Unifying design-based and model-based sampling inference by estimating a joint population distribution for weights and outcomes*

Andrew Gelman†

9 Aug 2023

Abstract

A well-known rule in practical survey research is to include weights when estimating a population average but not to use weights when fitting a regression model—as long as the regression includes as predictors all the information that went into the sampling weights. But it is not clear how to apply this advice when fitting regressions that include only some of the weighting information, nor does it tell us what to do when analyzing already-collected surveys where the weighting procedure has not been clearly explained or where the weights depend in part on information that is not available in the data. It is also not clear how one is supposed to account for clustering in such analyses. We propose a quasi-Bayesian approach using a joint regression of the outcome and the sampling weight, followed by poststratifcation on the two variables, thus using design information within a model-based context to obtain inferences for small-area estimates, regressions, and other population quantities of interest.

1. Background

1.1. Survey weights

One of the central challenges of statistics is generalizing from sample to population. The natural first step here is to adjust for known, expected, or assumed discrepancies between sample and population[1]—but even this basic level of correction can be challenging, especially when sample and population diverge in many dimensions (for example, age, sex, education, ethnicity, geography, and political affiliation in social surveys).

Weighting is a way to summarize an adjustment: each item in the sample gets a nonnegative weight which is intended to be proportional to its representation in the population. Population estimates can then be obtained as weighted averages of the sample.

Four difficulties arise with classical weighting: construction of weights, uncertainty estimates, small-area estimation, and regression modeling.

Construction of weights is difficult because real-world surveys will require adjustment for many factors, and simple approaches based on poststratification or estimated probabilities of sampling result in highly noisy weights. This in turn motivates more complicated approaches based on smoothing weights or modeling outcomes, which can be done but at the cost of many choices in modeling and estimation.

Standard errors or other uncertainty measures with weighted averages are challenging because a set of weights is sufficient to define a weighted average but does not specify a full probability model; additional assumptions must be added beyond those implied by the weights.

*We thank the U.S. National Science Foundation, National Institutes of Health, and Office of Naval Research for partial support of this work. This is a first draft and it's gotta be full of typos and probably some conceptual errors as well. Also I'd like to come up with a snappier title.

†Department of Statistics and Department of Political Science, Columbia University, New York.

[1]An example of a *known* discrepancy between sample and population would be a sample of 60 women and 40 men that is intended to represent a population that is 52% women and 48% men. An example of an *expected* discrepancy would be clusters sampled with probability proportional to a known measure of size. These discrepancies become *assumed* if the population proportions and sampling probabilities are approximate and not known.

Small-area estimation using weights is difficult because a small area may have so few observations that no weighted estimate will be reasonable. Consider, for example, a national political survey that contains five responses from Wyoming, all of whom support the Republican candidate for president. Any weighted average would result in an obviously wrong estimate of 100% Republican support in the state. Weighting is defeated by data granularity, and modeling is required.

Regression modeling with weights can work in simple settings, replacing least squares or maximum likelihood with weighted versions of these methods, but it becomes more difficult when moving to more advanced multilevel, Bayesian, or regularized methods that are needed to answer complex questions in the presence of data granularity.

This is not to say that weighting-based methods are useless. Much work has gone into population inference using survey weights. Our point here is that there are no generally applicable or easy solutions to the above problem, and so there are theoretical, methodological, and applied reasons for wanting a generally-applicable and unified approach to regression modeling and small-area estimation using survey weights.

1.2. Multilevel regression and poststratification

Multilevel regression and poststratification (MRP) or, more generally, regularized regression and poststratification, is an approach to survey analysis that combines modeling of the data with adjustment for nonrepresentativeness of the sample. In the basic MRP setup, an outcome y and background variables x are observed in the sample, and the distribution of x is known in the population. If the variables in x are discrete, then their interactions define poststratification cells. If the observed data are independently sampled with probabilities of selection that do not vary within poststratification cells, then we can perform population inference by fitting a regression model of y on x and then averaging over the cells in proportion to their known population counts.

So far, this is simply regression and poststratification. The multilevel part comes in because, given the implicit assumption of constant probability of inclusion within cells, there is a desire to poststratify on as many factors as possible, and a regression model with a large number of predictors and interactions cannot be estimated stably using least squares. Multilevel modeling is a good way to fit a regression with many predictors such as arise when modeling survey responses given demographic and geographic factors. Other approaches are possible, hence we can also use the more general term, "regularized regression and poststratification." A key attribute of MRP (or RRP) is that it allows predictions for y given values of x that are not observed in the sample, or which have such small counts in the sample that it would be impossible to make predictions for them from the data alone.

There is a growing literature on MRP and its generalizations. Challenges include obtaining good group-level predictors for multilevel regressions (so that, for example, inferences for small states in a national survey are partially-pooled toward reasonable state-level estimates rather than to a national baseline); adjusting for non-census variables (in which case the population counts of the poststratification cells themselves must be estimated from the data); analyzing cluster samples when the cluster sizes in the population are unknown; and, with particular relevance to the present research, modeling unequal sampling probabilities within poststratification cells. One quick way to incorporate survey weights is to replace the observed mean response within each cell by its weighted mean, but this approach fails when data are sparse and many cells have only a single respondent, in which case important variation in the weights can be missed.

1.3. Analyzing surveys collected by others

Textbooks on survey sampling focus on the scenario in which the data are analyzed by the same team that conducted the survey. There is some literature on the construction of sampling weights, but not much on the analysis of surveys collected by others, even though this is a common mode of social science research. Publicly-available surveys typically come with weights but often do not fully explain how the weighting scheme was chosen or exactly how the weights are computed, hence it can be difficult or impossible to reproduce the procedure starting from the data.

The standard advice for analysis of data collected by others is to use the weights when estimating population averages but not when fitting regressions, as long as all the variables that went into the weighting are included as predictors. This advice is useful where it can be followed, but it does not resolve the question of what to do when fitting a regression whose predictors do not include all the variables that went into the weights. In addition, it is awkward to consider averaging and regression as different problems, given that averaging is a special case of regression. For example, when estimating the average within a subgroup (for example, average responses for women or men), we might simply use weighted average from the relevant group in the sample, but if the subgroup is small enough (for example, individual states or geographic/demographic categories in a national survey), we would want to perform small-area estimation using regression.

The goal in the present paper is to share a general approach to analyzing surveys with weights, under the scenario that the weights have already been constructed before we see the data. This is similar to the idea in multiple imputation for public surveys, in which the organization in charge of the survey uses sophisticated methods to construct imputations, and then users can analyze the imputed datasets, taking the imputations as given. Dividing the problem in two parts—first the construction of the weights, then the analysis of the weighted dataset—entails an inevitable loss of statistical efficiency (except in some special cases), but, as with imputation, offers practical gains of division of labor and facilitates comparability of analyses by different users of the same survey.

2. A quasi-Bayesian approach to regression with survey weights

2.1. Model

Suppose we have a vector of background variables x that are observed in the sample and whose distribution is known in the population, and a weight variable $w > 0$ and scalar outcome y that are known only in the sample. Assume the data have been sampled independently from the population with probabilities inversely proportional to the weights.[2] The poststratification cells $j = 1, \dots, J$ correspond to the possible values of x in the population; we label these as x_j, with N_j being the size of cell j in the population.

Our goal is to perform Bayesian inference for the population values of y, given their known background variables x. Inference for these individual units can then be combined to get inference for the entire population or for subgroups of interest; this is the poststratification step. For example, if we are poststratifying a national poll into 4 age categories, 2 sex categories, 5 education categories, and 50 states, then the number of cells is $J = 4 \cdot 2 \cdot 5 \cdot 50 = 2000$, and the population mean value of the outcome for white people in Alabama, for example, is the weighted average of $\mathrm{E}(y|x_j)$ over the 10 cells corresponding to that group.

If there were no survey weights and we could assume equal-probability sampling, we would simply regress y on x in the sample and then use the fitted models to make predictions (with

[2] We use the term "sampling" here to include all factors relating to inclusion in the sample, including nonresponse.

uncertainty) for the rest of the population. The challenge is that the data are sampled with unequal probabilities. We use the notation p and p_{sample} for the distributions of the population and sample, respectively; that is, we are considering the items in the population to be drawn at random from an infinite superpopulation with distribution $p(y, x, w)$, so that the sample can be considered a draw from the distribution $p_{\text{sample}}(y, x, w) \propto p(y, x, w)/w$.

We handle the problem of unequal sampling probabilities by modeling the process in two steps:

$$\begin{aligned} \text{Model for the outcome:} \quad & p(y|x, w, \theta) \\ \text{Model for the weights:} \quad & p(w|x, \phi) \propto w\, p_{\text{sample}}(w|x, \phi), \end{aligned} \tag{1}$$

where θ and ϕ represent the parameters in the outcome and weight models. Both the models in (1) are conditional on x, which is fine because x is assumed to be known in the population. The advantage of the above formulation is that it makes clear how both models can be estimated from the sample data sample.

In effect, we are poststratifying on (x, w), which requires estimation of $p(w|x)$ so that we can construct the joint distribution of x and w in the population.

Three key aspects of this approach are:

- The outcome model in the first line of (1), which, following the principles of MRP, can include many predictors x and their interactions;
- The adjustment for the sampling weights in the transition from sample to population distributions in the first line of (1), which captures the adjustment information in the weights;
- The adjustment for w is performed using a model, rather than simply reweighting individual data points. Using a model allows the method to work with sparse data, using MRP, the observed data are used to estimate a complete population distribution.

Finally, we assume we are interested in the overall population mean $\overline{Y}$ and functions of the population mean within poststratification cells, $\overline{Y}_j$, and for simplicity we assume an essentially infinite population, so that we can approximate each $\overline{Y}_j$ by its expectation, $\text{E}(y|x_j)$, Our method also applies to finite populations and other summaries (for example, when estimating the variance of attitudes within states, which can be of interest when studying the geography of political polarization). We restrict the focus to means (or proportions, which is a special case of means with a binary outcome) only for simplicity in this short paper.

2.2. Inference and computation

Before getting into details of Bayesian inference, uncertainty, and computation, let us consider how to fit (1) using point estimation. The first step is to regress y on x and w, yielding some $p(y|x, w, \theta)$. The second step is to regress w on x, again using the observed data, thus yielding some $p_{\text{sample}}(w|x, \phi)$. Here we are simply taking θ and ϕ as their point estimates. Next we convert from sample to population distribution,

$$p(w|x, \phi) = \frac{w\, p_{\text{sample}}(w|x, \phi)}{\int w\, p_{\text{sample}}(w|x, \phi) dw}. \tag{2}$$

This latter expression needs to be evaluated for each value of x in the population (that is, for all the poststratification cells), hence the integral in (2) must either be determined analytically or though some fast approximation. Conceptually, though, the problem is now solved: for each

poststratification cell j, we determine $p(w|x_j, \phi)$ from (2) and then average over this distribution to get the predictive distribution of y in cell j:

$$p(y|x_j, \theta, \phi) = \int p(y|w, x_j, \theta) p(w|x_j, \phi) dw. \tag{3}$$

This integral can be performed analytically or else approximated in some way. In any case, we now have estimated the population predictive distribution within each cell and can then poststratify by averaging over the assumed-known cell counts in the population.

Bayesian inference is performed the same way, with the only difference being that inferential inference about θ and ϕ is propagated through (2) and (3). Here is a computational implementation:

1. Fit the model $p(y|x, w, \theta)$ to the sample data; obtain posterior simulations $\theta^s, s = 1, \ldots, S$.
2. Fit the model $p_{\text{sample}}(w|x, \phi)$ to the sample data; obtain posterior simulations $\phi^s, s = 1, \ldots, S$. If θ and ϕ share parameters or are dependent in their prior distribution, these two models would be fit together in one step.
3. For each draw (θ^s, ϕ^s):
   (a) For each poststratification cell j:
      i. Draw weights $w^l, l = l, \ldots, L$ from $p_{\text{sample}}(w|x_j, \phi^s)$
      ii. For each w^l, compute $\text{E}(y|x_j, w^l, \theta^s)$ from the regression model. Then compute $\sum_{l=1}^L w^l \text{E}(y|x_j, \theta^s, \phi^s) / \sum_{l=1}^L w^l$. Label this weighted average as $\widehat{Y}_j^s$; it is a Monte Carlo estimate of $\text{E}(y|x_j, \theta^s, \phi^s)$, the population mean within cell j under the model.
   (b) Compute the population mean $\widehat{Y}^s = \sum_{j=1}^J N_j \widehat{Y}_j^s / \sum_{j=1}^J N_j$ and any subpopulation means or comparisons of interest.
4. Approximate the posterior distribution of all quantities saved in the previous step by their S simulations.

The workflow would then be continued with the usual steps of checking computational accuracy, model fit, and sensitivity, and altering or expanding the model as necessary.

This approach should automatically give stable small-area estimates, as long as the factors defining the small areas are included in x, and as long as multilevel modeling or some other form of regularization is used to fit the regression models in (1). Indeed, this is the main selling point of our approach, that it seamlessly performs weighting adjustment within a modeling context that allows small-area estimation and poststratification.

If there is interest in within-cell population summaries other than averages, then step 3b of the above algorithm must be made more general. Instead of simply computing a weighted average over the draws w^l, we can use Pareto-smoothed importance resampling to draw a subset $M < L$ of these weights with probabilities proportional to w^l. Collect the M resampled draws and renumber them as $w^m, m = 1, \ldots, M$. These approximate a set of draws from the population model, $p(w|x_j, \phi^s)$. For each w^m, we can then continue by sampling one value y from the predictive distribution, $p(y|x_j, w^m, \theta^s)$. We can then complete the process by computing whatever summaries are desired using the M draws of y within that cell.

2.3. Closed-form solution with a lognormal or gamma model for the weights

The algorithm just described has a cumbersome nested design requiring a new draw of $w^1, \ldots, w^L$ for each posterior draw of the model parameters, along with a potentially unstable weighted averaging step.

One way to speed the computation is to use a model for the weights where the denominator of (2) can be evaluated in closed form. One such model is lognormal regression.

Suppose we define $v = \log w$ and fit the model, $p_{\text{sample}}(v|x) = \text{normal}(v|g(x,\beta),\sigma)$, where g is some family of regression functions given parameter vector β, so that $\phi = (\beta, \sigma)$. Then (2) can be written as,

$$p(v|x,\phi) = \frac{e^v p_{\text{sample}}(v|x,\phi)}{\int e^v p_{\text{sample}}(v|x,\phi)dv}, \tag{4}$$

and we can simplify the expression that appears in the numerator and denominator:

$$\begin{aligned} e^v p_{\text{sample}}(v|x,\phi) &= e^v \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(v-g)^2} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}((v-g)^2 - 2\sigma^2 v)} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}((v-(g+\sigma^2))^2 - \sigma^4 - 2g\sigma^2)} \\ &= e^{g+\frac{1}{2}\sigma^2}\, \text{normal}(v \,|\, g+\sigma^2, \sigma), \end{aligned}$$

so that (4) becomes,

$$\begin{aligned} p(v|x,\phi) &= \frac{e^{g+\frac{1}{2}\sigma^2}\, \text{normal}(v \,|\, g+\sigma^2, \sigma)}{e^{g+\frac{1}{2}\sigma^2}} \\ &= \text{normal}(v \,|\, g+\sigma^2, \sigma). \end{aligned}$$

Thus, under the lognormal model, the population distribution of the weights is identical to the sample distribution except that it is shifted to the right by σ^2. This makes sense. First, a large weight corresponds to more representation in the population, so we should expect higher weights to be more common in the population than in the sample. Second, σ^2 is the residual variance of the log weights, so the higher the value of σ, the more consequential will be the weighting, hence the larger the shift. At the extreme of $\sigma = 0$, the weights do not vary within poststratification cells at all, and no adjustment is needed.

The above calculation took advantage of a conjugacy property of e^v with the normal density. Closed-form computation is available under other models as well. For example, if the weights follow a gamma regression, then multiplying the density function by w has the effect of adding 1 to the shape parameter of the model and correspondingly shifting the mean upward, so that if the mean of the distribution of $p_{\text{sample}}(w|x,\phi)$ in the sample is g, then the mean in the population distribution $p(w|x,\phi)$ becomes $\frac{\alpha+1}{\alpha}g$, which again makes sense, both in that it is an increase compared to the sample and that the increase goes to zero in the limit of $\alpha \to \infty$, which corresponds to a gamma distribution with zero variance.

2.4. Closed-form solution with a mixture of lognormals or gammas

For various reasons, the distribution of weights can be far from normal or gamma. But we can retain the clean computation of these conjugate forms using a mixture model. We demonstrate with the lognormal.

Start with the model, $p_{\text{sample}}(v|x,\phi) = \sum_{k=1}^{K} \lambda_k \,\text{normal}(v|g_k(x,\beta),\sigma_k)$. Then similar algebra as above yields the population distribution,

$$p(v|x,\phi) = \frac{\sum_{k=1}^{K} e^{g_k+\frac{1}{2}\sigma_k^2}\,\lambda_k \,\text{normal}(v \,|\, g_k+\sigma_k^2,\sigma_k)}{\sum_{k=1}^{K} e^{g_k+\frac{1}{2}\sigma_k^2}\,\lambda_k},$$

which again is a mixture of lognormals. In addition to each mean being shifted by σ^2 as before, the mixture proportions change, with modes with higher values counting more in the population, which makes sense.

2.5. Using the closed-form models as approximations

In practice it may be enough to model the weights using lognormal or gamma error terms or perhaps mixtures of one of these. But if a more general model is desired, we should be able get much of the computational benefits by first fitting the closed-form model and then using it as an approximation for the desired model. Instead of importance ratios w^l in step 3b of our algorithm, we could use the ratio of the exact and approximate densities, which should be more stable.

3. Concerns

3.1. Unrealistic assumptions of the model

We call our method quasi rather than fully Bayesian because it is based on a generative model in which the weights w are defined in the population and are drawn to create the sample, but in real surveys the weights are constructed from the sample and do not have a population distribution to be estimated. In that way, our approach is similar to many applications of statistics in which a probability model is used even in the absence of any superpopulation or physical randomization.

Our model assumes independent sampling with probabilities proportional to $1/w$, but survey weights are often constructed by raking and do not represent sampling probabilities at all. Even when weights are intended to represent inverse sampling probabilities, they generally do not because the construction of weights is only approximate.

Why would we purposely construct a model that is wrong in these crucial ways? The short answer is that, to the extent that weights are well constructed in a practical sense and used as intended, an item with weight w in the sample is intended to represent w items in the population. Our procedure can be viewed as a smoothed version of applying weights to items in the sample. To the extent that it is intended to be a model-based adaptation or generalization of existing practice, it makes sense to take the weights seriously and consider them as being inversely proportional to the probability of inclusion in the sample, even if they are not. Similarly, the assumption of independent sampling can be viewed as an instantiation of the recommended methods in which weights are attached to individual units. As we have written elsewhere, we fit a model consistent with standard practice because we want our approach to be an improvement upon rather than merely a replacement for standard weighted analysis of sample surveys.

Another potential concern is the use of a regression, $p(w|x)$ that implies a continuous distribution of weights in the population, even though weights in real surveys typically take on only a finite possible number of values. With a reasonable model, this should not be such an issue—a lognormal regression should be a reasonable fit to weights that are constructed by multiplying many individual factors, mixture modeling can capture the discreteness that can arise if weights are dominated by one or two factors, also to the extent that the weights depend on variables in x, much of their

variation will be explained by the deterministic part of the weighting regression anyway—but it can represent a modeling challenge.

3.2. Sensitivity to large weights

In addition, we need to be concerned about the right tail of the fitted weight distribution, for two reasons. First, survey weights are often smoothed or trimmed to reduce variability, which can make sense as a variance-reduction tool but invalidates their interpretation as inverse probabilities of selection. Second, large weights correspond to low sampling probabilities, so they represent "dark matter" in the sampling procedure: potentially large chunks of the population that are expected to appear rarely or not at all in the data. Any resolution of this problem requires strong assumptions, such as a hard cap on the maximum weight in the population or a short upper tail that limits the total proportion of the population that would have large weights. In a finite-population analysis there is also a bound on the other end, because the probability of inclusion in the sample can never exceed 1.

When considering these aspects of sensitivity to model assumptions, remember that the ultimate goal is to estimate $p(y|x)$; the weights are just a means to this end, a way of adjusting for the biases that would occur if one were to attempt to extrapolate from a fitted model without adjusting for known discrepancies between sample and population. What is relevant, then, is the dependence of $p(y|x, w)$ on w. If this model is a smooth function of w, then approximating a discrete distribution of w by a continuous distribution might not cause serious problems. If the model behaves calmly for large values of w, then the "dark matter" problem of very large weights in the population might not be such a concern. It should be possible to do some theoretical analysis, looking at the tails of the model for w along with the functional form of $p(y|x, w)$ for large w to ensure bounded influence from the unobserved items with large weights.

4. Theoretical examples

We can work through some simple simulated-data examples to understand where the method works and where it breaks. Here are some ideas:

- No background variables x, only weights w, so the goal is to estimate the population mean. How does our approach compare to the simple weighted average?
- Weights w that depend entirely on the background variables x. Our approach should be identical to unweighed MRP.
- Simple stratified sample with weights; then perform the analysis ignoring the strata. How does this differ from the standard stratified analysis? How does it differ from a weighted-average analysis ignoring strata.
- Finite-population sample including a certainty stratum and thus a hard lower bound on weights.
- Poststratification weights modeled as inverse-probability weights: how much does our approach inflate the variance estimate compared to the correct poststratification analysis?
- Simple small-area estimation without or with a group-level predictor. Result will depend on the dependence between weights and expected outcome, so try different possibilities in the simulations.

- Small-area estimation with a huge number of cells so $n_j = 0$ or 1 in almost all cells and there is no observable variation in weights within each cell, but the weights still matter.

These examples raise conceptual challenges. Consider, for example, a national survey that is poststratified by geography, in such small areas (for example, zip codes) that there are no cells with more than one respondent in the sample. Also suppose that the survey weights are not based on geography but are instead based on the number of people living in the respondent's household, a variable that is not otherwise included in the analysis. The weights will still vary by geography even though they are not defined explicitly in geographic terms. But with only one respondent per poststratification cell in the data we cannot estimate the within-cell variance in log weights (the crucial parameter σ^2 in model (5)), and the only way forward, short of including the "number of people in the household" variable in the analysis, might be to combine cells to allow the estimation of within-cell variation of the weights.

Add some structure to the problem, though, and it becomes easier to solve. Take the same example, with the same number of poststratification cells, but suppose they are formed by the intersection of several variables, for example age, sex, ethnicity, education, and congressional district. In this case, a multilevel regression of log weight on these factors will yield a nonzero residual variance, as long as the model does not include the fully-saturated interaction of all the predictors.

5. Applications

We want as soon as possible to apply our method to real surveys. Two examples we have immediately at hand are the Cooperative Election Survey and Pew Research pre-election polls. Both these surveys include weights, and we have applied MRP to both of them in the past. We can also see how our method works when including post-election weighting based on vote preference.

We can also look into the Fragile Families Study, a survey of at-risk births for which we were involved in construction of the weights.

Evaluating the method in applied examples can be difficult because for most survey questions we do not know the true population values. One setting where we do know the truth, and which we have used to evaluate MRP in the past, is U.S. election polling; however, challenges arise there too given problems of differential nonresponse by party.

That said, we still think much can be learned by applying our procedure to real problems. The method could run into computational problems, it could give completely unreasonable results, the model could have problems fitting to the data. More positively, we could get a sense of distributions of weights in real surveys and compare different approaches to small-area estimation and regression modeling in the presence of survey weights.

We can also conduct simulation studies by subsampling from real survey data. In that case, the "population" (a large existing survey) is completely known, we have full control over the sampling procedure, we can define weights however we want, and we can compare our inferences to the population values, checking accuracy of estimates and coverage of uncertainty intervals.

6. Looking forward

Some loose ends in our procedure include model checking, workflow, and sensitivity to large weights in the population.

Some future challenges include inference with known margins (poststratification with marginal or lower-dimensional joint distributions, for example post-election adjustments based on local vote

totals); cluster sampling; inference for non-census variables (for example, religion); and inverse-probability weighting in causal inference.

For now, our practical recommendations depend on where you stand in the process of data preparation and analysis:

- If you have access to the raw data and relevant population information: Then we would not typically recommend the methods in this paper. Instead of creating weights and then incorporating them into the analysis, it should be better to just model the data directly conditional on all information that might go into weighting. This might require augmenting the poststratification table (if there are relevant non-census variables: information predictive of the outcome and predictive of inclusion in the sample that is not available at the population level), but that modeling could be done directly, with no need to create survey weights as intermediate quantities.

- If you have conducted a survey and want to create weights for others to use: In this case it could make sense to anticipate the methods discussed in the present paper when forming the weights. It could help future users of the survey if the weights contain relevant information to help the model-based analysis perform well. Some research is needed here, given that probabilities of inclusion in the sample are generally not known, only estimated, and also given that the goal is adjustment to the population, not estimation of inclusion probabilities.

- If you are analyzing a survey collected by others where the weights have been supplied: Here, we hope our theoretical and applied examples give some sense of when it would make sense to follow the approach presented here.

Our theoretical and practical challenge is to design a procedure that unifies existing design-based weighting methods and existing model-based approaches to small-area estimation. We will learn more when we try out the method on some examples.

References

Ansolabehere, S., Schaffner, B., and Luks, S. (2019). Guide to the 2018 Cooperative Congressional Election Survey. `https://cces.gov.harvard.edu/`

Ben-Michael, E., Feller, A., and Hartman, E. (2023). Multilevel calibration weighting for survey data. *Political Analysis*.

Bisbee, J. (2019). BARP: Improving Mister P using Bayesian additive regression trees. *American Political Science Review* **113**, 1060–1065.

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics* **29**, 329–353.

Brick, J. M., and Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics* **33**, 735–752.

Broniecki, P., Leemann, L., and Wüest, R. (2021). Improved multilevel regression with post-stratification through machine learning (autoMrP). *Journal of Politics* **84**, 597–601.

Chen, Ch., Duan, N., Meng, X. L., and Alegria, M. (2006). Power-shrinkage and trimming: Two ways to mitigate excess weights. *Proceedings of Section on Survey Research Methods, American Statistical Association*, 2839–2846.

Chen, Ci., Wakefield, J., and Lumley, T. (2014). The use of sampling weights in Bayesian hierarchical models for small-area estimation. *Spatial and Spatio-temporal Epidemiology* **11**, 33–43.

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

DuMouchel, W. H., and Duncan, G. J. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association* **78**, 535–543.

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

Elliott, M. R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science* **32**, 249–264.

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**, 269–277.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science* **22**, 153–188.

Gelman, A. (2018). Regularized prediction and poststratification. Statistical Modeling, Causal Inference, and Social Science, 19 May. `https://statmodeling.stat.columbia.edu/2018/05/19/regularized-prediction-poststratification-generalization-mister-p/`

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*, third edition. London: CRC Press.

Gelman, A., and Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* **23**, 127–135.

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

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P., and Modrák, M. (2020). Bayesian workflow. `https://arxiv.org/abs/2011.01808`

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**, 762–776.

Gopelrud, M. (2023). Re-evaluating machine learning for MRP given the comparable performance of (deep) hierarchical models. *American Political Science Review*.

Graubard, B. I., and Korn E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science* **17**, 73–96.

Haziza, D., and Beaumont, J. F. (2017). Construction of weights in surveys: A review. *Statistical Science* **32**, 206–226.

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

Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review* **51**, 175–188.

Kennedy, L. A., and Gelman, A. (2020). Year 15 Fragile Families survey weight adjustment. Princeton Center for Research on Child Wellbeing. `https://fragilefamilies.princeton.edu/sites/g/files/toruqf2001/files/ff_const_wgtsy15.pdf`

Kennedy, L. A., and Gelman, A. (2021). Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample. *Psychological Methods* **26**, 547–558.

Kish, L., (1992). Weighting for unequal P_i. *Journal of Official Statistics* **8**, 183–200.

Korn, E. L., and Graubard, B. I. (1999). *Analysis of Health Surveys.* New York: Wiley.

Kuriwaki, S., Ansolabehere, S., Dagonel, A., and Yamauchi, S. (2023). The geography of racially polarized voting: Calibrating surveys at the district level. *American Political Science Review.*

Lei, R., Gelman, A., and Ghitza, Y. (2017). The 2008 election: A preregistered replication analysis. *Statistics and Public Policy* **4** (1), 1–8.

Léon-Novelo, L. G., and Savitsky, T. D. (2019). Fully Bayesian estimation under informative sampling. *Electronic Journal of Statistics* **13**, 1609–1645.

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

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

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

Little, T. C., and Gelman, A. (1998). Modeling differential nonresponse in sample surveys. *Sankhya* **60**, 101–126.

Lohr, S. (2022). *Sampling: Design and Analysis*, third edition. London: CRC Press.

Lopez-Martin, J., Phillips, J. H., and Gelman, A. (2022). Multilevel regression and poststratification case studies. `https://bookdown.org/jl5522/MRP-case-studies/`

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software* **9** (8), 1–19.

Lumley, T., and Scott, A. (2017). Fitting regression models to survey data. *Statistical Science* **32**, 265–278.

Makela, S., Si, Y., and Gelman, A. (2018). Bayesian inference under cluster sampling with probability proportional to size. *Statistics in Medicine* **37**, 3849–3868.

Miratrix, L. W., Sekhon, J. S., and Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society B* **75**, 369–396.

O'Muircheartaigh, C., and Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society C* **63**, 195–210.

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

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B* **60**, 23–40.

Potthoff, R., Woodbury, M., and Manton, K. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association* **87**, 383–396.

Rao, J. N. K. (2003). *Small Area Estimation.* New York: Wiley.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91**, 473–520.

Särndal, C. E. (1978). Design-based and model-based inference in survey sampling (with discussion). *Scandinavian Journal of Statistics* **5**, 27–52.

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

Si, Y., Trangucci, R., Gabry, J., and Gelman, A. (2020). Bayesian hierarchical weighting adjustment and survey inference. *Survey Methodology* **46**, 181–214.

Solon, G., Haider, S. J., and Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human Resources* **50**, 301–316.

Stanek, E. J., and Singer, J. M. (2004). Predicting random effects from finite population clustered samples with response error. *Journal of the American Statistical Association* **99**, 1119–1130.

Su, Y. S., and Gelman, A. (2023). Who wants school vouchers in America? A comprehensive study using multilevel regression and poststratification. *Social Sciences* **12**, 430.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2015). Pareto smoothed importance sampling. `https://arxiv.org/abs/1507.02646`

Voss, D. S., Gelman, A., and King, G. (1995). Pre-election survey methodology: Details from nine polling organizations, 1988 and 1992. *Public Opinion Quarterly* **59**, 98–132.

Winship, C., and Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods and Research* **23**, 230–257.

Xie, H., Barker, L. E., and Rolka, D. B. (2020). Incorporating design weights and historical data into model-based small-area estimation. *Journal of Data Science* **18**, 115–131.