Multilevel Regression and Poststratification

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

October 10, 2019

Acknowledgements

- Organizing effort by James Wagner (Univ. of Michigan), ASA, SRMS
- Grant support from NSF-SES 1760133
- Comments and partial materials shared by
  - Andrew Gelman (Columbia University)
  - Lauren Kennedy (Columbia University)
  - Jonah Gabry (Columbia University)
  - Douglas Rivers (Stanford University)

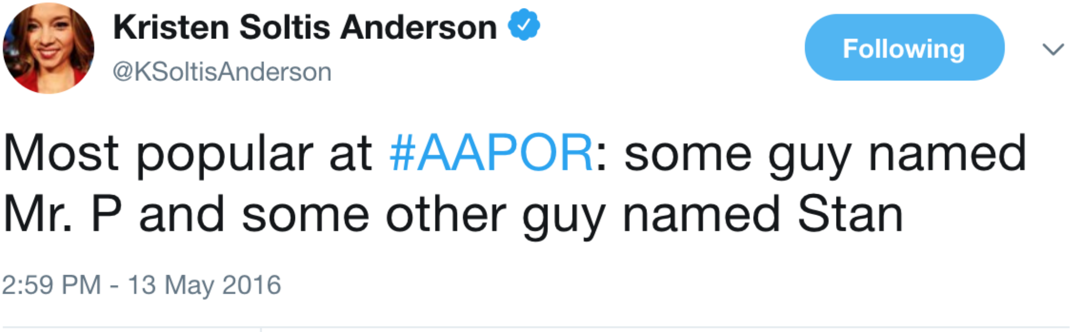
Outline

1. Overview and examples
2. Methodology and practice
3. Applications in survey research
4. Recent developments and challenges

- We will have a 10-min break at 2pm EST.
- All materials can be downloaded from Github: `https://github.com/yajuansi-sophie/MrP-presentations`

1. Overview and Examples

What is MRP?



Formally, **M**ultilevel **R**egression and **P**ost-stratificaiton

Informally, **Mr. P**

Behind MRP

Andrew Gelman

Gelman proposed MRP (A. Gelman and Little 1997) and has demonstrated its success in public opinion research, especially on subgroup and trend analysis, e.g., Ghitza and Gelman (2013); Shirley and Gelman (2015).

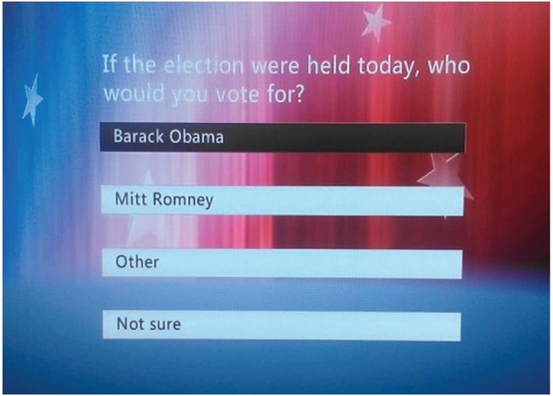
Stan made MRP generally accessible as an open source software project for statistical modeling and high-performance statistical computation.

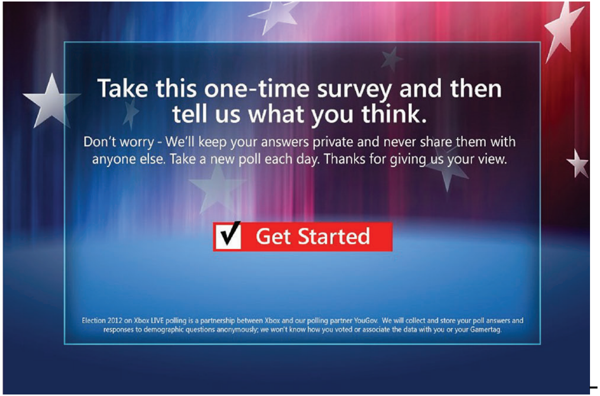
What problems does MRP address?

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.*)

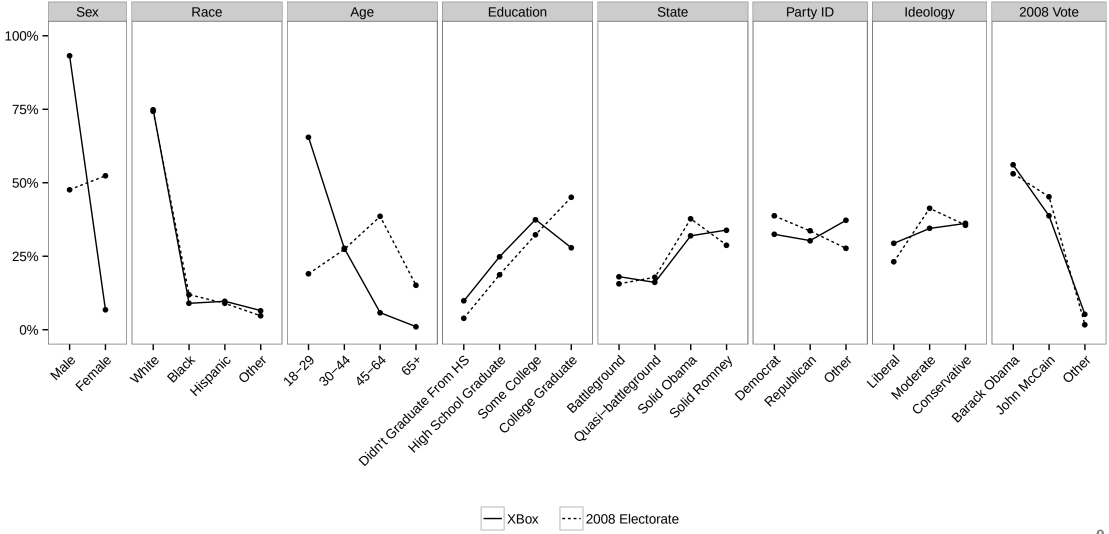
Example: the Xbox Poll





Wang et al. (2015) used MRP to obtain estimates of voting behavior in the 2012 US Presidential election based on a sample of 350,000 Xbox users, empaneled 45 days prior to the election.

Selection bias in the Xbox panel



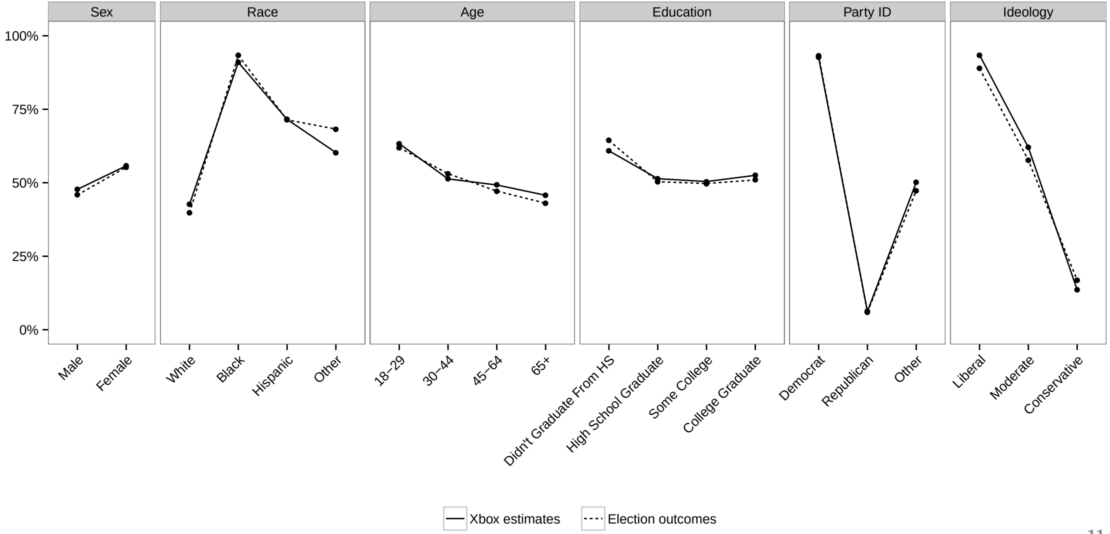
Apply MRP to big data

- Used detailed highly predictive covariates about voting behavior: sex, race, age, education, state, party ID, political ideology, and reported 2008 vote, resulting in 176,256 cells, 2 gender x 4 race x 4 age x 4 education x 4 party x 3 ideology x 50 states .
- Fit multilevel logistic regression:

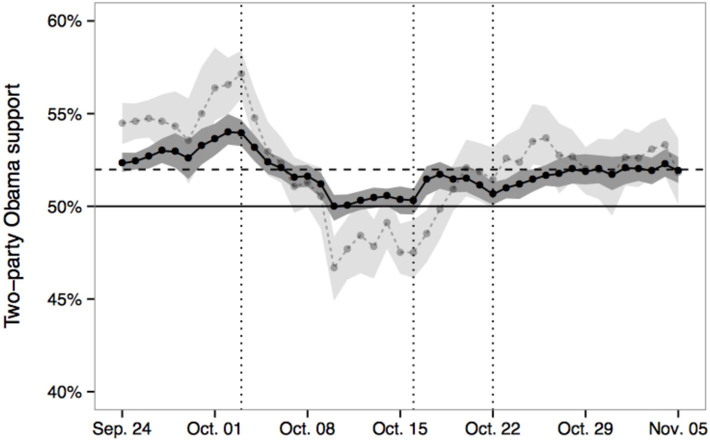
$$\Pr(Y_i = 1) = \text{logit}^{-1}(\alpha_0 + \alpha_1 * sh + \alpha^{state}_{j[i]} + \alpha^{edu}_{j[i]} + \alpha^{sex}_{j[i]} + \alpha^{age}_{j[i]} + \alpha^{race}_{j[i]} + \alpha^{party}_{j[i]}),$$

  where $j[i]$ refers to the cell j that unit i belongs to.
- Introduce prior distributions $\alpha^{var}_{j[i]} \sim N(0, \sigma^2_{var})$, $\sigma^2_{var} \sim inv - \chi^2(\nu_0, \sigma^2_0)$.

MRP estimates of 2012 voting from Xbox panel



The power of poststratefication adjustments



[1]

[1] The light gray line (with SEs) shows the result after adjusting for demographics; the dark gray line shows the estimates after also adjusting for day-to-day changes in the party identification of respondents. The vertical dotted lines show the dates of the presidential debates.

Examples: MRP for public health, economics research

- CDC has recently been using MRP to produce county, city, and census tract-level disease prevalence estimates in the 500 cities project (https://www.cdc.gov/500cities/).
- A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System (Zhang et al. 2014; Zhang et al. 2015).
- MRP used the relationships between demography and vote choices to project state-level election results (https://www.economist.com/graphic-detail/2019/07/06/if-everyone-had-voted-hillary-clinton-would-probably-be-president).

MRP can also fail

Use MRP with caution

Statistical Modeling, Causal Inference, and Social Science

HOME | BOOKS | BLOGROLL | SPONSORS | AUTHORS | FEED

« Scientific communication by press release

Nate Silver's website »

President of American Association of Buggy-Whip Manufacturers takes a strong stand against internal combustion engine, argues that the so-called "automobile" has "little grounding in theory" and that "results can vary widely based on the particular fuel that is used"

Posted by Andrew on 6 August 2014, 2:45 pm

Some people pointed me to this official statement signed by Michael Link, president of the American Association for Public Opinion Research (AAPOR). My colleague David Rothschild and I wrote a measured response to Link's statement which I posted on the sister blog. But then I made the mistake of actually reading what Link wrote, and it really upset me in that it reminded me of various anti-innovation attitudes in statistics I've encountered over the past few decades.

If you want to oppose innovation, fine: there are a lot of reasons why it can make sense to go with old methods and to play it slow. Better the devil you know etc. And on the other side there are reasons to go with the new. Open discussion and debate can be helpful in establishing the zones of application where different methods are more useful.

What I really *don't* like, though, is when someone takes a position and then just makes things up to support it, as if this is some kind of war of soundbites and it doesn't matter what you say as long as it sounds good. That's what Link did in his statement. He just made stuff up. AAPOR is a serious professional organization and this statement was a serious mistake on its part.

After reading Link's article, I wrote a long sarcastic post blasting it. But then I deleted my post: really, what was the point? Instead, I'll say things as directly as possible.

In his article, Link criticizes the recent decision of the New York Times to work with polling company YouGov to conduct an opt-in internet survey. Link states that "these methods have little grounding in theory and the results can vary widely based on the particular method used."

But he's just talking out his ass. Traditional surveys nowadays can have response rates in the 10% range. There's no "grounding in theory" that allows you to make statements about those missing 90% of respondents. Or, to put it another way, the "grounding in theory" that allows

Search

RECENT COMMENTS

- Anoneuoid on Alison Mattek on physics and psychology, philosophy, models, explanations, and formalization
- Daniel Lakeland on "Guarantee" is another word for "assumption"
- elin on A rise in premature publications among politically engaged researchers may be linked to Trump's election, study says
- elin on A rise in premature publications among politically engaged researchers may be linked to Trump's election, study says
- Daniel Lakeland on "Guarantee" is another word for "assumption"
- Corey on "Guarantee" is another word for "assumption"
- Daniel Lakeland on Alison Mattek on physics

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 (D. 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 $j = 1, \ldots, J$, 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.

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

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

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

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

Partial pooling with MRP

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

- The approximated MRP estimator is given by

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

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 tradeoff for the MRP estimator (Si et al, in preparation)

The key steps

1. **Multilevel regression** Fit a model relating the survey outcome to covariates across poststratication 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.

Ingredients for MRP and the running example

Survey Pew Research Organization's *October 2016 Political Survey* (2,583 interviews, conducted October 20-25, 2016.)

Survey variable 2016 Presidential voting intention

Covariates Individual characteristics (from the survey) and group level predictors (2012 state vote)

Post-strata Age x Gender x Race x Education x State

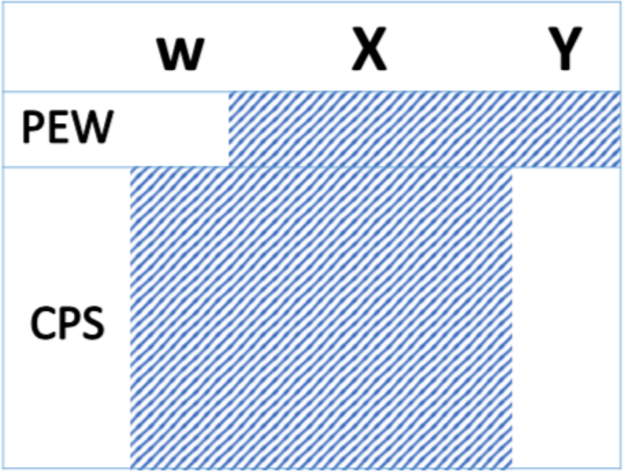
Stratum counts from the November 2016 Voting and Registration Supplement to the *Current Population Survey*

Data sources

The file `cleaned.RData` contains four R dataframes:

- `pew` - Pew Research Organization's **October 2016 Political Survey**. The original data can be found at `http://www.people-press.org/dataset/october-2016-political-survey/`.
- `cps` - the November 2016 Voting and Registration Supplement to the **Current Population Survey**. The full dataset can be downloaded from <www.nber.org/cps/>.
- `votes12` and `votes16` - votes cast for major presidential candidates, turnout, and voting age population by state. **Vote counts** are from `https://uselectionatlas.org/` and **population counts** are from `https://www2.census.gov/programs-surveys/cps/`.

Data structure



Recode Pew data

Variables should be `factors` (R's version of categorical variables) with the same `levels` (categories) *in the same order*.

```
suppressMessages(library("tidyverse"))
load("data/cleaned.RData")
pew <- pew %>%
  filter(
    complete.cases(age, raceeth, gender, educ, vote16),
    vote16 != "nonvoter") %>%
  mutate(
    demvote = ifelse(vote16 == "clinton", 1, 0),
    age4 = factor(case_when(age < 30 ~ "18-29",
      age < 45 ~ "30-44", age < 65 ~ "45-64",
      TRUE ~ "65+")),
    race3 = fct_collapse(raceeth,
      white = c("white", "other")),
    educ4 = fct_collapse(educ,
      "hs" = c("grades 1-8", "hs dropout", "hs grad"),
      "some col" = c("some col", "assoc")))
```

...then do the same for CPS

```
cps <- cps %>%
  filter(
    complete.cases(age_top_codes,
      raceeth, gender, educ, turnout),
    turnout == "yes") %>%
  mutate(
    age4 = factor(case_when(
      age_top_codes == "<80" & age < 30 ~ "18-29",
      age_top_codes == "<80" & age < 45 ~ "30-44",
      age_top_codes == "<80" & age < 65 ~ "45-64",
      TRUE ~ "65+")),
    race3 = fct_collapse(raceeth,
      white = c("white", "other")),
    educ4 = fct_collapse(educ,
      "hs" = c("grades 1-8", "hs dropout", "hs grad"),
      "some col" = c("some col", "assoc")))
```

Check that the datasets are consistent – mistakes will be made!

Time spent cleaning the data at this stage is time well spent.

```r
compare_distributions <- function(var, data1, data2, wgt1, wgt2, digits = 1) {
  stopifnot(all(levels(data1[[var]]) == levels(data2[[var]])))
  formula1 <- as.formula(paste(wgt1, "~", var))
  formula2 <- as.formula(paste(wgt2, "~", var))
  tbl <- rbind(round(100 * prop.table(xtabs(formula1, data1)), digits),
      round(100 * prop.table(xtabs(formula2, data2)), digits))
  row.names(tbl) <- c(substitute(data1), substitute(data2))
  tbl
}
compare_distributions("race3", pew, cps, "", "weight")
```

```
##      white black hispanic
## pew   83.3   8.9      7.8
## cps   78.9  11.9      9.2
```

Compare variables in pew and cps

```
compare_distributions("educ4", pew, cps, "", "weight")
```

```
##        hs some col col grad postgrad
## pew 22.0      26.9      29.7      21.3
## cps 29.6      30.8      25.0      14.6
```

```
compare_distributions("age4", pew, cps, "", "weight")
```

```
##     18-29 30-44 45-64  65+
## pew  12.8  19.3  40.6 27.3
## cps  15.7  22.5  37.6 24.2
```

```
compare_distributions("gender", pew, cps, "", "weight")
```

```
##     male female
## pew 53.5   46.5
## cps 46.4   53.6
```

Estimating the model in R

```r
install.packages(c("tidyverse", "lme4", "survey", "arm", "maps", "mapproj",
  "gridExtra"))
```

```r
library(tidyverse); library(maps); library(mapproj); library(gridExtra);
```

```
##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

Add group-level covariates

```
obama12 <- votes12 %>%
  mutate(obama12 = obama / turnout) %>%
  select(state, obama12)
pew <- left_join(pew, obama12, by = "state")
cps <- cps %>%
  mutate(female = ifelse(gender == "female", 1, 0),
    female.c = female - 0.5) %>%
  left_join(obama12, by = "state")
X <- model.matrix(~ 1 + age4 + gender + race3 + educ4 +
  region + qlogis(obama12), data = pew)
data <- list(n = nrow(X), k = ncol(X), X = X, y = pew$demvote,
  J = nlevels(pew$state), group = as.integer(pew$state))
```

Stan codes

```
model_code <- "data {
  int n; // number of respondents
  int k; // number of covariates
  matrix[n, k] X; // covariate matrix
  int<lower=0, upper=1> y[n]; // outcome (demvote)
  int J; // number of groups (states)
  int<lower=1, upper=J> group[n]; // group index
}
parameters {
  vector[k] beta; // fixed effects
  real<lower=0> sigma_alpha; // sd intercept
  vector[J] alpha; // group intercepts
}
model {
  vector[n] Xb;
  beta ~ normal(0, 4);
  sigma_alpha ~ normal(0.2, 1); // prior for sd
  alpha ~ normal(0, 1); // standardized intercepts
  Xb = X * beta;
  for (i in 1:n)
    Xb[i] = Xb[i] + sigma_alpha * alpha[ group[i] ];
  y ~ bernoulli_logit(Xb);
}"
```

```
sims <- stan(model_code = model_code, data = data,
  seed = 1234)
```

Rename the coefficients for easier reading

```
coef.names <- c(colnames(X), "sigma_alpha", levels(pew$state), "lp__")
names(sims) <- coef.names
```

```
##  [1] "(Intercept)"         "age430-44"           "age445-64"           "age465+"
##  [5] "genderfemale"        "race3black"          "race3hispanic"       "educ4some col"
##  [9] "educ4col grad"       "educ4postgrad"       "regionSouth"         "regionNorth Central"
## [13] "regionWest"          "qlogis(obama12)"     "sigma_alpha"         "AK"
## [17] "AL"                  "AR"                  "AZ"                  "CA"
## [21] "CO"                  "CT"                  "DC"                  "DE"
## [25] "FL"                  "GA"                  "HI"                  "IA"
## [29] "ID"                  "IL"                  "IN"                  "KS"
## [33] "KY"                  "LA"                  "MA"                  "MD"
## [37] "ME"                  "MI"                  "MN"                  "MO"
## [41] "MS"                  "MT"                  "NC"                  "ND"
## [45] "NE"                  "NH"                  "NJ"                  "NM"
## [49] "NV"                  "NY"                  "OH"                  "OK"
## [53] "OR"                  "PA"                  "RI"                  "SC"
## [57] "SD"                  "TN"                  "TX"                  "UT"
## [61] "VA"                  "VT"                  "WA"                  "WI"
## [65] "WV"                  "WY"                  "lp__"
```

Summary of fixed effect estimates

```r
print(sims, par = "beta")
```

```
## Inference for Stan model: 8cd7e7f51310865be8782b9d9386f08c.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##                      mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## (Intercept)         -0.81    0.00 0.25 -1.29 -0.97 -0.81 -0.65 -0.31  2616    1
## age430-44           -0.14    0.00 0.20 -0.53 -0.27 -0.14  0.00  0.24  3887    1
## age445-64           -0.35    0.00 0.17 -0.69 -0.47 -0.35 -0.23 -0.01  3716    1
## age465+             -0.17    0.00 0.18 -0.53 -0.30 -0.18 -0.05  0.18  3790    1
## genderfemale         0.64    0.00 0.11  0.41  0.56  0.64  0.71  0.86  4000    1
## race3black           3.11    0.01 0.32  2.51  2.89  3.09  3.32  3.78  4000    1
## race3hispanic        1.13    0.00 0.21  0.72  0.99  1.13  1.28  1.55  4000    1
## educ4some col        0.09    0.00 0.16 -0.22 -0.02  0.09  0.20  0.40  4000    1
## educ4col grad        0.48    0.00 0.16  0.16  0.37  0.47  0.59  0.78  4000    1
## educ4postgrad        1.07    0.00 0.17  0.73  0.96  1.07  1.19  1.42  4000    1
## regionSouth         -0.26    0.00 0.23 -0.71 -0.42 -0.27 -0.11  0.18  2652    1
## regionNorth Central -0.07    0.00 0.22 -0.51 -0.21 -0.08  0.08  0.36  2747    1
## regionWest           0.11    0.00 0.23 -0.35 -0.04  0.12  0.27  0.56  2345    1
## qlogis(obama12)      0.95    0.00 0.22  0.53  0.80  0.95  1.09  1.38  4000    1
##
## Samples were drawn using NUTS(diag_e) at Tue Oct  8 15:15:59 2019.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Predictive distributions: imputation of survey variables for the population

- The final step in MRP is to **impute** vote for the entire population.
  - The sample is a trivial proportion of the population.
  - We need to impute the survey variable to everyone **not** surveyed.
- The **posterior predictive distribution** $p(\tilde{y}|y)$ is the conditional distribution of a **new** draw $\tilde{y}$ from the model, conditional upon the **observed** data y.
- This requires averaging $p(\tilde{y}|\theta)$ over the posterior distribution $p(\theta|y)$, *i.e.*, over the uncertainty in both $\tilde{y}$ *and* θ.
- Contrast this with
  - **Regression imputation** the expected value of $\tilde{y}$ is used
  - **Plug-in methods** a point estimate is substituted for the unknown parameter.

Imputation in Stan

Munge the population data in R

```
X0 <- model.matrix(~ 1 + age4 + gender + race3 + educ4 +
    region + qlogis(obama12), data = cps)
data <- list(n = nrow(X), k = ncol(X), X = X, y = pew$demvote,
  J = nlevels(pew$state), group = as.integer(pew$state),
  N = nrow(X0), X0 = X0, group0 = as.integer(cps$state))
```

and add to the Stan data block:

```
data {
  ...
  // add population data definitions
  int N; // number of rows in population (cps)
  matrix[N, k] X0; // covariates in population
  int<lower=1, upper=J> group0[N]; // group index in population
}
```

The generated quantities block in Stan

Tell Stan what you want to impute and how to create the imputations.

```
generated quantities {
  int<lower=0, upper=1> yimp[N];
  {
    vector[N] Xb0;
    Xb0 = X0 * beta;
    for (i in 1:N)
      yimp[i] = bernoulli_logit_rng(Xb0[i] + sigma_alpha * alpha[ group0[i] ]);
  }
}
```

Note the use of the `bernoulli_logit_rng` (random number generator) function to draw from the posterior predictive distribution. The `generated quantities` block cannot contain any distributions (indicated by ~).

The complete Stan program

```
model_code <- "data {
  int n; // number of respondents
  int k; // number of covariates
  matrix[n, k] X; // covariate matrix
  int<lower=0, upper=1> y[n]; // outcome (demvote)
  int J; // number of groups (states)
  int<lower=1, upper=J> group[n]; // group index
  int N; // population size
  matrix[N, k] X0; // population covariates
  int group0[N]; // group index in population
}
parameters {
  vector[k] beta; // fixed effects
  real<lower=0> sigma_alpha; // sd intercept
  vector[J] alpha; // group intercepts
}
```

```
"model {
  vector[n] Xb;
  beta ~ normal(0, 4);
  sigma_alpha ~ normal(0.2, 1);
  alpha ~ normal(0, 1);
  Xb = X * beta;
  for (i in 1:n)
    Xb[i] += sigma_alpha * alpha[ group[i] ];
  y ~ bernoulli_logit(Xb);
}
generated quantities {
  int<lower=0, upper=1> yimp[N];
  {
    vector[N] Xb0;
    Xb0 = X0 * beta;
    for (i in 1:N)
      yimp[i]=bernoulli_logit_rng(Xb0[i]+sigma_alpha*alpha[group0[i]])
  }
}"
```

Extracting the simulations

Stan has imputed 4000 values for each of the rows in `cps`. We sample 500 (much more than necessary, but it's still fast).

Now we can perform any analyses we wish on the imputed `cps` data and average the results over the 10 imputed datasets to get point estimates.

The easy way with `rstanarm`

- `Rstanarm` is an R package that writes and executes Stan code for you.
- It uses the same notation as `lme4` for specifying multilevel models.

```
library(rstanarm)
fit <- stan_glmer(demvote ~ 1 + age4 + gender + race3 + educ4 +
    region + qlogis(obama12) + (1 | state), data = pew, family = binomial)
```

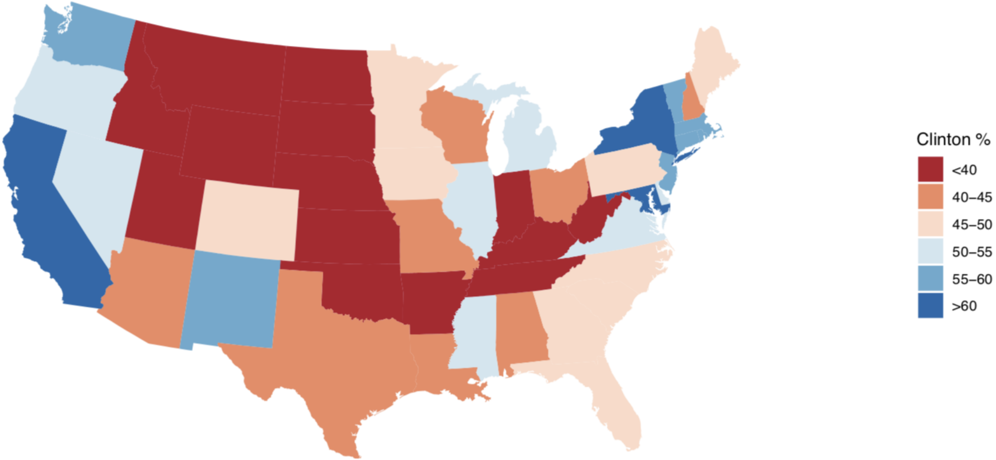
- The function `posterior_predict` in `rstanarm` substitutes for the usual `predict` function in R:

```
imputations <- posterior_predict(fit, draws = 500,
  newdata = select(cps, age4, gender, race3, educ4, region, obama12, state))
```

(This creates a matrix `imputations` of dimension `draws` x `nrow(newdata)`.)

- Extract the estimates using `get_state_estimates`.

What the map looks like



3. Applications in survey research

A unified MRP framework

- "Survey weighting is a mess" (A. 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)
- MY 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

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

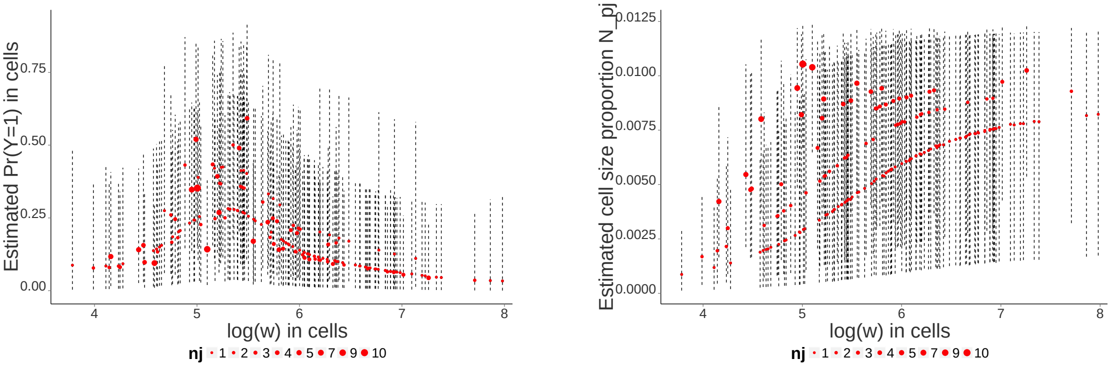
   where $\mu(x_j)$ is a mean function of x_j.

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

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

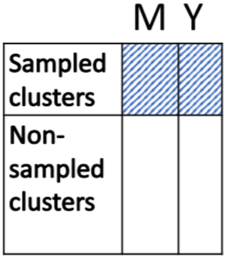
   where Σ_{xx} denotes the covariance function of the distances for any $x_j, x_{j'}$.

Estimates of cell means and cell size proportions



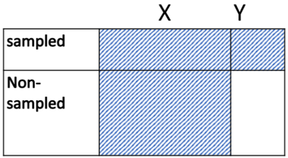
Proportion estimation of individuals with public support based on the Fragile Families and Child Wellbeing Study.

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. 2018)



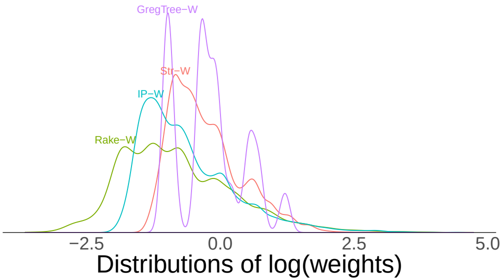
- Handle deep interactions among weighting variables
- The population cell mean θ_j is modeled as

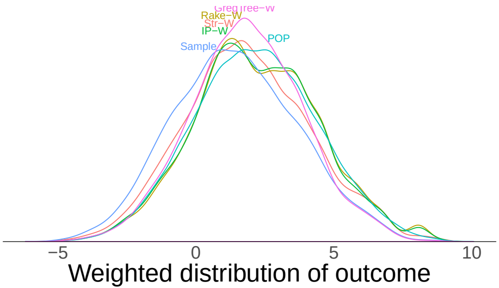
$$\theta_j = \alpha_0 + \sum_{k \in S^{(1)}} \alpha_{j,k}^{(1)} + \sum_{k \in S^{(2)}} \alpha_{j,k}^{(2)} + \cdots + \sum_{k \in S^{(q)}} \alpha_{j,k}^{(q)}, \qquad (4)$$

  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

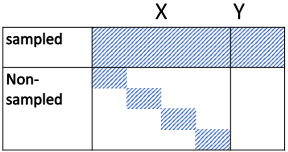
Stan fitting under structured prior in rstanarm

```r
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 2018)



- 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

4. Recent developments and challenges

Structural, spatial, temporal prior specification

- We developed structured prior distributions to reflect the hierarchy in deep interactions (Si et al. 2018)
- 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 (A. Gelman et al. 2019)

MRP framework for data integration (Si et al, 2019)

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 $c = \sum_j N_j\psi_j$, 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
- . . . .

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

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.

Challenges

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

References

Thank you

yajuan@umich.edu

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, and Thomas C. Little. 1997. "Poststratifcation into Many Cateogiries Using Hierarchical Logistic Regression." *Survey Methodology* 23: 127–35.

Gelman, Andrew, Jeffrey Lax, Justin Phillips, Jonah Gabry, and Robert Trangucci. 2019.