MAXIMIZING RETURN AND FORECASTING THE 2020 US PRESIDENTIAL ELECTIONS

Cesar Y. Villarreal Guzman

31 October 2020

Abstract

In this paper develop two models with the purpose of forecasting the 2020 United States presidential election. Multilevel regression with postratification (MRP) and Stacked regression with postratification models where constructed using ADDNAME survey data and postsratified using the ACS census dataset. We find that both our models forecast the victory of Joe Biden both in the popular vote and the electoral college vote.

Keywords: Forecasting; U.S. 2020 Election; Trump; Biden; Multilevel Regression with Poststratification

1. Introduction

It is a common theme in the social sciences and statistics literature to attempt constructing a model that accurately predicts the outcome of a presidential election. Often the purpose of these experiments ranges from finding a novel statistical model to a commentary of the impact at the economical, social or political level that the expected winner would cause. All such experiments, however, share the same underlying problem, predicting the outcome of the elections.

Isidore (2020) discusses how the 2020 United States presidential elections are now the most bet-on event in history. As of Wednesday morning, a week prior to the elections, about $284 million USD had already been wagered on the Betfair Exchange by British bettors. Betfair Exchange is one of the largest betting exchanges in the world based on London. Moreover, in this article he predicts that by the time the election is held approximately $520 million USD could be wagered.

Gambling and information theory tell us that if we where to know for certain the outcome of the 2020 U.S. presidential election we could double, our money. It is with this in mind that for this paper we set ourselves the premise of theoretical bet. We attempt to forecast the outcome of 2020 U.S. presidential elections and comment, based on our results, the optimal strategy that would maximize the return of the bet i.e. comment on the most likely to win candidate.

Two models are used to forecast, a multilevel regression with poststratification (MRP) model and a stacked regression with poststratification model (SRP). To train these models we used the Democracy Fund + UCLA Nationscape survey data set to construct the models and the ACS U.S. census data set for poststratification.

This paper is structured in the following manner. Section two contains a commentary on both data sets, what they are and where they come from. This is followed by a brief discussion on our data cleaning methodology. In section three we elaborate on how MRP and SRP work and how they are implemented. Section four is where we present our estimates for both models and finally in section five we comment on this estimates and argue on the most likely winner of the U.S. presidential election.

2. Data

2.1 Democracy Fund + UCLA Nationscape Data Set

Nationscape (Tausanovitch and Vavreck 2020) is a survey conducting 500,000 interviews of Americans from July 2019 through December 2020, covering the 2020 presidential election. The survey includes online interviews with roughly 6,250 people per week starting July 10, 2019. This is the survey used to fit both models.

As in almost all contemporary survey research, the Nationscape survey is not a random sample of the population of interest. In particular, the Nationscape survey is a convenience sample selected on a set of demographic criteria by a market research platform that runs an online exchange for survey respondents. Such samples where provided by the market research company Lucid.

With this description we can classify this survey as an online non-probability sample. At its core a non-probability sample is obtained by non-random selection. This contrasts with probability samples, where, a full list of the target population is available from which respondents are selected uniformly at random, although the method for random sampling may vary depending on the study.

The survey is divided into two phases. Phase 1 of the data, released in January of 2020, includes approximately 156,000 cases collected over 24 weeks, beginning from the week of July 18, 2019 and concluded with the week of December 26, 2019. Phase 2 of the data, released in September of 2020, includes a re-release of Phase 1 data and new data from January 2020 to July 2020 (Phase 2 data). Each weekly survey is released as its own data set, and combining all data set results in 318,697 cases.

2.2 ACS Census Data Set

The American Community Surveys (ACS) (Steven Ruggles and Sobek 2020) is a project of the U.S. Census Bureau that has replaced the decennial census as the key source of information about American population and housing characteristics. This survey has been conducted

since 2000 and the most recent on 2018. We used the 2018 version as our poststratification data set.

An account from IPUMS USA is required to access the data. The database allows for the creation of a customized data set. In particular we chose the 2018 ACS survey and selected the following variables: sex, age, state, race, and hispanic. Automatically ten other variables are appended to the selection. Out of these ten there is one in particular that was also used. According to the ACS code book, "PERWT indicates how many persons in the U.S. population are represented by a given person in an IPUMS sample." We use this variable in the poststratification step to obtain better population estimates.

2.3 Methodology

To train our models we used data from June 2020, since this are the closest to the elections. Moreover, it is well known that political views have shifted because of the U.S. treatment of the pandemic and therefore surveys conducted in 2019 and early 2020 do not take this important topic into consideration. The resulting data set contained $20,157$ entries. Next we selected the columns for: sex, age, state, race, hispanic and vote choice. This resulted in a rectangular data set of $20,157$ rows and 6 columns.

To clean the data we first removed 333 rows with missing values. Then we applied the common method of data binning to the age column. We created seven age groups that range from 18 to 93. A similar procedure was used for the race category where we reduced the number of race categories. The hispanic category was transformed into a binary feature, either 'hispanic' or 'not hispanic'. The same procedure was done on the vote choice column where we used only data which contained 'Joe Biden' or 'Donald Trump' as values.

The ACS data set followed a very similar cleaning procedure. We first selected the study columns, created and categorized age into age groups, reduced the number the race categories, and transformed the hispanic column to a binary variable. We performed this cleaning procedure while being consistent with the values of used in the Nationscape data set. Variable renaming was used for this purpose.

Proportions for each variable can be seen in Figure 1 and 2. Overall the sample represents very well the distributions for race, gender, and hispanics. However, there are some noticeable inconsistencies on the distribution of age groups and state population. This issue is solved by the poststratification step of our model.

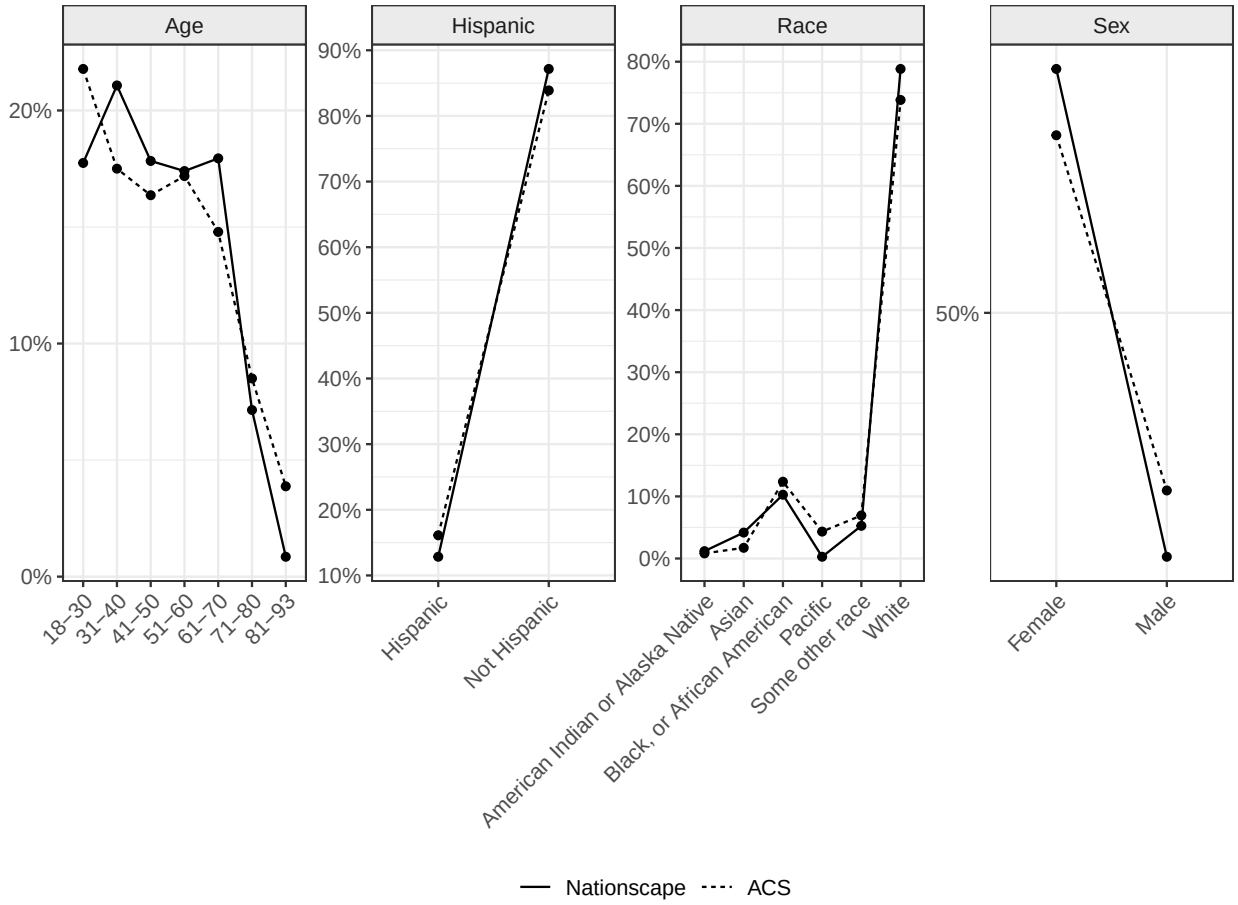

Figure 1: ACS and Nationscape Raw data

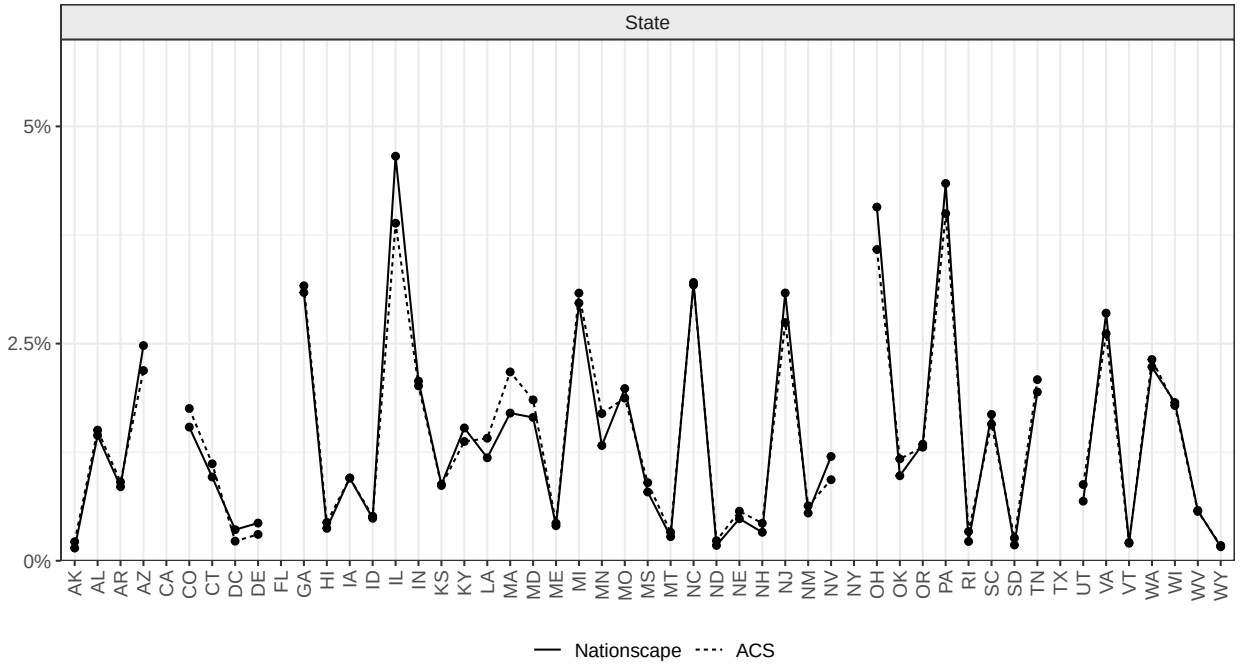

Figure 2: ACS and Nationscape Raw Data

3. Model

3.1 Multilevel Regression and Poststratification

To transform the Nationscape convenience sample into accurate estimates of voter intent, we make use of demographic information provided by respondents, then we poststratify raw responses to mimic a representative sample of likely voters. The core idea of MRP is to partition the population into cells based on combinations of various demographic attributes. Next, use the sample to fit a model which is then used to generate estimate the response variable within each cell. Finally, we aggregate the cell-level estimates up to a population-level estimate by correcting each cell using the relative proportion in the population.

The statistical technique MRP or multilevel regression and poststratification (Little 1993; Park, Gelman, and Bafumi 2004) allows researchers to infer quantities in the population from a sparse and/or non-representative sample (Wang et al. 2015), accomplished by combining two techniques: small-area estimation and nonresponse adjustment (Kennedy and Gelman 2020). Furthermore, MRP is widely used in the political science literature and has been proven to be an effective technique for non-probability based convenience samples, and the reason why we choose this technique as our main analysis tool.

Applying MRP in our setting comprises of the two mentioned steps. However, being specific to our problem, first, we fit a Bayesian hierarchical model to obtain estimates for our poststratification cells; second, we average over the cells, weighting the values by its relative proportion to the population. To generate the cells we consider all possible combinations of gender (2 categories), Hispanic origin (2 categories), age (7 categories), race (6 categories)

and state (51 categories), thus partitioning the data into 8568 cells.

We fit a multilevel regression model relating Biden support with age, race, Hispanic origin and state. Symbolically, for an individual i we denote $y_i = 1$ if the respondent support Joe Biden in the upcoming elections, and 0 otherwise. The non-hierarchical part of the model can be written as

$$P(y_i = 1) = \text{logit}^{-1}(\beta X_i)$$

where X contains indicator variables for gender and Hispanic origin and an interaction term with age and race. Adding the varying intercept, the final model can be written as

$$P(y_i = 1) = \text{logit}^{-1}\left(\beta X_i + \alpha^{\text{state}}_{j[i]}\right)$$

$$\alpha^{\text{state}} \sim N(0, \sigma_{\text{state}})$$

To implement this model we use the **brms** (Bürkner 2018) package for the R (R Core Team 2020) language. One important aspect of the Bayesian framework is the selection of priors. By default brms normalizes and re-scales the data and sets priors that reflect this transformation, this will be our prior of choice.

3.2 SRP

4. Results

5. Discussion

Appendix

References

Bürkner, Paul-Christian. 2018. "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal* 10 (1): 395–411. https://doi.org/10.32614/RJ-2018-017.

Isidore, Chris. 2020. "The Us Election Is Likely the Most Bet-on Event in History." https://edition.cnn.com/2020/10/30/business/us-presidential-election-wagering-record/index.html.

Kay, Matthew. 2020. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

Kennedy, Lauren, and Andrew Gelman. 2020. "Know Your Population and Know Your Model: Using Model-Based Regression and Poststratification to Generalize Findings Beyond the Observed Sample." http://arxiv.org/abs/1906.11232.

Little, R. J. A. 1993. "Post-Stratification: A Modeler's Perspective." *Journal of the American Statistical Association* 88 (423): 1001–12. http://www.jstor.org/stable/2290792.

Ornstein, Joseph T. 2019. "Stacked Regression and Poststratification." *Political Analysis* 28 (2): 293–301. https://doi.org/10.1017/pan.2019.43.

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. https://doi.org/10.1093/pan/mph024.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. 2020. "IPUMS Usa: Version 10.0 [American Community Surveys 2018]." https://doi.org/0.18128/D010.V10.0.

Tausanovitch, Chris, and Lynn Vavreck. 2020. "Democracy Fund + Ucla Nationscape, October 10-17, 2019 (Version 20200814)." https://www.voterstudygroup.org/publication/nationscape-data-set.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31 (3): 980–91. https://doi.org/10.1016/j.ijforecast.2014.06.001.

Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1–29. http://www.jstatsoft.org/v40/i01/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.