Predicting the Popular Vote of the 2020 American Federal Election

Pamela De Vera, Hyoeun Park

November 2, 2020

I. Model

I-i. Model Specifics

In this report, we are interested in forecasting whether Donald Trump or Joe Biden will win the popular vote in the 2020 American Federal Election. In order to do so, we are going to implement the post-stratification technique with two multilevel logistic regression models. Details of the model and the technique will be further discussed in the following subsections.

Model Selection and Variables

We will carry out two multilevel logistic regression models using the `lme4` package in `R` (programming language). The models will be *multilevel* because we want to model the survey data into two different levels: individual levels, defined as characteristics of individuals, and group levels, which groups individuals that share similar traits. With group levels in mind, we would then assume that individuals within the same group are not independently and identically distributed in the sample data set. Using such a type of model would yield more accurate inferences since it considers the hierarchical structure (Rasbash, n.d.). Additionally, the *logistic* model allows us to predict the probability of an event happening which will be denotes as 1 and its complement as 0. In this case, the event in question is voting for a specific candidate.

Since we know most Americans, in general, are supporting either Donald Trump or Joe Biden (Pew Research Center, 2020), we expect that either of the two would win the election. Hence, we are going to employ two different models, one for Trump and one for Biden, to forecast what voting intentions individuals have in the 2020 election. Conducting only one model will only allow us to see the probability of voting for one candidate and we note that the complement of this is not the probability of voting for the other candidate, since there are more than 2 candidates in the running.

For our models, we will be using the Nationscape Survey data set provided by the Democracy Fund Voter Study Group. This survey data set includes multiple questionnaires including information about demographics and voting intentions, which are practical to predict the popular vote outcome of the election. There will be particular demographic variables chosen as predictors and voting intentions for the 2020 election will be the response variable.

Specifically, we will be examining how voting intention for the 2020 election can be predicted from age, sex, household income, education, ethnicity/race, being a Hispanic[1], and region. We have chosen such demographic features since they are known to determine your political view (Knoke & Haut, 1974; Prato et al., 1997; Weakliem, 2002). Especially, we believe someone being Hispanic or not would have an impact on their voting outcome, due to Donald Trump's policy of constructing a wall in the southern U.S. border and his disparaging speech towards Mexico (BBC News, 2016). Furthermore, we will have the regions functioning as a variable in the group level while the rest are considered to be individual levels. We can indicate the regions as the group level since the area one lives in will have an impact on their political views, and thus the

[1] whether an individual is Hispanic or not

voting outcome. For instance, the election results in 2016 revealed that Hillary Clinton earned the majority of her votes in the Northeastern and far Western part of the U.S. (Kiersz, 2016), meaning the region potentially affects the vote.

Moreover, most of the variables were split into categories using the `tidyverse` package in `R`[2]. That is, all variables are now categorical variables.

Equation of the Model

Using such variables, the equation of the model which predicts the proportion of Americans intending to vote for Joe Biden is the following:

$$
\begin{aligned}
y_{ij}^B = \beta_{0j}^B &+ \beta_{\text{age26-35}}^B x_{\text{age26-35},ij} + \beta_{\text{age36-45}}^B x_{\text{age36-45},ij} + \beta_{\text{age46-55}}^B x_{\text{age46-55},ij} \\
&+ \beta_{\text{age56-65}}^B x_{\text{age56-65},ij} + \beta_{\text{age66-75}}^B x_{\text{age66-75},ij} + \beta_{\text{age76-85}}^B x_{\text{age76-85},ij} \\
&+ \beta_{\text{age86-95}}^B x_{\text{age86-95},ij} + \beta_{\text{male}}^B x_{\text{male},ij} + \beta_{\text{black}}^B x_{\text{black},ij} + \beta_{\text{indig}}^B x_{\text{indig},ij} \\
&+ \beta_{\text{other}}^B x_{\text{other},ij} + \beta_{\text{white}}^B x_{\text{white},ij} + \beta_{\text{hispan}}^B x_{\text{hispan},ij} \\
&+ \beta_{\text{M/D}}^B x_{\text{M/D},ij} + \beta_{\text{P/S}}^B x_{\text{P/S},ij} + \beta_{\text{low}}^B x_{\text{low},ij} + \beta_{\text{mid}}^B x_{\text{mid},ij} + r_{ij}^B
\end{aligned}
$$

Here, $y_{ij}^B = \log\left(\frac{p^B}{1-p^B}\right)$ represents the log odds of intending to vote for Joe Biden based on specific demographic factors for individual i in the j^{th} group. Therefore, we need to compute the following in order to find the actual probability p^B of an individual supporting Biden[3]:

$$
p^B = \frac{\exp\left(y_{ij}^B\right)}{1 + \exp\left(y_{ij}^B\right)}
$$

In addition, we have 17 different individual-level predictors, denoted by $x_{\text{k},ij}$, where k represents age groups (26-35, 36-45, 46-55, 56-65, 66-75, 76-85, 86-95), sex (male), races (Black, White, Indigenous, other), an individual identified as Hispanic, education level (Masters/Doctorate (M/D), Primary/Secondary (P/S)), and class of income (low, middle (mid)). [4] Then, the 17 β_l^B's are the slopes for each predictor. That is, the average increase or decrease of log-odds of one unit increase in the $x_{\text{k},ij}$'s will depend on the β_l^B values.

We also have β_{0j}^B as the random intercept value, which can be written as:

$$
\beta_{0j}^B = r_{00}^B + r_{0\text{NE}}^B W_{\text{NE}j} + r_{0\text{MW}}^B W_{\text{MW}j} + r_{0\text{S}}^B W_{\text{S}j} + r_{0\text{W}}^B W_{\text{W}j} + u_{0j}^B
$$

where $W_{\text{m}j}$ ($m \in \{\text{NE, MW, S, W}\}$) symbolizes the predictor variable for each region: Northeast (NE), Midwest (MW), South (S), and West (W), and $r_{0\text{n}}^B$ are the slopes of each $W_{\text{m}j}$. In other words, the value of β_{0j}^B depends on which of the four regions individual i is in. Moreover, β_{0j}^B is the log odds of voting for Biden given all $x_{\text{k},ij}$'s and r_{ij}^B equal to 0. Similarly, r_{00}^B, namely, the intercept of intercept β_{0j}^B, is the value of β_{0j}^B given that all $W_{\text{m}j}$'s and u_{0j}^B equal 0.

Last, we say r_{ij}^B and u_{0j}^B are the random error terms[5] of the individual levels and group level, respectively.

Similarly, we can apply the same logic for the model predicting the proportion of Americans intending to vote for Donald Trump, denoted as the following:

[2] Age: 8 categories; Sex: 2 categories (same as before); Household Income: 3 categories; Education: 3 categories; Race: 5 categories; Hispanic: 2 categories; Voting intentions for the 2020 election: 2 categories; Region: 4 categories

[3] Note that $0 \leq p^B \leq 1$, where $p^B = 1$ means it is certain that an individual will plan to vote for Biden, while $p^B = 0$ means it is impossible for an individual to do so.

[4] Note that we simply set $x_{\text{male}ij} = 0$ if we want to find the voting intention of a female; and the same logic applies to other predictors.

[5] Error of measurement

$$
\begin{aligned}
y_{ij}^T &= \log\left(\frac{p^T}{1-p^T}\right) \\
&= \beta_{0j}^T + \beta_{\text{age26-35}}^T x_{\text{age26-35},ij}^T + \ldots + \beta_{\text{mid}}^T x_{\text{mid},ij}^T + r_{ij}^T \\
\text{where } \beta_{0j}^T &= r_{00}^T + r_{0\text{NE}}^T W_{\text{NE}j} + r_{0\text{MW}}^T W_{\text{MW}j} + r_{0\text{S}}^T W_{\text{S}j} + r_{0\text{W}}^T W_{\text{W}j} + u_{0j}^T
\end{aligned}
$$

Note that the two models above are going to be trained on a survey data set, which does not embrace the actual population of interest (American citizens eligible for voting in the 2020 American Federal Election). Hence, we are going to apply these trained models to a census data with the post-stratification technique, so that it gives a better representation of the population.

Diagnostics: Binned Residual Plot

Using the `arm` package, we are going to use binned residual plots to grasp an idea of how well our models fit the survey data (Webb, 2017). Essentially, the plots will tell us how close our estimates are to the actual values observed in the survey data. We can verify that our models have reasonable fits when 95% or more of the points fall within the standard error bound[6], which is a measure of how much our data is spread out.

I-ii. Post Stratification

In our analysis, we will be using the post stratification technique to find estimates for the proportion of people voting for Joe Biden and Donald Trump in the 2020 election. This technique weighs each possible log probability, as calculated by our model, in a way that the proportions of each match the proportions of each cell in the census data. This will give us a more accurate representation of the whole population since our sample was collected from volunteers, which does not capture the true proportion of the demographics within the population.

For this technique, each cell is divided by sex, age group, income class, education level, race, if an individual is Hispanic or not, and the region they live in. The cells were split up in this way so that only one probability value can be calculated for each cell since each makes up one possible combination of our predictor variables. We also chose the variables as such since each holds some pressure against the party an individual identifies with (Knoke & Haut, 1974; Prato et al., 1997; Weakliem, 2002).

With our cells defined we can calculate the estimated proportion of people voting for Trump and Biden denoted as $\hat{y}_T^{PS}$ and $\hat{y}_B^{PS}$, respectively, using the following formula:

$$\hat{y}_k^{PS} = \frac{\sum N_{ij}\hat{y}_{ij}}{\sum N_{ij}} \text{ where } k = B \text{ or } k = T$$

The formula takes the probability that a single individual in a cell will be voting for either Trump or Biden and scales it to match the proportion that cell has relative to the census population. It then adds all of the weighted probabilities together to get our estimate. We will repeat this process twice; once for the model estimating the proportion of Trump voters and another for the model estimating the proportion of Biden voters.

II. Results

II-i. Regression Output

According to the estimates given by the summary output displayed in Table 1, we can write the equation predicting the proportion of people intending to vote for Joe Biden as the following:

[6] The space between the two lines symmetrical about the y-axis

Table 1: Model Summary

	Joe Biden		Donald Trump	
Coefficients	**Estimates**	**P-values**	**Estimates**	**P-values**
Intercept: Region				
Intercept	0.9681	4.24x10^(-7)	-1.1977	3.13x10^(-9)
Northwest	0.1114	-	-0.1272	-
Midwest	0.0564	-	-0.0492	-
South	-0.2238	-	0.2672	-
West	0.0562	-	-0.0906	-
Age Group				
26-35	-0.3371	0.0030	0.4406	0.0002
36-45	-0.3931	0.0005	0.5927	5.52^10x(-7)
46-55	-0.6572	4.70x10^(-8)	0.7820	3.80x10^(-10)
56-65	-0.3418	0.0034	0.5896	1.16x10^(-6)
66-75	-0.2956	0.0172	0.5363	3.06x10^(-5)
76-85	-0.5506	0.0072	0.7913	0.0001
86-95	-0.7272	0.3120	1.0882	0.1304
Sex				
Male	-0.4008	3.15x10^(-11)	0.3749	8.77x10^(-10)
Race				
Black	1.0339	4.90x10^(-9)	-1.3722	2.24x10^(-12)
Indigenous	-1.0264	0.0004	0.8021	0.0059
Other	-0.3567	0.0589	0.2186	0.2668
White	-0.7664	6.43x10^(-8)	0.7680	1.65x10^(-7)
Hispanic				
Hispanic	0.3433	0.0004	-0.3454	0.0005
Education Level				
Master/Doctorate	0.0235	0.8034	-0.0150	0.8746
Primary/Secondary	-0.3217	1.85x10^(-6)	0.2885	2.46x10^(-5)
Income Class				
Low Income	0.3341	0.0006	-0.4685	1.68x10^(-6)
Middle Income	0.2344	0.0100	-0.3252	0.0004

$$
\begin{aligned}
\hat{y}_{ij}^{B} = \hat{\beta}_{0j}^{B} & - 0.3371x_{\text{age26-35},ij} - 0.3931x_{\text{age36-45},ij} - 0.6572x_{\text{age46-55},ij} \\
& - 0.3418x_{\text{age56-65},ij} - 0.2956x_{\text{age66-75},ij} - 0.5506x_{\text{age76-85},ij} \\
& - 0.7272x_{\text{age86-95},ij} - 0.4008x_{\text{male},ij} + 1.0339x_{\text{black},ij} - 1.0264x_{\text{indig},ij} \\
& - 0.3567x_{\text{other},ij} - 0.7664x_{\text{white},ij} + 0.3433x_{\text{hispan},ij} \\
& + 0.0235x_{\text{M/D},ij} - 0.3217x_{\text{P/S},ij} + 0.3341x_{\text{low},ij} + 0.2344x_{\text{mid},ij} \\
\text{where } \hat{\beta}_{0j}^{B} & = 0.9681 + 0.1114W_{\text{NE}j} + 0.0564W_{\text{MW}j} - 0.2238W_{\text{S}j} + 0.0562W_{\text{W}j}
\end{aligned}
$$

Likewise, we can arrive at the equation predicting the proportion of those intending to vote for Donald Trump:

$$
\begin{aligned}
\hat{y}_{ij}^{T} = \hat{\beta}_{0j}^{T} & + 0.4406x_{\text{age26-35},ij} + 0.5927x_{\text{age36-45},ij} + 0.7820x_{\text{age46-55},ij} \\
& + 0.5896x_{\text{age56-65},ij} + 0.5363x_{\text{age66-75},ij} + 0.7913x_{\text{age76-85},ij} \\
& + 1.0882x_{\text{age86-95},ij} + 0.3749x_{\text{male},ij} - 1.3722x_{\text{black},ij} + 0.8021x_{\text{indig},ij} \\
& + 0.2186x_{\text{other},ij} + 0.7680x_{\text{white},ij} - 0.3454x_{\text{hispan},ij} \\
& - 0.0150x_{\text{M/D},ij} + 0.2885x_{\text{P/S},ij} - 0.4685x_{\text{low},ij} - 0.3252x_{\text{mid},ij} \\
\text{where } \hat{\beta}_{0j}^{T} & = -1.1977 - 0.1272W_{\text{NE}j} - 0.0492W_{\text{MW}j} + 0.2672W_{\text{S}j} - 0.0906W_{\text{W}j}
\end{aligned}
$$

A positive slope estimate implies someone with that demographic feature would likely to support Biden or Trump, and a negative slope estimate means the likelihood of it would decrease. From Table 1, we observe that all signs of the estimates in Biden's model are the opposite of that of the corresponding predictors in the other model. For instance, the slope estimates for males in Biden's model is negative (-0.4008), whereas it is positive (0.3749) for Trump's model. This means someone who is male is more likely to support Donald Trump than Joe Biden.

Next, we are going to use $\alpha = 0.05$ as the benchmark significance level when assessing the p-values. That is, a p-value lower than $\alpha = 0.05$ indicates the significance of the estimate values. Looking at Table 1, we notice that the p-values for both models have the same significance value for each of their corresponding predictors. For instance, the p-values of low income class in Biden's model and Trump's model are 0.0006 and $1.68 \cdot 10^{-6}$, respectively, and they are both lower than the benchmark significance level α. Furthermore, we also notice that not all p-values say that their corresponding estimates of the coefficients are significant. For example, p-values for the masters/doctorate, age from 86-95, and other race's slope estimates of both models are greater than α, providing evidence that the estimates don't hold significance. Some elaboration on the significance of the slope estimates will be discussed in **III-ii. Weaknesses**.

II-ii. Diagnosing the Model Fit: Binned Residual Plot

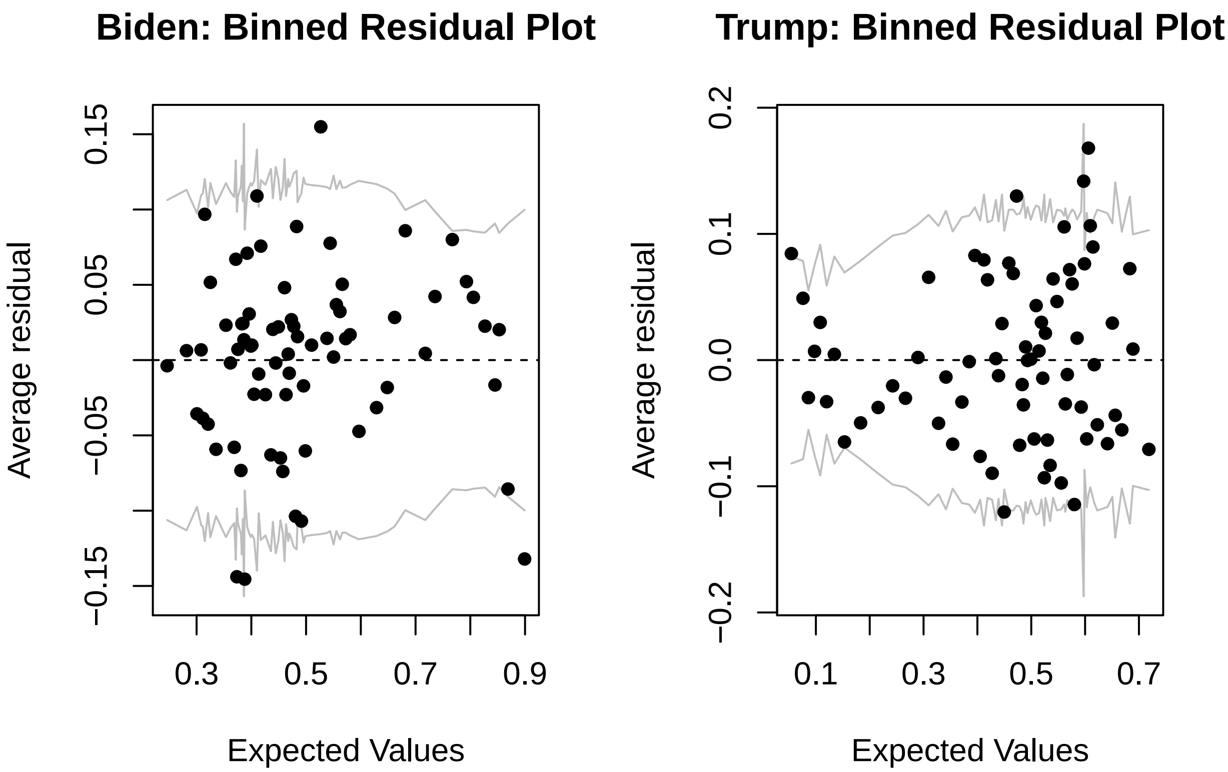

Figure 1

Figure 2

Figure 1 and Figure 2 above are two binned residual plots assessing the fit of each model. We can clearly discern that both models are good fits for the survey data set, since both plots show that more than 95% of the points fall within the error bounds.

II-iii. Post Stratification Values

	Donald Trump (k=T)	Joe Biden (k=B)
$\hat{y}_k^{PS}$	0.4918956	0.4680149

Table 2: Post Stratification Values for Trump and Biden Models

The post stratification values of our logistic models predicting the probability of voting for Trump and Biden are 0.4922043 and 0.4681613 respectively. From this calculation, we can predict that Donald Trump will have the support of 49.2% of voters, while Joe Biden will have the support of 46.8% of the voting population. The remaining voter population are individuals who are predicted to be voting for neither.

III. Discussion

III-i. Conclusion

Based on the proportions of voters voting for Trump and Biden, as estimated from the logistic model, we predict that the popular vote of the 2020 Federal Election will be Donald Trump. Although he does not hold the majority of votes from the voter population at 49.2%, Trump still holds the highest proportion when compared to the proportion of voters in support of Joe Biden at 46.8% and to the proportion voting for another candidate at 4%.

III-ii. Weaknesses

From Figures 1 and 2, we note that these models are good fits given our sample data. However, there are some setbacks when extrapolating our models to the census data. We must first note that the cleaned census data only accounts for 1.6 million Americans out of the voting age population (VAP) of 253,768,092 (Federal Register, 2019). So, it is possible that the proportions of each cell as defined in the **I. Model** section may not be the actual proportion of the entire VAP. Even so, IPUMS holds the largest collection of public census data (IPUMS USA, n.d.), so it is not feasible to perform post-stratification using a larger population.

Furthermore, we must also consider the years our census and survey data was collected. The survey data was collected in 2020 while the census data was collected in 2018. This means that the population used to calculate the post-stratification value may have slight differences than the population in question, which is the 2020 VAP.

Looking into the significance of each predictor variable, we notice in our model predicting the proportion of Trump and Biden voters, all slope estimates hold some significance with the exception of the education level,$\hat{\beta}_{\text{M/D}ij}$, the race, $\hat{\beta}_{\text{other}ij}$, and the age group, $\hat{\beta}_{\text{age86-95}ij}$. Although having a Master's degree or Doctorate does not appear to be statistically significant, we include it in our model in general, the higher education level an individual has, the more likely their views are liberal (Pew Research Centre, 2016). Furthermore, the category, $\hat{\beta}_{\text{other}ij}$, is ambiguous as it contains individuals who identify as 2 or more races. Lastly, we notice a large p-value in the age group 86-95 possibly due having only 9 observations in our survey data within that age group. We still consider it in both models since the other categories of education level and race are significant.

In terms of the winner of the presidential election, our model prevents us from predicting who will win as the winner of the election is based on winning a majority of votes from the electoral college, instead of having the popular vote.

III-iii. Next Steps

There are a few improvements and/or expansions we can make based on the weaknesses of our analysis.

First, the candidate winning the 2020 United States Federal Election can help us determine whether our predictions were accurate, and see if we can make any improvements on our algorithm. For example, we can add more predictors (e.g., religion) or eliminate predictors that have larger p-values than $\alpha = 0.05$.

On top of that, we can also be conducting a follow-up voluntary survey briefly asking whether the Americans have changed their voting intentions within the last few months before the election date. Note that the survey data was conducted from June 25 - July 1, 2020, so there is a possibility that one's voting intention could have changed within the 4 months before November 3, 2020. If our survey reveals that a substantial amount of Americans have decided to change their decisions, we should make a note that we have to use the latest survey data that is accessible, in order to yield more accurate predictions.

Lastly, as mentioned in our weaknesses, our method majorly predicting the outcome of the popular vote instead of the electoral vote might not accurately predict the winner of the 2020 American Federal Election. Hence, we can make alterations on our model so that we can predict the electoral vote of the election instead. That is, we would need to find data including some information about the electoral colleges of the United States.

III-iv. Summary

There is an ongoing debate about which of the two popular candidates, Donald Trump or Joe Biden, will win the 2020 United States Presidential Election. Likewise, our main objective of this paper was to predict the popular vote outcome of the American election, scheduled on November 3, 2020. We used the Nationscape Survey data set from *Democracy Fund Voter Study Group* to perform two multilevel logistic regression models, each predicting one's support for Joe Biden and Donald Trump based on specific demographic features including age, sex, income, education, being Hispanic, race, and region. Then, the models were applied to the U.S. Census data provided by *IPUMS* in order to post-stratify the predicted overall popular vote of both candidates. The post-stratification results suggested that Donald Trump would receive more votes from each individual American compared to Joe Biden. Furthermore, it revealed that a small proportion of Americans would vote for neither.

IV. References

Austin, P.C., & Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. *Statistics in Medicine*, 36(20), 3257–3277. https://dx.doi.org/10.1002%2Fsim.7336

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

BBC News. (2016, August 31). *'Drug dealers, criminals, rapists': What Trump thinks of Mexicans.* https://www.bbc.com/news/av/world-us-canada-37230916

Gelman, A., et al. (2020). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models.* https://cran.r-project.org/package=arm

Kiersz, A. (2016, November 28). *Here's the final 2016 Electoral College map.* Business Insider. https://www.businessinsider.com/final-electoral-college-map-trump-clinton-2016-11

Knoke, D., & Hout, M. (1974). Social and Demographic Factors in American Political Party Affiliations, 1952-72. *American Sociological Review*, 39(5), 700-713. https://doi.org/10.2307/2094315

Kochhar, K. (2018, September 6). *The American middle class is stable in size, but losing ground financially to upper-income families.* Pew Research Centre: Fact Tank. https://www.pewresearch.org/fact-tank/2018/09/06/the-american-middle-class-is-stable-in-size-but-losing-ground-financially-to-upper-income-families/#:~:text=In%202016%2C%20the%20national%20middle,(incomes%20in%202016%20dollars)

Lynden Educational Services. (n.d.). *U.S. System of Education.* https://www.lindentours.com/united-states-system-of-education/

Pew Research Centre. (2016, April 26). *A Wider Ideological Gap Between More and Less Educated Adults.* https://www.pewresearch.org/politics/2016/04/26/a-wider-ideological-gap-between-more-and-less-educated-adults/

Pew Research Centre. (2020, October 9). *Amid Campaign Turmoil, Biden Holds Wide Leads on Coronavirus,Unifying the Country.*https://www.pewresearch.org/politics/2020/10/09/the-trump-biden-presidential-contest/

Pratto, F., Stallworth, L., & Sidanius, J. (1997). The gender gap: Differences in political attitudes and social dominance orientation. *British Journal of Social Psychology*, 36(1), 49-68. https://doi.org/10.1111/j.2044-8309.1997.tb01118.x

Rasbash, J. (n.d.). *What are multilevel models and why should I use them?.* University of Bristol: Centre for Multilevel Modelling. http://www.bristol.ac.uk/cmm/learning/multilevel-models/what-why.html#:~:text=Multilevel%20models%20recognise%20the%20existence,the%20child%20and%20school%20level

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. (2020). *IPUMS USA* (Version 10.0) [dataset]. Minneapolis,MN: IPUMS. https://doi.org/10.18128/D010.V10.0

Tausanovitch, C., & Vavreck, L. (2020). *Democracy Fund + UCLA Nationscape, October 10- 17, 2019* (Version 20200625) [dataset]. https://www.voterstudygroup.org/publication/nationscape-data-set

United States Census Bureau. (2019, June 20). *2018 Population Estimates by Age, Sex, Race and Hispanic Origin*. https://www.census.gov/newsroom/press-kits/2019/detailed-estimates.html

Webb, J. (2017, September 03). *Course Notes for IS 6489, Statistics and Predictive Analytics.* https://bookdown.org/jefftemplewebb/IS-6489/logistic-regression.html

Wickham, H., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H. & Miller, E. (2020). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files.* R package version 2.3.1. https://CRAN.R-project.org/package=haven

Xie, Y. (2020). *knitr: A General-Purpose Package for Dynamic Report Generation in R.* R package version 1.30. https://CRAN.R-project.org/package=knitr

Zhu, H. (2020, October 22). *Create Awesome HTML Table with knitr::kable and kableExtra.* The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html

Zhu, H. (2020). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax.* R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra

V. Appendix

V-i. Github Respository

Here is the link to our Github Repository: https://github.com/hynprk/Predicting-2020-American-Federal-Election