Prediction of the 2020 Presidential Election*

Spoiler Alert: Biden Wins

Anees Shaikh, Jaffa Romain, Lu Mu, and Cameron Fryer

02 November 2020

Abstract

The purpose of this analysis is to forecast the results of the upcoming United States 2020 presidential election. In this paper, we first consider the Democracy Fund + UCLA Nationscape Wave 50 dataset, which contains the results of a survey (conducted June 25-July 1, 2020) on American voter attitudes. Thereafter, the survey data is used to train a model relating voter intent to a few explanatory variables. The model is then applied to the post-stratification dataset; namely, the results of the 2018 1-year American Community Survey (ACS). Since the ACS data pertains to individual persons and their characteristics, the associated use of our model allows us to conclude that Joe Biden will be the next President of the United States.|

1 Introduction

To accurately predict the winner of the United States 2020 presidential election is difficult, as many different factors are involved. . . . ..

2 Data

As previously mentioned, our data is from two different datasets; the Democracy Fund + UCLA Nationscape Wave 50 dataset, and the 2018 1-year American Community Survey (ACS) dataset. They each contain both qualitative and quantitative data. In regard to the Nationscape Wave 50 dataset, a few of its strengths are the variety in variables considered, and the quality of the survey responses. With respect to its weaknesses, it is likely that the data contains sampling error and/or coverage error. One of the various strengths of the ACS dataset is its ability to be customized with pre-specified (or pre-selected?) variables and sample size before download. Although there is a wide range of variables to choose from, being able to select only those which we needed made the exploratory data analysis process much more manageable than it could have otherwise been (with all possible variables). Furthermore, since sample size can be pre-specified to a limit of about 3.2 million people, the larger the sample size is set to, the greater the odds that the target population is represented with the data. In terms of weaknesses of the ACS dataset, there are none all that significant. Some minor concerns, however, are potential measurement error, processing error, and/or adjustment error in the data. The Democracy Fund + UCLA Nationscape is "one of the largest public opinion surveys ever conducted – interviewing people in nearly every county, congressional district, and mid-sized U.S. city in the leadup to the 2020 election." Consequently, interviewing "roughly 6,250 per week" was part of the survey methodology employed by Nationscape. As such, the Nationscape Wave 50 dataset contains results for 6,479 interviews conducted during the week of June 25-July 1, 2020. It can be accessed by visiting https://www.voterstudygroup.org/publication/nationscape-data-set and submitting a request to access the

*Code and data are available at: https://github.com/aneesshake/elections_prediction

data. By providing your full name and email address, you will then be emailed a link to the page where it can be downloaded. At first glance of the Wave 50 data, we noted that it contains 265 variables. This is noteworthy because it allows a more complete picture of voter attitude to be derived from the data compared to if there were less variables. Another strength of this dataset is the overall quality of responses from those surveyed. More specifically, the data does not contain any duplicate records, inappropriate entries, or poor entries (e.g. typos, misspellings); hence, making it easier to work with. This high quality of survey responses can be partially attributed to the fact that "Nationscape samples are provided by Lucid, a market research platform that runs an online exchange for survey respondents." What is more, the individuals belonging to the sample are required to "complete an attention check before completing the [online] survey." With this method of sample selection, along with the precaution taken to ensure their attentiveness, it is unlikely for survey respondents to submit low-quality responses. The main weakness of the Nationscape Wave 50 dataset is the likelihood that it includes sampling error and/or coverage error. While the goal of the Nationscape survey is to determine American voter attitudes, making inferences based only on the Wave 50 dataset would result in sample bias. This is because the target population is all American citizens age 18 years and older, whereas the sample frame is Lucid's online exchange for survey respondents, and the sample is a set of 6,479 individuals from said online exchange which match "a set of demographic criteria." It is very unlikely that a sample of 6,479 American citizens during a single week is representative of American voter attitudes as a whole. This may lead one to wonder why Nationscape chose to use the described sampling approach. Apart from the fact that the survey was conducted every week, the reason is trade-offs. For example, although there is potential coverage error in the data–a common result of online surveys–some trade-offs are a lesser rate of non-response, lower cost, convenience and accuracy. Whilst on the topic of non-response, it is worth mentioning that the variable with greatest amount of non-response in the Wave 50 dataset is household income at rouhly 5.5%. Nonetheless, this is a much smaller percentage of non-response in household income than is typically seen in political surveys, and Nationscape handles it by not weighting income for non-respondents. Similarly, a statement from Nationscape in its "Representativeness Assessment" implies an overall lack of concern for non-response in the data, as "previous evaluations of the samples Lucid provides have found them to be of high quality." The last aspect of the Nationscape Wave 50 dataset to be considered is the survey questionnaire itself. Strengths of the questionnaire include its structure, the design and wording of the questions being asked, and the variety of choices for closed-ended questions. As to the structure of the questionnaire, it contains an extensive amount of questions from which differing political views can be revealed. Furthermore, the survey questions are asked in an order that does not influence the answer of respondents to subsequent questions. In terms of the design and wording of the questions, they are closed-ended, and avoid the use of tricky wording. This is a strength of the questionnaire because it makes comparing the responses of individuals and thus, statistical analysis, much simpler than it would be with open-ended questions and/or troublesome language. The diverse number of options available to select as answers to the questions is a strength of the questionnaire because they are both mutually exclusive and exhaustive; hence, ensuring survey respondents are able to answer every question. With respect to weaknesses of the questionnaire, it seems that there are none other than its sheer length.

3 Model

In order to forecast the popular vote of the U.S 2020 elections, multilevel logistic regression with post-stratification (MRP) was used. MRP is useful for when generalizing from a possibly non-representative poll kennedy2020know. The individual survey gives sample data on voters of the general U.S population, and post-stratification allows us to re-weight estimates, adjusting bias between our sample and the target population, so that we have a representative sample of likely voters **?**. To achieve this, cells are constructed using variables in the cleaned Nationscape survey such as age, household income, and race. The model is then trained on the survey using the proportions given by the ACS post-stratification data set. This approach gives us an advantage when attempted to forecast voting, as we can use a broad survey to speak to subsets in the population, and also tends to be less expensive to collect than non-probability samples **?**. However, using this approach places limitations on how much we can interpret from the estimates. Although we can estimate voting intention in different demographic groups, it doesn't tell us any qualitative information on

voting patterns. For example, we might be able to see how different age categories vote, but we won't have a measure of the policy preferences, or their views on the candidates.

Using the Nationscape data, a multilevel logistic regression model can be fitted to the survey data set to model the proportion of voters who will vote for Trump. Logistic regression can be used to model a binary dependent variable, which is the choice between the Republic candidate, Donald Trump, and the Democratic candidate, Joe Biden in our case. This would be the main reason for choosing this model, over another model such as a linear regression model, where the output isn't necessarily binary. However, this imposes limitations. As we cannot take into consideration any other candidates running in the election due to the binary outcome. Another weakness is **** The variable 'vote_2020' in the data set was used as dependent variable, with the variable return 1 for voters intending to vote for Trump, and 0.

Our model uses the variables pertaining to age, state, race, sex, household income, and whether a voter is Hispanic to predict whether a respondent will vote for Trump or Biden in the elections. The model takes the form: *** TO DO: Regression and Other Stories to cite equation **

$$Pr(y_i = 1) = logit^{-}1(X_i * \beta)$$

where $y_i = 1$ represents a voter voting for Trump in the election. $Pr(y_i = 1)$ is the probability of a voter choosing to vote for Trump. The $X_i\beta$ are the linear predictors, where β represents the fitted coefficients for each independent variable X_i in the model. The coefficient values signify the mean variable changes in a one unit shift in the given variable β. The logit function $logit(x) = log(x/1 - x)$ maps the range (0, 1) to $(-\infty, \infty)$. Its inverse function $logit^{-}1(x) = e^x/(1 + e^x)$ maps back to the unit range. The model's output is bounded between 0 and 1, mapping the outputs into a binary outcome. The output tells us the probability that some person i, will vote for Trump depends on their age, sex, state, household income, and whether the voter is Hispanic.

We fit our regression model using the `glm()` function in [R]. We can see from table 1, which was made using `kable` from `knitr`, that the fitted model results in the following coefficients:

```
# loading data sets and cell counts for MRP
cell_counts <- readRDS(here::here("inputs/cleaned_data", "cell_counts.rds"))
individual_survey <- readRDS(here::here("inputs/cleaned_data","individual-survey.rds"))
post_strat <- readRDS(here::here("inputs/cleaned_data","post-strat.rds"))


#Loading in the final model and the coefficents of the model in this cell. Please note that the outputs



final_model <- readRDS(here::here("outputs/model","final_model.rds"))
coefs <- readRDS(here::here("outputs/model","coefficients.rds"))
knitr::kable(coefs, caption = " Table 1: Model Coefficients", digits = 2)
```

In logistic regression, collinearity implies that we have predictor variables that are highly correlated; that is, they have a linear relationship. This is possible when you have a larger number of variables in a model, but having collinearity in multiple variables can lead to unstable estimates and inaccurate variances, which can affect confidence intervals. This could lead to incorrect inferences about the relationship between our response variable and explanatory variables **?**. Thus, it is imperative to check for any correlated variables in our model to avoid these issues. The value of the Variance Inflation Factor(VIF) in for the explanatory variables can give us a measure of collinearity our model. VIF values should be less than 5 to guarantee that collinearity is not an issue. We can see in table 2 below that all VIF values are less than 5, so collinearity will not be a concern for our model.

Table 1: Table 1: Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	1.52	0.71	2.15	0.03
sexmale	0.40	0.07	5.45	0.00
age25-44	0.35	0.20	1.73	0.08
age45-64	0.46	0.21	2.22	0.03
age65+	0.38	0.21	1.81	0.07
raceblack	-3.13	0.45	-6.98	0.00
racechinese	-2.29	0.59	-3.87	0.00
racejapanese	-1.97	0.79	-2.50	0.01
raceother asian or pacific islander	-1.38	0.47	-2.95	0.00
raceother race, nec	-1.07	0.44	-2.42	0.02
racewhite	-0.55	0.41	-1.33	0.18
hispanmexican	-1.86	0.57	-3.26	0.00
hispannot hispanic	-1.22	0.55	-2.22	0.03
hispanother	-1.45	0.58	-2.51	0.01
hispanpuerto rican	-1.73	1.47	-1.18	0.24

Table 2: Checking Model Collinearity

	GVIF	Df	GVIF^(1/(2*Df))
sex	1.08	1	1.04
age	1.24	3	1.04
race	1.64	6	1.04
household_income	1.81	23	1.01
hispan	1.43	4	1.05
state	2.24	50	1.01

```
cor <- readRDS(here::here("outputs/model","cor.rds"))
knitr::kable(cor, caption =  "Checking Model Collinearity", digits = 2)
```

```
 anova_model <- anova(final_model, test="Chisq")
knitr::kable(anova_model, caption =  "Analysis of Deviance Table", digits = 2)
```

```
# make predictions using our model with data from post-strat data set

cell_counts
```

```
## # A tibble: 7,131 x 8
```

Table 3: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	3540	4907.75	NA
sex	1	67.34	3539	4840.41	0
age	3	29.72	3536	4810.69	0
race	6	360.92	3530	4449.77	0
hispan	4	23.61	3526	4426.16	0

```
##    state_name  sex   age   race          hispan  num_records    prop state_count
##    <fct>       <fct> <chr> <chr>         <fct>         <int>   <dbl>       <int>
##  1 connecticut male  18-24 american ind~ not hi~           1 4.00e-5       24986
##  2 connecticut male  18-24 american ind~ puerto~           2 8.00e-5       24986
##  3 connecticut male  18-24 american ind~ other             1 4.00e-5       24986
##  4 connecticut male  18-24 black         not hi~          89 3.56e-3       24986
##  5 connecticut male  18-24 black         puerto~           4 1.60e-4       24986
##  6 connecticut male  18-24 chinese       not hi~           5 2.00e-4       24986
##  7 connecticut male  18-24 japanese      not hi~           1 4.00e-5       24986
##  8 connecticut male  18-24 other asian ~ not hi~          31 1.24e-3       24986
##  9 connecticut male  18-24 other race, ~ mexican           8 3.20e-4       24986
## 10 connecticut male  18-24 other race, ~ puerto~          20 8.00e-4       24986
## # ... with 7,121 more rows
```

```
cell_counts$estimate <- predict(final_model, newdata = cell_counts, type = "response")

# post-stratified estimates - number of trump votes
estimate_votes <- cell_counts %>%
  mutate(predict_prop = prop * estimate) %>%
  group_by(state_name) %>%
  summarise(predicted_vote = sum(predict_prop)) %>% ungroup()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
estimate_votes
```

```
## # A tibble: 51 x 2
##    state_name    predicted_vote
##    <fct>                  <dbl>
##  1 connecticut            0.525
##  2 maine                  0.578
##  3 massachusetts          0.534
##  4 new hampshire          0.573
##  5 rhode island           0.545
##  6 vermont                0.577
##  7 delaware               0.497
##  8 new jersey             0.503
##  9 new york               0.490
## 10 pennsylvania           0.545
## # ... with 41 more rows
```

```
estimate_votes %>% mutate(status = if_else(predicted_vote>=0.50,"win","loss")) %>% group_by(status) %>%
)
```

```
## # A tibble: 2 x 2
##   status     n
##   <chr>  <int>
## 1 loss      15
## 2 win       36
```

*** add another diagnostic test

4 Results

Looking at (Figure **??**), we observe a shift in age distributions between the two datasets. There is a slight peak in the 25-44 age group in the nationscape data. We see this flatten and then observe a significant increase in the 65+ age group, and a minor increase in the 18-24 age groups.

All our analysis was done using R[R Core Team, 2020]

Comparisons to ACS data and Nationscape data

throw a few distributions in here to show the differences in distributions, and then tie each one in the discussion section to the potential implications of such a distribution. For instance, ACS data shows that there are more higher income earners than in Nationscape. Our model shows a strong effect between income and voting for a certain candidate.

4.1 comparison on age

4.2 comparison on race

4.3

5 Discussion

5.1 First discussion point

Talk about correlations between education and result

To begin this discussion, we begin with a quick explanation of the US Federal voting system. This context is important for us to understand some key points and drawbacks with interpreting our model.

Talk about previous research that has been done with similar models and the output of this model. Our model seems to indicate that Biden is going to win the popular vote. Throw some crap in here about 538 has certain polls that show a slight trump win. Show falling trump approvals and other things like that. This model has been thoroughly researched and has shown to be effective in multiple other contexts(needs citation)

One of the key things we have to explain here is how the electoral college works. Readers may remember the 2016 US election where Hillary Clinton won the popular vote, however, this didn't mean that she won the federal election. This was due to her not winning the electoral college. The electoral college is a body that is elected to determine who will be the next president of the US. This happened in 2000 when Bush didn't win the popular vote, but did end up winning the electoral college ## Second discussion point

5.2 Third discussion point

5.3 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

6 References

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [URL].

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

References

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