TBD*

TBD

Arjun Dhatt, Benjamin Draskovic, Yiqu Ding, Gantavya Gupta

02 November 2020

Abstract

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

2 Data

In this analysis, we collected data from the Nationscape dataset — a collaboration between the Democracy Fund Voter Study Group and UCLA Political Scientists. The Nationscape dataset that we will be analyzing was collected from June 25th, 2020 to July 1st, 2020. Using the data from Nationscape, we would like the predict the overall popular vote of the 2020 American presidential election using multilevel regression with post-stratification.

The samples are provided by Lucid, "a market research platform that provides access to authentic, targeted audiences". In order to ensure that an appropriate sample is selected, samples are selected based on a set of demographic quotas based on a multitude of factors such as age, gender, ethnicity, etc. Respondents of the survey are sent directly to an online survey (conducted in English) operated by the Nationscape team and forced to complete an attention check to ensure the answers they provide are meaningful. Lastly, the data is weighted to be representative of the American population using factors such as — gender, four major census regions, race, etc.

The population in this dataset is all voting residents in the USA that are eligible to vote. This does not include a large population of those that reside in the US such as non-citizens or those who are incarcerated, but this is representative of a real election. The population frame is all voting residents in USA that have access to a computer in order to complete the online survey. As mentioned before, the survey is taken online, therefore it excludes those without access to a device that can take the online survey. The sample in this case is all respondents who complete this survey online.

The data is collected using a convenience sample selected based on a set of demographic criteria. Because convenience sampling was used over random sampling, this can result in some biases in our dataset.

The respondents of the survey are found through Lucid, which "runs an online exchange for survey respondents". The respondents from the survey on Lucid are sent to the online Nationscape survey where they are prompted to complete the survey. Respondents who complete the survey not choose not to disclose their household income; this results in a large number of non-respondents because they may not be comfortable providing sensitive information. In an effort to account for the respondents who choose not to provide income, non-respondents are not weighed for income.

*Code and data are available at: https://github.com/STA304-PS4/Trump-vs-Biden.

The variables that we used in our dataset are as follow: - age o “What is your age” o It provides an integer value greater than or equal to 18 - education o “What is the highest level of education you have completed?” o An extensive list of options are provided - Hispanic o “Are you of Hispanic, Latino, or Spanish origin?” o An extensive list of options are provided - household_income o “What is your current annual household income before taxes? o Provides an extensive list of options beginning with $5,000 increments and then increases to $25,000 increments o This is the variable in our dataset that many people did not feel comfortable answering and was left as N/A. When cleaning the dataset, we removed all responses that left this answer as N/A - race_ethnicity o “What is your race” o An extensive list of options are provided. - extra_covid_worn_mask o “Have you done any of the following in the past week? - Worn a mask when going out in public” o Yes or No options are available for the respondent to fill out - state o “Respondent’s state, as a two-character postal abbreviation. Calculated based on entered ZIP code.” o Manually filled out by respondent - trump_biden o “If the general election for president of the United States was a contest between Joe Biden and Donald Trump, who would you support? “ o Options are either ‘Joe Biden’, ‘Donald Trump’, or ‘Don’t Know’

We also created another variable called ‘binary’ which outputs the result of the variable ‘trump_biden’ as a 1 or a 0 making it binary, therefore easier to analyze.

Key features of data?

The dataset we used contains many strengths and weaknesses. One of the strengths is due to the fact that the survey is distributed online. In today’s age where most things are online, this ensures that we can reach a wide part of the United States. At the same time, this can also be viewed as a weakness because certain age groups of the United States may not be tech-savvy enough to complete an online survey. The older population of the united states (e.g. 75+) may not be able to complete the survey because it is online which means that our data may not accurately reflect the older populations views regarding the 2020 election. A potential solution to this weakness is to use administer multiple forms of the survey to ensure that all people can access the survey.

Another weakness of the data comes from the fact that we use convenience-based sampling which inherently contains bias. Because the sampling method is based on convenience, we may be unlucky when sampling and get a sample group that inaccurately represents the population of the US making our results meaningless.

Lastly another weakness of our dataset is due to the fact that the survey data we are analyzing is collected from June 25th, 2020 to July 1st, 2020. Between July and November (the month of the election), many voters may not have an informed vote and their opinions may still be swayed after watching crucial debates and hearing what the leaders have to say; our data doesn’t take into consideration the people whose votes may change in the most crucial months of the election. A solution to this weakness can be to analyze data that may have occurred closer to the actual election so that the participants have an informed vote.

3 Model

The model we use to predict the result of the election is MRP, multilevel regression with post-stratification. The advantage of this model is that it has better performance when estimating behaviors of a particular subgroup of the population. One can interpret MRP as a combination of multilevel logistic regression and post-stratification. Logistic regression allows us the simplicity of analyzing the categorical response variable while incorporating both numerical and categorical explanatory variables; while combined with post-stratification, it yields a much narrower confidence interval for the estimation.

We are interested in how variables such as age, education, race, family income, geographic region, and mask-wearing decision affect citizens’ voting intentions, particularly between the major party candidates, Donald Trump for the Republicans and Joe Biden for the Democratic Party. The response variable that we are interested in is the chances of a citizen voting for Trump in the coming presidential election. It is a categorical variable with two levels: 1 represents intending to vote for Donald Trump, and 2 means planning to vote for Joe Biden.

Ideally, we ought to run a pre-analysis to determine the variables of interest. However, we will leave that open for further analysis and future elections due to the limited time and budget. We chose the explanatory

variables based on our understanding of a respondent's significant characteristics that could potentially affect his/her voting intention. Except for respondents' age, education, income level, and state(common in statistical studies), we are very interested in the respondent's habits towards mask-wearing. Firstly because the US is the country with the most COVID cases globally, the influence of COVID on the electoral votes is inevitable. Secondly, since Trump and Biden hold quite opposites opinions towards handling the pandemic, this parameter will likely represent part of the respondent's voting intention in the same way as the more familiar political indicators.

First, we train a multilevel logistic regression model using raw data from the voter study group towards the variable of interest using the seven explanatory variables. We then apply this model to our post-stratification data set to get the estimates.

Logistic regression estimates $\beta_0...\beta_k$ in (1)

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{1}$$

where p is the probability of event A that we are interested in, $\beta_0$is the intercept, $x_1...x_K$ are our variables of interest and $\beta_1...\beta_k$ are parameters for each of these variables. Based on the result, we are able to estimate p for a particular case given all the variables. We use `as.factors()` to incorporate dummy variables for all the categorical variable.

The main logic behind post-stratification is that we divide the sample data into strata based on a few attributes such as state, income, education level. Then we come up with a weight for each stratum to adjust its influence towards our prediction. Namely, if a stratum is over-represented, we want to reduce its impact on the prediction result; if a stratum is under-represented, we want to increase its influence. In this sense, the combination with logistic regression allows subgroups with a relatively small size to 'reference' from other subgroups with similar characteristics.

To mimic the population as closely as possible, we need a post-stratification dataset large enough to refer to. In this report, we use the ACS because the census happens less frequently than the ACS which is updated annually.

The post-stratification estimate is defined by $\hat{y}_{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$ where $\hat{y}_j$ is the estimate of y in cell j, and N_j is the size of the j-th cell in the population. An illustration using the education variable is $\hat{y}_{edu}^{PS} = \frac{\sum_{j \in J_{edu}} N_j \hat{y}_j}{\sum_{j \in J_{edu}} N_j}$, we can get an estimation given any level of education of respondents, J_{edu} denotes all subsets of education levels, $\cup_{i \in J} J_i^{edu}$ = all possible education levels.

(1) produces a proportion for the post-stratification based on regression, instead of merely averaging the sample in each stratum. Specifically, all possible cells are determined from combining:

- 4 different age groups;
- 6 education levels
- 5 races;
- 6 income levels;

That is 720 cells in total. We fit the logistic regressions for estimating candidate support in each cell. We used `function` from `package` to fit (2).

$$P(Trump) = logit^{-1}(\beta_0 + \beta_a ge + \beta_{j[i]}^{edu} + \beta_{j[i]}^{race} + \beta_{j[i]}^{income} + \beta_{j[i]}^{state} + \beta_{j[i]}^{mask}) \tag{2}$$

(2) is the model that we fit using the survey data, with $\beta_{j[i]}^{var} \sim N(0, \sigma_{var}^2)$. β_0 is the intercept,each of $\beta_{j[i]}^{var}$ represents the parameter for the i-th respondent in j-th cell. For example, $\beta_{j[i]}^{edu}$ can take values from β from any of the six education levels. Then we predict the results by summing up estimates for each cell to the population. The results of the logidtic regression is summarised in table 1.

4 Results

4.0.1 Summary Table

Table 1: statistics summary for logistic regression

variable	$\hat{\beta}$	Std. Error	p-value
intercept	-0.00	0.26	0.99
$36 < age < 55$	0.42	0.07	$6.06e^{-9}$
$56 < age < 75$	0.37	0.07	$8.91e^{-7}$
$age > 76$	0.62	0.18	0.00
Asian or Pacific Islander	-0.75	0.27	0.01
black	-1.89	0.26	$1.25e^{-12}$
nec	-0.75	0.26	0.01
white	0.09	0.24	0.68
$income > 150k$	0.14	0.11	0.22
$income < 25k$	-0.38	0.10	0.01
$25k < income < 50k$	-0.27	0.10	0.01
$50k < income < 75k$	-0.19	0.10	0.06
$75k < income < 100k$	-0.31	0.11	0.01

5 Discussion

education_grouped	president_predict
Completed 1 Higher Education Degree	0.0954821
Completed 1 Higher Education Degree or more	0.0634028
Completed HS	0.1879795
Completed Some HS	0.0207523
Less than HS	0.0190661
Participated in Higher Education	0.1142789

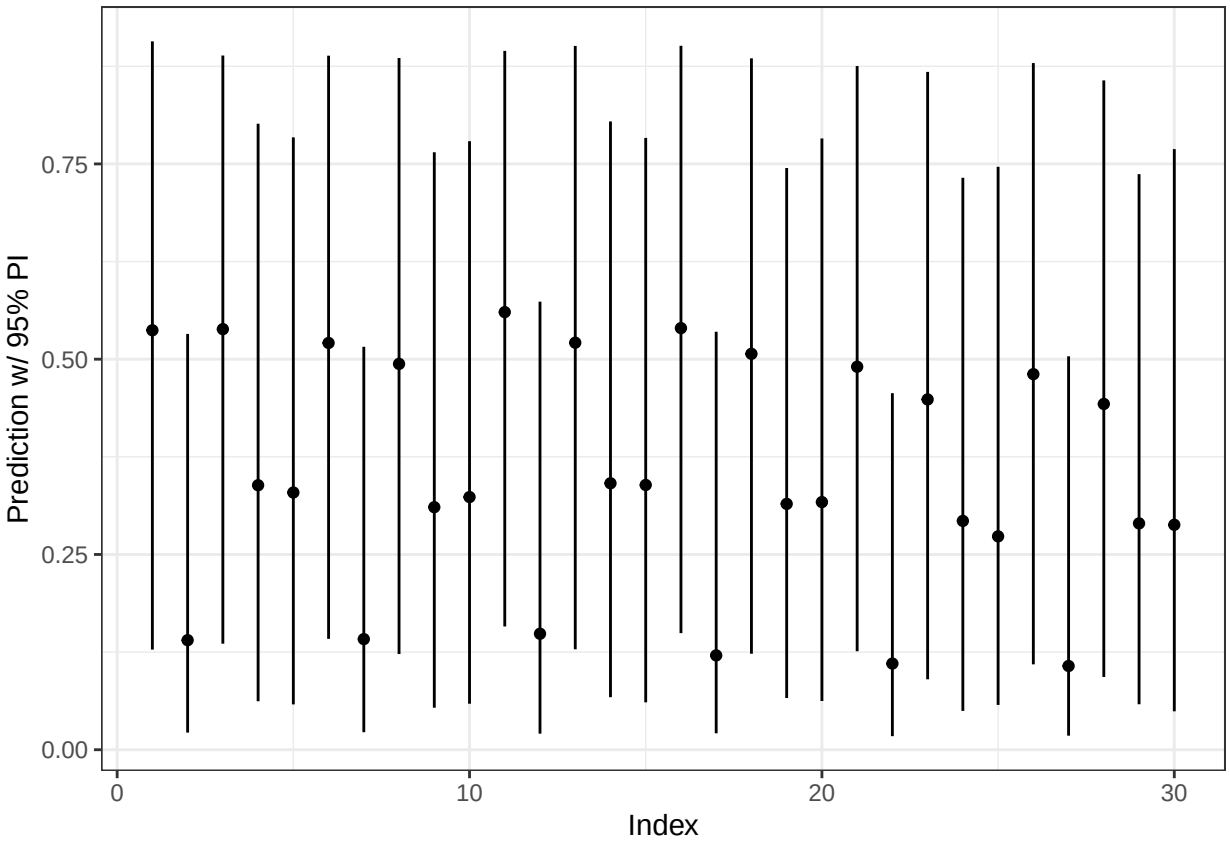

Figure 1: 95% Prediction Intervals for forcasting the Chances that Donald Trump wins the election

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

5.4.1 1: Nested Bayesian Model

5.4.2 2: Pre-analysis

Appendix

6 References