Who will win Popular Vote in 2020 US Election?*

Joe Biden predicted to win with 52.5% vote +/- 5.2% confidence by a multilevel logistic regression with post-stratification

Yinuo Zhang

10 April 2022

Abstract

US election is always a big deal attract people's attention all around the world, for the upcoming US 2020 election, it hits the world again. This study aims to build a multilevel logistic regression with post-stratification to forecast the outcome for 2020 US Election. The study was conducted based on Democracy Fund + UCLA ationscape survey data and American Community Surveys (ACS) Post-stratification data. The findings show Joe Biden will win the Popular Vote in 2020 US Election over Donald Trump (52.5% vs. 47.5%, +/- 5.2% confidence).

Keywords: US 2020 Election; Joe Biden; Donald Trump; post-stratification;

1 Introduction

The question like "Who will win Popular Vote in US Election?" would become hot when the Popular Vote in US Election coming once again. The 2020 United States presidential election was on November 3rd, 2020 and it indeed became a hot topic in that period. In the end of year 2019, due to COVID-19, the 2020 United States presidential election became more interesting as it would determine the two different ways of US in treating COVID-19 too. Donald Trump who won the 2016 election represented the Republican party and the vice-president of Barack Obama represented the Democratic party. People around the world were attracted by the US election, most of them wanted to predict the outcome before the outcome of election.

This study would also review the issue of US 2020 election and aims to build a multilevel logistic regression with post-stratification to forecast the outcome for 2020 US Election to investigate whether the predicted outcome matched the true outcome which Joe Biden won. The multilevel logistic regression with post-stratification is a widely used method that first it builds a model on a small survey data and then applied to a larger census data to obtain a nationwide outcome. The study was conducted based on Democracy Fund + UCLA ationscape survey data which was token by lots of scholars and analysts to investiget the behaviours of voters in American (Tausanovitch and Vavreck (2020)) and American Community Surveys

*Code and data are available at: https://github.com/Amy527/paper4.

(ACS) Post-stratification data which was published by Integrated Public Use Microdata Series (IPUMS) (Ruggles et al. (2020)). The key variables used are age, sex, race, education, income, hispanic or not. The study built model on Democracy Fund + UCLA ationscape survey data and applied the model on American Community Surveys (ACS) Post-stratification data, the study obtained the overall vote percentage would be won by Joe Biden with confidence interval as well as vote percentage won across all of the US states individually.

The study is important because once we have a good experience in using multilevel logistic regression with post-stratification and got an accurate forecasts of Popular Vote in 2020 US Election, we could apply the similar method in the next 2024 US Election, and we might be able to predict the outcome as there are lots of beneficials in predicting the outcome of US Election.

The study was organized as following: In data section, we introduce the Democracy Fund + UCLA ationscape survey data and American Community Surveys (ACS) Post-stratification data. In model section, we introduce the multilevel logistic regression with post-stratification and interpret model results. In results section, we show the forecasting results of votes with tables, graphs and maps. Finally, in the discussion section, we discuss the results and findings as well as weaknesses and next steps of the study. The study is carried out using R (R Core Team (2020)), Rmarkdown (Allaire et al. (2021)), tidyverse (Wickham et al. (2019)), ggplot2(Wickham (2016)), dplyr (Wickham et al. (2021)), scales (Wickham and Seidel (2020)), broom (Robinson, Hayes, and Couch (2021)), cowplot (Wilke (2020)), knitr (Xie (2021)), maps(Richard A. Becker, Ray Brownrigg. Enhancements by Thomas P Minka, and Deckmyn. (2021)).

2 Data

This study used the U.S.presidential election 2020 survey data from Nationscape on June 25, 2020 to train the multilevel logistic regression and then used the model fitted on the American Community Surveys (ACS) Post-stratification data.

According to Scheibe, Mata, and Carstensen (2011), there was found that age differences played an important role in the 2008 US presidential election, commonly, age was decied into the 4 groups: 18-35, 36-49, 50-65 and 65+. Valentino, Wayne, and Oceno (2018) claimed there was difference in gender attitudes in the 2016 US presidential election, it was claimed that males were consistently to be more conservative positions than females. Deckman and Cassese (2021) discussed the education level role in 2016 US presidential election, the study pointed out education level was an important part in voting behavior. Greenwald et al. (2009) discussed the effects of race in affecting the 2008 US presidential election, race was an important factor. Also, it was believed the vote willness was different between diferent income levels, the voting behavior between rich and poor voters was different. Based on the above researches, this study used age, sex, race, education, income, hispanic or not as explanatory variables and the dummy binary variable created from the variable

'vote_2020' in the survey data that it was coded 1 for Joe biden win and 0 for Donald trump win, other vote outcomes were ignored in this study.

2.1 Survey Data

The survey data comes from Democracy Fund + UCLA Nationscape (Tausanovitch and Vavreck (2020)). The sampling method of the survey is the stratified sampling as the Democracy Fund groups potential voters by several key demographics such as age, gender, race and education level. The population of the survey is people live in the United States, the sampling frame of the survey is dividing all people live in the United States into groups by key demographics such as age, gender, race and education level. The sample of the survey is people who conducted the survey. The survey was conducted online, the survey from June 25, 2020 was used in this study which includes 6046 observations and 266 variables.

The survey data was designed to cover as many ass possible factors could show the voteing behavior of voters, it includes age, gender, race, income, education, job, region, religion and so on. Due to different groups of people would support different political parties due to issues of gun policy, health care, income tax and so on, for year 2020, covid-19 insurance would be another important issue.

The survey was reliable to be representative of the American population, as it was designed weekly by Nationscape on the Lucid market research platform which targeted on specified group of people based on key demographics such as age, gender, race and education level. People with speeding answers (i.e. completed survey too quick) or people use similar answers for all questions in the questionnaires were dropped to make the results of survey as accurate as possible. So the Nationscape survey is useful due to the procedures to make sure the data obtained are accurate and could be representative of the American population.

The survey also encounter some weaknesses, first, it was conducted online from randomly selected counties and cities, there might be selection bias. Second, there might be non-response bias, people might missed or did not response to the survey.

2.2 Post-Stratification Data

The Post-Stratification Data comes from IPUMS America Census Service (ACS) (Ruggles et al. (2020)), U.S. Census Bureau designed the projects of ACS and the IPUMS hold data of ACS.

The sampling method of the ACS is the stratified sampling based on subgroups from several key demographics such as age, gender, race, education level and so on. The population of the ACS is United States population, the sampling frame of the ACS is people completed the census that about 1 out of 100 persons. The sample of the ACS is household not the individual as information like household income and topics are more easily conducted in households than individuals, the samples would be selected from subgroups.

Table 1: Votes percentages in Polling data

Candidate	Votes	Percentage
Donald Trump	2094	46.052
Joe Biden	2453	53.948

ACS has lots of strengths, first, it is a national-wide census, the population is the whole United States population. Second, it is more reliable as it is a nationwide census, the trends in different groups are accurate as information are from millions people. The weakness of the IPUMS data mainly are that the data is expensive as the census is nationwide, the cost is huge, also, there were some confusion records in the census data such as extremely high income, age and etc. Also, some data might be fixed using imputation methods which are something like "black box," people can not know what exactly be repaired and adjusted. IPUMS data also focus on the accuracy, they try to make sure the survey would be completed with less biasness.

The Post-Stratification Data has millions records, to keep consistent with the survey data, similar key attributes like age , gender, race, income level and education level were selected. Also, due to different sources of data, data cleanning procedures were conducted mainly to make the data variables have consistent names, levels of categories. For age groups, we have 4 groups: 18-35, 36-49, 50-65 and 65+. For race, we have 3 groups: white, black and other. For income level, we have median level or below and above median level where median level is defined as 36000 dollars computed from the ACS data. For education level, we have 3 groups: High school or below, BA or below, Abova BA. Also, we have groups whether haspanic or not.

Figure 1 shows the data consistency between the survey data and the Post-Stratification data for the key attributes age , gender, race, hispanic or not, income level and education level. Figure 1 shows the data consistency between the survey data and the Post-Stratification data across states.

Table 1 shows the overall vote percentage for Biden is about 54% which is much higher than that of Trump. Figure 1 shows that for most of parts, the polling survey data matches the Post-Stratification data well. The major difference is there is more proportion of voters aged over 65 and less roportion of voters aged 18-49 in the Post-Stratification data compared with thoses in the survey data. Also survey data appears to collect more data from BA education level while Post-Stratification data contains more data from High school or below level. The other variables gender, hispanic, race and income level are very similar between the survey data and Post-Stratification data.

Figure 2 shows that the population distributions by states that the polling survey data matches the Post-Stratification data well overall, but for states like arizona, florida, the gaps are not small.

Figure 3 shows the Joe Biden supports by demographics, clearly, it can be seen that younger ones 18-35, median income or beloe and blacks tend to be more willing in voting Biden. Figure 4 shows the Votes

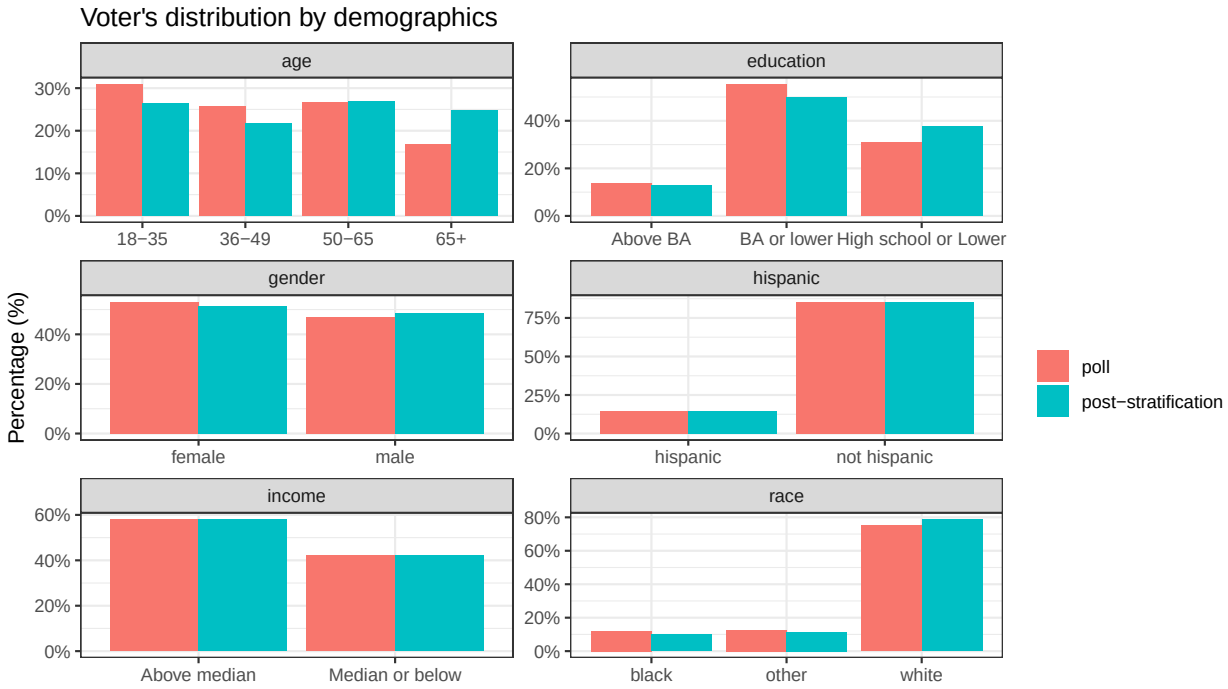

Figure 1: Voter's distribution by demographics

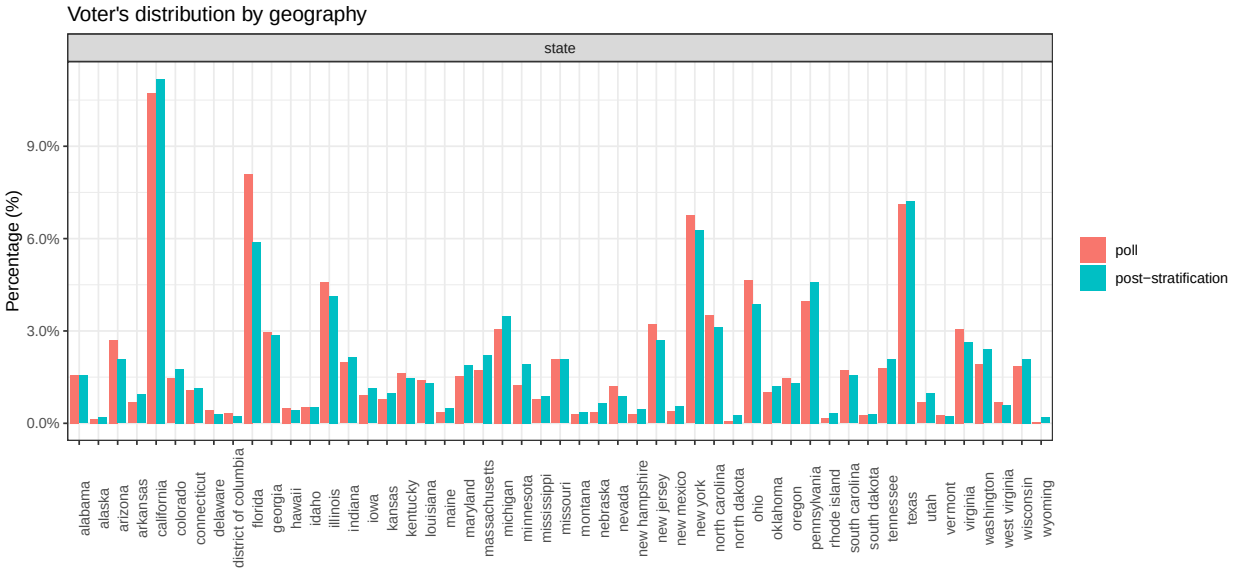

Figure 2: Voter's distribution by geography

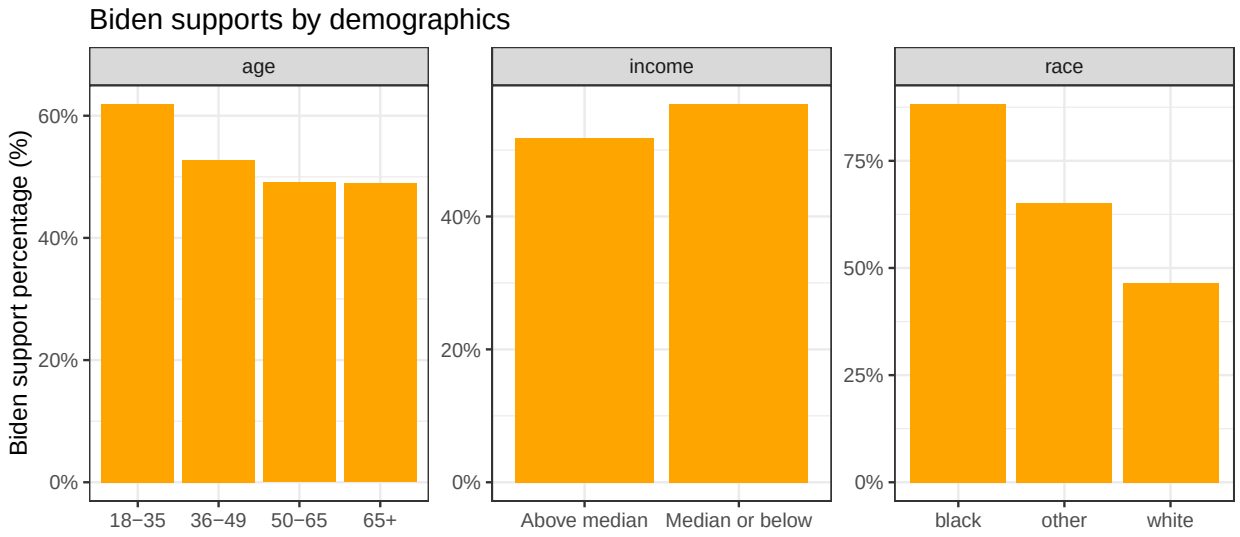

Figure 3: Joe Biden supports by demographics

percentages for Joe Biden across states, clearly, it shows the support rates declined by states, the states like district of columbia, vermont, connecticut, massachusetts, new mexico, hawaii, california support Joe biden clearly but states like alaska, south carolina,pennsylvania, indiana, texas support Trump clearly.

Overall, we can say the survey data and the post-stratification data are consistent, there is no big gap between them. And the differences of supports among variables indicating the variables could be used to make predictions of votes as they have predictive powers. Thus, we can build model on survey data and apply it on post-stratification data.

3 Model

This study plans to use the multilevel logistic regression with post-stratification which is a method that first fit on a smaller data and then applied to a larger data. This method is very useful not only becasue it is simple but also due to it could be applied to large data from survey data, large data set costs much higher than a survey, but with a careful designed survey and obtain similar variables as large data set, one can use multilevel regression with post-stratification method to estimate results in the large data set efficiently. However, this method has some weaknesses too, this method requires data consistent between the survey data and post-stratification, and even the data sets are consistent, the outcomes on post-stratification data using model built on survey data might be biased.

In this study, the multilevel logistic regression with post-stratification was first build on the the Democracy Fund + UCLA ationscape survey data (Tausanovitch and Vavreck (2020)) and then applied to the American Community Surveys (ACS) Post-stratification data (Ruggles et al. (2020)). The equation (1) shows the

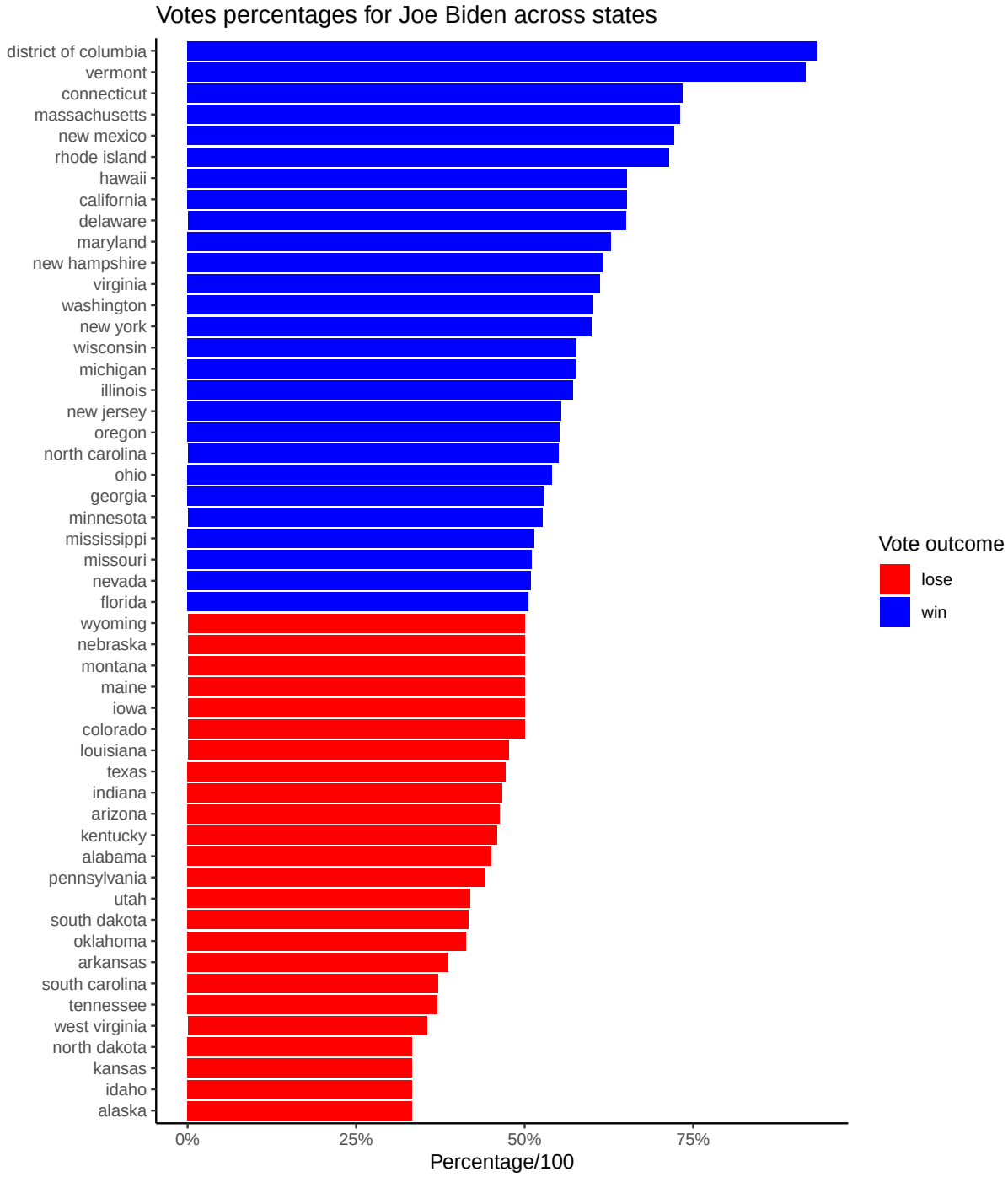

Figure 4: Joe Biden supports across states

form of the logistic model used. The logistic model is appropriate due to we are only interested in whether a voter will vote for Joe Biden or Donald Trump which is a binary outcome, the outcome 1 stands for Joe Biden and 0 for Donald Trump in this study. The logistic model assume the voters are independent and the data is large. For the assumption of independent, the Nationscape group already use unique IDs and for large data, the survey cleaned data is over 4000 which is large enough.

$$logit(p) = \beta_0 + \beta_1 sex + \beta_2 race + \beta_3 hispanic + \beta_4 age + \beta_5 income + \beta_6 education + \epsilon \tag{1}$$

where p is the probability of voting for Joe Biden, β_0 is intercept, other coefficients are for related variables, note that, here these coefficients could be multiple ones as the variable like age used in this study is age group with 4 groups that it has 3 dummies. The error term is ϵ.

The model would be built on the Democracy Fund + UCLA ationscape survey data (Tausanovitch and Vavreck (2020)) using `glm` function in R (R Core Team (2020)) and then applied to the American Community Surveys (ACS) Post-stratification data (Ruggles et al. (2020)). For Post-stratification data, we first group the data by the key variables which formed cells or bins, and we predict the probabilities of voting Joe biden for voters in that cells or bins, also, we can predict the results with 95% confidence interval. This could be done use `predict` function in R (R Core Team (2020)).

The predict formula of the logistic model is shown in equation (2):

$$\hat{p} = \frac{e^{\hat{y}}}{1 + e^{\hat{y}}} \tag{2}$$

where $\hat{y}$ comes from the (3) as following:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 sex + \hat{\beta}_2 race + \hat{\beta}_3 hispanic + \hat{\beta}_4 age + \hat{\beta}_5 income + \hat{\beta}_6 education \tag{3}$$

Where it is the linear combinations of estimated coefficients and variables.

With the probabilities of voting Joe biden obtained for voters in bins, we can use equation (4) to compute overall predicted probability of voting for Joe biden, this would be done overall in US all states as well as across states respectively.

$$\hat{p}^{PS} = \frac{1}{N} \sum_{i=1}^{L} N_i \hat{p}_i \tag{4}$$

Where $\hat{p}_i$ is predicted probability in the i-th cell or bin, N_i is voters number in the i-th cell or bin, N is overall number. In this study, bins are formed by subgroups from key variables age, sex, education and so on. The results are calculated for both US as whole and each state in US.

At last, the multilevel logistic regression with post-stratification model has some weakness here, first, we do not consider individual level in the post-stratification data but subgroups, second, we do not consider other candidates due to the logistic model here is only for binary outcome. Also, the model depend on survey data seriously, a diferent survey might lead to different predictions on the post-stratification data.

4 Results

Table 2 shows the Logistic model estimated coeficients, the results then would be applied to the American Community Surveys (ACS) Post-stratification data (Ruggles et al. (2020)). The coefficients are log-odds that positive ones means voters will likely vote for Joe Biden while the negative ones means voters will likely vote for Trump. The 95% confidence interval of the estimates are also shown in the table 2, when 95% confidence intervals contain 0s, then the variable's effect is not significant, so from the estimates and 95% confidence intervals, we can find the signs, the magnitude as well as significance of variables on the voting outcome clearly.

For the prediction example, use some voter comes from the connecticut state, aged 18-35 who are males and have income level above the median with white race not hispanic, the education level is above BA, the prediction from the model in the study would be 75.4% with a 95% confidence interval (68.9%, 82.0%). For another prediction example, use some voter comes from the pennsylvania state, aged 50-65 who are males and have income level above the median with white race not hispanic, the education level is High school or below, the prediction from the model in the study would be 24.0% with a 95% confidence interval (20.7%, 27.3%).

Figure 5 shows the estimated coeficients with 95% confidence intervals more directly sorted in desc order of estiamted coefficients, the error bars stand for confidence intervals while the points are the estimated coefficients. So it is more easily to find out which variables are positive and negative ones and whether the 95% confidence intervals include 0s for these variables.

Figure 7 shows the model diagnotics in details in the appendix. It shows the pearson residuals vs. fitted values from the model and the cook's distance plot. Pearson residuals vs. fitted values plot show there is no big issuse that the results formed two clusters clearly which are for those voting for Biden and Trump respectively. And as the higher cook's distance the large impact on the model, the cook's plot shows there are very few points with large cook's distance which could affect the results of model seriously, almost all of points have low cook's distance. so overall, the model diagnostics show the model is relatively strong.

Table 3 shows the voting percetanges for Joe Biden across states with 95% confidence intervals as well as the overall outcome in US. It could be seen clearly that lots of states with 95% confidence intervals including 50% which means these states are most hardest ones to predict and also, these states are most important one

Table 2: Logistic model estimated coeficients

term	estimate	std.error	statistic	conf.low	conf.high
(Intercept)	2.595	0.323	8.038	1.961	3.228
sexmale	-0.382	0.065	-5.854	-0.510	-0.254
raceother	-1.732	0.169	-10.228	-2.069	-1.405
racewhite	-2.235	0.143	-15.616	-2.523	-1.961
hispanicnot hispanic	-0.376	0.102	-3.704	-0.576	-0.178
agegroup36-49	-0.289	0.088	-3.303	-0.461	-0.118
agegroup50-65	-0.369	0.088	-4.205	-0.542	-0.197
agegroup65+	-0.273	0.101	-2.706	-0.471	-0.075
incomegroupMedian or below	0.197	0.070	2.797	0.059	0.335
educationBA or lower	-0.222	0.098	-2.256	-0.415	-0.029
educationHigh school or Lower	-0.648	0.111	-5.835	-0.866	-0.431
stateicpalaska	-0.283	0.913	-0.310	-2.325	1.447
stateicparizona	0.262	0.323	0.811	-0.369	0.900
stateicparkansas	-0.236	0.475	-0.496	-1.185	0.686
stateicpcalifornia	0.900	0.281	3.205	0.354	1.458
stateicpcolorado	0.452	0.366	1.236	-0.263	1.173
stateicpconnecticut	1.519	0.426	3.566	0.701	2.378
stateicpdelaware	1.136	0.556	2.044	0.065	2.267
stateicpdistrict of columbia	2.428	1.088	2.231	0.666	5.377
stateicpflorida	0.265	0.284	0.932	-0.288	0.830
stateicpgeorgia	0.058	0.326	0.178	-0.578	0.702
stateicphawaii	0.829	0.522	1.589	-0.174	1.888
stateicpidaho	-0.159	0.521	-0.305	-1.220	0.840
stateicpillinois	0.679	0.300	2.261	0.095	1.275
stateicpindiana	0.385	0.343	1.123	-0.286	1.063
stateicpiowa	0.542	0.411	1.318	-0.265	1.353
stateicpkansas	-0.272	0.457	-0.595	-1.191	0.611
stateicpkentucky	0.502	0.354	1.417	-0.191	1.201
stateicplouisiana	0.202	0.380	0.532	-0.544	0.948
stateicpmaine	0.691	0.574	1.204	-0.445	1.830
stateicpmaryland	0.580	0.372	1.560	-0.143	1.317
stateicpmassachusetts	1.504	0.372	4.048	0.787	2.247
stateicpmichigan	0.773	0.317	2.439	0.156	1.402
stateicpminnesota	0.597	0.379	1.574	-0.144	1.346
stateicpmississippi	-0.330	0.463	-0.713	-1.242	0.580
stateicpmissouri	0.513	0.339	1.514	-0.149	1.183
stateicpmontana	0.531	0.599	0.887	-0.659	1.723
stateicpnebraska	0.497	0.580	0.857	-0.656	1.646
stateicpnevada	0.141	0.392	0.359	-0.627	0.912
stateicpnew hampshire	1.175	0.638	1.841	-0.057	2.490
stateicpnew jersey	0.564	0.316	1.786	-0.051	1.190
stateicpnew mexico	1.263	0.603	2.095	0.124	2.525
stateicpnew york	0.744	0.289	2.575	0.182	1.318
stateicpnorth carolina	0.442	0.312	1.418	-0.165	1.059
stateicpnorth dakota	-0.324	1.279	-0.254	-3.455	2.125
stateicpohio	0.558	0.300	1.862	-0.026	1.152
stateicpoklahoma	0.008	0.409	0.020	-0.799	0.808
stateicporegon	0.752	0.363	2.073	0.045	1.470
stateicppennsylvania	0.262	0.304	0.860	-0.331	0.864
stateicprhode island	1.404	0.901	1.557	-0.270	3.436
stateicpsouth carolina	-0.264	0.366	-0.721	-0.985	0.453
stateicpsouth dakota	0.311	0.648	0.480	-1.014	1.576
stateicptennessee	-0.323	0.365	-0.885	-1.041	0.392
stateicptexas	0.026	0.288	0.089	-0.536	0.598
stateicputah	0.159	0.458	0.346	-0.752	1.054

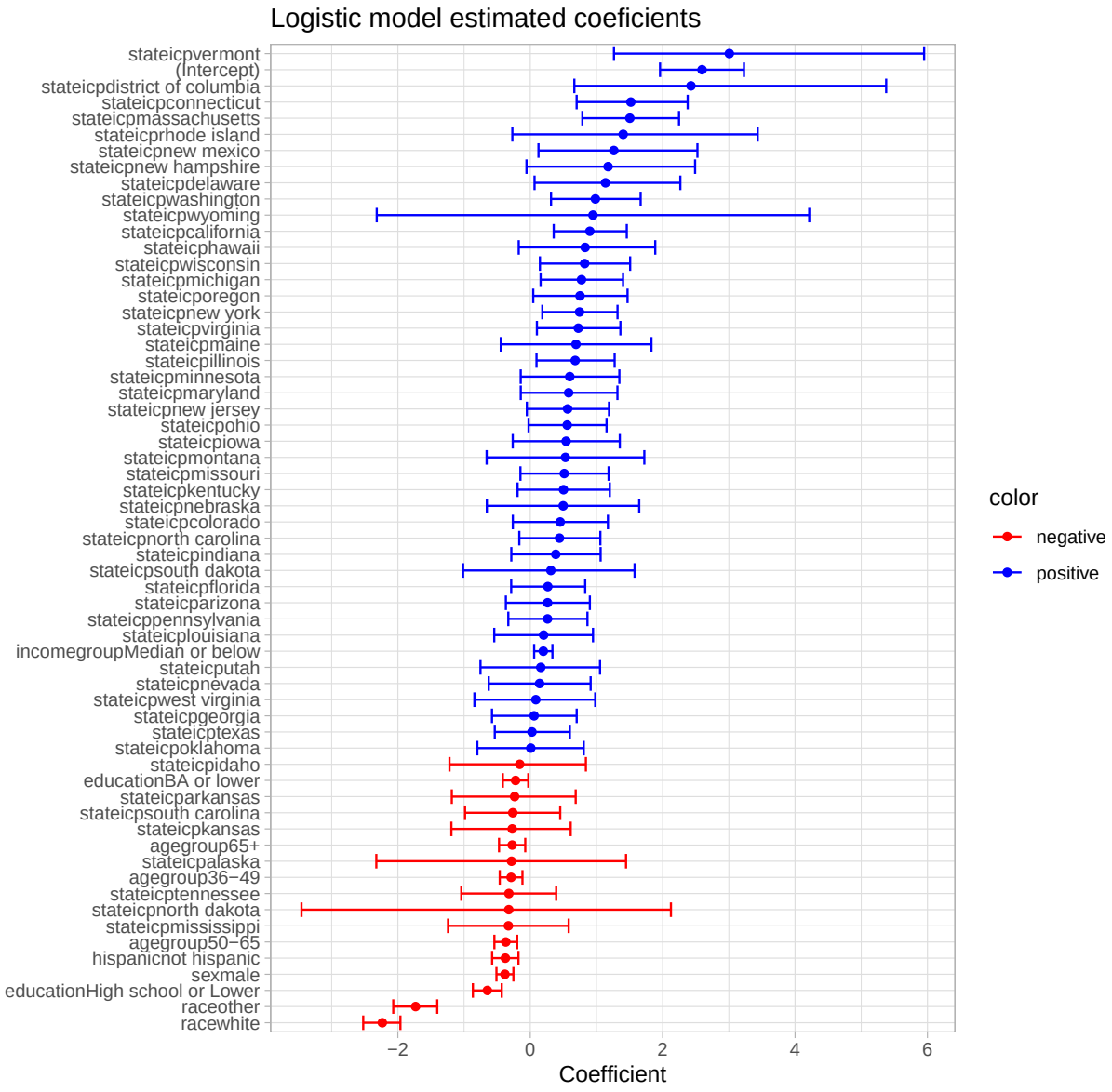

Figure 5: Logistic model estimated coeficients

Table 3: Final predictions in US for supports of Joe Biden

region	mean	low	upper
US	0.5248252	0.4727365	0.5769138
connecticut	0.7338181	0.6690070	0.7986292
maine	0.5065833	0.3818214	0.6313452
massachusetts	0.7298427	0.6769664	0.7827189
new hampshire	0.6290010	0.4964941	0.7615080
rhode island	0.6992810	0.5253918	0.8731702
vermont	0.9092795	0.8234377	0.9951213
delaware	0.6703004	0.5696073	0.7709935
new jersey	0.5519290	0.5059371	0.5979208
new york	0.5970611	0.5620241	0.6320981
pennsylvania	0.4300599	0.3888913	0.4712285
illinois	0.5627816	0.5227228	0.6028404
indiana	0.4593044	0.4039567	0.5146520
michigan	0.5624009	0.5164110	0.6083908
ohio	0.5094621	0.4696215	0.5493028
wisconsin	0.5452148	0.4873574	0.6030722
iowa	0.4760884	0.3976098	0.5545671
kansas	0.3155015	0.2376321	0.3933710
minnesota	0.4936178	0.4247046	0.5625311
missouri	0.4974876	0.4433233	0.5516519
nebraska	0.4788774	0.3544872	0.6032676
north dakota	0.2838756	0.0396659	0.5280852
south dakota	0.4234306	0.2829811	0.5638802
virginia	0.5976962	0.5520252	0.6433672
alabama	0.4487945	0.3918754	0.5057137
arkansas	0.3569104	0.2758879	0.4379328
florida	0.4821832	0.4482515	0.5161150
georgia	0.4893918	0.4440320	0.5347515
louisiana	0.5136076	0.4536458	0.5735694
mississippi	0.4412596	0.3662756	0.5162435
north carolina	0.5360630	0.4930803	0.5790456
south carolina	0.3989521	0.3444435	0.4534607
texas	0.4351170	0.3995892	0.4706448
kentucky	0.4833656	0.4239206	0.5428106
maryland	0.6039323	0.5456712	0.6621933
oklahoma	0.3875884	0.3159868	0.4591901
tennessee	0.3388009	0.2859838	0.3916181
west virginia	0.3669217	0.2813946	0.4524488
arizona	0.4584264	0.4083015	0.5085513
colorado	0.4899606	0.4252124	0.5547089
idaho	0.3240489	0.2270000	0.4210979
montana	0.4746414	0.3431723	0.6061104
nevada	0.4485476	0.3788434	0.5182517
new mexico	0.7012546	0.5902425	0.8122668
utah	0.4109558	0.3208997	0.5010118
wyoming	0.5805643	0.2474358	0.9136928
california	0.6417018	0.6097563	0.6736473
oregon	0.5463707	0.4826303	0.6101111
washington	0.6122564	0.5569719	0.6675410
alaska	0.3234547	0.1409099	0.5059994
hawaii	0.6245466	0.5219732	0.7271201
district of columbia	0.9191276	0.8433855	0.9948696

that both Biden and Trump want to win. For states like connecticut, the low voting percentage is already much above 50% which is about 67% for Biden, these states are supposed to vote Biden for sure, but for other states like west virginia, the upper voting percentage is already much below 50% which is about 45% for Biden, these states are supposed not to vote Biden for sure, these results are consistent with facts that some states are consistently vote the parties in history of US election.

Figure 6 shows the prediction outcome of vote percentage win by Joe Biden across states in 3 situations: average, worst and best. For average situation, we use mean predicted probability for Biden across states, for worst situation, we use lower bound predicted probability for Biden across states and for best situation, we use upper bound predicted probability for Biden across states. Clearly, we can find that even in worst situations, there are states will vote for Biden and even in best situations, there are states will not vote for Biden.

Overall, table 3 shows that Joe Biden predicted to win with 52.5% vote +/- 5.2% confidence by the multilevel logistic regression with post-stratification in US.

5 Discussion

5.1 Votes with confidence intervals

This study not only give estimates and forecasts for vote percentage of Biden across states and in whole US, this study also give confidence intervals correspondingly. Cconfidence intervals of forecastings are important due to they show how we are confident about our forecasting results.

As the table 3 shows that Joe Biden predicted to win with 52.5% vote +/- 5.2% confidence by the multilevel logistic regression with post-stratification in US. Actually, the 95% confidence interval of the overall outcome for Joe Biden in US includes 50% which is in the range (47.3%, 57.7%). This is a wide interval indicating the all possible outcomes that Joe Biden lose or win the US 2020 election. It indicates the competition would be very close in fact between Biden and Trump. Considering the new event COVID-19 started in the end of 2019, the unforecasting issues become more and more which could affect the final outcome of US 2020 election, and make the outcome hard to be predicted.

5.2 Votes by states

For states, table 3 shows the results in details combined with the Figure 6 which shows the average, worst and best situations for Biden. It is already know states in US have consistently voting behaviors, lots of former studies such as Levendusky and Pope (2011) discussed the red states and blue states, Donald Trump who won the 2016 election represented the Republican party while the vice-president of Barack Obama

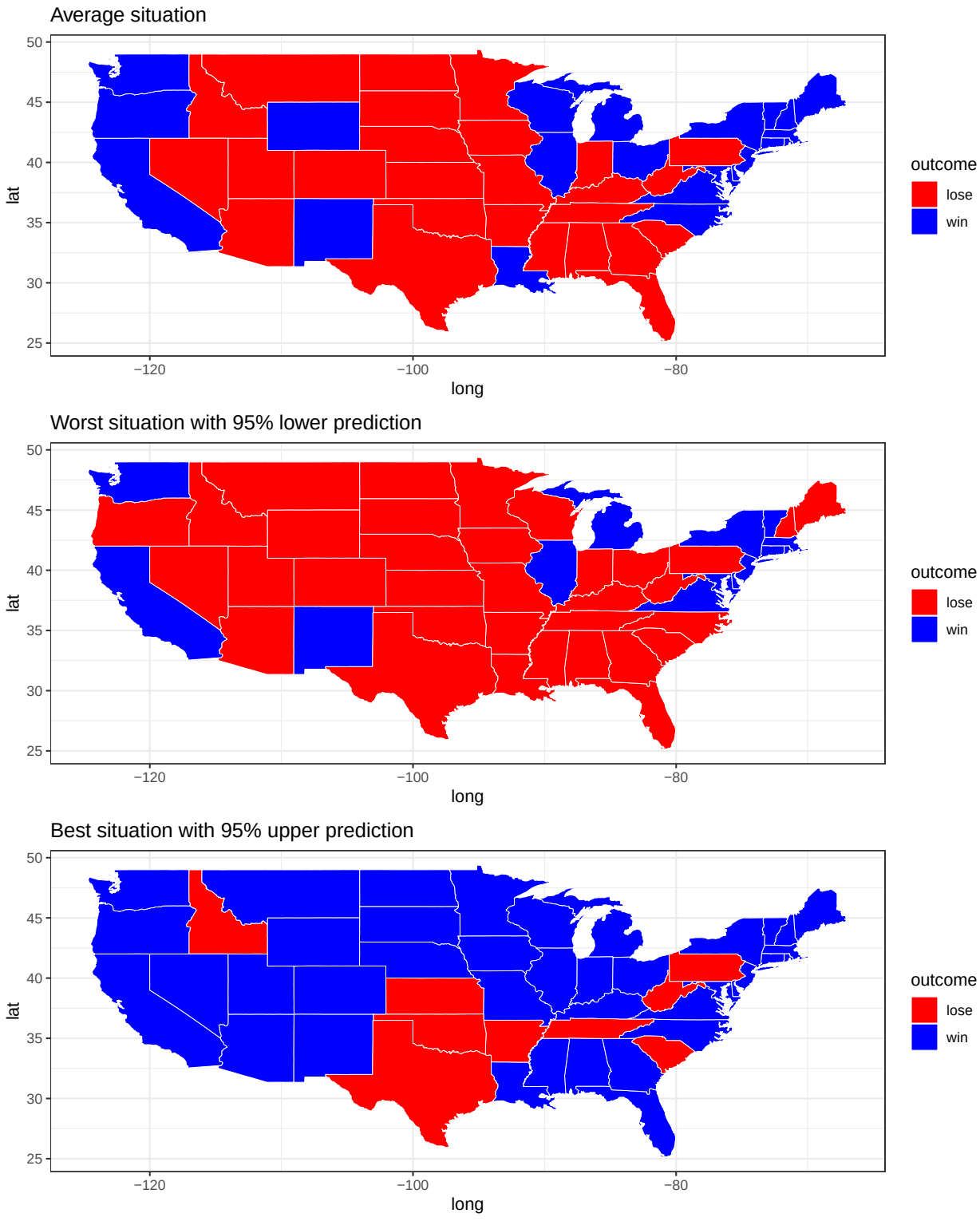

Figure 6: Prediction outcome of vote percentage win by Joe Biden across states

represented the Democratic party. The red states are the states carried by Republicans while the blue states are the states carried by Democrats. In history, voters in red states and blue states are deeply polarized, they have consistent vote behaviors. Red states like Alaska voted George H. W. Bush, George W. Bush, John McCain and Donald Trump in history for Republican candidates while blue states like District of Columbia voted Bill Clinton, Barack Obama, John Kerry and Hillary Clinton for Democratic candidates. The vote behaviors are almost not changed.

Focused on the average situation using the mean forecasts by states in Figure 6, it can be seen Joe Biden mainly won supports from states in the west coast and north-eastern coastal states while Trump mainly won supports in the middle states. For the worst and best situations, Joe Biden would be lost or win for sure as the map would be either red or blue overall. But the two situations are almost not possible to happen.

5.3 Votes by demographics

For this study, the key variables used are age, sex, race, education, income, hispanic or not. Based on the table 2 and figure 5 results, it could be found that males tend to be more willing to vote for Trump instead of Biden, fixed others, the odds of voting for Biden among males is about 32% lower compared with females. Due to confindence interval does not include 0 indicating it is statiscally significant in affecting the voting result. It could be found that whites tend to be more willing to vote for Trump instead of Biden, fixed others, the odds of voting for Biden among whites is about 90% lower compared with blacks. This might due to Trump policy is nicely for whites but not nicely for blacks. The confindence interval does not include 0 indicating race is statiscally significant in affecting the voting result.

Compared with young age group 18-35, the older age groups all show negative coefficients indicating that older voters are more likely to vote Trump but young ones are more likely to vote Biden, this result might be related with heath care of old people by Trump. Also, hispanic voters supports Biden more than non-hispanic voters. The higher income group of voters are more likely to vote Trump while poorer ones are more likely to vote Biden. For the education level, the higher the level, the high chance to vote Biden. At last, across states, the estimated coefficients indicate the signs of supports for Biden or Trump, they are consistent with the results from table 3.

5.4 Weaknesses and next steps

There are some weakness for this study. First based on experiences in past US election such as 2016 US election, the predictions of models or other ways could be inaccurate, Trump won the 2016 US election which was not predicted by most forecasts before the outcome. So there are limitations of the models in forecasting vote outcome in 2020 US election or in similar elections.

The survey data has weaknesses that there might be selection bias and non-response bias, especially in the COVID-19 period, survey might be affected by COVID-19 there could be missing values and non-responses. The model also depend on the survey data seriously, the post-stratification needs to be consistent with survey data. Data cleanning procedures used in this study might not be the most appropriate. For examples, age was divided into 4 groups, income level was divided into 2 groups, education level was divided into 3 groups in this study. All of the results could be changed under another cleanning plans which could lead to totally different forecasts on the post-stratification data.

The could be omitted important variables bias in the model results, there might be factors not considered in the study such as employment status, martial status, job types and so on. Also, policies like how to treat COVID-19 could affect the results too, for example, Trump not require a wearing mask policy could be an important factor in the COVID-19 pandemic period in the US 2020 election.

In future studies, more other important variables like employment status, martial status, job types could be considered in the models to give better estimates of voters' voting behaviors. More advanced models could be applied to compare with the model used in this study. It would be expected that the results could be improved in future studies based on all of the work in this study.

Appendix

A Additional details

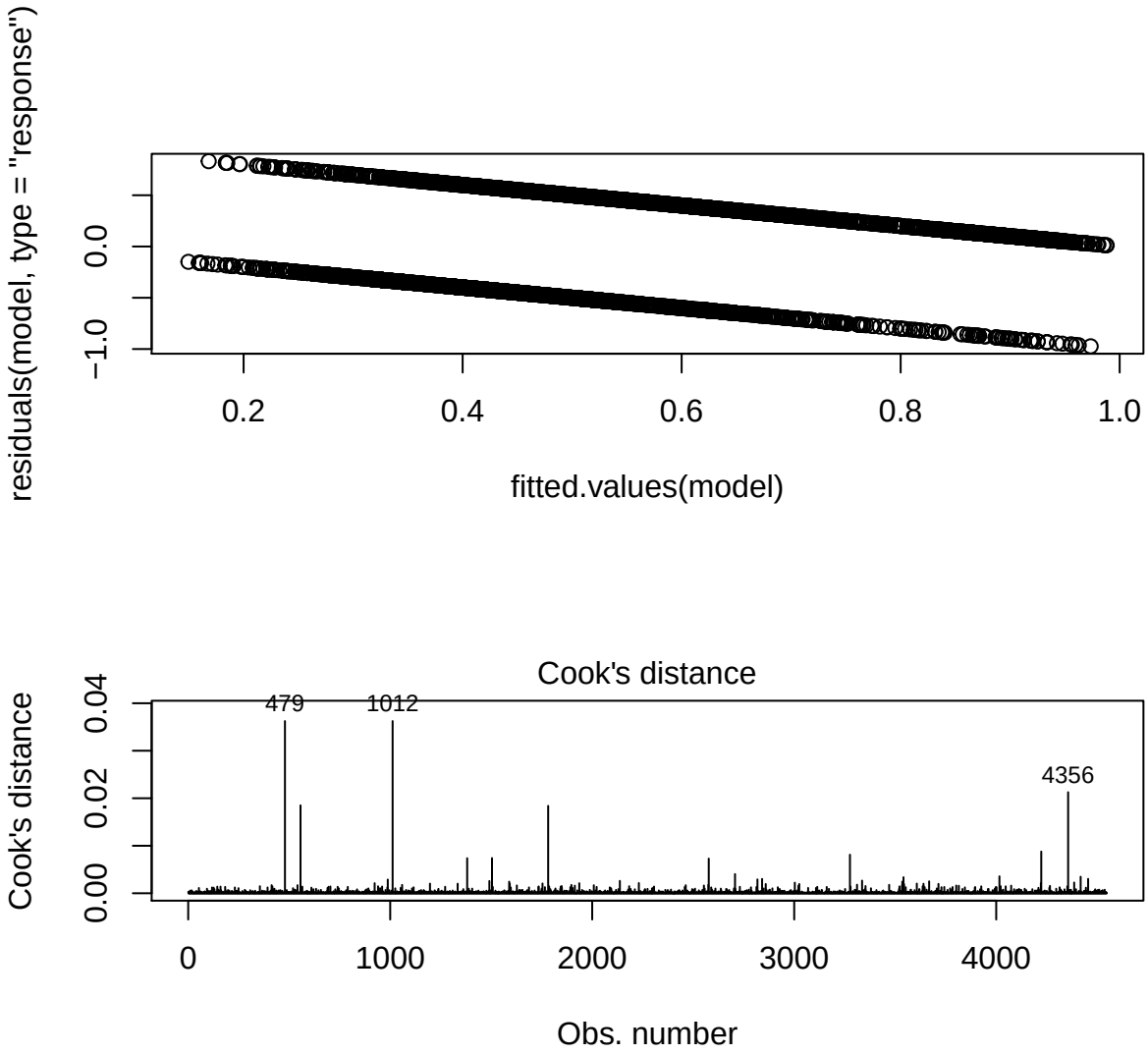

Figure 7: Model diagnostics

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. *Rmarkdown: Dynamic Documents for r.* https://github.com/rstudio/rmarkdown.

Deckman, Melissa, and Erin Cassese. 2021. "Gendered Nationalism and the 2016 US Presidential Election: How Party, Class, and Beliefs about Masculinity Shaped Voting Behavior." *Politics & Gender* 17 (2): 277–300.

Greenwald, Anthony G, Colin Tucker Smith, Nilakanta Sriram, Yoav Bar-Anan, and Brian A Nosek. 2009. "Implicit Race Attitudes Predicted Vote in the 2008 US Presidential Election." *Analyses of Social Issues and Public Policy* 9 (1): 241–53.

Levendusky, Matthew S, and Jeremy C Pope. 2011. "Red States Vs. Blue States: Going Beyond the Mean." *Public Opinion Quarterly* 75 (2): 227–48.

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

Richard A. Becker, Original S code by, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka, and Alex Deckmyn. 2021. *Maps: Draw Geographical Maps.* https://CRAN.R-project.org/package=maps.

Robinson, David, Alex Hayes, and Simon Couch. 2021. *Broom: Convert Statistical Objects into Tidy Tibbles.* https://CRAN.R-project.org/package=broom.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2020. *IPUMS USA: Version 10.0 [ACS 2018].* Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0.

Scheibe, Susanne, Rui Mata, and Laura L Carstensen. 2011. "Age Differences in Affective Forecasting and Experienced Emotion Surrounding the 2008 US Presidential Election." *Cognition & Emotion* 25 (6): 1029–44.

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

Valentino, Nicholas A, Carly Wayne, and Marzia Oceno. 2018. "Mobilizing Sexism: The Interaction of Emotion and Gender Attitudes in the 2016 US Presidential Election." *Public Opinion Quarterly* 82 (S1): 799–821.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.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.

Wickham, Hadley, Romain Fran?ois, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data*

Manipulation. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Wilke, Claus O. 2020. *Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'*. https://CRAN.R-project.org/package=cowplot.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.