Prediction: Joe Biden Wins the 2020 Election with Popular Vote and Minor Lead in Electoral College Votes

Chien-Che Hung

11/2/2020

Abstract

In this report, we will be predicting the winner for the 2020 United States Election. We utilized two datasets, Democracy Fund + UCLA Nationscape Data and American Community Survey (ACS) Data to perform Multilevel Modeling with Post-Stratification (MRP). This analysis is useful and important because we included all the basic yet crucial information that can easily be obtain from the public. With our model, broader predictions can be made. The model that we are using in this analysis is Logistic Regression, since the response variable that we have is binary, Joe Biden or Donald Trump. After our prediction on the post-stratification data (ACS) data, we can conclude that Joe Biden wins 53% the popular vote with 0.17 variance and 52% the electoral vote. In addition to the voting prediction, we also find out that Race, Income Level, and State Residency contribute primarily to a person's voting choice.

Keywords: Forecasting; US 2020 Election; Trump; Biden; Multilevel Regression

Introduction

As the 2020 United States Election is approaching, we would like to know who is more likely to be the next president of the United States, Donald Trump or Joe Biden. Even though whom to vote for in the election is a very subjective question, we can use some facts, such as demographic questions, to predict each individual's response. Each individual's eventual choice could depend on this background information to some extend.

This analysis process is fortunate enough to obtain two credible datasets, Democracy Fund + UCLA Nationscape Data (Tausanovitch, Chris and Lynn Vavreck. (2020)) and American Community Survey (ACS) Data (U.S. Census Bureau. (2012)). This allows us to perform logistic regression with Multilevel Modeling with Post-Stratification, which requires two datasets for this statistical method. The details for these methods will be elaborated in the following sections in the report. Browsing through numerous standard variables in both datasets, we decided to select the ones that would influence a person's choice of voting: State Residency, Gender, Age, Education, Income Level, and Employment Status. These variables would be positively correlated to their choice of candidate. Other than those six variables, we are going to use the answer to the question "If the election were going to held now and the Democratic nominee was Joe Biden, and the Republican nominee was Donald Trump, would you vote for. . . .?" in the Nationscape Data to construct the model for the prediction on the ACS Data.

Several interesting conclusions can be drawn from our analysis. Firstly, the effects of different variables on probability for a person to vote. For example, do races contribute positively or negatively to the probability of voting for Joe Biden? After obtaining the prediction based on the Nationscape data, we will determine the popular vote for this election. Secondly, as the race factor predominantly affects the election, it could suggest the candidates' policy direction. For instance, construct policies towards certain income levels of people. We will also consider which candidate wins in the Electoral College.

In the Data section of the report, we will talk about the data thoroughly on how the data was obtained, the surveys were asked, and the reasoning behind the variables that we choose. After the Data section, we will discuss the modeling method that we use and why we choose this specific model. The Result section and the

Discussion section come after. We talk about the results of our statistical analysis and what conclusions can be made through the results.

Data

Reasons for Two Datasets

There are two different types of datasets being used in this report and analysis. In this analysis, we utilize the statistical technique called Multi-level Modeling with Post-Stratification (MRP). We can think of this modeling method instead of making assumptions about how the observed sample was produced from the population. We make assumptions about how the observed sample can be used to reconstruct the population. Also, it uses an individual model (Nationscape Data) to adjust the population (ACS Data) estimates, which can be understood as a weighted average from all possible combinations of the attributes as we will talk about the variables being used in the following sub-sections. Our total combinations (cells) of $51 * 2 * 9 * 7 * 10 * 3 * 24 = 4626720$; these cells will be predicted after the model for Nationscape Data is built. After introducing the brief idea for MRP, the following are the steps to construct the MRP model (Lauren Kennedya and Andrew Gelman (2020)):

1. Collect demographic features during the survey collection stage and identify the post-stratification data.
2. Pick out the shared variables and the variable that we are going to train on.
3. Estimate the parameters in the model.
4. Estimate the post-stratification values from the estimated model.

Despite the convenience of the MRP model, there are still some weaknesses. Some of the weaknessess and limitations also occur in our analysis. Firstly, our model would fail to have good prediction due to insufficient amount of data, or the individual level data (sample data) creates a certain bias towards the population. Secondly, since individual level data and post-stratification level data need to have identitical variables and responses categories, the generalization of data from one to another often loses some detailed information during the process.

This report uses the datasets from Democracy Fund + UCLA Nationscape as our sampled dataset to reconstruct the American Community Survey Datasets. In this section, we will thoroughly discuss the purpose of each data and details about the data.

Democracy + UCLA Nationscape Data

The first data that we are going to use is from Democracy Fund + UCLA Nationscape (Tausanovitch, Chris and Lynn Vavreck. (2020)). It is a partnership between the Democracy Fund Voter Study Group and the University of California Los Angeles Political Scientist. Lucid provides the samples. Lucid is a market research platform that runs an online exchange for survey respondents. The population from the data covers from every county, congressional district to mid-sized U.S. cities. Nationscape aims to understand people's opinions on the 2020 election. It conducts weekly surveys and an estimated 500,000 interviews of Americans from July 2019 through December 2020. While the population for this survey is every American eligible to vote, the sampling frame would be the "suppliers" on the Lucid Marketplace Platform. Normally, if the interviews of this sruvey conducted through random sampling through phone numbers, the cost of could be unimaginable. However, since Nationscape conduct the survey and obtains the survey from Lucid, the cost would be relatively low.

In the study, the sample would be those who have access to the Lucid Marketplace Platform on a networked computer or mobile device as the mode of the interview is the online survey. The sampling method on this platform is called Programmatic Sampling, which automates buying and selling the sample (Patrick Comer (2019)). Using this method, Nationscape could get an enormous amount of data with less cost and human resources. The platform user accept the selection from the party that creates the survey and answer the

survey survey on the platform. There are roughly 12% of the selected people or groups declined to do the survey, and 5% of the people stopped doing the survey half-way. These respondents are dropped and not included in the data. However, even though the survey asks the respondents about their household income levels, they have the option not to answer the question. To deal with this question's non-responses, targets for response categories are based on American Community Survey responses multiplied by the proportion chosen to answer the income question.

As we look deeper into the survey, we can see that it covers a wide range of questions, including questions about respondents' attitudes, behaviors, and facts about their lives. These detailed questions could be a strength during the step of the analysis. However, the advantages could also be weaknesses. For instance, "How much do you trust the people in your neighborhood?" The answer to this question could depend on how the respondent defines their neighborhoods. Alternatively, some questions might make the respondents unwilling to tell the truth about unwillingness or embarrassment, such as "Did anyone in your household get food stamps or use a food stamp benefit card at any time during 2018?". Even if people speak the truth regarding these topics, the Lucid Marketplace Platform sampling method might not get a sample that represents the population.

As mentioned above, the sets of data are collected weekly. Here, we are going to pick the results gathered during the week of June 25th, 2020. As mentioned in the Introduction section, this research aims to predict the winner of the upcoming 2020 United States Election. There are several fundamental elements from the respondents that would affect the election outcome, for example, the necessary information such as gender, age, race, and individual information such as education level and income level. There are two reasons why we decide to use these variables. Firstly, it is general enough to cover the respondents' situations, yet not too detail to make the analysis too complicated. Secondly, since we perform multilevel modeling with post-stratification, the variables from two datasets have to match.

he following table (Table 1) is how the data looks like. It contains the variables that we are using and the variables that are mentioned in the report.

Table 1: First five observations of the Nationscape Data

vote_2020	vote_2020_employment	gender	race_ethnicity	household_income	education	state	age
Donald Trump	Not Asked	employed	Female	white	$75,000 to $79,999	Associate Degree	WI	40-49
Donald Trump	Donald Trump	employed	Female	white	$100,000 to $124,999	College Degree (such as B.A., B.S.)	VA	30-39
Donald Trump	Not Asked	employed	Female	white	$175,000 to $199,999	College Degree (such as B.A., B.S.)	VA	40-49
Donald Trump	Not Asked	unemployed	Female	white	$65,000 to $69,999	High school graduate	TX	70-79
Donald Trump	Not Asked	not in labor force	Female	white	Less than $14,999	High school graduate	WA	50-59
Joe Biden	Joe Biden	unemployed	Female	white	Less than $14,999	Other post high school vocational training	OH	40-49

The details of the data will be introduced after we talk about the American Community Survey (Post-Stratification Data).

American Community Survey

The post-stratification dataset that we are going to use is from the American Community Survey (ACS) Operations Plan (U.S. Census Bureau. (2012)). ACS aims to provide information to federal, state, and local governments. The ACS samples include about 3 million households nationwide and one percent of group

quarters population (places such as nursing homes, prisons, college dormitories, military barracks, juvenile institutions, and emergency and transitional shelters for people experience homelessness). However, due to hardware restrictions, we cannot predict 3 million households at once based on our Nationscape model. We cut down the sample size to 1550789 households. To accomplish this, smaller random subsets (around 50% of the original dataset in this case) of the entire data is being selected.

The sampling method for ACS is called systematic sampling. This method is usually done through the paper. Unlike Simple Random Sample Without Replacement (SRSWOR), systematic sampling considers the respondents' physical settings, where the respondents are arranged in a sequence. After choosing a random starting point of sampling, the samples will be collected once after a specific interval. The interval is decided by $\frac{Population}{Sample}$. For ACS, to reach the respondents, it utilized the addresses from Master Address File (MAF). After obtaining the addresses, the samples would receive an ACS survey at the beginning of the month. To deal with the non-respondents, Computer Assisted Telephone Interview would be conducted one month later. If there are still no responses from the selected respondent, a computer-assisted personal interview (CAPI) would be conducted for one-third of the non-respondents to either physical mail or telephone. While the American Community Survey population is all the Americans, since this survey type is like a census survey, the sampling frame would be the addresses in the MAF. Other than those strategies mentioned earlier to deal with the non-response, ACS also over-samples the low mail response areas and small population groups to reduce the variation. Since there is a correlation between low mail response and minority populations, oversampling for low mail responses may address providing reliable estimates for small population groups.

One of the ACS survey's critical features is that ACS Operation Plan detailed assesses the needs for each question and work with other agencies on their data needs. Thus, the ACS survey provides a wide range of questions ranging from household information, family relationship, migration status to occupation status. As a detail, as the data is, there are still some flaws. For example, for the race, it provides options such as "Two major races" and "Three or more major races." If the analysts want to analyze the race in detail, it would be hard for them to do the work.

Table 2: First five observations of the ACS Data

employment	gender	race_ethnicity	household_income	education	state	age
not in labor force	Female	black/african american/negro	Less than $14,999	Completed some college, but no degree	AL	10-19
not in labor force	Male	white	Less than $14,999	Other post high school vocational training	AL	50-59
not in labor force	Female	white	Less than $14,999	Completed some high school	AL	20-29
not in labor force	Female	white	Less than $14,999	High school graduate	AL	30-39
not in labor force	Male	white	Less than $14,999	Completed some high school	AL	30-39
not in labor force	Female	white	Less than $14,999	Completed some college, but no degree	AL	10-19

In Table 2, we can see that to match with the variables in the Individual Level Nationscape data, we also pick the ones that are being used for the Nationscape. Thus, state, gender, race, birthplace, education, employment status, and household income will be picked for post-stratification prediction. The modification applied to the data will be discussed in the next subsection.

Modifications and Visualization on Nationscape Data and ACS Data

This section will talk about the modifications being made on both data and the comparison of both datasets. The main reason for the modification is that to make the prediction, the variables and the datasets' categories

need to be identical.

- **Race and Ethnicity (Nationscape; Figure 1)**: In the Nationscape Dataset, race and ethnicity are separated into White, American Indian or Alaska Native African American, Asian with seven categories, Pacific Islander with four categories, and some other race, which has a total of 15 categories. Initially, there are categories such as "three or more major races" and "two major races" in the ACS data. However, since we do not know the races' specifics in those options, we drop the observations that contain two responses. This leaves race and ethnicity in the ACS data with seven categories. Because ACS's race and ethnicity variable only have seven categories, we re-distributed the 15 categories that Nationscape data has into these seven categories. The graph below shows the proportion of each race with different data types. We can see that both datasets follow a similar trend of the distribution for race and ethnicity.

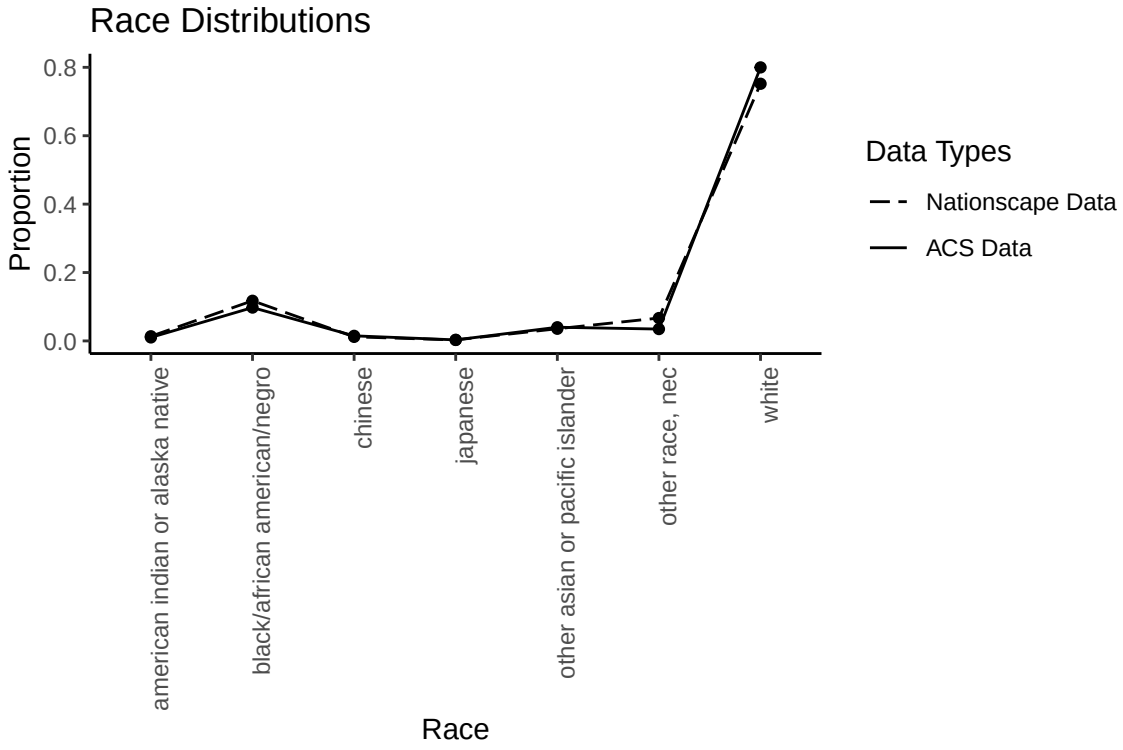

Figure 1: Distribution of Race and Ethnicity in both datasets

- **Voting in 2020 (Nationscape; Figure 9 lower level)**: Next, we will discuss the essential modification on the Nationscape data. The Nationscape data also talks about whom the respondents will vote for, Donald Trump or Joe Biden. Those who cannot decide whom to vote for during the time taking the survey, which shows up as "I am not sure/do not know" and "I would not vote," will be asked which candidate they are leaning towards. Thus, as we combined these two columns into one, we can know more about the support rate and find out some hidden voters who do not voice their opinions. The distribution of this variable will be discussed after the prediction from the post-stratification data is being made.

- **Age (Nationscape and ACS Data; Figure 2)**: For the respondents' age in both datasets, they are presented by integers rather than categories. Thus, we put the respondents' age in both datasets into their corresponding age groups ranging from 10-99 with an interval of ten years. The graph below shows the proportions of different age groups for different datasets. We can see that post-stratification data (ACS data) has a lower relatively lower proportion in the age range "20-29", "30-39", and "40-49." The difference in proportion could affect post-stratification prediction since there tend to have a partisan and ideological gap between younger and older generations.

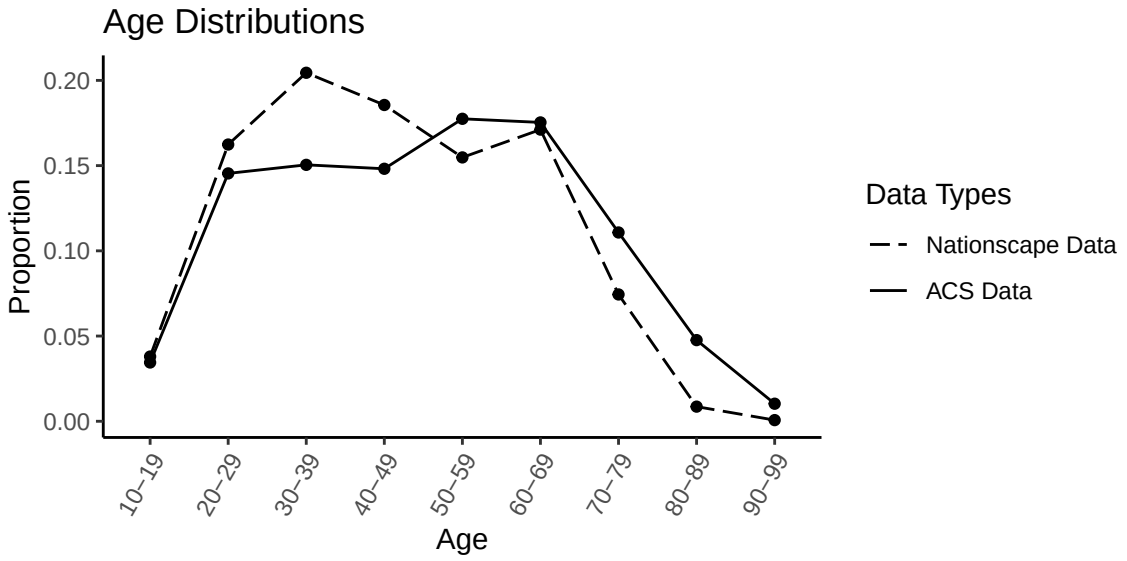

Figure 2: Distribution of age groups

- **Household Income (ACS Data; Figure 3)**: In the household income section, the Nationscape data's income is in categories. Because the ACS data's income is in integer form, we put the income according to the levels in the Nationscape data. The below graph shows the proportion of each household income for different Data Types.

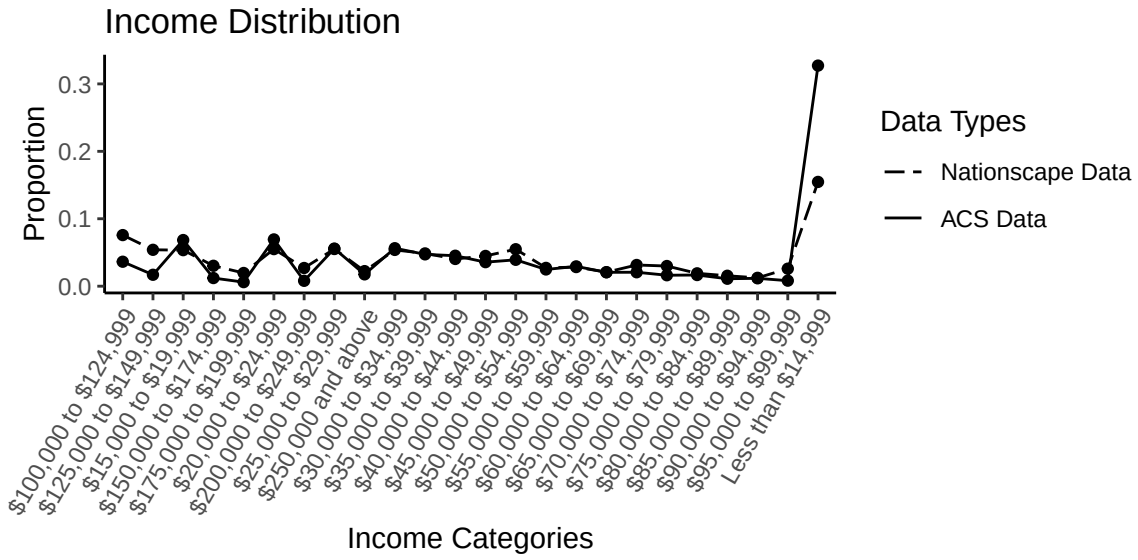

Figure 3: Distribution of Household Income for both datasets

- **Education (ACS Data; Figure 4)**: In the Nationscape data, it has a broader definition of categories. For example, there are separate categories for ACS Data for grades, from grade 1 to grade 11. On the other hand, Nationscape generalizes it into "3rd Grade or Less" and "Middle School - Grades 4-8." Since there is no way for us to know the specifics inside "3rd Grade or Less" and "Middle School - Grade 4-8", we can only make the ACS education categories into a more general form. The graph below shows the proportion of each education types for both datasets. We can also see that both distributions have similar distributions. Noticeably, the categories provided in the ACS data cannot be categorized in the "completed some graduate, but no degree." However, we decide to keep this category as this might also influence the voting results.

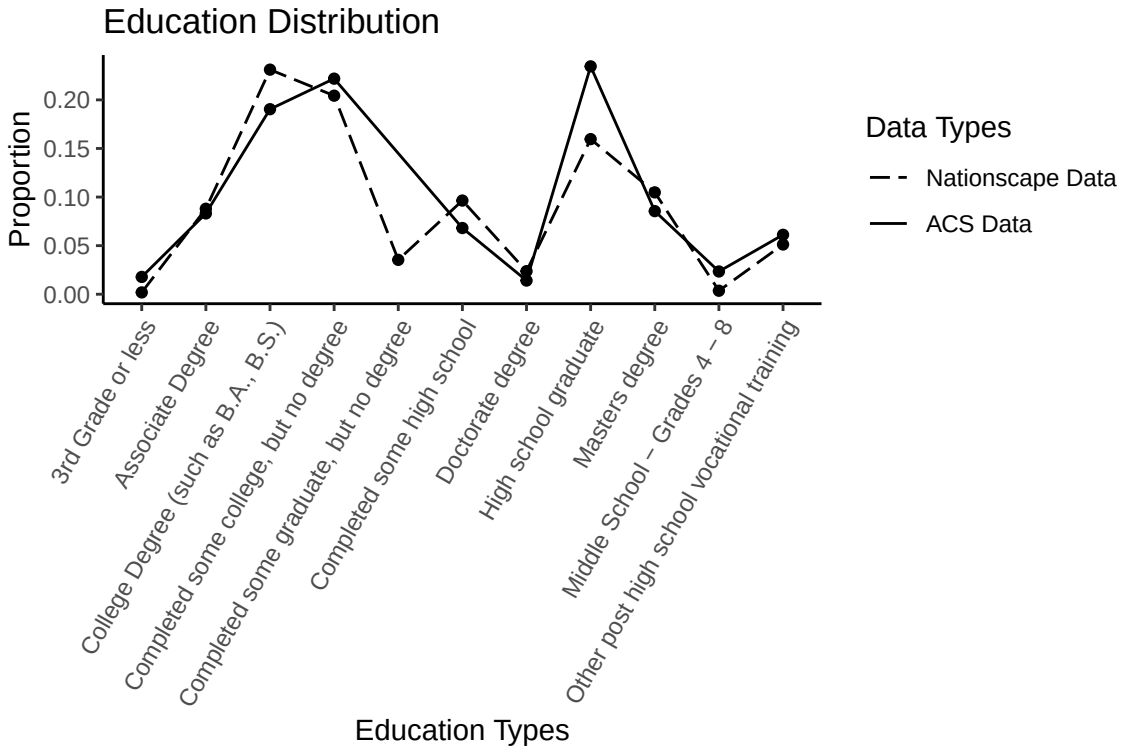

Figure 4: Distribution for education level for both datasets

- **State (ACS Data; Figure 5)**: In the ACS Data, the states' representation is in their full name lower case form. Thus, we need to convert it to the states' abbreviations, which would fit the representation in the Nationscape Data and easy to represent. Lastly, the below graph is the distribution of states in two datasets. We can see that they are identical.

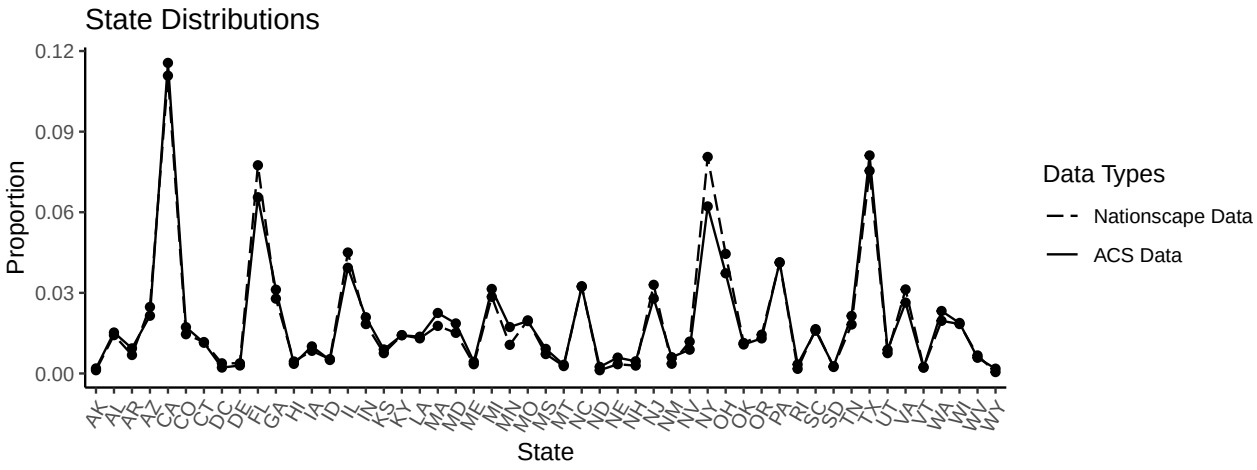

Figure 5: Distribution of each state for both datasets

Model

From the previous sections, we mentioned that we would use multi-level modeling with post-stratification to predict who will win the 2020 United States Election. Before the post-stratification stage, we are going to model the individual level (Nationscape Data). Since the response variable result would be binary, Donald Trump or Joe Biden, we will use logistic regression. We are interested in how state, gender, age, race and ethnicity, education, employment status, and household income affect people's decision on voting for candidates. Since different candidates would have different policies for different gender, ages, races, and social statuses, these are the critical factors that would affect how people cast their votes. Most importantly, the respondents' states cannot be neglected because some states support Republicans (Donald Trump) and some states that support Democratic (Joe Biden).

The formula below is the general formula for logistic regression. The p represents the mean probability for the event that we are interested in. β_0 to β_n are the estimates for the coefficients of the regression, where n represents number of the variables that we are going to use.

$$log(\frac{p}{1-p}) = \beta_0 + \beta^T X = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

In addition to logistic regression, we are going to utilize Bayesian Inference. Bayesian Regression Model is based on the application Bayes Theorem:

$$P(\theta|X,\alpha) = \frac{P(X|\theta,\alpha)P(\theta,\alpha)}{P(X|\alpha)P(\alpha)} \propto p(X|\theta,\alpha)$$

From the equation above, $P(\theta|X,\alpha)$ is the posterior probability. In order worlds the probability of θ given X,α. $P(X|\theta,\alpha)$ is the probability of observing X given θ,α, which is also known as the likelihood. Most importantly, $P(\theta,\alpha)$ is the prior probability, which means that we can use the estimate of the probability of θ,α that we obtain outside the research and use it in the current research.

The following formula is the model that we are using (Model 1):

$$
\begin{aligned}
log(\frac{p}{1-p}) =& \beta_0 + \beta_{state} x_{state} \\
&+ \beta_{gender} x_{gender} + \beta_{age} x_{age} + \beta_{race} x_{race} \\
&+ \beta_{edu} x_{edu} + \beta_{employ} x_{employ} + \beta_{income} x_{income} \\
\beta_{gender} &\sim N(0, 10) \\
\beta_{age} &\sim N(0, 10) \\
\beta_{race} &\sim N(0, 10) \\
\beta_{edu} &\sim N(0, 10) \\
\beta_{edu} &\sim N(0, 10) \\
\beta_{employ} &\sim N(0, 10) \\
\beta_0 &\sim student - T(3, 0, 2.5)
\end{aligned}
$$

In the equation, p is the probability of the respondents vote for the candidates. In this case, if p is higher than 0.5, then the respondents vote for Joe Biden. On the other hand, if p is lower than 0.5, then the respondents vote for Donald Trump. For the parameters, β_{state} is the parameter for the state variable, $\beta_{gender}$is the parameter for respondents' gender, β_{age} is the parameter for respondents' age, β_{edu} is the respondents education level, β_{employ} is the parameter for the respondents' employment status, and β_{income} is the parameter for the respondents' income level. For the priors, we pick weakly informed priors for the parameters that we just mentioned. As the coefficients in logistic regression could be in a broad range, we make the priors for the coefficients distributed with the variance of 10 and 0 mean. By picking weakly informed priors, we can prevent our data from being too sensitive to our prior. For the intercept β_0, we use the default prior, which is student t-distribution with degrees of freedom = 3, mean = 0, and standard deviation of 2.5. After deciding the model, we will run and interpret this model in R Programming Language (R Core Team (2020)) with library Bayesian Regression Models using 'Stan' (brms) (Paul-Christian Bürkner (2017)).

However, before we jump into interpreting the data, we have to check for the model convergence. Model convergence is an essential factor to consider since the parameters would not be trustworthy and should not be interpreted. The assessing method that we are going to use is Posterior Predictive Checks. How this checking method is going to work is that after simulated random draws from posterior predictive distribution, if the model has converged, we would expect that the lines of y_{rep} will be roughly similar to the observed y data. From the graph below (Figure 6), we can see the lines appear in a roughly similar pattern. Thus we can say that this model converges, and we can move on to interpret the parameters. For Generalized Linear Model, there is an assumption on constant error variance. However, since the model we are using here is logistic regression and the error variance is not a parameter in the Bernoulli Distribution, we will not consider this assumption.

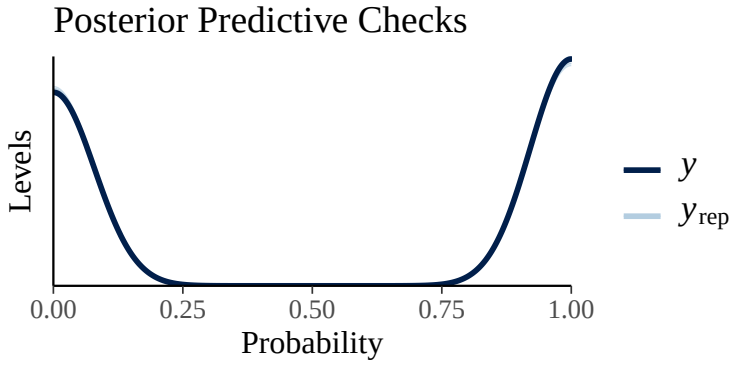

Figure 6: Posterior Predictive Checks for convergence

Originally, we were going to use regularization/partial pooling regression model. In brief, this regression

model allows us to assess the underlying effects contributed by the variables that we are interested in. The model is regarded as follows (Model 2):

$$
\begin{aligned}
log(\frac{p}{1-p}) =& \alpha_{state[i]} + \beta_0 \\
&+ \beta_{gender} x_{gender} + \beta_{age} x_{age} + \beta_{race} x_{race} \\
&+ \beta_{edu} x_{edu} + \beta_{employ} x_{employ} + \beta_{income} x_{income} \\
\beta_{gender} &\sim N(0, 10) \\
\beta_{age} &\sim N(0, 10) \\
\beta_{race} &\sim N(0, 10) \\
\beta_{edu} &\sim N(0, 10) \\
\beta_{edu} &\sim N(0, 10) \\
\beta_{employ} &\sim N(0, 10) \\
\beta_0 &\sim student - T(3, 0, 2.5) \\
\alpha_{state[i]} &\sim student - T(3, 0, 2.5)
\end{aligned}
$$

Other than the $\alpha_{state[i]}$is different from Model 1, the rest is the same. Here the use $\alpha_{state[i]}$ means that there might be underlying state effects in the dataset and i indicates the order of different states.

However, during the model diagnostic stage, it has a lower correctly-predicted rate (Model 1: **64.7%** vs Model 2: **63.1%**). The method for doing this is to predict the original data that we fit the model in and see the right classificationrate (True Positive Rate). Because of the lower classification rate, we discard Model 2. Also, we can reasonably conclude that States Factor is more than an underlying effect. Being in different states would have effects on whom the respondents pick to vote.

Other than the classification rate, we also assess the Area Under the Curve. It is a measure of the ability of a classifier to distinguish between classes. The way to interpret the result is when the AUC value is between 0.5 and 1, there is a high chance that the classification will predict the value correctly. In our model, the AUC is **0.7078**, whish is in the range of 0.5 and 1. Thus we can continue and use this model.

Post-Stratification

After fitting the data to the Multi-Level Logistic Regression, we can start applying the post-stratification ACS data into this model. The coefficients for the model represents the same parameters as the logistic regression model mentioned above. However, they would be the estimated parameters since we create the model based on the observed data. Different from the Nationscape Data, $X = (x_{state}, ...x_{income})$ are the observations from the ACS data.

$$
\begin{aligned}
log(\frac{\hat{p}}{1-\hat{p}}) =& \hat{\beta}_0 + \hat{\beta}_{state} x_{state} \\
&+ \hat{\beta}_{gender} x_{gender} + \hat{\beta}_{age} x_{age} + \hat{\beta}_{race} x_{race} \\
&+ \hat{\beta}_{edu} x_{edu} + \hat{\beta}_{employ} x_{employ} + \hat{\beta}_{income} x_{income}
\end{aligned}
$$

After fitting the model, we will have the probability ($\hat{p}$) of the willingness to vote for either candidate.

Results

Model Results

In this section, we will present the results of the model. The parameters has a hat on them because they are the estimated parameters from our logistic regression.

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) = \hat{\beta}_0 + \hat{\beta}_{state}x_{state} \\ + \hat{\beta}_{gender}x_{gender} + \hat{\beta}_{age}x_{age} + \hat{\beta}_{race}x_{race} \\ + \hat{\beta}_{edu}x_{edu} + \hat{\beta}_{employ}x_{employ} + \hat{\beta}_{income}x_{income} \end{aligned}$$

Table 3: Head and Tails of the coefficients of the estimates

	Estimate	Est.Error	Q2.5	Q97.5
stateVA	0.724	0.874	-0.970	2.460
stateVT	2.505	1.243	0.292	5.092
stateWA	0.859	0.880	-0.805	2.571
stateWI	0.934	0.880	-0.760	2.700
stateWV	0.378	0.924	-1.406	2.216
stateWY	0.067	1.738	-3.539	3.284
Intercept	-0.322	1.138	-2.580	1.850
genderMale	-0.410	0.059	-0.525	-0.295
age20M29	-0.473	0.193	-0.856	-0.095
age30M39	-0.908	0.189	-1.276	-0.551
age40M49	-1.090	0.191	-1.467	-0.725
age50M59	-1.040	0.196	-1.426	-0.653

However, due to the reason that there are 99 coefficients in total (Intercept + Gender + Age + Race + Household Income + Education + State = 99), we will have a quick glance of the coefficients here in Table 3 and the full version would be in the Appendix Table 6. To understand the coefficients, we will have to understand the concept of log odds. The log odds here is defined as $log(\frac{P_{Biden}}{P_{Trump}})$, which is the log of probability of the respondents of the Nationscape data that would vote for Joe Biden over the probability of the respondents of the Nationscape data that would vote for Donald Trump. As we take the exponential of the coefficients, it will become $\frac{P_{Biden}}{P_{Trump}}$. If the exponentiated coefficients are smaller than one, that means denominator would have higher probability. On the other hand, if the exponentiated coefficients are bigger than one, it means that the numerator probability would be higher.

As we mentioned in the Model Section, we also tested on the accuracy of the model. The confusion matrix (Table 4) below represents the accuracy and wrongly categorized predictions.

Table 4: Confusion Matrix for the model

	Donald Trump	Joe Biden
Donald Trump	1733	1102
Joe Biden	947	2038

Post-Stratification Results

After fitting the ACS data with the model, we will use the probability of voting for either Donald Trump or Joe Biden to decide whom the respondent would have voted. If the probability of the result is larger than 0.5, then we will categorize that the respondent would vote for Joe Biden and if it is less than 0.5, we will categorize that the respondent would vote for Donal Trump. After categorizing them, we can see the total voting counts for Donald Trump and Joe Biden (Table 5).

Table 5: Vote Counts

Candidates	Vote Counts
Donald Trump	720835
Joe Biden	829954

In addition, we also plot out the voting counts in each states (Figure 7) and the heat map (Figure 8) for each state to have a general and quick glance of the competitiveness in each state.

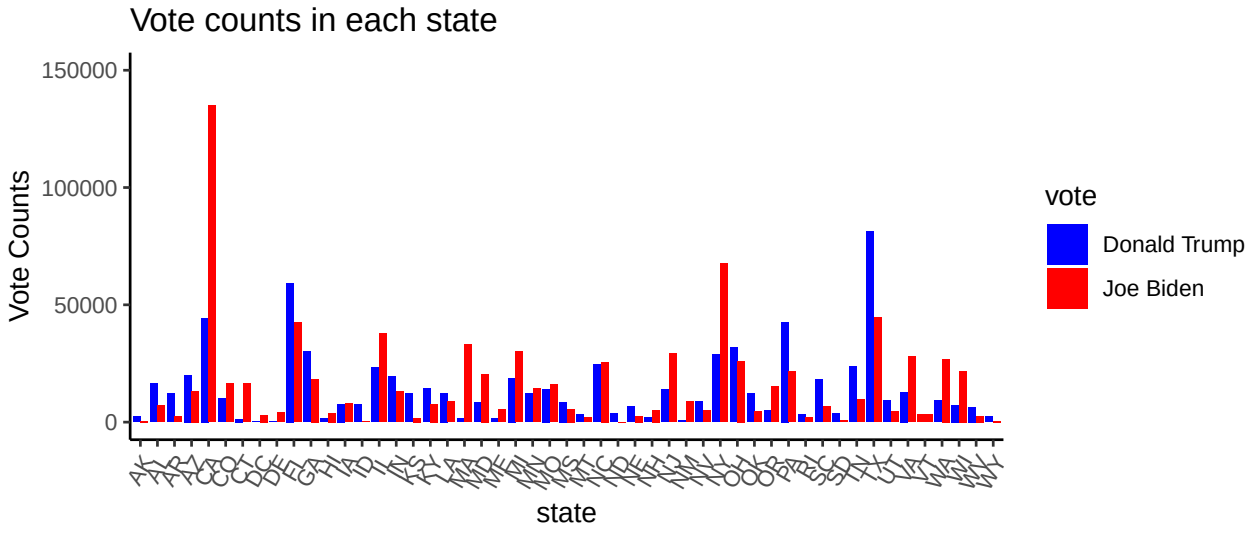

Figure 7: Vote counts in each state

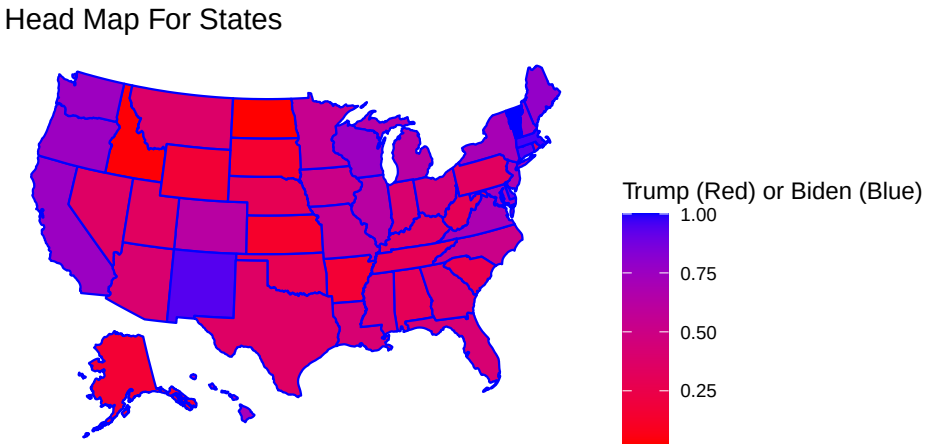

Figure 8: Head Map for different states

Model vs. Post-Stratification

Since we are using Nationscape data to create the model and apply the model on ACS data as post-stratification for our prediction. We would want to take a look of the difference between the two datasets in the states (Figure 9) and the race level (Figure 10).

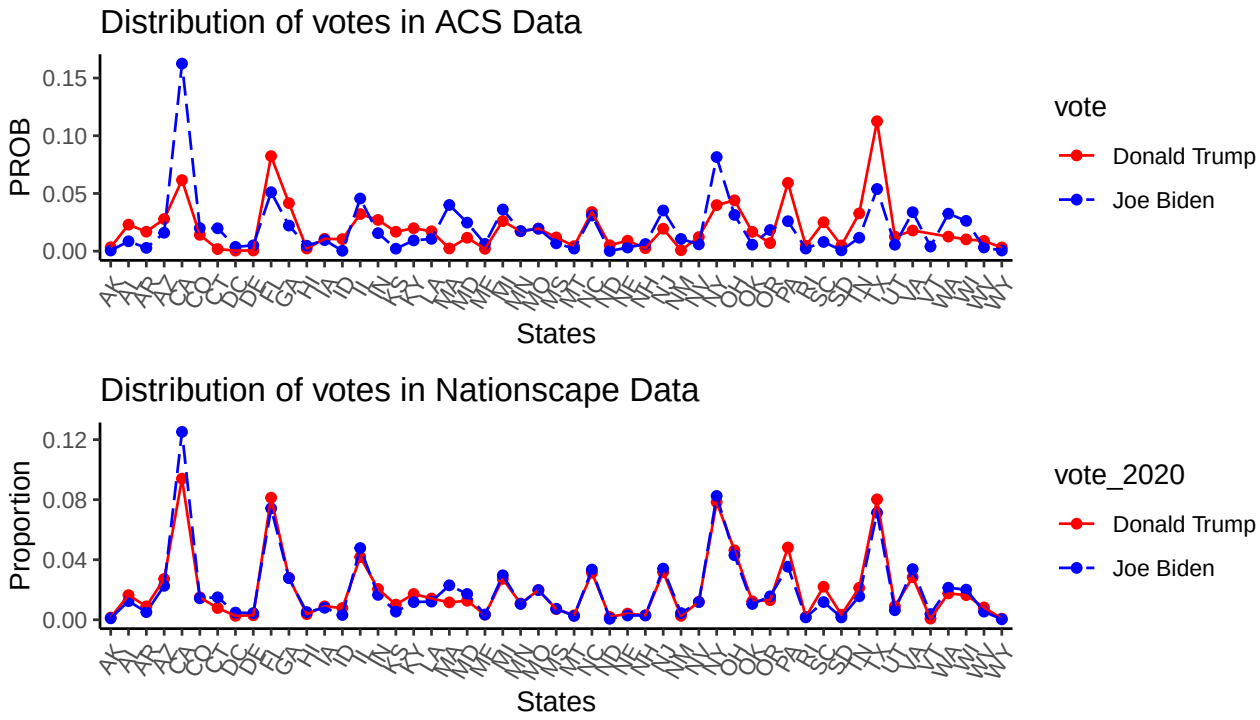

Figure 9: Distribution of votes in different states for two datasets

All the details of each results will be discussed later on in the Discussion Section.

Discussion

First of all, we want to assess the precision of our model from the confusion matrix (Table 4). The accuracy of the data is around 64.7%. Even though we cannot say that this is an excellent performance model, the accuracy rate is higher than chance (50%). Along with the accuracy rate and the model convergence test, we can say that this model is adequate but not optimal.

Second of all, from the table of exponentiated coefficients (Appendix Table 5), we will discuss the variables separately to have some perspectives on people's opinions in each group. If we solely look at the age groups, people from the ACS survey data, regardless of the age group, would more likely to support Donald Trump than Joe Biden. However, as we add in the factors of race and ethnicity, most races would lean towards Joe Biden other than White Americans. Expressly, if we take a look at the coefficients of African Americans, people turn towards Joe Biden more intensively. This result corresponds to some racial sensitive comments Donald Trump makes publicly. In the income section, while most of the income levels would support Joe Biden, the relatively high-income level respondents such as people who have income level "175000USD to 199999USD" and "250000USD and above" would lean towards Donald Trump. In the education section, regardless of the education degrees, people in each education degree would support Joe Biden. Most of the states lean toward supporting Joe Biden other than Arizona, Idaho, Kansas, Mississippi, North Dakota, and South Carolina. Interestingly, for Vermont (VT), it has the highest effects (more weight) with 12.24 to the

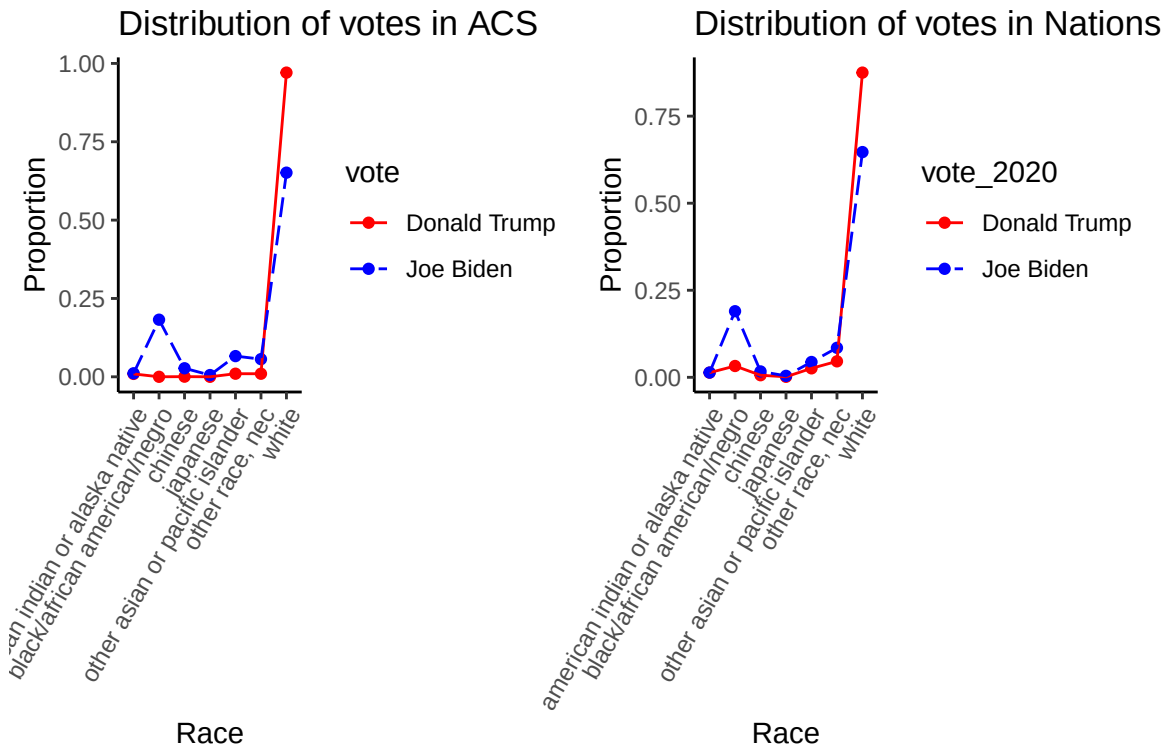

Figure 10: Distribution of votes in races for two datasets

probability of voting for Joe Biden. This has corresponded to the fact that Vermont has been Democratic Party's most loyal state.

After predicting the votes by running the model on the ACS data, we can get the total popular vote of the election (Table 5). In terms of population vote, Joe Biden wins the election by more than 100000 votes. However, after we take the average of the probability of voting for each candidate, the average is 0.53, with a 95% confidence interval (0.247, 0.906). It means that we are not confident that Biden will win the popular vote since the probability include 0.5 (if the probability is lower than 0.5, it means that Trump wins the popular vote). It is well known that each state has its stance on the election parties. Thus, Figure 7 talks about the distribution based on states. We can see that Biden wins most of the votes in California, and Trump wins a lot more votes than Biden in Texas. The heat maps (Figure 8)shows the political positions in each state. Other than some dark red states (Republican)and dark blue states (Democratic), which means that we are confident that each candidate will get the electoral votes corresponding to their parties, some ambiguous states/swing states.

Given that in the 2016 election, Hilary Clinton won the popular vote but lost the election because of the Electoral College system in the United States. Each state has its electoral votes according to their population obtained from the 2010 census. Here, we assume that Maine and Nebraska do not split their electoral votes. **Biden wins the election by getting 280 electoral votes, while Trump gets 258 electoral votes.**

Nationscape Voting vs. ACS Voting Prediction

Firstly we will take a look at the proportion of votes from different states for each candidate (Figure 10). We can see that in the Nationscape Data, the distribution of proportion is almost identical for Donald Trump and Joe Biden. When we look at ACS data's predict on the voting situation, few states have some discrepancy, where California takes up a considerable proportion of the votes that Biden gets. This means that our post-stratification successfully adjusts the imbalanced in the data. However, this also means that the

Nationscape data probability did not capture the proportion correctly that each candidate will get in each state. Or in McElreath's word, the information that our Bayesian model learned is not doing an excellent job of approximating the large world ((Mathias Harrer, Doing Meta-Analysis in R)).

For the distribution of votes in different race in Figure 10, both types of data get similar proportion of votes in different race and ethnicities.

Weakness and Future Works

Several weaknesses and future works can be done in this analysis, which revolves around the insufficient amount of data in this analysis.

First of all, the size of the Post-Stratification Data is significantly decreased, and the potential information is lost. Due to the hardware constraint, we have to lower the observation size before actually working on the prediction. The 8GB RAM will get overload while conducting the prediction. The final observation number for our prediction is 1550789, which is obtained from random subsets of 3 million observations.

Second of all, we introduced the modifications made in the dataset because we are using multi-level modeling with post-stratification. For both datasets, most of the modifications are done by generalizing the categories. This generalization process could cause some precious information to be lost in the process. For example, in the Nationscape Data, there are different categories of races, especially for the Asian and Pacific Islands. Different races could pose different elections in the election. However, due to the survey's different purposes and nature, those elements need to be taken out.

For the survey, even though both organizations have done a tremendous job, there are still some improvements to be made if the budget is not concerning. Be as detail as possible for each question being asked. For example, just like the Race problem we just mentioned, when the data is more detailed, analysis of any fields would be more accurate and efficient.

As the data section stated, there are weekly surveys for the Nationscape Data. In this analysis, we only utilize one week of the numerous survey. For future work, we could combine a certain number of surveys to get a more accurate classification from the model and the full ACS datasets. Regarding the model, since the weakly and the default is being used and could be not informative to some extend, some future work can be done on researching the useful priors in our models.

References

Data Cleaning Code Adapted From

- Rohan Alexander and Samantha Caetano (2020) 01-data_cleaning-survey.R
- Rohan Alexander and Samantha Caetano (2020) 01-data_cleaning-survey-post-strat.R

Survey/Data Source

- Tausanovitch, Chris and Lynn Vavreck. (2020). Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814) . Retrieved from [URL] .
- U.S. Census Bureau. (2012). 2009-2011 American Community Survey 3-year Public Use Microdata Samples [STATA Data file]. Retrieved from https://usa.ipums.org/usa/index.shtml
- 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

Library References

- Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
- Jona Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, Andrew Gelman (2019). "Visualization in Bayesian workflow." *J. R. Stat. Soc. A*, *182*, 389-402. doi: 10.1111/rssa.12378 (URL: https://doi.org/10.1111/rssa.12378).
- Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr
- Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.
- Hadley Wickham (2016). ggplot2: Elegant Graphics for Data Analysi. Springer-Verlag New York
- JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL https://rmarkdown.rstudio.com.
- Matthew Kay(2020). *tidybayes: Tidy Data and Geoms for Bayesian Models*. doi: 10.5281/zenodo.1308151 (URL: https://doi.org/10.5281/zenodo.1308151), R package version 2.1.1, <URL: http://mjskay.github.io/tidybayes/>.
- Lauren Kennedya and Andrew Gelman (2020). Know your population and know your model: Using model-based regression and post-stratification to generalize findings beyond the observed sample. arXiv: https://arxiv.org/pdf/1906.11323.pdf
- Mathias Harrer, M., Cuijpers, P., Furukawa, P., & Ebert, A. (n.d.). Doing Meta-Analysis in R. Retrieved November 02, 2020, from https://bookdown.org/MathiasHarrer/Doing_Meta_Analysisin_R/bayesian-meta-analysis-in-r-using-the-brms-packag.html
- Patrick Comer (2019). Sampling in the Digital Age. Retrieved November 03, 2020, from https://luc.id/blog/sampling-in-the-digital-age/
- Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. doi:10.18637/jss.v080.i01
- Paul-Christian Bürkner (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017
- Paolo Di Lorenzo (2020). usmap: US Maps Including Alaska and Hawaii. R package version 0.5.1. https://CRAN.R-project.org/package=usmap
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). "ROCR: visualizing classifier performance in R." *Bioinformatics*, *21*(20), 7881. <URL: http://rocr.bioinf.mpi-sb.mpg.de>.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

Appendix

Table 6: Parameter Estimate for our model

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.725	3.119	0.076	6.363
genderMale	0.664	1.060	0.592	0.745
age20M29	0.623	1.213	0.425	0.909
age30M39	0.404	1.208	0.279	0.577
age40M49	0.336	1.211	0.231	0.484
age50M59	0.354	1.217	0.240	0.521
age60M69	0.367	1.213	0.252	0.535
age70M79	0.358	1.233	0.237	0.535
age80M89	0.298	1.419	0.149	0.582
age90M99	0.600	3.218	0.059	6.345
race_ethnicityblackDafricanamericanDnegro	6.097	1.298	3.612	10.049
race_ethnicitychinese	2.393	1.450	1.164	4.979
race_ethnicityjapanese	2.993	1.946	0.879	11.887
race_ethnicityotherasianorpacificislander	1.478	1.325	0.858	2.584
race_ethnicityotherracenec	1.615	1.300	0.971	2.689
race_ethnicitywhite	0.799	1.268	0.507	1.263
household_income$125000to$149999	1.052	1.168	0.771	1.422
household_income$15000to$19999	1.794	1.182	1.301	2.488
household_income$150000to$174999	1.129	1.210	0.773	1.637
household_income$175000to$199999	0.716	1.261	0.456	1.124
household_income$20000to$24999	1.325	1.178	0.960	1.841
household_income$200000to$249999	0.600	1.238	0.391	0.908
household_income$25000to$29999	1.414	1.174	1.037	1.937
household_income$250000andabove	0.760	1.246	0.494	1.166
household_income$30000to$34999	1.534	1.173	1.123	2.073
household_income$35000to$39999	1.476	1.181	1.075	2.045
household_income$40000to$44999	1.701	1.190	1.215	2.400
household_income$45000to$49999	1.445	1.183	1.047	2.006
household_income$50000to$54999	1.196	1.173	0.877	1.637
household_income$55000to$59999	1.055	1.222	0.716	1.556
household_income$60000to$64999	1.545	1.210	1.068	2.239
household_income$65000to$69999	1.216	1.248	0.787	1.836
household_income$70000to$74999	1.432	1.210	0.979	2.075
household_income$75000to$79999	1.239	1.212	0.844	1.784
household_income$80000to$84999	1.708	1.254	1.086	2.682
household_income$85000to$89999	1.394	1.283	0.855	2.284
household_income$90000to$94999	1.179	1.303	0.699	1.982
household_income$95000to$99999	1.620	1.223	1.096	2.415
household_incomeLessthan$14999	1.739	1.145	1.339	2.271
educationAssociateDegree	2.475	2.003	0.614	9.817
educationCollegeDegreesuchasB.A.B.S.	2.402	1.996	0.599	9.686
educationCompletedsomecollegebutnodegree	2.082	1.998	0.522	8.311
educationCompletedsomegraduatebutnodegree	2.437	2.024	0.591	9.866
educationCompletedsomehighschool	1.425	1.998	0.360	5.507
educationDoctoratedegree	1.828	2.059	0.443	7.753
educationHighschoolgraduate	1.587	1.991	0.405	6.165
educationMastersdegree	2.601	2.004	0.646	10.461
educationMiddleSchoolMGrades4M8	1.072	2.321	0.204	5.800
educationOtherposthighschoolvocationaltraining	1.585	2.008	0.396	6.239
stateAL	1.033	2.446	0.184	6.173
stateAR	0.742	2.536	0.128	4.689

	Estimate	Est.Error	Q2.5	Q97.5
stateAZ	1.472	2.395	0.274	8.380
stateCA	2.251	2.372	0.435	12.722
stateCO	1.970	2.427	0.363	11.202
stateCT	3.782	2.468	0.649	21.979
stateDC	2.926	2.697	0.413	20.158
stateDE	3.210	2.666	0.504	22.613
stateFL	1.491	2.374	0.284	8.236
stateGA	1.111	2.402	0.207	6.307
stateHI	1.535	2.639	0.238	10.469
stateIA	1.801	2.485	0.313	10.985
stateID	0.654	2.602	0.104	4.393
stateIL	1.946	2.391	0.368	11.343
stateIN	1.548	2.418	0.285	9.112
stateKS	0.926	2.519	0.161	5.885
stateKY	1.472	2.448	0.270	8.780
stateLA	1.213	2.459	0.223	7.301
stateMA	4.035	2.417	0.698	23.150
stateMD	2.050	2.440	0.374	12.388
stateME	2.662	2.650	0.407	18.650
stateMI	1.997	2.397	0.379	11.564
stateMN	1.874	2.465	0.331	11.696
stateMO	1.815	2.406	0.344	10.670
stateMS	0.990	2.561	0.164	6.925
stateMT	1.569	2.786	0.215	11.881
stateNC	1.580	2.395	0.302	8.799
stateND	0.545	3.746	0.036	6.879
stateNE	1.309	2.640	0.193	8.739
stateNH	2.487	2.731	0.344	18.352
stateNJ	2.091	2.396	0.385	12.029
stateNM	3.427	2.725	0.502	25.720
stateNV	1.283	2.468	0.235	7.959
stateNY	2.077	2.377	0.393	11.722
stateOH	1.652	2.379	0.308	9.316
stateOK	1.254	2.455	0.223	7.333
stateOR	2.427	2.431	0.443	14.157
statePA	1.416	2.384	0.270	7.985
stateRI	1.378	3.043	0.158	12.122
stateSC	0.775	2.431	0.141	4.398
stateSD	1.041	2.865	0.136	8.208
stateTN	1.177	2.423	0.213	7.226
stateTX	1.299	2.380	0.244	7.267
stateUT	1.280	2.536	0.208	7.912
stateVA	2.063	2.396	0.379	11.703
stateVT	12.244	3.467	1.339	162.791
stateWA	2.361	2.412	0.447	13.073
stateWI	2.546	2.411	0.468	14.873
stateWV	1.460	2.519	0.245	9.173
stateWY	1.069	5.687	0.029	26.683

Git Repository

The files to the R scripts are in this GitHub Repository: https://github.com/frankkhung/us_election