Predicting 2020 State Presidential Election Results with a National Tracking Poll and MRP*

Luyuan Hu, Jiayi Wang, Linwei Yao & Bihan Lu

Abstract

This study utilizes public opinion survey data before the 2020 general election, and USES multi level Regression and Poststratification (MRP) estimation model to predict the election results of different regions.Specifically, the prediction model used in this paper consists of three steps. First, it estimates the probability that different types of voters will support Donald Trump and Joe Biden, respectively, through basic demographic variables (gender, age, and education level) supplemented by characteristics at the district level.Secondly, we used national census data to obtain the joint probability distribution of different types of voters in each constituency.Finally, the total number of adults voting for Donald Trump and Joe Biden in each district was calculated by dividing each district's total annual population by the total number of adults voting for Donald Trump and Joe Biden.

1 Introduction

One of the most fascinating aspects of election research is election prediction.Regardless of the size of the election, the pre-election focus is always on who will win and who will win.The uncertainty of the election also attracts many experts and scholars to build various election prediction models by using the data of the population or individuals.These prediction models not only examine relevant election theories, but also contribute a lot to the practice of election operation. This model of election prediction, based on individual polling data, is fairly common, has been studied around the world, and is, by and large, accurate.

The model consists of two parts: first, the national population is divided into different segments, and a small sample of national individuals is used to estimate the voting intention of each segment of the population, supplemented by regional characteristics.Secondly, the distribution of population strata in each region (or sub-level) is obtained in accordance with the census data to predict the behavior and attitude of each region or sub-level, such as the vote share of each county and city in the country.The core advantages of this model can be summarized as the following three points:

Secondly, the distribution of population strata in each region (or sub-level) is obtained in accordance with the census data to predict the behavior and attitude of each region or sub-level, such as the vote share of each state in the country.The core advantages of this model can be summarized as the following three points: The model first USES the human location variable and the actual observed values of the country to form a statistical regression model, and then the estimated values of the public opinion of the state and the city can be obtained according to the census data.Since the regression coefficients of the model are based ona national sample, not by region, the approach takes into account the national "voting trends" at the time of the election and USES information from other counties to assist in obtaining estimates for a particular state.Such an analytical framework is in line with the electoral context, in which each district has its own candidate or regional factors, but national factors still have some influence.

*Code and data are available at: https://github.com/LuyuanHu/-Stat304-PS4; https://ipums.org/projects/ipums-usa; https://www.voterstudygroup.org/publication/nationscape-data-set.

2 Data

The national survey data chiefly used in the analysis is from Democracy Fund + UCLA Nationscape, one of the largest public opinion survey projects ever conducted. It interviewed people in nearly every county, congressional district, and mid-sized U.S. city in the leadup to the 2020 election. We use data collected in phrase 2 from June 25 to July 01, 2020 (N = 6,479 US Adults). Figure 1 plots the sample size by state, ranging from 5 respondents in Wyoming to 717 in California, with the median state (Los Angeles) including 81 respondents.

As noted, for the turnout models, respondents were classified as voters if they (1) reported being registered to vote at their current address, (2) said they definitely would vote or already had voted and (3) said they voted in the 2016 election (n=5,656). Candidate preference questions were asked only of respondents who said they had voted or definitely would vote (n=5,170). Of these, those who said they preferred/voted for Donald Trump were coded as 1 for the Vote Donald Trump variable (n=2,481), with all others (supporters of Joe Biden, third-party candidates and undecideds) coded as 0. Similarly, for the Vote Joe Biden variable, Biden early voters and supporters were coded 1 (n=2,719) and all others 0.

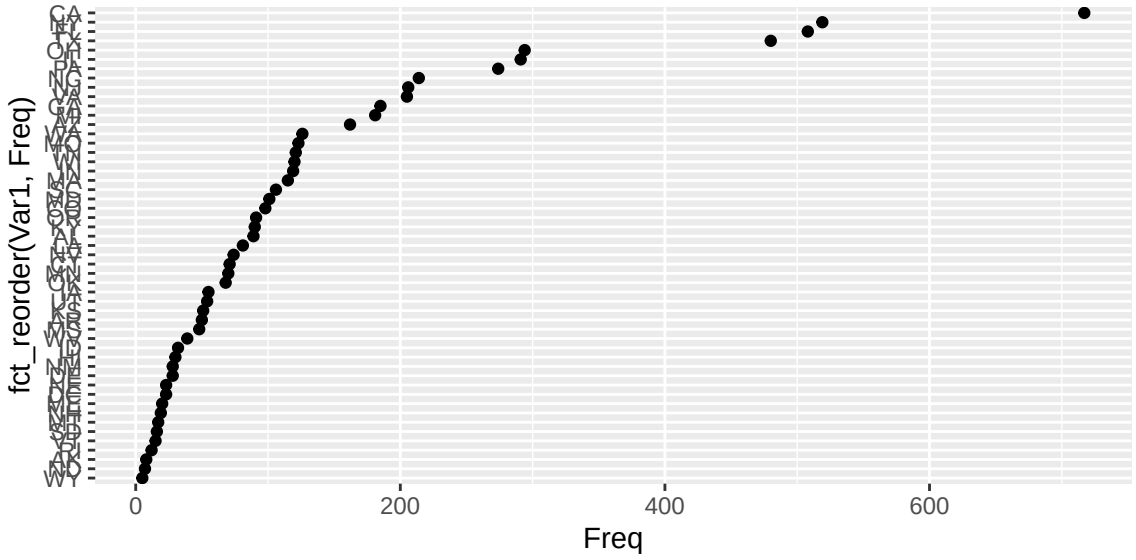

Figure 1: Sample size per state

Demographic group variables included as random effects were gender (male, female), age (18-29, 30-39, 40-49, 50-64, 65+), race/ethnicity (white non-Hispanic, black non-Hispanic, Hispanic, and other non-Hispanic),8 and education (less than high school, high school graduate, some college, four-year college graduate, post-graduate).

IPUMS USA (originally, the "Integrated Public Use Microdata Series") is a website and database providing access to over sixty integrated, high-precision samples of the American population drawn from sixteen federal censuses, from the American Community Surveys of 2000-present, and from the Puerto Rican Community Surveys of 2005-present. For estimates of uncertainty, MRP models can be estimated with full Bayesian methods; this analyses uses a more approximate maximum likelihood estimator since the focus is on point estimates. Models were estimated in `R` using the `naiveBayes` function in the `e1071` package(Meyer et al. 2020) and the glmer function in the `lme4` package(Bates et al. 2015).

3 Model

In this section, we will use MRP for state-level estimate. National-level surveys proved at least as accurate in 2016 as they have been in past elections – only slightly misstating the popular vote (overestimating Clinton's victory over Trump by 1-2 points), and more accurate on average than in 2012 ((Silver 2017); (Cohn, Katz, and Quealy 2016)). But state-level polling painted a different and ultimately much more inaccurate picture of the race.

MRP is a promising alternative that avoids the uncertain rigour of state polls and the prescience needed to predict where polls will be conducted.Instead, in this analysis, MRP relies on survey data at the national level, combined with statistical models and census data, to generate estimates of voter turnout and candidate selection at the state level.This approach utilizes MRP to provide highly accurate estimates of attitudes and behavior at the state level, even though the number of observations in each state is relatively small.

The statistical nature and substantial advantages of MRP can be found in papers(Park, Gelman, and Bafumi (2004); Ghitza and Gelman (2013)).The researchers started with a national survey data set, preferably with quite a few observations.Using basic demographic variables available in the state (or other subnational) level census data, multilevel statistical models can be used to predict outcomes of interest.Other state-level variables can be included in the model to improve the accuracy of the estimates.The coefficients of continuous variables are usually unmodeled (that is, fixed), while group variables are modeled as classified random effects.In the case of election preference estimation, multiple models are required, first estimating the probability of voting, then additional models estimating the preference of the candidate.

In the second stage of MRP analysis, the model estimates are used to predict the outcome variable for groups defined in a poststratification dataset. This dataset has an observation corresponding to each group defined for all combinations of the demographic variables included in the model. After predicting the outcome variable for each of the groups in the poststratification dataset, estimates can be aggregated to the state level, with the subgroup population sizes determining the relative weight of each group's estimate in the state-level estimate.

MRP provides a powerful way to generate opinion estimates at the state level by pooling information from similar groups in other states, in effect exploiting the homogeneity of attitudes at the subgroup level (as established in regression analysis).Through multi-level modeling, the results between multiple groups are partially aggregated.

First, we use naive bayes to fit a model. Assuming our survey sample

$(x_1^{(1)}, x_2^{(1)}, \dots x_n^{(1)}, y_1); (x_1^{(2)}, x_2^{(2)}, \dots x_n^{(2)}, y_2), \dots; (x_1^{(m)}, x_2^{(m)}, \dots x_n^{(m)}, y_m)$. There are 2 categories of feature outputs, defined as C_1, C_2, which denotes to vote Trump or Biden. From the sample we can learn the prior distribution of naive Bayes $P(Y = C_k)(k = 1, 2, ...K)$. And then we learned about conditional probability distributions$P(X = x|Y = C_k) = P(X_1 = x_1, X_2 = x_2, ...X_n = x_n|Y = C_k)$.The joint distribution P(X,Y) is defined as

$$P(X, Y = C_k) = P(Y = C_k)P(X = x|Y = C_k) \tag{1}$$
$$P(Y = C_k)P(X_1 = x_1, X_2 = x_2, ...X_n = x_n|Y = C_k) \tag{2}$$

Then

$$P(X_1 = x_1, X_2 = x_2, ...X_n = x_n|Y = C_k) = P(X_1 = x_1|Y = C_k)P(X_2 = x_2|Y = C_k)...P(X_n = x_n|Y = C_k)$$

Since it's a Bayesian model, it's certainly a posteriori probability maximization to determine the classification. We just need to calculate all K conditional probabilities $P(Y = C_k|X = X^{(test)})$, and then find the category corresponding to the largest conditional probability, which is the naive bayes' prediction.

4 Results

Here's what the model output looks like:

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     Biden     Trump
## 0.5340659 0.4659341
##
## Conditional probabilities:
##        race3
## Y            Black      Other      White
##   Biden 0.19615912 0.14769090 0.65614998
##   Trump 0.02725367 0.08962264 0.88312369
##
##        agegroup
## Y           18-29     30-44     45-54     55-64       65+
##   Biden 0.2363969 0.3036123 0.1326017 0.1641518 0.1632373
##   Trump 0.1200210 0.3197065 0.1687631 0.1933962 0.1981132
##
##        gender
## Y          Female      Male
##   Biden 0.5697302 0.4302698
##   Trump 0.4386792 0.5613208
##
##        income.6
## Y       $100k-200k  $30k-50k     $30k>  $50k-70k $70k-100k
##   Biden  0.1833562 0.2066758 0.3493370 0.1440329 0.1165981
##   Trump  0.2458071 0.2007338 0.2919287 0.1483229 0.1132075
##
##        educ6
## Y       College Degree Doctorate degree High School Graduate Masters degree
##   Biden     0.27800640       0.01920439           0.24737083     0.10745313
##   Trump     0.25419287       0.01886792           0.28563941     0.11792453
##        educ6
## Y       Some College
##   Biden   0.34796525
##   Trump   0.32337526
##
##        census_region
## Y       Northeast   Midwest     South      West
##   Biden 0.1998171 0.2144490 0.3516232 0.2341107
##   Trump 0.1755765 0.2117400 0.4077568 0.2049266
##
##        labor_force
## Y       in the labor force not in the labor force
##   Biden          0.5450389              0.4549611
##   Trump          0.5660377              0.4339623
```

Here's what the predicted probability of voting for Trump looks for the first ten test cases:

```
NB_Predictions[,1][1:10]
```

```
##  [1] 0.3168074 0.6632437 0.3591652 0.6794123 0.5263014 0.5163568 0.8246222
##  [8] 0.4590887 0.4164077 0.6794123
```

Below, we specify and train a logistic regression model using `glm()` and evaluate the predictions.

```
##
## Call:
## glm(formula = vote.2.binary ~ ., family = binomial(link = "logit"),
##     data = survey.logit %>% as.data.frame() %>% select(-vote.2))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8097  -1.1124  -0.4138   1.0798   2.5332
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -3.00358    0.20804 -14.437  < 2e-16 ***
## race3Other                         1.63743    0.17980   9.107  < 2e-16 ***
## race3White                         2.27848    0.15563  14.640  < 2e-16 ***
## agegroup30-44                      0.57708    0.10656   5.415 6.12e-08 ***
## agegroup45-54                      0.81637    0.12294   6.641 3.12e-11 ***
## agegroup55-64                      0.69657    0.11849   5.879 4.13e-09 ***
## agegroup65+                        0.75237    0.12398   6.068 1.29e-09 ***
## genderMale                         0.41036    0.06875   5.969 2.39e-09 ***
## income.6$30k-50k                  -0.36001    0.10948  -3.288  0.00101 **
## income.6$30k>                     -0.35053    0.10860  -3.228  0.00125 **
## income.6$50k-70k                  -0.20128    0.11674  -1.724  0.08467 .
## income.6$70k-100k                 -0.35706    0.12197  -2.927  0.00342 **
## educ6Doctorate degree              0.09340    0.25772   0.362  0.71706
## educ6High School Graduate          0.64005    0.10239   6.251 4.08e-10 ***
## educ6Masters degree               -0.10100    0.11914  -0.848  0.39661
## educ6Some College                  0.18692    0.09056   2.064  0.03901 *
## census_regionMidwest               0.12142    0.10656   1.140  0.25448
## census_regionSouth                 0.48647    0.09609   5.063 4.13e-07 ***
## census_regionWest                  0.05757    0.10488   0.549  0.58306
## labor_forcenot in the labor force -0.20697    0.07881  -2.626  0.00864 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5657.9  on 4094  degrees of freedom
## Residual deviance: 5088.2  on 4075  degrees of freedom
## AIC: 5128.2
##
## Number of Fisher Scoring iterations: 4
```

5 Discussion

While higher-quality polling in swing states likely would have improved predictions in the 2020 election, MRP provides an attractive alternative. As this paper demonstrates, even with relatively small state-level sample sizes, our MRP approach substantially outperforms leading polling aggregators in the 2020 election, and analyses of previous elections indicate the robustness of the technique.

This performance is likely related to factors including the quality of the underlying data and attributes specific to our approach. First, by using a single national-level survey, our MRP estimates are based on data collected with the same methods across states, while state-level surveys averaged by aggregators vary widely in methods and quality. To the extent that lowerquality or poorly devised polling methods produce inaccurate estimates, the presumed cancelingout benefits of aggregation can lead to biased and misleading results. Second, and relatedly, the analysis reported here is based on one of the most methodologically sound probability-based RDD surveys of its type in the country. These data may present advantages over non-probability data or voter registration lists; the latter suffer from sizable noncoverage and noisy weighting variables.

MRP also offers an alternative to traditional survey weighting and likely voter modeling that overcomes some of the challenges faced by standard survey weighting techniques – either iterative proportional fitting, which does not guarantee precise subgroup sizes, or cell weighting, which can be compromised by limited sample sizes. MRP is analogous in many ways to cell weighting, without the troubles associated with zero- or small-n cells. In the analysis presented here, the model estimates were poststratified on 10,200 cells, essentially a much finergrained weighting scheme than either rake or typical cell weighting

Future research may lead to additional improvements in the accuracy of the MRP approach employed here. Other strategies for estimating turnout (e.g., other deterministic operationalizations, continuous turnout variables or CPS-based models) could enhance subgrouplevel turnout estimates. Future research also could examine whether and how to poststratify on variables such as partisan identification or past vote (Wang et al. 2014; YouGov 2016), albeit with an eye to the inherent risks of doing so. Finally, the analysis could be conducted using full Bayesian methods, which would facilitate the calculation of uncertainty for the estimates.

References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Cohn, Nate, Josh Katz, and Kevin Quealy. 2016. "Putting the Polling Miss of the 2016 Election in Perspective." *New York Times*.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with Mrp: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57 (3): 762–76.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2020. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), Tu Wien*. https://CRAN.R-project.org/package=e1071.

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

Silver, Nate. 2017. "The Real Story of 2016." *FiveThirtyEight. Com, January* 19.