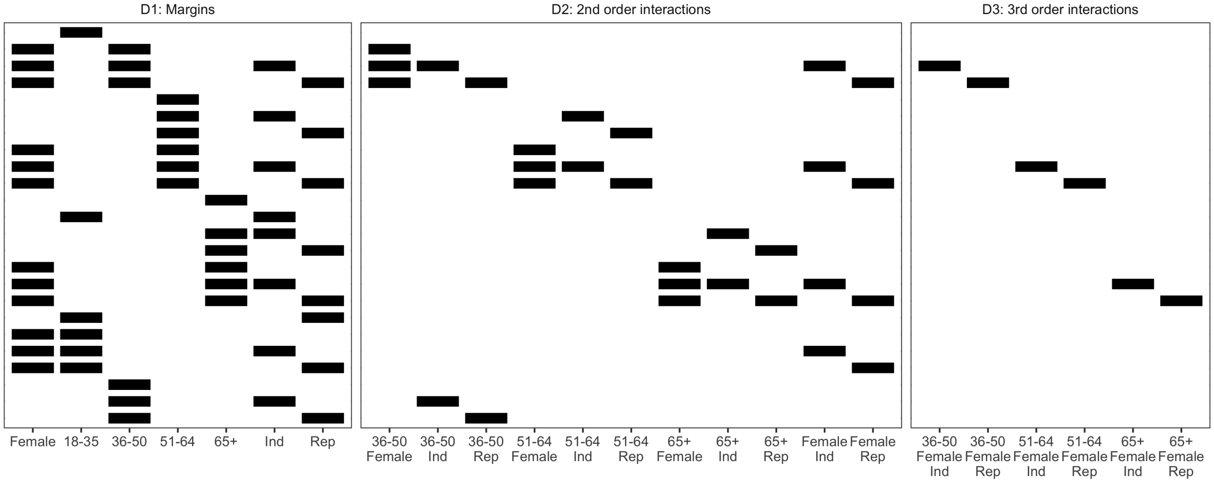

Figure A.1: $\boldsymbol{D}$ with three covariates: a binary for self-reported female, discretized age, and party identification (2, 4, and 3 levels, leading to $J = 24$ distinct cells). The $(1 + 4 + 2 = 7)$ columns of $\boldsymbol{D}^{(1)}$ represent the margins of the three covariates, while the $(3 \times (1 + 2) + 1 \times 2 = 11)$ columns of $\boldsymbol{D}^{(2)}$ represent the 2nd order interaction terms, and the $(24 - 7 - 11 = 6)$ columns of $\boldsymbol{D}^{(3)}$ represent the remaining 3rd order interactions required to uniquely identify the cells. Each row corresponds to a distinct interaction between these three covariates, where the black areas represent elements of this matrix that are equal to 1.

B Comparison to inverse propensity score weighting via multilevel modelling

We now show that the multilevel calibration approach is a form of inverse propensity score weighting with a multilevel non-response model. This connection is instructive, especially for DRP, because traditional propensity score models can have steep data requirements, often requiring detailed individual-level data for both the sample and the target population (see Chen, Li, and C. Wu,

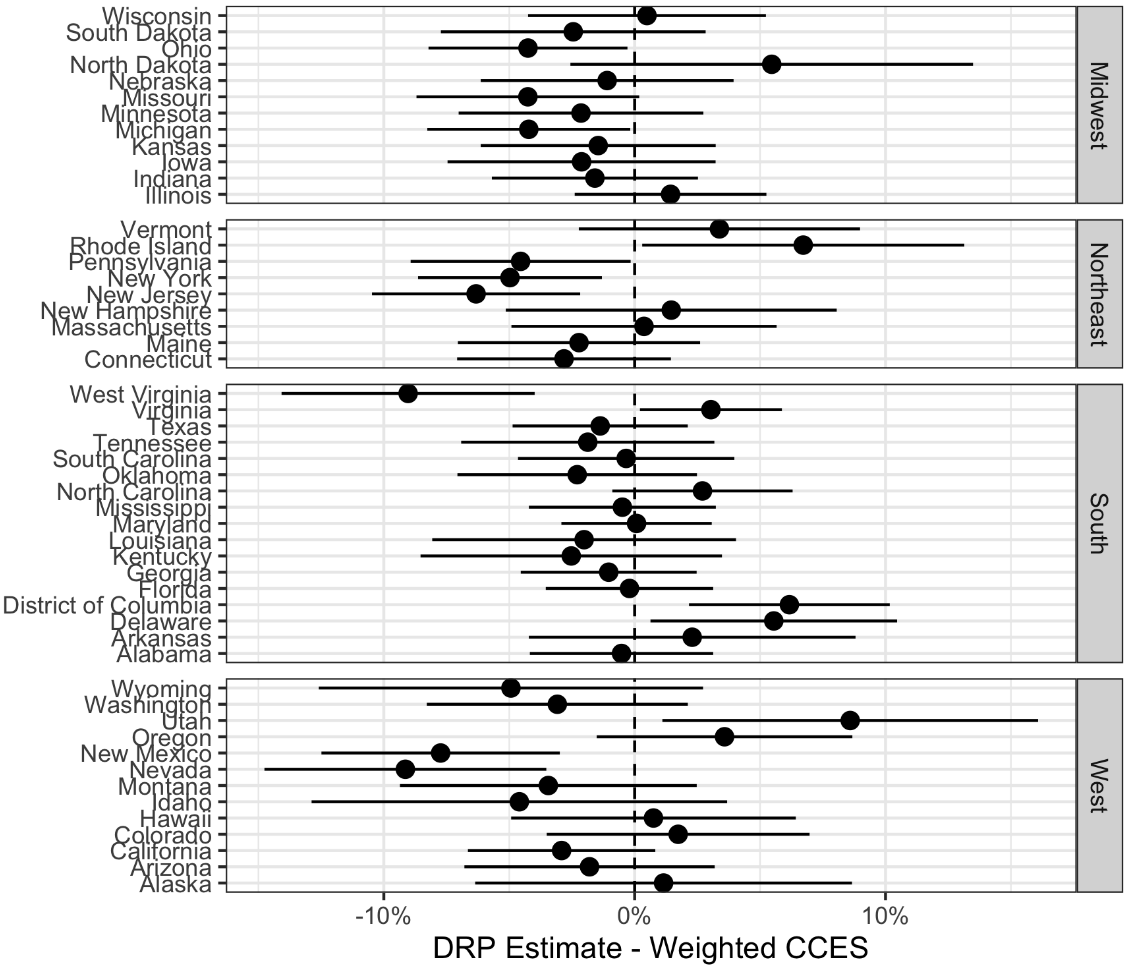

Figure A.2: DRP estimates and approximate 95% confidence intervals of state-level Republican vote share including up to 6$^{\text{th}}$ order interactions and gradient boosted trees, restricting to respondents in the same region.

2020). By contrast, the data requirements for multilevel calibration weights are somewhat weaker, requiring aggregate data on all interactions of interest.

In particular, when we enforce *exact* balance on all interactions, multilevel calibration weights are equivalent to IPW with propensity scores estimated via a fully-saturated generalized linear model (GLM) — and both our proposed weights and traditional IPW weights are equivalent to post-stratification weights. As we show, the primary difference between the multilevel calibration approach and a multilevel GLM is in how the propensity score coefficients are regularized. Through the Lagrangian dual, we will see that the multilevel calibration approach implicitly regularizes the coefficients on interactions to guarantee balance while the multilevel GLM approach does not.

B.1 Dual relation to multilevel non-response modelling

We begin by deriving the Lagrangian dual to the optimization problem (9). By inspecting the dual, we can characterize the implicit propensity score model associated with the weights, moving smoothly between raking on margins and post-stratification. This builds on recent results noting the connection between approximate balancing weights estimators and calibrated regularized propensity score estimation (e.g. Wang and Jose R Zubizarreta, 2020; Hirshberg, Maleki, and J. Zubizarreta, 2019; Zhao, 2019; Chattopadhyay, Christopher H. Hase, and Jose R. Zubizarreta, 2020; Tan, 2020; Ben-Michael, Feller, and Rothstein, 2020) as well as a long history linking raking weights to IPW with a propensity score that is log-linear in the first-order marginals (Little and M. M. Wu, 1991).

The dual problem involves optimizing over a series of Lagrange multipliers. The raking constraint induces one set of Lagrange multipliers $\beta^{(1)}$. In the same way, the approximate post-stratification objective induces an additional set of Lagrange multipliers $\beta^{(k)}$— one for each group of higher order interactions. These dual variables are then chosen to optimize a regularized objective function.[1]

Proposition A.1. If a feasible solution to (9) exists, the Lagrangian dual problem with $L = 0$ and $U = \infty$ is

$$\min_{\beta} \underbrace{\frac{1}{2N}\sum_{i=1}^{n}\left[R_i \max\left\{0, \sum_{k=1}^{d} D_{S_i}^{(k)} \cdot \beta^{(k)}\right\}^2 - \sum_{k=1}^{d} D_{S_i}^{(k)} \cdot \beta^{(k)}\right]}_{\text{loss function } q(\beta)} + \underbrace{\sum_{k=2}^{d} \frac{\lambda_k}{2}\|\beta^{(k)}\|_2^2}_{\text{regularization}}, \tag{A.1}$$

where $\beta = (\beta^{(1)}, \ldots, \beta^{(d)})$. If $\hat{\beta}$ solves (A.1), the primal weights are recovered as

$$\hat{\gamma}(s) = \max\left\{0, \sum_{k=1}^{d} D_{s_i}^{(k)} \cdot \hat{\beta}^{(k)}\right\} \equiv \gamma(s; \hat{\beta}). \tag{A.2}$$

To connect this to propensity score estimation, we can inspect the minimizer of the expected loss, $\mathbb{E}[q(\beta)]$. The zero gradient condition for the expected loss is

$$\nabla \mathbb{E}[q(\beta)] = 0 \iff \sum_{s} N_s^{\mathcal{P}} \pi(s) \gamma(s; \beta) D_s = \sum_{s} N_s^{\mathcal{P}} D_s.$$

The unique weights that solve the expected zero gradient condition are precisely the inverse propensity weights $\gamma(s; \beta) = \frac{1}{\pi(s)}$. Therefore, the dual solution is a *regularized* M-estimator for the propensity score, with a fully saturated propensity score model that includes all interactions.

B.2 The role of regularization

We now compare regularization in multilevel calibration weights versus more traditional multilevel GLM estimation for the propensity score. These two models have the same starting point: both are M-estimators for the propensity score and, in the special case without regularization, both are equivalent to post-stratification weights and to each other. Both estimators also partially pool the

[1] For ease of exposition we have derived the Lagrangian dual for the the usual case where $L = 0$ and $U = \infty$. For general $L < U$, $\gamma(s; \hat{\beta})$ will be truncated at L and U, and the loss function will change slightly.

propensity score estimates across cells. However, in practical settings where full post-stratification is infeasible, regularization affects the two approaches differently. For multilevel calibration, the regularization in the dual problem (A.1) ensures a level of balance on interaction terms. By contrast, for the multilevel GLM, the regularization instead controls a different quantity that is only indirectly relevant for estimating the population average.

To see this, we can examine the zero gradient conditions for the two approaches. First, for multilevel calibration weights, the level of partial pooling directly relates to balance in the higher order interactions. The zero gradient condition for the regularized dual problem (A.1) implies that the imbalance in the k^{th} order interactions is

$$\frac{1}{N}\left\|\sum_s D_s^{(k)} n_s^{\mathcal{R}} \gamma(s;\hat{\beta}) - \sum_s D_s^{(k)} N_s^{\mathcal{P}}\right\|_2 = \lambda_k \|\hat{\beta}^{(k)}\|_2,$$

Therefore λ_k directly controls the level of balance in the k^{th} order interactions, and so the level of regularization controls how far the re-weighted sample is from the target.

We can compare this to the zero gradient condition of the propensity score $\pi(s;\hat{\beta})$ estimated via a multilevel GLM with equivalent hyper-parameters:

$$\frac{1}{N}\left\|\sum_{s=1}^{J} D_s^{(k)} n_s^{\mathcal{R}} - \sum_{s=1}^{J} D_s^{(k)} \pi(s;\hat{\beta}) N_s^{\mathcal{P}}\right\|_2 = \lambda_k \|\hat{\beta}^{(k)}\|_2.$$

Here the hyper-parameter λ_k instead controls the difference between the observed sample counts and the expected counts under the model. This difference is only indirectly related to estimating the population means, in essence estimating the propensity score $\pi(s)$ rather than the inverse propensity score $\frac{1}{\pi(s)}$. Therefore, while both approaches are estimators of a fully-interacted propensity score, the regularization in multilevel calibration controls an upper bound on the bias when estimating the population average μ. In contrast, regularization in the multilevel GLM provides a condition on a quantity that is incidental to estimating μ.

C Asymptotic normality

We consider an asymptotic framework with a sequence of finite populations of size N, and let $N \to \infty$. In this framework, we make several modifications to our setup. First, we strengthen Assumption 2 to hold strictly for all population sizes N, so that $\min_s \pi(s) \geq \pi^* > 0$, and we allow the lower bound π^* to change with the population size N. This ensures that we have a strictly non-zero probability of having a respondent in each cell in all the populations we consider. We then allow the number of cells J to grow with the population size N. Denoting $\kappa \equiv \|D^{-1}\|_2 \|D\|_2$ as the condition number of the $J \times J$ matrix D, we restrict the number of cells so that $\frac{\kappa^2 J}{(\pi^* N)^\alpha}$ converges to a constant for a rate $0 \leq \alpha < 1$. We also adjust the multilevel calibration procedure to approximately (rather than exactly) rake on margins without regularization, ensuring that there is always a feasible solution for every finite population. Finally, we restrict the response variables R_i to be independent. We detail these and other regularity assumptions on the design in the following assumption.

Assumption A.1. There is a sequence of populations of size N with $N \to \infty$ such that

(a) The condition number of D, $\kappa \equiv \|D^{-1}\|_2 \|D\|_2$, and the number of cells J satisfy $\frac{\kappa^2 J}{(\pi^* N)^\alpha} \to c$ for some constant c and for an $0 \leq \alpha < 1$

(b) The response variables R_i are independent.

(c) $\pi(s) \geq \pi^* > 0$ for all N, where $\frac{1}{\pi^{*2} N} \to 0$ as $N \to \infty$.

(d) The residuals $\varepsilon_i \equiv Y_i - \mu_{S_i}$ satisfy $\frac{1}{N}\sum_{i=1}^N \varepsilon_i^2 < \infty$ for all population sizes N.

(e) The maximum variance across cells conditional on the cell counts, $\sigma^2 \equiv \max_s \sigma_s^2 = \max_s \operatorname{Var}\left(\bar{\varepsilon}_s \mid n_s^{\mathcal{R}}\right)$ is $o_p\left((\pi^* N)^{-\frac{\alpha}{2}}\right)$.

(f) For a random variable $Z = o_p\left(\frac{1}{\pi^* \sqrt{N}}\right)$, the variance of the oracle estimator $V = \frac{1}{N^2}\sum_i \frac{\pi_i(1-\pi_i)}{\pi(S_i)}\varepsilon_i^2$ satisfies $\frac{Z}{\sqrt{V}} = o_p(1)$.

(g) We find $\hat{\gamma}$ via the modified problem

$$\begin{aligned} \min_{\gamma \in \mathbb{R}^J} \; & \sum_{k=1}^{d} \left\| \sum_s D_s^{(k)} n_s^{\mathcal{R}} \gamma(s) - D_s^{(k)} N_s^{\mathcal{P}} \right\|_2^2 \\ \text{subject to } \; & 0 \leq \gamma(s) \leq 1 \quad \forall s = 1, \ldots J. \end{aligned} \tag{A.3}$$

Theorem A.1. If $\frac{\kappa^2 J}{(\pi^* N)^\alpha}$ converges to a constant for some $0 \leq \alpha < 1$, and $\sum_s (\hat{\mu}_s - \mu_s)^2 = o_p\left((\pi^* N)^{-\alpha/2}\right)$, then under the regularity conditions in Assumption A.1, the DRP estimator $\hat{\mu}^{\text{drp}}(\hat{\gamma})$ is

$$\hat{\mu}^{\text{drp}}(\hat{\gamma}) - \mu = \frac{1}{N}\sum_{i=1}^N \frac{R_i}{\pi(S_i)}\varepsilon_i + o_p\left(\frac{1}{\pi^* \sqrt{N}}\right).$$

Furthermore, if $\frac{1}{N\sqrt{V_N}}\sum_{i=1}^N \frac{R_i}{\pi(S_i)}\varepsilon_i \Rightarrow N(0,1)$ for $V_N = \operatorname{Var}\left(\frac{1}{N}\sum_{i=1}^N \frac{R_i}{\pi(S_i)}\varepsilon_i\right)$, then

$$\frac{\hat{\mu}^{\text{drp}}(\hat{\gamma}) - \mu}{\sqrt{V_N}} \Rightarrow N(0,1).$$

Theorem A.1 shows us that as long as the modelled cell averages estimate the true cell averages well enough, the model and the calibration weights combine to ensure that the bias will be negligible relative to the variance, asymptotically. The rate at which the number of cells grows with the population size affects how well the modelled cell averages need to perform. If the number of cells is constant, the model needs only to be consistent. On the other hand, if the number of cells grows quickly then Theorem A.1 implicitly requires more structure on the outcomes, so that the model can estimate the cell averages well enough. As we discuss in Section 3.1, we expect main effects to be much stronger than higher order interaction terms in practice. Therefore, including a new covariate or using a finer discretization of continuous covariates will primarily impact the outcome model through these main effects, leading to a substantial amount of underlying structure even though the total number of cells is increasing. As we discuss in Section 6, there are alternative ways to account for an increasing number of cells J that may allow J to grow more quickly relative to the population size N. We leave a thorough investigation of these alternatives to future work.

Note that the minimum cell response probability π^* affects the quality of the asymptotic approximation; if individuals in some cells are very unlikely to respond, it will be difficult both to model those averages and achieve good balance from the respondents. Theorem A.1 is analogous to recent double-robustness results in survey estimation, such as from Chen, Li, and C. Wu (2020). Rather than estimating a parametric outcome and non-response model, we instead consider all interactions.

D Proofs and derivations

Lemma A.1. Let $\kappa \equiv \|D^{-1}\|_2 \|D\|_2$ be the ratio of the maximum and minimum singular values of D. The solution to (A.3) satisfies

$$\sqrt{\sum_s \left(n_s^{\mathcal{R}} \hat{\gamma}(s) - N_s^{\mathcal{P}}\right)^2} \leq \kappa \sqrt{\sum_s \left(\frac{n_s^{\mathcal{R}}}{\pi(s)} - N_s^{\mathcal{P}}\right)^2}$$

Proof of Lemma A.1. Slightly abusing notation, denote $\frac{1}{\pi} \in (0,1)^J$ as the vector of inverse response probabilities for each cell. $\frac{1}{\pi}$ is feasible for optimization problem (A.3), and so

$$\begin{aligned}
\frac{1}{\|D^{-1}\|_2} \|\text{diag}(n^{\mathcal{R}})\hat{\gamma} - N^{\mathcal{P}}\|_2 &\leq \|D'(\text{diag}(n^{\mathcal{R}})\hat{\gamma} - N^{\mathcal{P}})\|_2 \\
&\leq \left\|D'\left(\text{diag}(n^{\mathcal{R}})\frac{1}{\pi} - N^{\mathcal{P}}\right)\right\|_2 \\
&\leq \|D\|_2 \left\|\text{diag}(n^{\mathcal{R}})\frac{1}{\pi} - N^{\mathcal{P}}\right\|_2
\end{aligned}$$

Multiplying by $\|D^{-1}\|_2$ gives the result. □

Lemma A.2. Let $\pi^* = \min_s \pi(s)$. For any $\delta > 0$,

$$\frac{1}{N}\sqrt{\sum_s \left(N_s^{\mathcal{P}} - \frac{n_s^{\mathcal{R}}}{\pi(s)}\right)^2} \leq \frac{1}{\pi^* \sqrt{N}}\left(\sqrt{J \log 5} + \delta\right),$$

with probability at least $1 - \exp\left(-2\pi^{*2} N \delta^2\right)$.

Proof of Lemma A.2. Since $R_i \in \{0,1\}$ is bounded, it is sub-Guassian with scale parameter $\frac{1}{2}$. and because they are independent, $\frac{N_s^{\mathcal{P}}}{N} - \frac{n_s^{\mathcal{R}}}{N\pi(s)} = N_s^{\mathcal{P}} - \frac{1}{\pi(s)} \sum_{S_i = s} R_i$ is a mean-zero sub-Gaussian random variable with scale parameter $\frac{\sqrt{N_s^{\mathcal{P}}}}{2\pi(s)N} \leq \frac{1}{2\pi^*\sqrt{N}}$. Now by a discretization argument (Wainwright, 2019, § 9.6), we have that

$$\frac{1}{N}\sqrt{\sum_s \left(N_s^{\mathcal{P}} - \frac{n_s^{\mathcal{R}}}{\pi(s)}\right)^2} \geq \frac{1}{\pi^* \sqrt{N}}\left(\sqrt{J \log 5} + \delta\right)$$

with probability at most $\exp\left(-2\pi^{*2} N \delta^2\right)$. This completes the proof. □

Lemma A.3. If $\sqrt{\sum_s \left(\hat{\mu}_s - \mu_s\right)^2} = o_p\left((\pi^* N)^{-\frac{\alpha}{2}}\right)$, then

$$\frac{1}{N}\sum_s \left(\hat{\mu}_s - \mu_s\right)\left(n_s^{\mathcal{R}}\hat{\gamma}(s) - N_s^{\mathcal{P}}\right) = o_p\left(\frac{1}{\pi^*\sqrt{N}}\right)$$

Proof of Lemma A.3. First, note that by Cauchy-Schwartz,

$$\frac{1}{N}\sum_s \left(\hat{\mu}_s - \mu_s\right)\left(n_s^{\mathcal{R}}\hat{\gamma}(s) - n_s^{\mathcal{P}}\right) \leq \sqrt{\sum_s \left(\hat{\mu}_s - \mu_s\right)^2}\frac{1}{N}\sqrt{\sum_s \left(n_s^{\mathcal{R}}\hat{\gamma}(s) - N_s^{\mathcal{P}}\right)^2}$$

From Lemma A.2, the term on the right is $O_p\left(\frac{\kappa}{\pi}\sqrt{\frac{J}{N}}\right)$. By Assumption A.1a, this is $O_p\left(\frac{1}{\pi^{1-\alpha/2}N^{1/2-\alpha/2}}\right)$. Now since $\sqrt{\sum_s \left(\hat{\mu}_s - \mu_s\right)^2} = o_p\left((\pi^* N)^{-\frac{\alpha}{2}}\right)$, the product is $o_p\left(\frac{1}{\pi^*\sqrt{N}}\right)$ □

Lemma A.4. Under Assumption A.1e, the solution to (A.3), $\hat{\gamma}$ satisfies

$$\frac{1}{N}\sum_s \hat{\gamma}(s) n_s^{\mathcal{R}}\bar{\varepsilon}_s = \frac{1}{n}\sum_{i=1}^{N}\frac{R_i}{\pi(S_i)}\varepsilon_i + o_p\left(\frac{1}{\pi^*\sqrt{N}}\right)$$

Proof of Lemma A.4. First, we write the noise term as

$$\frac{1}{N}\sum_s \hat{\gamma}(s) n_s^{\mathcal{R}}\bar{\varepsilon}_s = \frac{1}{n}\sum_{i=1}^{N}\frac{R_i}{\pi(S_i)}\varepsilon_i + \frac{1}{n}\sum_{i=1}^{N} R_i\left(\hat{\gamma}(s) - \frac{1}{\pi(S_i)}\right)\varepsilon_i.$$

The variance of the second term, conditional on the cell counts $n^{\mathcal{R}}$ is

$$\begin{aligned}\text{Var}\left(\frac{1}{n}\sum_{i=1}^{N} R_i\left(\hat{\gamma}(s) - \frac{1}{\pi(S_i)}\right)\varepsilon_i \mid n^{\mathcal{R}}\right) &= \frac{1}{N^2}\sum_s\left(\hat{\gamma}(s) - \frac{1}{\pi(s)}\right)^2 n_s^{\mathcal{R}2}\sigma_s^2 \\ &\leq \frac{\sigma^2}{N^2}\sum_s\left(\hat{\gamma}(s) - \frac{1}{\pi(s)}\right)^2 n_s^{\mathcal{R}2}\end{aligned}$$

So by Chebyshev's inequality, conditional on the cell counts $n^{\mathcal{R}}$ we have that with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{N} R_i\left(\hat{\gamma}(s) - \frac{1}{\pi(S_i)}\right)\varepsilon_i\right| \leq \frac{\sigma}{N\sqrt{\delta}}\sqrt{\sum_s\left(\hat{\gamma}(s) - \frac{1}{\pi(s)}\right)^2 n_s^{\mathcal{R}2}}.$$

Now notice that

$$\begin{aligned}\sqrt{\sum_s\left(\hat{\gamma}(s) - \frac{1}{\pi(s)}\right)^2 n_s^{\mathcal{R}2}} &= \left\|\text{diag}(n^{\mathcal{R}})\left(\hat{\gamma} - \frac{1}{\pi}\right)\right\|_2 \\ &= \left\|\text{diag}(n^{\mathcal{R}})\hat{\gamma} - N^{\mathcal{P}} + N^{\mathcal{P}} - \text{diag}(n^{\mathcal{R}})\frac{1}{\pi}\right\|_2 \\ &\leq \left\|\text{diag}(n^{\mathcal{R}})\hat{\gamma} - N^{\mathcal{P}}\right\|_2 + \left\|N^{\mathcal{P}} - \text{diag}(n^{\mathcal{R}})\frac{1}{\pi}\right\|_2\end{aligned}$$

From Lemma A.1 we can further bound this by

$$\sqrt{\sum_s \left(\hat{\gamma}(s) - \frac{1}{\pi(s)}\right)^2 n_s^{\mathcal{R}2}} \leq (1+\kappa)\left\|N^{\mathcal{P}} - \text{diag}(n^{\mathcal{R}})\frac{1}{\pi}\right\|_2$$

Following the Proof of Lemma A.3, by Lemma A.2 and Assumption A.1a this is $O_p\left(\frac{1}{\pi^{1-\alpha/2}N^{1/2-\alpha/2}}\right)$. Noting that by Assumption A.1e $\sigma = o_p\left((\pi^* N)^{-\frac{\alpha}{2}}\right)$, shows that this remainder term is $o_p\left(\frac{1}{\pi^*\sqrt{N}}\right)$. □

Proof of Theorem A.1. First, we write $\hat{\mu}^{\text{drp}}(\hat{\gamma}) - \mu$ as

$$\hat{\mu}^{\text{drp}}(\hat{\gamma}) - \mu = \frac{1}{N}\sum_s \left(\hat{\mu}_s - \mu_s\right)\left(n_s^{\mathcal{R}}\hat{\gamma}(s) - N_s^{\mathcal{P}}\right) + \frac{1}{N}\sum_{i=1}^N R_i\hat{\gamma}(S_i)\varepsilon_i$$

From Lemma A.3, the first term is $o_p\left(\frac{1}{\pi^*\sqrt{N}}\right)$ and from Lemma A.4 the second term is $\frac{1}{N}\sum_i \frac{R_i}{\pi(S_i)}\varepsilon_i +$ $o_p\left(\frac{1}{\pi^*\sqrt{N}}\right)$. Combining these gives the first result. Assumption A.1f combined with an application of Slutsky's theorem gives the second result. □

Proof of Proposition A.1. We begin by re-writing the optimization problem (9) with $L = 0$ and $U = \infty$ in terms of auxiliary covariates $\mathcal{E}^{(k)} \equiv \sum_s D_s^{(k)} n_s^{\mathcal{R}}\gamma(s) - D_s^{(k)}N_s^{\mathcal{P}}$. The optimization problem becomes

$$\begin{aligned}
\min_{\gamma\in\mathbb{R}^J} \quad & \sum_{k=2}^d \frac{1}{2\lambda_k}\left\|\mathcal{E}^{(k)}\right\|_2^2 + \frac{1}{2}\sum_s n_s^{\mathcal{R}}\gamma(s)^2 \\
\text{subject to} \quad & \sum_s D_s^{(1)} n_s^{\mathcal{R}}\gamma(s) = \sum_s D_s^{(1)}N_s^{\mathcal{P}} \\
& \sum_s D_s^{(k)} n_s^{\mathcal{R}}\gamma(s) - D_s^{(k)}N_s^{\mathcal{P}} - \mathcal{E}^{(k)} = 0 \\
& 0 \leq \gamma(s) \quad \forall s = 1, \ldots J.
\end{aligned}$$

The Lagrangian is

$$\mathcal{L}(\gamma, \mathcal{E}, \beta) \equiv \sum_{s=1}^J \frac{1}{2}n_s^{\mathcal{R}}\gamma(s)^2 - n_s^{\mathcal{R}}\gamma(s)\sum_{k=1}^d D_s^{(k)}\cdot\beta^{(k)} + N_s^{\mathcal{P}}\sum_{k=1}^d D_s^{(k)}\cdot\beta^{(k)} + \sum_{k=2}^d \frac{1}{2\lambda_k}\|\mathcal{E}^{(k)}\|_2^2 - \mathcal{E}^{(k)}\cdot\beta^{(k)}$$

The dual problem maximizes the Lagrangian over the domain of γ and $\mathcal{E}$, so

$$\begin{aligned}
q(\beta) &= -\min_{0\leq\gamma(s),\mathcal{E}} \mathcal{L}(\gamma,\mathcal{E},\beta) \\
&= \sum_{s=1}^J n_s\mathcal{R} \min_{0\leq\gamma(s)}\left\{\frac{1}{2}\gamma(s)^2 - \gamma(s)\sum_{k=1}^d D_s^{(k)}\cdot\beta^{(k)}\right\} + N_s^{\mathcal{P}}\sum_{k=1}^d D_s^{(k)}\cdot\beta^{(k)} + \sum_{k=2}^d \min_{\mathcal{E}^{(k)}}\left\{\frac{1}{2\lambda_k}\|\mathcal{E}^{(k)}\|_2^2 - \mathcal{E}^{(k)}\cdot\beta^{(k)}\right\} \\
&= \frac{1}{2}\sum_{s=1}^J n_s^{\mathcal{R}} \max\left\{0, \sum_{k=1}^d D_{S_i}^{(k)}\cdot\beta^{(k)}\right\}^2 - N_s^{\mathcal{P}}\sum_{k=1}^d D_s^{(k)}\cdot\beta^{(k)} + \sum_{k=2}^d \frac{\lambda_k}{2}\|\beta^{(k)}\|_2^2
\end{aligned}$$

Since there exists a feasible solution to (9) by assumption, by Slater's condition $\min_\beta q(\beta)$ is equivalent to the solution to the primal problem. The solution to the inner minimization shows that the primal and dual variables are related by $\hat{\gamma}(s) = \max\left\{0, \sum_{k=1}^d D_s^{(k)} \cdot \hat{\beta}^{(k)}\right\}$.

□

E Simulation study calibrated to the 2016 U.S. presidential election

We now evaluate the statistical behavior of the multilevel calibration and DRP estimators on simulated data based on our application to the 2016 United States Presidential election, described in Section 1.1. We calibrate two non-response models to the response structure in this population. First, we fit a random forest model to predict response (i.e., inclusion in the Pew sample) with $B = 500$ trees, so that the probability of responding in cell s is

$$\pi^{\text{rf}}(s) = \sum_{s'} \frac{n_{s'}^{\mathcal{R}}}{B} \sum_{b=1}^{B} \frac{\mathbb{1}\{s' \in L_b(s)\}}{|L_b(s)|}.$$

We also consider a fourth-order model, where the response probability for cell s is

$$\pi^{(4)}(s) = \text{logit}^{-1}\left(\sum_{k=1}^{4} \hat{\beta}^{(k)} \cdot D_s^{(k)}\right),$$

and the coefficient vector $\hat{\beta}$ is under-regularized so that there is poor overlap. We similarly consider two different outcome models for presidential candidate vote choice. First, we fix the outcomes to be unchanged from the original data; second, we model the probability that unit i votes Republican ($Y_i = 1$) as a fourth order logistic regression model as above, similarly under-regularized.

To generate simulation runs, we re-sample from the population with replacement, so the total number of units N is fixed while the number of units within each cell $N_s^{\mathcal{P}}$ varies. We then generate responses and outcomes according to the probabilities above using two pairs: (a) fourth order models for both the response and the outcome, and (b) a random forest response model with the true, deterministic outcomes. We consider using multilevel calibration weighting in Equation (9), balancing first, second, third, and fourth order interactions with $\lambda^{(k)} = 1$ and setting $\lambda^{(k)} = 0$ for interactions of higher order. We also consider the DRP estimator, bias correcting with either a third-order ridge regression or a random forest, as well as MRP with these outcome models. Finally, we compare to the oracle Horvitz-Thompson estimator with the true response probabilities.

Figure E.1 shows the bias and root mean square error (RMSE) of these approaches across simulation runs. First looking at the bias, we see that under both data generating processes (DGPs) it is not enough to rake on margins, and there are substantial gains to balancing second and higher order interactions. Next, bias correction can provide large improvements: under both DGPs, DRP reduces the bias relative to raking on the margins alone by nearly the same degree as directly balancing higher order interaction terms. Even in the under-regularized fourth order DGP—where the oracle Horvitz-Thompson estimator performs poorly—we can significantly reduce the bias. DRP also has reduced bias relative to MRP alone with the same outcome model. Focusing on RMSE, we see that the decrease in bias from balancing higher order interactions outweighs the increase in variance only when balancing second order interactions, with third and fourth order

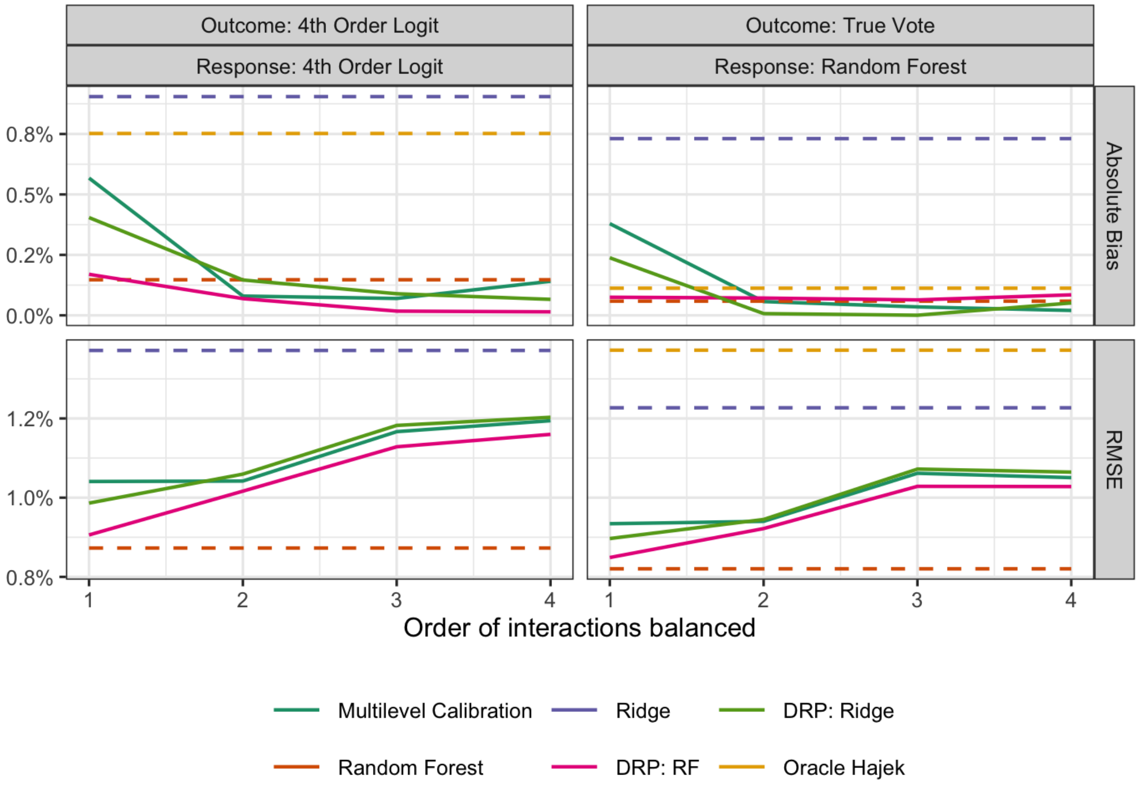

Figure E.1: Bias and RMSE across 1000 simulation runs. RMSE for the oracle Horvitz-Thompson estimator in the under-regularized fourth order model (6%) omitted for scale.

interactions having a worse bias-variance trade-off. We see however, that the bias-variance trade-off for including an outcome model through DRP is favorable under both outcome models and DGPs, with the DRP estimator with a random forest outcome model and raking weights having the lowest RMSE. Finally, MRP with ridge regression has higher RMSE than multilevel calibration and DRP, while MRP with random forest (the oracle estimator for one of the DGPs) has lower RMSE. Finally, Figure E.2 shows the empirical coverage for 95% confidence intervals constructed as $\hat{\mu}(\hat{\gamma}) \pm z_{1-0.025}\sqrt{\hat{V}}$ for multilevel calibration and DRP with both bias correction methods. The intervals for multilevel calibration are fairly conservative, while the intervals for both bias-correction approaches achieve close to nominal coverage.

References

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein (2020). “Variation in impacts of letters of recommendation on college admissions decisions: Approximate balancing weights for treatment effect heterogeneity in observational studies”. In: arXiv: arXiv:2008.04394v1.

Chattopadhyay, Ambarish, Christopher H. Hase, and Jose R. Zubizarreta (2020). “Balancing Versus Modeling Approaches to Weighting in Practice”. In: *Statistics in Medicine* in press. DOI: 10.1002/sim.0000.

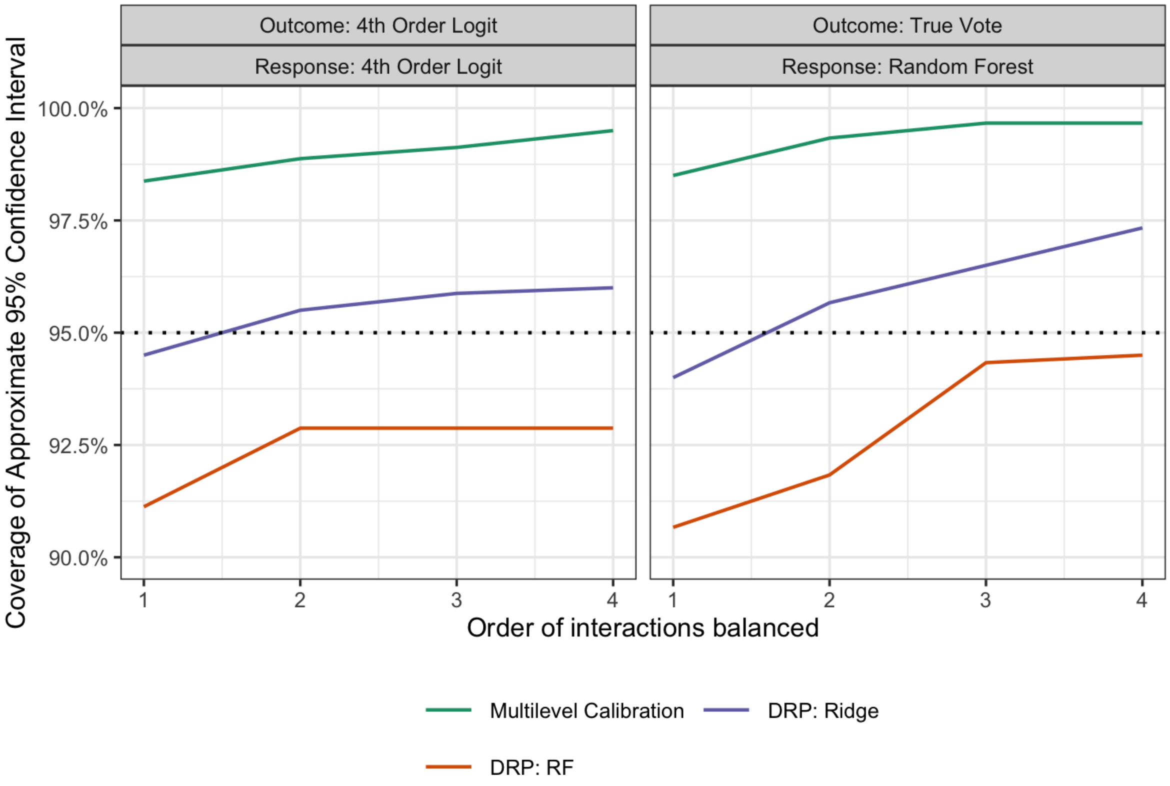

Figure E.2: Empirical coverage rate of approximate 95% confidence intervals over 1000 simulation runs for multilevel calibration alone and DRP with a 3$^{\text{rd}}$ order ridge regression and a random forest outcome model.

Chen, Yilin, Pengfei Li, and Changbao Wu (2020). “Doubly robust inference with nonprobability survey samples”. In: *Journal of the American Statistical Association* 115.532, pp. 2011–2021.

Hirshberg, David A, Arian Maleki, and José Zubizarreta (2019). “Minimax Linear Estimation of the Retargeted Mean”. arXiv: 1901.10296v1. URL: https://arxiv.org/pdf/1901.10296.pdf.

Little, Roderick J.A. and Mei Miau Wu (1991). “Models for contingency tables with known margins when target and sampled populations differ”. In: *Journal of the American Statistical Association* 86.413, pp. 87–95. ISSN: 1537274X. DOI: 10.1080/01621459.1991.10475007.

Tan, Z. (2020). “Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data”. In: *Biometrika* 107.1, pp. 137–158. ISSN: 14643510. DOI: 10.1093/biomet/asz059.

Wainwright, Martin J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. DOI: 10.1017/9781108627771.

Wang, Yixin and Jose R Zubizarreta (2020). “Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations”. In: *Biometrika* 107.1, pp. 93–105. ISSN: 14643510. DOI: 10.1093/biomet/asz050. arXiv: 1705.00998.

Zhao, Qingyuan (2019). "Covariate balancing propensity score by tailored loss functions". In: *Annals of Statistics* 47.2, pp. 965–993. ISSN: 00905364. DOI: 10.1214/18-AOS1698. arXiv: 1601.05890.