Bayesian hierarchical weighting adjustment and survey inference

Yajuan Si, Rob Trangucci, Jonah Sol Gabry and Andrew Gelman[1]

Abstract

We combine weighting and Bayesian prediction in a unified approach to survey inference. The general principles of Bayesian analysis imply that models for survey outcomes should be conditional on all variables that affect the probability of inclusion. We incorporate all the variables that are used in the weighting adjustment under the framework of multilevel regression and poststratification, as a byproduct generating model-based weights after smoothing. We improve small area estimation by dealing with different complex issues caused by real-life applications to obtain robust inference at finer levels for subdomains of interest. We investigate deep interactions and introduce structured prior distributions for smoothing and stability of estimates. The computation is done via Stan and is implemented in the open-source R package `rstanarm` and available for public use. We evaluate the design-based properties of the Bayesian procedure. Simulation studies illustrate how the model-based prediction and weighting inference can outperform classical weighting. We apply the method to the New York Longitudinal Study of Wellbeing. The new approach generates smoothed weights and increases efficiency for robust finite population inference, especially for subsets of the population.

Key Words: Weighting; Prediction; Multilevel regression and poststratification; Structured prior.

1 Introduction

1.1 Background

Design-based and model-based approaches have long been contrasted in survey research (Little, 2004). The former automatically takes into account survey design, while the latter can yield robust inference for small sample estimation. Rao (2011) provides an appraisal of frequentist and Bayesian methods on survey sampling practice. Classical design-based approaches use weights to adjust the sample to the population; see Chen, Elliott, Haziza, Yang, Ghosh, Little, Sedransk and Thompson (2017) for a review of various weighted estimators for a population mean. However, classical survey weighting usually relies on many user-defined choices so that the process of weighting can be difficult to codify in real-world surveys (Gelman, 2007). The Bayesian approach for finite population inference (Ghosh and Meeden, 1997) allows prior information to be incorporated, when appropriate, but is subject to model misspecification.

In the present paper we combine Bayesian prediction and weighting in a unified approach to survey inference, applying scalable and robust Bayesian regression models to account for complex design features under the framework of multilevel regression and poststratification (MRP, Gelman and Little (1997); Park, Gelman and Bafumi (2005); Ghitza and Gelman (2013); Si, Pillai and Gelman (2015)). MRP adjusts for complex design and response mechanisms and improves small area estimation (Fay and Herriot, 1979; Rao and Molina, 2015). We deal with different complex issues caused by real-life applications and much finer levels for subdomain inference of interest. Our method yields efficient and valid finite population inference, especially for subgroups, and constructs model-based weights after smoothing.

1. Yajuan Si, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor. E-mail: yajuan@umich.edu; Rob Trangucci, Department of Statistics, University of Michigan; Jonah Sol Gabry, Department of Statistics, Columbia University; Andrew Gelman, Departments of Statistics and Political Science, Columbia University.

The contributions of this paper are two folded: 1) as innovative Bayesian methodology developments we develop a new structured prior setting to handle high-order interaction terms; and 2) to improve survey research and operation, we combine Bayesian prediction and weighting as a unified approach to survey inference, accounting for design features in the Bayesian modeling. We generalize MRP for finite population inferences and construct stable and calibrated model-based weights to solve the problems of classical weights. We disseminate the R package `rstanarm` implementing the proposed methods for public use, promoting the model-based approaches in survey research and operational practice. More importantly, the paper builds the groundwork to use MRP in the survey weighting adjustment and data integration, for example, to make inferences with nonprobability surveys. Our proposed methods offer one important and practical tool for designing and weighting survey samples (Valliant, Dever and Kreuter, 2018).

1.2 Framework

For a finite population of N units, we denote the variable of interest as $y = (y_1, \ldots, y_N)$ and the inclusion indicator variable as $I = (I_1, \ldots, I_N)$, where $I_i = 1$ if unit i is included in the sample and $I_i = 0$ otherwise. Here, inclusion refers to selection and response. The general inference framework considers the joint distribution for I and y. Design-based inference considers the distribution of I and treats y as fixed. Under probability sampling, model-based inferences can be based on the distribution of y alone given the variables that affect the inclusion mechanisms are included in the model (Royall, 1968), that is, under the ignorable inclusion mechanism when the distribution of I given y is independent of the distribution of y (Rubin, 1976, 1983).

To account for the factors that affect inclusion, classical design weights adjust for unequal probabilities of sampling, with subsequent weighting adjustments accounting for coverage problems and nonresponse during data collection or data cleaning. Classical weights are thus generated as a product of multiple adjustment factors: inverse probability of selection, inverse propensity score of response, and poststratification (also called calibration or benchmarking; Holt and Smith (1979)). Each of these adjustments can be approximate when the probability of selection, the probability of response, or population totals are estimated from data. Beyond any approximation issues, even if the inclusion model is known exactly, extreme values of weights will cause high variability and then inferential problems, especially when the weights are weakly correlated with the survey outcome variable (Rao, 1966a, b; Hájek, 1971; Särndal, Swensson and Wretman, 1992). When the weighting process involves poststratification or nonresponse adjustment – where the weights themselves are random variables – the variance estimation will be different from the cases only with fixed design weights. It is nontrivial to analytically derive a variance estimator under the multi-stage weighting adjustment or complex sampling design.

In practice, the construction of survey weights requires somewhat arbitrary decisions of the selection of variables and interactions, pooling of weighting cells, and weight trimming. It can be unclear whether and how to incorporate auxiliary information (Groves and Couper, 1995). Discussion of smoothing and trimming in the survey weighting literature (e.g., Potter, 1988, 1990; Elliott and Little, 2000; Elliott, 2007;

Xia and Elliott, 2016) has focused on estimating the finite population total or mean, with less attention to subdomain estimates. Beaumont (2008) proposes to regress weights on the survey variables and use the predicted values as smoothed weights, where the direction is inspiring but tangential to the inference objective where good inference properties are desired for the survey variable of interest rather than the weights. Borrowing information on survey outcomes potentially increases efficiency and calls for a general framework.

Gelman (2007) recommends regression models including as covariates any variables that affect selection and response, including stratification variables, clusters, and auxiliary information. Any of these approaches can be sensitive to the prior specification for stable estimation; this is the model-based counterpart to the decisions required for smoothing or trimming classical survey weights. Flexible prediction techniques, such as spline functions, penalized regression and tree-based models, have been proposed to accommodate model-assisted survey estimation (Särndal et al., 1992; Wu and Sitter, 2001; Breidt and Opsomer, 2017; McConville and Toth, 2018).

Model-based and model-assisted weighting adjustment methods for finite population total estimation have been compared by Henry and Valliant (2012). The model-based weighting methods in the superpopulation perspective (Valliant, Dorfman and Royall, 2000) use predictions from regression models to derive case weights, where the predictions are based on hierarchical linear regression models with various bias corrections (Chambers, Dorfman and Wehrly, 1993; Firth and Bennett, 1998). Based on the finite population total estimation, model-assisted methods derive case weights mainly from calibration on benchmark variables (Kott, 2009) via the generalized regression estimator (GREG, Deville and Särndal (1992)). However, the case weights derived from regression predictions can be highly variable and even negative and may damage some domain estimates. Model-based approaches play a vital role in small area estimation but are subject to misspecification and need new developments when the number of domains is large and the inclusion mechanism is not simply random.

To protect against model misspecification, Little (1983) recommends modeling differences in the distribution of outcomes across classes defined by differential probabilities of inclusion. Si et al. (2015) construct poststratification cells based on the unique values of inclusion probabilities and build hierarchical models to smooth cell estimates as advocated by Little (1991, 1993).

We propose to use Bayesian hierarchical models accounting for survey design to generate weights that can be used in design-based inference. The inference is well calibrated and valid with good frequentist properties (Little, 2011). For large samples, the inference will parallel with design-based inference. For small samples, the hierarchical model smoothing will stabilize domain estimation and generate robust weighting adjustment.

We use the intrinsic variables that are used for design weight construction, nonresponse adjustment and calibration, assume they are discretized, and construct poststratification cells based on the cross-tabulation. Weights are derived through the regressing survey outcome on variables used for weighting given the poststratification. The inclusion of the outcome variable into weighting and poststratification can

avoid model misspecification and potentially increase efficiency (Fuller, 2009). Multilevel model estimates shrink the cell estimates towards the prediction from the regression model. The MRP framework combines multilevel regression and poststratification, accounts for design features in the Bayesian paradigm, and is then well equipped to handle complex design features. Our proposal distinguishes from the model-based weights in the literature by using the poststratification cell structure and improves by smoothing, thus avoiding negative weight values.

Si et al. (2015) incorporate weights into MRP, increasing flexibility and efficiency comparing to the pseudo-likelihood approach (Pfeffermann, 1993). In the present paper we go further, starting from the variables that are used for weighting and constructing model-based weights as byproducts under MRP. We develop a novel prior specification for the regularization to handle potentially large numbers of poststratification cells. The prior setting allows for variable selection and keeps the hierarchical structure among main effects and high-order interaction terms for categorical variables. That is, if one variable is not predictive, then the high-order interactions involved with this variable are also likely to be not predictive, to facilitate model interpretation. McConville and Toth (2018) use tree-based methods to automatically select poststrata based on auxiliary variables that are potentially correlated with the survey outcome. Our proposed structured prior plays a similar role with the recursive partitioning algorithm to facilitate poststrata selection but improves efficiency by partial pooling. We use the smoothed weights and estimates that are more stable than the regression tree estimator, and the Bayesian framework propagates all sources of uncertainty while McConville and Toth (2019) ignore the variance for tree growing and use mean squared error to approximate the variance.

We have implemented the computation in the R package `rstanarm` (Goodrich and Gabry, 2017). The fully Bayesian inference is realized via Stan (Stan Development Team, 2018, 2017), which uses Hamiltonian Monte Carlo sampling with adaptive path lengths (Hoffman and Gelman, 2014). Stan promotes robust model-based approaches by reducing the computational burden of building and testing new models. The `rstanarm` package allows for efficient Bayesian hierarchical modeling and weighting inference. The codes are publicly available and reproducible. Our developed computation software provides the accessible platform and has the potential to support the unified framework for survey inference.

Section 2 introduces the motivating problem of weighting for an ongoing social science survey. We discuss the method in detail Section 3. Section 4 describes the statistical evaluation of model-based prediction and weighting inference, and demonstrate the efficiency gains in comparison with classical weighting. We apply the proposal to the real-life survey in Section 5. Section 6 summarizes the improvement and discusses further extension.

2 Motivating application

Our methodological research is motivated by operational weighting practice for ongoing surveys. Our immediate goal is to construct weights for the New York City (NYC) Longitudinal Study of Wellbeing

(LSW; Si and Gelman (2014); Wimer, Garfinkel, Gelblum, Lasala, Phillips, Si, Teitler and Waldfogel (2014)), a survey organized by the Columbia University Population Research Center, aiming to provide assessments of income poverty, material hardship, and child and family wellbeing of city residents.

We use the LSW as an example to illustrate practical weighting issues and our proposed improvement, with the understanding that similar concerns arise in other surveys. The survey includes a phone sample based on random digit dialing and an in-person respondent-driven sample of beneficiaries from Robin Hood philanthropic services and their acquaintances. We focus on the phone survey here as an illustration. The LSW phone survey interviews 2,002 NYC adult residents, including 500 cell phone calls and 1,502 landline telephone calls, where half of the landline samples are from low-income areas defined by zipcode information. The collected baseline samples are followed up every three months. We match the samples to the 2011 American Community Survey (ACS) records for NYC. The discrepancies are mainly caused by the oversampling of the low-income neighborhoods and nonresponse.

The baseline weighting process (Si and Gelman, 2014) adjusts for unequal probability of selection, coverage bias, and nonresponse. Classical weights are products of estimated inverse probability of inclusion and raking ratios (Deville, Särndal and Sautory, 1993). However, practitioners have to make arbitrary or subjective choices on the selection and values of weighting factors. For example, to construct weights for individual adults, we have to weight up respondents from large households, as just one adult per sampled household is included in the sample. Gelman and Little (1998) recommend the square root of the ratio of household sizes to family sizes for this weighting adjustment because using household sizes as weights (for example, ACS Weighting Method, 2014) tend to overcorrect in telephone surveys. The raking operation procedure in practice adjusts for socio-demographic factors without tailoring for particular surveys.

The survey organizers are interested in the aspects of life quality of city residents, such as the percentage of children who live under poverty and material hardship. Thus, it is important to get accurate estimates for subpopulations. We would like to develop an objective procedure and let the collected survey data determine the weighting process. The basic principle is to adjust for all variables that could affect the selection and response into weighting. Ideally, we expect that variables used for weighting should include phone availability (number of landline/cell phones and duration with interrupted service), family structure, household structure, socio-demographics and potentially their high-order interaction terms. However, the ACS records only provide information on family size, age, ethnicity, sex, education and poverty gap (a family poverty measure). Meanwhile, considering the substantive analysis goal, the variables describing the number of elder people in the family, the number of children in the family, and the family size, as well as their interactions with poverty gap are recommended by the survey organizers to be included into the weighting process to balance the distribution discrepancy with the population.

To generate classical weights, we select the raking factors that could affect the selection and response, including sex, age, education, ethnicity, poverty gap, the number of children in the family, the number of elder people in the family, the number of working aged people in the family, the two-way interaction

between age and poverty gap, the two-way interaction between the number of persons in the family and poverty gap, the two-way interaction between the number of children in the family and poverty gap, and the two-way interaction between the number of elder people in the family and poverty gap. We collect the marginal distributions from ACS and implement raking adjustment. The generated weights have to be trimmed due to some extreme values.

However, it is possible that the subjective weighting adjustment includes some variables or interactions that are not essentially predictive or does not take account for all the important factors that could be of substantive interest later. The raking adjustment assumes that these factors are independent. This will cause biased domain inference bases on the cross-tabulation if the correlation structure in the sample is different from that in the population. Ideally, we should match based on the joint distribution of these weighting related variables. However, small cell sizes or empty under the deep interactions will lead to extremely large weights that need cell collapsing.

The problems we face with classical weighting for the LSW baseline survey are reflective of problems for most operational weighting practice in real-life surveys, which are often complicated with complex designs, longitudinal structure or multi-stage response mechanisms. The ad-hoc decisions that often go into classical weighting schemes can result in different practitioners generating different sets of weights for the same survey. In order to avoid the need for subjectivity, it is important to propose a model-based weighting procedure that allows the data to select weighting factors. We would like to incorporate these variables used for weighting into the model for survey outcomes for efficiency gains, model their high-order interaction terms under regularized prior setting and generate the weights that can be equally treated as classical weights. A large number of variables used for weighting and deep interactions will cause small weighting cells based on the cross-tabulation. The small weighting cells call for statistical adjustment for smoothness and stability.

MRP have achieved success for domain estimation at much finer levels. Borrowing the strength of hierarchical modeling framework with an informative prior distribution, we should be able to obtain the estimate after smoothing the sparse cells. Poststratification via census information will match the estimate from the sample to the population. The combination of regression and poststratification is similar to the endogenous poststratification concept (Breidt, 2008; Dahlke, Breidt, Opsomer and Keilegom, 2013). We introduce the MRP framework in detail.

3 Method

3.1 Multilevel regression and poststratification

In the basic setting, we are interested in estimating the population distribution of the survey outcome y. When the weighting process is transparent, we can directly include the auxiliary variable X into regression modeling for the survey outcome y. Here X is a q-dimensional vector of variables that affect the sampling design, nonresponse and coverage. Conditional on X, the distribution of inclusion indicator I is ignorable.

The selection of the auxiliary variables and the availability of their joint distributions in the population are the key to success for MRP, and also for all other methods to adjust for the sampling selection and nonresponse bias and yield valid population inferences. We recommend including all variables that potentially affect the sample inclusion, such as design information, paradata, and socio-demographics. One advantage of MRP is to perform variable selection and stabilize weights in contrast to noisy classical weights.

Another practical challenge is that the population distribution of the calibration variables may be unknown. We obtain the joint population control distribution from ACS in our application study. Wang, Rothschild, Goel and Gelman (2015) used the aggregated exit polls, Zhang, Holt, Yun, Lu, Greenlund and Croft (2015) used the census tract-level information and Yougov (Yougov, 2017) used the Current Population Survey to directly obtain such information for the poststratification adjustment. In practice, we recommend to obtain the population information either directly from census or large studies with minimal errors or estimated based on available information in related studies. Some auxiliary variables' population distribution may not be available in the census database, such as the number of phones, and we can estimate from other surveys as reference samples. Reilly, Gelman and Katz (2001) applied models to predict the unknown population poststratification information. When marginal distributions are available, Little and Wu (1991) discuss an equivalent model approach for raking and Si and Zhou (2020) develop a Bayes raking estimation in the population cell size estimation. We discuss extensions to develop an integrative framework accounting for the estimation uncertainty of unknown control information in Section 6. The availability of population control information with high quality and predictive power directly affects inferential validity, either for model-based or design-based approaches.

Under MRP, the auxiliary variables X are discretized, and their cross-tabulation constructs the poststratification cells j, with population cell sizes N_j and sample cell sizes n_j, for $j = 1, \ldots, J$, where J is the total number of poststratification cells (Little, 1991, 1993; Gelman and Little, 1997; Gelman and Carlin, 2001). Then the total population size is $N = \sum_{j=1}^{J} N_j$, and the sample size is $n = \sum_{j=1}^{J} n_j$.

Poststratification inference is different from design-based inference under stratified sampling by the fact that n_j's are now random functions of the sampling distribution I. In the repeated sampling of I, there is a nonzero probability that $n_j = 0$ for some j. The usual resolution of this problem is to condition on n_j's observed in the realized sample, however, the sample inference is not design-unbiased conditionally on n_j's. The MRP framework assumes a model for n_j's to account for the design feature.

The poststratification implicitly assumes that the units in each cell are included with equal probability. Suppose θ is the population estimand of interest, such as the overall or domain means, and it can be expressed as a weighted sum over any subset or domain D of the poststrata,

$$\theta = \frac{\sum_{j \in D} N_j \theta_j}{\sum_{j \in D} N_j}, \tag{3.1}$$

where θ_j is the corresponding estimand in cell j. The proposed poststratified estimator will be of the general form,

$$\tilde{\theta}^{\mathrm{PS}} \;=\; \frac{\sum_{j\in D} N_j \tilde{\theta}_j}{\sum_{j\in D} N_j}, \tag{3.2}$$

where $\tilde{\theta}_j$ is the corresponding estimate in cell j. Various modeling approaches can be used to estimate the cell estimates, such as the flexible nonparametric Bayesian models and machine learning algorithms (Rasmussen and Williams (2006); Hastie, Tibshirani and Friedman (2009)). Here, we illustrate using a hierarchical regression model.

In practice, survey weights are attached to each unit, even though they are not attributes of individual units. It is natural to generate unit-level weights based on the entire survey design, and use the weighted averages of the form, such as $\tilde{\theta} = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i$. Our goal here is to obtain an equivalent set of unit-level weights w_i through a model-based procedure for the estimation of $\tilde{\theta}^{\mathrm{PS}}$ to connect weighting and poststratification. Therefore, regression models can be used to obtain $\tilde{\theta}_j$, poststratification accounts for the population information, and model-based weights are re-derived via the expression (3.2).

In classical regression models, full poststratification is a special case, where the cell estimates are computed separately for each cell without any pooling effect, i.e., no pooling. For example, if we are interested in the population mean, then the cell means will be used as the cell estimates. Generally, classical regression models are conducted on cell characteristics without going to the extreme fitting separately for each cell. If more interactions among the characteristics are included, the resulted weights become more variable. On the other side, complete pooling ignores the heterogeneity among cells. Hierarchical regression models will smooth the variable estimates under partial pooling.

Gelman (2007) uses the exchangeable normal model as an illustration and shows that the poststratification estimate $\tilde{\theta}^{\mathrm{PS}}$ for population mean can be expressed as a weighted average between the cell means and the global mean, which yields the unit weights, also as a weighted average between the completely smoothed weights, $w_j = 1$, and the weights from full poststratification, $w_j = (N_j/N)/(n_j/n)$. Hierarchical poststratification is approximately equivalent to shrinkage of weights through the shrinkage of the parameter estimates. The degree of shrinkage goes to zero as the sample size increases, which implies that estimates from the model are close to the truth under the sampling design. However, further developments are necessary to handle a large number of cells and deep interactions, and rigorously evaluate the performance of model-based weights.

In our application to the LSW study, the variables used for weighting include age (5 categories), ethnicity/race (5 categories), education (4 categories), sex (2 categories), poverty measure (5 categories), family size (4 categories), the number of elder people (3 categories) and the number of children (4 categories), in the family, and this results in $J = 5 \times 5 \times 4 \times 2 \times 5 \times 3 \times 4 \times 4 = 48,000$ poststrata. The majority of the poststratification cells will be empty or sparse due to the limited sample size (2,002). The sample cell sizes are unbalanced. Often cells are arbitrarily collapsed or combined (Little, 1993) without theoretical justification. Recent model-based weighting smoothing procedures across cells could not handle such sparse cases (Elliott and Little, 2000). Xia and Elliott (2016) introduced a Laplace prior

for weight smoothing across a modest number of poststrata based on inclusion probabilities but ignored the variables used for weighting and their hierarchy structure. Using the MRP framework, we account for the variable hierarchy structure to smooth and pool the estimates across the sparse and unbalanced cell sizes with a novel set of prior distributions.

3.2 Structured prior distribution

We introduce a structured prior distribution to improve MRP under the sparse and unbalanced cell structures, thus yielding stable model-based survey weights that account for design information. Suppose the population distribution of X is known, that is, we can obtain N_j's from the external data to describe a joint distribution of the variables used for weighting. Extension to unknown N_j's is discussed in Section 6. In practice, the number of poststratification cells J can be large, even much larger than the sample size n. The variables used for weighting could affect the inclusion through a complex relationship or a differential response mechanism. Deep interactions are essential for complex relationship structure, but we cannot include all and have to select the predictive main effects and interactions.

Suppose the collected survey response is continuous, y_i, for $i = 1, ..., n$, and we are interested in the population mean $\bar{Y}$ estimation. We use $\left(X^{1\mathsf{T}}, ..., X^{J\mathsf{T}}\right)^{\mathsf{T}}$ to represent the $J \times q$ predictor matrix in the population under the poststratification framework. For illustration, assume a normal distribution,

$$y_i \sim N\left(\theta_{j[i]}, \ \sigma_y^2\right), \tag{3.3}$$

where $j[i]$ denotes the cell j that unit i belongs to. We can also consider unequal variances, allowing the cell scale σ_y to vary across cells, indexed as σ_j. For the prior specification of θ_j, one choice can be $\theta_j = X^j\beta$, and β is assigned with some prior distribution. In the hierarchical regression example of Gelman (2007), a multivariate normal distribution is considered, $y_i \sim N\left(X_i\beta, \Sigma_y\right)$ and $\beta \sim N\left(0, \Sigma_\beta\right)$, where the covariates include all main effects and a few selected two-way interactions in X and the covariance matrix Σ_β is diagonal with different scales. However, the model is subject to misspecification, and the generated weights could be negative.

Since X^j consists different level indicators of the q discrete auxiliary variables, we can express the population cell mean θ_j as

$$\theta_j = \alpha_0 + \sum_{k \in S^{(1)}} \alpha_{j,k}^{(1)} + \sum_{k \in S^{(2)}} \alpha_{j,k}^{(2)} + ... + \sum_{k \in S^{(q)}} \alpha_{j,k}^{(q)}, \tag{3.4}$$

where $S^{(l)}$ is the set of all possible l-way interaction terms, and $\alpha_{j,k}^{(l)}$ represents the kth of the l-way interaction terms in the set $S^{(l)}$ for cell j. For example, $\alpha_{j,k}^{(1)}$'s with $k \in S^{(1)}$ refer to the main effects, $\alpha_{j,k}^{(2)}$'s with $k \in S^{(2)}$ being the two-way interaction terms, for cell j. This decomposition covers all possible interactions among the q variables. When the cell structure is sparse, variable selection is necessary. In practical applications, we recommend the initial inclusion of covariates and interactions with substantive importance and scientific interest in Model (3.4) and perform Bayesian variable selection under the proposed structured prior setting.

We induce structured prior distributions to be able to handle deep interactions and account for their hierarchy structure, where the high-order interaction terms will be excluded if one of the corresponding main effects is not selected. Larger main effects often lead to larger effects of the involved interaction terms. Ideally, more shrinkage should be put on the high-order interactions than that on the main effects, and the prior setting should reflect the nested structure. The challenge embodies the problem in Bayesian inference for group-level variance parameters in an ANOVA structure (Gelman, 2005, 2006). Volfovsky and Hoff (2014) introduce a class of hierarchical prior distributions for interaction arrays that can adapt to the potential similarity between adjacent levels, where the covariance matrix for the high-order interactions is assumed as a Kronecker product of the covariance matrices of main effects after adjusting relative magnitudes. Our proposal extends by inducing more structure among the variance parameters, more shrinkage and smoothing effect to handle an extremely large number of cells with unbalanced sizes than the generally balanced setting in Volfovsky and Hoff (2014), and improves the computation performance.

We start with independent prior distributions on the regression parameters α:

$$\alpha_{j,k}^{(l)} \sim N\left(0, \left(\lambda_k^{(l)}\sigma\right)^2\right),$$

where $\lambda_k^{(l)}$ represents the local scale and σ is the global error scale, for $k \in S^{(l)}$ and $l = 1, \ldots, q$. The error scale is the same across the main effects and high-order interactions, while the local scales are different. The shrinkage effect is induced through the specification of local scales. We assume the local scale of high-order interactions is the product of those for the corresponding main effects after adjusting relative magnitudes.

$$\lambda_k^{(l)} = \delta^{(l)} \prod_{l_0 \in M^{(k)}} \lambda_{l_0}^{(l)},$$

where $\delta^{(l)}$ is the relative magnitude adjustment and $M^{(k)}$ is the collection of corresponding main effects that construct the kth l-way interaction in the set $S^{(l)}$. For example, the local scale of the three-way interaction among age, sex, and education, middle-aged men with college education, will be the product of those for the main effects on age, sex, and education, that is, the product of the local scale parameters for middle-aged, men, and college educated, respectively.

We use the following hyperpriors on the scale parameters:

error scale: $\sigma \sim \text{Cauchy}_+ (0, 1)$

local scale for main effects: $\lambda_k^{(1)} \sim N_+(0, 1)$

local scale for high-order interactions: $\lambda_k^{(l)} = \delta^{(l)} \prod_{l_0 \in M^{(k)}} \lambda_{l_0}^{(1)}$ \hfill (3.5)

relative magnitude for high-order interactions: $\delta^{(l)} \sim N_+ (0, 1)$, for $l = 2, \ldots, q$.

Here Cauchy$_+$ and N_+ denotes the positive part of the Cauchy and normal distributions, respectively. Gelman (2006) proposes the half-Cauchy prior for the scale parameter in hierarchical models, which has the appealing property that it allows scale values arbitrarily close to 0, with heavy tails allowing large values when supported by the data. When $\lambda_k^{(l)}$ is close to 0, the posterior samples of $\alpha_{j,k}^{(l)}$ are shrunk towards 0. The scale parameter for the high-order interaction terms will be 0 if any of the related scale parameters for the main effects is 0. The overall regularization effect is determined by the error scale and the multiplicative scale parameters of the corresponding main effects. We assign a noninformative prior distribution to the intercept term and weakly informative prior distributions to the two global error scale parameters (σ_y, σ), where $\sigma_y \sim$ Cauchy$_+$ (0, 5).

The global-local shrinkage prior can stabilize random effects modeling in small area estimation (Tang, Ghosh, Ha and Sedransk, 2018). Our proposed prior specification features the global-local shrinkage and group selection of all possible level indicators for the same variable, similar to the group lasso (Yuan and Lin, 2006). We achieve the goal of variable selection under the similar specification with the Horseshoe prior distribution (Carvalho, Polson and Scott, 2010) and improve by setting up the group selection and multiplicative scales for high-order interactions for sparsity gains. We introduce weakly informative half-Cauchy prior distributions to error scales and informative half-normal prior distributions to the local scale parameters to improve parameter shrinkage estimation and computation efficiency. When the posterior estimation of the scale parameter is close to 0, indicating the variable is not predictive; post-processing can be done to exclude the variable from poststratification cell construction for dimension reduction. This class of priors allows for variable selection in high dimension and keeps the hierarchical structure among main effects and interactions.

Piironen and Vehtari (2016) recommend the prior choice for the global shrinkage hyperparameters based on prior beliefs about the number of nonzero coefficients in the model. The hierarchy setting with correlated variables requires further investigation. We use the default choice Cauchy$_+$ (0, 1) and conduct an extensive sensitivity analysis of the hyperparameter specification, where the results do not change.

3.3 Model-based weights

We can re-express (3.4) and (3.5) as the exchangeable normal model:

$$\theta_j \sim N(\alpha_0, \sigma_\theta^2), \quad \sigma_\theta^2 = \sum_{l=1}^{q} \sum_{k \in S^{(l)}} \left(\lambda_k^{(l)} \sigma \right)^2 . \tag{3.6}$$

Conditional on the variance parameters, the posterior mean in the normal model with normal prior distribution is a linear function of data; thus we can determine *equivalent weights* w_i^*'s so that one can re-express the smoothed estimate $\sum_{j=1}^{J} N_j \big/ N \tilde{\theta}_j$ as a classical weighted average, $\sum_{i=1}^{n} w_i^* y_i \big/ \sum_{i=1}^{n} w_i^*$. Combining the posterior mean estimates for θ_j and the model-based estimate given in Model (3.2), Gelman (2007) derives the equivalent unit weights in cell j that can be used classically.

$$w_j \approx \frac{n_j/\sigma_y^2}{n_j/\sigma_y^2 + 1/\sigma_\theta^2} \cdot \frac{N_j/N}{n_j/n} + \frac{1/\sigma_\theta^2}{n_j/\sigma_y^2 + 1/\sigma_\theta^2} \cdot 1, \qquad (3.7)$$

where the model-based weight is a weighted average between full poststratification without pooling (weights of $(N_j/N)/(n_j/n)$) and complete pooling (weights equal to 1). The pooling or shrinkage factor is $1/(1 + n_j\sigma_\theta^2/\sigma_y^2)$, which depends on the group and individual variances as well as sample size in the cell. The model-based weights are random variables, and fully Bayesian inference will propagate the corresponding variability. We collect the posterior mean values and treat as the weights that can be used the same as classical weights.

3.4 Computation

The Bayesian hierarchical prediction and weighting inference procedure is reproducible and scalable. We implement the proposed structured prior distributions in the open source R package `rstanarm` (Goodrich and Gabry, 2017). The computation codes are available online (Si, Trangucci and Gabry, 2020) for public use. We present the example code for the real data application in Appendix A to demonstrate the user-friendly and efficient computation interface, where survey practitioners can straightforwardly use and adapt. The fully Bayesian inference is realized via Stan. As open source and user-friendly software, Stan contributes to the wide application of Bayesian modeling. Survey practitioners resist model-based approaches mainly due to computation burden. However, model-based methods are ready to face the new challenges on big survey data, such as unbalanced cell structure, combining multiple surveys and analyzing streaming data. The development of Stan can improve the generalization of the model-based approach and provide the computational platform for the unified survey inference framework.

In our implementation, the Markov chain Monte Carlo samples mix well and the chains converge quickly. The fast computation speed widens the usability of model-based survey inference approaches. The proposed prior specification improves the stability for smoothed weights under partial pooling. We compare the model-based weights with classical weights in Section 4 and 5 to demonstrate the calibration for design-based properties (Little, 2011). Furthermore, we illustrate the proposed improvement for domain estimation under unbalanced and sparse sample cell structure.

4 Simulation studies

We evaluate the Bayesian procedure by the design-based properties and demonstrate the validity. We consider two main simulation scenarios: a slightly unbalanced structure with a moderate number of poststratification cells and a very unbalanced structure with a large number of poststratification cells. We evaluate the statistical validity of the model-based and weighted estimation for the finite population and domain inference to demonstrate the improved capability to solve the classical weighting problems. To illustrate the capability of variable selection and hierarchy maintenance and the resulting efficiency gains, we compare the posterior estimation with that under independent prior setting but without the

multiplicative scale constraint, which is similar with Horseshoe prior under group specification, called as independent prior distributions in the paper: $\lambda_k^{(l)} \sim N\left(0, \left(\sigma_k^{(l)}\right)^2\right)$.

We consider model-based predictions under the structured prior (Str-P) and the independent prior (Ind-P) distributions. For weighted inference, we evaluate the estimation after applying the model-based weights under structured prior (Str-W) setting, model-based weights under independent prior (Ind-W) distributions, weights obtained via raking adjustment (Rake-W), classical poststratification weights (PS-W), and inverse probability of selection weighting (IP-W). We present the graphical diagnosis tools to compare the weights and weighted inference.

We borrow 2011 ACS survey of NYC adult residents treating it as the "population", and randomly draw samples out of it according to a pre-specified selection model without nonresponse. We collect covariates from ACS and simulate the outcome variable to obtain the true distribution as a benchmark. The details of model specifications for the following scenarios are presented in Appendix B. We implement the raking procedure by balancing the marginal distributions of the calibration variables in the selection model and generate the raking weights. The classical poststratification weights N_j / n_j's are obtained by matching the selected sample cell indices with those of the population cells. The selection model can provide the inverse probability of selection weights by matching the sampled unit indices. We also generate model-based weights under independent prior distributions for the main effects and high-order interaction terms of the ACS variables. The generated weights are normalized to average 1 for comparison convenience.

4.1 Slightly unbalanced structure

We first handle slightly unbalanced structure when the number of poststratification cells and the sample cell sizes are moderate. We implement repeated sampling process to investigate the frequentist properties of model-based predictions and weighted inferences. With little shrinkage effect on high-order interactions, the model-based prediction and weighting with structured prior distributions have similar performance with that under independent prior distributions, while outperforming the classical weighting approaches.

Assume three variables are included in the selection and outcome models: age, ethnicity, and education. We discretize the three variables in ACS as *age* (18–34, 35–44, 45–54, 55–64, 65+), *eth* (non-Hispanic white, non-Hispanic black, Asian, Hispanic, other), and *edu* (less than high school, high school, some college, bachelor degree or above). The number of poststratification cells is $5 \times 5 \times 4 = 100$. We assume the outcome depends on deep interactions, including all the main effects, two-way and three-way interaction terms among the three variables; and the selection indicator depends on the three main effects. The specific values of the coefficients are given in Tables B.2-B.3 in Appendix B. The values are set to reflect the strong correlations between the covariate and dependent variables. And the effects are not necessarily similar across the adjacent factor levels, different from the scenarios in Volfovsky and Hoff

(2014). The error scale in the outcome model is set as 1, where the true value is always fully recovered from the posterior estimation. The data generation model is different from the estimation model, but the latter is flexible enough to cover the former since the dependency structure will be recovered by the estimation. The proposal is robust against model misspecification.

We repeat the sampling 500 times. The sample sizes vary between 2,141 and 2,393 with median 2,288. Empty sample cells occur with spread-out selection probabilities (ranging from 0.001 to 0.269) over the repeated sampling process. The number of occupied cells in the sample is between 80 and 93 with median 87. The slightly unbalanced cell structure is common in practical surveys with simple and clean sampling design. The population quantities of interest include the overall mean, domain means across the 13 (= 5 + 4 + 4) marginal levels of three variables and domain mean for nonwhite youths (an example of interaction between age and race/ethnicity). We examine the absolute value of estimation bias, root mean squared error (RMSE), standard error (SE) approximated by the average value of standard deviations (Ave. SD) and nominal coverage rate of the 95% confidence intervals.

The outputs in Figure 4.1 show that the model predictions have the smallest RMSE, the smallest SE with reasonable coverage rates, and comparable bias among all the methods. All variables affecting the outcome and selection mechanism are included in the modeling to satisfy the Bayesian principle for ignorable sampling mechanism. The model will predict all the cell estimates including the empty cells in the sample, fully using the population information and poststratification cell structure. The weighting inference is conditional on the observed units within occupied cells, and thus less efficient than the model predictions. Generally, the model-based weighting inference has smaller RMSE and SE but more reasonable coverage rates than that with classical weighting. Raking adjustment is not valid for the domain estimation with large bias, large RMSE, and poor coverage, even though the selection mechanism depends on only the main effects. The inverse probability of selection weighting inference tends to have large SE but low coverage rates, especially for domain estimation. The poststratification weighting inference is close to the model-based weighting estimation since the domain sizes are modestly large. The cell shrinkage effect towards no weighting is small (between 0 and 0.19 with mean 0.05) under slightly unbalanced design. The number of cases who are less than high school educated is small (around 80), resulting in large estimation bias and SE for the weighting inferences, but not in model-based predictions. The model-based predictions stabilize the small area estimation by smoothing, as shown in Table 4.1 that displays the numerical comparison for the subdomain inference.

Model prediction performs well and similarly under the structured prior distribution or independent prior distribution. This is expected due to the small shrinkage effect. The cell structure is slightly unbalanced, and the outcome and selection models depend on all the main effects and high-order interaction terms. But the structured prior setting yields more efficient inference than the independent prior setting with smaller SE. This improvement is obvious in the very unbalanced design as shown in the following simulation of Section 4.2.
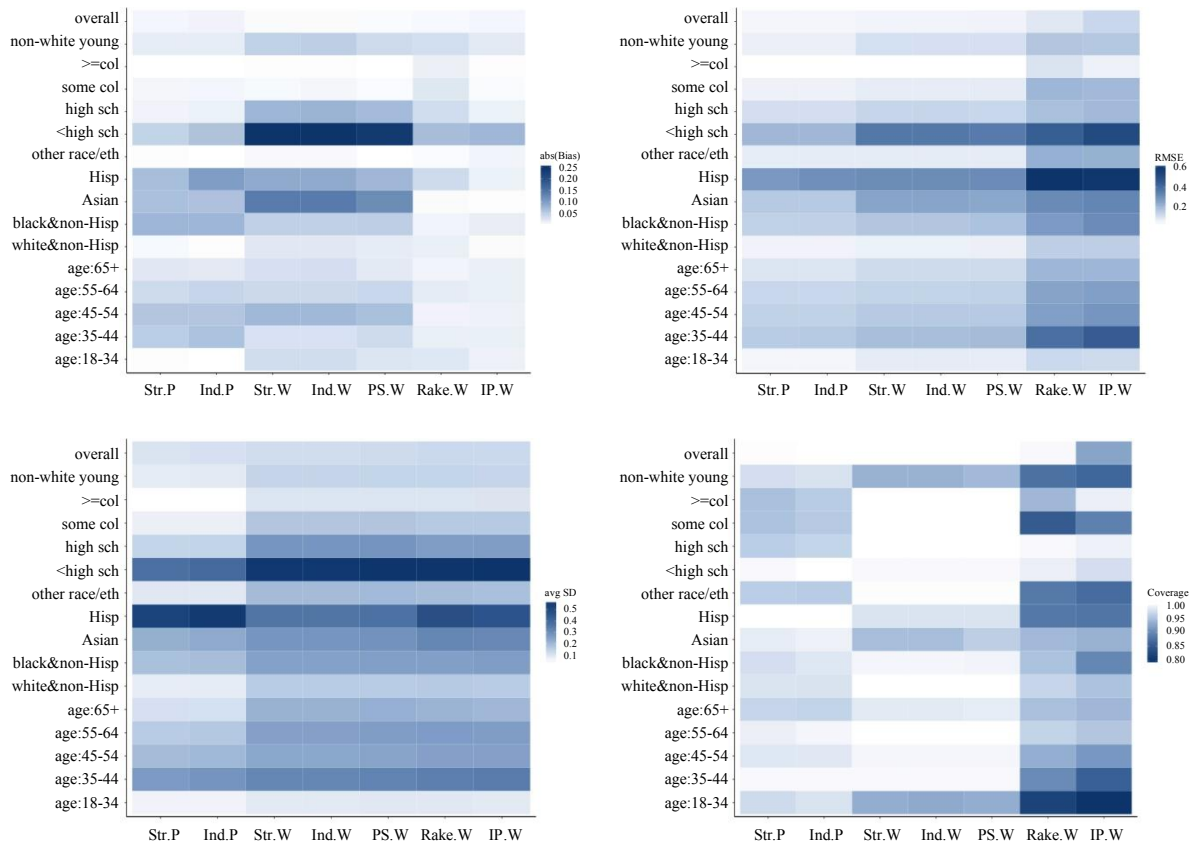
Figure 4.1 Comparison of prediction and weighting performances on the validity of finite population inference under slightly unbalanced design. The y-axis denotes different groups for the mean estimation. The x-axis includes two model-base prediction methods (Str-P, Ind-P), two model-based weighting methods (Str-W, Ind-W), and three classical weighting methods (PS-W, Rake-W, IP-W). Str-P: model-based prediction under the structured prior; Ind-P: model-based prediction under the independent prior distribution; Str-W: model-based weighting under structured prior; Ind-W: model-based weighting under independent prior distribution; Rake-W: weighting via raking adjustment; PS-W: poststratification weighting; and IP-W: the inverse probability of selection weighting. The plots show that the model-based predictions outperform weighting with the smallest RMSE, the smallest SE, reasonable coverage rates, and comparable bias among all the methods. Model-based weighting inference has smaller RMSE and SE but more reasonable coverage rates than that with classical weighting.

Table 4.1
Comparison of prediction and weighting performances on the subgroup mean for non-white youth under slightly unbalanced design

	Str-P	Ind-P	Str-W	Ind-W	PS-W	Rake-W	IP-W
Abs.Bias	0.02	0.02	0.05	0.05	0.04	0.03	0.02
RMSE	0.07	0.07	0.11	0.11	0.10	0.17	0.17
Ave.SD	0.08	0.08	0.13	0.13	0.13	0.13	0.13
Coverage	0.97	0.98	0.94	0.94	0.94	0.88	0.86

Additionally, we considered nine cases with different survey outcome models and sample selection models depending on various predictors as in Table B.1 in Appendix B. The specific values of the coefficients are given in Tables B.2-B.3. The conclusions are consistent that the model-based prediction and weighting yield more efficient and precise inference than that under classical weighting, in particular for domain estimation.

4.2 Very unbalanced structure

Complex sampling design and response mechanisms tend to create very unbalanced data structures where most poststratification cells are sparse and empty. The proposed structured prior setting brings in strong regularization effect to stabilize the model prediction and improves the estimation efficiency, especially for domain estimation, outperforming the independent prior distributions. The posterior inference on scale parameters can inform variable selection to improve model interpretation. When the main effects are not predictive, neither are the related high-order interactions. However, the posterior inference with independent prior distributions distorts the hierarchical structure between main effects and high-order interactions and hardly informs variable selection. The classical weighting inferences are highly variable in the sparse scenario.

Following the LSW, we collect eight variables in the 2011 ACS-NYC data that affect sample inclusion: *age* (18–34, 35–44, 45–54, 55–64, 65+), *eth* (non-Hispanic white, non-Hispanic black, Asian, Hispanic, other), *edu* (less than high school, high school, some college, bachelor degree or above), *sex* (male, female), *pov* (one household income or poverty measure, poverty gap under 50%, 50–100%, 100–200%, 200–300%, more than 300%), *cld* (0, 1, 2, 3+ young children in the family), *eld* (0, 1, 2+ elders in the family), and *fam* (1, 2, 3, 4+ individuals in the family). The number of unique cells occupied by this classification is 8,874, while the number of poststratification cells constructed by the full cross-tabulation is 48,000.

In the simulation described in Table B.4 and Table B.5, the selection probability depends on the main effects of all variables, while the outcome depends on the main effects of five variables. The cell selection probabilities will be clustered, where some cells have the same selection probabilities. The error scale in the outcome model is set as 1. The selection probabilities fall between 0 and 0.90 with average 0.12, and we select 6,374 units. Even though the sample sizes are large, the simulation creates a very unbalanced structure. The majority of the cells are empty, and 1,096 of 1,925 selected cells have one unit. Starting from an estimation model with sparsity, we assume the Model (3.4) for the cell estimations includes the main effects of the eight variables, eight two-way interactions, and two three-way interactions. These terms are potentially important factors for weighting from the survey organizer's view. Our proposal can provide the insight of variable selection and then facilitate dimension reduction.

When only the main effects are predictive, the posterior median values under the structured prior setting for the scales of the *cld*, *eld*, and *fam* are small (0.002, 0.003, 0.000), and the posterior median

values for the scales of all high-order interactions are close to 0 (with magnitude smaller than or around 0.0001). The posterior mean of the error scale is 0.99 with SE 0.008, close to the true value 1. This is consistent with the simulation design. With independent prior distributions, however, the hierarchical structure between the main effects and high-order interaction terms is ignored. The posterior samples of scale parameters of the high-order interactions can be larger than that of the main effects. It is unclear about their predictive power and then hard to decide which terms to be selected. The posterior samples of the variance parameters under the independent prior distributions tend to be highly variable with heavy tails. For example, the variances of the main effects of age and sex have extremely large sampled values (14,496 and 390,000) and skewed distributions. For variables with a small number of levels, such as sex, the group-level variance estimation is sensitive to the prior distribution, and the independent prior distribution cannot regularize well. The structured prior distribution performs better by assuming the prior distributions share some common parameter and using more information for estimation and then is able to stabilize the variance estimation. The structured prior setting yields more stable inference than the independent prior, and moreover can facilitate variable selection.

The proposed structured prior setting suggests that we exclude the nonpredictive main effects and high-order interactions from the regression model for cell estimates, by either post-processing the posterior samples of the corresponding scales and coefficients to be 0 or refitting the updated model. In the simulation design, three variables affect the selection probability but are not related to the outcome. The inclusion of these variables into the regression model will increase the inference variability. The poststratification cell structure accounts for the eight variables to meet the ignorable sampling assumption. A further modification could be the exclusion of the three variables from the poststratification, which could make the assumption of ignorable sampling vulnerable but have efficiency gains. This is a tradeoff between efficiency and robustness that needs balance based on substantive interest. The selection of survey outcome variables in the weighting process needs further investigation, which we will elaborate in Session 6. We compared the inference with that after excluding the nonpredictive terms and obtained similar outputs for the finite population and domain estimation since the parameter estimates are close to 0 for the nonpredictive terms. Here we present the outputs keeping such variables in the poststratification cell construction and the regression model.

First, we compare the generated weights by the model-based and classical methods. We collect the posterior samples of generated weights and present the posterior mean as the model-based weights. The model-based weights have smaller variability and narrower range than the classical weights, as shown in Figure 4.2. The iterative proportional fitting procedure does not converge after the default 10 iterations that need increasing. We examine the distribution of the outcome after accounting for the weights and compare with the population and sample distribution in the right plot of Figure 4.2. The sample distribution differs from the population distribution by underestimating the outcome values. The weighted distribution shifts towards the true population. The outcome distributions after weighting are similar among the model-based and classical methods, and the model-weights generate a smooth distribution of

outcomes. This is reasonable as we expect the model-based weights perform similarly with classical weights on point estimation but improve efficiency by reducing the variability. The shrinkage effect under the structured prior distribution is large, between 0.86 and 1.00 with mean 0.90. The very unbalanced cell structure needs a strong smoothing effect across cells. The model-based weights under the structured prior and independent distributions have similar distributions with the poststratification weights, so the latter two sets of weights are omitted in Figure 4.2.
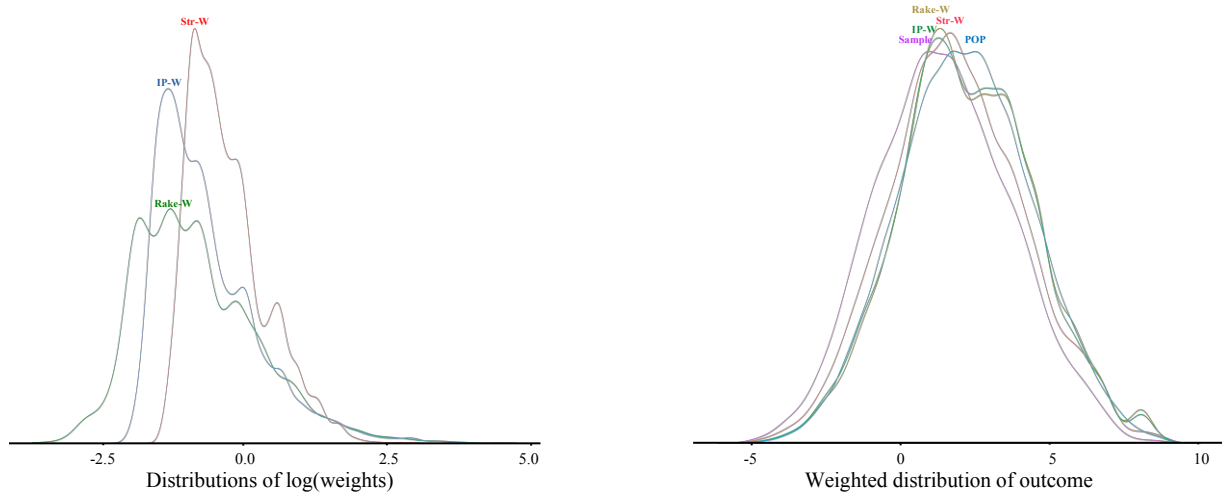

Figure 4.2   Comparison of generated weights after logarithmic transformation and weighted outcome distributions under a very unbalanced design. Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and IP-W: inverse probability of selection weighting. Sample: sample distribution of the outcome; and POP: population distribution of the outcome. The model-based weights are more stable and generate a more smoothed outcome distribution after weighting than the raking weights and the inverse probability of selection weights.

We examine the inference for the overall mean and domain means across the marginal levels and for nonwhite young adults. The conclusions are the same as that in Section 4.1. Model-based prediction outperforms weighting inference with smallest bias and SE. The benefit can be explained by that the model uses the population information for empty cell prediction under regularization. Model-based weighting inference has smaller SE than that with classical weighting. Even when the selection probabilities depend on only main effects, raking yields small bias but performs badly with large SE.

Under the very unbalanced design, the model-based weighting inference under structured prior setting is more efficient than that under independent prior setting or with poststratification weights. We compare the SE of the marginal mean estimates of the eight variables from the three weighting methods and plot the relative ratios in the left plot of Figure 4.3. The model-based weighting inference has smaller SE than the poststratification weighting, and the weighting under structured prior setting has the smallest SE. Because the sample sizes and the domain sizes are large and the data generation model is sparse, the

model-based weighting inference has a little but not much improvement over the poststratification weighting inference due to small smoothing effect.

The model-based prediction and inference under the structured prior setting are more efficient than that under the independent prior setting. The SEs are smaller with the structured prior than those with the independent prior in the right plot of Figure 4.3. To demonstrate the efficiency gain, we look at the SEs for the population cell estimates. The Bayesian structural inference generally has smaller variability than that with independent prior, especially in the sparse scenarios.
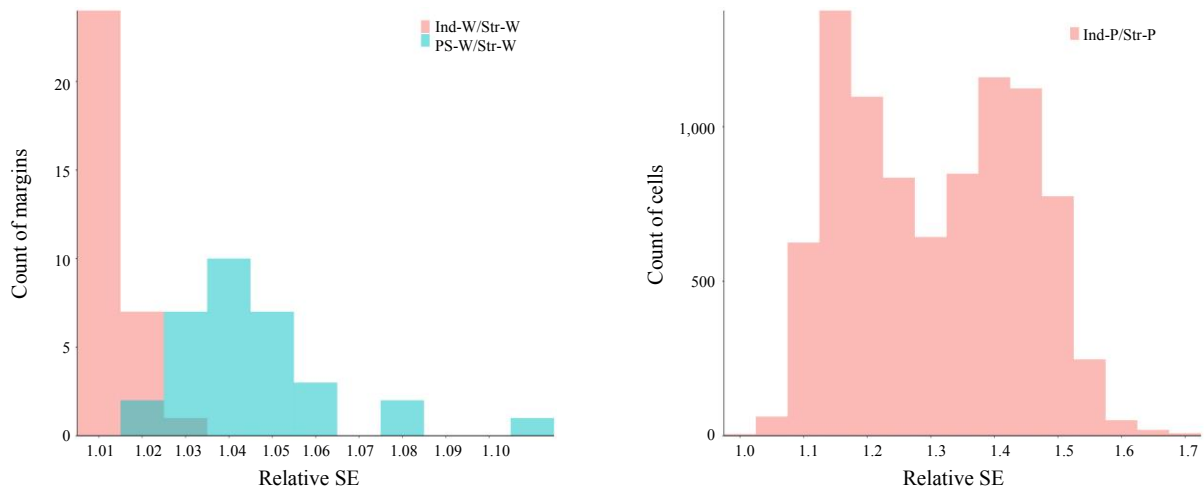
Figure 4.3 Efficiency comparison of prediction and weighting performances on finite population domain inference under a very unbalanced design. The left plot examines the mean estimation across the margins defined by the eight variables. The right plot presents the population cell mean estimation. The model-based weighting and prediction under the structured prior distribution yield smaller SE than those under independent prior. Model-based weighting yields smaller SE than poststratification weighting.

We assume different outcome and selection models with different covariates with scenarios summarized in Table B.4 and achieve the same evaluation conclusions.

5 Application to longitudinal study of wellbeing

With the background introduced in Section 2, we apply the prediction and weighting inference to the NYC Longitudinal Study of Wellbeing. We match the LSW to the adult population via the ACS. We would like to conduct finite population and domain inference and generate weights allowing for general analysis use. The outcome of interest is the self-reported score of life satisfaction on a 1–10 scale. We model the outcome as normally distributed, which is not quite correct given that the responses are discrete,

but should be fine in practice for the goal of estimating averages. We first include the same eight variables to construct the poststratification cells and use the same estimation model as those in Section 4.2 under the structured prior setting. The posterior inference shows that the variables *sex, cldx, eldx*, and *psx* are not predictive, and neither are the related high-order interactions. The scale estimates of such terms have posterior median values close to 0 and several large values as long tails. The posterior samples of scales for several high-order interactions among the remaining four variables concentrate around 0, showing these quantities are not predictive. Another complexity is that, for the sample cells of the LSW, the corresponding population cells are not available in the ACS data. This could happen because the sampling frame is not the ACS survey. The population information is unknown for such cells, and untestable assumptions have to be made. The model fitting improves after variable selection when we check the prediction errors for cell estimates.

Hence, we use four variables after selection, *age, eth, edu* and *pov*, which constructs 500 poststratification cells. The 2,002 units in the LSW spread out in 359 cells. The largest sample cell has 86 units, while 92 cells have only one unit. The covariates in the model (3.4) for cell estimates include the main effects of the four variables, five two-way interactions (*age * eth*, *age * edu*, *eth * edu*, *age * inc* and *eth * inc*), and two three-way interactions (*age * eth * edu* and *age * eth * inc*). We implement the fully Bayesian inference with the structured prior distributions. We are interested in estimating the average score of life satisfaction for overall and several subgroups of NYC adults and construct weights for general analysis purposes using the LSW.

The posterior median of the unit scale inside cells σ_y is 1.93 with 95% credible interval [1.87, 1.99]. The posterior median of the group scale σ_θ is 0.79 with 95% credible interval [0.65, 1.02]. These lead to moderately large shrinkage effects between 0.11 and 0.90 with mean 0.30 across cells. The moderate shrinkage effect makes sense based on the four variables and up to three-way interactions being included. The posterior mean values of the model-based weights are presented in the left plot of Figure 5.1. We can generate the raking weights after adjustment for the marginal distributions of the four variables and poststratification weights based on the ACS data. The population information is obtained after applying the ACS personal weights.

Comparing with the classical weights, our model-based weights have smaller variability with standard deviation 0.32 and the ratio of the maximum and minimum value 3.87, and these values are much smaller than those for the raking and poststratification weights, as shown in Table 5.1. The right plot in Figure 5.1 shows the distribution of the lift satisfaction score after weighting. The model-based weighted distributions and classically weighted distributions are similar as expected, which is consistent with the results in Section 4.2. The weighting process adjusts for the sample distribution by upweighting the high scores and downweighting the low scores. The LSW oversamples poor residents who tend not be satisfied with life, and the weighting adjustment balances the discrepancy.
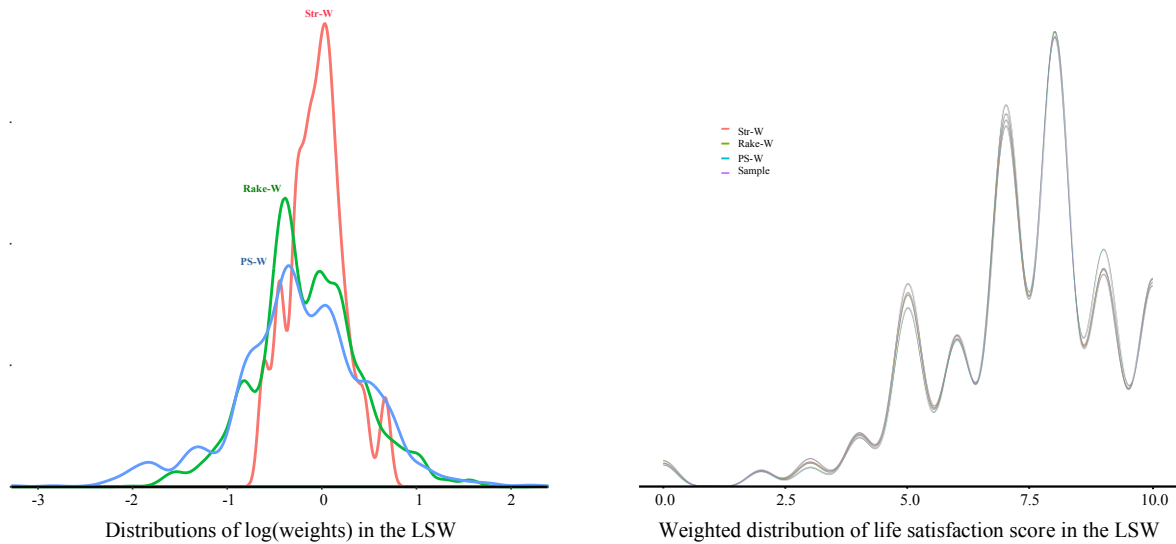
Figure 5.1 Comparison of generated weights after logarithmic transformation and weighted distributions of life satisfaction score in the LSW. Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; IP-W: inverse probability of selection weighting, and Sample: sample distribution of the outcome. The weighted distributions are similar between model-based weights and classical weights, but model-based weights are more stable than classical weights.

Table 5.1 and Figure 5.2 present the finite population and domain inference. The average score of life satisfaction for NYC adults is 7.24 with standard error 0.05, predicted by the structural model. The estimate is similar to that under model-based weighting and raking inferences, but lower than the poststratification weighting inference. However, the difference is not significant. For example, the structural model predicts the average score of life satisfaction for middle-aged, college-educated whites with income more than three times the poverty level as 7.40 with standard error 0.10, higher than that under weighting inferences. Nevertheless, the predicted scores for the elder with relatively low income (7.37 with SE 0.15) and low-income black New Yorkers (7.01 with SE 0.18) are lower than those under weighting inferences. The discrepancy could be explained by the nonrepresentativeness of the LSW and the deep interactions included by the model. The subgroup of individuals who are middle-aged, college-educated whites may be undercovered in the LSW − as empty poststratification cells occurring − with overcoverage among elderly poor blacks. Weighting the collected samples cannot infer or extrapolate inference on those who are not present in the survey. Though the differences are not significant, inferences conditioning on the collected samples cannot recover the truth, especially for the empty cell estimates. Figure 5.2 shows the model-based prediction yields a higher score for young, highly educated and Hispanic NYC adults, but a lower score for those with poverty gap $<50\%$, comparing with the weighted inference.
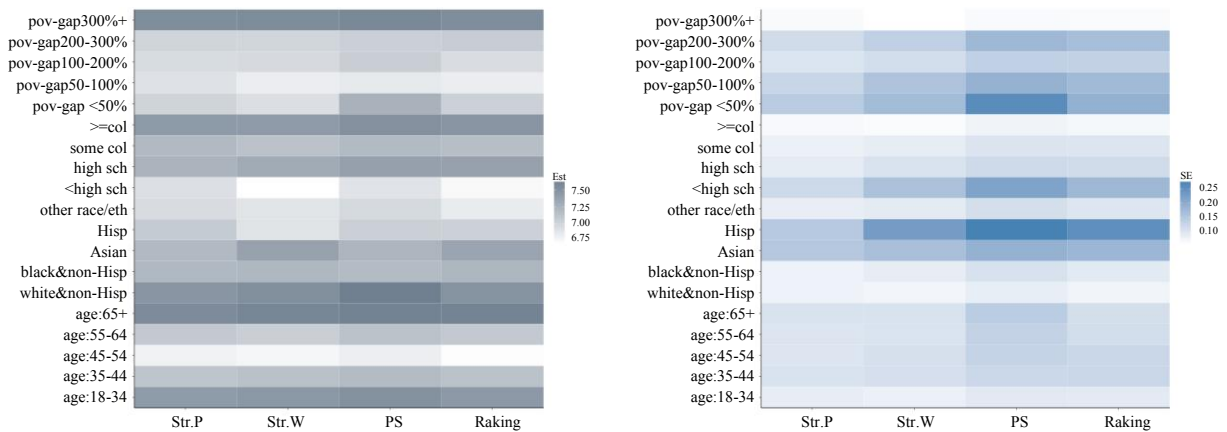
Table 5.1
Comparison of prediction and weighting performances on estimating various domain averages for life satisfaction in the LSW. Str-P: model-based prediction under the structured prior; Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and PS-W: poststratification weighting

	Str-P	Str-W	Rake-W	PS-W
SD of weights / mean of weights		0.32	0.66	0.80
Max weight / min weight		3.87	81.28	274.65
Overall average for NYC adults ($n = 2,002$)				
Est	7.24	7.23	7.24	7.30
SE	0.05	0.05	0.05	0.06
Average for middle-aged, college-educated whites with poverty gap $> 300\%$ ($n = 222$)				
Est	7.40	7.34	7.34	7.34
SE	0.10	0.11	0.11	0.11
Average for elders with poverty gap $< 200\%$ ($n = 154$)				
Est	7.37	7.52	7.49	7.53
SE	0.15	0.18	0.19	0.22
Average for blacks with poverty gap $< 50\%$ ($n = 57$)				
Est	7.01	7.16	7.30	7.16
SE	0.18	0.26	0.28	0.29

The SEs are similar for the overall mean estimation between predictions and various weighting inferences because of the large sample size. For domain estimation, the model-based prediction and weighting are more efficient than that with raking and poststratification weighting, and the model-based prediction has the smallest standard error. The efficiency gains of model-based prediction and weighting are further demonstrated by domain mean estimation for life satisfaction scores across the marginal levels of four variables, shown in Figure 5.2. The model-based prediction and weighting particularly improve small domain estimation and increase the efficiency.

Survey practitioners often compare the weighted distribution of socio-demographics with the population distribution to check the weighting. While weighting diagnostics need further research and management, we follow this routine to compare the model-based and classical weights. We calculate the Euclidean distances between the weighted distributions and the population distribution for the main effects and high-order interactions among the four variables in the LSW, shown in Table B.6 in Appendix B. The weighted distributions are generally close to the true distributions. Raking focuses on adjusting for the marginal distributions of calibration variables but distorts the joint distributions, where the dependency structure is determined only by the sample without calibration. The poststratification weighting adjusts for the joint distribution, but empty cells in the sample present from the exact matching. The unbalanced cell structure yields unstable inference. The model-based weighting smooths the poststratification weightings and outperforms raking to match the distributions of three-way and four-way interaction terms. Practitioners often rely upon the marginal distributions to evaluate weighting performances, thus in favor of raking. However, raking yields high variable and potentially biased inferences, shown in the Section 4, even in the cases when raking adjustment is correct. Modification of model-based weighting to satisfy such desire on matching marginal distributions will be a future extension to incorporate constraints.

Figure 5.2 Comparison of predictions and weighting performances on estimating life satisfaction score across the margins of four variables in the LSW. Str-P: model-based prediction under the structured prior; Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and PS-W: poststratification weighting. Model-based predictions and weighting generate different estimates for several subsets and are generally more efficient comparing with classical weighting.

6 Discussion

We combine Bayesian prediction and weighting as a unified approach to survey inference. Multilevel regression with structured prior distributions and poststratification on the population inference yield efficient estimation when accounting for the design feature. The computation is implemented via Stan and disseminated through the R package `rstanarm` for public use, and the software development promotes the model-based approaches in survey research and operational practice. We construct stable and calibrated model-based weights to solve the problems of classical weights. This article builds up the model-based prediction and weighting framework and serves as the first contribution to evaluate the statistical properties of model-based weights and compare the performances with classical weighting. Model-based weights are smoothed across poststratification cells and improve small domain estimation.

The structured prior uses the hierarchical structure between the main effects and high-order interaction terms to introduce multiplicative constraints on the corresponding scale parameters and informs variable selection. Model improvement can be done after post-processing the posterior inferences. The Bayesian structural model yields more stable inference than that with independent prior distributions. Such hierarchy assumption may not be valid for special cases, such as the Exclusive-Or problem where two variables show no main effects but a perfect interaction. However, we do not have strong evidence in the application studies against the plausibility of hierarchy. Furthermore, the unified prediction and weighting approach is well equipped to deal with complex survey designs and big data in surveys, such as streaming data and combining multiple survey studies.

The general MRP framework is open to flexible modeling strategies. In this article, we illustrate by a regression model with all variables of interest and the high-order interactions and incorporate structured prior distributions for regularization. Other approaches, such as nonparametric models and machine learning tools, can be implemented under the MRP framework, being robust against model misspecification. Si et al. (2015) use Gaussian process regression models to borrow information across poststratification cells based on the distances between the inverse inclusion probability weights. Further extensions include applying such flexible approaches to weight smoothing and deriving the model-based weights.

The broad application opportunities come with various challenges that need further investigation. The model-based weights are outcome dependent, which improves the efficiency but potentially reduces the robustness. Survey organizers prefer a set of weights that can be used for general analysis purpose, without being sensitive to outcome selection. We can compare different weights constructed by several important outcomes and conduct sensitivity analysis. When the model-based weights give different inference conclusions, we recommend choosing the set of weights that generate the most reasonable results, with scientific reasoning and be consistent with the population inference.

The weighted marginal distributions of the calibration variables are a bit different from the population inferences, as in Section 5, which does not meet the usual weighting diagnosis standard of survey organizers. The model-weights tend to match the joint distribution to that in the population, but weight smoothing may bring in bias. Tradeoff constraints can be induced to the model to match the marginal distributions.

Another practical challenge is that the population distribution of the calibration variables may be unknown, that is, the population poststratification cell sizes N_j's are unknown. A supplemental model is needed to allow estimation of this information from the sample and integrated with MRP to propagate all sources of uncertainty as an extension, similar to the framework in Si and Zhou (2020) by incorporating known margins. The model-based predictions and weighting inferences need further extensions to handle discrete outcomes, inference on regression coefficients and non-probability or informative sampling designs (Kim and Skinner, 2013). It will be useful to link these ideas on survey inference with the biostatistical and econometrics literature on inverse propensity score and doubly robust weighting (Kang and Schafer, 2007).

Acknowledgements

Appendix

A. Example code

Here we present code for the application described in the data. We have written a function `model_based_cell_weights` to calculate the model-based weights from a fitted `rstanarm` model.

```
model_based_cell_weights <- function(object, cell_table) {
  stopifnot(
    is.data.frame(cell_table),
    colnames(cell_table) == c("N", "n")
  )
  draws <- as.matrix(object)
  Sigma <- draws[, grep("^Sigma\\[", colnames(draws)), drop = FALSE]
  sigma_theta_sq <- rowSums(Sigma)
  sigma_y_sq <- draws[, "sigma"]^2
  Ns <- cell_table[["N"]] # population cell counts
  ns <- cell_table[["n"]] # sample cell counts
  J <- nrow(cell_table)
  N <- sum(Ns)
  n <- sum(ns)
  # implementing equation 7 in the paper (although i did some algebra first to
  # simplify the expression a bit)
  Nsy2 <- N * sigma_y_sq
  ww <- matrix(NA, nrow = nrow(draws), ncol = J)
  for (j in 1:J) {
    ww[, j] <-
      (Nsy2 + n * Ns[j] * sigma_theta_sq) / (Nsy2 + N * ns[j] * sigma_theta_sq)
  }
  return(ww)
}
# prepare population data: acs_ad has age, eth, edu and inc
acs_ad %>%
  mutate(
    cell_id = paste0(age, eth, edu, inc)
  ) -> acs_ad
acs_design <- svydesign(id = ~1, weights = ~perwt, data = acs_ad)
agg_pop <-
  svytable( ~ age + eth + edu + inc, acs_design) %>%
  as.data.frame() %>%
  rename(N = Freq) %>%
  mutate(
```

```r
    cell_id = paste0(age, eth, edu, inc)
  ) %>%
  filter(cell_id %in% acs_ad$cell_id)
# prepare data to pass to rstanarm
# SURVEYdata has 4 variables used for weighting: age, eth, edu and inc; and outcome Y
dat_rstanarm <-
  SURVEYdata %>%
  mutate(
    cell_id = paste0(age, eth, edu, inc)
  )%>%
  group_by(age, eth, edu, inc) %>%
  summarise(
    sd_cell = sd(Y),
    n = n(),
    Y = mean(Y),
    cell_id = first(cell_id)
  ) %>%
  mutate(sd_cell = if_else(is.na(sd_cell), 0, sd_cell)) %>%
  left_join(agg_pop[, c("cell_id", "N")], by = "cell_id")
# Stan fitting under structured prior in rstanarm
fit <-
  stan_glmer(
    formula =
      Y ~ 1 + (1 | age) + (1 | eth) + (1 | edu) + (1 | inc) +
      (1 | age:eth) + (1 | age:edu) + (1 | age:inc) +
      (1 | eth:edu) + (1 | eth:inc) +
      (1 | age:eth:edu) + (1 | age:eth:inc),
    data = dat_rstanarm, iter = 1000, chains = 4, cores = 4,
    prior_covariance =
      rstanarm::mrp_structured(
        cell_size = dat_rstanarm$n,
        cell_sd = dat_rstanarm$sd_cell,
        group_level_scale = 1,
        group_level_df = 1
      ),
    seed = 123,
    prior_aux = cauchy(0, 5),
    prior_intercept = normal(0, 100, autoscale = FALSE),
    adapt_delta = 0.99
  )
# model-based weighting
cell_table <- fit$data[,c("N","n")]
weights <- model_based_cell_weights(fit, cell_table)
```

```
weights <- data.frame(w_unit = colMeans(weights),
                      cell_id = fit$data[["cell_id"]],
                      Y = fit$data[["Y"]],
                      n = fit$data[["n"]]) %>%
        mutate(
            w = w_unit / sum(n / sum(n) * w_unit), # model-based weights
            Y_w = Y * w
        )
with(weights, sum(n * Y_w / sum(n)))# mean estimate
```

B. Simulation designs

Here we present the simulation designs, coefficient values, and comparison on the weighted distributions of socio-demographics as a supplement to Sections 4 and 5.

Table B.1

Covariates in the outcome (O) and selection (S) models for slightly unbalanced design

	Case 1		Case 2		Case 3		Case 4		Case 5		Case 6		Case 7	
	O	**S**	**O**	**S**	**O**	**S**	**O**	**S**	**O**	**S**	**O**	**S**	**O**	**S**
age	√	√	√	√	√	√	√	√	√	√	√	√	√	√
eth	√	√	√	√	√	√	√	√	√			√		√
edu	√	√	√	√	√		√	√	√		√	√	√	
age*eth	√			√	√	√				√				√
age*edu	√			√	√						√		√	
eth*edu	√			√	√									
age*eth*edu	√			√	√									

Table B.2

Assumed regression coefficients in the outcome model for the simulation using a slightly unbalanced design

	All	Main effects	Two variables
age	(0.5, 1.375, 2.25, 3.125, 4)	(0.5, 1.375, 2.25, 3.125, 4)	(0.5, 1.375, 2.25, 3.125, 4)
eth	(-2, -1, 0, 1, 2)	(2, -1, 0, 1, 2)	$\vec{0}$
edu	(3, 2, 1, 0)	(3, 2, 1, 0)	(3, 2, 1, 0)
age*eth	(4, 2, 1, 1, 3, 3, 2, 1, 1, 2, 3, 2, 2, 1, 4, 4, 3, 2, 3, 2, 4, 1, 4, 1)	$\vec{0}$	$\vec{0}$
age*edu	(-2, -1, 2, 2, 1, -2, 2, 1, 0, -2, 1, -2, -1, 2, 1, -1, -1, 2, 0, 2)	$\vec{0}$	(2, 0, -2, -2, 1, 1, -1, -2, -2, -1, -1, 1, 0, -1, -1, 2, 2, 1, -1, 0)
eth*edu	(1, -2, 0, -3, -1, 0, -1, -2, 0, -1, -3, -3, 0, -1, -1, 0, 0, -1, 0, -1)	$\vec{0}$	$\vec{0}$
age*eth*edu	(-1, -0.5, 0.5, -1, -1, -0.5, -1, 0, -1, 0, -1, 0, 1, 1, 0.5, 1, 1, -1, -1, 0, -1, -0.5, -0.5, -1, 1, -1, -0.5, -1, 1, 0, 0.5, 0.5, 1, 0.5, 1, 1, 1, 0.5, 1, 0, 0, -0.5, 0, 1, -1, -1, 0, -1, -1, -1, -0.5, -0.5, 0, 1, -1, 0, 0, -0.5, 1, -0.5, 0.5, -1, 1, 0, 1, 0, -1, 0, -0.5, 1, -0.5, -1, -0.5, 0, 0.5, -0.5, 1, 0.5, -0.5, 0.5, 0, 1, 0, 1, 0.5, 0.5, 0.5, 0, 0, -0.5, 1, -1, 0, 1, 1, 1, 1, -0.5, -1, -1)	$\vec{0}$	$\vec{0}$

Table B.3
Assumed regression coefficients in the selection model for the simulation using a slightly unbalanced design

	All	Main effects	Two variables
Intercept	-2	-2	-2
age	(-2, -1.75, -1.5, -1.25, -1)	(0, 0.5, 1, 1.5, 2)	(-2, -1.5, -1, -0.5, 0)
eth	(-1, -0.25, 0.5, 1.25, 2)	(-2, -1.5, -1, -0.5, 0)	(-1, -0.5, 0, 0.5, 1)
edu	(0, 0.67, 1.33, 2)	(0, 1, 2, 3)	$\vec{0}$
age × eth	(1, 1, -1, 1, -1, 1, -1, 0, 0, -1, 0, 0, -1, 1, 0, 0, -1, 1, 1, -1, -1, 0, 1, -1, 1)	$\vec{0}$	(-1, 1, 1, 1, -1, -1, -1, 0, -1, -1, -1, -1, 1, -1, -1, 0, 1, 1, -1, 1, -1, -1, 1, 0, 0)
age × edu	(0, 1, -1, -1, 0, 1, 1, 0, 1, 0, 1, -1, -1, 1, 1, -1, 0, -1, 1, 1)	$\vec{0}$	$\vec{0}$
eth × edu	(-1, -1, 0, -1, -1, 1, 1, 1, 1, 0, -1, 0, -1, 0, -1, 1, 0, -1, -1, -1)	$\vec{0}$	$\vec{0}$
age × eth × edu	(0.8, -0.4, 0.6, -0.2, 0.8, 0.2, 0.4, 0.8, 0.4, -0.6, -0.8, -0.4, -0.8, -0.4, 0.4, -1, 0.6, -0.8, -0.6, 0.6, -0.2, 0.2, 0.6, -0.6, 0, 0, -1, -0.2, 0.6, 0.8, -0.4, 0.2, -0.8, 0.4, 0.6, -0.6, 0.8, 0, 0.2, -1, 1, 0.4, 0, 0.8, -0.2, 0, 0, 0.6, -0.8, -0.8, -0.2, 0.4, -1, -0.8, 1, -0.2, 0, 0.8, 0.6, 0.8, -0.2, -0.2, -0.8, 1, 0.8, 0.8, -0.4, -0.8, 0.4, -0.4, 1, -0.6, -1, -0.6, -0.2, 1, 1, -0.2, 1, 0.6, 0.4, 0.8, 0.2, -0.2, -0.6, 0, 0.8, -0.4, 0.4, 0.4, 0.6, -1, -0.8, -0.8, 1, 1, 0.4, 0.6, 0.4, 0.8)	$\vec{0}$	$\vec{0}$

Table B.4
Covariates in the outcome (O) and selection (S) models for a very unbalanced design

	Case 1		Case 2		Case 3		Case 4	
	O	S	O	S	O	S	O	S
age	√	√	√	√	√	√	√	√
eth	√	√	√	√	√	√	√	√
edu	√	√	√	√	√	√	√	√
sex	√	√	√	√	√	√	√	√
pov	√	√	√	√	√	√	√	√
cld		√		√		√	√	√
eld	√	√		√	√	√	√	√
fam	√	√		√	√	√	√	√
age*eth	√	√			√			√
age*edu	√	√			√			√
eth*edu	√	√			√			√
eth*pov	√	√			√			√
age*pov	√	√			√			√
pov*fam	√	√			√			√
pov*eld	√	√			√			√
pov*cld		√						√
age*eth*edu	√	√			√			√
age*eth*pov	√	√			√			√

Table B.5
Assumed regression coefficients in the outcome (O) and selection (S) models for a very unbalanced design

	O	S
age	(2, 0, -2, -2, 1)	(0, 0.75, 1.5, 2.25, 3)
eth	(1, -1, -2, -2, -1)	(-1, -0.5, 0, 0.5, 1)
edu	(-1, 1, 0, -1)	(0, 0.67, 1.33, 2)
sex	(-1, 2)	(-1, 0)
pov	(2, 1, -1, 0, -1)	(0, 1, 2, 3, 4)
cld	$\vec{0}$	(-1, -0.33, 0.33, 1)
eld	$\vec{0}$	(-2, -1, 0)
fam	$\vec{0}$	(-1, -0.67, -0.33, 0)

Table B.6
Euclidean distances between the weighted distributions and the population distribution. Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and PS-W: poststratification weighting

	Str-W	PS-W	Rake-W
age	0.04	0.02	0.00
eth	0.08	0.06	0.00
edu	0.08	0.03	0.00
inc	0.02	0.02	0.00
age * eth	0.05	0.03	0.05
age * edu	0.05	0.02	0.05
age * inc	0.03	0.01	0.03
eth * edu	0.06	0.04	0.05
eth * inc	0.04	0.04	0.03
edu * inc	0.06	0.03	0.04
age * eth * edu	0.03	0.02	0.05
age * eth * inc	0.03	0.02	0.04
age * edu * inc	0.03	0.01	0.04
eth * edu * inc	0.04	0.02	0.04
age * eth * edu * inc	0.02	0.01	0.04

References

ACS Weighting Method (2014). *American Community Survey Design and Methodology*, Chapter 11: Weighting and Estimation. United States Census Bureau.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95, 539-553.

Breidt, F.J. (2008). Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *Annals of Statistics*, 36, 403-427.

Breidt, F., and Opsomer, J. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32, 190-205.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97, 465-480.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 260-269.

Chen, Q., Elliott, M.R., Haziza, D., Yang, Y., Ghosh, M., Little, R., Sedransk, J. and Thompson, M. (2017). Approaches to improving survey-weighted estimates. *Statistical Science*, 32(2), 227-248.

Dahlke, M., Breidt, F., Opsomer, J. and Keilegom, I.V. (2013). Nonparametric endogenous poststratification in surveys. *Statistica Sinica*, 23, 189-211.

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

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Journal of Official Statistics*, 33(1), 23-34.

Elliott, M.R., and Little, R.J. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16(3), 191-209.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a), 269-277.

Firth, D., and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, 60, 3-21.

Fuller, W. (2009). *Sampling Statistics*. Hoboken: John Wiley & Sons, Inc.

Gelman, A. (2005). Analysis of variance: Why it is more important than ever (with discusion). *Annals of Statistics*, 33(1), 1-53.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3, 515-533.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2), 153-164.

Gelman, A., and Carlin, J.B. (2001). Poststratification and weighting adjustments. In *Survey Nonresponse*, (Eds., R. Groves, D. Dillman, J. Eltinge and R. Little).

Gelman, A., and Little, T.C. (1997). Poststratifcation into many cateogiries using hierarchical logistic regression. *Survey Methodology*, 23, 2, 127-135. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997002/article/3616-eng.pdf.

Gelman, A., and Little, T.C. (1998). Improving on probability weighting for household size. *Public Opinion Quarterly*, 62, 398-404.

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.

Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman Hall/CRC Press.

Goodrich, B., and Gabry, J.S. (2017). `rstanarm`: Bayesian applied regression modeling via Stan.https://cran.r-project.org/web/packages/rstanarm/.

Groves, R., and Couper, M. (1995). Theoretical motivation for post-survey nonresponse adjustment in household surveys. *Journal of Offcial Statistics*, 11, 93-106.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling" by D. Basu. In *The Foundations of Survey Sampling*, (Eds., V.P. Godambe and D.A. Sprott), 236. Holt, Rinehart and Winston.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition. Springer.

Henry, K., and Valliant, R. (2012). Comparing alternative weight adjustment methods. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Hoffman, M.D., and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1351-1381.

Holt, D., and Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142(1), 33-46.

Kang, J.D.Y., and Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 523-539.

Kim, J.K., and Skinner, C.J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, 100, 385-398.

Kott, P. (2009). Calibration weighting: Combining probability samples and linear prediction models. In *Handbook of Statistics, Sample Surveys: Design, Methods and Application*, (Eds., D. Pfeffermann and C.R. Rao), Volume 29B. Elsevier.

Little, R. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", by M.H. Hansen, W.G. Madow and B.J. Tepping. *Journal of the American Statistical Association*, 78, 797-799.

Little, R. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

Little, R. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R. (2004). To model or not to model? Competing modes of inference for finite population sampling inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

Little, R. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, 26, 162-174.

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

McConville, K.S., and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2), 389-413.

Park, D.K., Gelman, A. and Bafumi, J. (2005). State-level opinions from national surveys: Poststratification using multilevel logistic regression. In *Public Opinion in State Politics*, (Ed., J.E. Cohen), Standord University Press.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2), 317-337.

Piironen, J., and Vehtari, A. (2016). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. https://arxiv.org/abs/1610.05559.

Potter, F.A. (1988). Survey of procedures to control extreme sample weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 453-458.

Potter, F.A. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 225-230.

Rao, J.N.K. (1966a). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā, Series A*, 28(1), 47-60.

Rao, J.N.K. (1966b). On the relative effciency of some estimators in PPS sampling for multiple characteristics. *Sankhyā, Series A*, 28(1), 61-70.

Rao, J.N.K. (2011). Impact of frequentist and bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science*, 26(2), 240-256.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rasmussen, C.E., and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. MITPress, Cambridge, MA.

Reilly, C., Gelman, A. and Katz, J. (2001). Poststratication without population level information on the poststratifying variable, with application to political polling. *Journal of the American Statistical Association*, 96, 1-11.

Royall, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.

Rubin, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, 63, 581-592.

Rubin, D.B. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", by M.H. Hansen, W.G. Madow and B.J. Tepping. *Journal of the American Statistical Association*, 78, 803-805.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Si, Y., and Gelman, A. (2014). Survey weighting for New York Longitudinal Survey on Poverty Measure. Technical report, Columbia University.

Si, Y., and Zhou, P. (2020). Bayes-raking: Bayesian finite population inference with known margins. *Journal of Survey Statistics and Methodology*, smaa008.

Si, Y., Pillai, N.S. and Gelman, A. (2015). Nonparametric Bayesian weighted sampling inference. *Bayesian Analysis*, 10, 605-625.

Si, Y., Trangucci, R. and Gabry, J.S. (2020). Computation codes for manuscript "Bayesian hierarchical weighting adjustment and survey inference". https://github.com/yajuansisophie/weighting.

Stan Development Team (2017). Stan modeling language user's guide and reference manual.http://mc-stan.org.

Stan Development Team (2018). Stan: A C++ library for probability and sampling. http://mcstan.org.

Tang, X., Ghosh, M., Ha, N.S. and Sedransk, J. (2018). Modeling random effects using global-local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, 0(0),1-14.

Valliant, R., Dever, J.A. and Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*, 2nd Edition. New York: Springer.

Valliant, R., Dorfman, A. and Royall, R. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

Volfovsky, A., and Hoff, P. (2014). Hierarchical array priors for ANOVA decompositions of cross-classified data. *Annals of Applied Statistics*, 8(1), 19-47.

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with nonrepresentative polls. *International Journal of Forecasting*, 31(3), 980-991.

Wimer, C., Garfinkel, I., Gelblum, M., Lasala, N., Phillips, S., Si, Y., Teitler, J. and Waldfogel, J. (2014). Poverty tracker − Monitoring poverty and well-being in NYC. Columbia Population Research Center and Robin Hood Foundation.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185-193.

Xia, X., and Elliott, M.R. (2016). Weight smoothing for generalized linear models using a Laplace prior. *Journal of Official Statistics*, 32(2), 507-539.

Yougov (2017). Introducing the Yougov referendum model. https://yougov.co.uk.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68, 49-67.

Zhang, X., Holt, J.B., Yun, S., Lu, H., Greenlund, K.J. and Croft, J.B. (2015). Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American Journal of Epidemiology*, 182(2), 127-137.