# Estimating State Public Opinion With Multi-Level Regression and Poststratification using *R*

Jonathan P. Kastellec
jkastell@princeton.edu
Department of Politics
Princeton University

Jeffrey R. Lax
Department of Political Science
Columbia University
JRL2124@columbia.edu

Justin Phillips
JHP2121@columbia.edu
Department of Political Science
Columbia University

September 6, 2019

**Abstract**

This paper provides a primer for estimating public opinion at the state level using the technique of Multilevel Regression and Postratification (MRP). We provide sample *R* code for creating estimates and give step-by-step instructions on setting up the data, running models, and collecting estimates. Replication datasets and code found in the paper can be accessed at `https://scholar.princeton.edu/sites/default/files/jkastellec/files/mrp_primer_replication_files.zip`

# 1 Introduction

Despite the proliferation of public opinion polls, state-level surveys remain quite rare. Finding comparable surveys across all (or even many) states is nearly impossible. To cope with this problem, scholars have devised techniques which allow them to use national surveys to generate estimates of state-level opinion. The dominant method is disaggregation, popularized by Erikson, Wright and McIver (1993). This method pools large numbers of national surveys and then disaggregates the data so as to calculate opinion percentages by state. While disaggregation is easily implemented, it has its drawbacks. Typically, surveys over many years, often 10 or more, must be pooled to guarantee sufficient samples sizes within each state. This constrains the number and types of issues for which scholars can estimate state opinion. Furthermore, disaggregation does not correct for sampling issues and may obscure temporal dynamics in state opinion. Indeed, if there are temporal dynamics, opinion estimates produced via disaggregation will be inaccurate.

We recommend, at least in some circumstances, that scholars estimate state-opinion by employing a technique that we refer to as multilevel modeling with poststratification (MRP). This method has a long history (see e.g. Pool, Abelson and Popkin (1965)), but its modern-day implementation can be traced to Park, Gelman and Bafumi (2004). Like disaggregation, MRP relies upon national survey data. MRP, however, begins by using multilevel regression to model individual survey responses as a function of *demographic and geographic predictors*, partially pooling respondents across states to an extent determined by the data. The final step is poststratification, in which the estimates for each demographic-geographic respondent type are weighted (poststratified) by the percentages of each type in the actual state populations. Why do we recommend this technique?

- MRP strongly outperforms disaggregation (i.e., produces opinion estimates that are more accurate and robust) when working with small and medium-sized samples. MRP

does slightly better in large samples, particularly when it comes to estimating opinion in small states (see Lax and Phillips 2009*b*, Figures 1 & 2)

- MRP has been shown to produce reasonably accurate estimates of state public opinion using as little as a single large national poll—approximately 1,400 survey respondents. (see Lax and Phillips 2009, Figures 1, 2, & 5)

- Poststratifcation corrects for clustering and other statistical issues that may bias estimates obtained via disaggregation.

- MRP can deal with temporal instability in public opinion

- MRP produces much more information than disaggregation. It provides insights about the determinants of public opinion and the degree to which state variation is based on demographic characteristics versus residual (cultural?) differences.

- MRP can be used to estimate opinion in states that are rarely surveyed. For example, respondents from Alaska and Hawaii are usually not included in national polls and therefore opinion in these states cannot be measured using disaggregation. Estimates for Alaska and Hawaii can be created using MRP.

- MRP can be used to estimate opinion in other subnational areas besides states (i.e., congressional districts).

We have used MRP to study both the relationship between public opinion and gay rights policies in the U.S. states (Lax and Phillips 2009*a*) and the relationship between state-level public opinion and senators' voting on Supreme Court nominees (Kastellec, Lax and Phillips 2010). We believe the method has the potential to open up several research avenues that have been closed to date. This paper discusses how to collect the data necessary to construct state-level estimates and how to implement MRP in *R*. We use public opinion data on same-sex marriage as a running example.

## 2 Steps for Implementing MRP:

In this section we describe how to implement MRP, providing annotated *R* code where appropriate.

**1) Gather national opinion polls**. These polls should include some respondent demographic information and some type of geographic indicator. If you are interested in estimating opinion at the state level (as we are), the surveys should include a respondent's state of residence (if you are interested in opinion at the level of congressional districts, the survey should include an indictor of a respondent's congressional district). We find that state-level opinion can be estimated fairly accurately using as little as a single large national poll (approximately 1,400 respondents). Here we use five national polls that were conducted in 2004.

**2) Recode these polls as necessary so that they can be combined into a single internally-consistent dataset**. For convenience, we call this dataset a "megapoll." Where possible you should use respondents' demographic and geographic characteristics to create group (i.e., categorical) variables. This will allow for a more efficient estimation and also means that you do not need to exclude a reference category. For example, in our research we use data on respondents' sex, race (white, Hispanic, or black), age, education, state, and region. We combine race and gender into a single variable with six possible categories (ranging from male-white to female-Hispanic). We also use group variables for age (18-29, 30-44, 45-64, and 65+), education (less than a high school education, high school graduate, some college, and college graduate), an interaction between our age and education measures, and state (Alabama through Wyoming). We treat Washington D.C. as a state. When identifying respondent demographic data in surveys, be sure to only use data that is also available from the census (otherwise you will not be able to properly post stratify). If you

are using survey responses from multiple polls or years you can also create group variables for these as well. This helps control for poll, question wording, and year effects (we do this below).

Loading the megapoll is the first step in *R*. We begin by loading the `arm` package, which contains several functions to implement and analyze multilevel models, including the `lmer` function, and the `foreign` package, to allow the importation of Stata datasets.

```
library(‘‘arm")
library(‘‘foreign")
```

We next load our megapoll into R:

```
marriage.data <- read.dta("gay_marriage_megapoll.dta",
    convert.underscore = TRUE) #convert variables names with underscores to periods
```

**3) You may also want to create a separate dataset of state-level predictors**. In a multilevel regression, state-level effects can be modeled using additional state-level predictors such as region or state-level (aggregate) demographics (e.g., those not available at the individual level in the survey or census). Adding group-level predictors usually reduces unexplained group level variation thus reducing group level standard deviation. This in turn increases the amount of pooling done by the multilevel model, giving more precise estimates, especially for groups with small populations. We use a group variable for region (Northeast, Midwest, South, West, and Washington D.C.) and a continuous measure for the share of the state's population that is evangelical Protestant or Mormon. At various times we also use the Republican vote share in the previous presidential election and state-level per-capita income.

We read the state-level dataset into *R*, sort it by the numeric order of the state's initials (e.g. AL = 1, DC = 8, WY = 51):

```
Statelevel <- read.dta("state_level_update.dta",convert.underscore = TRUE)
Statelevel <- Statelevel[order(Statelevel$sstate.initnum),]
```

**3) Collect census data to enable poststratification**. To poststratify one needs to have census data that corresponds to all of the individual-level demographic variables included in the opinion model. Be careful here. MRP requires knowing not just the simple state-level statistics reported in the Statistical Abstract, such as the number of females or African Americans in a state. If your model treats opinion as a function of gender, race, age, and education you will need to know, for instance, the number of African American females aged 18 to 29 years who are college graduates. The necessary data can be obtained from the Census Bureaus website using the "DataFerret" (at http://dataferrett.census.gov/). The DataFerret will help you get cross-tabs for state-level data using the 1% or 5% Public Use Microdata Sample from either the 2000 or 1990 census. Older census data can be obtained, though it is a bit more difficult to access (see www.census.gov/main/www/pums.html). Keep in mind that not all cross-tabulations are available, particularly for smaller geographic units (say, congressional districts). You are also limited by the type of data the census collects. For instance, the census does not gather data on an individual's religious affiliation, voting behavior, or partisan identification (all of which political scientists care about). Note, however, that our research suggests that you may be able to generate reasonably accurate estimates of opinion using simple models that include basic demographic and geographic information.

Ultimately, you need a dataset of the population counts for each demographic-state type (or "cell"). In our analysis, this table is 4,896 rows long (excluding the top row of labels). A sample of the table is shown below.

For same-sex marriage, we use the 5% Public Use Microdata Sample from the 2000 census. We use the "match" function to create a variable indicating the state initial number for each cell in the Census data:

```
Census <- read.dta("poststratification 2000.dta",convert.underscore = TRUE)
Census <- Census[order(Census$cstate),]
Census$cstate.initnum <-  match(Census$cstate, statelevel$sstate)
```

| | race.gender | age | edu | state | N |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 66177 |
| 2 | 1 | 1 | 2 | 1 | 32465 |
| 3 | 1 | 1 | 3 | 1 | 59778 |
| 4 | 1 | 1 | 4 | 1 | 27416 |
| 5 | 1 | 2 | 5 | 1 | 43032 |
| 6 | 1 | 2 | 6 | 1 | 81312 |
| 7 | 1 | 2 | 7 | 1 | 52699 |
| 8 | 1 | 2 | 8 | 1 | 90217 |
| 9 | 1 | 3 | 9 | 1 | 63155 |
| 10 | 1 | 3 | 10 | 1 | 68821 |
| 11 | 1 | 3 | 11 | 1 | 43127 |
| . . . | | | | | |
| 4894 | 6 | 4 | 2 | 51 | 2541 |
| 4895 | 6 | 4 | 3 | 51 | 2967 |
| 4896 | 6 | 4 | 4 | 51 | 1029 |

With all the data in hand, we can now create a series of index variables that we will use in the individual-level model and in the poststratification:

```
#At level of megapoll
marriage.data$race.female <- (marriage.data$female *3) + marriage.data$race.wbh
marriage.data$age.edu.cat <- 4 * (marriage.data$age.cat -1) + marriage.data$edu.cat
marriage.data$p.evang.full <- Statelevel$p.evang[marriage.data$state.initnum]
marriage.data$p.mormon.full <-Statelevel$p.mormon[marriage.data$state.initnum]
marriage.data$p.relig.full <- marriage.data$p.evang.full + marriage.data$p.mormon.full
marriage.data$p.kerry.full <- Statelevel$kerry.04[marriage.data$state.initnum]

#At census level (same coding as above for all variables)
Census$crace.female <- (Census$cfemale *3) + Census$crace.WBH
Census$cage.edu.cat <- 4 * (Census$cage.cat -1) + Census$cedu.cat
Census$cp.evang.full<-  Statelevel$p.evang[Census$cstate.initnum]
Census$cp.mormon.full <- Statelevel$p.mormon[Census$cstate.initnum]
Census$cp.relig.full <- Census$cp.evang.full + Census$cp.mormon.full
Census$cp.kerry.full <-  Statelevel$kerry.04[Census$cstate.initnum]
```

**4) Fit a regression model for an individual survey response given demographics and geography**. We are now ready to estimate an individual-level model of opinion on gay marriage rights. We treat each individual's response as a function of his or her demographics and state (for individual $i$, with indexes $j$, $k$, $l$, $m$, $s$, and $p$ for race-gender combination, age category, education category, region, state, and poll respectively, and including an age-education interaction):

$$\Pr(y_i = 1) = \text{logit}^{-1}(\beta^0 + \alpha^{race,gender}_{j[i]} + \alpha^{age}_{k[i]} + \alpha^{edu}_{l[i]} + \alpha^{age.edu}_{k[i],l[i]} + \alpha^{state}_{s[i]} + \alpha^{year}_{p[i]}) \quad (1)$$

The terms after the intercept are modeled effects for the various groups of respondents. Each is modeled as drawn from a normal distribution with mean zero and some estimated variance:

$$\begin{aligned}
&\alpha^{race,gender}_j \sim N(0, \sigma^2_{race,gender}), \text{ for } j = 1, ..., 6 \quad (2)\\
&\alpha^{age}_k \sim N(0, \sigma^2_{age}), \text{ for } k = 1, ..., 4\\
&\alpha^{edu}_l \sim N(0, \sigma^2_{edu}), \text{ for } l = 1, ..., 4\\
&\alpha^{age.edu}_{k,l} \sim N(0, \sigma^2_{edu}), \text{ for } k = 1, ..., 4 \text{ and } l = 1, ..., 4\\
&\alpha^{poll}_p \sim N(0, \sigma^2_{poll}), \text{ for } p = 1, ...
\end{aligned}$$

The state effects are in turn modeled as a function of the region into which the state falls and the state's conservative religious percentage and Democratic 2004 presidential vote share[1]:

$$\alpha^{state}_s \sim N(\alpha^{region}_{m[s]} + \beta^{relig} \cdot relig_s + \beta^{presvote} \cdot presvote_s, \sigma^2_{state}), \text{ for } s = 1, ..., 51 \quad (3)$$

[1]These are just some examples of group-level predictors—which reduce unexplained group-level variation, leading to more precise estimation (Gelman and Hill 2007, 271)—one might choose to employ

The region variable is, in turn, another modeled effect:

$$\alpha_m^{region} \sim N(0, \sigma^2_{region}), \text{ for } m = 1, ..., 5 \tag{4}$$

In the model we present below, we label the survey responses $y_i$ as 1 for supporters of same-sex marriage and 0 for opponents and those with no opinion. Depending on the situation, you might also be interested in public opinion among only those respondents who offer an opinion (that is, excluding observations with missing values.) While it is tempting to drop these observations, doing so would create problems, since the Census data on which we will poststratify takes into account all persons, not just those with an opinion. Thus, it is necessary to evaluate both the "yesses" among all respondents (including those who do not offer an opinion) and the "noes" among all respondents, then use both to create a proper estimate of state-level opinion among opinion holders. We discuss how to implement this procedure below.

The model we present below estimates an average response $\theta_j$ for each cross-classification $j$ of demographics and state. Thus$j$ = 1,..., J = 4,896 categories (96 per state). We fit our model in R using the LMER function (linear mixed effects in R (Bates 2005)). Note that multilevel modeling partially pools the group level parameters toward their mean level. There is more pooling when the group level standard deviation is small and more smoothing for groups with fewer observations.

The code for the individual-level model (which follows the structure of *R's* "glm" command) is:

```
individual.model <- glmer(formula = yes.of.all ~ (1|race.female) + (1|age.cat)
  + (1|edu.cat) + (1|age.edu.cat) + (1|state) + (1|region) + (1|poll) + p.relig.full
        + p.kerry.full,data=marriage.data, family=binomial(link="logit"))
```

We use the "display" command to obtain the following results:

```
             coef.est coef.se
(Intercept)  -1.41     0.54
p.relig.full -0.02     0.00
p.kerry.full  0.02     0.01
Error terms:
 Groups      Name        Std.Dev.
 state       (Intercept) 0.04
 age.edu.cat (Intercept) 0.09
 race.female (Intercept) 0.23
 poll        (Intercept) 0.21
 region      (Intercept) 0.20
 edu.cat     (Intercept) 0.36
 age.cat     (Intercept) 0.55
 Residual                NA
---
number of obs: 6341, groups: state, 49; age.edu.cat, 16; race.female, 6; poll, 5; region, 5;
edu.cat, 4; age.cat, 4 AIC = 7459.4, DIC = 7439.4 deviance = 7439.4
```

Of more interest are the coefficients and standard errors on our random effects; here, for example, are those for "race.female":

```
ranef(individual.model)$race.female
  (Intercept)
1      -0.210
2      -0.087
3       0.049
4       0.230
5      -0.226
6       0.246
se.ranef(individual.model)$race.female
     [,1]
[1,] 0.11
[2,] 0.15
```

```
[3,] 0.15
[4,] 0.11
[5,] 0.14
[6,] 0.15
```

Since we do not have any respondents from Alaska or Hawaii, we have to create a vector of state random effects that accounts for these states. We choose to set their random effects to zero.

```
state.ranefs <- array(NA,c(51,1))
dimnames(state.ranefs) <- list(c(Statelevel$sstate),"effect")
for(i in Statelevel$sstate){
    state.ranefs[i,1] <- ranef(individual.model)$state[i,1]
}
state.ranefs[,1][is.na(state.ranefs[,1])] <- 0
```

**5) Poststratify the demographic-geographic types**. The logistic regression above now gives the probability that any adult will support same-sex marriage given the person's sex, race, age, education, and state. We now need to compute weighted averages of these probabilities to estimate the proportion of same-sex marriage supporters in each state.
For any specific cell $j$, specifying a set of individual demographic and geographic values, the results of the opinion model above allow us to make a prediction of pro-gay support, $\theta$j. Specifically, $\theta j$ is the inverse logit given the relevant predictors and their estimated coefficients.

Since we controlled for poll effects, one could choose a specific poll coefficient when generating these predicted values using the inverse logit. We simply use the average across the polls. Since poll effects are centered at zero, like all random effects, we simply plug in zero. The following code creates a prediction for each demographic-state type (that is, each cell in the Census data):

```
cellpred <- invlogit(fixef(individual.model)["(Intercept)"]
            +ranef(individual.model)$race.female[Census$crace.female,1]
            +ranef(individual.model)$age.cat[Census$cage.cat,1]
            +ranef(individual.model)$edu.cat[Census$cedu.cat,1]
            +ranef(individual.model)$age.edu.cat[Census$cage.edu.cat,1]
            +state.ranefs[Census$cstate,1]
            +ranef(individual.model)$region[Census$cregion,1]
            +(fixef(individual.model)["p.relig.full"] *Census$cp.relig.full)
            +(fixef(individual.model)["p.kerry.full"] *Census$cp.kerry.full)
               )
```

The prediction in each cell needs to be weighted by the actual population frequency of that cell, $Nj$ (that is, by how many such people are in the state). For each state, we then can calculate the average response, over each cell $j$ in state $s$:

$$y^{\text{MRP}}_{\text{state } s} = \frac{\sum_{c \in s} N_c \theta_c}{\sum_{c \in s} N_c} \tag{5}$$

To accomplish this, we use the following code

```
cellpredweighted <- cellpred * Census$cpercent.state #weight the
    prediction by the frequency of each cell
#now calculate the percent within each state (weighted average of responses)
statepred <- 100* as.vector(tapply(cellpredweighted,Census$cstate,sum))
statepred
```

If done properly, the result will be a set of state-level opinion estimates. While these estimates are interesting by themselves, they can easily be used as explanatory variables in an empirical analysis of government responsiveness.

**Additional Recommendations:**

- Make sure that you have a good model of individual-level opinion that includes both demographic and geographic variables. The demographic variables included might vary across policy areas.

- When constructing your models, be sure to use your subject-area expertise. You need to construct a good model of individual-level opinion, but not a perfect one.

- If estimating your individual-level model using LMER, confirm that the AIC looks normal and that the standard errors on your coefficients look normal. If the variance on a random effect is zero you can actually just drop it.

- If the effects of demographic variables differ across states, you may want to consider using a varying-intercepts varying-slopes model. This may, however, require a larger number of survey responses.

- If the number of groups in your model is small or the multilevel model is complicated (with many varying intercepts and slopes), you may want to use a full Bayesian approach to estimation.

# References


Bates, Douglas. 2005. "Fitting Linear Models in R Using the lme4 Package." *R News* 5(1):27–30.

Erikson, Robert S., Gerald C. Wright and John P. McIver. 1993. *Statehouse Democracy Public Opinion and Policy in the American States*. Cambridge: Cambridge University Press.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel-Hierarchical Models*. Cambridge: Cambridge University Press.

Kastellec, Jonathan P., Jeffrey R. Lax and Justin H. Phillips. 2010. "Public Opinion and Senate Confirmation of Supreme Court Nominees." *Journal of Politics* 72:767–84.

Lax, Jeffrey R. and Justin H. Phillips. 2009*a*. "Gay Rights in the States: Public Opinion and Policy Responsiveness." *American Political Science Review* 103(3):367–86.

Lax, Jeffrey R. and Justin H. Phillips. 2009*b*. "How Should We Estimate Public Opinion in the States?" *American Journal of Political Science* 53(1):107–21.

Park, David K., Andrew Gelman and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12(4):375–85.

Pool, Ithiel de Sola, Robert P. Abelson and Samuel Popkin. 1965. *Candidates, Issues, and Strategies*. Cambridge, MA: M.I.T. Press.