Trump expected to win 53.5% popular vote in 2020 US Election*

Lan Cheng, Liyuan Cao, Shabier Zulihumaer, Zijian Wang

2020-11-02

Abstract

This study mainly applied a GLMM model with random effect intercept using logit family to forecasting the 2020 US Presidential Election based on a nationscape individual survey and post-stratification data. It was found that Trump expected to win 53.5% popular vote in 2020 US Presidential Election and we have 95% confident that the final proportion of popular voting for Trump would fall from 52.1% to 54.8%. This study also found that poor black people is the group least likely to vote for Trump while high degree level educated rich white males is the main group most likely to support Trump. . And the findings also show voting rates could be much different across states.

Keywords: forecasting; us 2020 election; Trump; Biden; multilevel regression with post-stratification;

1 Introduction

The founders of the United States designed the electoral system two hundred years ago when the United States was founded. At that time, the main purpose of this electoral system was to prevent politicians from falsely promising voters to cheat for their votes, therefore the "electors" indirectly elected the president, so as to avoid malpractices. But today this kind of sense has disappeared. This kind of system is mainly to respect the rights of the every State within the United State. It is the embodiment of the decentralization and respect of local state power in American democracy. There are 538 electoral votes in the United States, which is the sum of the total number of senators (100), representatives (435) and representatives of Washington, D.C. (3). Senators are allocated by state, with two in each of the 50 states; and representatives are elected by population, with one of the representatives is elected from about half a million of American citizens. For example, with a population of more than 16 million, therefore New York has 31 representatives and with those two senators together, there are 33 electoral votes in total within New York. With the exception of Maine and Nebraska, if any presidential candidate wins the majority vote of this state, in other words, he has already won all the electoral votes of the state, which is called "winner take all". According to the electoral college system, a candidate wins more than half of 538 electoral votes which is equal to (270) in each state is elected as a president.

It is upcoming for the 2020 US Presidential Election soon, and the election is so important that it would not only affect the US itself but it would also affect the whole world. Especially, there might be changes in the relations between China and US if Biden win the election. And it is known that lots of factors affect the outcome of the election, some well known factors including race of voters, age groups of voters, social levels of voters. It is already found by lots of former studies that the different attitudes of black and white poeple, young and old people as well as poor and rich people could affect the final outcome of the election. Pople would try their best to vote for the candidate who would supposed to bring benifits to them.

Under this background, this study is mainly aimed to forecast the 2020 US Presidential Election based on a nationscape individual survey and post-stratification data using a formal statistical model. The model mainly investigated is the generalzied linear mixed model with logit family and random effect intercept, this study

*Code and data are available at: https://github.com/tong304/PS4.

uses this model by modeling the probability of voting for Trump through several important covariates like age, gender, income, race, education level and etc as well as random effect inercepts across states in US.

The structure of the report is organized as following: First, an introduction of the whole study was given. Second data sets and the GLMM model used in this study are discussed in the data and model sections respectively. Then the main results of models would be shown in the results section. At last, the discussion would discuss the results and findings, weakness as well as future work and next steps. A link to the source of the report file could be found in https://github.com/tong304/PS4/blob/main/paper.pdf.

2 Data

The data used in this study mainly contain two differents. The first one used comes from the Democracy Fund + UCLA Nationscape 'Full Data Set', the study used is a subset interested one, and the second one used is the 2018 5-year American Community Surveys (ACS) data which comes from the IPUMS USA. And we call the first one the nationscape survey data and the second one the post-stratification data in this study.

For the questionnaire of the surveys, they are good because they are nationscape surveys that lots of expense have been spent to collec them. The data sets are collected from large sample sizes which are reliable and in high quality. The questionnaires used in the nationscape are well designed and well tested in well selected smaples before used to the wholte large population.

For the methodology of the survey, the target population of the survey is all of the people who are living in the US across different states and no younger than 18 years old that be able to vote. The frame in the survey is the various source with dwelling frame. The original samples contain surveys in many different periods, this study uses a recent one which is several months before the upcoming election including individuals about 5000 units after data cleanning process.

The survey took a stratified sampling design, the states in US are divided into the straficiations. And then samples are selected from the straficiations with the sample unit to be the individuals in the household associated with the dwelling frame. Figure 1 illustrates the densities of ages grouped by gender and income for NS suverty data, we can find that the data in these groups are approximately balanced as the distributions of ages are similar overall across different groups.

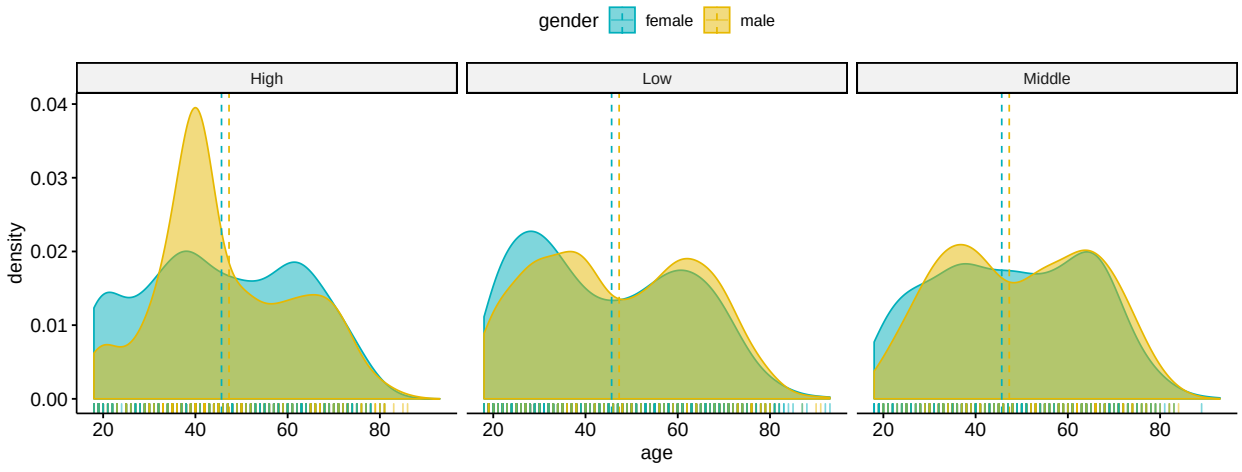

Figure 1: Densities of ages grouped by gender and income for NS suverty data

3 Model

The study mainly uses a generalized linear mixed model with logit family as well as random effects for states in US based on the survey data and evaluated on the post-stratification data to obtain estimated probability of voting for Trump. The GLMM model is described in details as following:

$$Y_{ij} \overset{\text{ind}}{\sim} Bernoulli(\lambda_{ij})$$
$$logit(\lambda_{ij}) = \mu + X_{ij}\beta + S_i + U_{ij}$$
$$S_i \overset{\text{ind}}{\sim} N(0, \sigma_S^2)$$
$$U_{ij} \overset{\text{ind}}{\sim} N(0, \sigma_U^2)$$

The variables are:

- Y_{ij} is the response whether an invidual would vote for Trump or Biden, 1 = Trump, 0 = Biden, it stands for i-th state, j-th invidual
- λ_{ij} is the probability of voting for Trump, it stands for i-th state, j-th invidual
- μ is the intercept to be estimated
- β are the parameteric coefficients to be estimated
- X_{ij} are covariates mainly include gender, age, income, education, race, employment
- S_i is the independent random effect for ith state
- U_{ij} is the independent random effect for ith state, jth invidual

For alternative models, as in this study, the response is whether an invidual would vote for Trump or Biden, 1 = Trump, 0 = Biden which is a binary outcome, so linear regression model is not considered in this study. And it is suitable for using both logit and probit families. However, it is rather hard to interpret the results uisng a probit family and interpretion of effects of factors on voting is very improtant so we use the logit family. Also, as the supporting rates for Trump and Biden are known to be different across states, this study also considers a random effect intercept for state, thus, the generalized linear mixed model with logit family and random effect intercepts for states is used rather than a generalized linear model. At last, bayesian models are not used due to the unknown information of priors and not efficiency as well as not stable for convergency in estimating based on a relative big survey data. We use the R programming software to run a GLMM with logit family and random effect intercept model. Also, this study conduct model checks and diagnostics mainly using creterions like AIC, deviance tests to ensure the inferences based on the model are reliable.

Table 1: Fixed effect estimates of GLMM model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.98	0.20	-14.94	0.00
age	0.01	0.00	5.96	0.00
gendermale	0.45	0.06	7.41	0.00
raceOther	1.52	0.16	9.63	0.00
raceWhite	2.17	0.14	15.90	0.00
educationDoctoral degree	0.46	0.20	2.26	0.02
educationMaster's degree	0.01	0.11	0.10	0.92
educationOther	0.31	0.07	4.13	0.00
employmentOther	0.04	0.25	0.18	0.86
employmentYes	0.35	0.07	5.08	0.00
incomeLow	-0.26	0.08	-3.35	0.00
incomeMiddle	-0.25	0.08	-3.02	0.00

Table 2: Model measures including AIC, BIC

criterion	value
AIC	6576.520
BIC	6661.754
logLik	-3275.260
deviance	6550.520
df.resid	5187.000

4 Results

In this section, we introduce the results based on the fitted GLMM model and the predicted probability of voting for Trump in US 2020 election. As GLMM model mainly includes two parts - fixed effects and random effects, both are presented. First, the fixed effects for gender, age, income, education, race, employment are illustrated in table 1. It can be found all of the variables included in the GLMM model are significant at 5% level due to p values of them are all lower than 0.05. This means all of these covariates could be very helpful in predicting probability of voting for Trump in US 2020 election. For examples, table 1 shows for each additonal one year old in age, the odds of chance voting for Trump would increase 1%; Males have an average 56.8% times higher in the odds of chance voting for Trump compared with Females. Also, we can find the people with DOctoral degree, employment and High income have higher chance voting for Trump. At last, cearly, white people is abouve 8.76 times of the odds of chance voting for Trump compared with black people in average. These findings show that all of the coviartes are important and could be used to determine whether a vote is for Trump or Biden.

The random effects for the states are also not ignorable, it is known that the support rates for Trump and Biden are much different across states. The GLMM model with random effects for states could reflect this fact, the states located at the right side of the figure 2 tend to have positive mean random effect intercepts, so that they are more likley to vote for Trump while the states located at the left side of the plot tend to have negative mean random effect intercepts, so that they are more likley to vote for Biden.

To evalute the performances of our GLMM model, Table 2 shows some important model measures criterions including AIC, BIC, Log likelihood, Deviance and Degrees freedom of residuals. These criterions show that the model is generally appropriate and could be used to compare with the performances of alternative models.

To investigate the distributions of probability of voting Trump among different groups, figure 3 shows the density of probability of voting for Trump grouped by race. Clearly, we can find the white poeple has a much higher average probability supporting Trump compared with black people, the other people is in the middle.

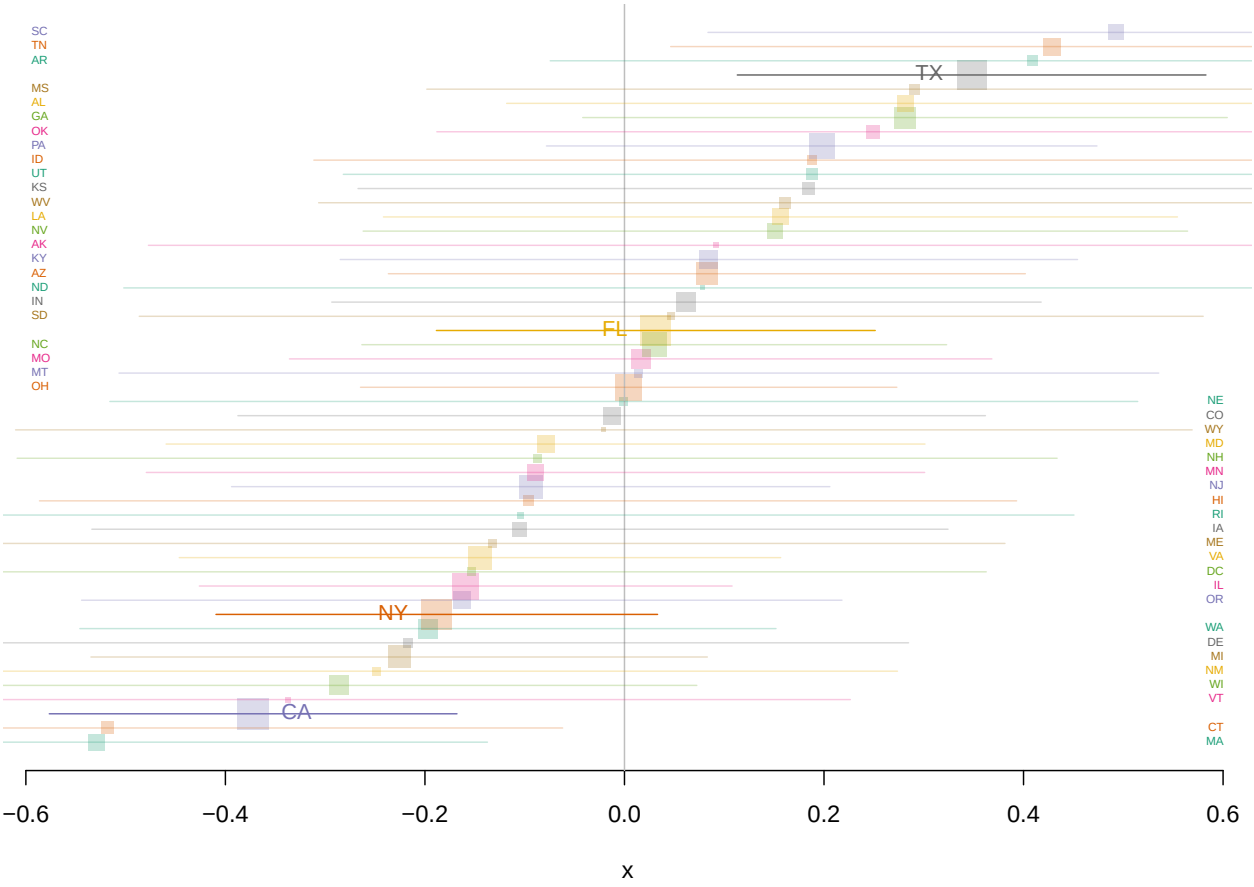

Figure 2: Random effects plot for states estimated by GLMM model

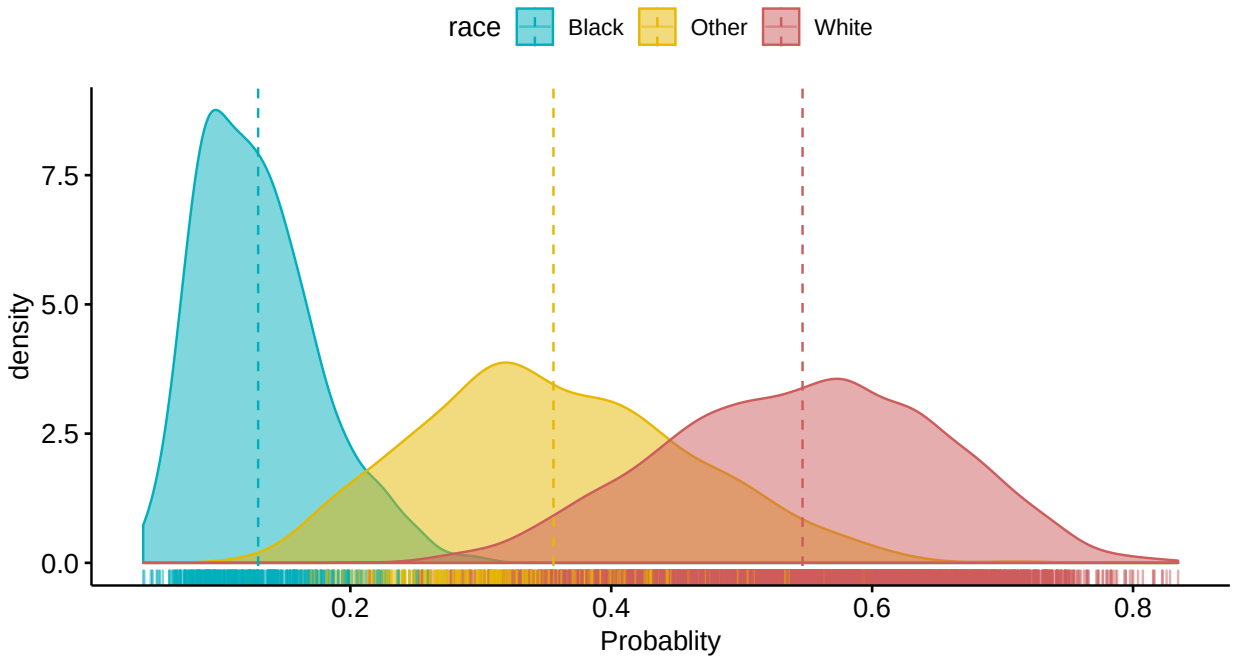

Figure 3: Density of probability of voting for Trump grouped by race

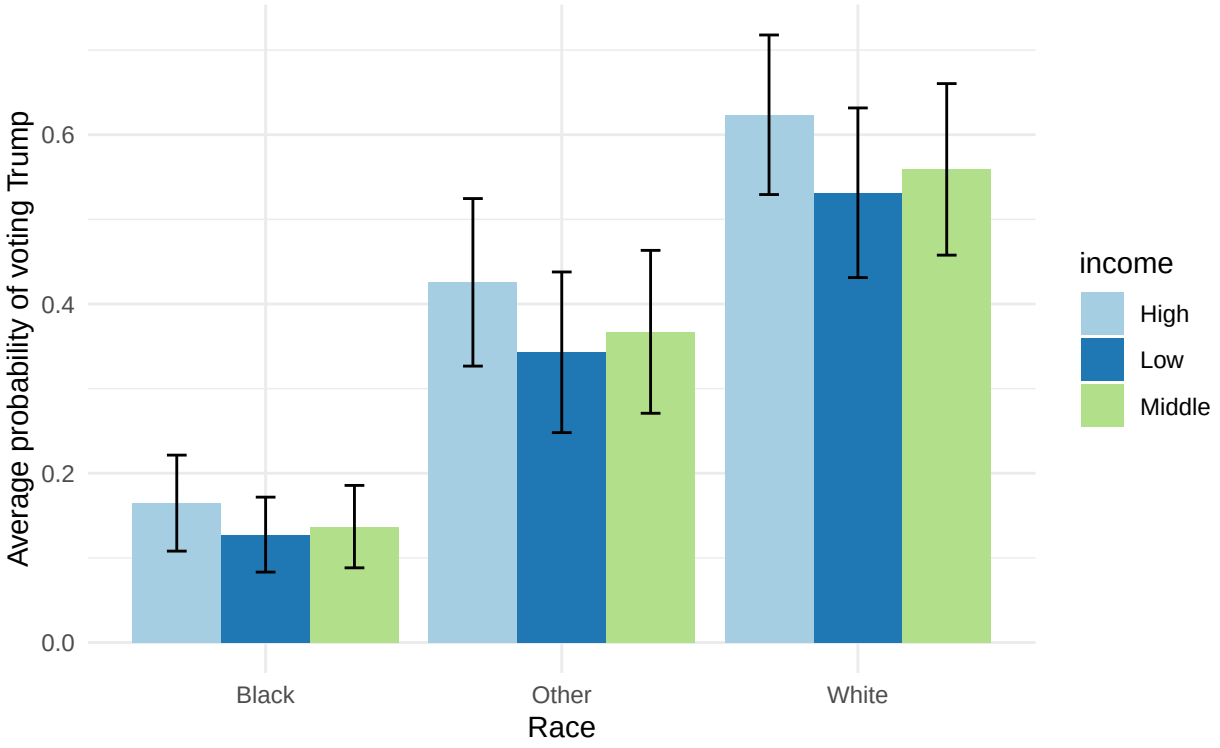

Figure 4: Distribution of probability of voting for Trump grouped by race and income

Thus, it is clearly that white people is more likely to vote Trump than black people.

To investigate the relations between probability of voting for Trump with race and income, figure 4 shows the average probability of voting for Trump with estimated standard errors grouped by race and income. We can find from the dodged error bar plot that overall while race is more likely to voting for Trump, more importantly, we can find among all of the three race groups, people with high income are tend to be more likley to voting for Trump compared with people with low and middle income level.

To investigate the relations between probability of voting for Trump with race and gender, figure 5 shows the average probability of voting for Trump with estimated standard errors grouped by race and gender. We can find from the dodged error bar plot that for all of the three race groups, males are tend to be more likley to voting for Trump compared with females obviously.

To investigate the relations between probability of voting for Trump with education and gender, figure 6 shows the average probability of voting for Trump with estimated standard errors grouped by education and gender. We can find from the dodged error bar plot that for both male and female, only people with doctoral degree tends to show clearly higher average probability of voting for Trump, this difference is not found between other education groups obviously.

To investigate the variations of probability of voting for Trump across states, figure 7 shows the average probability of voting for Trump in each of the 51 states in US for 2020 election. Using a cutoff 0.5, the states with probabilities higher than 0.5 are shown in the top which are support for Trump more than Biden, and the states which are in the bottom of the plot are support for Biden more than Trump. Clearly, we can find the numbers of states appear to be very close in supporting Trump and Biden correspondingly. A more interesting finding is that the state DC tends to be show very high probability in supporting Biden compared with Trump that no such state obviously.

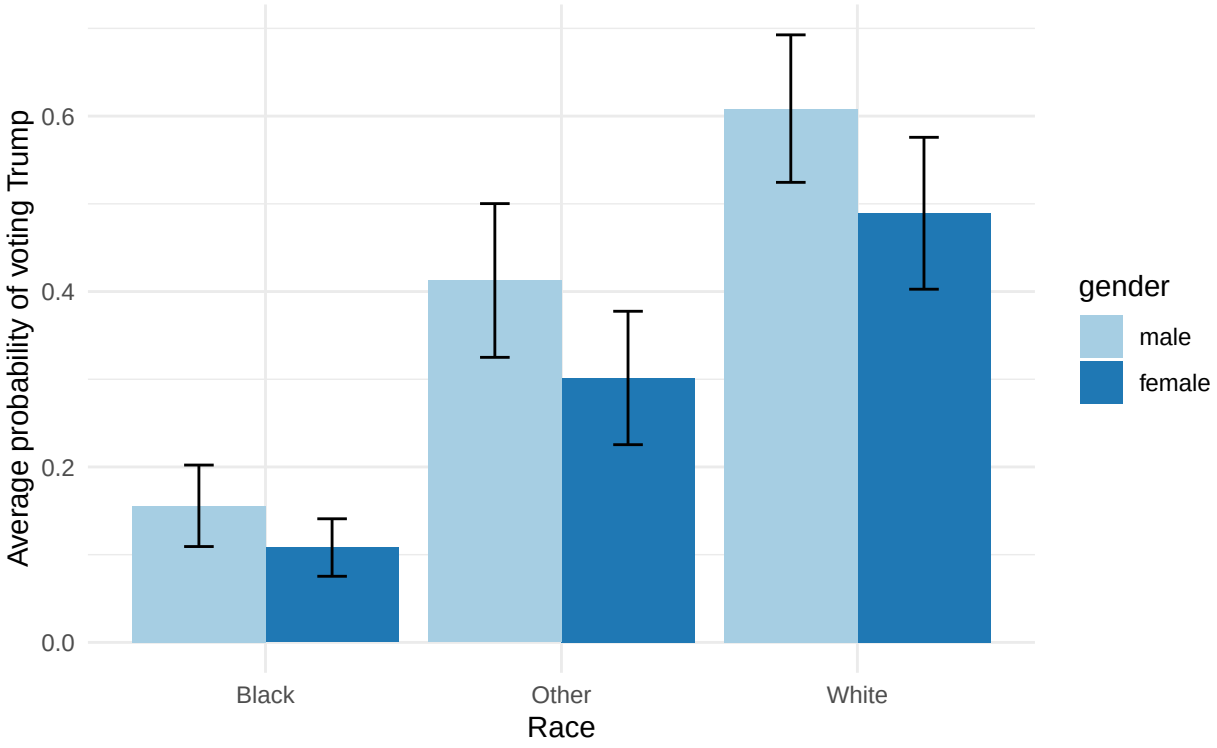

Figure 5: Distribution of probability of voting for Trump grouped by race and gender

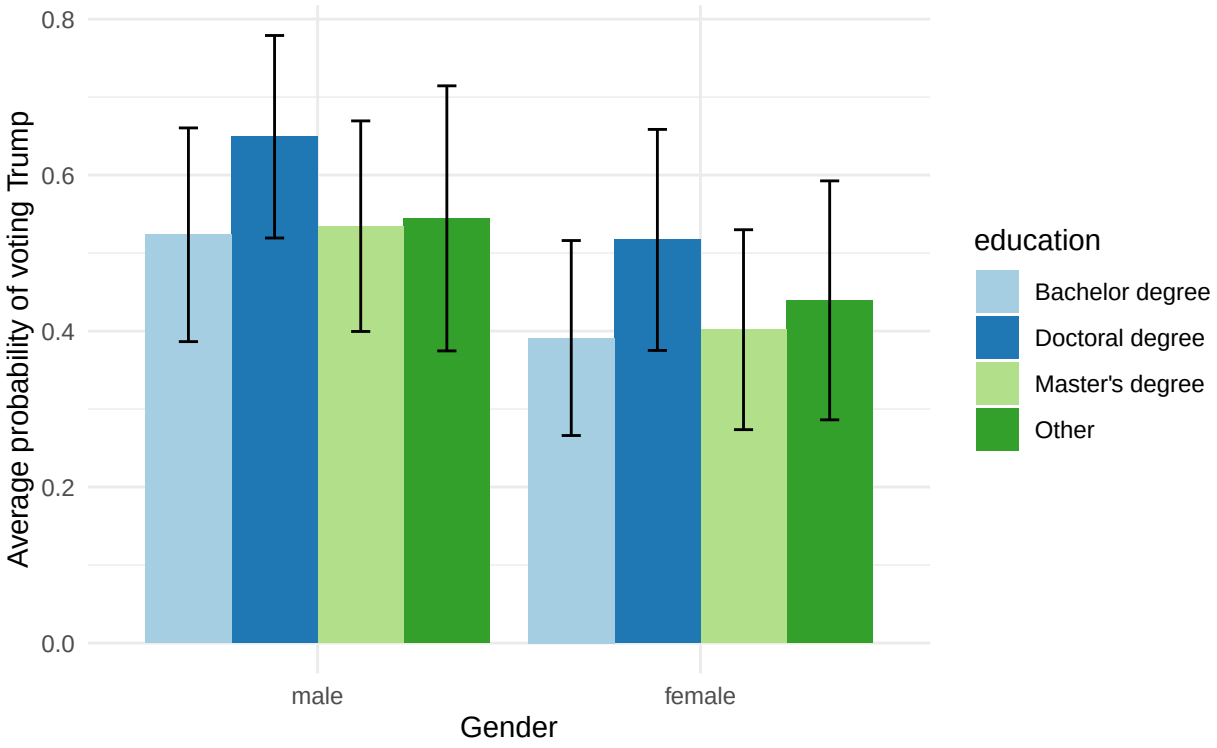

Figure 6: Distribution of probability of voting for Trump grouped by education and gender

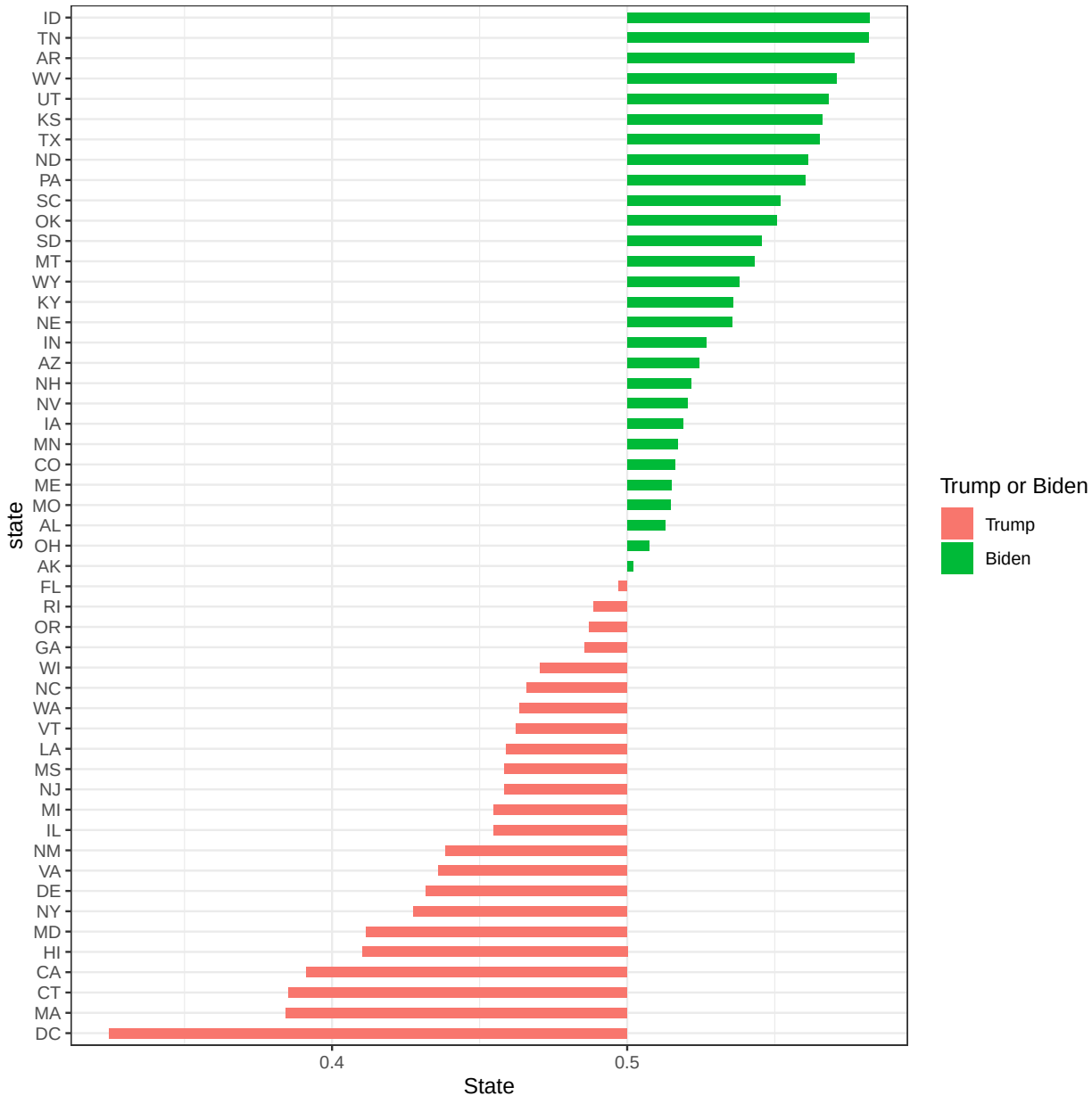

Figure 7: Distribution of probability of voting for Trump vs. Biden grouped by state

Table 3 Estimated Trump's winning probability with 95% wald confidence interval

Name	Estimates
Estimated Probability	0.535
Estimated s.e.	0.007
Eestimated 95% Wald CI	(0.521,0.548)

Finally, the forecasting for the US 2020 election is shown in table 3. It shows that Trump expected to win 53.5% popular vote in 2020 US Presidential Election and we have 95% confident that the proportion of popular vote would fall from 52.1% to 54.8%. This forecasting indicates that Trump will win the 2020 US Presidential Election with a slightly higher voting rate.

5 Discussion

5.1 Model findings

The Generalized linear model is mainly used to model the probability of voting for Trump with binomial logistic family as well as random effects for states. Based on the model, the most important findings are that white people is about 8.76 times of the odds of chance voting for Trump compared with black people in average; males is 56.8% times higher in the odds of chance voting for Trump compared with Females. And we also find that people with Doctoral degree with high income and in employment is more likely to vote for Trump. It is no big surprise that poor black people turn out to show such a low voting rate for Trump compared with rich white people, there are lots of accidients of black people shooting to dealth by white officers in Trump's government and the attitude of Trump's government is to do nothing about it, this should be one of the key reasons that poor black people does not support Trump. However, this study also find high educated people, rich people and people have jobs are the main group to vote for Trump, and most of them are males.

5.2 Forecasting

This study shows that Trump expected to win 53.5% popular vote in 2020 US Presidential Election and we have 95% confident that this percentage of vote would fall from 52.1% to 54.8%. As the lower bound of the 95% confidence interval of forecasting is still much higher than 50%, this indicates that Trump will win the 2020 US Presidential Election with a slightly higher voting rate. However, this forecasting could be affected by lots of factors. Besides the accuracy of the responses of voting for Trump or Biden, other important factors including the divisions of groups, balance of data, omitted variables biasness, non-responses biasness as well as sample size and quaility of post-stratification data.

The divisions of groups means that if we change the recoding of bins for income levels, education levels and etc, the forecasting would be change too. The balance of data means that the proportions of voting for Trump and Biden in the training survey data should be close enough otherwise the prior distribution of whether vote for Trump is biased. Also both of omitted variables biasness and non-responses biasness could affect forecasting results. The quality of data is also important.

Also, as we estimated the random effects of states as well as average probability of voting for Trump based on the model, we can find that the probability of voting for Trump could be more confident in states like TX, ID and etc as shown in figure 2 and figure 8, and the probability of voting for Trump could be least confident in the state DC as DC state supports Biden very obviously.

5.3 Weaknesses and next steps

Finally, there are some weaknesses in this study besides the above findings. First, as the study is performed on a survey data, there are issues of biasness in estimation such as non-response biasness. Second, the forecasting are based on a survey several months ago which might not be so reliable as people could change their mind easily due to lots of factors. For example, a person who is tend to vote for Trump might be turn to vote for Biden after a call from Obama who is aimed to persuade the person to vote for Biden. Third, the model used in this study does not use all of the possible information included in the survey data, only a very small subset of features like age, gender, income are used, this could cause omitted variables biasness in estimation. And also for the GLMM model, this study considers a logistic family with random effects for states, more advanced models might be investigated in next steps. For examples, we can also include random effects for inviduals, interaction among covariates, time series, clusters in errors for GLMM model. And after investigations of these works, model comparisions and model selection could be conducted to find a better model to make predictions for US 2020 election.

6 References

1. Alboukadel Kassambara (2019). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.2.4. https://CRAN.R-project.org/package=ggpubr
2. Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS','Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven
3. Hadley Wickham, Romain Franois, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr
4. Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
5. Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler and Benjamin M. Bolker (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. The R Journal, 9(2), 378-400.
6. Patrick Brown (2020). Pmisc: Various Utilities for knitr and inla. R package version 0.3.2/r2380. https://R-Forge.R-project.org/projects/diseasemapping/
7. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
8. Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020.https://doi.org/10.18128/D010.V10.0
9. Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=196fe68b-363c-46f1-880b-75b48cd5dc4d
10. 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
11. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.27.