# Will Trump the 2020 Presidential Election? Prediciting the Chance of Donald Trump Winning using a Poststratification Prediction Model

Yuyu Fei, Biqi Jiang, Jiayi Yang, Yuwen Wu

Nov 2, 2020

## Abstract

We used a logistic regression model with poststratification to predict this outcome for different combinations of different demographic variables. We predicted the probability to be 39%. Next steps include collecting required data to calculate the margin of error and more variables to increase prediction accuracy. The results are of great use to those interested.

Keywords: America, 2020 United States Presidential election, Donald Trump, popular vote

# Introduction

On Tuesday, November 3, 2020, the 2020 United States Presidential election will be held. The outcome is of interest to the world. We analyzed individual-level survey data and poststratified census data to predict the probability of Trump winning the election. The 2020 presidential election is scheduled to occur simultaneously with elections to the Senate and the House of Representatives. Also, Gubernatorial and legislative elections will take place in several states. After the election, the United States House will reassign the seats among the 50 states using the results of the 2020 United States Census. Most of the time a party that wins the presidential election experiences a tornado effect that also helps other candidates of the smae party win elections. Thus, the party that wins the presidential election could win a significant advantage in getting new Congressional and state legislative districts that will remain in effect until the 2032 elections.

The aim of this analysis is to estimate the probability of presidential candidate getting the majority of votes for the 2020 American presidential election using statistical technique: multivariable regression with post-stratification. The outcome variable we were particularly interested in was whether a vote would vote for Donald Trump, which was a binary outcome. First, we fit a multivariable logistic regression model to fit our outcome variable using a few demographic demographic characteristics. Next, we poststratified the selected sample with the variables in the logistic regression model. We then assigned sampled units into different cells based on combinations of the variables. We used the logistic regression model to predict the probability of giving the vote to Donald Trump for each cell. Ultimately, we combined the estimated probabilities of all cells to compute the probability of Donald Trump getting the popular vote.

The estimated probability of Trump winning from our poststratification model is 39%. We roughly estimated the margin of error but should be taken lightly since a naive estimation was conducted. The generalizability of our results are limited due to low prediction power. Next steps will be gathering more data to estimate the margin of error and more variables to increase prediction power. We could possibly pool data from multiple surveys for estimation to decrease bias and obtain more accuracy in the estimates of the desired probability.

# Data

### Survey dataset

We also used individual level data from the wave 50 Democracy Fund UCLA Nationscape dataset collected and compiled by UCLA Democracy Fund.Nationscape conducts weekly surveys. The first wave went out on July 18, 2019. Each survey was in the field one week. The target population is American residents in 50 states. The methodology of sampling for this dataset is stratified sampling with inverse weighting. The frame were online surveys as interviews were conducted online and the respondents had access to a computer connected to the Internet. The biggest strength of this dataset is its complexity and variety of variables collected. The weaknesses is non-response bias. Non-response bias was corrected. Standard two-phase sampling weights can be developed to eliminate non-response bias.

### Survey dataset: exploratory data analysis

We see from Figure @ref(fig:fig2) that less than 40% of the sample would vote for Trump.

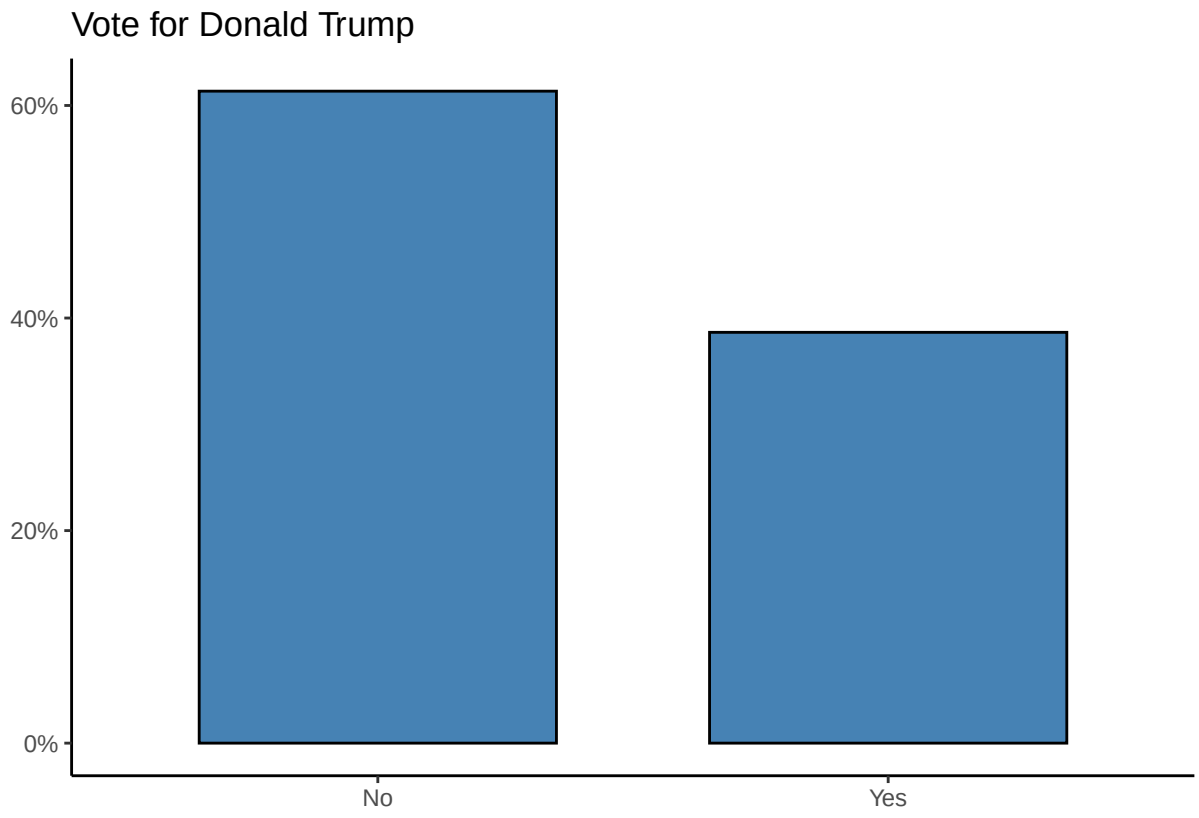


Figure 1: Proportion of sampled individuals in the survey data who would vote for Donald Trump

We see from Figure @ref(fig:fig3) men are more likely to vote for Trump than women.

We see from Figure @ref(fig:fig3) that minorities are less likely to vote for Trump, especially Black people.

We see from Figure @ref(fig:fig4) that younger people are less likely to vote for Trump.

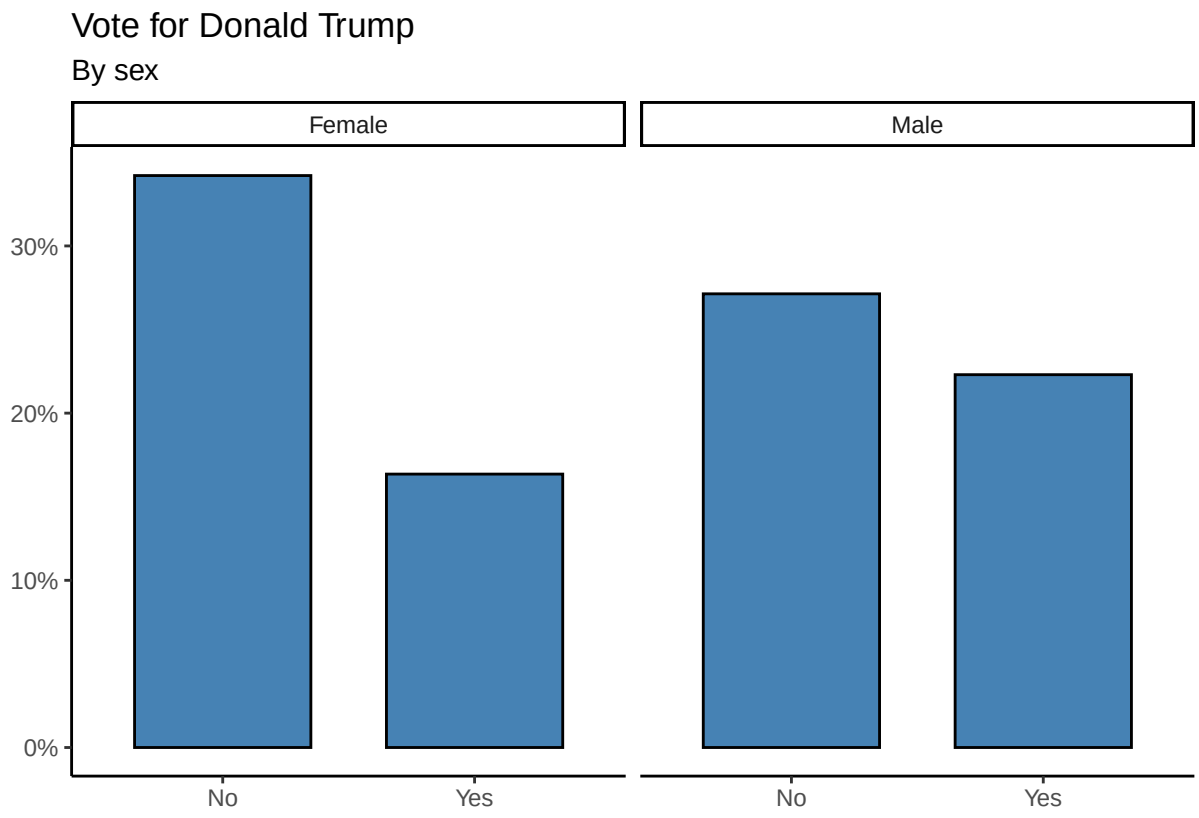


Figure 2: Proportion of sampled individuals in the survey data who would vote for Donald Trump by Sex

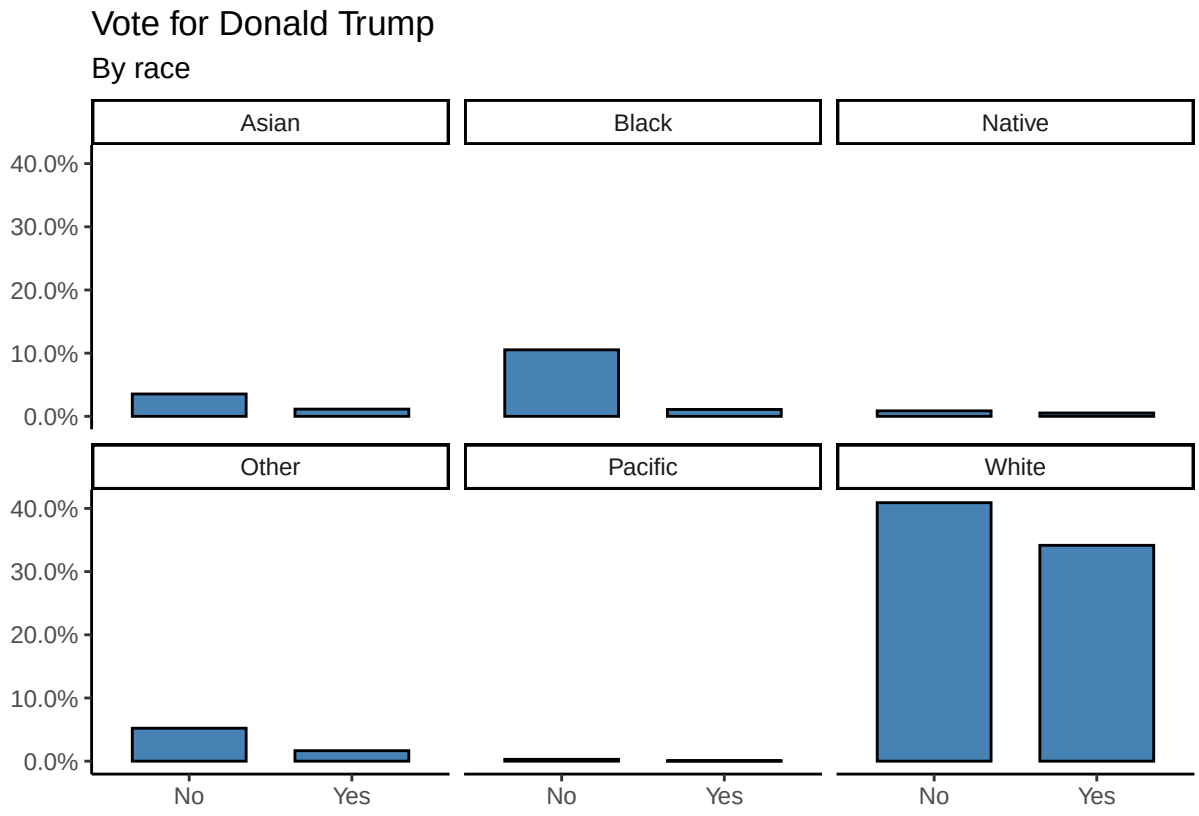


Figure 3: Proportion of sampled individuals in the survey data who would vote for Donald Trump by Race

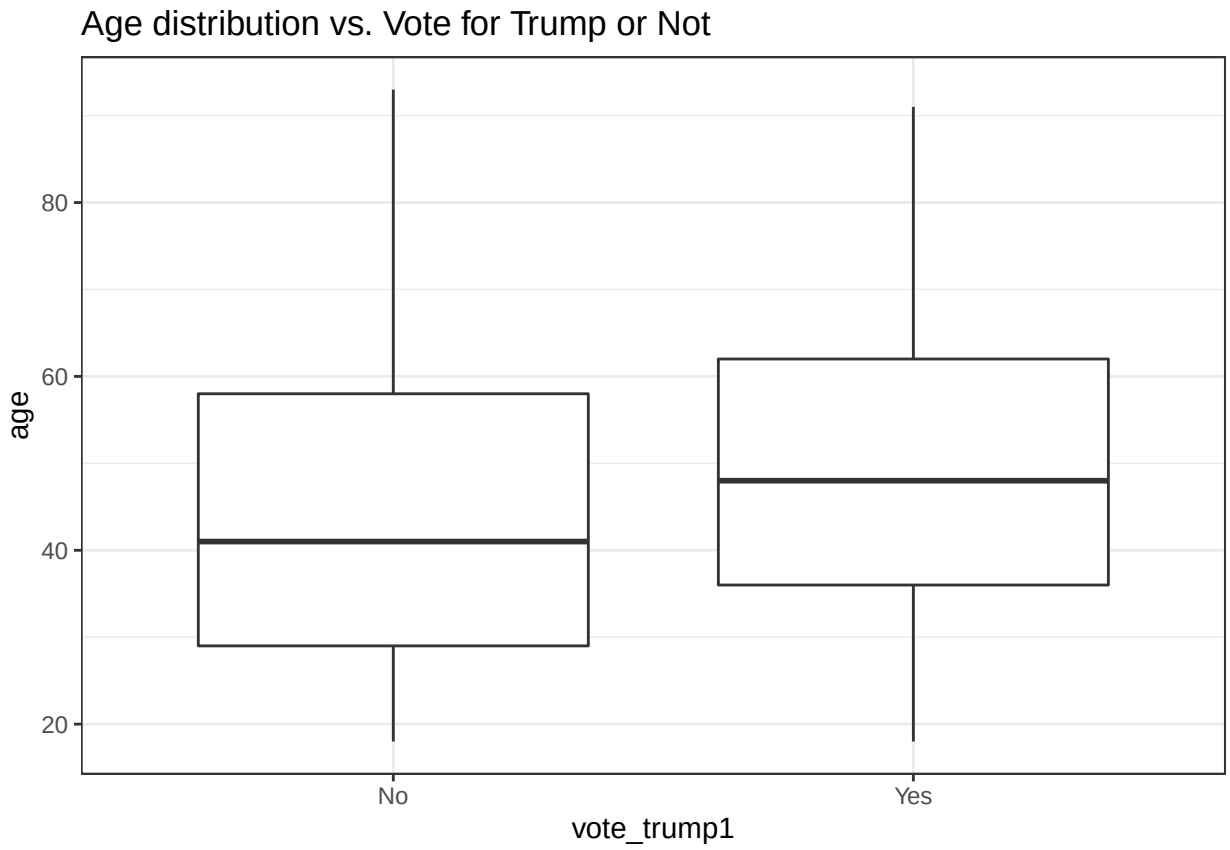


Figure 4: Proportion of sampled individuals in the survey data who would vote for Donald Trump by age

**Post-stratification dataset**

We used poststratification data from the census dataset American Community Surveys (ACS) from 2018. The source of the data is IPUMS USA. IPUMS USA collects, preserves and harmonizes U.S. census microdata and provides easy access to this data with enhanced documentation. Data includes decennial censuses from 1790 to 2010 and American Community Surveys (ACS) from 2000 to the present. The American Community Survey (ACS) is a demographics survey program produced by the the United States Census Bureau. It constantly collects data previously included only in the long form of the decennial census, such as ancestry, citizenship, educational attainment, income, language proficiency, migration, disability, employment, and housing characteristics. The target population is American residents in 50 states. The methodology for data colleciton is stratified sampling with inverse weighting probability. Non-response bias was corrected. Standard two-phase sampling weights can be developed to eliminate non-response bias. The most outstanding strength of this dataset is its clarity and completeness of variables collected.

Multilevel regression with poststratification corrects for differences between survey samples and the target population. National surveys, although representative at the national level, are often problematic and biased in terms of representativeness or locational coverage at the state or congressional district level, because of clustering and other survey strategies used by polling companies. With the rising popularity of internet survey strategies and mobile devices, it is becoming challenging to even find data with a random sample of the countrywide population, let alone random samples of subgroups of interest like demographic slices or residents of particular states. Multilevel regression with poststratification addresses these concerns. But one disadvantage of this technique is that a large number of sample units (big data) is required.

**Post-stratification dataset: exploratory data analysis**

We see from Figure @ref(fig:fig5) that the distribution of age is even.

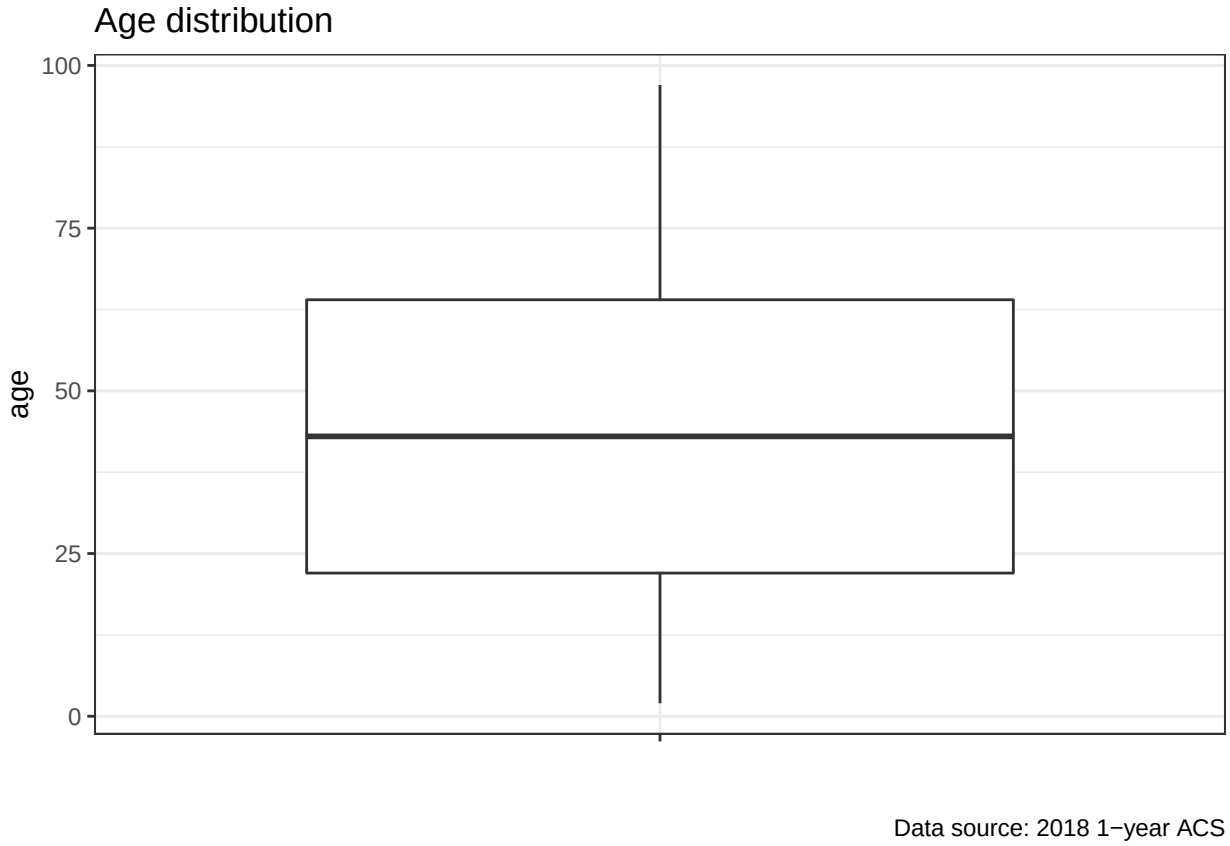


Figure 5: Distribution of Age

We see from Figure @ref(fig:fig6) that the distribution of sex is even.

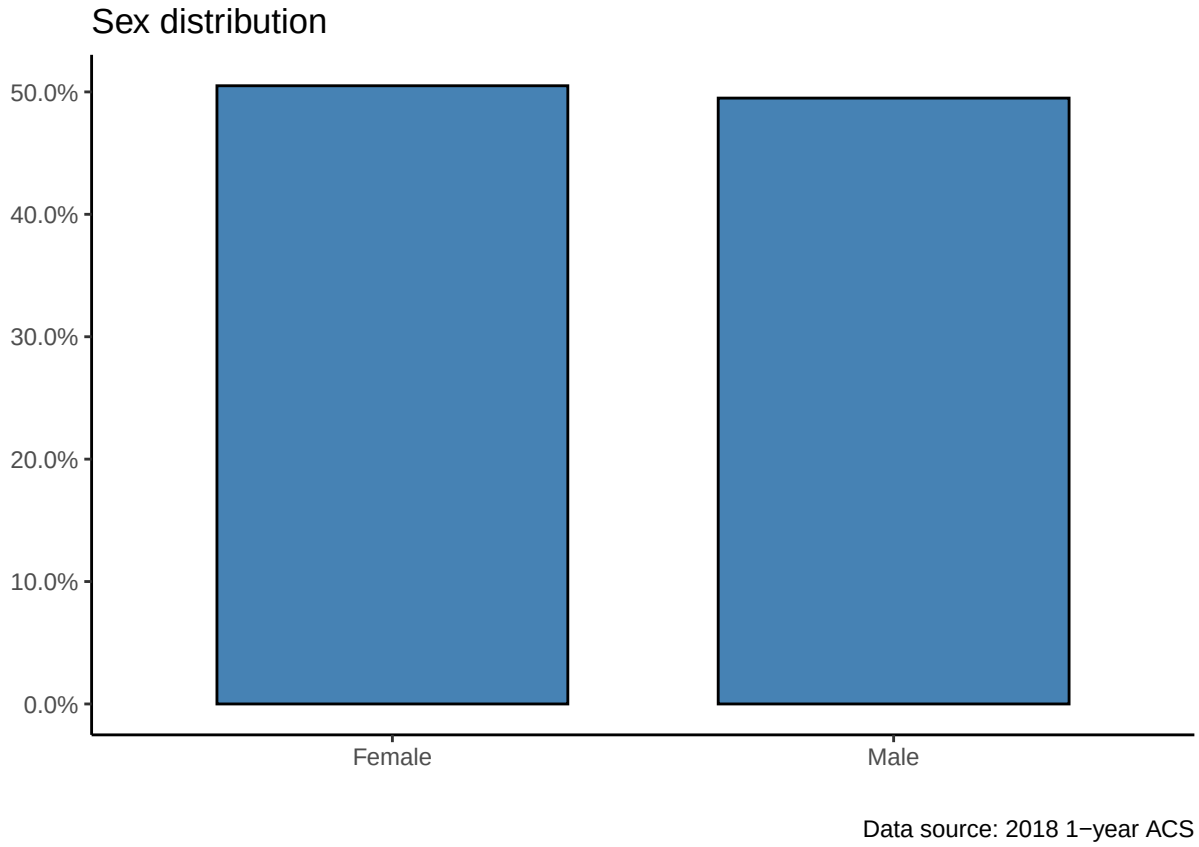


Figure 6: Proportion of Sex

We see from Figure @ref(fig:fig7) that the distribution of state is not super even. Some states are more populated than others.

We see from Figure @ref(fig:fig8) that the distribution of race is fairly even. As expected, there are more whites than other ethnic groups in America.

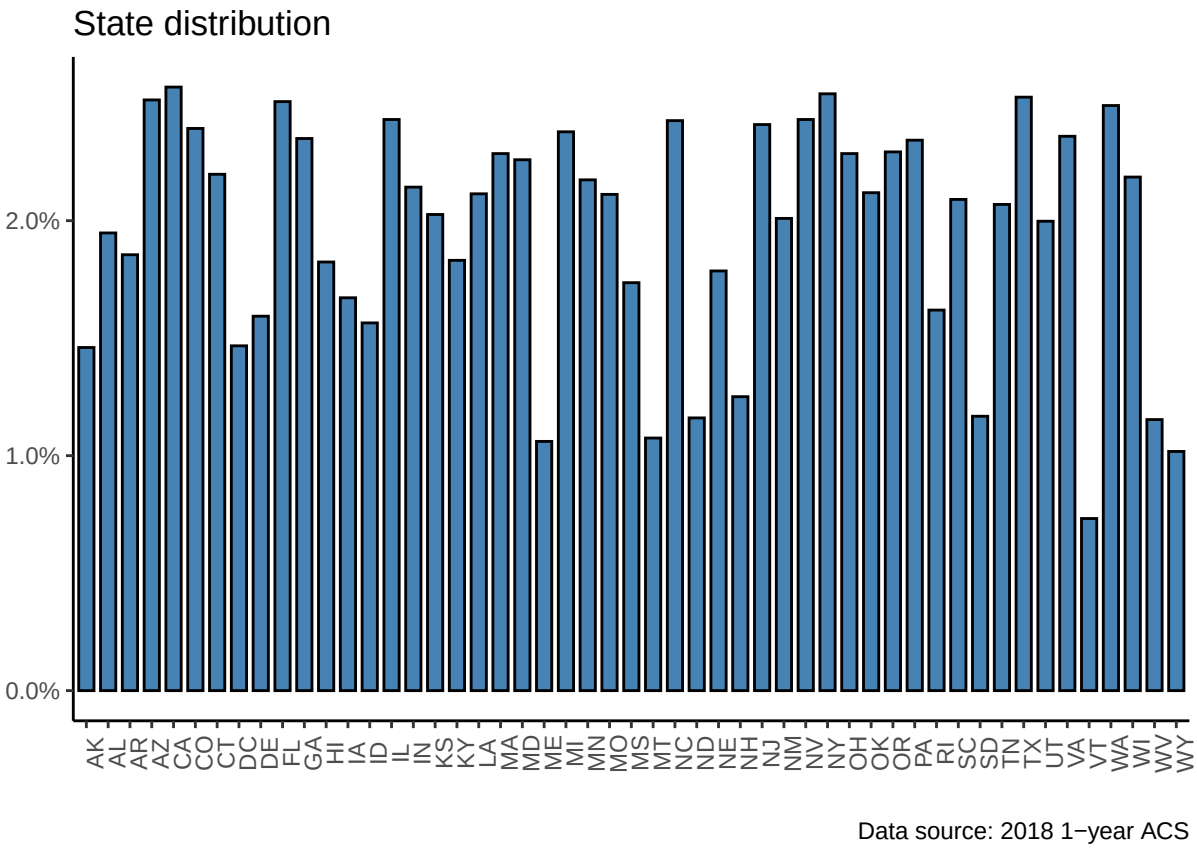


Figure 7: Proportion of State

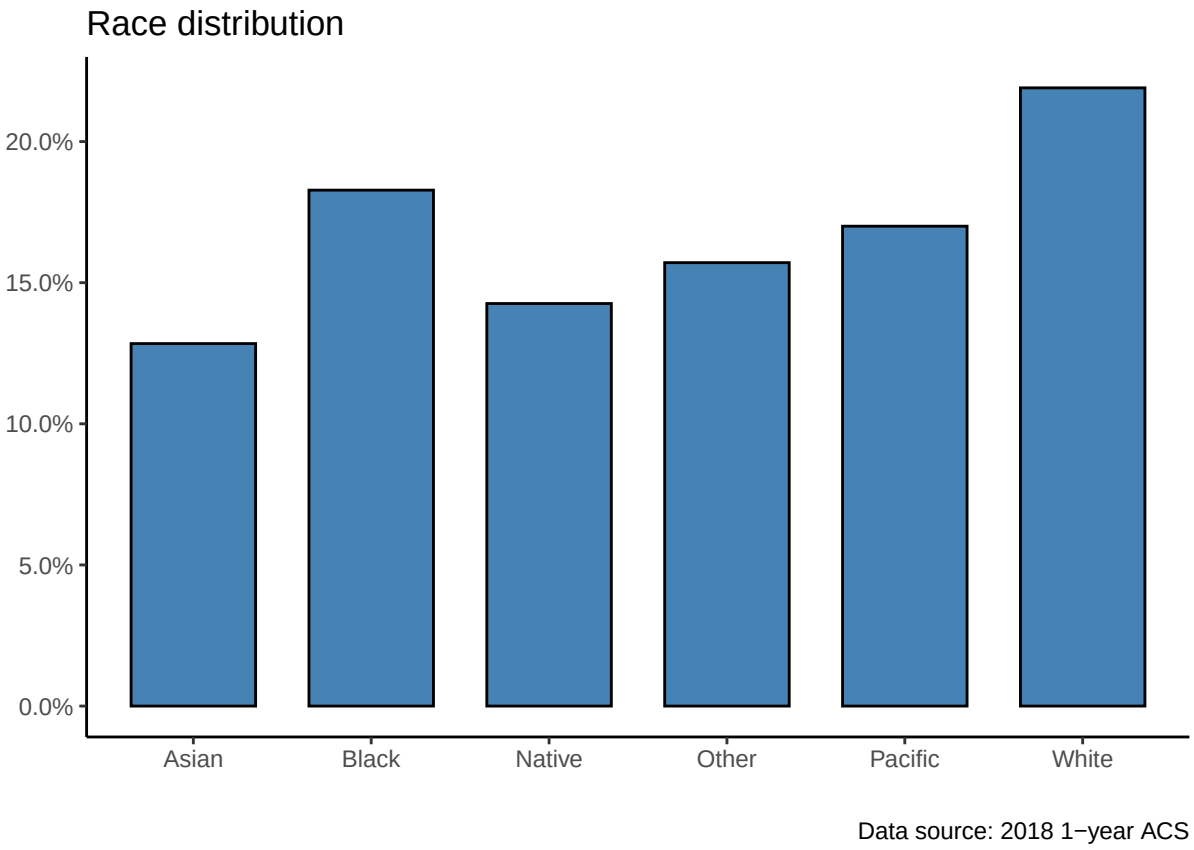


Figure 8: Proportion of Race

# Model

We used a logistic regression model with the response variable being vote for Donald Trump or not. The logistic regression model models the probability of a particular class or event happening such as pass/fail, win/lose, alive/dead or healthy/sick. In our case, it was voting for Trump/ not voting for Trump. The regression coefficients could be interpreted as log odds ratios and the actual statistic being modeled is the log of the odds of voting for Trump. The covariates we included in the model were age, sex, state and race.

First, we fit a multilevel logistic regression model to our survay dataset to estimate how likely a voter is to vote for Donald Trump based their demographic characteristics and the state where they live. The major advantage of using multivariable models for estimating state-level opinion is that it allows us to incorporate extra information beyond the polling sample dataset into our analysis. Things like past election results, income statistics, crime rates, and any number of other state-level data can help us get reliable estimates of the opinion to vote for Trump, given a meaningful relationship between those data and opinion.Next, we conducted post-stratification, where we used the logistic regression model we built to predict the probability of voting for Trump each variable combination. We post-stratified the survey sample and the population. In summary, the logistic regression allowed us to estimate trustworthy relationships between opinion and demographic and geographic variables of a survey sample, and poststratification corrects for differences that exist etween the survey sample and the target population.

## Results

The ROC curve for the logistic regression model is shown in Figure @ref(fig:figure1) The AUC is 0.69 indicating sufficiently good model prediction power.

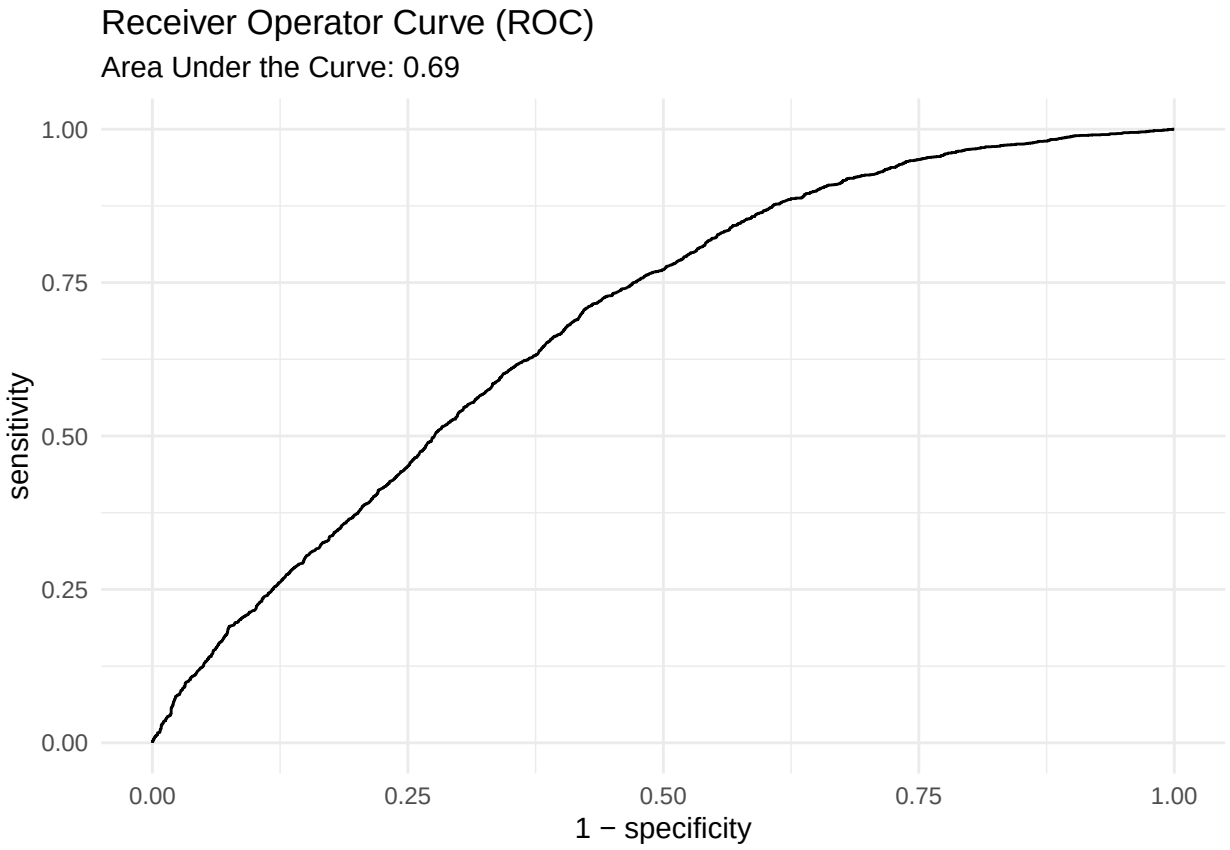


Figure 9: ROC curve of the logistic regression model

The point estimate of the probability of Donald Trump winning the election, the margin of error around it, and the 95% confidence interval around it, calculated from the poststratification conducted are presented in Table @ref(tab:table) below.

Table 1: Estimated Probability of Trump getting a Vote

| Estimated.Probability |
|---|
| 0.387 |

# Discussion

We used the survey data with individual data from UCLA's Democracy Fund's Nationscape wave 50 survey and poststratification data from the 2018 1-year American Community Surveys (ACS) to estimate the probability of Donald Trump winning the American 2020 Presidential Election. We used a multivarible logistic regression model with demographic variables age, sex, and race and geographic variable state as covariates. We then produced a final estimate from the model through poststratification to correct for differences between the sample and the target populaton. We estimated he probability of Trump winning the election to be 39%.

**Weaknesses**

There are two major weaknesses revolving around our analysis. The first one being our model has limited prediction ability; it only makes the correct prediction about 70% of the time. This is due to several reasons including not having enough variables, not having random effect terms (random intercepts or random slopes), and not having interactions amongst variables. The second major weakness is that we were unable to produce a margin of error and the associated confidence interval to quantify the uncertainty around our probability estimate. We were unable to do so because appropriate techniques have not been validated and that we would need more individual level, more granular data to compute it.

The generalizability of our results is limited for the following reasons: 1) we only used two survey datasets and these datasets were designed specifically for different reasons, 2) we the prediction ability of our model is limited and 3) multilevel modeling is probably more appropriate here. The major limitation of logistic Regression is the assumption of linearity between the outcome variable and the independent variables. In such cases, non-linearity as found in our analysis will limit generalizability.

**Next steps**

We could try to model state as a random effect variable and use a multilevel model. This way we could account for randomness in which state the sample unit comes from. Moreover, we could look for variables that could affect our outcome variable to decrease the amount of confounding and enhance the prediction ability of our model. Furthermore, we could fit different models with different combinations of variables and use appropriate techniques to compare models and choose the best one for our data. All these could make our estimate more precise. We could also quantify the uncertainty around our estimated probability by calculating a margin of error. But this is very complicated and could only be calculated with individual-level data for both survey and census data, which was complicated for our analysis.

## Dataset citations

Press, C., Finance, Y., & Newsweek. (2020, October 30). New: Second Nationscape Data Set Release. Retrieved November 03, 2020, from https://www.voterstudygroup.org/publication/nationscape-data-set

Team, M. (n.d.). U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. Retrieved November 03, 2020, from https://usa.ipums.org/usa/index.shtml

## References

Gelman, Andrew, and Thomas C Little. 1997. "Poststratification into Many Categories Using Hierarchical Logistic Regression."

Holt, David, and TM Fred Smith. 1979. "Post Stratification." *Journal of the Royal Statistical Society: Series A (General)* 142 (1): 33–46.

Little, Roderick JA. 1993. "Post-Stratification: A Modeler's Perspective." *Journal of the American Statistical Association* 88 (423): 1001–12.

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

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frederique Lisacek, Jean-Charles Sanchez, and Markus Muller. 2011. "PROC: An Open-Source Package for R and S+ to Analyze and Compare Roc Curves." *BMC Bioinformatics* 12: 77.

Smith, Terence MF. 1991. "Post-Stratification." *Journal of the Royal Statistical Society: Series D (the Statistician)* 40 (3): 315–23.

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