Biden Expected to Win the 2020 US Popular Vote

Colin Conant, Sara Hamdy, Jack Smith

November 2, 2020

GitHub Repo Link: https://github.com/Saraahamdy/STA304H1-PS3-Group-70.git

Model

Here we are interested in predicting the popular vote outcome of the 2020 American federal election (Silver, 2016). To do this we are employing a post-stratification technique. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

Model Specifics

To best analyze this data, we used a frequentist approach, developing two binary multiple logistic regression models using R. These models were based on the Democracy Fund + UCLA Nationscape Date Set and were created to explore the popular vote for the American 2020 national election (Tausanovitch & Vavreck, 2020). The first model explores the proportion of voters who will vote for Trump, while the second model explores the proportion of voters who will vote for Joe Biden. These two multivariable logistic models are appropriate as all assumptions required for a logistic model are fulfilled: a binary response variable, independent data, the predictor variables are uncorrelated with each other, and the model has a linear relationship with the logit function and the variables (Equation 1). Multiple logistic models are more appropriate to use than a linear model for our analysis because our response variable was binary.Each model has a single binary response variable that takes a value of 1 if a person intends to vote for the respective candidate, and a 0 for any other outcome, whether that be voting for another candidate, or undecided. Our models have the same three predictor variables. These variables are age, gender, and race. These variables were selected for the model as they are the underlying factors that affect social and personal beliefs, and therefore will influence voting. When evaluating the explanatory variables, race was reduced from 13 categories to 4 categories of similar races to gather better sample size representation for race groups with extremely few observations. As well, age was left as a continuous variable, and was not grouped by age ranges, in order to allow for more detailed data on demographics.

The logistic regression models used follow the equation:

Equation 1: Model Equation

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0+\beta_1 x_{age}+\beta_2 x_{male}+\beta_3 x_{Asian/PacificIslander}+\beta_4 x_{Black,orAfricanAmerican}+\beta_5 x_{SomeOtherRace}+\epsilon$$

Where $\hat{p}$ represents the proportion of voters who will vote for Donald Trump in the first model and Joe Biden in the second. The betas and intercept have the same interpretations with respect to both candidates. β_0 represents the intercept of the model, and represents the log odds of an 18 year old White female voting for the respective candidate. Additionally, β_1 represents the slope coefficient for the age predictor variable. So, for everyone one-unit increase in age, we expect a β_1 increase in the log odds of voting for the respective candidate. β_2 represents the slope coefficient for the gender predictor variable. The log odds change by β_2 when the voter is Male. β_3, β_4 and β_5 represent the slope coefficients for the race predictor variable. β_3 represents the slope coefficient and log odds change for an Asian/Pacific Islander voter. β_4 represents the slope coefficient and log odds change for a Black or African American voter. Finally β_5 represents the slope for Some other race. If the voter is any other race than White, Black or African American, or Asian/Pacific Islander, the log odds change by β_5. ϵ represents the error term in the function.

term	estimate	std.error	statistic	p.value
(Intercept)	-1.0250029	0.0896841	-11.429039	0
age	0.0126838	0.0016582	7.649137	0
as.factor(gender)Male	0.4455057	0.0538410	8.274475	0
as.factor(race_ethnicity)Asian/Pacific Islander	-0.8459145	0.1336238	-6.330566	0
as.factor(race_ethnicity)Black, or African American	-1.9931967	0.1304565	-15.278627	0
as.factor(race_ethnicity)Some other race	-0.6730536	0.1044641	-6.442915	0

The Trump model is as follows: Equation 2: Trump Model

$$log(\frac{\hat{p}}{1-\hat{p}}) = -0.8504 + 0.0107x_{age} + 0.4439x_{male} - 0.8816x_{Asian/PacificIslander} - 2.0282x_{Black,orAfricanAmerican}$$
$$-0.6452x_{SomeOtherRace} + \epsilon$$

term	estimate	std.error	statistic	p.value
(Intercept)	-0.4828032	0.0856325	-5.638081	0.0000000
age	0.0022534	0.0015937	1.413966	0.1573720
as.factor(gender)Male	-0.2917982	0.0519107	-5.621154	0.0000000
as.factor(race_ethnicity)Asian/Pacific Islander	0.6026399	0.1143061	5.272158	0.0000001
as.factor(race_ethnicity)Black, or African American	1.2012152	0.0836531	14.359490	0.0000000
as.factor(race_ethnicity)Some other race	0.2501206	0.0935041	2.674971	0.0074736

The Biden model is as follows: Equation 3: Biden Model

$$log(\frac{\hat{p}}{1-\hat{p}}) = -0.2840 - 0.0001x_{age} - 0.3214x_{male} + 0.7686x_{Asian/PacificIslander} + 1.3489x_{Black,orAfricanAmerican}$$
$$+0.3773x_{SomeOtherRace} + \epsilon$$

Post-Stratification

We used a post-stratification technique in our analysis, utilizing the models based on the Nationscape Data set, and a larger data set of the American voting population given by the American Community Surveys(ACS) (Ruggles et al., 2020)(Tausanovitch & Vavreck, 2020). In post-stratification, the population is divided into groups called cells, based on various factors, which allows us to make estimates on our topic of interest based on distinct groups of people, rather than the general population. Then with the estimates of specific groups of people, a larger data set can be used to weigh, or standardize, the groups according to a more realistic representation of the true population.

This process is important in our particular analysis, as it allows for a statistically sound way to 'extrapolate' our estimate of the popular vote to a larger, much more representative sample of the population, which will increase the accuracy of our prediction of the popular vote. In this analysis, our cell-level estimates are the proportions of each cell's population that will vote for the respective candidates. Cells were split on a personal level using age, gender, and race. Each of these variables are likely to influence voter outcome because they affect social, personal, and societal frames of reference. Therefore after the split, cells are created for every race, age, and gender combination in the model, for a total of 638 cells. By restricting the splitting process to only three variables, we ensured that every cell would have a sufficient sample size to make estimates on.

After using the logistic model to make estimates of voting proportions for every cell, we then used the ACS data set to reweigh our cells, by evaluating the proportion of each individual cell population compared to the total population of the census data set, which was 12,736,954 voting age U.S. citizens (Ruggles et al., 2020). This allowed us to make a prediction of the popular vote for both Biden and Trump with a larger and more representative sample of the population.

```
## [1] "From the Biden Model we conculed that 43.98% of votes are for Biden."
```

```
## [1] "From the Trump Model we concluded that 41.83% of votes are for Trump."
```

Results

After fitting our models, we derived two model equations, Equation 2 and Equation 3.

Equation 2:

$$log(\frac{\hat{p}}{1-\hat{p}}) = -0.8504 + 0.0107x_{age} + 0.4439x_{male} - 0.8816x_{Asian/PacificIslander} - 2.0282x_{Black,orAfricanAmerican} \\ -0.6452x_{SomeOtherRace} + \epsilon$$

Equation 3:

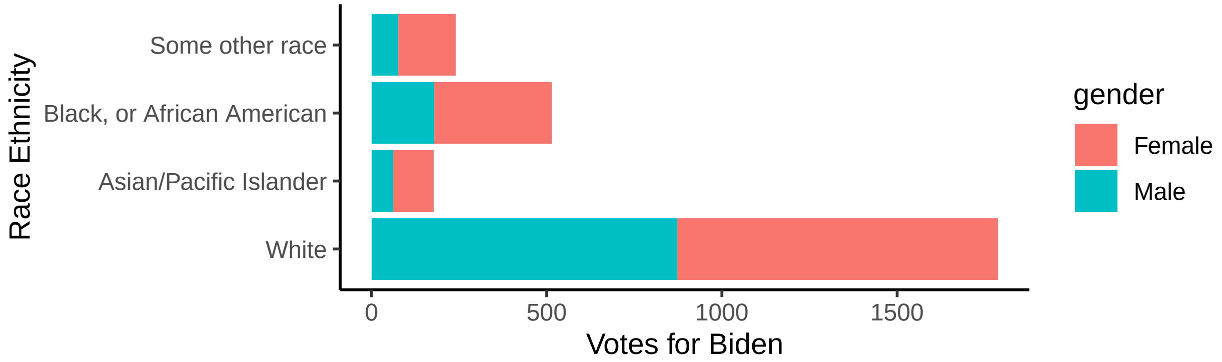
$$log(\frac{\hat{p}}{1-\hat{p}}) = -0.2840 - 0.0001x_{age} - 0.3214x_{male} + 0.7686x_{Asian/PacificIslander} + 1.3489x_{Black,orAfricanAmerican} \\ +0.3773x_{SomeOtherRace} + \epsilon$$

The models highlight that Biden receives more support from Black or African Americans, Asians and Pacific Islanders when compared to Trump. It also should be noted that men are more inclined to vote for Trump rather than Biden. Biden's β_1 is -0.0001 and has a high p-value of 0.95. This can be a result of the age distribution for Biden being multimodal (Fig. 2).

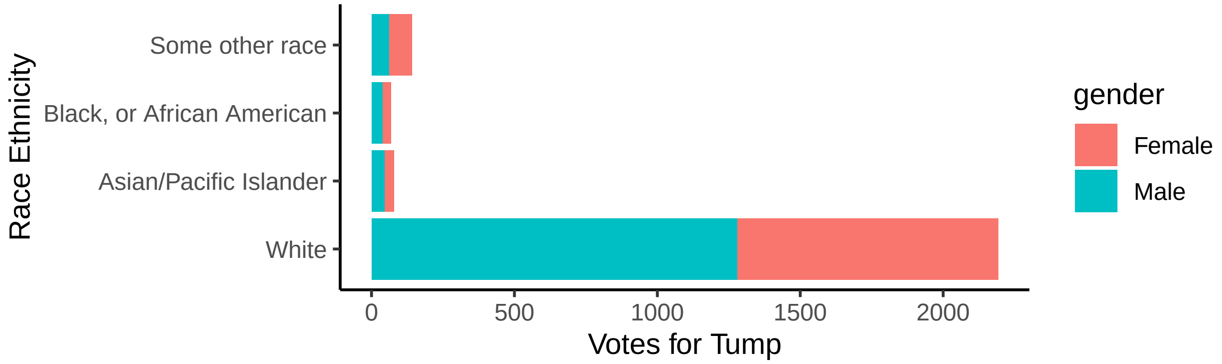
Based on our multiple logistic models of voting probability for Donald Trump and Joe Biden, in conjunction with a post stratification analysis on the proportion of voters intending to vote for Donald Trump and Joe Biden, we predicted that 41.83% of voters will vote for President Trump and 43.98% of voters will vote for Vice-President Biden. Our models and post stratification analysis accounted for a voter's age, gender, and race when determining the odds of voting for either of the candidates.

As we can see below, the majority of voters for Joe Biden as well as Donald Trump are White (Fig.1 and Fig.3). But, it should be noted that Biden is predicted to gain more votes than Trump from the other demographics, most substantially from Black or African American voters and women

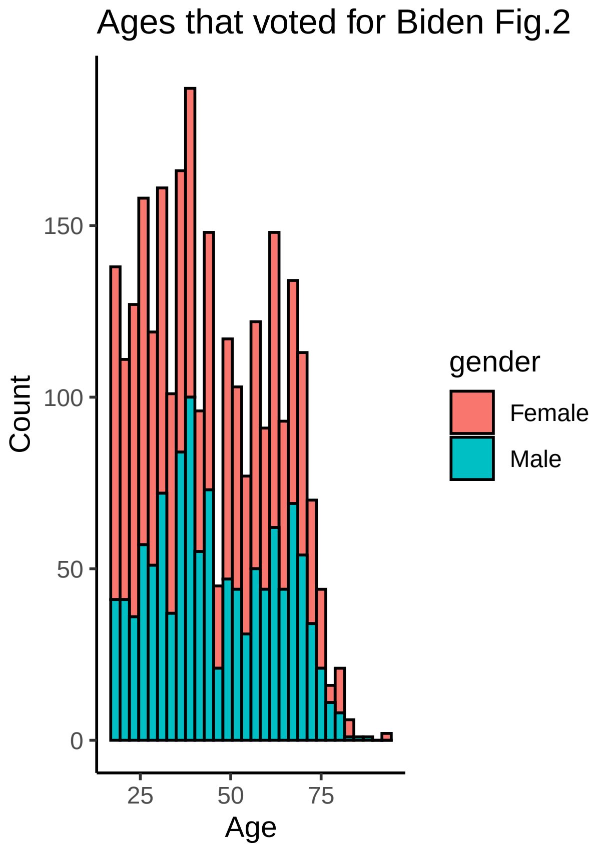
Races that voted for Biden Fig.1

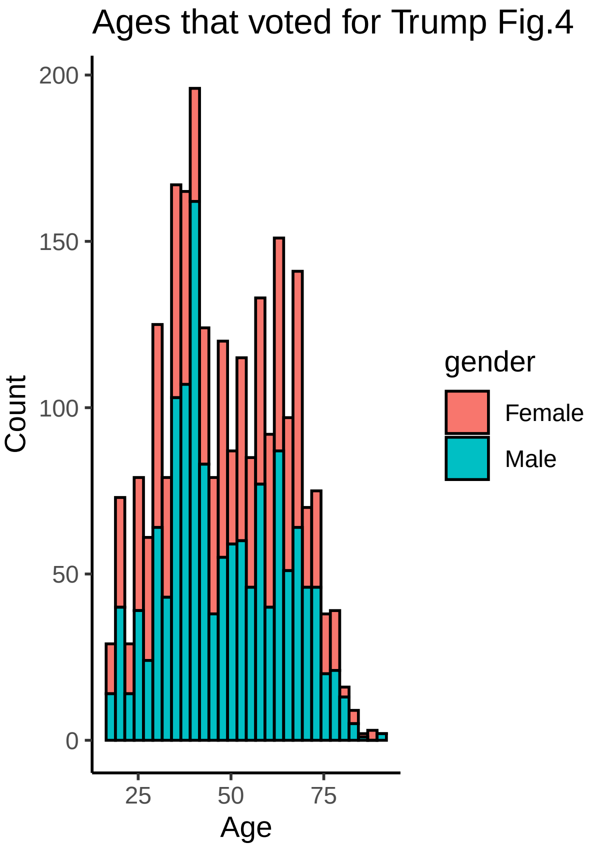


Races that voted for Trump Fig.3



As well, Biden receives more votes from younger demographics when compared to Trump (Fig.2 and Fig.4).





Discussion

Our model was designed to predict the winner of the 2020 US presidential popular vote. We used logistic regression models based on the Nationscape Data set to measure the proportion of voters who will vote for Trump and Biden using predictor variables age, race and gender (Tausanovitch & Vavreck, 2020) . To expand our results to the greater population of the United States, we used a post stratification technique with the ACS (Ruggles et al., 2020). In the post stratification stage, we separated the population into distinct groups based on the predictor variables in the logistic regression(gender/age/race). With these groups we were able to expand on our results and predict the winner of the popular vote.

Our analysis leads us to conclude that the winner of the 2020 US popular vote will be Joe Biden. Our Trump model predicts that Trump holds 41.83% of votes while our Biden model predicts that Biden holds 43.98% of votes. The rest of the voters are either undecided or have chosen not to vote. Further analysis after post-stratification concludes that overall, Biden has more votes from women, people of color, and younger voters - citizens in their twenties. Although Biden does not have the majority of White votes, his popularity among the other demographics makes up for his deficit in the White votes. This is what we believe to be the deciding factor in allowing Biden to win our predicted popular vote.

Weaknesses

A weakness in the model is that the popular vote is not always representative of the electoral college. We believe that Joe Biden will win the popular vote and this normally results in a win in the electoral college. Although we think Biden will win, there is a possibility that Biden does really well in large states like California and New York but loses the electoral college. In the 2016 American election, the predictions based on polls were incorrect by historic margins. States in which Hillary Clinton was predicted to win by large percentages of the votes ended up going to Donald Trump, showing there's always a chance the polls may not be representative of the population. It seemed like an all but guarantee that Trump would lose but pulled through in key states with by thin margins to win the presidency, many undecided voters ended up going to Trump and that might be the case here as well. In states like Wisconsin and Michigan pollsters had him losing by anywhere between 5 to 11 percent but he ended up winning (Silver, 2016). We found that 41.83% of decided voters will go Trump and 43.98% will go to Biden, since this is so close, it could indicate that undecided voters could swing the polls again to the underdog, Donald Trump.

Another weakness that could be improved upon in the Joe Biden logistic regression model is that the age variable is not a statistically significant predictor in this model. An extremely large p-value of ~0.95 in this case means that we can not conclude that Age is a factor in determining if someone votes for Biden. This is reflected in the beta 1 value for the Biden model, which has a value of -0.0001, having near negligible impact on the log odds and probability of the Biden model.

Next Steps

In our next steps we should look at the state level and see how the vote count in each state is expected to finish. Most US elections are won and lost in key battleground states so we should look to see how the votes shape up in battleground states as the popular vote does not always go to the winner of the election.

We should also compare our post stratification prediction to the actual election results. We predicted Biden would win the popular vote as we got 43.98% support for Biden and 41.83% support for Trump. We can check if our prediction was correct, or if our estimates based on the ACS survey were misleading. When comparing the results, we should do a post-hoc analysis to improve the accuracy of future models and to find specific differences between our model and the final results. Specifically in our post-hoc analysis, we would be interested in discrepancies between the voting distributions of our cell demographics and the true proportion of those same demographics. To obtain this data, we could conduct another survey in which we ask for who the respondent intended to vote for before the election and who they ended up voting for, as well as their age, gender and race. This would allow us to compare and contrast which cells had inaccurate predictions from our model and post-stratification.

References

Caetano, S., (2020) , 01-data_cleaning-survey1.R. 01-data_cleaning-post-strat1.R. ProblemSet3 - template-logistic.Rmd., https://q.utoronto.ca/courses/184060/modules.

Kassambara, A. , (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0.https://CRAN.R-project.org/package=ggpubr

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria., https://www.R-project.org/.

Robinson et al., (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.1., https://CRAN.R-project.org/package=broom

Ruggles et al., (2020), IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. Retrieved from [https://doi.org/10.18128/D010.V10.0]

Silver, N. (2016, November 08). 2016 Election Forecast. Retrieved November 03, 2020, from https://projects.fivethirtyeight.com/2016-election-forecast/?ex_cid=2016-forecast

Tausanovitch, C. & Vavreck, L., (2020), Democracy Fund + UCLA Nationscape, October 10- 17, 2019 (version 20200814). Retrieved from [https://www.voterstudygroup.org/publication/nationscape-data-set].

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham et al., (2020). devtools: Tools to Make Developing R Packages Easier. R package version 2.3.2., https://CRAN.R-project.org/package=devtools

Wickham et al., (2020). haven: Import and Export ‘SPSS’, ‘Stata’ and ‘SAS’ Files. R package version 2.3.1., https://CRAN.R-project.org/package=haven

Wickham et al., (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2., https://CRAN.R-project.org/package=dplyr

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29.,http://www.jstatsoft.org/v40/i01/.

Xie, Y., (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R