Title of Your Report

Names of your Group Members

Due Date

Title of your Report

Name(s) of Author(s)

Date

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

From the 2018 dataset and () we have selected age, census_region, employment, and race_ethnicity to model the probability of voting for Donald Trump. Based on the label dictionary obtained along with the data, each variables represent the following:

- *vote_trump* - "A binary column for voting trump or not in the 2020 election. Present 1 if yes, otherwise shows 0"
- age - "Age of respondent in integer at time of interview older than 18 years old"
- census_region - "Respondent's census region, calculated based on entered ZIP code"
- employment - "Respondent's current employment status"
- race_ethnicity - "Respondent's race"

We will use a logistic regression model to model the proportion of voters who will vote for Donald Trump. As mentioned above, we will be using age as a numeric variable, census_region, employment, and race_ethnicuty as a character object to model the probability of voting for Donald Trump. Rstudio is used to run the following logistic regression model. The logistic regression function I am using is shown below:

$$
\begin{aligned}
log(\frac{(p)}{1-(p)}) =& \beta_0 + \beta_1 X_{age} + \beta_2 X_{census\_region:Northeast} + \\
& \beta_3 X_{census\_region:South} + \beta_4 X_{census\_region:West} + \\
& \beta_5 X_{employment:n/a} + \beta_6 X_{employment:not\ in\ labor\ force} + \\
& \beta_7 X_{employment:unemployed} + \beta_8 X_{race\_ethnicity:Asian\ or\ pacific\ islander} + \\
& \beta_9 X_{race\_ethnicity:Black,\ or\ African\ American} + \beta_{10} X_{race\_ethnicity:Someother\ race\ or\ multiraces} + \\
& \beta_{11} X_{race\_ethnicity:White} + \beta_{12} X_{marital\_status:Living\ common-law}
\end{aligned}
$$

Where y represents the proportion of voters who will vote for Donald Trump. Similarly, β_0 represents the intercept of the model, and is the probability of voting for Donald Trump at age 0. Additionally, β_1 represents the slope of the model. So, for everyone one unit increase in age, we expect a β_1 increase in the probability of voting for Donald Trump.

Post-Stratification

I performed a post-stratification analysis to estimate the proportion of voters who will vote for Donald Trump in 2020. I create cells in census_data based on combinations of different ages, races, employment status, and living regions. By using the model described in the previous sub-section, I will get the log estimates and then divided by one plus the exponential of the log estimates which will give me each proportion estimate. Finally, use each proportion estimate times the respective population size of that bin and sum those values and divide that by the entire population size. This will give us the probability for president Donald Trump to win the 2020 election.

We did not include gender because we thought gender was not likely to influence the voter outcome that much.

Since We are looking at the proportion of people voting for Trump, we want to avoid anything that might be a potential lurking variable. Hence, we chose citizenship as one factor from the census_data and then removed people who are "not a citizen" of America because they do not have the right to vote.

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.378
```

Results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.1775685	0.2406411	-4.8934640	0.0000010
age	0.0175288	0.0018603	9.4225055	0.0000000
as.factor(census_region)Northeast	-0.0728052	0.0846179	-0.8604002	0.3895685
as.factor(census_region)South	0.2998697	0.0734102	4.0848528	0.0000441
as.factor(census_region)West	-0.0575314	0.0826211	-0.6963277	0.4862237
as.factor(employment)n/a	-0.3055560	0.2293608	-1.3322070	0.1827922
as.factor(employment)not in labor force	-0.4673032	0.0751731	-6.2163650	0.0000000
as.factor(employment)unemployed	-0.2878783	0.0691670	-4.1620741	0.0000315
as.factor(race_ethnicity)Asian or pacific islander	-0.5764070	0.2574954	-2.2385137	0.0251876
as.factor(race_ethnicity)Black, or African American	-1.8475597	0.2558306	-7.2218084	0.0000000
as.factor(race_ethnicity)Some other race or multiraces	-0.5805640	0.2483863	-2.3373429	0.0194214
as.factor(race_ethnicity)White	0.2492494	0.2243073	1.1111961	0.2664840

Figure 1: Key statistical summaries of the logistic regression model.

Before observing the relationships between our chosen predictor variables with whether to vote for Donald Trump or not, we will have to confirm the existence of the relationship first. Upon close inspection of the table above that summarize the results of our model, we can see that the following variables have p-values greater than 0.05: all categories for census_region except people living in the South region, unknown employment status, and white people. This result tells us that we cannot reject the null hypothesis such that there is no

effect. For each categorical variable, one category is omitted. The slope of every omitted category is 0. And the slope of other categories of the same variable indicate how they affect the response compared with the missing category.

To see whether people are likely or not to vote for Trump, we can check the estimate column. This column provides the coefficients for our logistic regression function specified in the beginning of the model section of this report. Specifically, we can substitute the corresponding variable's β symbol. When filling in the values we can observe that the probability of voting for Trump will decrease with factors that have negative estimates. Therefore we can conclude the following characteristics will decrease Trump's winning possibility: All other races beside white people, people living in Northeast and West region, and all employment status except employed. The rest of the independent variables have positive estimates and as such we can conclude that people who fall in such categories have a likelier outcome of owning a place. For instance, we can see in the plot below that people are more likely to vote for Trump as age increases before a certain age around 68.

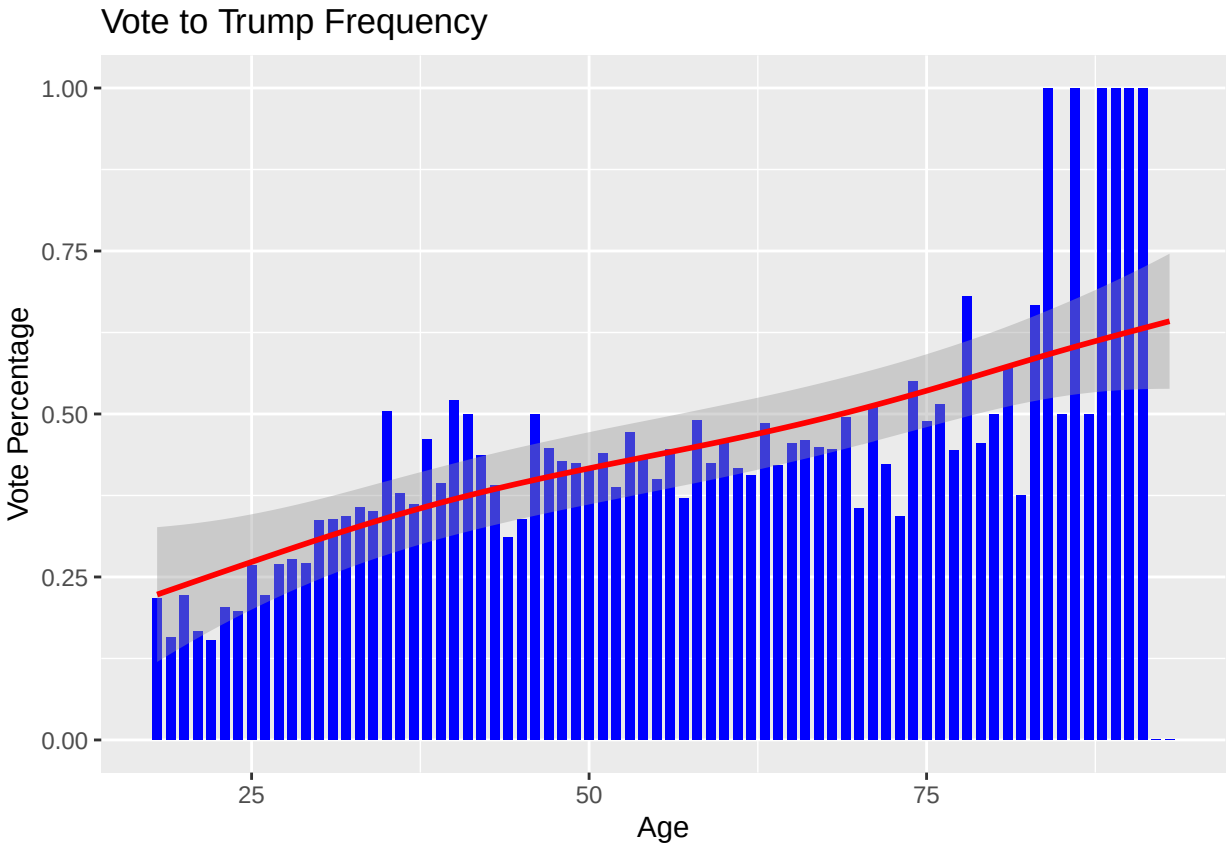

Figure 2: Rate of voting for Trump in different ages.

you will include all results. This includes descriptive statistics, graphs, figures, tables, and model results. Please ensure that everything is well formatted and in a report style. You must also provide an explanation of the results in this section.

Please ensure that everything is well labelled. So if you have multiple histograms and plots, calling them Figure 1, 2, 3, etc. and referencing them as Figure 1, Figure 2, etc. in your report will be expected. The reader should not get lost in a sea of information. Make sure to have the results be clean, well formatted and digestible.

Discussion

Here you will summarize the previous sections and discuss conclusions drawn from the results. Make sure to elaborate and connect your analysis to the goal of the study.

Weaknesses

Since the data we chose are from 2018, data may be outdated and results may vary. Due to the large outbreak of COVID-19 in 2020, people's voting intentions might change according to Donald Trump's government's response to COVID-19 pandemic and policy changes.

In both census_data and survey_data, rows that have "Asian or pacific islander" in the race_ethnicity column may affect the proportion outcome. Since, not all parts of Asian and pacific islander can be considered as one group.

Next Steps

As mentioned above, there's the possibility of lurking variables that we have yet to observe. Therefore a similar analysis will need to be conducted on various other factors in the American Community Survey 2018 Sample.

Since the election is scheduled on Nov/ 3rd, 2020, we will search on the internet afterwards to compare the support rate of each candidate and compare it with other analysis. An online survey will be created related to people's voting intentions in 2018 and who they vote for in 2020.

References

1. AMERICAN COMMUNITY SURVEY 2018 SAMPLE. (n.d.). Retrieved from https://usa.ipums.org/usa-action/samples
2. Tidyr. (n.d.). Retrieved from https://www.rdocumentation.org/packages/tidyr/versions/0.8.3/topics/drop_na
3. kableExtra. (n.d.) Retrieved from https://www.rdocumentation.org/packages/kableExtra/versions/1.2.1