Predicting the US 2020 election vote[*]

My subtitle if needed

Mohammad Sardar Sheikh, Danur Mahendra, Justin Teng

10 April 2022

Abstract

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

You can and should cross-reference sections and sub-sections. For instance, Section 2. R Markdown automatically makes the sections lower case and adds a dash to spaces to generate labels, for instance, Section **??**.

This paper predicts the popular vote result of the 2020 United States presidential election using multilevel regression and post-stratification in R. As one of the most influential countries in the world, the election does not only affect American citizens; but rather, includes those nations who depend on the United States for aid, security, or trade. This presidential election is between the incumbent Republican president Donald Trump and Democratic candidate Joe Biden, who was the former vice-president serving from 2008-2016. Trump's presidency began with his inauguration following the 2016 presidential elections against Democratic candidate Hillary Clinton. Nevertheless, Trump's victory shocked major news outlets who considered him a significant underdog against Clinton (cite). Trump will now attempt to win a second term of presidency against Biden, who many analysts again see Trump as the underdog following his controversial time in office (). His tenure was met with criticism following his controversial stance or remarks on racial inequality, diplomatic relations, and inefficient spending (cite). Trump's turbulent presidency can be attributed to his lack of political experience and thus, many believe that a second term will not see much improvement. Ultimately, poll analysis and forecast outlets will once again favour the Democrat over the Republican in this presidential election. This paper will attempt to use R to forecast our own prediction and analyse the main driving factors behind each vote.

In this report, we will be utilizing the data from U.S.presidential election 2020 survey data from Nationscape

2 Data

This task requires multiple datasets to accurately predict the results of the 2020 United States popular vote. We first used Democracy Fund + UCLA Nationscape's December 2020 (Wave 76) data set to construct our multilevel regression method and followed it up by applying a post-stratification method using the IPUMS American Community Survey 2020 dataset. The ACS dataset would allow us to use our findings from the smaller Nationscape dataset to more accurately represent a much larger population.

[*] Code and data are available at: LINK.

Nationscape is a weekly online survey conducted by LUCID for Democracy Fund and UCLA researchers. Data for this wave's dataset was collected between December 24 - 30, 2020 and received 6,692 samples. Each wave must collect a set of demographic quotas based on the respondents' age, gender, race/ethnicity, region, income, and education. The quotas are based on the U.S. adult population in 2017 provided by the U.S. Census Bureau. Respondents submit their responses through online survey software provided by LUCID.

The ACS survey is a monthly rolling survey used to update census estimation for the Census Bureau. The ACS uses two sampling frames both provided by the Census Bureau, housing unit (HU) addresses and residents of group quarter (GQ) facilities. Samples were collected by a method of stratified sampling. Respondents were then contacted to complete the survey via either Computer-Assisted Personal Interview (CAPI) or Computer-Assisted Telephone Interview (CATI). The ACS samples include roughly 3 million households with each sampling unit representing a household and all persons residing in the household.

The original datasets that we received needed to be cleaned so that they could be used effectively in our analysis. We used separate scripts to clean the datasets. First, we read in the datasets and then we choose the variables that we require. Since we are planning to do a regression with the data that we have, we cannot have any NA (missing) values in the cleaned datasets, so we remove all the rows that contain NA values. For the UCLA dataset, we create a new variable that equals 1 if a respondent chooses to vote for Donald Trump, and 0 if the respondent plans to vote for Joe Biden. We filter out all the observations that do not refer to either of these two candidates, for the sake of more accurate results. We then need to make sure that the common variables in both the datasets have equivalent values. We factorise race as a variable and we categorise it according to research and looking at the codebook for both the datasets. We categorise race into 5 parts, white, african american or black, asian and pacific islanders, native american, and others.

Regarding education, we split it into 3 categories. The first one is pre-high school (the respondent has not received a high school diploma), the second is high school diploma or equal, and the third is college diploma or higher. Since we are interested in how the Hispanic community votes, we convert being hispanic into a binary variable, much like voting for Trump, where a value of 1 signals that the respondent is Hispanic, and 0 says otherwise. We do a very similar thing with gender, choosing to classify females as 1 and males as 0.

Finally, we want to make sure that the value of States are the same for both the datasets. For this, all the unique strings need to match up accordingly, so "New York" has to be the same in both the datasets. It wont work if its "New York" for one and "new york" for the other.

3 Model

We are interested in forecasting the popular vote result of the 2020 United States presidential election. In Particular, we want to predict the proportion of voters that will vote for the Republican candidate, Donald Trump. To achieve this, we used the December 2021 Nationscape dataset to find a relationship between population characteristics and their vote intention. We then used the 2020 ACS dataset to apply a post-stratification method. This technique allows us to fit a smaller data set to match one that would more accurately represent a much larger population. In this case, the population we are trying to represent is the American voting population.

We used a logistic regression model to estimate the probability of a voter voting for Trump given certain characteristics (represented by predictor variables). An individual's voting intention is represented by a binary response variable in our data set thus, we believe that a logistic regression model best suits this task. The reason we don't choose a linear regression model is because we believe that a straight line will not accurately represent the data. A logistic regression line has a curved 'S' shape, and this is more suited to predict and calculate binary values that can only either take 0 or 1 as responses. This is crucial as the post-stratification process is contingent on choosing the most appropriate model. Choosing a model that breaks assumptions may render our post-stratification and prediction inaccurate. The response variable in our model denotes whether the voter intends to vote for Republican candidate Trump. We get a probability (p) as the response so we assume that (1-p) is the probability that a respondent intends to vote for Joe Biden.

The predictor variables used are the voter's age, state of residence, gender, race, income, whether the voter is Hispanic or not, and highest level of education attained as we believe they are key factors that form one's political views. We believe that age is an important variable as most people's political views change with age, with some particular age groups preferring one candidate over the other. States are a big part of the model, there are some states that are very pro Trump, and some states that are very anti-Trump. We feel like incorporating these differences in views is a crucial part for the model. Another variable that we feel is important to include is gender. A respondents gender can influence the way they perceive society and can affect their political views. There is a slight difference in the gender make-up of the country so we feel like we should include this as a variable. Education is another important variable that we want to include in the model. Educated people are more aware of the way society works and what is going on around them. They are less likely to make decisions based on passion and feelings and are more likely to think before they do something important, such as decide which candidate to support. The difference in viewpoints that they bring with them seems like an important thing to not include in our model. We decide to include whether a respondent is Hispanic or not in the model as we want to see whether these specific people will choose to vote for Trump, even after all the statements that he has made against them such as "I would build a great wall, and nobody builds walls better than me, believe me, and I'll build them very inexpensively. I will build a great great wall on our southern border and I'll have Mexico pay for that wall." (Donald Trump). We want to see whether statements like these have had any effect on his popularity. Finally, we incorporate race into the model. People from different racial backgrounds have their own personal agendas, choosing to side with a specific candidate over the other.

Incorporating all of these variables, our logistic regression model can be represented by the following formula:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{gender} + \beta_2 x_{age} + \beta_3 x_{race} + \beta_4 x_{stateicp} + \beta_5 x_{education_category} + \beta_6 x_{hispanic} \qquad (1)$$

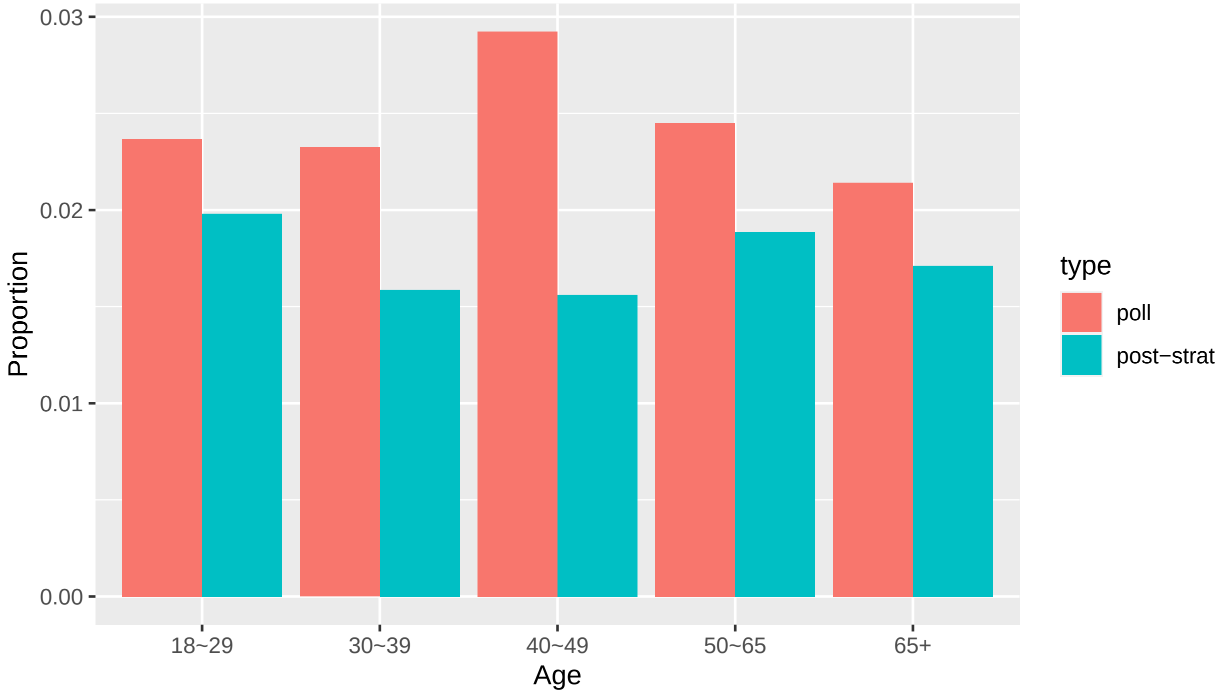
Equation (1) is not the full representation of the model as we have not shown the full model to account for space. $\beta_4 x_{stateicp}$ actually refers to all the different states that we have in the data set (50 of them), and each state has its own unique β value. Similar for education and race. The intercept for a logistic regression model, β_0, while included, does not offer much in terms of information and interpretability

$$Pr(\theta|y) = \frac{Pr(y|\theta)Pr(\theta)}{Pr(y)} \qquad (2)$$

Equation (2) seems useful, eh?

4 Results

4.0.1 Age



We can see from **??** the way that age was distributed in the UCLA and ACS data. We have a smaller proportion for ACS because there were a number of observations that were from respondents that had not yet turned 18 and hence could not vote.

4.0.2 Gender

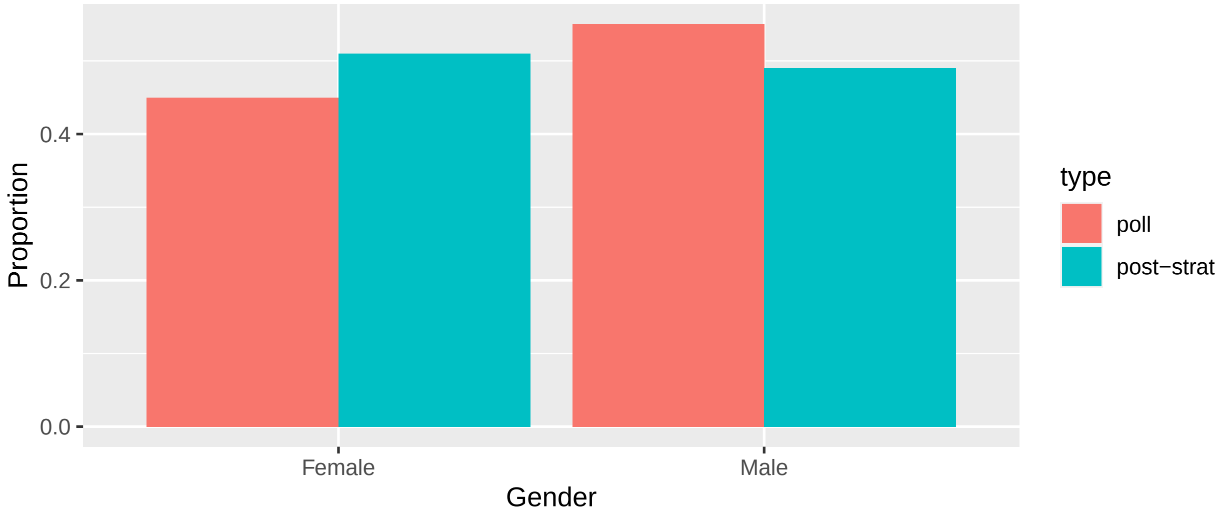


Figure 1: Voter's Demographic: Gender

We can see from ref graph that the proportion of males and females in the two data sets are different, with the post stratification data having a higher male population as compared to the polling data.

4.0.3 Education

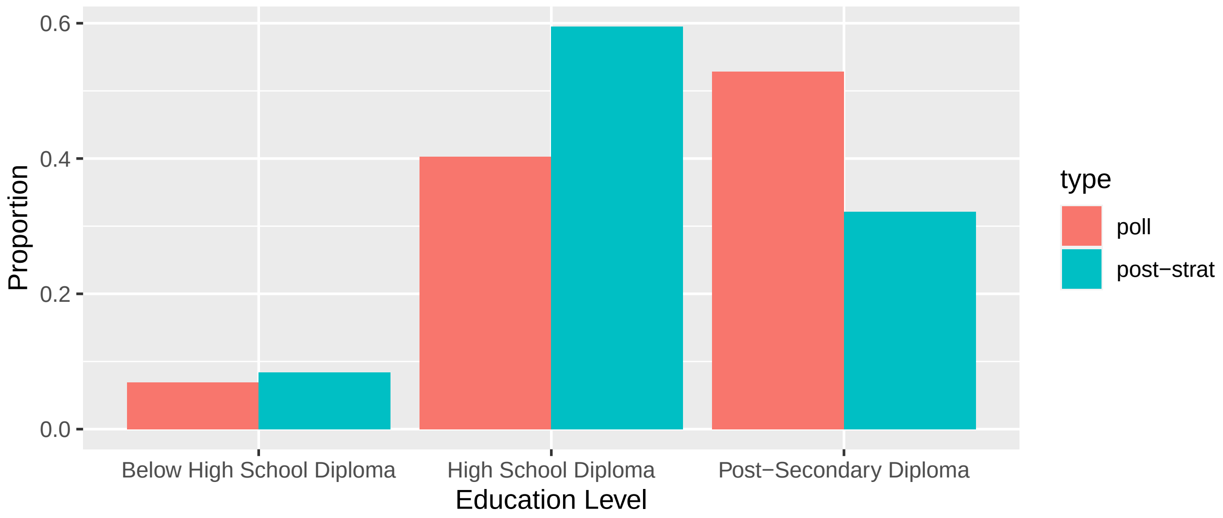


Figure 2: Voter's Demographic: Education

We can see from the education graph that most of the people in the polling data have an education that is college degree or higher, whereas most of the people in the post stratification data have an education only comparable to high school level.

4.0.4 Hispanic

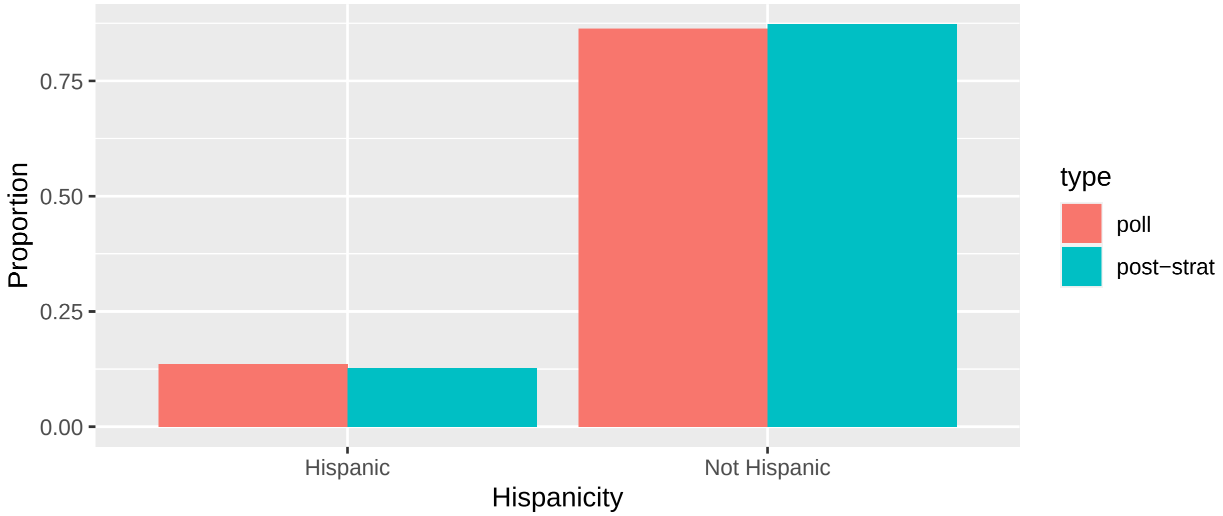


Figure 3: Voter's Demographic: Gender

(ref) We can see from the graph above that the distribution for a voter being hispanic or not is the same in both of the datasets.

4.0.5 Race

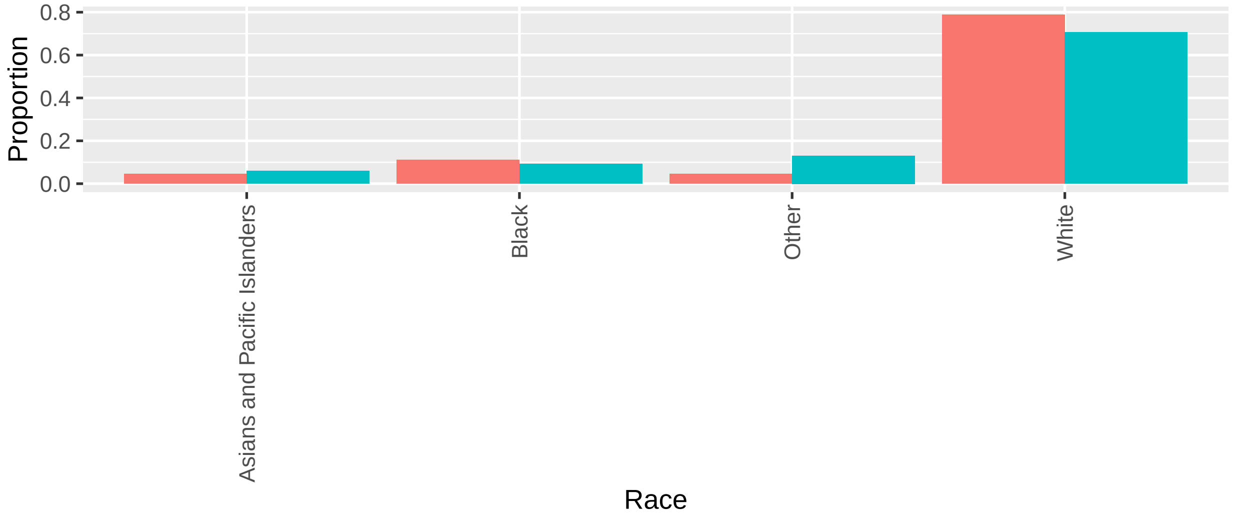
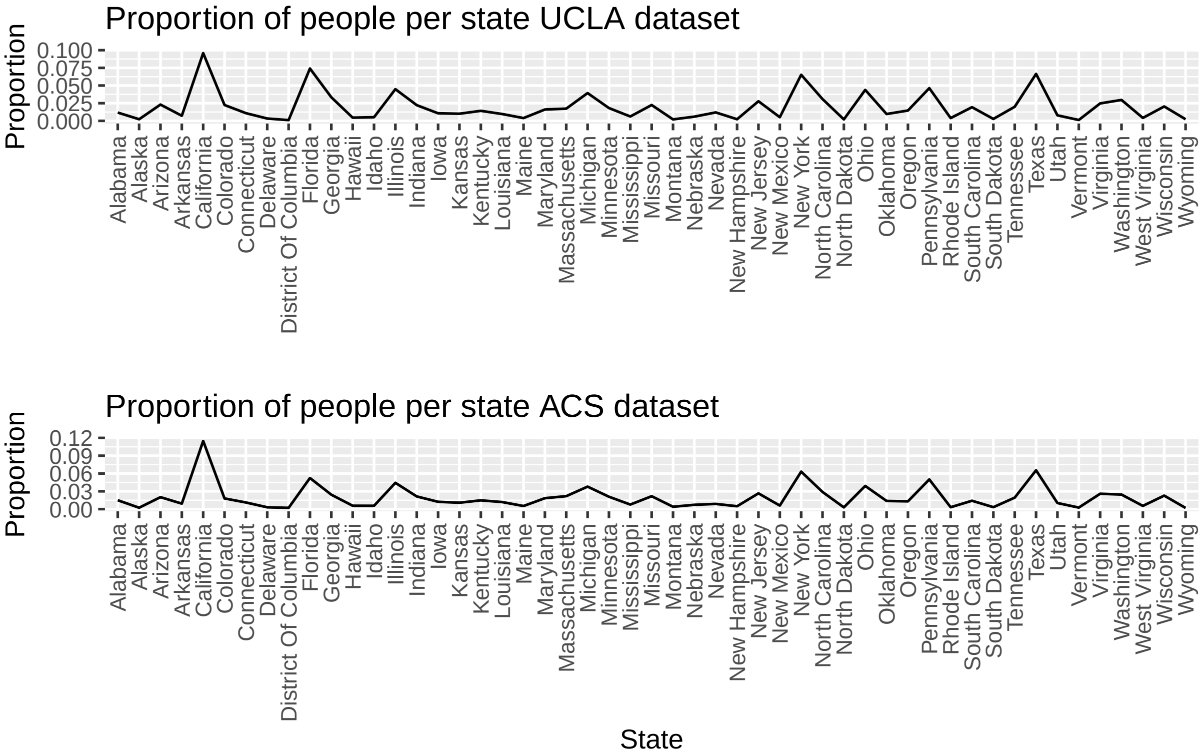


Figure 4: Voter's Demographic: Race

From the graph above we can see that respondents race is equally distributed across the two datasets.

4.0.6 States



The above graph shows us the relative proportion of people that belong to each of the states. While the concentrations appear to be approximately the same, we do have some states that have different proportions. States such as Florida, New York and Nevada appear to have subtle differences in proportions.

4.0.7 Regression and Results

Have to ref the table:

We can see from the results of the logistic regression model above that most of the states are not in favor of voting for Trump. This can be seen from the negative values that we get in the regression table that we have. Since we have done a logistic regression model, the coefficient values that we get cant be interpreted normally, they are the log odd ratio's. What this means is that if the value that we get is negative, then the voter is more likely to vote for Joe Biden, whereas if the value that we get is positive, then the voter is more likely to vote for Donald Trump. Based on the results above, we can see that age is the only variable that has a positive effect on Trump, as the voter gets older, there is a higher chance for them to vote for Trump.

We can get odds ratios for the model simply by exponentiating the coefficients. What these odd ratios tell us are the relative chance that a particular group votes for Trump as compared to the baseline for that group. We can see from the table above that African American people (race2) are very unlikely to vote for Trump. the coefficient tells us that the probability that an African American votes for Trump is 0.17 the probability that a white person votes for Trump. Similarly, for education we can see that the more educated that you are, the less likely it is that you vote for Donald Trump, as can be seen from the coefficients. While all of the states are slightly against voting for Trump, some of them are very heavily against him. States like Columbia, Rhode Island and Vermont are very likely to vote for Joe Biden instead of Trump.

Table 1: Coefficients from the Model

term	estimate	standard error	statistic	p-value
(Intercept)	1.165	0.340	3.431	0.001
age	0.014	0.002	6.777	0.000
as.factor(race)2	-2.009	0.146	-13.777	0.000
as.factor(race)3	-0.305	0.322	-0.947	0.344
as.factor(race)4	-1.077	0.184	-5.844	0.000
as.factor(race)5	-0.886	0.165	-5.356	0.000
as.factor(education_category)2	-0.288	0.131	-2.193	0.028
as.factor(education_category)3	-0.769	0.130	-5.894	0.000
sex	-0.349	0.063	-5.528	0.000
stateicpAlaska	-1.152	0.690	-1.669	0.095
stateicpArizona	-1.297	0.365	-3.555	0.000
stateicpArkansas	-0.426	0.484	-0.880	0.379
stateicpCalifornia	-1.406	0.323	-4.350	0.000
stateicpColorado	-0.959	0.364	-2.632	0.008
stateicpConnecticut	-1.867	0.435	-4.292	0.000
stateicpDelaware	-1.740	0.636	-2.736	0.006
stateicpDistrict Of Columbia	-1.971	1.185	-1.663	0.096
stateicpFlorida	-1.354	0.326	-4.152	0.000
stateicpGeorgia	-0.901	0.350	-2.573	0.010
stateicpHawaii	-0.423	0.555	-0.761	0.446
stateicpIdaho	-0.601	0.520	-1.155	0.248
stateicpIllinois	-1.316	0.338	-3.889	0.000
stateicpIndiana	-1.294	0.366	-3.534	0.000
stateicpIowa	-1.314	0.420	-3.127	0.002
stateicpKansas	-0.876	0.427	-2.053	0.040
stateicpKentucky	-0.678	0.399	-1.701	0.089
stateicpLouisiana	-1.075	0.451	-2.383	0.017
stateicpMaine	-1.209	0.563	-2.148	0.032
stateicpMaryland	-1.624	0.399	-4.069	0.000
stateicpMassachusetts	-2.126	0.401	-5.297	0.000
stateicpMichigan	-1.386	0.342	-4.051	0.000
stateicpMinnesota	-1.267	0.379	-3.345	0.001
stateicpMississippi	-0.652	0.511	-1.276	0.202
stateicpMissouri	-0.924	0.366	-2.524	0.012
stateicpMontana	-0.720	0.731	-0.986	0.324
stateicpNebraska	-0.981	0.490	-2.001	0.045
stateicpNevada	-1.711	0.426	-4.017	0.000
stateicpNew Hampshire	-1.762	0.712	-2.477	0.013
stateicpNew Jersey	-1.657	0.361	-4.584	0.000
stateicpNew Mexico	-1.363	0.530	-2.572	0.010
stateicpNew York	-1.520	0.331	-4.598	0.000
stateicpNorth Carolina	-1.244	0.353	-3.522	0.000
stateicpNorth Dakota	-0.241	0.760	-0.317	0.751
stateicpOhio	-1.168	0.338	-3.453	0.001
stateicpOklahoma	-0.479	0.437	-1.096	0.273
stateicpOregon	-1.029	0.393	-2.620	0.009
stateicpPennsylvania	-1.216	0.337	-3.609	0.000
stateicpRhode Island	-2.403	0.655	-3.670	0.000
stateicpSouth Carolina	-0.614	0.379	-1.619	0.105
stateicpSouth Dakota	-1.298	0.643	-2.019	0.043
stateicpTennessee	-0.678	0.375	-1.811	0.070
stateicpTexas	-1.179	0.328	-3.596	0.000
stateicpUtah	-0.258	0.468	-0.552	0.581
stateicpVermont	-14.314	216.133	-0.066	0.947
stateicpVirginia	-1.208	0.362	-3.338	0.001

Table 2: Exponentiated Coefficients from the Model

	coefficient	2.5 %	97.5 %
(Intercept)	3.207	1.661	6.308
age	1.014	1.010	1.018
as.factor(race)2	0.134	0.100	0.177
as.factor(race)3	0.737	0.386	1.374
as.factor(race)4	0.340	0.235	0.484
as.factor(race)5	0.412	0.296	0.567
as.factor(education_category)2	0.750	0.579	0.970
as.factor(education_category)3	0.463	0.359	0.598
sex	0.706	0.623	0.798
stateicpAlaska	0.316	0.078	1.230
stateicpArizona	0.273	0.132	0.555
stateicpArkansas	0.653	0.254	1.711
stateicpCalifornia	0.245	0.129	0.458
stateicpColorado	0.383	0.186	0.778
stateicpConnecticut	0.155	0.065	0.358
stateicpDelaware	0.175	0.047	0.589
stateicpDistrict Of Columbia	0.139	0.007	1.109
stateicpFlorida	0.258	0.135	0.485
stateicpGeorgia	0.406	0.202	0.802
stateicpHawaii	0.655	0.216	1.928
stateicpIdaho	0.548	0.198	1.538
stateicpIllinois	0.268	0.137	0.517
stateicpIndiana	0.274	0.132	0.558
stateicpIowa	0.269	0.117	0.608
stateicpKansas	0.416	0.179	0.959
stateicpKentucky	0.508	0.231	1.107
stateicpLouisiana	0.341	0.139	0.821
stateicpMaine	0.298	0.097	0.901
stateicpMaryland	0.197	0.089	0.427
stateicpMassachusetts	0.119	0.053	0.259
stateicpMichigan	0.250	0.127	0.485
stateicpMinnesota	0.282	0.133	0.588
stateicpMississippi	0.521	0.190	1.420
stateicpMissouri	0.397	0.192	0.809
stateicpMontana	0.487	0.117	2.191
stateicpNebraska	0.375	0.142	0.979
stateicpNevada	0.181	0.077	0.411
stateicpNew Hampshire	0.172	0.039	0.672
stateicpNew Jersey	0.191	0.093	0.384
stateicpNew Mexico	0.256	0.088	0.712
stateicpNew York	0.219	0.113	0.415
stateicpNorth Carolina	0.288	0.143	0.572
stateicpNorth Dakota	0.786	0.188	4.056
stateicpOhio	0.311	0.158	0.599
stateicpOklahoma	0.619	0.263	1.468
stateicpOregon	0.357	0.164	0.768
stateicpPennsylvania	0.296	0.152	0.570
stateicpRhode Island	0.090	0.022	0.305
stateicpSouth Carolina	0.541	0.255	1.133
stateicpSouth Dakota	0.273	0.075	0.966
stateicpTennessee	0.508	0.242	1.053
stateicpTexas	0.308	0.160	0.580
stateicpUtah	0.773	0.311	1.957
stateicpVermont	0.000	NA	573.241
stateicpVirginia	0.299	0.146	0.603

4.0.8 Predicting

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1     156306.
```

When we use our regression model that we have fit using the UCLA polling data, on the ACS post stratification data, the result that we get is that only 43% of the population will vote for Donald Trump in the 2020 US presidential elections. This results tell us that Trump will lose the elections and that Joe Biden will win (which is exactly what happened). # Discussion

For this research paper we took two different sources of data and we used those to predict the outcome of the 2020 US general elections. We cleaned the datasets first to make all the variables equal and we then created and fit a logistic regression model, taking a vote for Donald trump as the response variable. A logistic regression model has a binary response and gives us relative probabilities as the coefficients of the predictors. We fit this model on the UCLA dataset as the dataset had a respondents vote intention as one of the variables that was recorded. We then use the model that we have fit, on the post stratification dataset (the ACS data) and we find predictions on the outcome of the 2020 US election based on the results of the prediction. We can see from our model results, that our model predicts that Donald Trump is expected to receive 43% of the votes for the 2020 US elections. We unfortunately are not completely certain whether this grants him a victory or not as we are not aware of the percentage of the other respondents that vote for Joe Biden vs the percentage that vote for a third party (though we can make a strong prediction that Joe Biden will win that election based on the results). We can see that there are certain factors that make it more likely for a candidate to vote for Trump. These are RESULTS to be discussed here along with some literature review. Based on the results of our logistic regression model, we can see that most of the states have a higher tendency to vote for Joe Biden as compared to Donald Trump. This does not exactly seem to make sense as their are some states that are very pro Trump, yet the output of the model tells us otherwise. We can also se from the results that females are slightly more likely to vote for Joe Biden as compared to Donald Trump. Based on common knowledge, people that belong to the Hispanic community are not major fans of Donald Trump. They are very unlikely to vote for Donald Trump in the 2020 elections. From the model results we can also see that the odds of voting for Trump decrease if the respondent is not 'white'. All of the other races are less likely to vote for Donald trump as compared to white people. There are some things that we need to take into account while presenting our results. The survey data that we have was gathered in December 2020, while the US elections took place in November 2020. Essentially we have used data from beyond the election to predict the results of the election. This can pose some problems as Americans were regretting electing Trump very soon after he was sworn in. Another potential reason as to why we are getting biased results may be because in our survey data, the majority of the respondents (around 60%) were intending to vote for Joe Biden. This may have slightly biased the results of the model.

4.1 Weaknesses and next steps

There are some limitations that we have to consider. Our model and prediction are heavily relying on the data that we have from UCLA, that is the polling data, since we make predictions about the population based on those. In the weekly polling data that we had, the people that wanted to vote for Joe Biden were of a higher proportion than those that wanted to vote for Donald trump. It was roughly a 60-40 split respectively. This has a chance to bias the results that we have from the post-stratification data on the population as the initial groups were uneven. Another possible limitation is that the logistic regression model that we use can only have a binary response, so either we know if a respondent voted for Donald Trump, or whether they voted for Joe Biden. The model does not take into account any third party that someone might want to vote for.

Appendix

A Additional details

B References