Mapping subntational public opinion estimates: a tutorial of multi-level regression with post-stratification and Leaflet

2023-04-09

In the fall, Brian Schaffner and I shared state-level estimates of belief in Trump's Big Lie in *The Washington Post*. Given pollsters did not systematically survey each state to probe its voting age population's belief in the integrity of the 2020 Presidential Election, we were forced to make estimates of state-level opinion using a multilevel modeling method known as multilevel regression with post-stratification (MrP).

MrP is commonly used by social-science to estimate subnational public opinion from national surveys — on issues such as support for free speech the marriage equality. YouGov even uses MrP to forecast the partisan vote of parliamentary districts in the UK!

To demonstrate how to use MrP to estimate subnational public opinion, I will use the 2022 Cooperative Election Study — a national survey of nearly 60,000 adults — to estimate state-level support for keeping abortion legal in all circumstances as a matter of choice. The CES can be downloaded from the Harvard Dataverse at this link. The survey is large, but MrP can be used with smaller-N surveys — and even to produce estimates of opinion in less populated geographic units, such as state house districts.

```
dat <- read_csv("CES22_Common.csv")
```

Within the CES, I will look at **CC22_332a** which asks if respondents support or oppose always allowing a woman to obtain an abortion as a matter of choice. Responses are coded as 1 = "support" and 2 = "oppose" which I will recode to a binary of [0,1] wherein 1 = support. This will allow us to estimate the percentage of state populations who suppot always allowing abortion as a matter of choice.

```
dat$prochoice[dat$CC22_332a==1] <- 1 #support
dat$prochoice[dat$CC22_332a==2] <- 0 #oppose
```

Next, we need to prepare the demographics data which I will use in the multilevel model to arrive at state-level estimates of public opinion. My model will control for education (**educ**), race (**race**, **hispanic**), income (**faminc_new**) and age (obtained via **birthyr**). However, these variables need to be recoded into binary, given the Census data we use to weight make our prediction are in proportions of the target population — in this case, states.

First, we will recode my education variable to a binary for if a respondent has *not* pursued more than a high school degree. This is due to the education polarization surrounding both unplanned pregnancies and the two American political parties. As such, we will recode **educ** to a new variable **hsmax** in which if a responded has pursued a higher degree, they are coded as 0, and respondents who have only received at most a high school diploma, who are coded as 1.

```
dat$hsmax <- 0
dat$hsmax[dat$educ < 3] <- 1 #For those with high school degree or less
```

Next, we should recode race into two different binary variables: one for if a respondent identifies as Black (**black**), and another for if a respondent is Hispanic (**hispanx**).

```
dat$black <- 0
dat$black[dat$race==2] <- 1 #where 2 is "Black or African-American"

dat$hispanx <- 0
dat$hispanx[dat$race==3] <- 1 #for those who said their race is "Hispanic of Latino"
dat$hispanx[dat$hispanic==1] <- 1 #for those who said they are Spanish, latino or Hispanic
```

I then recode age into a binary variable for if a respondent is over the age of 45 (**over45**). In early 2023, this dummy variable will be true for respondents born before 1977, as indicated in **birthyr**.

```
dat$over45 <- 0
dat$over45[dat$birthyr < 1977] <- 1
```

I will also recode income to a binary for those whose household incomes are over $100,000 (**over100k**).

```
dat$over100k <- 0
dat$over100k[dat$faminc_new > 10 & dat$faminc_new != 97] <- 1
```

Ultimately, and perhaps surprising given the focus of this analysis, we do not need to control for gender, given the limited variation in gender composition of states. However, I will include a dummy variable for if respondents voted for former-President Donald Trump in the 2020 election (**trump**).

```
dat$trump <- 0
dat$trump[dat$presvote16post==2] <- 1
```

Now, we can run a regression model to estimate the effect of causal these demographic variables have on support for keeping abortion legal in circumstances as a matter of personal choice. To do this, we can use a logistic regression.

```
Model <- lm(prochoice ~ hsmax + black + hispanx + over100k + trump + over45, data=dat)
summary(Model)
```

```
##
## Call:
## lm(formula = prochoice ~ hsmax + black + hispanx + over100k +
##     trump + over45, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8407 -0.3198  0.2121  0.2357  0.7176
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.787931   0.003611  218.222  < 2e-16 ***
## hsmax       -0.037414   0.003856   -9.703  < 2e-16 ***
## black        0.036047   0.005346    6.743 1.57e-11 ***
## hispanx      0.009842   0.005363    1.835 0.066470 .
## over100k     0.016676   0.005065    3.292 0.000995 ***
## trump       -0.444499   0.004010 -110.838  < 2e-16 ***
## over45      -0.023618   0.003710   -6.366 1.95e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4322 on 59951 degrees of freedom
##   (42 observations deleted due to missingness)
## Multiple R-squared:  0.1919, Adjusted R-squared:  0.1918
## F-statistic:  2373 on 6 and 59951 DF,  p-value: < 2.2e-16
```

However, as our intention is to estimate public opinion in the states by constructing this multilevel model with state-level estimates, we need to use an inverse logit model. In this model, we will also have to employ state random effects. To do this, we need to load the **lme4** package.

However, before we can run this model, we first need to convert the respondent's state (**inputstate**) which is their state's FIPS code, to an index for the state in alphabetical order. To do this easily, I created a dataframe, found here on my GitHub repository for this tutorial which can be used inported to convert states identified by FIPS codes to an index between 1-50 corresponding to each state's alphabetical order, so that we can employ a simple loop when controling for state-level random effects.

```
library(lme4)
statesfips <- read_csv("https://raw.githubusercontent.com/BrendanTHartnett/MRP_demo_abortion/main/fipstostates.cs
v")
dat <- merge(dat, statesfips, by.x="inputstate", by.y="fips")
```

This does remove Washington, D.C., which, given to its overwhelmingly liberal voting record, is fine. We can now run a generalized linear mixed-effects model.

```
state_model <- glmer(formula = prochoice ~  (1 | STATE) + hsmax + black + hispanx + over100k + trump + over45, da
ta=dat, family=binomial(link="logit"))
summary(state_model)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: prochoice ~ (1 | STATE) + hsmax + black + hispanx + over100k +
##     trump + over45
##    Data: dat
##
##      AIC      BIC   logLik deviance df.resid
##  66449.5  66521.5 -33216.8  66433.5    59785
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7142 -0.6894  0.4810  0.5755  2.1864
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  STATE  (Intercept) 0.08536  0.2922
## Number of obs: 59793, groups:  STATE, 50
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.28947    0.04704  27.412  < 2e-16 ***
## hsmax       -0.17749    0.02069  -8.576  < 2e-16 ***
## black        0.25753    0.03092   8.330  < 2e-16 ***
## hispanx      0.01003    0.03013   0.333    0.739
## over100k     0.04202    0.02766   1.519    0.129
## trump       -1.92281    0.02072 -92.784  < 2e-16 ***
## over45      -0.13903    0.02015  -6.900  5.2e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) hsmax  black  hispnx ovr100 trump
## hsmax    -0.161
## black    -0.117 -0.034
## hispanx  -0.120 -0.025  0.115
## over100k -0.111  0.185  0.068  0.036
## trump    -0.125 -0.013  0.170  0.064 -0.013
## over45   -0.254  0.024  0.063  0.152 -0.009 -0.160
```

Once we obtain state-level population data for these fixed-effects, we can fit our model to each state's demographic composition.

Obtaining state-level population data

We can use the **tidycensus** package to import state-level population metrics used to predict state-level opinion. My lab mate Julian Perry recently published a great tidycensus tutorial (cite) which I encourage those of you unfamiliar with **tidycensus** to consult.

Using **tidycensus** I will import state-level Census estimates of education, race, income, age and labor force participation rates. The code for this can be found here, but we will just work with the resulting Census data, the file with the resulting Census data can be imported from my GitHub.

```
Census <- read_csv("https://raw.githubusercontent.com/BrendanTHartnett/MRP_demo_abortion/main/state_census_data.c
sv")
head(Census)
```

```
head(Census)
```

```
## # A tibble: 6 × 11
##    ...1 GEOID NAME   not_i…¹ over45 over1…² highs…³  black hispa…⁴ total…⁵ STATE
##   <dbl> <dbl> <chr>    <dbl>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl> <dbl>
## 1     1     1 Alaba…   0.424  0.561   0.342   0.435 0.263   0.0207  3.71e6     1
## 2     2     2 Alaska   0.325  0.490   0.519   0.353 0.0327  0.0587  5.33e5     2
## 3     3     4 Arizo…   0.399  0.548   0.408   0.359 0.0448  0.237   5.05e6     3
## 4     4     5 Arkan…   0.419  0.555   0.312   0.467 0.150   0.0376  2.22e6     4
## 5     5     6 Calif…   0.363  0.520   0.520   0.364 0.0665  0.305   2.58e7     5
## 6     6     8 Color…   0.318  0.513   0.502   0.291 0.0391  0.160   4.15e6     6
## # … with abbreviated variable names ¹not_in_labor_force, ²over100k,
## #   ³highschool_only, ⁴hispanic, ⁵totalVAPcitizens
```

Obviously, the Census does not ask about one's voting history nor their political leanings. Therefore, in order to discern the percentage of each state's voting age population that voted for Trump in 2020 or did not vote, we will need to use data from the MIT Election Lab to get state-level election returns from the 2020 Presidential Election. To do this, we will access presidential election results from the Harvard Dataverse using the **dataverse** package.

```
library(tidyverse)
library(dataverse)
Sys.setenv("DATAVERSE_SERVER" = "dataverse.harvard.edu")

#Call the specific file of the dataset
election_dat.tab <- get_dataframe_by_name(
  filename = "1976-2020-president.tab",
  dataset = "10.7910/DVN/42MVDX",
  server = "dataverse.harvard.edu")
```

I then just wrangle the results into Trump's votes count as a percentage of all votes in each state.

```
results2020 <- subset(election_dat.tab, year==2020)
trump.results <- subset(results2020, candidate=="TRUMP, DONALD J.")

trump.results$trumpN <- trump.results$candidatevotes
trump.results$trumpP <- trump.results$trumpN/trump.results$totalvotes
trump.results$NAME <- str_to_title(trump.results$state)

election_data <- trump.results[, c("NAME", "trumpN", "trumpP", "totalvotes")]
```

Finally, we can merge this data with my national survey results.

```
Census1 = merge(Census, election_data, by="NAME")
Census$STATE <- Census$NAME
Census <- Census1
```

Using MrP

Now, we can estimate state-level public opinion, using MrP. To begin, we need to create an array to contain state random effects.

```
state_ranefs <- array(NA, c(50, 1))
```

Next, we need to assign state random effects to each respondent using the model obtained from the CES. To do this, we will run a loop through each state.

```
# assign state random effects to array while preserving NAs
for (i in Census$STATE) {

  state_ranefs[i, ] <- ranef(state_model)$STATE[i, 1]

}
state_ranefs[, 1][is.na(state_ranefs[, 1])] <- 0
```

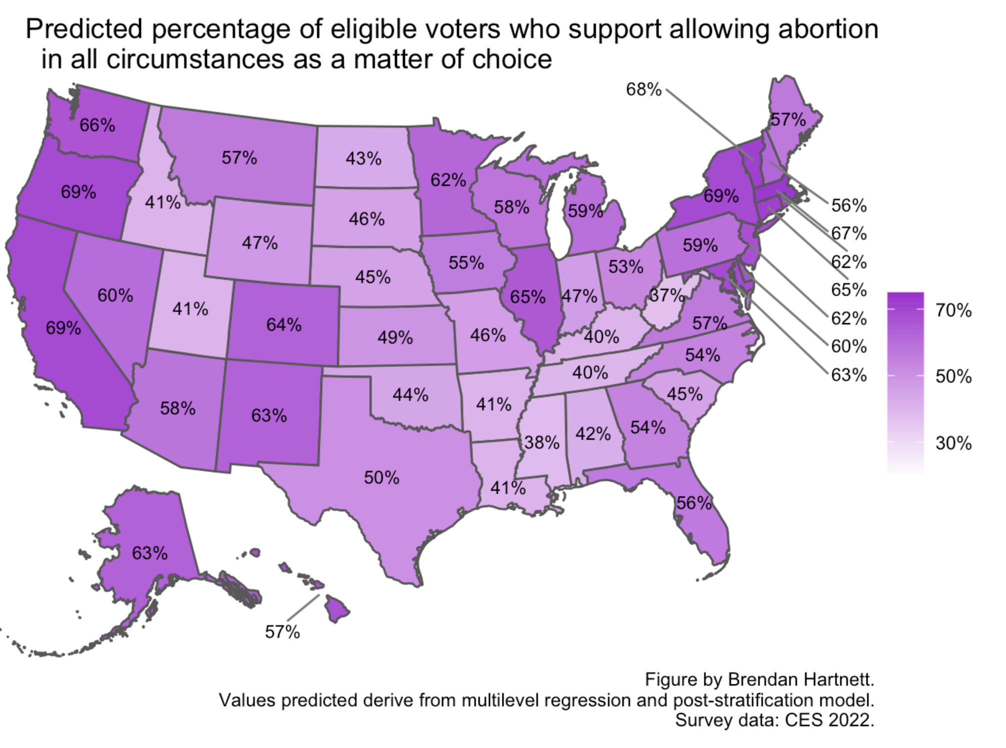
We can then use the invert logit function to model state-level predictions of support for keeping abortion legal in all circumstances with random effects for states and fixed effects from their demographics. We will need to do this using the **arm** package.

```
library(arm)
Census$prediction <- invlogit(fixef(state_model)['(Intercept)'] +
                              state_ranefs[Census$STATE, 1] +
                              (fixef(state_model)['hsmax'] * Census$highschool_only) +
                              (fixef(state_model)['black'] * Census$black) +
                              (fixef(state_model)['hispanx'] * Census$hispanic) +
                              (fixef(state_model)['over100k'] * Census$over100k) +
                              (fixef(state_model)['over45'] * Census$over45) +
                              (fixef(state_model)['trump'] * Census$trumpP))

summary(Census$prediction)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3672  0.4556  0.5700  0.5541  0.6470  0.7278
```

And there you have it! Now we have estimated support for abortion among voting age adults in each state.



You can find the code that is used to make the above visualization here.