A statistical analysis on 2020 American federal election

Biden slightly ahead of Trump

Zhipeng Zhou

November 2, 2020

Model

Here we are interested in predicting the popular vote outcome of the 2020 American federal election (include citation). To do this we are employing a post-stratification technique. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

Model Specifics

To predict the popular vote outcome of the 2020 American federal election over Donald Trump and Joe Biden, I build statistical models for them separately, since different voter can support their presidential candidate for different reasons. So modeling them separately can reflect the actual situation better.

For both Donald Trump and Joe Biden, I will use logistic regression model to model the proportion of voters who will vote them. The reason why I choose logistic regression model is logistic regression is a classification algorithm, which works well when the target variable is categorical in nature. And this time, both models' target variables are whether vote certain presidential candidate or not which is binary. So logistic regression model can be a good choice here. And for the entire process of analysis, I will use R in Rstudio to perform.

Model for Donald Trump

For Donald Trump, I will use age, family income, race, and gender to model the probability of voting for Donald Trump. And all variables are recorded as categorical variables. The multiple logistic regression model I am using is:

$$\ln\frac{y_T}{1-y_T} = -0.31 - 0.6 \cdot x_{age} + 0.73 \cdot x_{income} - 0.53 \cdot x_{mexican} - 1.95 \cdot x_{black} - 1.38 \cdot x_{chinese} - 0.49 \cdot x_{other} + 0.39 \cdot x_{sex} + \epsilon$$

Where y_T represents the proportion of voters who will vote for Donald Trump. Similarly, x_{age} represents whether the voter's age is under 30 or not. x_{income} represents whether the family income is within the range of \$200,00 and \$249,999. $x_{mexican}$, x_{black}, $x_{chinese}$, and x_{other} show the race of the voter. And x_{sex} represents the gender of the voter, where 1 is male and 0 is female. And -0.31 represents the probability of voting for Donald Trump when the voter is female, white, over 30 years old, and family income not in that range.

According to the multiple logistic regression model we have, we can find that. First, voters whose age greater than 30 may not likely to vote to Trump. Except white voter, all the other voter also may not vote to Trump. And male voters have high chance to vote to Trump.

The major reason why I choose these variables is the banner book for my survey data set. Refer to that book [1], we can find that there were already some analysis over the 2020 American federal election. So this time, we just use the same variable they used to do our further analysis. However, there are many subdivisions for each variable, like age which has over 80 subdivisions. So I first divide the age into different age groups. Since same generation always have similar way of thinking, I divide the age into 10-20, 20-30, and so on. Then combining the Post-Stratification data we have and P-value for each variables, I first filter the variables and only keep some important subdivision variables. Then I use $regsubset()$ function in R [2] to find the final variables, which performs best subset selection by identifying the best model that contains a given number of predictors, where best is quantified using RSS.

Model for Joe Biden

For Joe Biden, I will use family income, race, and gender to model the probability of voting for Joe Biden. And all variables are recorded as categorical variables. The multiple logistic regression model I am using is:

$$\ln \frac{y_B}{1-y_B} = -0.35 - 0.54 \cdot x_{income} + 0.34 \cdot x_{mexican} + 1.17 \cdot x_{black} + 0.75 \cdot x_{chinese} + 1.42 \cdot x_{japanes} - 0.28 \cdot x_{sex} + \epsilon$$

Where y_B represents the proportion of voters who will vote for Joe Biden. Some variables are already shown above and they have the same meaning as that in previous model. And $x_{japanes}$ represents whether the voter is Japanese or not. And we can find there is an interesting phenomenon that many variables cross in these two models and they are opposite influence on voters' decision. The voters whose family income in the range of \$200,00 and \$249,999 are not likely to vote Biden, but Trump. And Voters who are mexican, black, Chinese, or Japanes are likely to vote Biden, but not Trump. For Biden's model, I use the same method to choose variables.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump and Joe Biden. I need to perform a post-stratification analysis. The post-stratification technique can reduce the bias of our model. Since the population of voter for 2020 American federal election are huge and various, applying the model to the whole population can lead to great bias. The post-stratification technique can divide the whole sample population into several small groups and each group has their own feature. And the bias for each group can be smaller than consider them as a whole. Then we add the result for each group up, weight them by size, and eventually calculate the average value for the whole population. In this case, we are not summarize the whole population, but also take feature of different groups into consideration.

Here I create cells based off different ages, gender, race, and family income. First, I only keep useful data from data set and found that there are some unusual data type in variable age ("less than 1 year old" and "90 (90+ in 1980 and 1990)"). Then I removed both data and also those who age is less than 18, since only those whose age are greater than 18 can vote [3] and data "90 (90+ in 1980 and 1990)" already cover by other 90+ years old data. Then I split the data into different cells based on the rule I used for both Trump and Biden model. Like whether age under 30, whether the voter's race is black, and so on.

Using the model described in the previous sub-section I will estimate the proportion of voters in each cell. I will then weight each proportion estimate (within each cell) by the respective population size of that cell and sum those values and divide that by the entire population size.

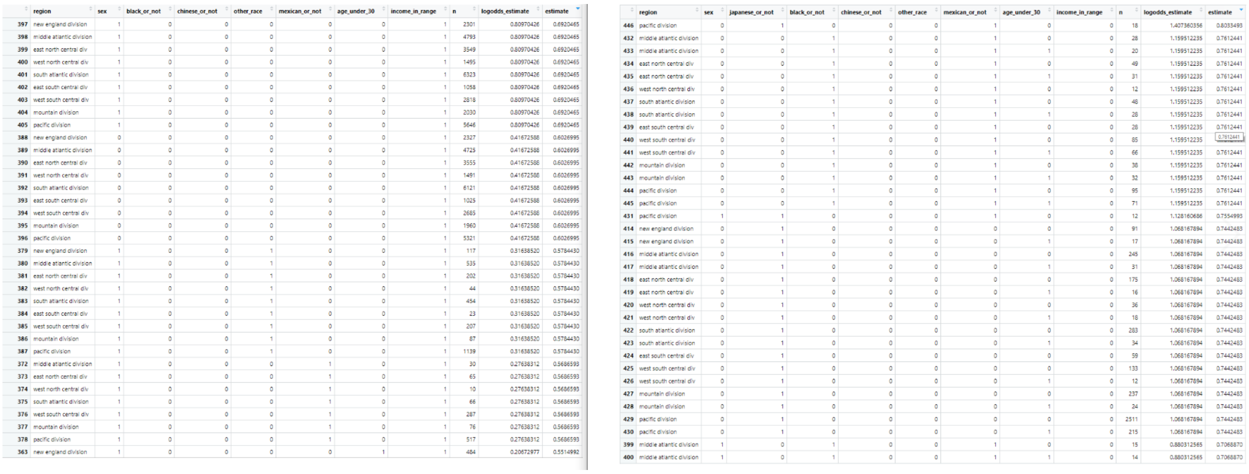
Results

The result for each model is calculated by

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

Where $\hat{y_j}$ is the estimate in each cell. And N_j is the population size of the j^{th} cell based off demographics. Then we have the result,

Donald Trump	Joe Biden
39.5%	41.4%



	region	sex	black_or_not	chinese_or_not	other_race	mexican_or_not	age_under_30	income_in_range	n	logodds_estimate	estimate
397	new england division	1	0	0	0	0	0	1	2301	0.80970426	0.6920465
398	middle atlantic division	1	0	0	0	0	0	1	4793	0.80970426	0.6920465
399	east north central div	1	0	0	0	0	0	1	3549	0.80970426	0.6920465
400	west north central div	1	0	0	0	0	0	1	1495	0.80970426	0.6920465
401	south atlantic division	1	0	0	0	0	0	1	6323	0.80970426	0.6920465
402	east south central div	1	0	0	0	0	0	1	1058	0.80970426	0.6920465
403	west south central div	1	0	0	0	0	0	1	2818	0.80970426	0.6920465
404	mountain division	1	0	0	0	0	0	1	2030	0.80970426	0.6920465
405	pacific division	1	0	0	0	0	0	1	5646	0.80970426	0.6920465
388	new england division	0	0	0	0	0	0	1	2327	0.41672588	0.6026995
389	middle atlantic division	0	0	0	0	0	0	1	4725	0.41672588	0.6026995
390	east north central div	0	0	0	0	0	0	1	3555	0.41672588	0.6026995
391	west north central div	0	0	0	0	0	0	1	1491	0.41672588	0.6026995
392	south atlantic division	0	0	0	0	0	0	1	6121	0.41672588	0.6026995
393	east south central div	0	0	0	0	0	0	1	1025	0.41672588	0.6026995
394	west south central div	0	0	0	0	0	0	1	2685	0.41672588	0.6026995
395	mountain division	0	0	0	0	0	0	1	1960	0.41672588	0.6026995
396	pacific division	0	0	0	0	0	0	1	5321	0.41672588	0.6026995
379	new england division	1	0	0	1	0	0	1	117	0.31638520	0.5784430
380	middle atlantic division	1	0	0	1	0	0	1	535	0.31638520	0.5784430
381	east north central div	1	0	0	1	0	0	1	202	0.31638520	0.5784430
382	west north central div	1	0	0	1	0	0	1	44	0.31638520	0.5784430
383	south atlantic division	1	0	0	1	0	0	1	454	0.31638520	0.5784430
384	east south central div	1	0	0	1	0	0	1	23	0.31638520	0.5784430
385	west south central div	1	0	0	1	0	0	1	207	0.31638520	0.5784430
386	mountain division	1	0	0	1	0	0	1	87	0.31638520	0.5784430
387	pacific division	1	0	0	1	0	0	1	1139	0.31638520	0.5784430
372	middle atlantic division	1	0	0	0	1	0	1	30	0.27638312	0.5686593
373	east north central div	1	0	0	0	1	0	1	65	0.27638312	0.5686593
374	west north central div	1	0	0	0	1	0	1	10	0.27638312	0.5686593
375	south atlantic division	1	0	0	0	1	0	1	66	0.27638312	0.5686593
376	west south central div	1	0	0	0	1	0	1	287	0.27638312	0.5686593
377	mountain division	1	0	0	0	1	0	1	76	0.27638312	0.5686593
378	pacific division	1	0	0	0	1	0	1	517	0.27638312	0.5686593
363	new england division	1	0	0	0	0	1	1	484	0.20672977	0.5514992

	region	sex	japanese_or_not	black_or_not	chinese_or_not	other_race	mexican_or_not	age_under_30	income_in_range	n	logodds_estimate	estimate
446	pacific division	0	1	0	0	0	1	0	0	18	1.407360356	0.8033493
432	middle atlantic division	0	0	1	0	0	1	0	0	28	1.159512235	0.7612441
433	middle atlantic division	0	0	1	0	0	1	1	0	20	1.159512235	0.7612441
434	east north central div	0	0	1	0	0	1	0	0	49	1.159512235	0.7612441
435	east north central div	0	0	1	0	0	1	1	0	31	1.159512235	0.7612441
436	west north central div	0	0	1	0	0	1	0	0	12	1.159512235	0.7612441
437	south atlantic division	0	0	1	0	0	1	0	0	48	1.159512235	0.7612441
438	south atlantic division	0	0	1	0	0	1	1	0	28	1.159512235	0.7612441
439	east south central div	0	0	1	0	0	1	0	0	28	1.159512235	0.7612441
440	west south central div	0	0	1	0	0	1	0	0	85	1.159512235	0.7612441
441	west south central div	0	0	1	0	0	1	1	0	66	1.159512235	0.7612441
442	mountain division	0	0	1	0	0	1	0	0	38	1.159512235	0.7612441
443	mountain division	0	0	1	0	0	1	1	0	32	1.159512235	0.7612441
444	pacific division	0	0	1	0	0	1	0	0	95	1.159512235	0.7612441
445	pacific division	0	0	1	0	0	1	1	0	71	1.159512235	0.7612441
431	pacific division	1	1	0	0	0	1	0	0	12	1.128160686	0.7554993
414	new england division	0	1	0	0	0	0	0	0	91	1.068167894	0.7442483
415	new england division	0	1	0	0	0	0	1	0	17	1.068167894	0.7442483
416	middle atlantic division	0	1	0	0	0	0	0	0	245	1.068167894	0.7442483
417	middle atlantic division	0	1	0	0	0	0	1	0	31	1.068167894	0.7442483
418	east north central div	0	1	0	0	0	0	0	0	175	1.068167894	0.7442483
419	east north central div	0	1	0	0	0	0	1	0	16	1.068167894	0.7442483
420	west north central div	0	1	0	0	0	0	0	0	36	1.068167894	0.7442483
421	west north central div	0	1	0	0	0	0	1	0	18	1.068167894	0.7442483
422	south atlantic division	0	1	0	0	0	0	0	0	283	1.068167894	0.7442483
423	south atlantic division	0	1	0	0	0	0	1	0	34	1.068167894	0.7442483
424	east south central div	0	1	0	0	0	0	0	0	59	1.068167894	0.7442483
425	west south central div	0	1	0	0	0	0	0	0	133	1.068167894	0.7442483
426	west south central div	0	1	0	0	0	0	1	0	12	1.068167894	0.7442483
427	mountain division	0	1	0	0	0	0	0	0	237	1.068167894	0.7442483
428	mountain division	0	1	0	0	0	0	1	0	24	1.068167894	0.7442483
429	pacific division	0	1	0	0	0	0	0	0	2511	1.068167894	0.7442483
430	pacific division	0	1	0	0	0	0	1	0	215	1.068167894	0.7442483
399	middle atlantic division	1	0	1	0	0	1	0	0	15	0.880312565	0.7068870
400	middle atlantic division	1	0	1	0	0	1	1	0	14	0.880312565	0.7068870

Figure 1: The estimate result for Trump and Biden

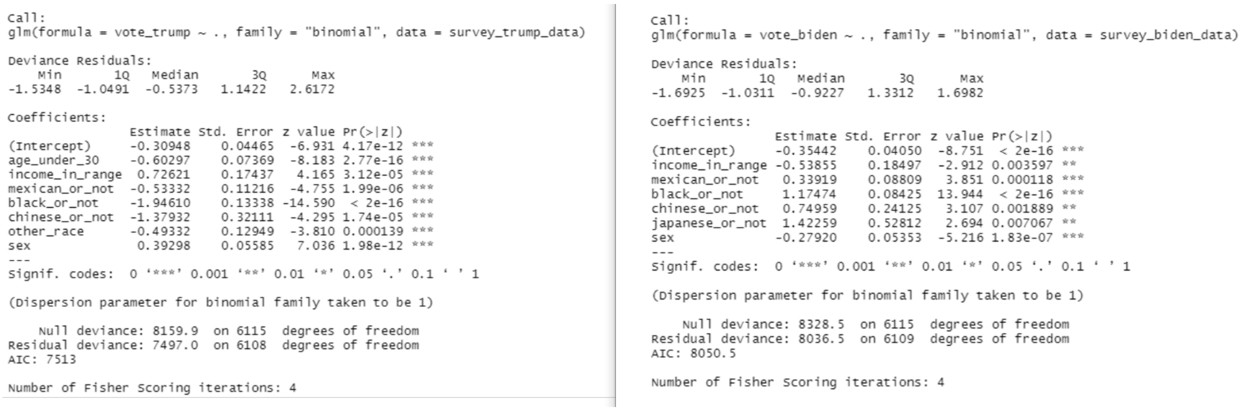
The left result in [Figure 1] is for Trump and the right on is for Biden, then we can find that the the proportion of voters who will vote Trump and the proportion of Biden are very close. And Joe Biden is slightly ahead of Trump.

Discussion

For this analysis, the survey data we used to build model is Nationscape Data Set from Voter Study Group [4], which was released in January 2020 and included around 156,000 cases. And the census dats is from American Community Surveys (ACS) which is 2018 5-year ACS. According to the banner book of the survey, we selected only few variables from 265 variables to build our multiple regression model for Trump and Biden.

But according to the P-value for each variables [Figure 2], we can say that both model are close relative to the target variable. Since the lower the P-value, the higher the relativity to the target variable. And when we select the variables, we also use the *regsubset*() function in R which give us the best subset selection by identifying the best model. So the model we built are somehow relative the true result. The following [Figures 3] show the result of *regsubset*() function for both Trump model and Biden model.

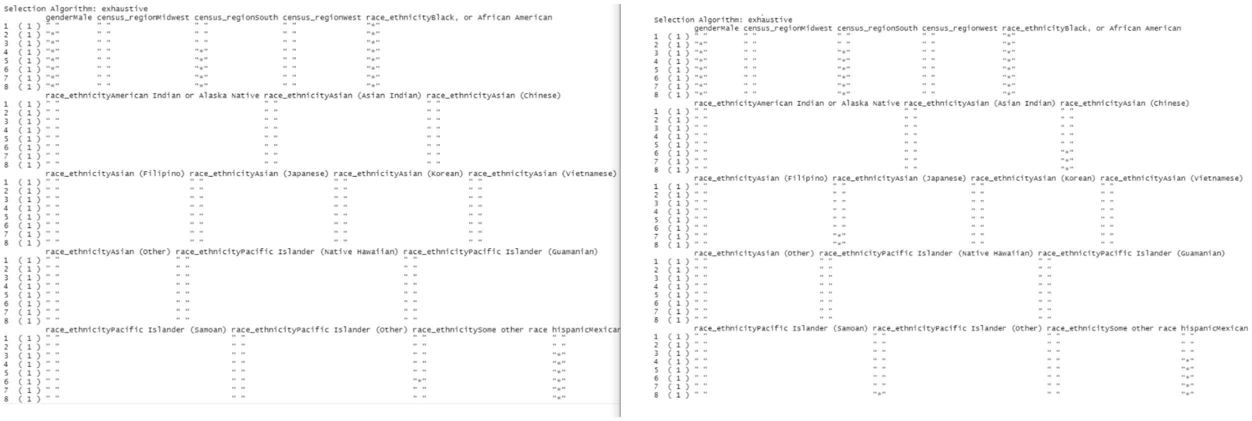
Then we apply post-stratification technique into our multiple logistic model to reduce the bias of our models. And eventually get the result of our analysis. Based on the current model we built and the current data we have, Joe Biden is slightly ahead of Donald Trump in the 2020 American federal election.



```
Call:
glm(formula = vote_trump ~ ., family = "binomial", data = survey_trump_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.5348  -1.0491  -0.5373   1.1422   2.6172  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -0.30948    0.04465  -6.931 4.17e-12 ***
age_under_30    -0.60297    0.07369  -8.183 2.77e-16 ***
income_in_range  0.72621    0.17437   4.165 3.12e-05 ***
mexican_or_not  -0.53332    0.11216  -4.755 1.99e-06 ***
black_or_not    -1.94610    0.13338 -14.590  < 2e-16 ***
chinese_or_not  -1.37932    0.32111  -4.295 1.74e-05 ***
other_race      -0.49332    0.12949  -3.810 0.000139 ***
sex              0.39298    0.05585   7.036 1.98e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8159.9  on 6115  degrees of freedom
Residual deviance: 7497.0  on 6108  degrees of freedom
AIC: 7513

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = vote_biden ~ ., family = "binomial", data = survey_biden_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6925  -1.0311  -0.9227   1.3312   1.6982  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -0.35442    0.04050  -8.751  < 2e-16 ***
income_in_range -0.53855    0.18497  -2.912 0.003597 ** 
mexican_or_not   0.33919    0.08809   3.851 0.000118 ***
black_or_not     1.17474    0.08425  13.944  < 2e-16 ***
chinese_or_not   0.74959    0.24125   3.107 0.001889 ** 
japanese_or_not  1.42259    0.52812   2.694 0.007067 ** 
sex             -0.27920    0.05353  -5.216 1.83e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8328.5  on 6115  degrees of freedom
Residual deviance: 8036.5  on 6109  degrees of freedom
AIC: 8050.5

Number of Fisher Scoring iterations: 4
```

Figure 2: The summary result for Trump and Biden



```
Selection Algorithm: exhaustive
         genderMale census_regionMidwest census_regionSouth census_regionWest race_ethnicityBlack, or African American
1  ( 1 ) " "        " "                  " "                " "               "*"                                   
2  ( 1 ) "*"        " "                  " "                " "               "*"                                   
3  ( 1 ) "*"        " "                  " "                " "               "*"                                   
4  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
5  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
6  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
7  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
8  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
         race_ethnicityAmerican Indian or Alaska Native race_ethnicityAsian (Asian Indian) race_ethnicityAsian (Chinese)
1  ( 1 ) " "                                            " "                                " "                          
2  ( 1 ) " "                                            " "                                " "                          
3  ( 1 ) " "                                            " "                                " "                          
4  ( 1 ) " "                                            " "                                " "                          
5  ( 1 ) " "                                            " "                                " "                          
6  ( 1 ) " "                                            " "                                " "                          
7  ( 1 ) " "                                            " "                                " "                          
8  ( 1 ) " "                                            " "                                " "                          
         race_ethnicityAsian (Filipino) race_ethnicityAsian (Japanese) race_ethnicityAsian (Korean) race_ethnicityAsian (Vietnamese)
1  ( 1 ) " "                            " "                            " "                          " "                             
2  ( 1 ) " "                            " "                            " "                          " "                             
3  ( 1 ) " "                            " "                            " "                          " "                             
4  ( 1 ) " "                            " "                            " "                          " "                             
5  ( 1 ) " "                            " "                            " "                          " "                             
6  ( 1 ) " "                            " "                            " "                          " "                             
7  ( 1 ) " "                            " "                            " "                          " "                             
8  ( 1 ) " "                            " "                            " "                          " "                             
         race_ethnicityAsian (Other) race_ethnicityPacific Islander (Native Hawaiian) race_ethnicityPacific Islander (Guamanian)
1  ( 1 ) " "                         " "                                              " "                                       
2  ( 1 ) " "                         " "                                              " "                                       
3  ( 1 ) " "                         " "                                              " "                                       
4  ( 1 ) " "                         " "                                              " "                                       
5  ( 1 ) " "                         " "                                              " "                                       
6  ( 1 ) " "                         " "                                              " "                                       
7  ( 1 ) " "                         " "                                              " "                                       
8  ( 1 ) " "                         " "                                              " "                                       
         race_ethnicityPacific Islander (Samoan) race_ethnicityPacific Islander (Other) race_ethnicitySome other race hispanicMexican
1  ( 1 ) " "                                     " "                                    " "                           " "            
2  ( 1 ) " "                                     " "                                    " "                           " "            
3  ( 1 ) " "                                     " "                                    " "                           "*"            
4  ( 1 ) " "                                     " "                                    " "                           "*"            
5  ( 1 ) " "                                     " "                                    " "                           "*"            
6  ( 1 ) " "                                     " "                                    "*"                           "*"            
7  ( 1 ) " "                                     " "                                    " "                           "*"            
8  ( 1 ) " "                                     " "                                    " "                           "*"            
```

```
Selection Algorithm: exhaustive
         genderMale census_regionMidwest census_regionSouth census_regionWest race_ethnicityBlack, or African American
1  ( 1 ) " "        " "                  " "                " "               "*"                                   
2  ( 1 ) "*"        " "                  " "                " "               "*"                                   
3  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
4  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
5  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
6  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
7  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
8  ( 1 ) "*"        " "                  "*"                " "               "*"                                   
         race_ethnicityAmerican Indian or Alaska Native race_ethnicityAsian (Asian Indian) race_ethnicityAsian (Chinese)
1  ( 1 ) " "                                            " "                                " "                          
2  ( 1 ) " "                                            " "                                " "                          
3  ( 1 ) " "                                            " "                                " "                          
4  ( 1 ) " "                                            " "                                " "                          
5  ( 1 ) " "                                            " "                                " "                          
6  ( 1 ) " "                                            " "                                "*"                          
7  ( 1 ) " "                                            " "                                "*"                          
8  ( 1 ) " "                                            " "                                "*"                          
         race_ethnicityAsian (Filipino) race_ethnicityAsian (Japanese) race_ethnicityAsian (Korean) race_ethnicityAsian (Vietnamese)
1  ( 1 ) " "                            " "                            " "                          " "                             
2  ( 1 ) " "                            " "                            " "                          " "                             
3  ( 1 ) " "                            " "                            " "                          " "                             
4  ( 1 ) " "                            " "                            " "                          " "                             
5  ( 1 ) " "                            " "                            " "                          " "                             
6  ( 1 ) " "                            " "                            " "                          " "                             
7  ( 1 ) " "                            "*"                            " "                          " "                             
8  ( 1 ) " "                            "*"                            " "                          " "                             
         race_ethnicityAsian (Other) race_ethnicityPacific Islander (Native Hawaiian) race_ethnicityPacific Islander (Guamanian)
1  ( 1 ) " "                         " "                                              " "                                       
2  ( 1 ) " "                         " "                                              " "                                       
3  ( 1 ) " "                         " "                                              " "                                       
4  ( 1 ) " "                         " "                                              " "                                       
5  ( 1 ) " "                         " "                                              " "                                       
6  ( 1 ) " "                         " "                                              " "                                       
7  ( 1 ) " "                         " "                                              " "                                       
8  ( 1 ) " "                         " "                                              " "                                       
         race_ethnicityPacific Islander (Samoan) race_ethnicityPacific Islander (Other) race_ethnicitySome other race hispanicMexican
1  ( 1 ) " "                                     " "                                    " "                           " "            
2  ( 1 ) " "                                     " "                                    " "                           " "            
3  ( 1 ) " "                                     " "                                    " "                           " "            
4  ( 1 ) " "                                     " "                                    " "                           "*"            
5  ( 1 ) " "                                     " "                                    " "                           "*"            
6  ( 1 ) " "                                     " "                                    " "                           "*"            
7  ( 1 ) " "                                     " "                                    " "                           "*"            
8  ( 1 ) " "                                     "*"                                    " "                           "*"            
```

Figure 3: The result of regsubset function

Weaknesses

1. In this analysis, the variable selection was based on the banner book, which limit the possibilities for research. Although it makes our analysis and model easier, but it also makes the bias larger which is not good for research. So if possible, we can try some more variables to low the bias.
2. In this analysis, we use data set from different sources, which makes it hard to manage the data. For example, in our survey, we only have 4 regions. But there are 17 regions in census data, which means they have different rules to decide the region. And these differences can also lead to a big bias.
3. The way we handle the cases who contain NA (missing value) is just remove that case, which is really greedy. Although we have a relative large data set time, those removed cases can also be useful. So we may can find a better way to handle those cases.
4. When I split census data into different cells, there are many cells with only 1 case. And I just removed those cells, which can also lead to a big bias, since those data can be useful. They represent some unique group in our census data. To solve this question, we can try to find a larger data set.

Next Steps

In our analysis, we used the data released in January 2020 to build model. And after the result of the 2020 American federal election come out, we can gather the data again to compare the difference between January and now. And analysis why the change happened. What's more, in January 2020, the situation of COVID-19 was not that bad. Did the development of COVID-19 affect the American federal election? If so, how does it influence the result and why. After we get the result of American federal election, all these topic are worth to analysis.

References

[1] LUCID. (n.d.). If the general election for president of the United States was a contest between Joe Biden and Donald Trump, who would you support? Nationalscape Wave 50, June 25 -July 01, 2020, 86-87.

[2] Regsubsets. (n.d.). R Documentation and manuals | R Documentation. https://www.rdocumentation.org/packages/leaps/versions/2.1-1/topics/regsubsets

[3] Who can and can't vote in U.S. elections. (2020, May 7). Official Guide to Government Information and Services | USAGov. https://www.usa.gov/who-can-vote

[4] New: Second Nationscape data set release. (2020, September 10). Democracy Fund Voter Study Group. https://www.voterstudygroup.org/publication/nationscape-data-set