Forecasting the 2020 U.S. Presidential Election: Biden will Defeat Trump By A Slim Margin

Boyu Sheng,Jiaxiang Miao,Qiyue Zhang,Tianxiao Ma

02 November, 2020

Abstract

Keywords: forecasting; Multilevel Regression with post-stratification; US 2020 election; Trump; Biden

Thanks to the electoral college system, Republican Donald J. Trump won the presidential election in the year of 2016 and defeated Democrat candidate Hillary Clinton even though Hillary won the popular vote by a slim margin. However, as the difference between the popular approval rate of two candidates Joe Biden and Donald Trump increases and the constituencies are getting polarized, it is hard to tell whether or not Donald Trump will be re-elected when he has a relatively lower approval rate. Hence in our research, we used the multilevel regression with post stratification and found out that Biden will have a winning rate of 50.13% with a margin of error of 5% and Trump will lose the election by a very slim proportion about 0.26%. Such findings forecast the result of the 2020 US presidential election and furthermore tell the trending ideology of the general American people is going more left or progressive rather than conservative so that it will indirectly show the possible policies introduced by the US government in the near future.

Contents

1 Introduction

For millions of Americans, November 3rd, 2020 will be an election day that is anything but similar to the ones before. In fact, for many Americans, election day has already happened—due to the global COVID-19 pandemic, an unprecedented number of voters have casted their votes through mail-in ballots or in the early-voting process.

"United" in name, the U.S. politics has become increasingly chaotic and divisive in recent years. The 2020 presidential race, coming down to current president Donald J. Trump and former vice-president Joe R. Biden, is deemed to be one of the most consequential elections in many people's lifetimes. Americans have a crucial choice to make (and many have already made the choice at the time of writing) after a summer of reckoning on systematic racism, a wildfire season that ravaged the West Coast which is an ecological catastrophe resulted by Climate change, not to mention the ongoing pandemic and its massive impact on the healthcare system and economy.

Current polls and predictions are showing on the national level and many battleground states, Joe Biden is the leading candidate. To assess the seemingly high voter support of Biden, especially which factor affect the decision the most, we use the results of Democracy Fund + UCLA Nationscape survey on June 25th, 2020 to build a prediction model based on people's current candidate preference as the dependent variable, and their 2016 vote, with their demographic background as explanatory variables.

Key demographic divisions reflected in the contemporary American political landscape are what we are focusing on in our model. R Core Team (2020) is used to create Data cleaning and model code. We include the education level variable, which is simplified into two groups: people who have received some college education and those who have not. This division has been an important voting predictor since at least the 2016 general election. Age, gender, race, Hispanic identity and geography have also been considered in our model.

Yet we should also acknowledge the fact that by using one poll to build our prediction on, we inevitably amplify the bias in the results of that one poll. To reduce the bias while avoiding complicating our model-building process, we decided to adapt the post-stratification method: using more representative data on our population to reweight the model results. In this study, we used the results of the 2018 American Community Survey to post-stratify. On the national popular vote, besides the general forecast, we also look into the specific demographic groups. Meanwhile, regarding the electoral college vote, while our binary outcome setup restricts our ability to give a prediction.

Ultimately, regardless of our or others' predictions, we will know the results on Nov 3rd (or more likely some days, even weeks later given the number of mail-in ballots this year). But we do hope our model and its results capture the current US political landscape and forecast the outcome as close as possible.

2 Data

2.1 Individual-level survey data

Individual-level survey data originally adapted from The June 25 to July 01 Nationscape wave 50 survey results of Democracy Fund + UCLA Nationscape project [Tausanovitch, Chris and Lynn Vavreck. 2020]. The Nationscape is a 16-month election study conducted by UCLA researchers, the project conducting interviews of roughly 6,250 Americans per week from July 10, 2019 through December 2020, covering the 2020 campaign and election.

The target population included all adults living in the US. The desired number of respondents is 6250 per week while the actual number of respondents collected on the week of June 25th, 2020 was 6479.

Survey samples are provided by Lucid Inc., a market research platform that runs an online exchange for survey respondents. Lucid directly sends respondents to survey software operated by the Nationscape team. Respondents were interviewed in English and the online questionnaires are designed for a 15 minutes median administration time. The frame for individual-level survey is a list of Americans adults registered with Lucid Inc. platform. To avoid non-responses and response errors, all respondents must complete an attention check before taking the survey. In addition, Nationscape removes the survey that is completed in fewer than 6 minutes and respondents selecting the same response for every question. The overall participation rate is around 75% for all waves. Such a design acts as a strength Nationscape survey since the participation rate of Nationscape survey exceeds other online survey participation rate significantly.

However, non-sampling errors still occurred as a weakness when the respondent did not understand or misinterpreted a question, or could not recall the requested information. Response Bias also acts as a weakness because the respondent's unwillingness to answer the questions honestly.

Then the frame is reduced by convenience sampling, which is a type of non-probability sampling involving the sample being drawn from that part of the population that is convenient for Lucid to reach out. Then this convenience sample was selected on a set of demographic criteria to be representative of the American population.

The weight of the survey sample is adjusted to be a purposive sampling that is derived by directly comparing responses in the Nationscape survey with responses to the 2018 American Community Survey(ACS) of the U.S. Census Bureau,with 2016-vote as an exception which is derived from the official election results released by the Federal Election Commission.

A subset of the Nationscape dataset including demographics was selected and retrieved for this investigation. We use the results of Democracy Fund + UCLA Nationscape survey on June 25th, 2020 to build a prediction model based on people's current candidate preference as the dependent variable, and their 2016 vote, with their demographic background as explanatory variables.

Key demographic divisions reflected in the contemporary American political landscape are what we are focusing on in our model. Table 1 shows the variables with descriptions used in the subdataset, the subset is cleaned by selecting wanted variables and reconstructing some variables to help build a model. Vote_2016 and Vote_2020 are two variables that only exist in the Nationscape surveym where Vote_2016 indicates who did the respondent vote for in 2016; Vote_2020 shows who the respondent voted for if the election for president is held today. To be noticed, gender variables and five reconstructed variables are included in both individual-level survey data and Post-stratification data. Below is the explanation of 5 reconstructed variables.

The education level variable, which is a binary variable containing two groups of respondents: people who have received some college education(1) and those who have not(0). This division has been an important voting predictor since at least the 2016 general election. The constructed age variable refers to seven age groups, without respondents under 18 years old and over 84 years old: "18" refers to respondents aged from 18-24, and the rest age group contains respondents with 10 years age differences such as "25" represents respondents with age 25-34. Regrouping age groups instead of every single age is more convenient when building a model. In addition, the hispanic identity variable is constructed into binary variables where 1

Table 1: Variables in Democracy Fund + UCLA Nationscape survey collected from June 25 - July 01, 2020

Variables	Description
state	51 states + DC in postal abbreviation
vote_2016	The candidate that the respondent voted for in 2016
vote_2020	If the election for president is hold now(2020), who will the respondent vote for
gender	Gender of the respondent
hisp	If the respondent is hispanic, 1 = yes, 0= No
race	Race: Native, Black, White, API ("Asian or Pacific Islander"), Other
edu	Education level, 1 = has received some college education, 0 otherwise
age	Age group: "18" means 18-24, "25" means 25-34, and so on.
reg	Registered to vote, 1 = yes,0 = no
intention	Intention to vote in 2020 election

Table 2: Age group proportion in Nationscape wave 50 survey

age	count	proportion
18	782	0.1206976
25	1219	0.1881463
35	1397	0.2156197
45	948	0.1463189
55	1071	0.1653033
65	855	0.1319648
75	207	0.0319494

represents the respondent meets the condition, 0 otherwise. State variable here represents 51 states + DC in postal abbreviation(Details in Appendix 1). Race variable is regrouped into five groups: Native(american indian or alaska native), Black(black/african american/negro), White(White), API (Chinese, Japanese and other asian or Pacific Islander), and Other.

Table 2-5 are frequency tables for variable age group, education level, Hispanic identity, and gender respectively. In the Democracy Fund + UCLA Nationscape wave 50 survey, there are about two thirds respondents who have taken at least some post-secondary education, and respondents are mainly composed by people who are 25-44. Besides it, the majority of respondents have a non-hispanic identity. Female respondents compose 51.07% of the respondent population.

Due to the length of the frequency table for the state variable, see Appendix 4 Table 15 for the frequency table of state distribution in Nationscape survey. Among these respondents, 7.4% people from the state of Texas(TX) compose the largest proportion while only 0.77% respondents come from the State of Wyoming(WY) compose the smallest proportion.

Table 6 is a frequency table showing respondents' 2020 candidate preference focusing on Biden and Trump only. Biden gains a slight advantage than Trump.

Table 3: If the respondent's education level meets some college degree, No(0) vs Yes(1) in Nationscape wave 50 survey

edu	count	proportion
0	2078	0.3207285
1	4401	0.6792715

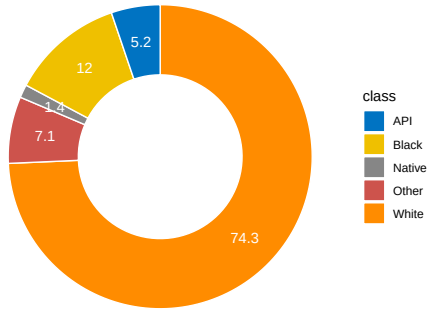
Table 4: If the respondent's ethnical/cultural identity is hispanic. No vs Yes in Nationscape wave 50 survey

hisp	count	proportion
0	5498	0.8485877
1	981	0.1514123

Table 5: Gender proportion in Nationscape wave 50 survey

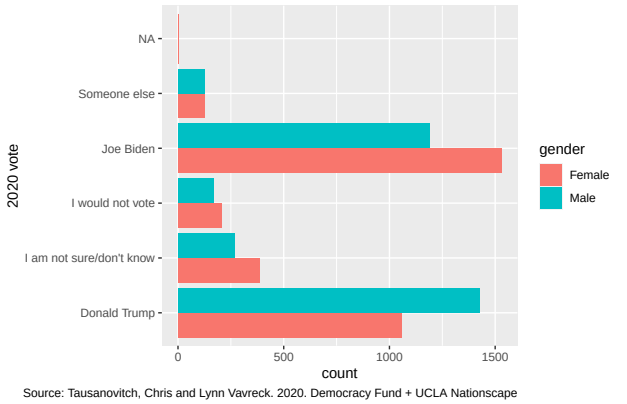
gender	count	proportion
Female	3309	0.510727
Male	3170	0.489273

Figure 1: Race distribution in Nationscape wave 50 survey



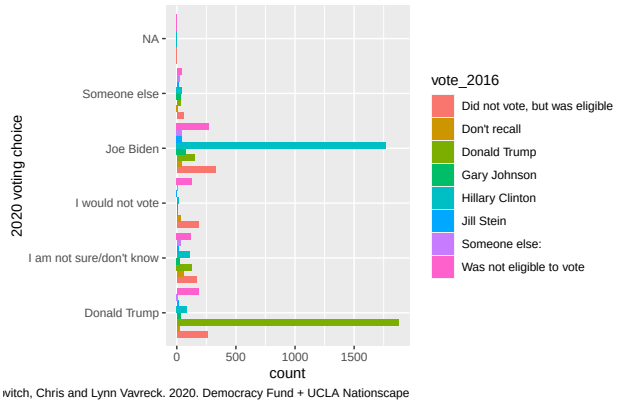
ce: Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape

Figure 2: 2020 voting choice based on gender in Nationscap



Source: Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape

Figure 3: How 2016 voters move their choices in Nationscap



witch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape

The donut chart in figure 1 displays the race distribution in percentage among the respondents. White respondents take up the largest proportion and native respondents take up the smallest proportion.

Figure 2 illustrates the 2020 candidate preference respondents answered grouped by respondent's gender.

Table 6: 2020 voting preference focusing on two candidates - Biden and Trump in Nationscape wave 50 survey

candidate	count	proportion
Biden	2719	0.5228846
Trump	2481	0.4771154

We can observe female respondents have a preference on Biden while male respondents are more likely to vote for Trump.

Figure 3 shows how respondents move their candidate preference from 2016 to 2020, where the color difference represents 2016 voting choices. The majority respondents who voted for Hillary Clinton in 2016 shift to Joe Biden in 2020. Most Respondents voted for Trump in 2016 still hold their ground for Trump in 2020. However, there is a small group of respondents who voted for Trump in 2016 and shifted to Biden in 2020.

2.2 Post-stratification data

Post-stratification data is the 2018 American Community Survey(ACS) conducted by the U.S. Census Bureau, obtained from IPUMS USA. (IPUMS USA,2018 ACS). The Primary purpose of ACS is to measure the changing social and economic characteristics of American population yearly. ACS collects data through sending questionnaires that design on the Internet and paper form along with an instruction hyperlink/booklet on a monthly basis to selected American households through either the paper or internet that accepts principles of respondent friendliness and navigation. Respondents are able to call a toll-free TQA line answering questions as well.

If the Census Bureau is unable to get a response by Internet, mail, or TQA, then one out of three non-respondent households may be selected for an in-person interview. The proportion of non responding households selected for in-person interviews is higher in areas with lower predicted response rates.

The sampling unit for ACS is the household and all persons residing in the household in the US. Sampling frame for ACS is a national Master address file (MAF: see details in Appendix 2). Census Bureau selects a systematic sample of addresses from the most current MAF on a monthly basis to use as the ACS sample. Target population for ACS is all U.S. residents. The sample size in the subset of 2018 ACS for the analysis is 11733.

This systematic sampling approach eliminates the phenomenon of clustered selection and a low probability of contaminating data. However, the weakness of 2018 ACS data subjects to sampling error since the data is derived from a sample of the population. Nonresponse error occurs since some respondents provide data considered as invalid or inconsistent with other answers, or simply skip answers. Measurement error is another weakness if the vague/ambiguous questions misinterpreted by respondents.

A subset of the 2018 American Community Survey obtained from IPUMS USA including demographics was selected and retrieved to post-stratify. Variables used in the subset contain 'state', 'gender','hisp','race','edu' and 'age'. Education level('edu'), age group('age'). hispanic identity('hisp'), Race('race'), State('state') and gender('gender') variables are constructed to keep consistent with variables in individual-level survey data. Details see explanations of the individual-level survey located above Table 1.

Table 7 shows the variables with descriptions used in the sub dataset based on 2018 ACS, the subset is cleaned by selecting wanted variables and reconstructing some variables to help build a model.

Table 8-11 are frequency tables for variable age group, education level, Hispanic identity and gender gender respectively. In the subset of 2018 ACS data, there is an even distribution trend of each group within gender, education and age variables. Respondents who have taken at least some post-secondary education account for about 50.7% and who have not compose 49.3%. Without the consideration of people who did not answer their age in the 2018 ACS questionnaire, the population of each age group is almost evenly distributed. The number of respondents in each state is almost the same. The ratio of male respondents to female respondents is basically the same. However, respondents with non-hispanic ethical/cultural identity account for 58% while the percentage for respondents with hispanic identity is only 42% .

Due to the length of the frequency table for the state variable, see appendix 5 16 for the Frequency table of state distribution in a subset of 2018 ACS survey. We can observe the number of respondents in each state is almost the same with very similar proportions.

Table 7: Variables in ACS 2018 as Post-stratification data

Variables	Description
state	51 states + DC in postal abbreviation
gender	Gender of the respondent
hisp	If the respondent is hispanic, 1 = yes, 0= No
race	Race: Native, Black, White, API ("Asian or Pacific Islander"), Other
edu	Education level, 1 = has received some college education, 0 otherwise
age	Age group: "18" means 18-24, "25" means 25-34, and so on.

Table 8: Age group proportion in 2018 ACS

age	count	proportion
18	1726	0.1471065
25	1801	0.1534987
35	1751	0.1492372
45	1723	0.1468508
55	1685	0.1436120
65	1589	0.1354300
75	1458	0.1242649

Table 9: If the respondent's education level meets some college degree, No(0) vs Yes(1) in 2018 ACS

edu	count	proportion
0	5952	0.5072871
1	5781	0.4927129

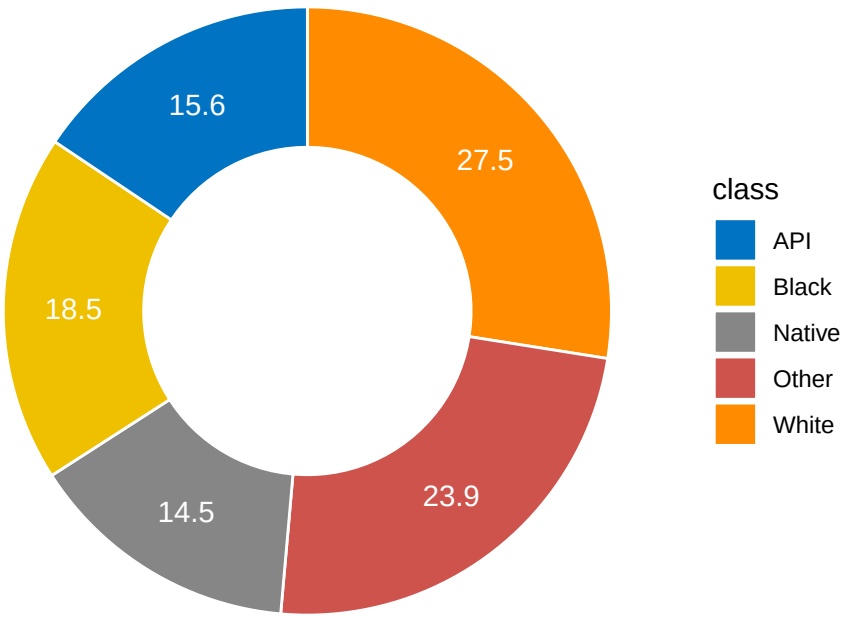
Table 10: If the respondent's ethnical/cultural identity is hispanic. No vs Yes in 2018 ACS

hisp	count	proportion
0	6807	0.5801585
1	4926	0.4198415

Table 11: Gender proportion in 2018 ACS

gender	count	proportion
Female	5896	0.5025143
Male	5837	0.4974857

Figure 4: Donut chart of Race distribution in 2018 ACS



Source: IPUMS USA, University of Minnesota, www.ipums.org.

The donut chart in figure 4 displays the race distribution in percentage among the respondents in the subset of 2018 ACS. White respondents take up the largest proportion, 24.2%, and API respondents account for the smallest proportion, 16.28%.

3 Model

3.1 Multilevel Logistic Regression Model

We are interested in predicting the popular vote outcome for Trump of the 2020 American presidential election by using a multilevel logistic regression based on the factors of age group, gender, education level, hispanics, and race for different intercept for each state.

Multilevel logistic regression model:

$$\log(\frac{p_{ij}}{1-p_{ij}}) = \beta_0 + \beta_1 x_{ij}^{age25\ 34} + \beta_2 x_{ij}^{age35\ 44} + \beta_3 x_{ij}^{age45\ 54}$$

$$+\beta_4 x_{ij}^{age55\ 64} + \beta_5 x_{ij}^{age65\ 74} + \beta_6 x_{ij}^{age75above} + \beta_7 x_{ij}^{genderM} + \beta_8 x_{ij}^{edu}$$

$$+\beta_9 x_{ij}^{hisp} + \beta_{10} x_{ij}^{raceBlack} + \beta_{11} x_{ij}^{raceNative} + \beta_{12} x_{ij}^{raceOther} + \beta_{13} x_{ij}^{raceWhite}$$

where $i = 1, ..., 11733$(number of observations), $j = 1, ..., 51$(number of states),

p represents the probability of voting for Trump,

$\frac{p}{1-p}$ represents the odds of voting for Trump,

β_1 coefficient represents the change in log odds for respondent age between 25 to 34,

β_2 coefficient represents the change in log odds for respondent age between 35 to 44,

β_3 coefficient represents the change in log odds for respondent age between 45 to 54,

β_4 coefficient represents the change in log odds for respondent age between 55 to 64,

β_5 coefficient represents the change in log odds for respondent age between 65 to 74,

β_6 coefficient represents the change in log odds for respondent age between 75 and above,

β_7 coefficient represents the change in log odds for male respondents,

β_8 coefficient represents the change in log odds for respondent who has received some college education,

β_9 coefficient represents the change in log odds for respondent who is hispanic,

β_{10} coefficient represents the change in log odds for respondent whose race is black,

β_{11} coefficient represents the change in log odds for respondent whose race is native,

β_{12} coefficient represents the change in log odds for respondent whose race is others,

β_{13} coefficient represents the change in log odds for respondent whose race is white,

x represents each factor respectively.

The general aim of multilevel logistic regression is to estimate the odds that an event will occur (the yes/no outcome) while taking the dependency of data into account. Practically, it will allow you to estimate such odds as a function of lower level variables, higher level variables, and the way they are interrelated (cross-level interactions). Specifically, a multilevel logistic regression can be used when the outcome variable describes the presence/absence of an event or a behavior.[A]

We considered using a linear regression model, but finally we kept with logistic regression model. As in our situation, the response variable, whether vote for Trump, is a binary variable where 1 represent vote for Donald Trump and 0 represent vote for Joe Biden. Thus, it is more appropriate to use a logistic regression model in predicting the popular vote outcome for Trump of the 2020 American presidential election. Moreover, layers were added to our model, so we can get different intercepts for each state in order to make a preciser prediction.

Whereas linear regression gives the predicted mean value of an outcome variable at a particular value of a predictor variable, logistic regression gives the conditional probability that an outcome variable equals one at a particular value of a predictor variable (e.g. the likelihood of vote for Trump for a male hispanic respondent aged at 35 who has received some college education with race native).

The expression on the left hand side of the equation is often called logit function, it is used to predict such a probability. In our model, it describes the relationship between a series of predictor variables(age, gender, education level, hispanics, and race) and the conditional probability that an outcome variable Y_i equals one(vote for Trump). Also, a multilevel regression is used to smooth noisy estimates in the cells with too little data by using overall or nearby averages.[B]

We use `glmer()` from `lme4` package in `R` to fit the model to our data. We use `as.factor()` for age variable, because even though age is numerical in the original dataset, but it becomes categorical as we group them into age groups during the data cleaning process. For each categorical variable(age, gender, education level, hispanic, race) with n levels, we need $n-1$ dummy variables to fully study its influence on our response variable(vote for Trump).

3.2 Post-stratification Calculation

$$\hat{y}^{PS} = \frac{\sum N_j \hat{y_j}}{\sum N_j}$$

where $\hat{y_j}$ is the estimate in each cell and $\sum N_j$ is the population size of j^{th} cell based off demographics.

Table 12: Summary of Model Estimates

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.1656	0.1799	-6.4803	0.0000
as.factor(age)25	0.4406	0.1276	3.4535	0.0006
as.factor(age)35	0.8163	0.1245	6.5565	0.0000
as.factor(age)45	0.8924	0.1323	6.7453	0.0000
as.factor(age)55	0.6717	0.1285	5.2289	0.0000
as.factor(age)65	0.6051	0.1331	4.5446	0.0000
as.factor(age)75	0.8771	0.1896	4.6253	0.0000
genderMale	0.4314	0.0613	7.0382	0.0000
edu	-0.3794	0.0686	-5.5281	0.0000
hisp	-0.4327	0.1005	-4.3042	0.0000
raceBlack	-1.4042	0.1940	-7.2366	0.0000
raceNative	0.9512	0.3073	3.0957	0.0020
raceOther	0.3219	0.1956	1.6455	0.0999
raceWhite	0.8187	0.1462	5.6007	0.0000

Table 13: Vote Outcome Prediction for Trump of the 2020 American Presidential Election

alp_predict
0.4987

In order to estimate the proportion of voters who will vote for Donald Trump, we performed a post-stratification analysis. We make predictions using our model above with census data from 2018 ACS, specifically estimate y from each cell using our multilevel model, meaning use demographics to extrapolate how the entire population will vote.

Response recorded basic demographics: age(7 categories), gender(2 categories), education level(2 categories), hispanics(2 categories), race(5 categories), thus partitioning the data into 280 cells.

We weight each proportion estimate by the respective population size and sum those values and divide by the entire population size. The post-stratification weights are a sophisticated weighting strategy that help to reduce sampling error and potential non-response bias.(ESS Methodology)

4 Result

4.1 Model Result

Multilevel logistic regression model estimates by interpreting regression coefficients:

$$\log\left(\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right) = -1.17 + 0.44x_{ij}^{age25\ 34} + 0.82x_{ij}^{age35\ 44} + 0.89x_{ij}^{age45\ 54}$$

$$+0.67x_{ij}^{age55\ 64} + 0.61x_{ij}^{age65\ 74} + 0.88x_{ij}^{age75above} + 0.43x_{ij}^{gender} - 0.37x_{ij}^{edu}$$

$$-0.43x_{ij}^{hisp} - 1.40x_{ij}^{raceBlack} + 0.95x_{ij}^{raceNative} + 0.32x_{ij}^{raceOther} + 0.82x_{ij}^{raceWhite}$$

4.2 Post-stratification Calculation Result

The result table 14 of Vote Outcome Prediction for Trump of the 2020 American Presidential Election by State.

Table 14: Vote Outcome Prediction for Trump of the 2020 American Presidential Election by State

state	alp_predict
AK	0.5688
AL	0.5231
AR	0.5982
AZ	0.5493
CA	0.4039
CO	0.5387
CT	0.3949
DC	0.3286
DE	0.4519
FL	0.5166
GA	0.4987
HI	0.4209
IA	0.5396
ID	0.6015
IL	0.4741
IN	0.5455
KS	0.5779
KY	0.5475
LA	0.4711
MA	0.3960
MD	0.4217
ME	0.5365
MI	0.4777
MN	0.5313
MO	0.5224
MS	0.4751
MT	0.5681
NC	0.4804
ND	0.5747
NE	0.5525
NH	0.5365
NJ	0.4749
NM	0.4791
NV	0.5243
NY	0.4477
OH	0.5221
OK	0.5725
OR	0.5016
PA	0.5811
RI	0.5037
SC	0.5534
SD	0.5776
TN	0.6047
TX	0.5737
UT	0.5739
VA	0.4440
VT	0.4893
WA	0.4735
WI	0.4964
WV	0.5904
WY	0.5589

5 Discussion

5.1 Implication

Although the election is going to be tight, Biden, as we analyzed, is going to win the election. From the above statistical summary output for our model, we can conclude that there are only three factors that have negative correlations with Trump's election result, especially African Americans who have the strongest negative impact. From which we can further conclude that Donald Trump has tougher and more stringent policies about racial groups and the strategy that Trump and the Republicans use is to mainly attract white voters by giving them more interests and reducing the privileges of the racial minorities. While Joe Biden's policies are in favour of racial minorities who are particularly triggered by the recent police brutality against black people and the "Black Lives Matter"(BLM) movement has pushed the racial problem to a new peak and exacerbate the polarization on the political spectrum. Moreover, respondents whose age is between 35-54 and above 75 are more likely to vote for Trump in the election which are the two most essential groups in the American society. The 35-54 group represents the core and backbone of the society who generate the most wealth and capital in the country so that their choices have a larger weight than other age groups. For example, young students may not agree with what the Republicans and President Trump has announced but they may still vote for Trump as they are aware that their family income can benefit from his policies. For people who are over 75 mostly represent the retired senior group of the US, as Republicans and Trump are conservative, it is easier for the seniors to accept and agree with conservative policies. Plus, retired seniors usually have more free time than young college students and people who are busy with their full-time job so that they are more willing to and more capable of getting involved in this political activity. In state perspective, due to the electoral college, Trump will win approximately 31 states and Biden 19 states with Washington D.C. which their electoral college votes are respectively 278 and 260 only if Trump wins roughly all the battleground states or he will lose. Based on American history, recent election results and more importantly the model, we can certainly predict that Democrats and Biden will win dominantly in larger, more urbanized states such as New York, California, New England Area. While Trump and Republicans will lock the win in the small middle American rural states where natives, seniors and less educated votes are concentrated. The swing states are mostly located near the Great Lake area and the most essential Florida and those states are the key factors that will dominate the result of the presidential election. Under this system, what a candidate really hopes is not to acquire overwhelming and dominant support in one state but to win a small margin in every state as the excess votes can be considered a waste when the result is certain. Hence, one of the advantages that Biden has that we previously analyzed can also be a shortcoming for him. As it is a common sense that people who share the same culture, race or ethnics would like to dwell together instead of living separately. Hispanics who constitute nearly 20% of American population, as one of the biggest Biden's supporters, mainly located in the South and West which are closer to the border of Mexico and Latino America. Unlike the Caucasian American people who are located equally elsewhere in the US, the population advantage of Hispanics may not be as effective as we assume which can only ensure Biden to win some "blue states' " overwhelmingly. However, the result can be overturned by other factors since the differences of the votes between those two candidates are not huge. In this COVID-19 pandemic, old people are easier to be infected and more fatal to the retired seniors who are the primary supporters of Trump, thus the voting rate of the seniors will drop which is more unacceptable for Trump. Mail-in ballot also created a uncertainty to the election as people are less likely to ensure their votes are correctly and effectively registered. Those practical ambiguities can make a totally different outcome from our presumption.

5.2 Regarding the electoral college vote outcome:

The US president isn't directly elected by the people (the national popular vote results). Rather, it is elected by an electoral college formed by 538 electors from different states. In most of the states, the candidate that won the popular vote *of the state* takes all of its electoral college votes, rather than distribute the votes based on the proportion of the ballots that they received. So our national popular vote forecast doesn't necessarily forecast the winner of the race–in fact, Donald Trump himself won the election in 2016 without winning the

popular votes.
However, based on our state level post-stratification results, shouldn't we be able to predict the number of electors that would vote for each candidate? Problem: given that we build our model outcome in a binary way (vote for candidate A/not for A) and on data of everyone who is theoretically able to vote rather than only the likely voters, along with the fact that voting isn't mandatory for all adult US citizens, and the first-past-the-post system, we can't actually predict the electoral college vote. For example, if we estimate Biden wins the vote of 37% of all adult Floridian US citizens, it doesn't mean Trump wins Florida – it could be Trump wins the 35% of the adult Floridian US citizens and the rest of them either don't cast a vote (due to lack of interests or voter suppression), or vote for a third party, or their mail-in ballots get burnt by some Republican or are never counted. As long as Biden wins the majority or even just the plurality (with strong 3rd party candidate(s), that is) of the ballots casted (and counted) in that state, he wins the state and thus all its electoral college votes. So what our model really can do in this aspect is predicting the states, thus the number of electoral college vote, that are *safely* Biden's.

5.3 Summary of our forecast analysis

Although Trump is an incumbent and may have easier access to be re-elect, our research still gives out a prediction that Biden will win the election by a small margin as it is hard for Trump to win all the swing states. Nevertheless, the 0.3% of the winning rate difference can be easily overturned by any ambiguity related to the COVID or etc. Such as the 2000 US election Al Gore lost the election by only 5 electoral votes out of 538 as he lost only hundreds of popular votes in Florida. Hence Biden will win the election with a relatively large uncertainty.

5.4 Weakness and next step:

1. Filtered 2018 ACS only contains a sample of roughly 12000 respondents. Compared to 128,824,246 voters who voted in the 2016 election, the result cannot fully represent the opinion of all American people. More surveys should be given out by the polling companies.

2. The result of the presidential election is determined by the electoral college instead of a direct election such as proportional representation and this implies that if the candidate loses in one state he would lose all of the votes in that state instead of getting a certain proportion of the electoral college vote. Hence in the context of America, "swing states" where both candidates gain support about the same play an important role and should be weighted more when surveys are given out to predict the result of the election. However in the sample, the number of respondents are dispersed equally among all the regions of the US, especially the respondents who come from West and North East states which are not swing states and are generally Democratic constitute more than 40% of the sample. The data from those states are typically futile and more data from the swing states can better reflect the prediction. If we could allocate the budget of polling companies, more budgets should be distributed in the swing states such as Florida, Wisconsin and etc instead of California, New York or Texas where the results are certain.

3. As the world is suffering from the global pandemic COVID-19, it may more or less affect the result of the election. The data was collected before 2020 when the US had not been hit by the virus, yet the solution that President Trump gave to deal with the pandemic for sure shifted people's opinion of him. And COVID has become a hot issue in the presidential debate. Moreover, it altered the method of how people are going to vote. Mail-in ballot is, for the first time, used in a presidential election however the postal services in the US cannot be fully trusted as there might be postman throwing away people's ballot or it takes too long to be delivered on time before the election day and it also left the space for candidates to "cheat" as more people are afraid to catch the virus if they vote in person. For polling companies, the data should be collected and organized regularly, for example every two months, especially near the election day. Moreover, the weight of those voters who vote in person should increase as it is relatively more certain that their votes are counted in the election.

6 References

David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

ESS Methodology. (n.d.). Retrieved from https://www.europeansocialsurvey.org/methodology/ess_methodology/data_processing_archiving/weighting.html#:~:text=Post-stratification weights are a,gender, education, and region. Hadley Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

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

Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Multilevel regression with poststratification. (2020, October 14). Retrieved from https://en.wikipedia.org/wiki/Multilevel_regression_with_poststratification

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Ryan Hurl.(2016). Understanding America, Parties and Elections, pp101-121

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

Sommet, N., & Morselli, D. (2017). Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS. International Review of Social Psychology, 30(1), 203–218. DOI: http://doi.org/10.5334/irsp.90

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=84a68f37-86ac-4871-b68a-a57b4d9e31d2. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30, https://yihui.org/knitr/

“State Abbreviations and State Postal Codes” Fact Monster. © 2000–2017 Sandbox Networks, Inc., publishing as Fact Monster. 23 Oct. 2020 https://www.factmonster.com/us/postal-information/state-abbreviations-and-state-postal-codes/.

7 Appendices

1. State variable in datasets used to build the multilevel model is reconstructed to represent 51 states + DC in postal abbreviation. State Postal Codes are retrieved from State“State Abbreviations and State Postal Codes” Fact Monster. Website link: https://www.factmonster.com/us/postal-information/state-abbreviations-and-state-postal-codes/.

2. MAF is an inventory of known housing units, group quarters, and selected nonresidential units in the United States. MAF records contain mailing address, location information, and additional attributes of each residence.(National Research Council. 2010)

3. Github link which contains all the code, dataset(except for original ACS 2018 and Nationscape survey, method to download is attached in readme.md), and report for the project: https://github.com/rubytianxiaoma/2020-election-forecast

4. See Table 15 for the frequency table of state distribution in Nationscape survey.

5. See Table 16 for the Frequency table of state distribution in subset of 2018 ACS survey

Table 15: State proportion in Nationscape wave 50 survey

state	count	proportion
AK	8	0.0012348
AL	89	0.0137367
AR	50	0.0077172
AZ	162	0.0250039
CA	717	0.1106652
CO	98	0.0151258
CT	71	0.0109585
DC	23	0.0035499
DE	28	0.0043217
FL	508	0.0784072
GA	185	0.0285538
HI	30	0.0046303
IA	55	0.0084890
ID	32	0.0049390
IL	291	0.0449143
IN	119	0.0183670
KS	51	0.0078716
KY	90	0.0138910
LA	81	0.0125019
MA	115	0.0177497
MD	101	0.0155888
ME	20	0.0030869
MI	181	0.0279364
MN	70	0.0108041
MO	123	0.0189844
MS	48	0.0074086
MT	17	0.0026239
NC	214	0.0330298
ND	7	0.0010804
NE	23	0.0035499
NH	19	0.0029326
NJ	206	0.0317950
NM	28	0.0043217
NV	74	0.0114215
NY	519	0.0801050
OH	294	0.0453774
OK	68	0.0104954
OR	91	0.0140454
PA	274	0.0422905
RI	12	0.0018521
SC	106	0.0163605
SD	16	0.0024695
TN	121	0.0186757
TX	480	0.0740855
UT	54	0.0083346
VA	205	0.0316407
VT	15	0.0023152
WA	126	0.0194474
WI	120	0.0185214
WV	39	0.0060194
WY	5	0.0007717

Table 16: State proportion in 2018 ACS

state	count	proportion
AK	204	0.0173869
AL	231	0.0196881
AR	224	0.0190915
AZ	270	0.0230120
CA	280	0.0238643
CO	263	0.0224154
CT	242	0.0206256
DC	198	0.0168755
DE	209	0.0178130
FL	275	0.0234382
GA	259	0.0220745
HI	225	0.0191767
IA	201	0.0171312
ID	214	0.0182392
IL	268	0.0228416
IN	235	0.0200290
KS	233	0.0198585
KY	216	0.0184096
LA	239	0.0203699
MA	252	0.0214779
MD	251	0.0213927
ME	175	0.0149152
MI	249	0.0212222
MN	228	0.0194324
MO	233	0.0198585
MS	216	0.0184096
MT	173	0.0147447
NC	267	0.0227563
ND	168	0.0143186
NE	215	0.0183244
NH	185	0.0157675
NJ	273	0.0232677
NM	246	0.0209665
NV	269	0.0229268
NY	278	0.0236939
OH	251	0.0213927
OK	231	0.0196881
OR	242	0.0206256
PA	260	0.0221597
RI	214	0.0182392
SC	231	0.0196881
SD	175	0.0149152
TN	234	0.0199437
TX	275	0.0234382
UT	232	0.0197733
VA	258	0.0219893
VT	141	0.0120174
WA	261	0.0222450
WI	223	0.0190062
WV	168	0.0143186
WY	173	0.0147447