MRP for Statistical Data Integration and Inferences

Yajuan Si
Research Assistant Professor
Institute for Social Research, University of Michigan

April 4, 2020

Acknowledgements

- Grant support from NSF-SES 1760133
- Organizing effort by Lauren Kennedy
- Inspiration and mentorship from Andrew Gelman

Outline

1. Overview and motivation
2. Methodology and practice in survey research
3. Recent developments and challenges

1. Overview and Motivation

MRP is a statistical method

HOME BOOKS BLOGROLL SPONSORS

« Ed Sullivan (3) vs. Sid Caesar; DJ Jazzy Jeff advances

Babe Didrikson Zaharias (2) vs. Adam Schiff; Sid Caesar advances »

MRP (multilevel regression and poststratification; Mister P): Clearing up misunderstandings about

Posted by Andrew on 10 January 2019, 9:50 am

Someone pointed me to this thread where I noticed some issues I'd like to clear up:

David Shor: "MRP itself is like, a 2009-era methodology."

Nope. The first paper on MRP was from 1997. And, even then, the component pieces were not new: we were just basically combining two existing ideas from survey sampling: regression estimation and small-area estimation. It would be more accurate to call MRP a methodology from the 1990s, or even the 1970s.

Will Cubbison: "that MRP isn't a magic fix for poor sampling seems rather obvious to me?"

Yep. We need to work on both fronts: better data collection and better post-sampling adjustment. In practice, neither alone will be enough.

David Shor: 2012 seems like a perfect example of how focusing on correcting non-response bias and collecting as much data as you can is going to do better than messing around with MRP.

There's a misconception here. "Correcting non-response bias" is not an alternative to MRP; rather, MRP is a method for correcting non-response bias. The whole point of the "multilevel" (more generally, "regularization") in MRP is that it allows us to adjust for more factors that could drive nonresponse bias. And of course we used MRP in our paper where we showed the importance of adjusting for non-response bias in 2012.

And "collecting as much data as you can" is something you'll want to do no matter what. Yair used MRP with tons of data to understand the 2018 election. MRP (or, more generally, RRP) is a great way to correct for non-response bias using as much data as you can.

Also, I'm not quite clear what was meant by "messing around" with MRP. MRP is a statistical method. We use it, we don't "mess around" with it, any more than we "mess around" with any other statistical method. Any method for correcting non-response bias is going to require some "messing around."

In short, MRP is a method for adjusting for nonresponse bias and data sparsity to get better survey estimates. There are other ways of getting to basically the same answer. It's important to adjust for as many factors as possible and, if you're going for small-area estimation with sparse data, that you use good group-level predictors.

MRP is a 1970s-era method that still works. That's fine. Least squares regression is a 1790s-era method, and it still works too! In both cases, we continue to do research to improve and better understand what we're doing.

2

Filed under Multilevel Modeling, Political Science, Teaching, Zombies

Comment (RSS) | Permalink

What problems does MRP address?

1. **Poststratification** adjustment for selection bias. Correct for imbalances in sample composition, even when these are severe and can involve a large number of variables.
2. **Multilevel Regression** for small area estimation (SAE). Can provide stabilized estimates for subgroups over time (such as states, counties, *etc.*)

Two key assumptions under MRP

1. Equal inclusion probabilities of the individuals within cells.
2. The included individuals are similar to those excluded within cells.

2. Methodology and practice

Unify design-based and model-based inferences

- The underlying theory is grounded in survey inference: a combination of small area estimation (Rao and Molina 2015) and poststratification (Holt and Smith 1979).
- Motivated by R. Little (1993), a model-based perspective of poststratification.
- Suppose units in the population and the sample can be divided into J poststratification cells with population cell size N_j and sample cell size n_j for each cell $j = 1, \dots, J$, with $N = \sum_{j=1}^{J} N_j$ and $n = \sum_{j=1}^{J} n_j$.
- Let $\overline{Y}_j$ be the population mean and $\bar{y}_j$ be the sample mean within cell j. The proposed MRP estimator is,

$$\tilde{\theta}^{\mathrm{mrp}} = \sum_{j=1}^{J} \frac{N_j}{N} \tilde{\theta}_j,$$

where $\tilde{\theta}_j$ is the model-based estimate of $\bar{Y}_j$ in cell j.

Compare with unweighted and weighted estimators

1. The unweighted estimator is the average of the sample cell means,

$$\bar{y}_s = \sum_{j=1}^{J} \frac{n_j}{n} \bar{y}_j. \tag{1}$$

2. The poststratification estimator accounts for the population cell sizes as a weighted average of the sample cell means,

$$\bar{y}_{ps} = \sum_{j=1}^{J} \frac{N_j}{N} \bar{y}_j. \tag{2}$$

Bias and variance

Let the poststratification cell inclusion probabilities, means for respondents and nonrespondents be ψ_j, $\bar{Y}_{jR}$ and $\bar{Y}_{jM}$, respectively.

$$\mathrm{bias}(\bar{y}_s) = \sum \frac{\frac{N_j}{N}\bar{Y}_{jR}(\psi_j - \bar{\psi})}{\bar{\psi}} + \sum \frac{N_j}{N}(1-\psi_j)(\bar{Y}_{jR} - \bar{Y}_{jM}) \doteq A + B$$

$$\mathrm{bias}(\bar{y}_{ps}) = \sum \frac{N_j}{N}(1-\psi_j)(\bar{Y}_{jR} - \bar{Y}_{jM}) = B$$

$$Var(\bar{y}_s|\vec{n}) = \sum_j \frac{n_j}{n^2} S_j^2$$

$$Var(\bar{y}_{ps}|\vec{n}) = \sum_j \frac{N_j^2}{N^2}(1 - n_j/N_j)\frac{S_j^2}{n_j}$$

Partial pooling with MRP

- Introduce the exchangable prior, $\theta_j \sim N(\mu, \sigma_\theta^2)$.

- The approximated MRP estimator is given by

$$\tilde{\theta}^{\mathrm{mrp}} = \sum_{j=1}^{J} \frac{N_j}{N} \frac{\bar{y}_j + \delta_j \bar{y}_s}{1 + \delta_j}, \text{ where } \delta_j = \frac{\sigma_j^2}{n_j \sigma_\theta^2}, \tag{3}$$

  as a weighted combination of $\bar{y}_s$ and $\bar{y}_{ps}$, where the weight is controlled by $(n_j, \sigma_\theta^2, \sigma_j^2)$.

- The bias and variance trade-off for the MRP estimator (Si 2020, under review)

The key steps

1. **Multilevel regression** Fit a model relating the survey outcome to covariates across poststratification cells to estimate θ_j;
2. **Poststratification** Average the cell estimates weighted by the population cell count N_j; or **Prediction** Impute the survey outcomes for all population units.

A unified MRP framework

- "Survey weighting is a mess" (Gelman 2007).
- It depends on the goal of weighting adjustments (Bell and Cohen 2007; Breidt and Opsomer 2007; R. J. A. Little 2007; Lohr 2007; Pfeffermann 2007)
- Our goal is to unify design-based and model-based inference approaches as *data integration* to
    - Combine weighting and prediction
    - Unify inferences from probability- and nonprobability-based samples
- **Key quantities** : $j = 1, \ldots, J$, θ_j and N_j

Bayesian Nonparametric Weighted Sampling Inference (Si, Pillai, and Gelman 2015)

- Consider independent sampling with unequal inclusion probabilities.
- The externally-supplied weight is the only information available.
- **Assume the unique values of unit weights determine the poststratification cells via a 1-1 mapping.**
- Simultaneously predict $w_{j[i]}$'s and y_i's for $N-n$ nonsampled units.

Incorporate weights into modeling

1. We assume n_j's follow a multinomial distribution conditional on n,

$$\vec{n} = (n_1, \ldots, n_J) \sim \text{Multinomial}\left(n; \frac{N_1/w_1}{\sum_{j=1}^J N_j/w_j}, \ldots, \frac{N_J/w_J}{\sum_{j=1}^J N_j/w_j}\right).$$

   Here N_j's are unknown parameters.

2. Let $x_j = \log w_j$. For a continuous survey response y, by default

$$y_i \sim \mathrm{N}(\mu(x_{j[i]}), \sigma^2),$$

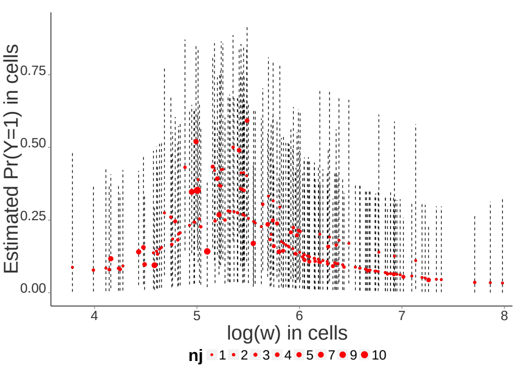
   where $\mu(x_j)$ is a mean function of x_j.

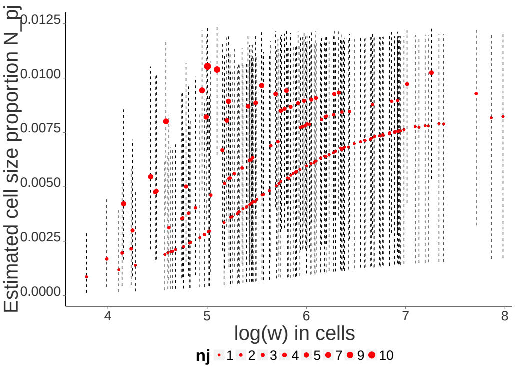
3. We introduce a Gaussian process (GP) prior for $\mu(\cdot)$

$$\mu(x) \sim \text{GP}(x\beta, \Sigma_{xx}),$$

   where Σ_{xx} denotes the covariance function of the distances for any $x_j, x_{j'}$.

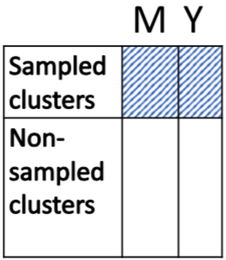
Estimates of cell means and cell size proportions





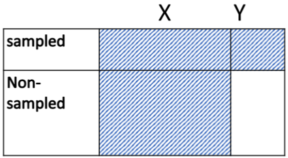
Proportion estimation of individuals with public support based on the Fragile Families and Child Wellbeing Study.

Bayesian inference under cluster sampling with probability proportional to size (Makela, Si, and Gelman 2018)



- Bayesian cluster sampling inference is essentially outcome prediction for nonsampled units in the sampled clusters and all units in the nonsampled clusters.
- However, the design information of nonsampled clusters is missing, such as the measure size under PPS.
- Predict the unknown measure sizes using Bayesian bootstrap and size-biased distribution assumptions.
- Account for the cluster sampling structure by incorporation of the measure sizes as covariates in the multilevel model for the survey outcome.

Bayesian hierarchical weighting adjustment and survey inference (Si et al. 2020)



- Handle deep interactions among weighting variables
- The population cell mean θ_j is modeled as

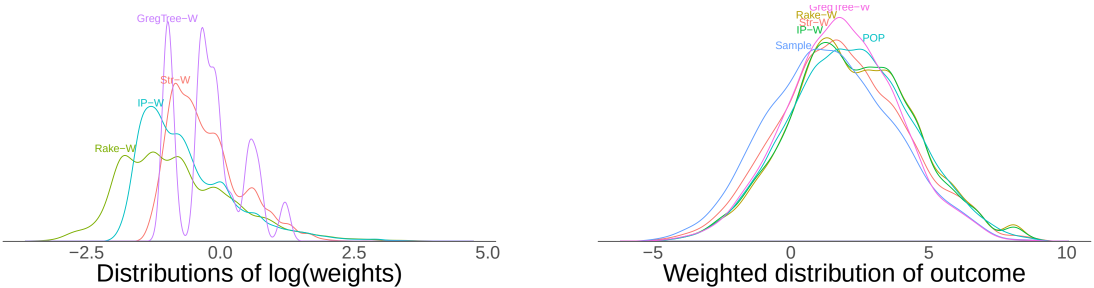
$$\theta_j = \alpha_0 + \sum_{k \in S^{(1)}} \alpha_{j,k}^{(1)} + \sum_{k \in S^{(2)}} \alpha_{j,k}^{(2)} + \cdots + \sum_{k \in S^{(q)}} \alpha_{j,k}^{(q)}, \qquad (4)$$

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.

- Introduce structured prior distribution to account for the hierarchical structure and improve MrP under unbalanced and sparse cell structure.
- Derive the equivalent unit weights in cell j that can be used classically

$$w_j \approx \frac{n_j/\sigma_y^2}{n_j/\sigma_y^2 + 1/\sigma_\theta^2} \cdot \frac{N_j/N}{n_j/n} + \frac{1/\sigma_\theta^2}{n_j/\sigma_y^2 + 1/\sigma_\theta^2} \cdot 1, \qquad (5)$$

Model-based weights and predictions



The model-based weights are stable and yield efficient inference. Predictions perform better than weighting with the capability to recover empty cells.[1]

[1] Greg-tree is based on the tree-based method in McConville and Toth (2017)

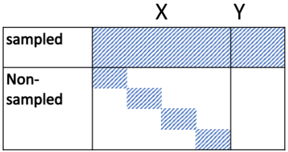
Stan fitting under structured prior in rstanarm

```
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
  )
```

Generated model-based weights

```
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
           )
```

Bayesian raking estimation (Si and Zhou 2020)



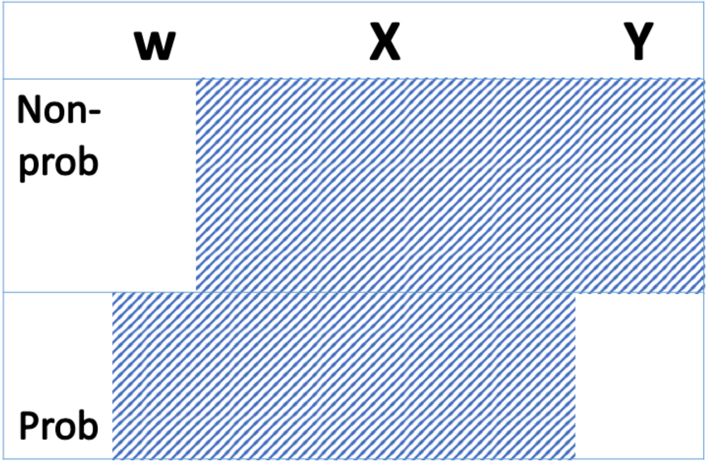
- Often the margins of weighting variables are available, rather than the crosstabulated distribution
- The iterative proportional fitting algorithm suffers from convergence problem with a large number of cells with sparse structure
- Incorporate the marginal constraints via modeling
- Integrate into the Bayesian paradigm, elicit informative prior distributions, and simultaneously estimate the population quantity of interest

3. Recent developments and challenges

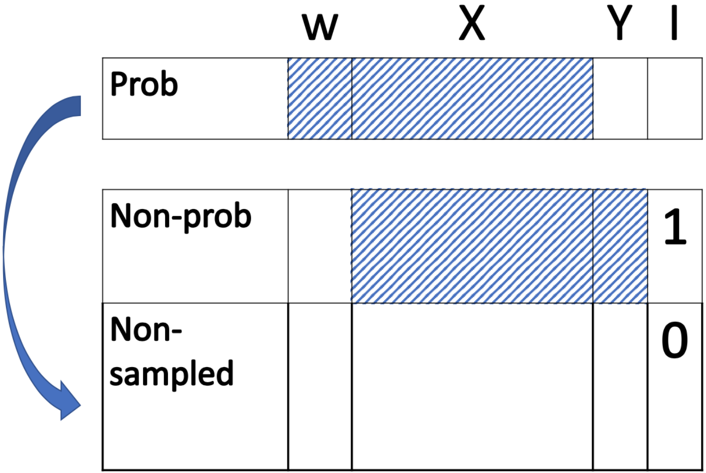
Structural, spatial, temporal prior specification

- We developed structured prior distributions to reflect the hierarchy in deep interactions (Si et al. 2020)
- Sparse MRP with LassoPLUS (Goplerud et al. 2018)
- Use Gaussian Markov random fields as a prior distribution to model certain structure of the underlying categorical covariate (Gao et al. 2019)
- Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion (Gelman et al. 2019)

Data integration and inferences with probability and nonprobability samples



More formally



MRP framework for data integration (Si 2020, under review)

1. Under the quasi-randomization approach, we assume the respondents within poststratum h are treated as a random sample of the population stratum cases,

$$\vec{n} = (n_1, \ldots, n_J)' \sim \mathrm{Multinomial}((cN_1\psi_1, \ldots, cN_J\psi_J), n), \tag{6}$$

   where $c = 1/\sum_j N_j\psi_j$, and the poststratification cell inclusion probabilities $\psi_j = g^{-1}(Z_j\alpha)$. With noninformative prior distributions, this will be equivalent to Bayesian bootstratp.

2. Under the super-population modeling, we assume the outcome follows a normal distribution with cell-specific mean and variance values, and the mean functions are assigned with a flexible class of prior distributions

$$\begin{aligned} y_{ij} &\sim N(\theta_j(\psi_j), \sigma_j^2) \\ \theta_j(\psi_j) &\sim f(\mu(\psi_j), \Sigma_\Psi) \end{aligned} \tag{7}$$

Manuscripts in preparation

- Noncensus variables in poststratification
- Adjust for selection bias in analytic modeling
- Compare MRP estimator with doubly robust estimators
- . . . .

Challenges

- Robust model specification for complicated data
- Multiple (types of) survey variables
- Missing not at random/non-ignorable/informative selection
- External validation
- Incorporate substantive knowledge

Thank you

yajuan@umich.edu

References

Bell, Robert M., and Michael L. Cohen. 2007. "Comment: Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 165–67.

Breidt, F. Jay, and Jean D. Opsomer. 2007. "Comment: Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 168–70.

Gao, Yuxiang, Lauren Kennedy, Daniel Simpson, and Andrew Gelman. 2019. "Improving Multilevel Regression and Poststratification with Structured Priors." https://arxiv.org/abs/1908.06716.

Gelman, Andrew. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 153–64.

Gelman, Andrew, Jeffrey Lax, Justin Phillips, Jonah Gabry, and Robert Trangucci. 2019. "Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion." http://stat.columbia.edu/ gelman/research/unpublished/MRT(1).pdf.

Goplerud, Max, Shiro Kuriwaki, Marc Ratkovic, and Dustin Tingley. 2018. "Sparse Multilevel Regression and Poststratification." https://scholar.harvard.edu/files/dtingley/files/sparsemultilevel.pdf.

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

Little, R. J. A. 1993. "Post-Stratification: A Modeler's Perspective." *Journal of the American Statistical Association* 88: 1001–12.

Little, Roderick J. A. 2007. "Comment: Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 171–74.

Lohr, Sharon L. 2007. "Comment: Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 175–78.

Makela, Susanna, Yajuan Si, and Andrew Gelman. 2018. "Bayesian Inference Under Cluster Sampling with Probability Proportional to Size." *Statistics in Medicine* 37 (26): 3849–68.

McConville, Kelly S., and Daniell Toth. 2017. "Automated Selection of Post-Strata Using a Model-Assisted Regression Tree Estimator." https://arxiv.org/abs/1712.05708.

Pfeffermann, Danny. 2007. "Comment: Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 179–83.

Rao, J.N.K., and Isabel Molina. 2015. *Small Area Estimation*. John Wiley & Sons, Inc.

Si, Yajuan, and Peigen Zhou. 2020. "Bayes-Raking: Bayesian Finite Population Inference with Known Margins." *Journal of Survey Statistics and Methodology* Forthcoming