A Prediction of the 2020 US Election

Shuyu Duan,Fanxi Zhou, Feixue Han, Zhiang Chen

2020-11-2

Abstract

The 2020 United States presidential election is scheduled for November 3, 2020. The result of it would impose a substantial influence on the USA and the world's economy as both presidential candidates have opposing views on the matters of pandemic and economic recovery plan. This study aims to predict the outcomes by using data from Democracy Fund and UCLA Nationscape dataset to build a binary logistic model. The findings include that the males are more likely to vote for Trump than the females, whereas the wealthy are more likely to vote for Biden than the poor. This study is significant since both candidates have to pay attention to their potential voters' needs in order to get elected in office for future terms.

Introduction

The 2020 United States presidential election is scheduled for Tuesday, November 3, 2020. It will be the 59th presidential election for the USA. The US political system is dominated by just two parties, the Democratic Party and the Republican Party. The two major candidates and parties are Republican incumbent President Donald Trump and Democratic former Vice President Joe Biden. The US president has a huge influence on the lives of people who are domestic and abroad. In the US, US citizens who are over 18 years old have the right to vote for the president. Recently, according to the poll prediction of CBC, the overall supporting rate of Biden is higher than Trump. Biden has a high supported rate in some states such as WA, OR,CA and so on. When the next election is held on 3 November, everyone will know the outcome.

In this survey, we analyzed subsets of survey and census which is part of ACE of UCLA and Nationscape Data Set. The American Community Survey (ACS) is a nationwide survey which provides communities with reliable and timely social, economic, housing, and demographic data. In the Census Bureau, at least two questionnaires were used to collect the census data. The form collected basic demographic information and detailed housing and socioeconomic information. The det

Data

Loading Data

Data Clean

```
## Warning: Unknown or uninitialised column: `sex`.

## Warning: Unknown or uninitialised column: `race`.

## Warning: Unknown or uninitialised column: `labforce`.

## Warning: Unknown or uninitialised column: `education`.

##      0%     25%     50%     75%    100%
##       1   13600   31000   64000 1423000

## Warning: Unknown or uninitialised column: `income_level`.
```

```
## Warning: Unknown or uninitialised column: `household_income_bottom`.
##      0%     25%     50%     75%    100%
##       0   20000   45000   95000 250000
## Warning: Unknown or uninitialised column: `income_level`.
```

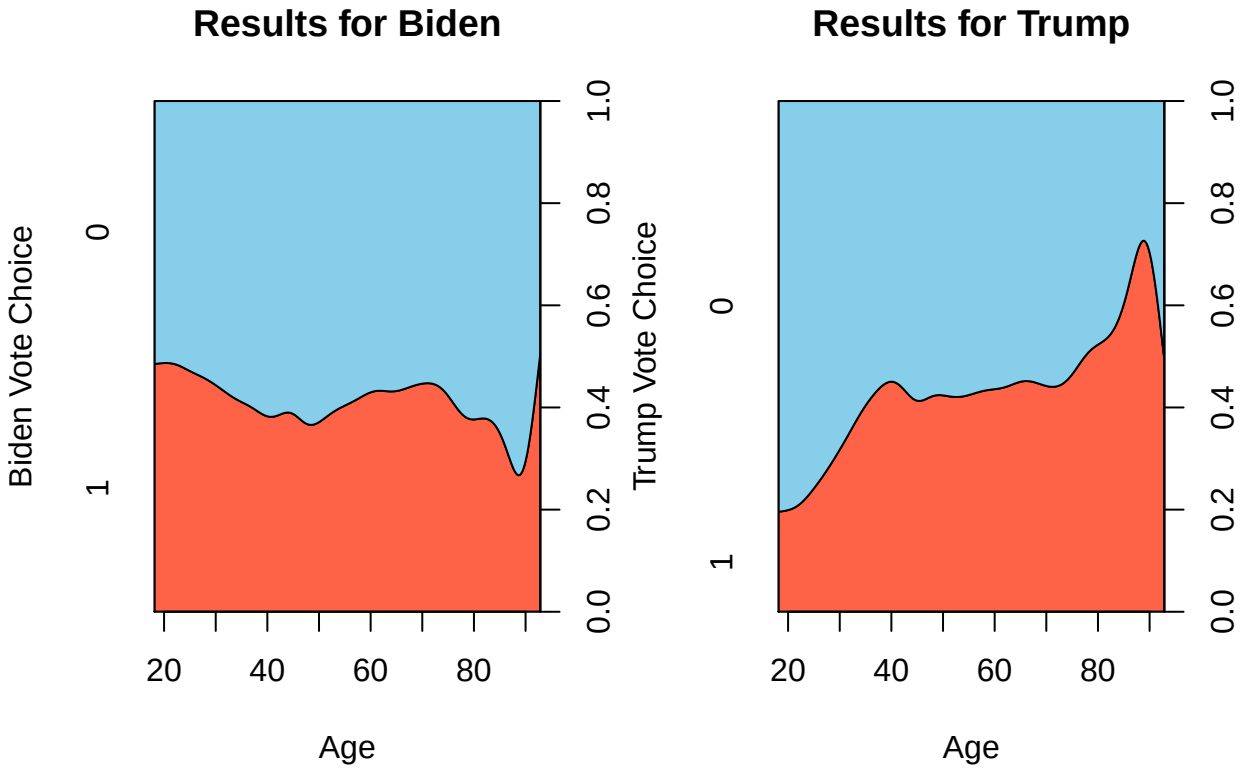
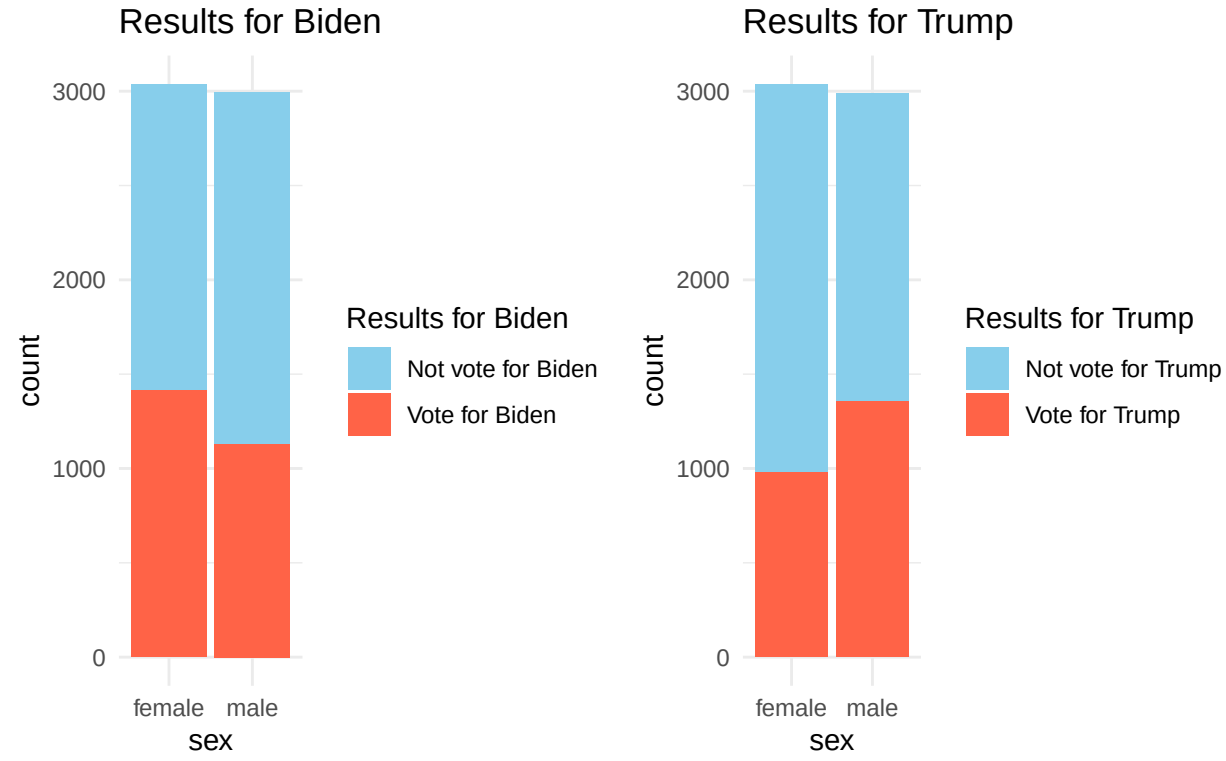


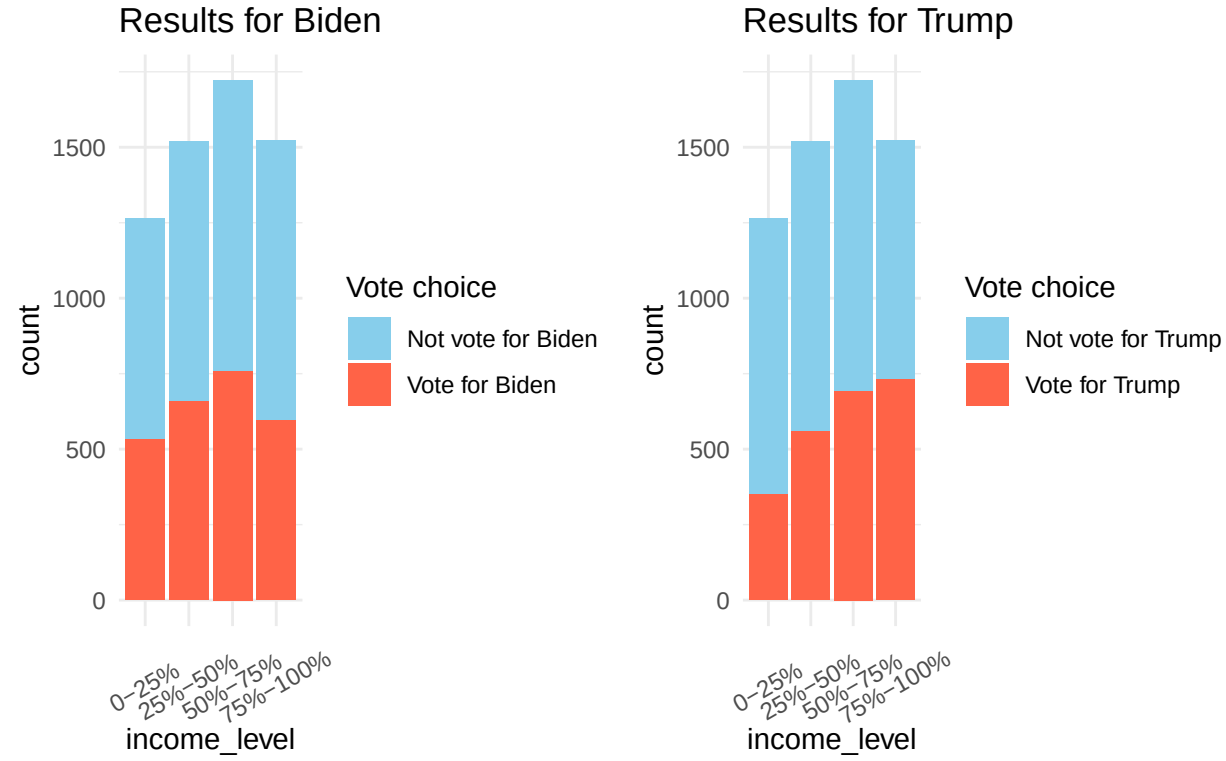
Figure 1: Age Distribution of Supporters

Figure 2: Boxplot of Sex and Voting



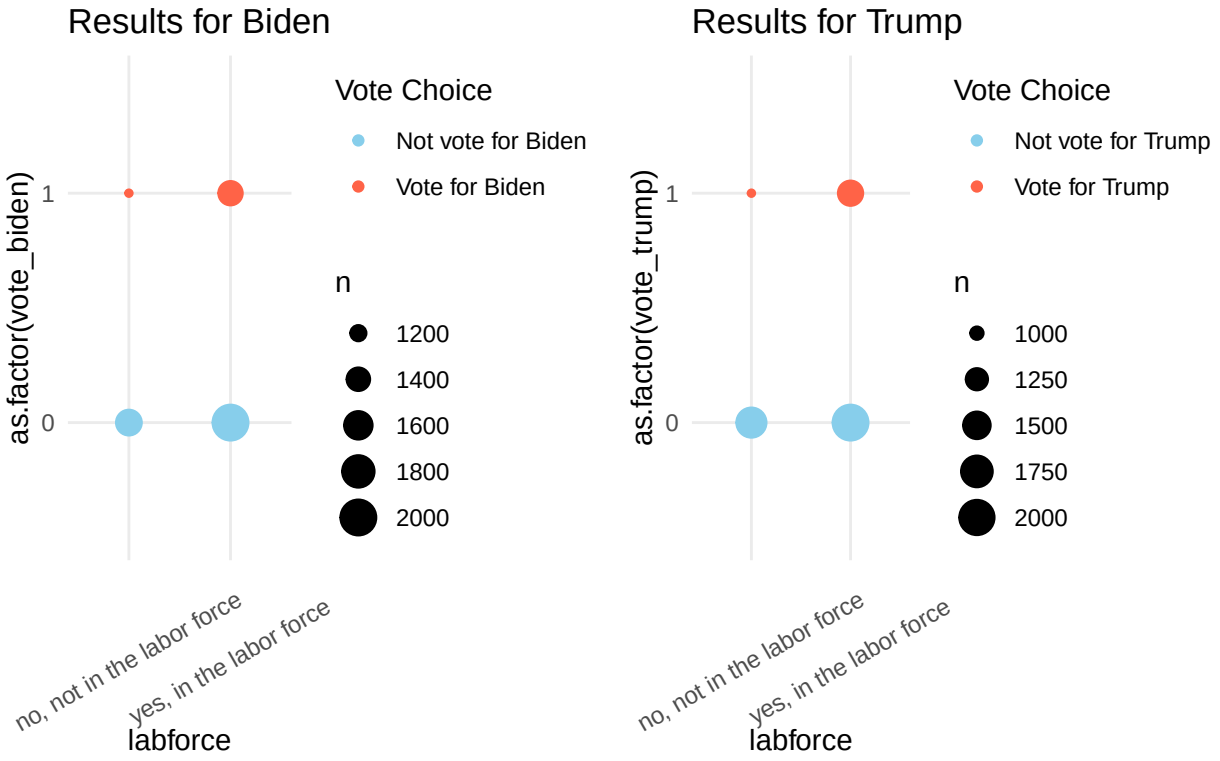
/ Fund + UCLA Nationscape Source: Democracy Fund + UCLA Nationscape

Figure 3: Boxplot of Income and Voting



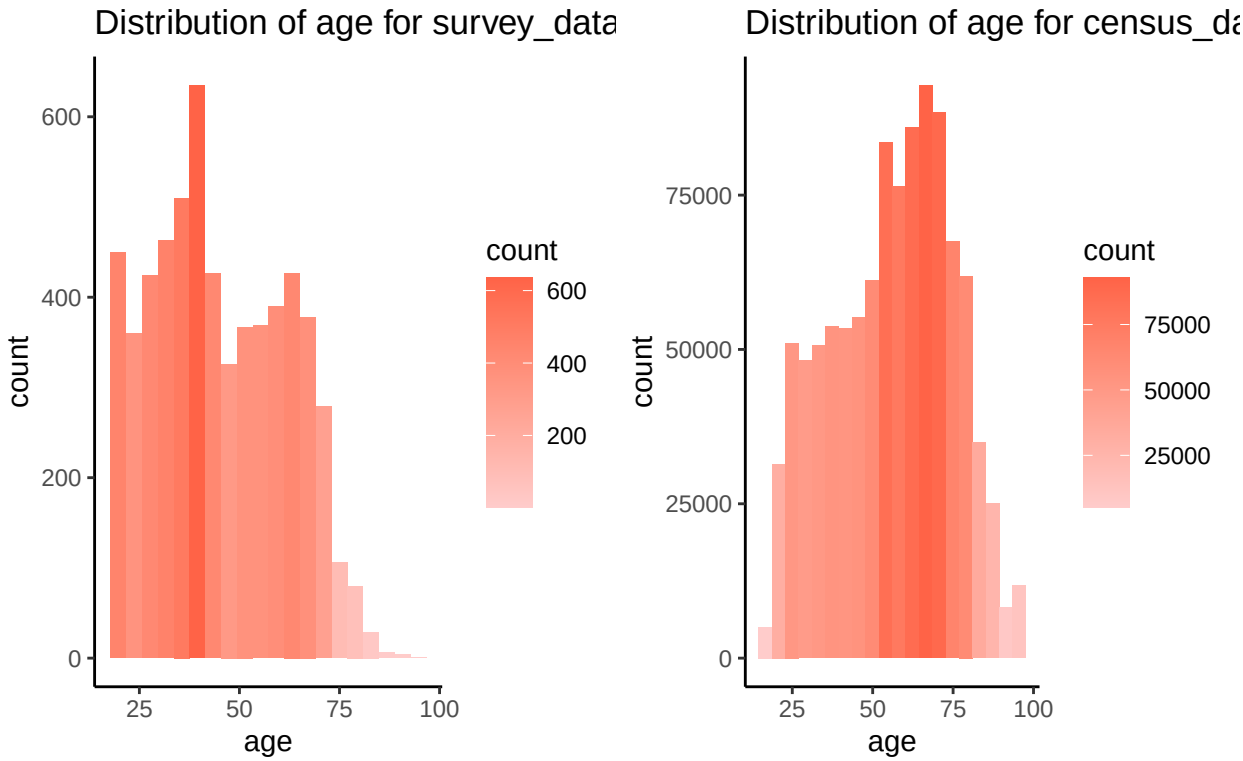
/ Fund + UCLA Nationscape Source: Democracy Fund + UCLA Nationscape

Figure 4: Mapping of Labforce and Voting



/ Fund + UCLA Nationscape Source: Democracy Fund + UCLA Nationscape

Figure 5: Distribution of Age



ce: Democracy Fund + UCLA Nationscape Source: Democracy Fund + UCLA Nationscape

The data:survey_data that we use to build the model is from the Democracy Fund + UCLA Nationscape dataset released in August 2020. (Tausanovitch, Chris and Lynn Vavreck, 2020) This data is collected by LUCID, Inc which is the partner of Nationscape personnel. And they interviewed people from almost every congressional district, and mid-sized U.S. cities across the country to collect 318,697 observations. And we choose the 6479 observations' data collected on June 25 2020 to build the model.

And the data:census_data that we use the built model to predict the election results is from the American Community Survey data.(Erica Gardner, Tomas kimpel, 2015) This data is collected by Census Bureau in April 2020, they collected randomly sampled households' data across the country. In our study, we use the 1,046,401 observations' data from this dataset as the input of our model to predict the election results. For the survey_data datasets, the observations that are lack key features that we need in our model have been deleted since it might reduce the accuracy of the model, however, we should notice it also might create bias. But generally speaking, after the data clean process, the number of observations in the survey_data reduce from 6479 to 6030. And for the census_data, non-response does not exist in the dataset.

In our model, we choose 6 predictors to predict the election results which is a binary variable, age, sex, whether or not in the labor force, race, education background, and income level.(citation survey dataset)(citation cenusu dataset) Age is numeric data we can directly get from the two data sets. From the Figure 5 the distribution of age we can see that the age in the survey_data is a right skew histogram which means the number of aged people's data is small. However, the age in the census_data is more likely normally distributed centered at the age of around 65. Sex is binary data, in the data cleaning process, we maintain the consistency from both datasets. Labforce is also binary data, we divide the labforce data from the survey_data into binary data to match the census_data. The race is a categorical variable, the race has been classified more specifically in survey_data, thus we regroup the race from the survey_data to match the census_data. Education is also a categorical variable but with the opposite situation, we change the census_data to match the survey_data since the first one has been divided into more detail. Finally, the income level is a categorical variable that we create from the original data. For the survey_data, we divide their income into 4 parts according to the bottom line of their household income since the categorical data gives a range, the richest 25%, above the average but not the richest, below the average but not the lowest, and the lowest 25%. And the census_data we directly divide them into 4 parts by the quantiles.

From the Figure 1, we can see that most of the people who are under 30 will choose not to vote for Trump, and there exists an extremely high support rate for Trump and will not vote for Biden for the people which is around 90. From the Figure 2, we can see that Biden is more popular in women and Trump is more popular in men, meanwhile the intersexual difference for Trump is much higher than for Biden. From the Figure 3, we can see that as people's income increase, they are more likely to support Trump, and Trump's approval rate among people with the lowest 25% income is rather low. The supporter for Biden is more evenly distributed in different income level, but people with high income gives a lower approval rate. From the Figure 4, we can see that people who are not in the labor force have a high probability that they will neither vote for Biden nor vote for Trump. This can be explained that people who do not have a job have a relatively small intention to participate in politics.

Model

The purpose of our study is to predict the vote outcome of the 2020 American federal election in ACS dataset(include citation). We used multilevel regression and post-stratification technique for this analysis. In the following sub-sections I will describe the model specifics and the calculation for the post-stratification process.

Model Specifics

In the beginning, we tried to use the linear regression model for this analysis. We made the model for predicting the proportion of voters who will vote for Donald Trump and the proportion of voters who will vote for Joe Biden separately. Since we have two models, the response variables for them is vote Biden or vote trump(1 for vote Biden or vote trump, 0 for not vote for them). For both models, we used the same

six predictors: age, sex, race, education, labor force, income level to model the probability of voting for Donald Trump, and the probability of voting for Joe Biden. The age is a numeric variable and the other five predictors are categorical variables. However, we found that our response variable is binary which means a binary logistic regression model may be better. So we made two binary logistic linear regression models as well. We decided to use the Akaike information criterion (AIC) for testing which model is better for fitting our data. AIC uses the number of independent variables that are used to make a model and the maximum likelihood estimate of the model to get a value(cite). Here, the maximum likelihood estimate tells us how well the model reproduces the data. By comparing the AIC for the linear regression model and binary logistic regression model, we observed that the AIC for our two linear regression models are 8263 and 7843 and the AIC for two binary logistic regression models are 7873 and 7406. The smaller AIC value tells that the model fits the data better. Thus, By comparing the AIC value, the binary logistic regression model would be a better model for our study. Then, we used R(cite) to run two binary logistic regression models to model the proportion of voters who will vote for Joe Biden and the proportion of who will vote for Donald Trump separately based on the Nationscape Dataset(cite). The equation below is the binary logistic regression model:

$$\pi_i = Pr(Y_i = 1|X_i = x_i) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$$

or

$$logit(\pi_i) = log(\frac{\pi_i}{1 - \pi_i}) = \beta_0 + \beta_1 x_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_q x_{iq}$$

We assume that Y_i is a binary response variable for i = 1,... n and takes on value 0 or 1 with P(Y_i = 1) = pi_i. Suppose X is a set of explanatory variables, x_i is the observed value of the explanatory variables for observation i = 1,... q. From the above formula, we can also get:

$$\frac{\pi}{1 - \pi} = e^{\beta_0} e^{\beta_1 x_1} ... e^{\beta_q x_q}$$

Then the β_0 is the baseline odds and β_1 can be interpreted as holding predictors constant, a one-unit increase in x_1 increases the probability of voting for Donald Trump or Joe Biden by a factor of e^{β_1}.

```
## [1] 8261.323
```

Post-Stratification

Multilevel regression and post-stratification (MRP) combines two statistical techniques to determine the relationship between the response variable of our interest and predictors we chose. Unlike the normal multilevel regression analysis, we add a post-stratification process base on the previous multilevel regression analysis. We used the sample data to train a regression model and then we would use this trained model to predict the outcome in the population dataset which would be a large population. The MRP requires the data to be demographic. In our study, we chose six predictors which are mentioned in the previous section as the key demographic features of the sample. However, MRP also has some limitations. As we mentioned before, the MRP requires the data to be demographic. Also, if the sample data is not sufficient enough or the demographic predictors are not enough, the outcome would be biased and can even be failed. In our sample data Nationscape Dataset, we could find some key demographic features as our predictors, so we chose to use MRP for our analysis.

In the post-stratification process, in order to estimate the proportion of voters who will vote for Donald Trump and the proportion of voters who will vote for Joe Biden. We performed a post-stratification analysis on the ACS dataset(citation). We created many cells based on different age, race, sex, education, labor force, and income level. Performing the model described in the above section, we estimated the proportion of voters in each cell. Then, we calculate the proportion of voters estimate for each cell by using the respective population size of that cell and sum those values and divide that by the whole population.

```
## # A tibble: 1,046,401 x 8
##      age sex   race  labforce        education       income_level     n estimate
```

```
##    <dbl> <chr> <chr> <chr>           <chr>           <chr>        <dbl>    <dbl>
##  1    18 male  white no, not in the~ 3rd Grade or l~ 0-25%           1    0.397
##  2    18 male  white no, not in the~ 3rd Grade or l~ 0-25%           1    0.397
##  3    18 male  white no, not in the~ 3rd Grade or l~ 0-25%           1    0.397
##  4    18 male  white no, not in the~ 3rd Grade or l~ 25%-50%         1    0.423
##  5    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.276
##  6    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.276
##  7    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.276
##  8    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.276
##  9    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.276
## 10    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.276
## # ... with 1,046,391 more rows
```

```
## # A tibble: 1 x 1
##   win_prob
##      <dbl>
## 1    0.400
```

```
## # A tibble: 1,046,401 x 8
##      age sex   race  labforce        education       income_level     n estimate
##    <dbl> <chr> <chr> <chr>           <chr>           <chr>        <dbl>    <dbl>
##  1    18 male  white no, not in the~ 3rd Grade or l~ 0-25%           1    0.429
##  2    18 male  white no, not in the~ 3rd Grade or l~ 0-25%           1    0.429
##  3    18 male  white no, not in the~ 3rd Grade or l~ 0-25%           1    0.429
##  4    18 male  white no, not in the~ 3rd Grade or l~ 25%-50%         1    0.485
##  5    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.283
##  6    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.283
##  7    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.283
##  8    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.283
##  9    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.283
## 10    18 male  white no, not in the~ Middle School ~ 0-25%           1    0.283
## # ... with 1,046,391 more rows
```

```
## # A tibble: 1 x 1
##   win_prob
##      <dbl>
## 1    0.423
```

Results

Table 1: Summary of Biden Model Results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.379	0.655	-0.579	0.563
age	-0.002	0.002	-0.936	0.350
sexmale	-0.305	0.056	-5.489	0.000
labforceyes, in the labor force	-0.138	0.061	-2.275	0.023
raceblack/african american/negro	1.591	0.255	6.244	0.000
racechinese	0.953	0.343	2.779	0.005
racejapanese	1.594	0.591	2.697	0.007
raceother asian or pacific islander	0.663	0.280	2.370	0.018
raceother race, nec	0.743	0.261	2.842	0.004
racewhite	0.299	0.245	1.220	0.223
educationAssociate Degree	0.001	0.626	0.002	0.998

	Estimate	Std. Error	z value	Pr(>|z|)
educationCollege Degree (such as B.A., B.S.)	0.094	0.622	0.151	0.880
educationCompleted some college, but no degree	-0.189	0.623	-0.304	0.761
educationCompleted some high school	-0.599	0.626	-0.957	0.339
educationDoctorate degree	-0.194	0.646	-0.300	0.764
educationHigh school graduate	-0.500	0.623	-0.804	0.421
educationMasters degree	0.192	0.626	0.307	0.759
educationMiddle School - Grades 4 - 8	-0.546	0.765	-0.713	0.476
income_level25%-50%	0.105	0.083	1.278	0.201
income_level50%-75%	0.103	0.084	1.221	0.222
income_level75%-100%	-0.134	0.094	-1.428	0.153

Table 2: Summary of Trump Model Results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.040	0.667	-1.558	0.119
age	0.015	0.002	7.693	0.000
sexmale	0.376	0.057	6.542	0.000
labforceyes, in the labor force	0.210	0.064	3.301	0.001
raceblack/african american/negro	-1.820	0.264	-6.891	0.000
racechinese	-1.308	0.396	-3.302	0.001
racejapanese	-0.838	0.624	-1.343	0.179
raceother asian or pacific islander	-0.455	0.278	-1.635	0.102
raceother race, nec	-0.609	0.258	-2.364	0.018
racewhite	0.114	0.233	0.488	0.626
educationAssociate Degree	-0.757	0.643	-1.178	0.239
educationCollege Degree (such as B.A., B.S.)	-0.738	0.638	-1.155	0.248
educationCompleted some college, but no degree	-0.604	0.639	-0.945	0.345
educationCompleted some high school	-0.331	0.642	-0.515	0.606
educationDoctorate degree	-0.301	0.662	-0.455	0.649
educationHigh school graduate	-0.438	0.638	-0.686	0.493
educationMasters degree	-0.653	0.642	-1.016	0.309
educationMiddle School - Grades 4 - 8	-0.643	0.792	-0.812	0.417
income_level25%-50%	0.224	0.089	2.530	0.011
income_level50%-75%	0.329	0.090	3.676	0.000
income_level75%-100%	0.599	0.098	6.096	0.000

Table 3: Summary of prediction of vote outcome

	Probability
vote for Biden	0.400
Vote for Trump	0.423

Using a binary logistic model, we got two summary statistics tables and the probability of voting for Biden and Trump. The data "vote for Biden" was summarized in the first table. In the Pr category, the EducationAssociate degree has the highest value which is 0.998 while age, education level and so on have relatively large p-value. Only the P-value for the sexmale, labor force and races category are lower than 0.05. The std.error values are all lower than 0.65 except for the std.error for educationMiddle School category. The maximum z value is 6.244 which is from the raceBlack category while the minimum is -5.489 which is from

the SexMale category. The coefficients for the categories are half positive and half negative. The maximum value is 1.594 which is from the race Japanese category. The minimum value for the coefficient is -0.599 which is from the EducationCompleted some high school category. The data for "vote for Trump" was summarized in the second table. In the Pr category, the Education Doctorate degree has the highest value which is 0.649 while education level and so on have relatively large p-value. Only the P-value for the age, sexmale, labor force and so on are lower than 0.05. The std.error values are all lower than 0.65 except for the std.error for educationMiddle School category. The maximum z value is 7.693 which is from the Age category while the minimum is -6.891 which is from the raceBlack category. The coefficients for the categories are mostly negative. The maximum coefficient is 0.599 which is from the income75%-100% category. The minimum value for the coefficient is -1.820 which is from the raceBlack category. Finally, The probability we got for "vote for Biden" is 0.400 and the probability we got for "vote for Trump" is 0.423.

Discussion

In this model, the p-value helps us to test the null hypothesis so that we can indicate whether the factors have a correlation with our predictor in the whole population. The null hypothesis for our model is that there is no correlation with our response variable. If the p-value is smaller than 0.05, it rejects the null hypothesis so there may be a correlation between that factor and our response variable. The smaller the p-value, the stronger evidence for rejecting the null hypothesis. However, if the p-value is greater than 0.05, it supports the null hypothesis, which means that there may not be a correlation between the factor and our response variable. In this survey, the p-value of education level and white is over 0.05 so that it is not significant. The p-value of age in the Trump model result is less than 0.05 while the one in Biden is more than 0.05. It shows that age has a significant effect on Trump and not significant effect on Biden. Each one-unit change in age of the Trump model will increase the log odds of getting admitted by 0.0015. In the sexmale catagory, the p-value of Trump and Biden model are less than 0.05 which has significant effect on both of them.

Each unit change in sexmale in the Trump model increases the log odds of getting admitted by 0.376 while each unit decrease in sexmale in the Biden model decreases the log odds of getting admitted by -0.305. It shows that females are likely to vote for Biden while male is likely to vote for Trump. In the Chinese and nergo category, the p-value of Trump and Biden model are less than 0.05 which has significant effect on both of them. Each unit change in Chinese and nergo category in the Trump model decreases the log odds of getting admitted by -1.308 and -1.820 correspondingly while each unit change in Chinese and nergo in the Biden model increases the log odds of getting admitted by 0.953 and 1.591. It shows that Chinese and nergo are more likely to vote for biden. For the labforce category, the p-value of Trump and Biden model are less than 0.05 which has significant effect on both of them. Each unit change in labforce category in the Trump model increases the log odds of getting admitted by 0.210 while each unit change in labforce category in the Biden model decreases the log odds of getting admitted by -0.138.

Weaknesses

- Some weaknesses exist in the data analysis part. While we are cleaning the data, the na which refers to non- response questions was deleted so that bias would exist in the result. In the race section, the " three or more major races" and "two major races" were deleted. While analysing the labforce, there are several original options such as Full- time employed, retired, students and so on. However, only two options which are "yes, in the labor force" and "no, not in the labor force" were used because we used binary models to predict the result. The predicted result could be biased. In the income_level, the personal income and household income are converted from numeric to percentile("0%-25%", "25%-50%", "50%-75%", "75%-100%"). When personal income and household income are directly compared, the result could be biased.
- The data set is lacking some demographic predictor, thus some important parts of the model were misspecified. For example, the data are not divided by states. Since there are 56 states in the US, data will be differ from states. If this part of data is not included, bias will exist. Also, the variable marriage is not mentioned in the survey data. Marriage is a possible variable to predict the winner of the president election. The data set was recorded in June 2020 and it has been a long time until now.

-In this survey, 6000 questionnaire samples were used to predict the result of the election. However, the overall sample data is more than 1000000, the insufficient data may result in the bias.

Next Steps

In the future, the model and data need to be improved. The marriage status needs to be added to the data set. Also, the data from each state should be added to reduce bias. Only one day of survey data which is on June 25th in 2020 was used to predict the result of the election. More data needs to be selected so that the result could be more accurate. Furthermore, more models can be used to predict the result such as mixed effects logistic regression, Bayesian Generalized and non- Linear Multivariate Multilevel Model. We can compare the results from different models to find the more possible result.

Appendix

Code and data supporting this analysis is available at: "https://github.com/zhoufanx/STA304-problem-set4".

References

[1] Choi, Matthew (October 31, 2019). "Trump, a symbol of New York, is officially a Floridian now". Politico. Retrieved October 31, 2019. [2] "3 U.S.C. § 7 – U.S. Code – Unannotated Title 3. The President § 7. Meeting and vote of electors", FindLaw.com [3] Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). https://www.voterstudygroup.org/downloads?key=8c1f266c-976d-493d-a6d1-4ee560f83785. [4] Erica Gardner, Tomas kimpel. 2015. American communication survey. https://www.census.gov/programs-surveys/acs

Yihui Xie (2014) 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. ISBN 978-1466561595

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

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

Zhu, Hao. 2020. KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax. https://CRAN. Rproject. org/package=kableExtra. 17