Binary Logistic Regression Model: Joe Biden predicted to win by popular vote

Mackenzie Qu, Bianca Pokhrel, Zhuoqian Li

03 November 2020

Abstract

The result of the 2020 presidential election in the US has been a pressing topic. In this paper, we aim to predict the cadidate who will win the popular vote by a binary logistic model using race, gender, age, state, and household income as predictors. Our result suggests that Joe Biden will surpass Donald Trump in popular vote significantly, thus further secure the presidency.

Keywords: forecasting; US 2020 election; Trump; Biden; Multilevel Regression with Post Stratification

Contents

1 Introduction

On November 3rd, voters will cast their last ballot for the 2020 U.S presidential election as either Donald Trump or Joe Biden will become the next president of the United States of America. Americans are eager to acquire more information of the upcoming election from various sources of election forecast. In this paper, we explored the data from Democracy Fund + UCLA Nationscape and American Community Surveys in a binary logistic regression, aiming to predict the election result by popular vote.

Upon fitting the datasets into our model, we have made some significant findings. Our model suggests that Biden is likely to win the popular vote by 16.32%, compared to the last election where Hilary Clinton won the popular vote by 2%, yet still lost the election. This finding is significant as it gives us more confidence that Joe Biden may win the election. We also looked at the geographic representations of popular vote. In general, East and West Coast shows higher preference of Biden in comparison to mid-west and south. We have also looked at some socio-demographic factors such as age, gender, race and income. Given the current pandemic, we have observed some interesting result from the variables above, which will be throughly explored in the discussion. The data wrangling and model for this paper is done using statistical language R Core Team [2020] in Rmarkdown(Allaire et al. [2020])(Xie et al. [2018])(Xie et al. [2020]). To reproduce the result, code can be accessed at: "https://github.com/bianca-pokhrel/forecasting-us-election-2020"

In conclusion, our findings suggest that Biden may win this election by popular vote. The remainder of this paper further discusses the two survey data respectively in section 2 and 3, model design and justification in section 4, and results in section 5, as well as discussion in section 6, intending to provide a explanation for our prediction. That being said, our results are limited to forecasting the popular vote, while having little ability to predict the winner of this election as it does not include information on electoral colleges.

2 Democracy Fund + UCLA Nationscape Survey Data

2.1 Survey

The individual-level data used is the result of the Democracy Fund + UCLA Nationscape survey(Tausanovitch and Vavreck [2020]) conducted on June 25th 2020. Nationscape conducts various interviews over 80 weeks, targeting Americans over the age of 18 to gain an insight of the 2020 election. The specific survey we have used contains 6,479 interview responses.

The Democracy Fund + UCLA Nationscape survey was conducted as a non-probability online survey, with samples provided by Lucid, a third party market research platform. Prior to publishing the survey, a demographic quota was first decided on age, gender, ethnicity, region, income, and education. Such a quota insures the representation of all American voters. Upon setting up the quota, the respondents are then sourced by Lucid Marketplace suppliers, each with different survey methodology including target emails, online portal, offerwalls, and SMS or in-app messaging(HQ [2018]), and sent directly to the survey software operated by Nationscape.

The survey contains 265 questions in three major catagories First of all the survey contains questions about the respondents' attitudes, such as whether they approve Donald Trump handling his job as president. Second type of questions asks about respondents' behavior, such as vote intention. Thirdly, the survey asks the respondents' to state the facts of their lives, such as their gender, age, or whether they have been sick with coronavirus. Each type of question serves its unique purpose. In particular, Nationscape used the results of the third type of question(i.e. facts) to compare with the results from large government surveys. The survey data are weighted from the 2017 American Community Survey to represent the American population using simple ranking techniques. More details of the weighting technique is discussed in Appendix 1. In addition, some adjustments are done to the data including handling non-response. First of all, voting rates in 2016 were adjusted. Some younger respondents reported voting in 2016 even though their age indicated that they were eligible to vote. As a result, a -2.7 percent point adjustment is made for young respondents. Moreover,

some respondents may choose not to disclose household income. In effect, non-respondents are not weighted for income.

Overall, the survey is self contained. Though Nationscape has chosen a third party for sampling, Lucid's sample has been proven high quality by previous evaluations(Coppock and Green [2016]). Moreover, Lucid has its own quality program to evaluate and ensure the sample quality and minimize sampling bias. The survey covers the majority of the predictors of the 2020 election in detail, which provides us an insight of the electoral forecast, as well as a wide range of explanatory variables to base the model off.

However, some weaknesses of the survey must be taken into account. First of all, not only has the survey taken a non-probability quota approach, the sampling method also differs depending on the suppliers, thus does not provide a consistent sampling technique. Consequently, the data may face selection bias and bias due to non-response. Secondly, though the simple ranking technique used for weighting has an advantage of time and cost efficiency, its explanatory power decreases quickly with an increasing number of criteria(Alliance [2016]). It is justified that more complicated weighting methods would not provide significant benefit(TAUSANOVITCH et al. [2019]), however, it still may increase the overall representativeness of the data. Moreover, the survey was completely voluntary via online software, yet, more than 200 questions were asked. Some may question the quality of responses for such a time consuming survey. Also the mid drop out rate may be accountable for non-response bias.

2.2 Dataset

Among over 200 survey questions, we have first used the vote intention to filter out respondents who are either not eligible to vote, or choose not to. We would like to determine whether Donald Trump or Joe Biden would gain popular votes, thus have chosen some specific variables, aiming to predict the election result. In particular, demographic variables such as state, census region, age, gender, race, as well as income are selected for the purpose of this paper. In addition, respondents' political stance and vote history are included in this section, even though it is not used for modeling. It serves the purpose of visualizing vote intentions based on history and ideologies, and may provide further insights of the election.
Our data wrangling is completed in the statistical language R(R Core Team [2020]), using haven(Wickham and Miller [2020]), tidyverse(Wickham et al. [2019]), dplyr(Wickham et al. [2020]), ggplot2(Wickham [2016]), usmap(Di Lorenzo [2020]), waffle(Rudis and Gandy [2017]), tydyr(Wickham [2020]), knitr(Xie [2020])(Xie [2015])(Xie [2014]), gridExtra(Auguie [2017]), and kabelExtra(Zhu [2020]).

2.3 Data Preview

Below(Table **??**) is a preview

registration	vote_2016	vote_intention	trump_biden	ideo5	gender
Registered	Donald Trump	Yes, I will vote	Trump	Conservative	Female
Registered	Others	Yes, I will vote	Trump	Conservative	Female
Registered	Donald Trump	Yes, I will vote	Trump	Conservative	Female
Registered	Donald Trump	Yes, I will vote	Trump	Conservative	Female
Registered	Donald Trump	Yes, I will vote	Trump	Very Conservative	Female
Not registered	Others	No, I am not eligible to vote	Biden	Liberal	Female

race_ethnicity	household_income	state	census_region	age	age_group	state_name
White	$75,000 to $99,999	WI	Midwest	49	45 to 59	Wisconsin
White	$100,000 to $124,999	VA	South	39	30 to 44	Virginia
White	$175,000 to $199,999	VA	South	46	45 to 59	Virginia
White	$50,000 to $74,999	TX	South	75	60+	Texas
White	Less than $24,999	WA	West	52	45 to 59	Washington
White	Less than $24,999	OH	Midwest	44	30 to 44	Ohio

2.4 Data Discussion

The graphs of our raw data contain both election forecasts by popular vote, and visualizations of predictor variables used in the model. We are going to discuss some interesting observations.

First of all, we plotted the expected vote counts for both Donald Trump and Joe Biden(Figure 1). The plot shows 3,068 confirmed voters among the respondents would vote for Biden, while 2,619 would vote for Trump, suggesting that Biden is expected to exceed Trump's vote by 7.89%.

Secondly, when looking at the expected popular vote for each state(Figure 2), it is noticeable that the west and east coast lean towards Biden and his Democratic Party. On the other hand, Trump is comparatively more popular in the Midwest and South region.

Thirdly, the political ideology of Trump/Biden supporters also provides crucial information for the upcoming election. It is indicated in Figure 3 that the majority of Biden's votes are gained from liberal/very liberal individuals. In comparison, Trump's majority votes are from a more conservative spectrum.

Also, Figure 4 not only suggests that Trump has a large number of repeated voters, but it also shows that Biden has gained the majority of voters who voted Hilary Clinton in 2016, as well as people who did not vote.(Figure 4) This finding is significant as it shows more people have decided to vote.

Figure 5 to Figure 8 visualizes the distribution of some explanatory variables for our model. Here are some interesting findings:

Biden has a majority of female voters, while Trump has a majority of male voters(Figure 5). Although the majority of both Trump and Biden voters are white, minority racial groups are biased towards Biden(Figure 6). Biden is preferred among younger voters(Figure 7). Despite that the distribution of Trump and Biden's votes by household income is similar, more low income households prefer Biden over Trump(Figure 8).

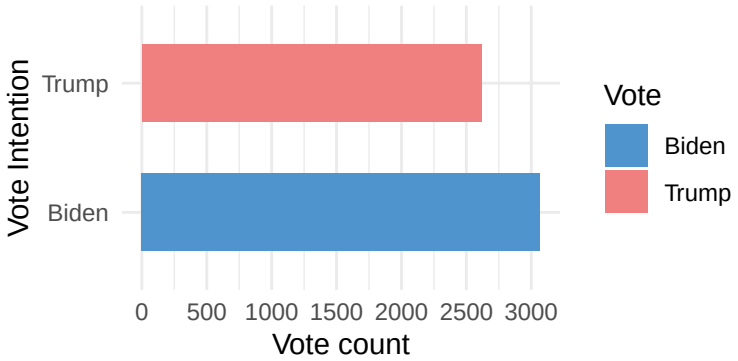


Figure 1: Vote intention 2020

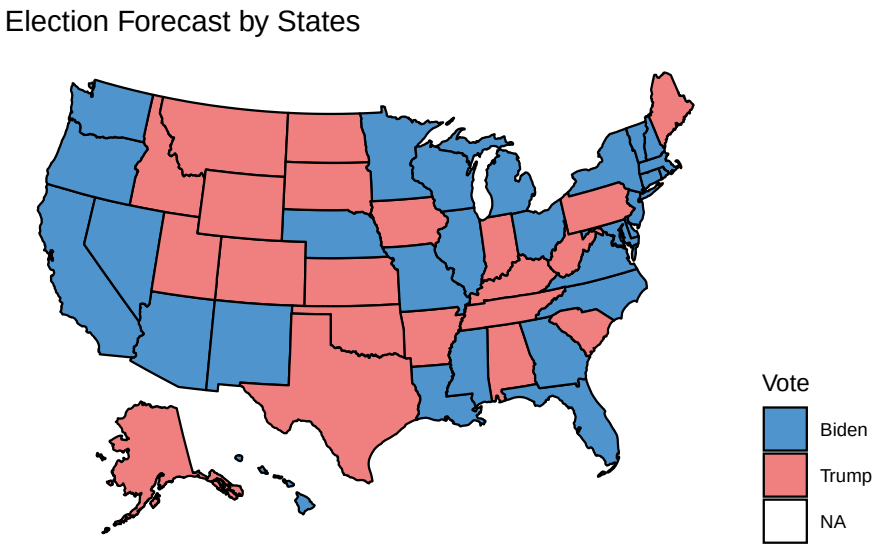


Figure 2: Election forecast by States

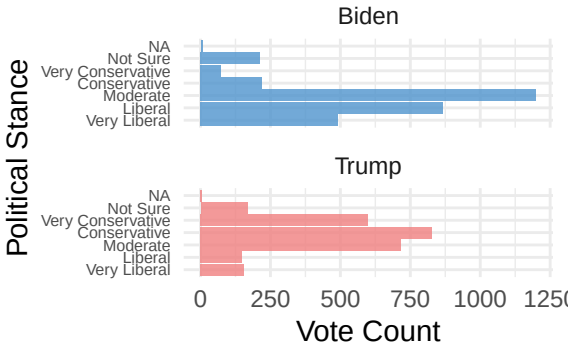


Figure 3: Vote intention by political stance

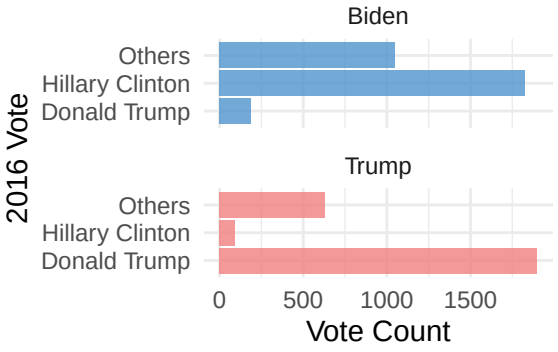


Figure 4: Vote intention by 2016 vote

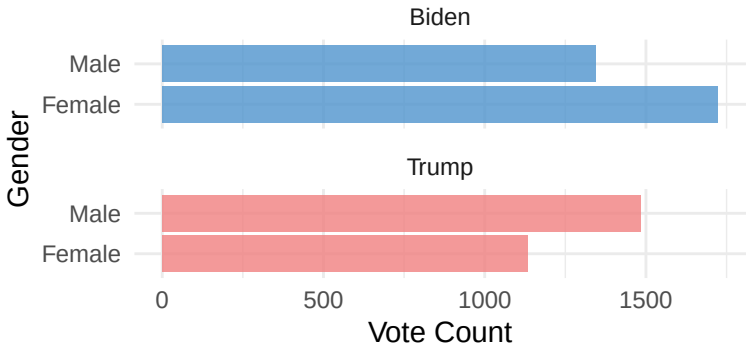


Figure 5: Vote intention by gender

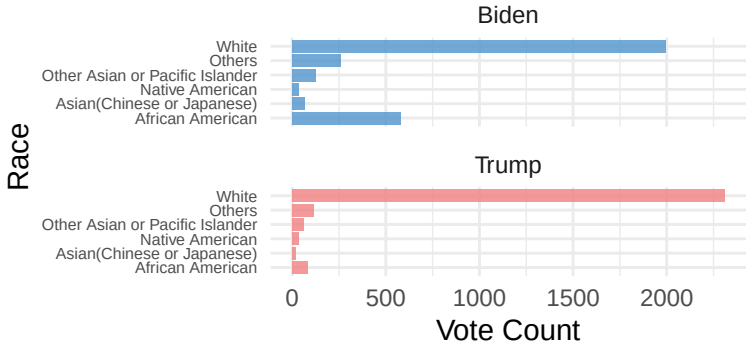


Figure 6: Vote intention by race

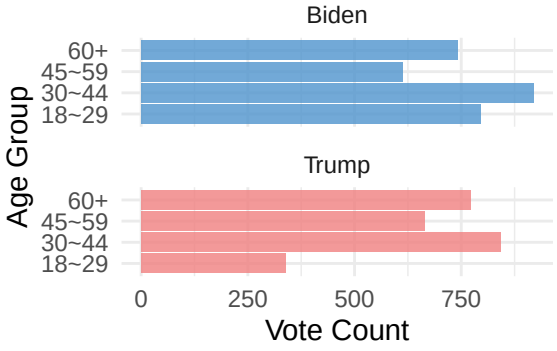


Figure 7: Vote intention by age

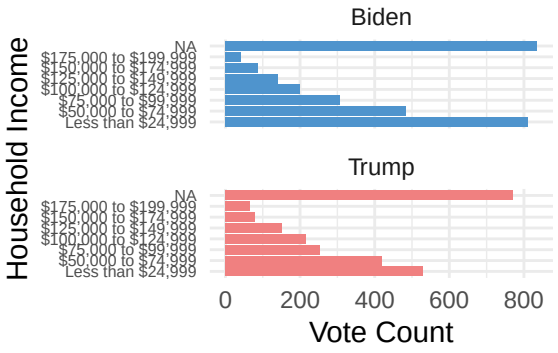


Figure 8: Vote intention by household income

3 Post-Stratification Data

3.1 Data Source and The American Community Survey (ACS)

The data used to construct the post-stratification dataset was a selected subset of the data collected through the American Community Surveys (Ruggles et al. [2018]). The data was retrieved from IPUMS with a density of 0.2 due to the full size of the dataset being quite large. Data from 2018 was chosen since the goal of the model is to predict 2020 election results, as a result, it was ideal to choose the most recent dataset that was available in IPUMS.

The ACS run by the Census Bureau is performed by sending a survey to a sample of the population in the 50 states every month of every year. The population of this survey would be the American population, and the frame would be the 140 million addresses of American citizens available to the Census Bureau. Respondents are chosen by their address to ensure geographical coverage, and no respondent is to be selected more than once within 5 years. According to the Census Bureau, the survey is mailed to approximately 295,000 respondents a month. Respondents are allowed to respond to the survey in multiple ways such as by mail, telephone, internet, or in-person interviews. Furthermore, it can be noted that the addresses selected for the ACS are required by law to accurately respond to all questions (cit [2017]). The multitude of methods that respondents can use to respond to the survey, as well as the legal component, allow for the Census Bureau to collect data on most of the 295,000 chosen addresses that make up the sample.

In addition, the guide released by the Census Bureau states that in person interviews are conducted for the addresses if there had been a non-response for 3 months. Moreover, the Census Bureau conducted telephone follow-ups for the questionnaires that were returned incomplete or needed clarification (cit [2017]).

The main strength of the survey is that it is conducted very often and the survey collects a large amount and variety of data on the respondents. As a result, the ACS can provide an accurate and current reflection of American communities. Moreover, the survey is conducted by an official government organization, which has access to a large amount of resources to facilitate the survey distribution and data collection. Furthermore, respondents are allowed to respond to the survey in multiple ways such as by mail, telephone, internet, or in-person interviews.

3.2 Dataset

The variables included in the post stratification dataset were the variables that were determined to have a larger impact in the model. As a result, the variables included were age, gender, income, census regions, state, and race. The age variable represents the age range which corresponds to the age of respondent. Originally the age variable was just the age of the respondent when extracted from the IPUMS data. However, it was changed to be a range for the development of a model since it is easier for the model to determine the impact of a range of ages in comparison to exact age. Similarly, income groups were built from household income for the same reason.

In addition, state was chose to represent geographical location over census region in the final model due to there being 50 states in the US. The large number of states could help draw more specific conclusions.

Below is a preview of the finalized cleaned dataset(Table **??**) with counts for each subgroup. However, it should be notified that the count for post stratification can be relatively small to the overall size of the dataset since we have state as our predictor, which can have 50 possible values.

gender	age_group	household_income	race_ethnicity	state_name	n
Female	18 to 29	$100,000 to $124,999	African American	new york	3
Female	18 to 29	$100,000 to $124,999	African American	pennsylvania	2
Female	18 to 29	$100,000 to $124,999	African American	ohio	1
Female	18 to 29	$100,000 to $124,999	African American	wisconsin	1
Female	18 to 29	$100,000 to $124,999	African American	minnesota	2
Female	18 to 29	$100,000 to $124,999	African American	florida	4

(Figure 9) shows the distribution of our predictor variables

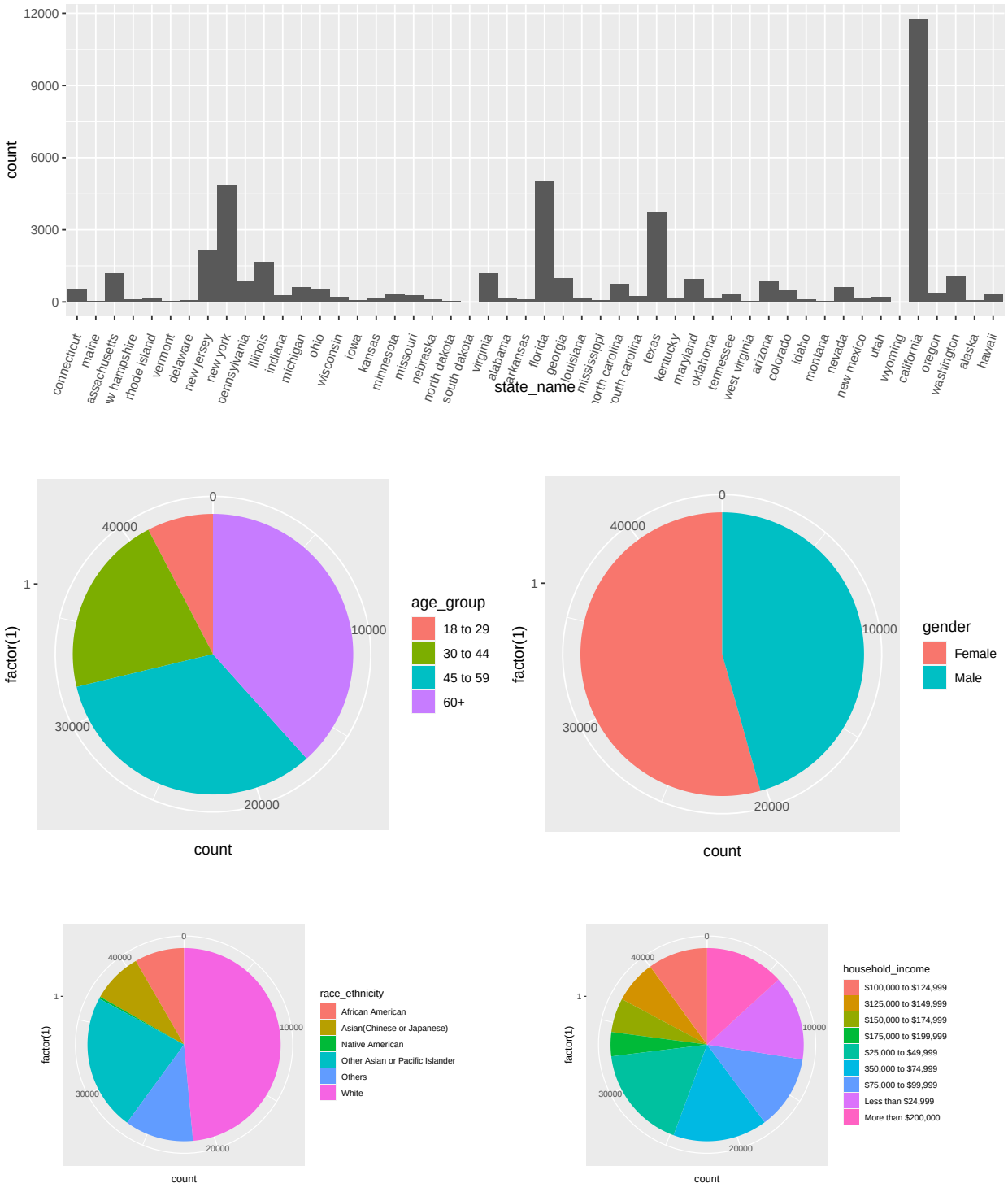


Figure 9: Distribution of Predictor Variables

(Figure 9) demonstrate the distribution of the variables used for the development of the model. We can notice that the distribution of the variables are different but close to the distribution obtained by the survey mentioned in the previous section where there are not any major discrepancies. It can be noted that a lot of respondents are from California, Texas, and Florida, which is representative of the distribution of the american population among the 50 states.

3.3 Overview of MRP

Multilevel modelling with post stratification (MRP) is often used to help bridge the difference between non-representative survey samples and the actual demographics of the population. MRP is often done by first training a model from the survey data where the individual response is estimated. The estimation of the effect of individual predictors is typically improved through the use of multilevel modeling. Then the model is applied onto another dataset whose percentages of each type of sub population is more accurate to re-weight the estimates (Gelman et al. [2018]). A benefit of this method is that it permits the use of broad surveys for analysis with regards to subgroups. However, it should be noted that there will be a larger uncertainty for the estimates. A caveat of MRP is that it requires the assumption that we know the composition of our population. Although this point is not as applicable to the point of research in this paper, it could cause an issue when trying to post-stratify a variable whose distribution is not clearly known (Simpson [2019]).

4 Model

4.1 Variable Selection

Our primary goal in this report is to be able to cast predictions on the upcoming U.S. presidential election. The importance of forecasting the 2020 election cannot be understated after the fiasco of the 2016 election results. In this section we will talk about our model development process and how we decided to use a logistic regression using the glm() function in R.

As discussed in the previous data section, we have taken particular interest in the following variables:

-gender (categorical)
-race/ethnicity (categorical) -household income (categorical)
-states (categorical)
-age group (categorical)

We have picked gender as a categorical variable of interest as from the 2016 election, the percentage of women who voted for Trump was 42% (Tyson and Maniam [2020]). We would like to see how over the course of his administration, that percentage has potentially changed. Gender is considered a categorical variable as there are specific categories of gender and there is no intrinsic ordering to the categorical- this is also called a nominal variable.

Secondly, we have picked race/ethnicity as a categorical variable of interest. In the current climate, the U.S. remains extremely racially divided. Since 2016 we have witnessed the White House's response to police brutality, xenophobia, islamophobia, etc. We would like to further investigate how much of a role race plays in whether Trump or Biden gets a vote.

Thirdly, we have chosen household income as a categorical variable of interest. Tax breaks and "taxing the rich" is a common rhetoric used by politicians. We want to know who this rhetoric is affecting and how one's income bracket affects who they vote for.

Fourth, we have chosen states as a categorical variable of interest. Historically, states are considered either 'red' or blue.' In election polling, an important variable to keep an eye on are the regions that may be 'swing' regions. For example, southern states such as Florida have been known to be a crucial state that can 'swing' the election spontaneously. Thus, we want to determine the significance of what regions people have answered this survey from and how crucial it is to forecasting an election. Rather than grouping states together by census region, we decided we want to be able to create categories for each of the states as we are then able to group by states and look at voting predictions in each of the states. It yields a more informative result and allows us to take note of potential 'swing-states' mentioned above.

Lastly, we have chosen age groups as a categorical variable of interest. Since the 2016 election, there has been a disparity between the younger and older populations.This is in part due to topics of political interest such climate change among others.This, we want to determine the significance of what the current voting demographic is, and if they have significant sway in the determination of the presidential winner. We decided to group age into categories rather than a continuous variable as this allows us to have larger cell counts in the post-stratification phase when we look at the percentage of Trump/Biden voters in age categories. Additionally, having categories increases the readability of the model predictions as we are able to give the average reader a better understanding of the results generated.

For our response variable, we have created a binary variable with 1 = Trump and 0 = Biden. Thus, to forecast the election, a vote percentage over 50% will result in the potential victory for Trump.

4.1.1 Building the model

To build the model, we will first start with the OLS Linear Regression Model and work our way up in complexity.

4.1.2 OLS Linear Regression Model

Generally, the OLS Linear Regression Model is used for continuous independent variables rather than a binary outcome variable.Despite this, it is generally good practice to check to see the fit of the data to a linear model (**?**). Therefore we will model our variables of interest in the context of linear regression.

Let x_1, X_2, X_3, X_4, X_5 represent gender, race/ethnicity, household income, states, and age group respectively. Let Y_i represent the i^{th} response variable observation- the binary variable of Trump or Biden. We can then model our regression as:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i$$

Where the beta coefficients represent the slope of the independent variables and the epsilon represents the error term for the i^{th} observation.

There are a few downfalls with this model. Firstly, the independent variables in our model are categorical rather than continuous. This means that our regression line will not be following a trend as the independent variables are discrete and will not follow a linear trend. Additionally, if we take a look at the following normal-QQ plot, we can see that our data shows a more S-shape rather than a linear relationship.

Clearly, we can see that this model is not fit for the data we are working with and cannot properly predict election results

4.1.3 Logistic Model Using glm

Next, we will use a generalized linear model. This model allows us to define the family of linear modeling we are using. As we are using a binary outcome variable and logistic regression, we will use the binomial family with the logit function.

For a binomial regression model, if Y_i for $i = 1, ..., n$ is independently and identically distributed binomial distribution, then we can model it as the following:

$$p(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

Where the n_i is the sample size of the ith observation and y_i is the response of the ith observation, p_i is the probability of observation y_i being 1.

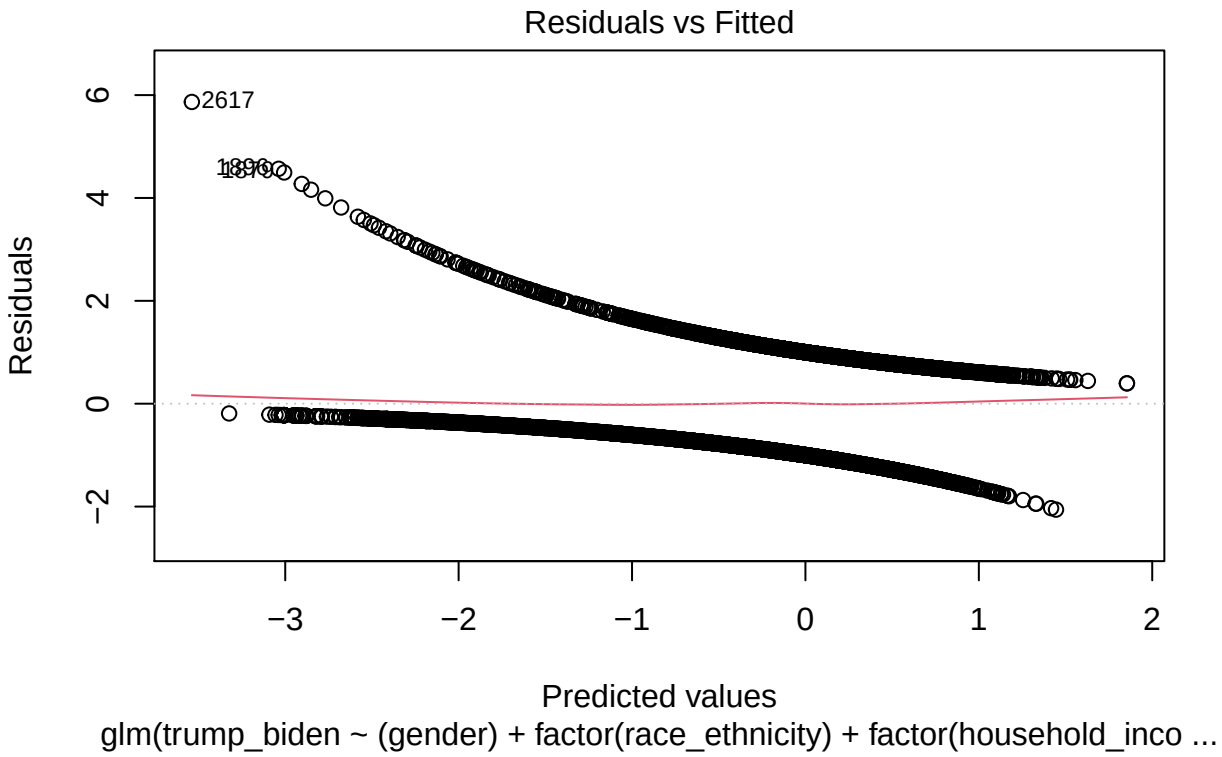
Additionally, assume these are affected by q predictors listed as the x values from above.

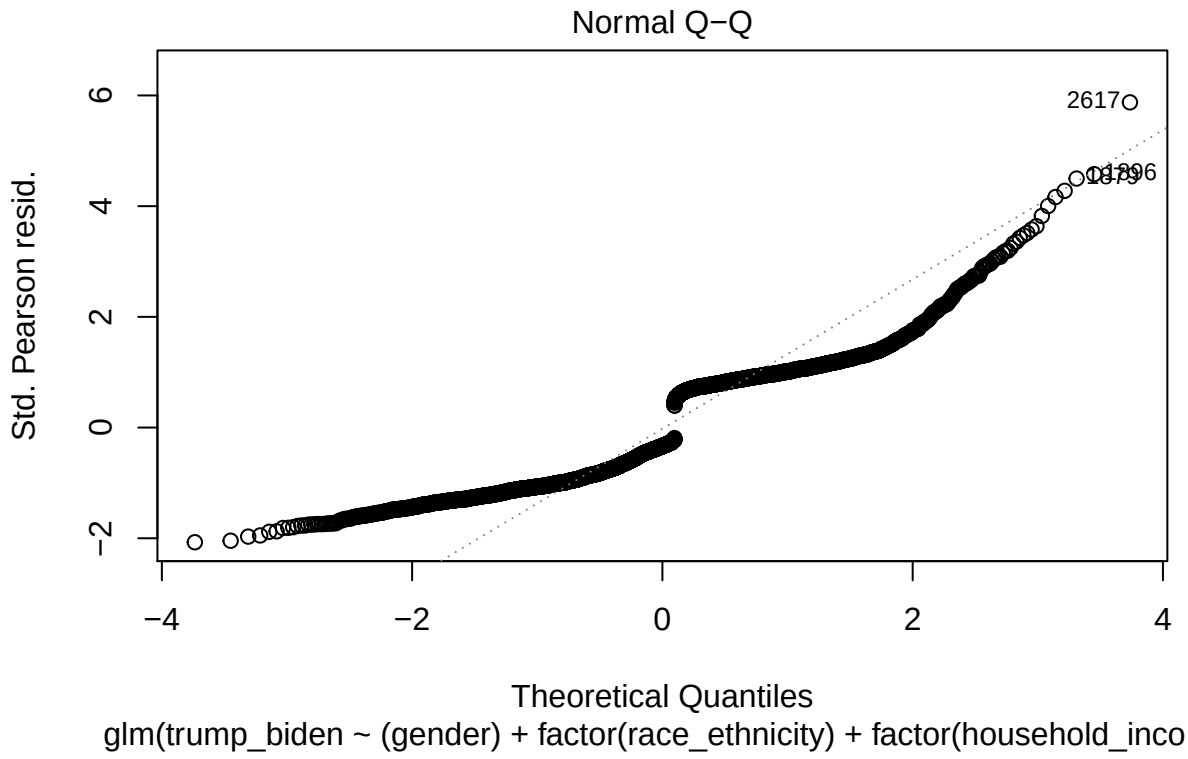
To connect them, we can use the linear model above as a predictor. To do this we require a function to link the two together. In the case of logistic regression we can use the logit function:

$$logit = log(\frac{p}{1-p}), p = \frac{exp(log(-log(1-p)))}{1 + exp(log(-log(1-p)))}[1 + exp(-(log(-log(1-p))))]$$

where the variable definitions remain the same as listed above.

To validate this model, we can perform a series of checks. First we can check the Normal QQ plot to check to see if the fit of our model is better than the OLS Linear Model:


Residuals vs Fitted
Residuals
Predicted values
glm(trump_biden ~ (gender) + factor(race_ethnicity) + factor(household_inco ...
2617
6
4
2
0
−2
−3
−2
−1
0
1
2



From the plot above, we can see that the model is clearly a better fit than the OLS estimator.

Additionally, we can perform a series of checks such as Anova chi-squared tests and AIC values.

The AIC value of the created model is 6796.309. We will use this for comparison and we are looking for AIC values lower than the full fitted model to decide if we should drop the variable or not.

4.1.3.1 Without State We can see that the anova checks resulted in a larger deviation after dropping the state variable. Additionally, the AIC value increased thus confirming that this is a signifcant variable

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
5341	6787.649	NA	NA	NA
5292	6662.309	49	125.3391	0

x
6823.649

4.1.3.2 Without Gender We can see that the anova checks resulted in a larger deviation after dropping the gender variable. Additionally, the AIC value increased thus confirming that this is a signifcant variable

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
5293	6705.052	NA	NA	NA
5292	6662.309	1	42.74256	0

x
6837.052

4.1.3.3 Without Race We can see that the anova checks resulted in a larger deviation after dropping the race variable. Additionally, the AIC value increased thus confirming that this is a signifcant variable

```
## Analysis of Deviance Table
##
## Model 1: trump_biden ~ (gender) + factor(household_income) + factor(state_name) +
##     factor(age_group)
## Model 2: trump_biden ~ (gender) + factor(race_ethnicity) + factor(household_income) +
##     factor(state_name) + factor(age_group)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5297     7025.3
## 2      5292     6662.3  5   362.94 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

x
7149.251

4.1.3.4 Without Income We can see that the anova checks resulted in a larger deviation after dropping the income variable. Additionally, the AIC value increased thus confirming that this is a signifcant variable

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
5300	6690.660	NA	NA	NA
5292	6662.309	8	28.35032	0.000412

x
6808.66

4.1.3.5 Without Age We can see that the anova checks resulted in a larger deviation after dropping the age variable. Additionally, the AIC value increased thus confirming that this is a signifcant variable

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
5295	6724.422	NA	NA	NA
5292	6662.309	3	62.11241	0

x
6823.649

After performing the checks, we can see that the fully fitted logistic model calculated above is the optimal one. However, it is not the only model possible. Other variants exist such as the Logistic model with Bayesian Estimation. It is recommended that if there are many indpendent variables, the Bayesian model should be used however, in the context of computational power, this was not feasible for us. Building the Bayesian model took near an hour with only 2 chains thus, we decided in order to keep this as reproducible as possible, we will use the GLM model instead.

```
## # A tibble: 1 x 3
##    mean lower higher
##   <dbl> <dbl>  <dbl>
## 1 0.404 0.315  0.464
```

```
## # A tibble: 1 x 3
##    mean lower higher
##   <dbl> <dbl>  <dbl>
## 1 0.434 0.391  0.477
```

```
## # A tibble: 1 x 3
##    mean lower higher
##   <dbl> <dbl>  <dbl>
## 1 0.364 0.148  0.533
```

```
## # A tibble: 1 x 3
##    mean lower higher
##   <dbl> <dbl>  <dbl>
## 1 0.471 0.275  0.659
```

```
## # A tibble: 1 x 3
##    mean lower higher
##   <dbl> <dbl>  <dbl>
## 1 0.435 0.376  0.528
```

5 Results

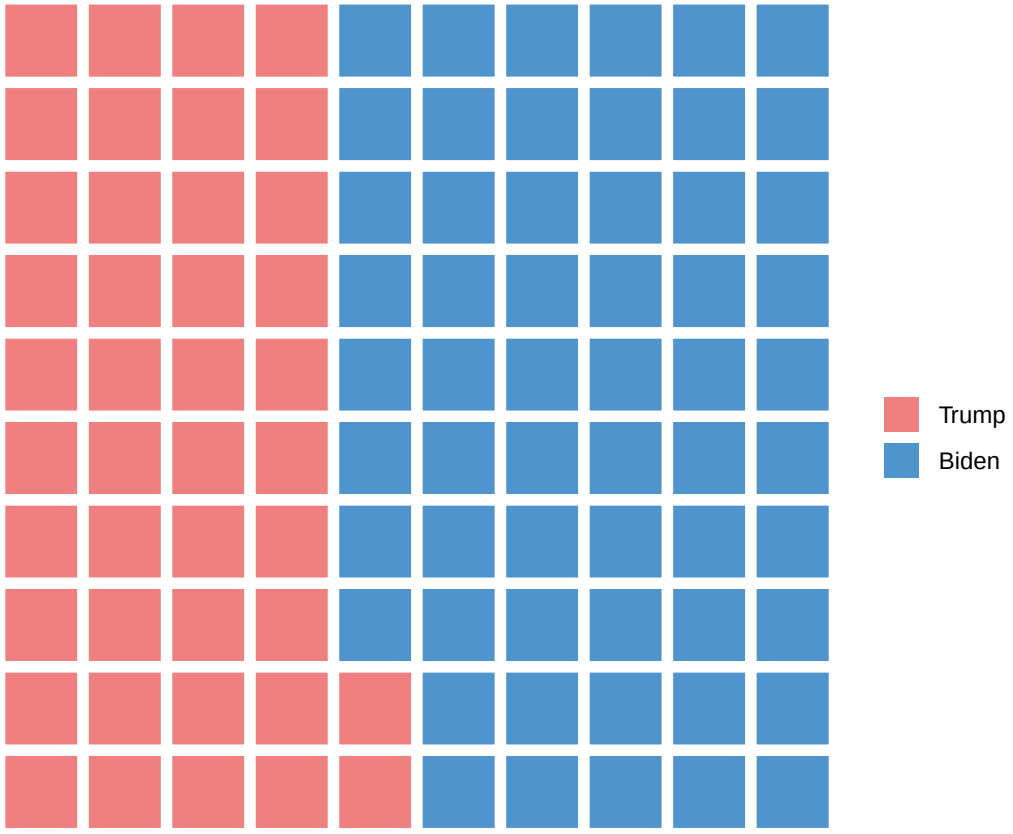


Figure 10: Visualization of Trump and Biden's expected vote in 100 people

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

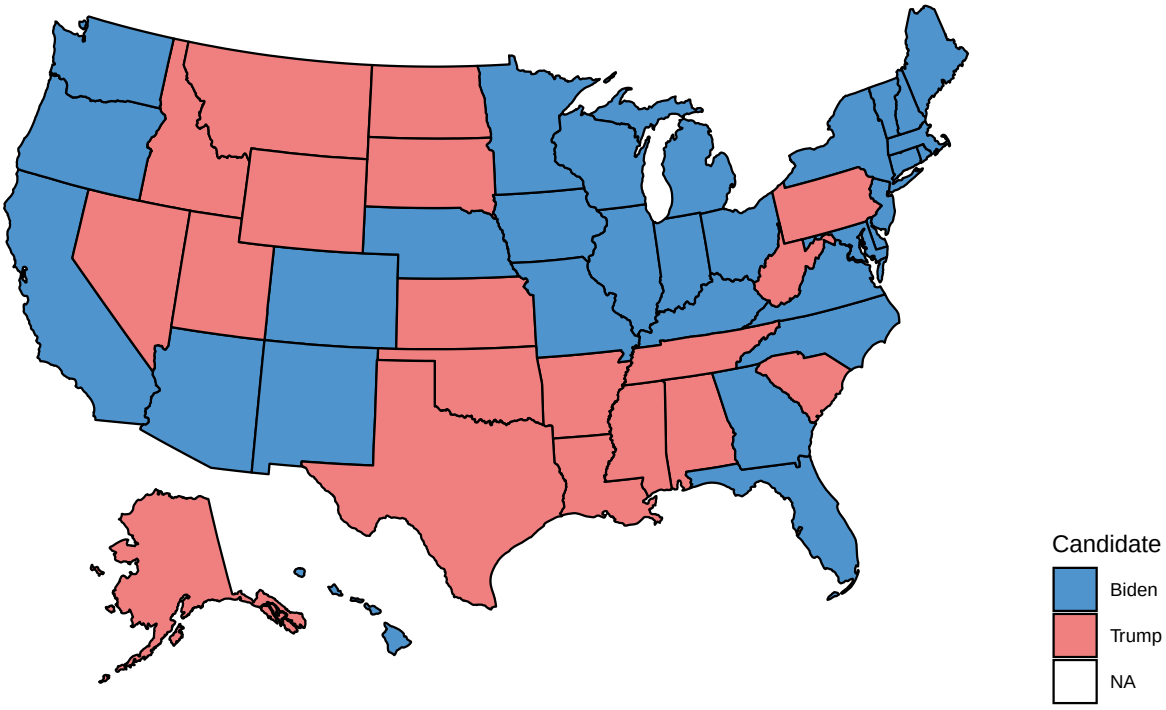


Figure 11: Popular vote perdiction by State

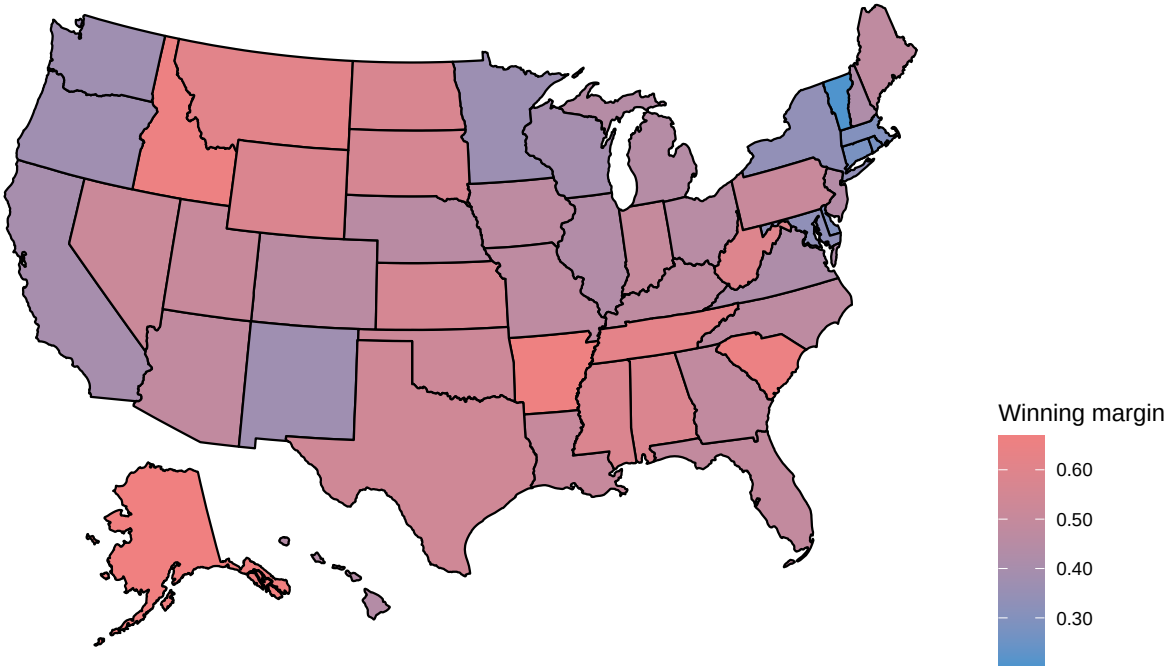


Figure 12: Winning Margin for Trump and Republican Party by popular vote

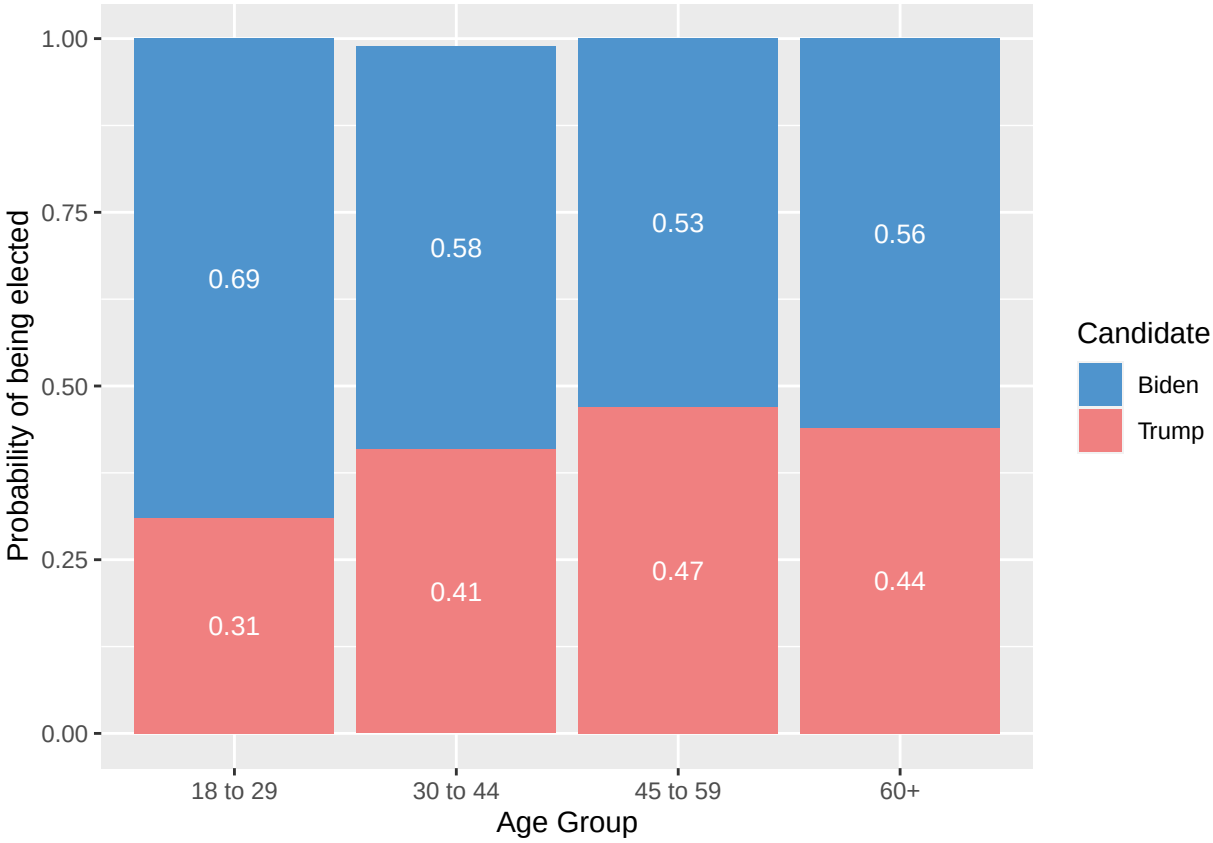


Figure 13: Probability of vote by age

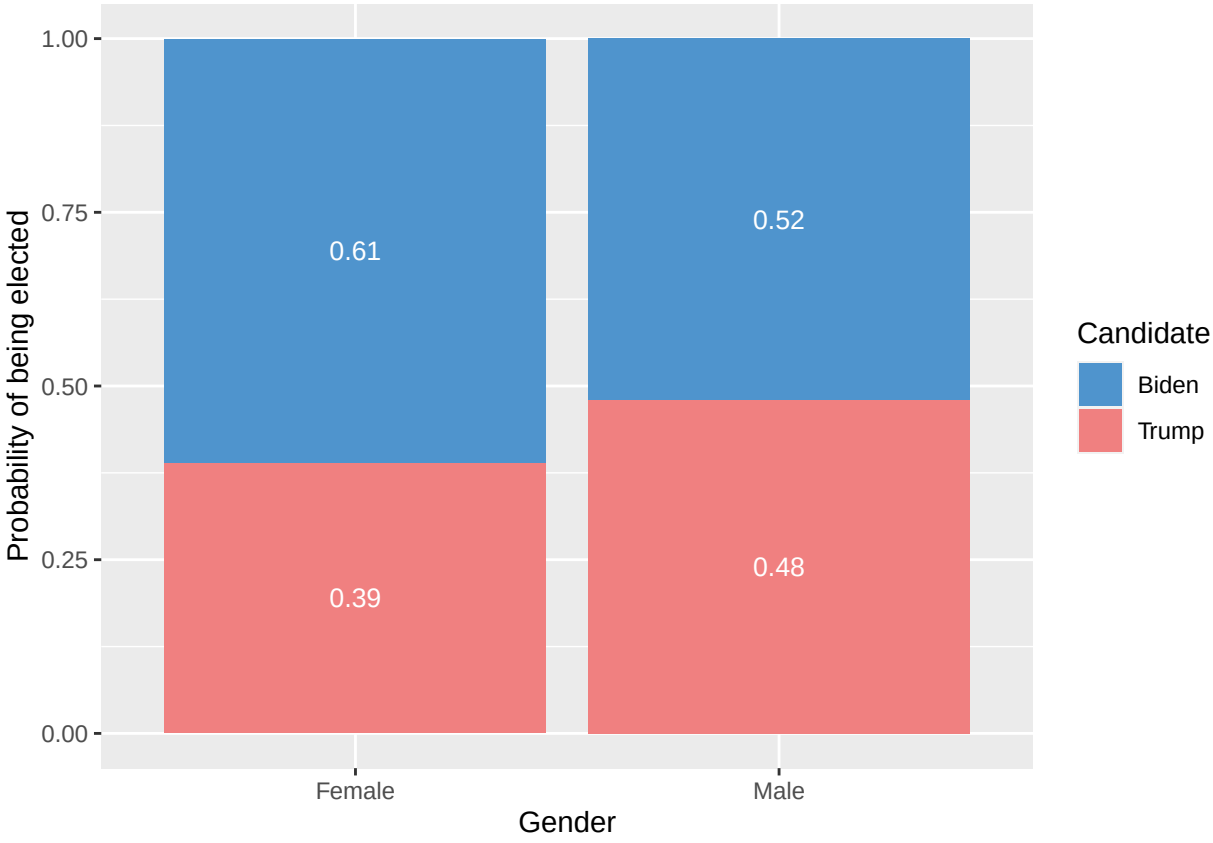


Figure 14: Probability of vote by gender

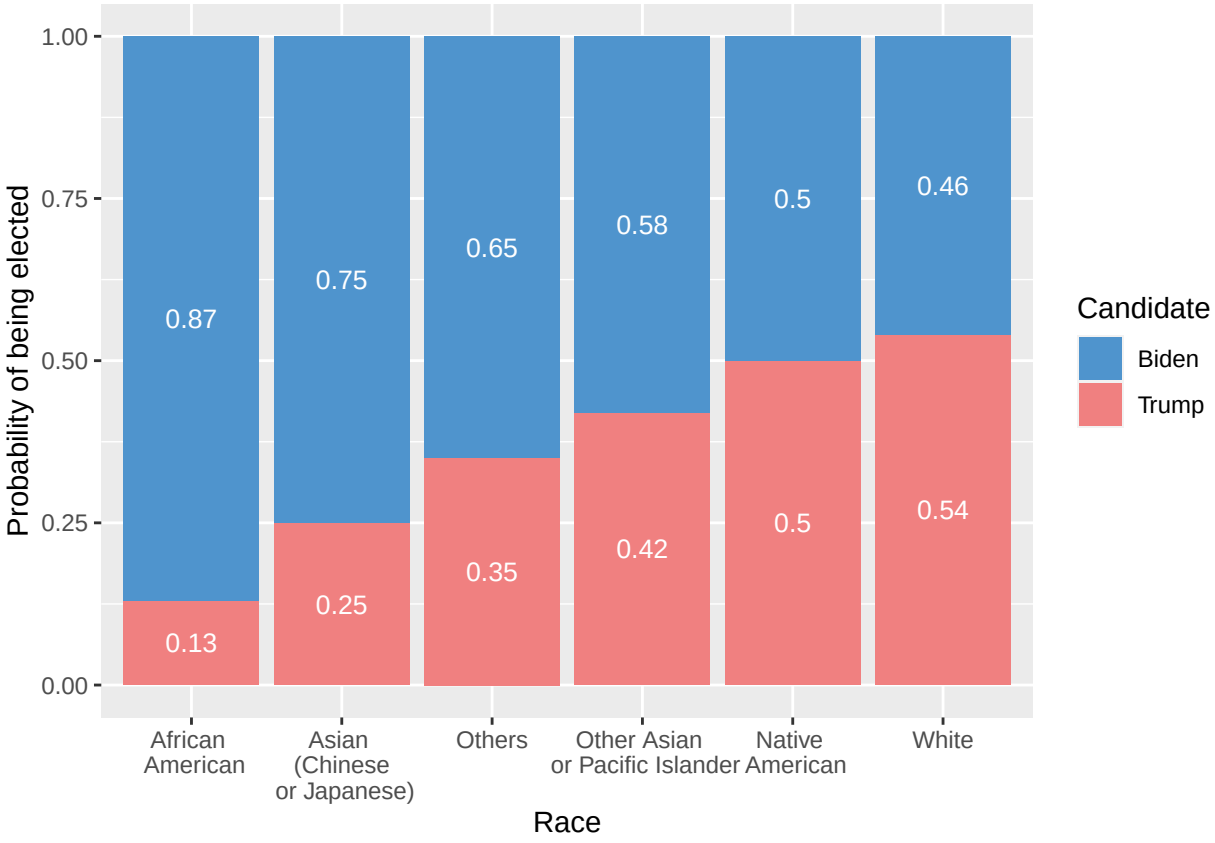


Figure 15: Probability of vote by race

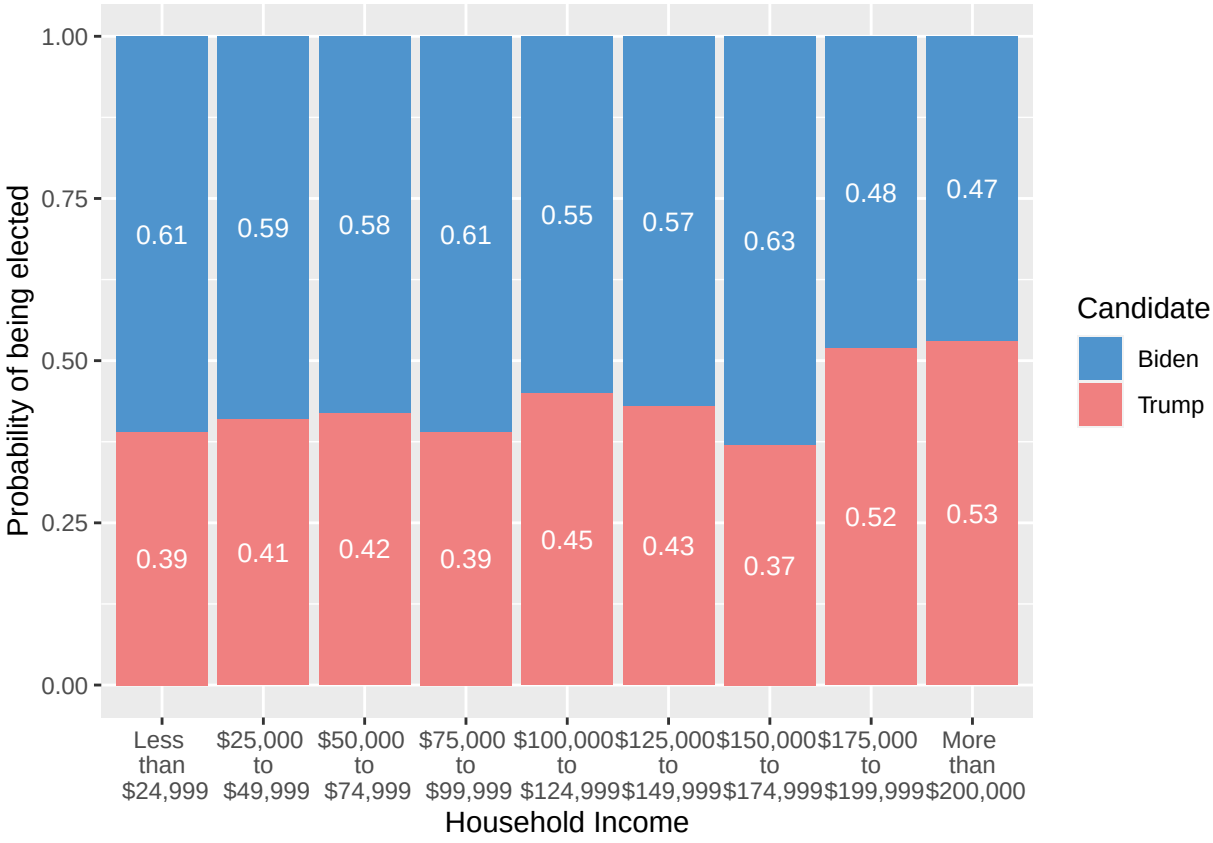


Figure 16: Probability of vote by income

6 Discussion

From the model, we are able to predict that biden is likely to win the popular vote by 16.32%. Figure 10 gives us a general understanding of the popularity of two candidates. It demonstrates that in every 100 people, 58 people are likely biased towards voting for Biden.

In Figure 11 and Figure 12, we have further demonstrated the predicted result of popular vote. Comparing the result in Figure 11 to Figure @red(fig:state), we have found that Iowa is the only state that changed its candidate after modelling. In the raw data, Trump has the popular vote in the state of Iowa, however, Biden has replaced the position. This gives us more confidence in Biden's popularity. Figure 12 is an important graph, serving its significance by providing us the details of the winning margin. While red is represented by Trump's Republican Party, and blue is Biden and his Democratic Party, the graph indicates how much of the popular vote they are likely to win in each state. For example, Vermont shows the darkest blue, meaning it is confident that Biden is expected to receive the popular vote in Vermont. And Idaho shows the brightest red, indicating the probability of Trump winning the popular vote in Idaho is higher than other stats.

With regards to the impact of the variables, it comes with little surprise that race(Figure 15) has a significant impact on one's choice of presidential candidate. Certain races such as African Americans and Asians (Chinese or Japanese) demonstrate very strong preferences towards Joe Biden. With the multitude of incidents that have occured over the current American president's term with relation to race, it is fitting to assume that many people of certain races might hold stronger feelings of dissatisfaction with Trump. Similar incidents have also occurred around gender(Figure 14) which makes the result where women show a stronger preference towards Biden unsurprising.

Moreover, another unsurprising outcome was that younger people have a stronger preference for Biden with regards to older age groups. Many of the younger generation are concerned with issues such as climate change, and racial and economic injustice (@cite_youth_poll). Many of the outcomes reflect the current sentiments of the public surrounding the current political climate and POTUS. A surprising outcome however, our predictions showed that income did not have a large impact on one's choice of presidential candidate despite their different approaches towards income tax being a subject of much debate. Our final model predicted that Biden wins with 58.16% of the votes.

By understanding the model and results, we can infer that a person's demographic information can have a significant impact on their preference of presidential candidate. This could be the result of certain policies being more beneficial towards certain demographics. Moreover, we can also reasonably draw the conclusion that comments made by politicians who receive much attention are remembered by the public. Their remarks, regardless of negative or not, are interpreted as indications towards their integrity and highly unlikely to be forgotten.

However, a caveat worth mentioning is that there was little consideration towards the electoral colleges, which is a large part of the American election. Moreover, it should be kept in mind that our model and predictions are very much contained in a small world. There can be many surprises in the large world of reality, where many surprising events can occur. This is further emphasized by how one's choice of presidential candidate is influenced by but not entirely determined by just a few demographic factors. There are many variances in how one perceives certain political policies such as the American government's methods of handling of the second wave of Covid19, which occurred months after the survey data was collected.

6.1 Weaknesses and Future Work

This paper suggests that Biden is expected to win by popular vote. However, some factors may take account to weaken the result. First of all, some limitations arise from the survey results. Our individual level survey from Democracy Fund + UCLA Nationscape uses an online portal to conduct non-probability quota surveys. Which may result in selection bias and non-response bias. Secondly, we have used the 2018 American Community Survey as post-stratification data. The survey is conducted every two years, meaning we have no access to the most recent data. Some factors such income, or political stance might have changed given

recent circumstances. Thus 2020 ACS would provide more accurate information if used. Upon modelling, the data has shown some more limitations. We have chosen to use a binary logistic regression with age, gender, income, state, and race as variables. However, due to the nature of two datasets, we are forced upon dropping or changing some valuable observations/variables. For example, racial groups are significantly different in the two datasets. We have to separate Chinese and Japanese, while other Asians are grouped together with Pacific Islanders to match the names. This may result in some inaccuracy of vote intentions by race. Moreover, ACS data does not include some factors such as 2016 vote history, which could potentially be a strong predictor of the 2020 election. In addition, a Bayesian binary regression model was our first choice of modeling, which may provide a more accurate result. However, due to excessively long run time and multiple crashes, we are forced to change it into logistic regression due to computational limitation. Another major limitation that directly impacts our result is that the result does not take accountability of people who would not vote. Our result simply demonstrates that among all registered votes, Biden's is likely to be 58.16% of them. If a significant number of people choose not to vote in this election, the difference between Biden and Trump's expected vote will be lessened. Despite the weakness mentioned above, this paper serves its purpose of predicting the election by popular vote. We are able to predict our result with a 95% confidence interval. Some future studies may be done with the addition of Bayesian to improve the accuracy. In addition, this paper can be reproduced as a forecast for other elections such as provincial elections.

7 Appendix

1. Simple ranking technique is to weight the criteria by rank in ascending/descending order. Ascending order means that the most important criteria is giving rank 1, vise versa. Upon assigning the ranks, the attributes are then weighted by the numerical weights corresponding to the ranks(Alliance [2016]). Simple ranking techniques are the more cost/time efficient way of ranking, yet drawbacks exist and are discussed.

References

American Community Survey Information Guide, 2017. URL https://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf.

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2020. URL https://github.com/rstudio/rmarkdown. R package version 2.5.

Geographic Information Technology Training Alliance. *Suitability analyis: Determining weights: Weighting by ranking*, 2016. URL http://www.gitta.info/Suitability/en/html/Normalisatio_learningObject1.html.

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. R package version 2.3.

A. Coppock and D. P. Green. *"What Can Be Learned From 500,000 Online Survey Responses About Party Identification?" American National Election Studies white paper*, 2016. URL https://static.lucidhq.com/ANES_White_Paper.pdf.

Paolo Di Lorenzo. *usmap: US Maps Including Alaska and Hawaii*, 2020. URL https://usmap.dev. R package version 0.5.1.

Andrew Gelman, Jeffrey Lax, Justin Phillips, Jonah Gabry, and Robert Trangucci. *Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion*, 2018. URL http://www.stat.columbia.edu/~gelman/research/unpublished/MRT(1).pdf.

Lucid HQ. *Sample Sourcing FAQs*, 2018. URL https://support.lucidhq.com/s/article/Sample-Sourcing-FAQs.

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/.

Bob Rudis and Dave Gandy. *waffle: Create Waffle Chart Visualizations in R*, 2017. URL https://github.com/hrbrmstr/waffle/tree/cran. R package version 0.7.0.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. *IPUMS USA: Version 10.0 [American Community Census 2018]*, 2018. URL https://doi.org/10.18128/D010.V10.0.

Dan Simpson. *Multilevel structured (regression) and post-stratification*, 2019. URL https://statmodeling.stat.columbia.edu/2019/08/22/multilevel-structured-regression-and-post-stratification/.

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

CHRIS TAUSANOVITCH, LYNN VAVRECK, TYLER RENY, ALEX ROSSELL HAYES, and AARON RUDKIN. *Democracy Fund + UCLA Nationscape Methodology and Representativeness Assessment*, 2019. URL https://www.voterstudygroup.org/uploads/reports/Data/NS-Methodolog-Representativeness-Assessment.pdf.

Alec Tyson and Shiva Maniam. *Behind Trump's victory: Divisions by race, gender and education*, 2020. URL https://www.pewresearch.org/fact-tank/2016/11/09/behind-trumps-victory-divisions-by-race-gender-education/.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham. *tidyr: Tidy Messy Data*, 2020. https://tidyr.tidyverse.org, https://github.com/tidyverse/tidyr.

Hadley Wickham and Evan Miller. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*, 2020. http://haven.tidyverse.org, https://github.com/tidyverse/haven, https://github.com/WizardMac/ReadStat.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2020. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Yihui Xie. knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. URL http://www.crcpress.com/product/isbn/9781466561595. ISBN 978-1466561595.

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL https://yihui.org/knitr/. ISBN 978-1498716963.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2020. URL https://yihui.org/knitr/. R package version 1.30.

Yihui Xie, J.J. Allaire, and Garrett Grolemund. *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida, 2018. URL https://bookdown.org/yihui/rmarkdown. ISBN 9781138359338.

Yihui Xie, Christophe Dervieux, and Emily Riederer. *R Markdown Cookbook*. Chapman and Hall/CRC, Boca Raton, Florida, 2020. URL https://bookdown.org/yihui/rmarkdown-cookbook. ISBN 9780367563837.

Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2020. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.